

Landscape of Policy Optimization for Finite Horizon MDPs with General State and Action

Xin Chen

ISyE, Georgia Institute of Technology, xin.chen@isye.gatech.edu

Yifan Hu

College of Management of Technology, EPFL, yifan.hu@epfl.ch

Minda Zhao

ISyE, Georgia Institute of Technology, mindazhao@gatech.edu

Policy gradient methods are widely used in reinforcement learning. Yet, the nonconvexity of policy optimization imposes significant challenges in understanding the global convergence of policy gradient methods. For a class of finite-horizon Markov Decision Processes (MDPs) with general state and action spaces, we develop a framework that provides a set of easily verifiable assumptions to ensure the Kurdyka-Łojasiewicz (KL) condition of the policy optimization. Leveraging the KL condition, policy gradient methods converge to the globally optimal policy with a non-asymptomatic rate despite nonconvexity. Our results find applications in various control and operations models, including entropy-regularized tabular MDPs, Linear Quadratic Regulator (LQR) problems, stochastic inventory models, and stochastic cash balance problems, for which we show an ϵ -optimal policy can be obtained using a sample size in $\tilde{O}(\epsilon^{-1})$ and polynomial in terms of the planning horizon by stochastic policy gradient methods. Our result establishes the first sample complexity for multi-period inventory systems with Markov-modulated demands and stochastic cash balance problems in the literature.

Key words: finite-horizon Markov Decision Processes (MDPs), Kurdyka-Łojasiewicz (KL) condition, policy gradient methods, inventory, cash balance, data-driven operations models

1. Introduction

Reinforcement Learning (RL) has achieved remarkable success in various real-world applications, including the game of Go ([Silver et al. 2016](#)) and robotics ([Hwangbo et al. 2019](#)). An important class of algorithms for solving these RL problems is policy gradient methods, which search over a parameterized policy space by applying (stochastic) gradient methods on the total expected cost of a Markov Decision Process (MDP). Despite wide applicability, our understanding of the global convergence and non-asymptotic convergence behavior of policy gradient methods remains limited due to the nonconvex landscape of the policy gradient optimization problem ([Agarwal et al. 2021](#), [Bhandari and Russo 2024](#)).

This paper seeks to understand the nonconvex landscape of the policy gradient optimization problem and establish non-asymptotic convergence rates for policy gradient methods in solving a class of finite-horizon MDPs with general state and action spaces. Specifically, we aim to identify structural properties shared by a host of applications such that policy gradient optimization problems would satisfy the Kurdyka-Łojasiewicz (KL) condition (Kurdyka 1998, Łojasiewicz 1963). Informally, the KL condition states that the norm of the gradient dominates the suboptimality gap. It is a relaxation of the strong convexity while maintaining a key property that any point satisfying the first-order necessary optimality condition (Nocedal and Wright 1999) is globally optimal. Policy gradient methods are designed to find these points and thus converge globally on nonconvex MDP problems.

For this purpose, we introduce a framework with several easily verifiable assumptions to establish the KL condition of the policy gradient optimization problem in finite-horizon MDPs with general state and action spaces. More specifically, we demonstrate that the policy gradient optimization problem satisfies the KL condition when (i) the objective function has bounded gradients, (ii) expected optimal Q-value functions satisfy the KL condition, and (iii) sequential decomposition inequalities hold. Roughly speaking, for any given period t and a policy π , sequential decomposition inequalities bound the difference between policy gradients at one policy using π for the period 1 to period t and switching to optimal policy parameters after period t and a neighboring policy using π for the period 1 to period $t - 1$ and switching to optimal policy parameters from period t onwards, by the suboptimality gap of the expected optimal Q-value functions of the policy π .

Leveraging the KL condition, we demonstrate that any point satisfying the first-order necessary optimality condition of the policy gradient optimization problem is globally optimal. Moreover, we show that exact policy gradient methods exhibit a linear convergence rate and stochastic policy gradient methods achieve an $\tilde{\mathcal{O}}(\epsilon^{-1})$ sample complexity to find an ϵ -optimal policy. We remark that the KL constant admits a polynomial dependence on the time horizon, and correspondingly, the non-asymptotic convergence rate of policy gradient methods scales polynomially with the time horizon.

Our framework applies to a variety of control and operations models, including (i) entropy-regularized tabular MDPs with the class of all stochastic policies, (ii) Linear Quadratic Regulator (LQR) problems with the class of affine policies, (iii) multi-period inventory systems with Markov-modulated demands that use state-dependent base-stock policies, and (iv) stochastic cash balance problems with the class of two-sided base-stock policies. Notably, (i)-(iii) are commonly seen in dynamic programming textbooks (Bertsekas 1995, Puterman 2014).

For entropy-regularized tabular MDPs and LQR problems, we establish a linear convergence rate for exact policy gradient methods to achieve an ϵ -optimal policy, consistent with existing results in

the literature (Bhandari and Russo 2024, Hambly et al. 2021). In the case of multi-period inventory systems with Markov-modulated demands and stochastic cash balance problems, we demonstrate an $\tilde{\mathcal{O}}(\epsilon^{-1})$ sample complexity for stochastic policy gradient methods to achieve an ϵ -optimal policy, which gives the first sample complexity results in the literature. It is worth noting that all these sample complexities exhibit a polynomial dependence on the time horizon.

1.1. Related Literature

This study intersects with three streams of literature: (i) nonconvex landscape conditions that ensure global convergence for algorithms, (ii) global optimality guarantees for policy gradient methods, and (iii) data-driven operations management.

Nonconvex Landscape Conditions In nonconvex optimization, several landscape conditions guarantee convergence to global optimality for algorithms, including the hidden convexity (Stern and Wolkowicz 1995, Ben-Tal and Teboulle 1996), i.e., the problem admits a convex reformulation, the Polyak-Łojasiewicz (PL) condition (Polyak et al. 1963, Łojasiewicz 1963), the Kurdyka-Łojasiewicz (KL) condition (Kurdyka 1998, Bolte et al. 2007), and others as summarized in Karimi et al. (2016).

Hidden convexity has emerged in numerous modern applications, such as policy optimization in convex RL (Zhang et al. 2020, Sun and Fazel 2021), supply chain and revenue management (Feng and Shanthikumar 2018, Chen et al. 2018, Chen and Gao 2019, Miao and Wang 2021, Chen and Shi 2023). In the optimization society, Stern and Wolkowicz (1995) and Ben-Tal and Teboulle (1996) studied hidden convexity in quadratic programming. Since then, several works have developed tools to identify hidden convexity (Ben-Tal et al. 2011, Ben-Tal and Den Hertog 2014) and studied the landscape of hidden convex functions (Levin et al. 2024). Subsequently, research focuses on algorithm design and non-asymptotic convergence guarantees under hidden convexity. For instance, Chen et al. (2024) addressed a specific problem arising from network revenue management. Building on a convex reformulation presented by Feng and Shanthikumar (2018), Chen et al. (2024) developed Mirror Stochastic Gradient Descent which achieves global convergence with a sample complexity of order $\tilde{\mathcal{O}}(\epsilon^{-2})$. Moreover, Fatkhullin et al. (2023b) extended the results to general-purpose stochastic optimization under hidden convexity, demonstrating an $\mathcal{O}(\epsilon^{-3})$ sample complexity for project Stochastic Gradient Descent (SGD).

Recently, several works have investigated the Polyak-Łojasiewicz (PL) condition, first introduced by Polyak et al. (1963) and Łojasiewicz (1963). The PL condition ensures a linear convergence rate for Gradient Descent (Karimi et al. 2016) and an $\mathcal{O}(\epsilon^{-1})$ sample complexity for SGD (Hu et al.

2024) to achieve an ϵ -optimal solution. Additionally, Karimi et al. (2016) and Liao et al. (2024) explored the connection between the PL condition and other nonconvex landscape conditions, such as the error bounds (Luo and Tseng 1993), the quadratic growth (Anitescu 2000), the weak strong convexity (Necoara et al. 2019), and the proximal PL condition (gradient dominance). These conditions ensure that any first-order stationary point is globally optimal, leading to a global convergence of first-order methods despite nonconvexity.

Our study builds on the Kurdyka-Łojasiewicz (KL) condition (Kurdyka 1998, Bolte et al. 2007), which extends the PL condition to handle more general settings, particularly for non-smooth functions and constrained optimization problems. Assuming the optimization problem satisfies the KL condition, Attouch and Bolte (2009) demonstrated the global convergence of the classic proximal-point algorithm. Furthermore, Attouch et al. (2013) proved a convergence result for descent methods satisfying several conditions, which includes the forward-backward splitting algorithm.

Global Convergence for Policy Gradient Methods Several works have analyzed the global convergence of policy gradient methods for MDPs with finite state and action spaces. Among these, Agarwal et al. (2021) provided a comprehensive analysis of policy gradient methods solving discounted infinite-horizon MDPs. They demonstrated iteration complexities of $\mathcal{O}(\epsilon^{-2})$ for exact policy gradient methods and $\mathcal{O}(\epsilon^{-1})$ for exact Natural Policy Gradient (NPG) methods with softmax parameterization. Cen et al. (2022) studied entropy-regularized NPG methods in conjunction with softmax parameterization. They established a linear convergence rate for exact entropy-regularized NPG methods solving entropy-regularized MDPs and further proved an $\tilde{\mathcal{O}}(\epsilon^{-2})$ sample complexity for approximate entropy-regularized NPG methods. Lan (2023) introduced stochastic Policy Mirror Descent (PMD) approaches and demonstrated $\mathcal{O}(\epsilon^{-1})$ (resp., $\mathcal{O}(\epsilon^{-2})$) sample complexity for solving RL problems with strongly convex (resp., convex) regularizers. Following these works, Klein et al. (2023) investigated the policy gradient methods with a softmax parameterization for finite-horizon MDPs. They established a weak PL condition of the policy gradient objective function and proved an $\mathcal{O}(\epsilon^{-1})$ iteration complexity for exact policy gradient methods. For more references, we refer the readers to Mei et al. (2020), Xiao (2022), and Fatkhullin et al. (2023a).

When solving MDPs with general state and action spaces, it is impractical to extend existing results directly as the complexity for policy gradient methods solving tabular MDPs depends on the cardinalities of state and action sets (Lan 2023, Klein et al. 2023). To address this issue, many works impose additional assumptions to ensure the global convergence of policy gradient methods. For instance, when dealing with discounted infinite-horizon MDPs, Lan (2022) demonstrated a linear convergence rate of PMD when the advantage function is convex and the regularizer is strongly convex. Bhandari and Russo (2024, Theorem 2) identified certain structural characteristics shared

by several discounted infinite-horizon control problems, which guarantee that the policy gradient objective function satisfies the so-called (c, μ) -gradient dominance condition (Bhandari and Russo 2024, Definition 2), a condition equivalent to the KL condition under some mild assumptions (Karimi et al. 2016). They established iteration complexities of $\mathcal{O}(\epsilon^{-2})$ and $\mathcal{O}(\log(\epsilon^{-1}))$ for exact policy gradient methods to achieve an ϵ -optimal policy under the $(c, 0)$ -gradient dominance condition and (c, μ) -gradient dominance condition with $\mu > 0$, respectively.

For finite-horizon MDPs, Bhandari and Russo (2024, Theorem 3) established conditions for a class of finite-horizon MDPs under which first-order stationary points are globally optimal, leading to an asymptotic convergence rate for policy gradient methods (Bhandari and Russo 2024, Lemma 2). Yet, they “leave the study of a gradient dominance condition for finite-horizon problems as future work”. Our work bridges this gap by establishing the KL condition of the policy gradient optimization for such problems.

In addition to the results for general MDPs, Fazel et al. (2018) established the gradient domination of the policy gradient objective function for the infinite-horizon discounted LQR problem. They demonstrated a linear convergence rate for exact policy gradient methods to achieve an optimal policy. Han et al. (2023) extended results to infinite horizon nearly linear quadratic regulators, where the dynamic system integrates linear and nonlinear components. They established a linear convergence rate of policy gradient methods using the linear policy class to achieve globally optimal policies. More closely related to our work, Hambly et al. (2021) studied policy gradient methods for solving the finite-horizon LQR problem. Their complexity admits a polynomial dependence on the time horizon. However, the applicability of their analysis is limited due to the specific structures inherent to LQR problems, making it challenging to generalize to other applications. Compared to Hambly et al. (2021), one can check that our framework works for general strongly convex per period costs, going beyond a quadratic form in the LQR problem.

Data-driven Operations Management Most of the literature studying multi-period operations models relies on the Sample Average Approximation (SAA) approach, which constructs an empirical objective using samples and then solves the corresponding empirical problem (Kleywegt et al. 2002). For instance, Levi et al. (2007) established the sample complexity required for SAA to find an ϵ -optimal base-stock policy of the multi-period inventory system. Cheung and Simchi-Levi (2019) applied the SAA method for multi-period capacitated stochastic inventory control problems and derived a sample complexity required to achieve a near-optimal expected cost. They further proposed a polynomial-time approximation scheme that also uses polynomially many samples to solve the empirical counterpart. Qin et al. (2022) investigated multi-period joint pricing and inventory control models and established the sample complexity for the

SAA approach. They applied a sparsification technique and proposed a polynomial-time approximation algorithm for the empirical problem. [Zhang et al. \(2022\)](#) applied the SAA method for managing inventories in an infinite-horizon series system with multiple stages and derived sample complexities. [Xie et al. \(2024\)](#) investigated uniform generalization errors of the SAA approach for different inventory policy classes and established a sample complexity non-dependent on the time horizon length for learning a base-stock policy which incurs an averaged cost no more than the optimal averaged cost plus ϵ .

In addition to SAA methods, some works apply value-based methods to solve multi-period operations problems. For example, [Qin et al. \(2023\)](#) introduced a variance-reduced value iteration algorithm for multi-period stochastic inventory control with independent demands, establishing matching upper and lower bounds on the sample complexity. [Gong and Simchi-Levi \(2023\)](#) developed online Q-learning methods for stochastic inventory models with cyclic demands. They considered two scenarios: the episodic model where inventory is discarded at the end of each cycle and the non-discarding case. We remark that the SAA approach and value-based methods usually rely on solving the empirical counterpart through dynamic programming, whereas policy gradient methods directly optimize a single objective.

Similar to policy gradient methods, some studies apply stochastic gradient methods to solve data-driven operations models. For instance, [Kunnumkal and Topaloglu \(2008\)](#) proposed a biased stochastic gradient method to solve finite-horizon inventory systems with independent demands and established an asymptotic convergence rate to achieve optimal base-stock levels. [Huh and Rusmevichientong \(2014\)](#) applied the same biased stochastic gradient method for a class of multistage stochastic optimization problems, where the objective satisfies a generalized convex condition called the sequentially convex condition.

Though our sequential decomposition inequalities are inspired by one of the conditions in sequential convexity [Huh and Rusmevichientong \(2014\)](#), our work is fundamentally different from theirs. First, [Huh and Rusmevichientong \(2014\)](#) (and [Kunnumkal and Topaloglu \(2008\)](#)) applied biased gradient methods to minimize cost-to-go functions, which differs from standard policy gradient methods that optimize the objective function directly as studied in the literature and this work. Second, the sequential convexity proposed in [Huh and Rusmevichientong \(2014\)](#) cannot be directly applied to the inventory system with Markov-modulated demands and stochastic cash balance problems. For example, the minimizers of cost-to-go functions differ from the optimal policy parameters of stochastic cash balance problems. On the other hand, our framework is readily applicable to the operations applications in [Huh and Rusmevichientong \(2014\)](#), including inventory control, capacity allocation, and lifetime buy decision problems. Third, [Huh and Rusmevichientong \(2014\)](#),

leveraging the sequential convexity, proved that biased stochastic gradient methods achieve a sample complexity with an exponential dependence on the planning horizon. In contrast, we show that policy gradient optimization satisfies the KL condition and policy gradient methods admit a sample complexity that scales polynomially with the planning horizon.

1.2. Organizations

The rest of this paper is structured as follows. Section 2 presents the MDP formulation and the policy gradient optimization problem. Section 3 outlines the definition and properties of the KL condition and identifies several easily verifiable conditions required to ensure the KL condition of the policy gradient optimization problem. In Section 4 and 5, we validate the KL condition for the policy gradient optimization problem of the entropy-regularized tabular MDPs and the LQR problem, respectively. Sections 6 and 7 establish the KL condition for policy gradient optimization problems of the inventory system with Markov-modulated demands and the stochastic cash balance problem, respectively. Utilizing the KL condition, we provide the first sample complexity results for these settings in the literature.

1.3. Notations and Definitions

We use the following notations throughout the paper. Let $\bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$. \mathbb{N}_+ denotes the set of all natural numbers. For $n \in \mathbb{N}_+$, denote $[n]$ as the set $\{1, \dots, n\}$. $\|x\|_2 = \sqrt{\sum_{i=1}^n |x_i|^2}$ denotes the l_2 -norm of a vector $x \in \mathbb{R}^n$. We use $\|A\|_2$ to denote the spectral norm of a matrix $A \in \mathbb{R}^{m \times n}$, the largest singular value of A . Let $\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2}$ denote the Frobenius norm of a matrix $A \in \mathbb{R}^{m \times n}$. We use $\lambda_{\max}(A)$ to denote the spectral radius of a square matrix $A \in \mathbb{R}^{n \times n}$, which is the largest eigenvalue of A . Let e denote $\exp(1)$. We use $\lfloor x \rfloor$ to denote the greatest integer less than or equal to x . For $D \in \mathbb{R}^n$, define $D_{[j,k]} := \sum_{i=j}^k D_i$ for $1 \leq j \leq k \leq n$. We say a point $x \in \mathcal{X}$ satisfies the first-order necessary optimality condition of the optimization problem $\min_{x \in \mathcal{X}} f(x)$ if $\langle \nabla f(x), x' - x \rangle \geq 0, \forall x' \in \mathcal{X}$ for a differentiable function f . A point \bar{x} is an ϵ -optimal solution of $\min_x f(x)$ if $f(\bar{x}) - \min_x f(x) \leq \epsilon$. Throughout the paper, we assume that the desired accuracy $\epsilon > 0$ is small enough so that $\epsilon^{-1} \gg \log(\epsilon^{-1}) \geq 1$. We use $\mathcal{O}(\cdot)$ to denote the order in terms of ϵ^{-1} and $\tilde{\mathcal{O}}(\cdot)$ to denote the order hiding the logarithmic dependency on ϵ^{-1} .

2. Problem Formulation

We specify a finite horizon MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, C, T, \rho)$ defined in Puterman (2014): the time horizon T ; the state space $\mathcal{S} = \mathcal{S}_1 \cup \dots \cup \mathcal{S}_T$, where $\mathcal{S}_t \subseteq \mathbb{R}^m$ is the feasible region for a state s at period t ; the action space $\mathcal{A} = \cup_{s \in \mathcal{S}} \mathcal{A}_s$, where $\mathcal{A}_s \subseteq \mathbb{R}^n$ is the set of feasible actions for state $s \in \mathcal{S} \subseteq \mathbb{R}^m$; the transition kernel $P: \mathcal{S} \times \mathcal{A} \times [T] \rightarrow \mathcal{S}$, where $P(s'|s, a, t)$ is the probability density

function (or probability mass function in the discrete setting) of transitioning into s' when taking action a in state s at period t ; the cost function $C : \mathcal{S} \times \mathcal{A} \times [T] \rightarrow \mathbb{R}$, where $C(s, a, t)$ is the immediate cost after taking action a in state s at period t ; and the initial state distribution ρ . For simplicity, we use $P_t(\cdot|s, a) := P(\cdot|s, a, t)$, and $C_t(s, a) := C(s, a, t)$ for all $s \in \mathcal{S}, a \in \mathcal{A}, t \in [T]$. The agent starts at state $s_1 \in \mathcal{S}_1$, which follows the initial state distribution ρ . At period t , the agent first observes the current state $s_t \in \mathcal{S}_t$ and then takes an action $a_t \in \mathcal{A}_{s_t}$. Afterwards, it receives an immediate cost $C_t(s_t, a_t)$ and proceeds to the next period with state $s_{t+1} \sim P_t(\cdot|s_t, a_t)$.

A non-stationary policy $\pi : \mathcal{S} \times [T] \rightarrow \mathcal{A}$ is a function that maps the current state s to a feasible action a at period t , e.g., $a = \pi(s, t)$. Similarly, we use $\pi_t(\cdot)$ to denote the policy at period t and $\pi_t(s) := \pi(s, t)$ for all $s \in \mathcal{S}, t \in [T]$. Let Π denote the set of feasible policies and Π_t denote the set of feasible policies at period t . For any $\pi \in \Pi$, the total expected cost starting from state s is

$$J^\pi(s) = \mathbb{E} \left[\sum_{t=1}^T C_t(s_t, \pi_t(s_t)) \middle| s_1 = s, \pi \right].$$

We take the expectation over a Markovian sequence (s_1, \dots, s_T) , where s_1 is the initial state and $s_{t+1} \sim P_t(\cdot|s_t, \pi_t(s_t))$ for all $t = 1, \dots, T-1$. A policy π^* is optimal if it minimizes the total expected cost $J(\pi)$ with the initial distribution ρ :

$$J(\pi) = \mathbb{E}_{s \sim \rho} [J^\pi(s)] = \mathbb{E} \left[\sum_{t=1}^T C_t(s_t, \pi_t(s_t)) \middle| s_1 \sim \rho, \pi \right].$$

2.1. Bellman Equation

We introduce several terminologies commonly used in the literature of MDPs. Let π be a given policy. We define $\rho_t(\cdot|\pi)$ as the cumulative distribution function of s_t incurred by policy π starting with the initial distribution ρ . By definition, we have $\rho_1(\cdot|\pi) = \rho$. Furthermore, we define the value function $V_t^\pi : \mathcal{S} \rightarrow \mathbb{R}$, which represents the total expected cost at time t starting with the initial state s and policy π :

$$V_t^\pi(s) = \mathbb{E} \left[\sum_{k=t}^T C_k(s_k, \pi_k(s_k)) \middle| s_t = s, \pi \right].$$

In the same manner, we define the function $Q_t^\pi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ as the action-value (or Q-value) function:

$$Q_t^\pi(s, a) = C_t(s, a) + \mathbb{E} \left[\sum_{k=t+1}^T C_k(s_k, \pi_k(s_k)) \middle| s_t = s, a_t = a, \pi \right].$$

By definition, the value function V^π and action-value function Q^π have the following relationships:

$$\begin{cases} V_t^\pi(s) = Q_t^\pi(s, \pi_t(s)), \\ Q_t^\pi(s, a) = C_t(s, a) + \mathbb{E}[V_{t+1}^\pi(s') | s' \sim P_t(\cdot|s, a)], \end{cases} \quad (1)$$

for all $s \in \mathcal{S}, a \in \mathcal{A}, t \in [T]$ and the boundary condition is $V_{T+1}^\pi(\cdot) = 0$ for all $\pi \in \Pi$. These are commonly known as the Bellman equations in the literature (Bellman 1952). By the principle of optimality (Puterman 2014), an optimal policy π^* solves the following Bellman equations:

$$\begin{cases} V_t^*(s) = \min_{\pi_t \in \Pi_t} Q_t^*(s, \pi_t(s)), \\ Q_t^*(s, a) = C_t(s, a) + \mathbb{E}[V_{t+1}^*(s') | s' \sim P_t(\cdot | s, a)], \end{cases} \quad (2)$$

for all $s \in \mathcal{S}, a \in \mathcal{A}, t \in [T]$ and the boundary condition is $V_{T+1}^*(\cdot) = 0$. Here V_t^* and Q_t^* denote the value function and the Q-value function corresponding with the optimal policy π^* , respectively.

2.2. Policy Gradient Formulation

Policy gradient methods apply first-order algorithms to minimize the total expected cost $J(\pi)$. Note that general policy optimization falls into functional optimization as we search over the function class Π , which is computationally intractable. To avoid functional optimization, it is common to parameterize the policy through finite-dimensional parameters $\theta = (\theta_1, \dots, \theta_T)$ (Sutton et al. 1999). At time t , the parameterized policy is $\pi_t(\cdot | \theta_t)$ and θ_t belongs to a convex and compact set $\Theta_t \subseteq \mathbb{R}^d$. The feasible region of θ is a product set $\Theta = \Theta_1 \times \dots \times \Theta_T$, which is also convex and compact. In such a case, the parameterized policy class is $\Pi_\Theta = \{\pi(s, t | \theta) : \mathcal{S} \times [T] \times \Theta \rightarrow \mathcal{A}\} \subseteq \Pi$ and we use π_θ to denote $\pi(\cdot | \theta)$ for simplicity.

For a given parameterized policy π_θ , we represent the total expected cost by $l(\theta) := J(\pi_\theta)$, called the policy gradient objective function. We define a policy π_θ to be ϵ -optimal if θ is an ϵ -optimal solution of $l(\theta)$. We denote $\nabla_t l(\theta) = \frac{\partial l(\theta)}{\partial \theta_t}$ as the partial derivative of $l(\theta)$ with respect to the t -th coordinate θ_t . We further denote the gradient $\nabla l(\theta) = (\nabla_1 l(\theta), \dots, \nabla_T l(\theta))$.

Let θ^* denote one of the minimizers of $\min_{\theta \in \Theta} l(\theta)$ and π_{θ^*} as the corresponding policy. We define $V_t^{\pi_{\theta^*}}$ and $Q_t^{\pi_{\theta^*}}$ as its corresponding value and Q-value function, respectively. It's worth noting that V_t^* (resp., Q_t^*) and $V_t^{\pi_{\theta^*}}$ (resp., $Q_t^{\pi_{\theta^*}}$) are identical when the parameterized policy class Π_Θ contains the optimal policy π^* . This often occurs when the optimal policy class is known, e.g., affine policies in the Linear Quadratic Regulator (LQR) problem and base-stock policies in the multi-period inventory control model.

3. Landscape Characterization

In this section, we first formally define the KL condition and discuss its properties. Next, we present the non-asymptotic convergence rate of algorithms for optimization problems satisfying the KL condition. Lastly, we provide a set of verifiable assumptions for the policy gradient optimization in Section 2.2 to satisfy the KL condition, enabling a non-asymptotic convergence rate for policy gradient methods to achieve globally optimal solutions.

3.1. Definition and Properties of KL Condition

Before formally introducing the KL condition, we provide some basic definitions from variational analysis.

DEFINITION 1 (FRÉCHET SUBDIFFERENTIAL (ROCKAFELLAR AND WETS 2009)). Let $f : \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$ be a proper lower semicontinuous function. For each $x \in \text{dom} f$, the Fréchet subdifferential of f at $x \in \mathbb{R}^n$, written as $\hat{\partial}f(x)$, is the set of $v \in \mathbb{R}^n$ such that

$$\liminf_{y \neq x, y \rightarrow x} \frac{f(y) - f(x) - \langle v, y - x \rangle}{\|x - y\|_2} \geq 0.$$

We set $\hat{\partial}f(x) = \emptyset$ if $x \notin \text{dom} f$.

DEFINITION 2 (LIMITING SUBDIFFERENTIAL (MORDUKHOVICH 1976)). The limiting subdifferential of f at $x \in \mathbb{R}^n$ is the set $\partial f(x) := \{v \in \mathbb{R}^n : \exists x_n \rightarrow x, f(x_n) \rightarrow f(x), v_n \in \hat{\partial}f(x_n), v_n \rightarrow v\}$.

Definitions 1 and 2 imply that $\hat{\partial}f(x) \subseteq \partial f(x)$. The limiting subdifferential $\partial f(x)$ reduces to the gradient $\nabla f(x)$ when f is continuously differentiable and reduces to the classic subdifferential when f is convex.

Although the KL condition usually describes the landscape of unconstrained problems, we can use it for constrained smooth optimization problems. In fact, for a constrained optimization problem $\min_{x \in \mathcal{X}} f(x)$ over a closed set \mathcal{X} , one can write it as an unconstrained optimization problem $\min_x f + \delta_{\mathcal{X}}$ over \mathbb{R}^n , where $\delta_{\mathcal{X}}(x)$ represents the indicator function for \mathcal{X} and is defined as follows:

$$\delta_{\mathcal{X}}(x) = \begin{cases} 0 & x \in \mathcal{X}, \\ +\infty & x \notin \mathcal{X}. \end{cases}$$

To characterize the landscape of constrained smooth optimization problems, we utilize a specific form of the KL condition (Karimi et al. 2016, Appendix G). For more general definitions, we refer the readers to (Attouch et al. 2013, Definition 2.4).

DEFINITION 3 (KL CONDITION). Consider a convex and compact set $\mathcal{X} \subseteq \mathbb{R}^n$ and a differentiable function f . Denote f^* as the optimal function value with $f^* := \min_{x \in \mathcal{X}} f(x)$. The function f satisfies the KL condition on \mathcal{X} if there exists $\mu > 0$ such that

$$f(x) - f^* \leq \frac{1}{2\mu} \min_{g \in \partial \delta_{\mathcal{X}}(x)} \|\nabla f(x) + g\|_2^2 \quad \forall x \in \mathcal{X},$$

where μ refers to the KL constant.

REMARK 1. When the decision variable x is a matrix, e.g., parameters in the LQR problem, we replace the l_2 -norm with the Frobenius norm.

The subdifferential of $\delta_{\mathcal{X}}(x)$ is the normal cone of \mathcal{X} at x . When $\mathcal{X} = \mathbb{R}^n$, the KL Condition reduces to the PL condition. Similar to the PL condition, the KL condition is a generalization of the strong convexity (Corollary 1 in Appendix A). It's well known that strongly convex functions exclude all suboptimal stationary or local optimal points. Optimization problems with the KL condition exhibit the same structural property.

PROPOSITION 1 (**Karimi et al. (2016)**). *Consider a convex and compact set $\mathcal{X} \subseteq \mathbb{R}^n$. Suppose a function $f: \mathcal{X} \rightarrow \mathbb{R}$ satisfies the KL condition with a KL constant $\mu > 0$ over \mathcal{X} . Then, any point satisfying the first-order necessary optimality condition of the optimization problem $\min_{x \in \mathcal{X}} f(x)$ is globally optimal.*

3.2. Convergence Rate under KL Condition

This subsection presents the global convergence results of the projected (stochastic) gradient descent for solving the stochastic optimization problem over a convex and compact set \mathcal{X} under the KL condition:

$$\min_{x \in \mathcal{X}} f(x) := \mathbb{E}_{\xi \sim \mathbb{P}(\xi)} [F(x, \xi)], \quad (3)$$

where x is the decision variable and ξ is the random variable with a cumulative distribution function $\mathbb{P}(\xi)$. Furthermore, we analyze the algorithm's sample complexity, which refers to the number of samples required to obtain an ϵ -optimal solution.

We assume f is differentiable and L -smooth to establish the convergence result. These assumptions are standard in the stochastic approximation literature (Nemirovski et al. 2009, Ghadimi and Lan 2013).

ASSUMPTION 1. *For the function $f: \mathcal{X} \rightarrow \mathbb{R}$, we assume that it is differentiable and is L -smooth, i.e., its gradient is L -Lipschitz:*

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y \in \mathcal{X}.$$

It is sometimes impractical to compute the gradient in stochastic optimization problems. We use the stochastic gradients instead of computing the true expectation in such settings. In the rest of this subsection, we will denote $\nabla F(x_k, \xi_k)$ as the stochastic gradient with one sample ξ_k drawn from $\mathbb{P}(\xi)$ and $\nabla \hat{f}(x_k)$ as the stochastic gradient with an empirical distribution constructed by N i.i.d samples $\{\xi_k^{(i)}\}_{i=1}^N$:

$$\nabla \hat{f}(x_k) := \frac{1}{N} \sum_{i=1}^N \nabla F(x_k, \xi_k^{(i)}).$$

ASSUMPTION 2. *For the gradient information, assume one of the two options holds:*

1. *At each iteration k , the gradient $\nabla f(x_k)$ is accessible.*

2. At each iteration k , only the stochastic gradient $\nabla F(x_k, \xi_k)$ is accessible. In addition, given x_k , the stochastic gradient $\nabla F(x_k, \xi_k)$ is unbiased and the variance is bounded by σ^2 ,

$$\begin{aligned}\mathbb{E}_{\xi_k}[\nabla F(x_k, \xi_k)] &= \nabla f(x_k), \\ \mathbb{E}_{\xi_k}[\|\nabla F(x_k, \xi_k) - \nabla f(x_k)\|_2^2] &\leq \sigma^2.\end{aligned}\tag{4}$$

From Assumption 2.2, the stochastic gradient $\nabla \hat{f}(x_k)$ is unbiased $\mathbb{E}_{\xi_k}[\nabla \hat{f}(x_k)] = \nabla f(x_k)$. Furthermore, its variance is bounded by $\mathbb{E}_{\xi_k}[\|\nabla \hat{f}(x_k) - \nabla f(x_k)\|_2^2] \leq \sigma^2/N$.

Extensive research has studied the convergence of first-order methods for optimization problems satisfying the KL condition. Attouch et al. (2013) presented a general framework for analyzing the convergence of a class of descent methods, in which the projected gradient descent is a special case. Following the same framework, one can easily extend the convergence results to the stochastic setting. We state the results for completeness and leave the proof in Appendix A.

LEMMA 1. Consider a constrained optimization problem $\min_{x \in \mathcal{X}} f(x)$ over a convex and compact set \mathcal{X} . Assume f satisfies the KL condition on \mathcal{X} with a parameter $\mu > 0$, and Assumptions 1 and 2 hold. Denote $f^* = \inf_{x \in \mathcal{X}} f(x)$.

1. (Attouch et al. 2013) The sequence of projected gradient descent $x_{k+1} = \text{Proj}_{\mathcal{X}}(x_k - \gamma_k \nabla f(x_k))$ with stepsizes $\gamma_k = \frac{1}{L}$ achieves a linear convergence rate, i.e.,

$$f(x_k) - f^* \leq \left(1 - \frac{\mu}{4L + \mu}\right)^k (f(x_0) - f^*).$$

2. The sequence of projected stochastic gradient descent $x_{k+1} = \text{Proj}_{\mathcal{X}}(x_k - \gamma_k \nabla \hat{f}(x_k))$ with stepsizes $\gamma_k = \frac{1}{L}$ admits a sublinear convergence rate, i.e.,

$$\mathbb{E}[f(x_k)] - f^* \leq \left(1 - \frac{\mu}{16L + \mu}\right)^k \left(\mathbb{E}[f(x_0)] - f^*\right) + \frac{17\sigma^2}{\mu N}.$$

REMARK 2. From Lemma 1.2, we need to set $N = \mathcal{O}(\epsilon^{-1})$ and $k = \mathcal{O}(\log(\epsilon^{-1}))$ to obtain an ϵ -optimal solution. As a result, the sample complexity of projected stochastic gradient descent is $\tilde{\mathcal{O}}(\epsilon^{-1})$.

3.3. KL Condition in Policy Gradient Formulation

Leveraging the KL condition, we can establish the global convergence of first-order methods for nonconvex smooth optimization problems. However, verifying the KL condition for the policy gradient optimization is challenging. To address this difficulty, we develop a general framework to validate the KL condition for a class of MDPs in the following theorem. To maintain consistency in notation, we present the theorem using the same terminology in Section 2.

THEOREM 1. Consider a Markov Decision Process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, C, T, \rho)$ and a policy class Π_Θ with a convex and compact set Θ . Suppose the following conditions hold.

1. **(Bounded Gradients)** For any $t \in [T]$, the expected Q -value function $\mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^{\pi_\theta}(s_t, \pi_t(s_t|\theta_t)) \right]$ is continuously differentiable on Θ_t with the 2-norm of its gradient upper bounded by G .
2. **(KL Condition of Expected Optimal Q -value Functions)** For any $t \in [T]$, the expected optimal Q -value function $\mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^{\pi_{\theta^*}}(s_t, \pi_t(s_t|\theta_t)) \right]$ satisfies the KL condition with a KL constant μ_Q on Θ_t .
3. **(Sequential Decomposition Inequality)** For any $\theta \in \Theta$ and $1 \leq t < k \leq T$, there exists $M_g > 0$:

$$\begin{aligned} & \left\| \nabla_{\theta_t} l(\theta_1, \dots, \theta_{k-1}, \theta_k, \theta_{k+1}^*, \dots, \theta_T^*) - \nabla_{\theta_t} l(\theta_1, \dots, \theta_{k-1}, \theta_k^*, \theta_{k+1}^*, \dots, \theta_T^*) \right\|_2 \\ & \leq M_g \left(\mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} \left[Q_k^{\pi_{\theta^*}}(s_k, \pi_k(s_k|\theta_k)) \right] - \mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} \left[Q_k^{\pi_{\theta^*}}(s_k, \pi_k(s_k|\theta_k^*)) \right] \right). \end{aligned} \quad (5)$$

Then the policy gradient objective function $l(\theta)$ satisfies the KL condition on Θ . Furthermore, the corresponding KL constant is $\mu_l = \frac{\mu_Q^3}{eM_g^2G^2T^2}$, i.e.,

$$l(\theta) - l(\theta^*) \leq \frac{eM_g^2G^2T^2}{2\mu_Q^3} \min_{g \in \partial \delta_\Theta(\theta)} \left\| \nabla l(\theta) + g \right\|_2^2, \quad \forall \theta \in \Theta.$$

REMARK 3. The parameters G , μ_Q , and M_g may depend on the time horizon T . The following sections show these parameters exhibit polynomial dependence on T for different applications.

Theorem 1 provides a set of assumptions to validate the KL condition of the policy gradient optimization problem in finite-horizon MDPs with general state and action spaces. Compared to Bhandari and Russo (2024, Theorem 3), which showed that $l(\theta)$ has no suboptimal stationary points, Theorem 1 characterizes a stronger landscape condition by demonstrating the KL condition. Therefore, we can obtain an $\tilde{\mathcal{O}}(\epsilon^{-1})$ sample complexity for stochastic policy gradient methods to attain an ϵ -optimal policy, which is an improvement upon the asymptotic convergence rate in Bhandari and Russo (2024).

We provide some intuitions as to why the conditions in Theorem 1 hold for a broad class of MDPs. The *bounded gradients* condition is standard and is likely to hold for many applications. For the *KL condition of expected optimal Q -value functions*, one can verify it using convex cost-to-go functions and strongly convex costs. We discuss more intuitions in the following sections for different cases. The less intuitive condition is the *sequential decomposition inequality*. A weaker result using standard assumptions can be easily derived. Suppose $l(\theta)$ is S_l -smooth, and expected optimal Q -value functions satisfy the KL condition. We have

$$\begin{aligned} & \left\| \nabla_{\theta_t} l(\theta_1, \dots, \theta_k, \theta_{k+1}^*, \dots, \theta_T^*) - \nabla_{\theta_t} l(\theta_1, \dots, \theta_{k-1}, \theta_k^*, \dots, \theta_T^*) \right\|_2 \leq S_l \|\theta_k - \theta_k^*\|_2 \\ & \leq S_l \sqrt{\frac{\mu}{2}} \sqrt{\mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} \left[Q_k^{\pi_{\theta^*}}(s_k, \pi_k(s_k|\theta_k)) \right] - \mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} \left[Q_k^{\pi_{\theta^*}}(s_k, \pi_k(s_k|\theta_k^*)) \right]}. \end{aligned}$$

The last inequality holds as the KL condition implies the quadratic growth condition (Karimi et al. 2016). In our analysis, this weaker condition leads to a suboptimal characterization of the KL constant with an exponential dependence on the time horizon T (see discussions in Appendix A.5). To remove such an exponential dependence, we instead use the stronger *sequential decomposition inequalities*. Interestingly, we demonstrate that *sequential decomposition inequalities* indeed hold for the control and operations models analyzed in the following sections.

Proof Sketch: To understand the high-level idea of Theorem 1, we illustrate the role of each condition. First, the differentiability of $l(\theta)$ is essential to establish the KL condition in Definition 3. Deterministic Policy Gradient Theorem (Silver et al. 2014) requires the *differentiability* condition of the expected Q -value function to ensure the differentiability of $l(\theta)$ and provides the expression of $\nabla l(\theta)$:

$$\nabla_{\theta_t} l(\theta) = \nabla_{\theta_t} \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^{\pi_\theta}(s_t, \pi_t(s_t|\theta_t)) \right]. \quad (6)$$

Based on the policy gradient formulation, our goal is to establish the relationship between the suboptimality gap $l(\theta) - l(\theta^*)$ and $\nabla l(\theta)$. However, $l(\theta) = J(\pi_\theta) = \mathbb{E} \left[\sum_{t=1}^T C_t(s_t, \pi_t(s_t|\theta_t)) | s_1 \sim \rho, \pi_\theta \right]$ has a nested formulation and is difficult for one to check the KL condition directly. Thanks to the Performance Difference Lemma (Kakade and Langford 2002), we have

$$l(\theta) - l(\theta^*) = \sum_{t=1}^T \left(\mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^{\pi_{\theta^*}}(s_t, \pi_t(s_t|\theta_t)) \right] - \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^{\pi_{\theta^*}}(s_t, \pi_t(s_t|\theta_t^*)) \right] \right).$$

The suboptimality gap $l(\theta) - l(\theta^*)$ can be decomposed as the differences of expected optimal Q -value functions under two different single-stage policies at each period t . Therefore, we only need to check the *KL condition of expected optimal Q -value functions* for different applications. Leveraging the *KL condition of expected optimal Q -value functions*, we have

$$l(\theta) - l(\theta^*) \leq \sum_{t=1}^T \frac{1}{2\mu_Q} \min_{g_t \in \partial \delta_{\Theta_t}(\theta_t)} \left\| \nabla_{\theta_t} \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^{\pi_{\theta^*}}(s_t, \pi_t(s_t|\theta_t)) \right] + g_t \right\|_2^2. \quad (7)$$

The right-hand side of (7) differs from the gradient formulation $\nabla_{\theta_t} l(\theta)$ presented in (6). To establish the KL condition of $l(\theta)$, we need to prove a *bounded gradient mismatch* inequality:

$$\begin{aligned} & \sum_{t=1}^T \min_{g_t \in \partial \delta_{\Theta_t}(\theta_t)} \left\| \nabla_{\theta_t} \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^{\pi_{\theta^*}}(s_t, \pi_t(s_t|\theta_t)) \right] + g_t \right\|_2^2 \\ & \leq M \sum_{t=1}^T \min_{g_t \in \partial \delta_{\Theta_t}(\theta_t)} \left\| \nabla_{\theta_t} \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^{\pi_\theta}(s_t, \pi_t(s_t|\theta_t)) \right] + g_t \right\|_2^2. \end{aligned} \quad (8)$$

This inequality captures the relationship of Q-value functions' gradients under two policies. To prove it, we note that

$$\begin{aligned}
& \left\| \nabla_{\theta_t} \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^{\pi_{\theta^*}}(s_t, \pi_t(s_t|\theta_t)) \right] + g_t \right\|_2 - \left\| \nabla_{\theta_t} \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^{\pi_\theta}(s_t, \pi_t(s_t|\theta_t)) \right] + g_t \right\|_2 \\
& \leq \left\| \nabla_{\theta_t} \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^{\pi_{\theta^*}}(s_t, \pi_t(s_t|\theta_t)) \right] - \nabla_{\theta_t} \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^{\pi_\theta}(s_t, \pi_t(s_t|\theta_t)) \right] \right\|_2 \\
& = \left\| \nabla_{\theta_t} l(\theta_1, \dots, \theta_t, \theta_{t+1}^*, \dots, \theta_T^*) - \nabla_{\theta_t} l(\theta_1, \dots, \theta_t, \theta_{t+1}, \dots, \theta_T) \right\|_2.
\end{aligned}$$

Applying the *sequential decomposition inequality* and the *KL condition of expected optimal Q-value functions*, we end up with the following inequalities:

$$\begin{aligned}
& \left\| \nabla_{\theta_t} l(\theta_1, \dots, \theta_t, \theta_{t+1}^*, \dots, \theta_T^*) - \nabla_{\theta_t} l(\theta_1, \dots, \theta_t, \theta_{t+1}, \dots, \theta_T) \right\|_2 \\
& \leq \sum_{k=t+1}^T \left\| \nabla_{\theta_t} l(\theta_1, \dots, \theta_k, \theta_{k+1}^*, \dots, \theta_T^*) - \nabla_{\theta_t} l(\theta_1, \dots, \theta_{k-1}, \theta_k^*, \dots, \theta_T^*) \right\|_2 \\
& \leq \sum_{k=t+1}^T M_g \left(\mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} \left[Q_k^{\pi_{\theta^*}}(s_k, \pi_k(s_k|\theta_k)) \right] - \mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} \left[Q_k^{\pi_{\theta^*}}(s_k, \pi_k(s_k|\theta_k^*)) \right] \right) \\
& \leq \sum_{k=t+1}^T \frac{M_g}{2\mu_Q} \min_{g_k \in \partial \delta_{\Theta_k}(\theta_k)} \left\| \nabla_{\theta_k} \mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} \left[Q_k^{\pi_{\theta^*}}(s_k, \pi_k(s_k|\theta_k)) \right] + g_k \right\|_2^2.
\end{aligned}$$

The following technical lemma is essential for us to prove the *bounded gradient mismatch* inequality.

LEMMA 2. Assume that the nonnegative sequences $\{X_t\}_{t=1}^T$ and $\{Y_t\}_{t=1}^T$ satisfy

$$|X_t - Y_t| \leq M_g \sum_{k=t+1}^T X_k^2, \quad (9)$$

with some positive constant M_g . If $X_T = Y_T$ and $X_t, Y_t \leq G$ for all $t = 1, \dots, T$, then

$$\sum_{t=1}^T X_t^2 \leq \max\{e, 4eM_g^2G^2T^2\} \sum_{t=1}^T Y_t^2.$$

Without loss of generality, we focus on the case with $e < 4eM_g^2G^2T^2$. Employing Lemma 2, we prove the *bounded gradient mismatch* inequality (8), thereby demonstrating how (6) relates to (7). This completes the proof of the KL condition. We refer readers to Appendix A.4 for a more rigorous proof.

Theorem 1 allows us to verify the KL condition for policy gradient optimization problems by verifying several more manageable conditions. In the following sections, we demonstrate the KL condition using Theorem 1 for several control and operations models, e.g., the entropy-regularized tabular MDPs, the LQR problem, the inventory systems with Markov-modulated demand, and the stochastic cash-balance problem. Leveraging the KL condition, we demonstrate the global convergence of exact and stochastic policy gradient methods for solving these problems and provide their sample complexities to achieve ϵ -optimal solutions.

4. Entropy-Regularized Tabular MDPs

Tabular MDP is one of the most popular models in Reinforcement Learning (RL) and many papers focus on this setting (Agarwal et al. 2021, Lan 2023, Klein et al. 2023). This section considers a finite horizon version of this problem with an entropy regularization to the per-period cost functions, a variant of infinite horizon regularized tabular MDPs analyzed in Bhandari and Russo (2024). The regularization smooths the objective function such that optimal policies are stochastic. See Geist et al. (2019) for a more detailed discussion about the regularized MDPs.

4.1. Problem Formulation

Consider an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, C, T, \rho)$ with a finite state space $\mathcal{S} = \{1, \dots, m\}$. We assume a finite set of actions $\mathcal{N} = \{1, \dots, n\}$ to choose and take $\mathcal{A} = \Delta(\mathcal{N})$ as the set of probability distributions over these actions. For the tabular setting, it is natural to work directly with a randomized policy $\pi_t(s_t|\theta_t) = \theta_t(s_t, \cdot) \in \mathcal{A}$, instead of using any parameterization as the functional optimization reduces to a finite-dimensional optimization problem. Therefore, the per-period costs and transition functions of π_θ are

$$C_t(s_t, \pi_t(s_t|\theta_t)) = \sum_{i \in \mathcal{N}} \theta_t(s_t, i) C_t(s_t, i), \quad P_t(s_{t+1}|s_t, \pi_t(s_t|\theta_t)) = \sum_{i \in \mathcal{N}} \theta_t(s_t, i) P_t(s_{t+1}|s_t, i).$$

We assume that per-period costs are non-negative and uniformly bounded by \bar{C} for any $s_t \in \mathcal{S}$ and $i \in \mathcal{N}$.

Define $\mathcal{R}(p) := D_{\text{KL}}(U||p) = \sum_{i=1}^n \frac{1}{n} \log(\frac{1/n}{p_i})$ as the Kullback-Leibler (KL) divergence between a uniform distribution and $p \in \Delta(\mathcal{N})$. $\mathcal{R}(p)$ is a strongly convex function with $\text{dom}(\mathcal{R}) = \{p \in \Delta(\mathcal{A}) : \mathcal{R}(p) < +\infty\} = \{p \in \Delta(\mathcal{A}) : \min_i p_i > 0\}$. The entropy-regularized per-period cost of π is

$$C_t^r(s_t, \pi_t(s_t|\theta_t)) = C_t(s_t, \pi_t(s_t|\theta_t)) + \lambda \mathcal{R}(\pi_t(s_t|\theta_t)).$$

Since $\theta_t(s_t, \cdot)$ denotes a probability vector, it automatically falls into the constraints $\sum_{i=1}^n \theta_t(s, i) = 1$, $\theta_t(s, i) \geq 0$, $\forall s \in \mathcal{S}$, $\forall i \in \mathcal{N}$. Adding regularizations requires that $\theta_t(s, i) > 0, \forall s \in \mathcal{S}, \forall i \in \mathcal{N}$. We can further restrict the feasible region by analyzing the property of θ^* . From the Bellman equation (2), θ_t^* minimizes the following function:

$$\begin{aligned} f_t(\theta_t) &:= C_t(s_t, \pi_t(s_t|\theta_t)) + \lambda \mathcal{R}(\pi_t(s_t|\theta_t)) + \sum_{s_{t+1} \in \mathcal{S}} P_{t+1}(s_{t+1}|s_t, \pi_t(s_t|\theta_t)) V_{t+1}^*(s_{t+1}) \\ &= \lambda \mathcal{R}(\theta_t(s, \cdot)) + \sum_{i \in \mathcal{N}} \theta_t(s, i) \left(C_t(s, i) + \sum_{s_{t+1} \in \mathcal{S}} P_{t+1}(s_{t+1}|s, i) V_{t+1}^*(s_{t+1}) \right). \end{aligned}$$

From the first-order optimality condition, we have $\nabla f_t(\theta_t^*) = 0$, which implies

$$\nabla_{\theta_t(s, i)} f_t(\theta_t^*) = -\frac{\lambda}{n\theta_t^*(s, i)} + C_t(s_t, i) + \sum_{s_{t+1} \in \mathcal{S}} P_{t+1}(s_{t+1}|s_t, i) V_{t+1}^*(s_{t+1}) = 0, \quad \forall s \in \mathcal{S}, i \in \mathcal{N}.$$

Therefore

$$\theta_t^*(s, i) \geq \frac{\lambda}{n(C_t(s, i) + \sum_{s_{t+1} \in \mathcal{S}} P_{t+1}(s_{t+1}|s, i) V_{t+1}^*(s_{t+1}))} \geq \frac{\lambda}{n\bar{C}T}.$$

The last inequality holds as the accumulated cost is upper bounded by $T\bar{C}$. Thus, we can add $\theta_t(s, i) \geq \lambda/(n\bar{C}T)$ into the feasible region without cutting off optimal solutions. We use $\underline{p} := \lambda/(n\bar{C}T)$. The feasible set of policy parameters is $\Theta_t = \{\theta_t \in \mathbb{R}^{m \times n} : \sum_{i=1}^n \theta_t(s, i) = 1, \theta_t(s, i) \geq \underline{p}, \forall s \in \mathcal{S}, \forall i \in \mathcal{N}\}$.

4.2. KL condition of Policy Gradient Objectives

Let π^* denote the optimal policy that minimizes the total expected cost and Q^* denote the corresponding Q-value function. For any $\pi \in \Pi$, the expected Q-value function satisfies the Bellman equation (1):

$$\begin{aligned} & \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^{\pi_\theta}(s_t, \pi_t(s_t|\theta_t)) \right] \\ &= \sum_{s_t \in \mathcal{S}} \rho_t(s_t|\pi_\theta) \left(C_t(s_t, \pi_t(s_t|\theta_t)) + \lambda \mathcal{R}(\pi_t(s_t|\theta_t)) + \sum_{s_{t+1} \in \mathcal{S}} P_{t+1}(s_{t+1}|s_t, \pi_t(s_t|\theta_t)) V_{t+1}^{\pi_\theta}(s_{t+1}) \right) \\ &= \sum_{s_t \in \mathcal{S}} \rho_t(s_t|\pi_\theta) \left(\underbrace{\lambda \mathcal{R}(\theta_t(s_t, \cdot))}_{(I)} + \underbrace{\sum_{i \in \mathcal{N}} \theta_t(s_t, i) \left(C_t(s_t, i) + \sum_{s_{t+1} \in \mathcal{S}} P_{t+1}(s_{t+1}|s_t, i) V_{t+1}^{\pi_\theta}(s_{t+1}) \right)}_{(II)} \right). \end{aligned}$$

Term (I) is (λ/n) -strongly convex in θ_t (Bhandari and Russo 2024) and (II) is linear in θ_t . The *differentiability* condition holds as the linear function and entropy regularization are smooth. If we replace $V_{t+1}^{\pi_\theta}$ with V_{t+1}^* in (II), we get the expression of expected optimal Q-value function at period t , which is (λ/n) -strongly convex in θ_t . From Corollary 1, expected optimal Q-value functions satisfy the KL condition.

LEMMA 3. *The expected Q-value function $\mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} [Q_t^{\pi_\theta}(s_t, \pi_t(s_t|\theta_t))]$ is continuously differentiable on Θ_t . Furthermore, the expected optimal Q-value function $\mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} [Q_t^*(s_t, \pi_t(s_t|\theta_t))]$ satisfies the KL condition with constant λ/n over Θ_t .*

To verify the *gradient mismatch* condition, we apply the Policy Gradient Theorem (Sutton et al. 1999, Theorem 1) and get

$$\begin{aligned} \nabla_{\theta_t(s_t, i)} l(\theta) &= \nabla_{\theta_t(s_t, i)} \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^{\pi_\theta}(s_t, \pi_t(s_t|\theta_t)) \right] \\ &= \rho_t(s_t|\pi_\theta) \left(\underbrace{-\frac{\lambda}{n\theta_t(s_t, i)}}_{(I)} + \underbrace{C_t(s_t, i)}_{(II)} + \sum_{s_{t+1} \in \mathcal{S}} P_{t+1}(s_{t+1}|s_t, i) V_{t+1}^{\pi_\theta}(s_{t+1}) \right). \end{aligned}$$

The absolute value of (I) is upper bounded by $\lambda/(n\underline{p})$. (II) is uniformly bounded by \bar{C} . The regularized per-period costs are bounded as well. Therefore, $l(\theta)$ has bounded gradients by mathematical induction.

LEMMA 4. *The policy gradient objective function $l(\theta)$ has bounded gradients*

$$\|\nabla_{\theta_t} l(\theta)\|_F \leq T\bar{C} + \frac{\lambda}{n\underline{p}} + \lambda T \log\left(\frac{1}{n\underline{p}}\right), \quad \forall t \in [T].$$

According to Theorem 1, the last condition we need to verify is *sequential decomposition inequalities*. This inequality characterizes how far the gradients under the two policies

$$\pi_\alpha := (\pi_1(\cdot|\theta_1), \dots, \pi_{k-1}(\cdot|\theta_{k-1}), \pi_k(\cdot|\theta_k), \pi_{k+1}(\cdot|\theta_{k+1}^*), \dots, \pi_T(\cdot|\theta_T^*))$$

and

$$\pi_\beta := (\pi_1(\cdot|\theta_1), \dots, \pi_{k-1}(\cdot|\theta_{k-1}), \pi_k(\cdot|\theta_k^*), \pi_{k+1}(\cdot|\theta_{k+1}^*), \dots, \pi_T(\cdot|\theta_T^*)).$$

The structure of tabular MDPs naturally builds the connection between the difference in gradients and the difference in optimal Q-value functions for $t < k$:

$$\begin{aligned} & \left\| \nabla_{\theta_t} l(\theta_1, \dots, \theta_{k-1}, \theta_k, \theta_{k+1}^*, \dots, \theta_T^*) - \nabla_{\theta_t} l(\theta_1, \dots, \theta_{k-1}, \theta_k^*, \theta_{k+1}^*, \dots, \theta_T^*) \right\|_F \\ & \leq \sum_{s_t \in \mathcal{S}, i_t \in \mathcal{N}} \left| \nabla_{\theta_t(s_t, i_t)} l(\theta_1, \dots, \theta_{k-1}, \theta_k, \theta_{k+1}^*, \dots, \theta_T^*) - \nabla_{\theta_t(s_t, i_t)} l(\theta_1, \dots, \theta_{k-1}, \theta_k^*, \theta_{k+1}^*, \dots, \theta_T^*) \right| \\ & = \sum_{s_t \in \mathcal{S}, i_t \in \mathcal{N}} \left| \rho_t(s_t|\pi_\theta) \sum_{s_{t+1} \in \mathcal{S}} P_t(s_{t+1}|s_t, i_t) \left(Q_{t+1}^{\pi_\alpha}(s_{t+1}, \pi_{t+1}(s_{t+1}|\theta_{t+1})) - Q_{t+1}^{\pi_\beta}(s_{t+1}, \pi_{t+1}(s_{t+1}|\theta_{t+1})) \right) \right| \\ & \quad \dots \\ & = \sum_{s_t \in \mathcal{S}, i_t \in \mathcal{N}} \left| \rho_t(s_t|\pi_\theta) \sum_{s_{t+1} \in \mathcal{S}} P_t(s_{t+1}|s_t, i_t) \sum_{i_{t+1} \in \mathcal{N}} \theta_{t+1}(s_{t+1}, i_{t+1}) \sum_{s_{t+2} \in \mathcal{S}} P_t(s_{t+2}|s_{t+1}, i_{t+1}) \dots \right. \\ & \quad \left. \sum_{i_{k-1} \in \mathcal{N}} \theta_{k-1}(s_{k-1}, i_{k-1}) \sum_{s_k \in \mathcal{S}} P_{k-1}(s_k|s_{k-1}, i_{k-1}) \underbrace{\left(Q_k^*(s_k, \pi_k(s_k|\theta_k)) - Q_k^*(s_k, \pi_k(s_k|\theta_k^*)) \right)}_{(I)} \right|. \end{aligned}$$

Term (I) is exactly the difference in optimal Q-value functions at period $k > t$. From the assumption that $\theta_t(s, i) \geq \underline{p}$ for any $s \in \mathcal{S}$, $i \in \mathcal{N}$, and $t \in [T]$, *sequential decomposition inequalities* hold.

LEMMA 5. *Sequential decomposition inequalities hold with $M_g = 1/\underline{p}$.*

We have checked all the required conditions in Theorem 1. The following theorem establishes the KL condition of the policy gradient optimization problem for entropy-regularized tabular MDPs.

THEOREM 2. *Consider the entropy-regularized tabular MDPs. The policy gradient optimization problem satisfies the KL condition with the corresponding KL constant $\mu_l =$*

$$\frac{\lambda^3 \underline{p}^2}{en^3 T^2 \left(T\bar{C} + \frac{\lambda}{n\underline{p}} + T\lambda \log\left(\frac{1}{n\underline{p}}\right) \right)^2}.$$

$$l(\theta) - l(\theta^*) \leq \frac{1}{2\mu_l} \min_{g \in \partial \delta_\Pi(\pi)} \left\| \nabla l(\theta) + g \right\|_F^2, \quad \forall \pi \in \Pi.$$

Proof of Theorem 2 Plugging Lemma 3, 4, and 5 into Theorem 1 can get the results. \square

Leveraging the KL condition, one can establish a linear convergence rate for exact policy gradient methods and $\tilde{\mathcal{O}}(\epsilon^{-1})$ sample complexity for stochastic policy gradient methods to achieve an optimal policy by Lemma 1. This is essentially the same as the result in Bhandari and Russo (2024), which demonstrated the gradient domination of $l(\theta)$, implying a linear convergence rate for exact policy gradient methods.

5. Linear Quadratic Regulator

The Linear Quadratic Regulator (LQR) is one of the fundamental problems in the optimal control theory. It seeks an optimal control for a linear dynamic system, in which the state's dynamic is a linear function of the current state and action while incurring a quadratic cost. We present the problem following most of the terminologies in Fazel et al. (2018) and Hambly et al. (2021), while keeping some differences to maintain consistency in Section 2.

5.1. Problem Formulation

Consider an MDP with $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, C, T, \rho)$. We aim to solve the following optimization problem over a finite time horizon T :

$$\min_{\{a_t\}_{t=0}^{T-1}} \mathbb{E} \left[\sum_{t=0}^{T-1} (s_t^\top Q_t s_t + a_t^\top R_t a_t) + s_T^\top Q_T s_T \mid s_0 \sim \rho \right], \quad (10)$$

such that for all $t = 0, \dots, T-1$,

$$s_{t+1} = A s_t + B a_t + w_t. \quad (11)$$

Here $s_t \in \mathbb{R}^m$ is the state of the system with an initial distribution ρ , $a_t \in \mathbb{R}^n$ is the action at period t , and $\{w_t\}_{t=0}^{T-1}$ are independent and identical distributed random variables with zero mean that are independent from the initial distribution. Dynamic (11) captures the transition kernel P_t and the cost function is

$$C_t(s_t, a_t) = s_t^\top Q_t s_t + a_t^\top R_t a_t, \quad \forall t = 0, \dots, T-1.$$

Our analysis can deal with time-dependent parameters in (11), i.e., A_t and B_t . For simplicity, we assume that these parameters are time-independent. To ensure the problem is well-defined, we make the following assumptions:

ASSUMPTION 3. *Assume that the following assumptions hold.*

1. **(Cost Parameters)** Assume that $Q_t \in \mathbb{R}^{m \times m}$ and $R_t \in \mathbb{R}^{n \times n}$ are positive definite matrices for all $t = 0, \dots, T-1$. Furthermore, define $\underline{\sigma}_Q$ and $\underline{\sigma}_R$ as the smallest eigenvalue of $\{Q_t\}_{t=0}^{T-1}$ and $\{R_t\}_{t=0}^{T-1}$ respectively:

$$\begin{aligned} \underline{\sigma}_Q &= \min_{t=0, \dots, T-1} \sigma_{\min}(Q_t) > 0, \\ \underline{\sigma}_R &= \min_{t=0, \dots, T-1} \sigma_{\min}(R_t) > 0. \end{aligned}$$

2. (**Randomness**) Assume that the second moments of x_0 and $\{w_t\}_{t=0}^{T-1}$ are finite. Furthermore, we assume that $\mathbb{E}[x_0 x_0^\top]$ and $\mathbb{E}[w_t w_t^\top]$ are positive definite matrices for all $t = 0, \dots, T-1$.

Similarly, define $\underline{\sigma}_X$ as the smallest eigenvalue of $\mathbb{E}[s_t s_t^\top]$:

$$\underline{\sigma}_X = \min_{t=0, \dots, T} \sigma_{\min}(\mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)}[s_t s_t^\top]).$$

Then, we have the following result that shows the well-definedness of the state covariance matrix.

LEMMA 6 (**Hambly et al. (2021), Lemma 3.2**). Suppose that Assumption 3 holds, the second moment of s_t is positive definite for any $t = 0, \dots, T$ under any policy $\pi \in \Pi$. Therefore, we have $\underline{\sigma}_X > 0$.

This lemma is essential for the landscape characterization and the global convergence of policy gradient methods. In this setting, it's well-known that the linear policy $\pi_t(s_t|\theta_t) = \theta_t s_t$ is optimal for some unknown parameters $\theta_t \in \mathbb{R}^{n \times m}$ (Bertsekas 1995). The policy gradient objective function using linear policies is

$$l(\theta) = \mathbb{E} \left[\sum_{t=0}^{T-1} (s_t^\top Q_t s_t + (\theta_t s_t)^\top R_t (\theta_t s_t)) + s_T^\top Q_T s_T \middle| s_0 \sim \rho \right]$$

with $s_{t+1} = (A + B\theta_t)s_t + w_t$. When the linear system is unstable, complexity in Hambly et al. (2021) has an exponential dependence on T . To stabilize the system, one needs to restrict to the set $\{\theta : \lambda_{\max}(A + B\theta_t) \leq 1, \forall t = 0, \dots, T-1\}$. However, this set is unbounded and non-convex, making the analysis difficult.

In the following, we only consider the landscape of $l(\theta)$ within a convex and compact set Θ such that $\lambda_{\max}(A + B\theta_t) \leq \|A + B\theta_t\|_2 \leq 1$ for all $\theta_t \in \Theta_t$ and $0 \leq t \leq T-1$. Furthermore, we denote $\bar{\sigma}_\Theta$ as the largest spectral norm of $\theta_t \in \Theta_t$ for all $0 \leq t \leq T-1$, i.e.,

$$\bar{\sigma}_\Theta = \max_{t=0, \dots, T-1} \left\{ \max\{\|\theta_t\|_2 : \theta_t \in \Theta_t\} \right\}.$$

For the positive definite matrices $Q_t, R_t, \mathbb{E}[x_0 x_0^\top], \mathbb{E}[w_t w_t^\top]$, we define $\bar{\sigma}_Q, \bar{\sigma}_R, \bar{\sigma}_X, \bar{\sigma}_W$ as the upper bound on their eigenvalues for all $t = 0, \dots, T-1$, respectively. In addition, we assume $\lambda_{\max}(Q_T) \leq \bar{\sigma}_Q$ as well.

5.2. KL Condition of Policy Gradient Objectives

To establish the KL condition of $l(\theta)$, we aim to verify all the conditions in Theorem 1. The differentiability condition and the KL condition of expected optimal Q -value functions come from the definition of Q -value functions. Given any policy $\pi_\theta \in \Pi_\Theta$, Q -value functions satisfy (1):

$$Q_t^{\pi_\theta}(s_t, \pi_t(s_t|\theta_t)) = \underbrace{s_t^\top Q_t s_t + s_t^\top \theta_t^\top R_t \theta_t s_t}_{(I)} + \underbrace{\mathbb{E}_{w_t} \left[V_{t+1}^{\pi_\theta}((A + B\theta_t)s_t + w_t) \right]}_{(II)}.$$

(I) is a quadratic function with a positive definite matrix R_t , thereby is continuously differentiable. For (II), the value function is continuously differentiable by mathematical induction. Since the composition of a continuously differentiable function and a linear function is continuously differentiable, we have the continuous differentiability of (II). If we plug π^* into (II), we get an explicit expression of the optimal Q-value function Q_t^* . Bertsekas (1995) demonstrated the convexity of V_t^* by mathematical induction. Therefore, (II) is a convex function of θ_t , which implies the strong convexity (and the KL condition) of the optimal Q-value function combined with the strong convexity of (I).

LEMMA 7. *Suppose that Assumption 3 holds. The expected Q-value function is continuously differentiable on Θ_t . Furthermore, the expected optimal Q-value function $\mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^*(s_t, \pi_t(s_t|\theta_t)) \right]$ satisfies the KL condition on Θ_t with KL constant $2\underline{\sigma}_X \underline{\sigma}_R$.*

To validate other conditions in Theorem 1, we need an explicit expression of the policy gradient $\nabla l(\theta)$. Hambly et al. (2021) established the formulation of the Q-value function and the policy gradient $\nabla l(\theta)$ through a recursive form. First, let us define $P_t(\theta)$ by the following recursive equations:

$$P_t(\theta) := Q_t + \theta_t^\top R_t \theta_t + (A + B\theta_t)^\top P_{t+1}(\theta) (A + B\theta_t), \quad \forall t = 0, \dots, T-1, \quad (12)$$

The boundary condition is $P_T(\theta) = Q_T$. In addition, define

$$L_t(\theta) := L_{t+1}(\theta) + \mathbb{E}[w_t^\top P_{t+1}(\theta) w_t], \quad \forall t = 0, \dots, T-1,$$

with $L_T(\theta) = 0$, and

$$E_t(\theta) := (R_t + B^\top P_{t+1}(\theta) B) \theta_t + B^\top P_{t+1}(\theta) A, \quad \forall t = 0, \dots, T-1. \quad (13)$$

The subsequent proposition presents an explicit formulation of the Q-value function and $\nabla l(\theta)$.

PROPOSITION 2 (Hambly et al. (2021), Lemma 3.5). *The Q-value function has an explicit expression:*

$$\mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^{\pi_\theta}(s_t, \pi_t(s_t|\theta_t)) \right] = \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} [s_t^\top P_t(\theta) s_t] + L_t.$$

Furthermore, the policy gradient objective function $l(\theta)$ has the following gradient form:

$$\nabla_t l(\theta) = 2E_t(\theta) \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} [s_t s_t^\top].$$

From the expression of the policy gradient $\nabla l(\theta)$, we verify the *bounded gradients* condition by showing the boundedness of P_t . It utilizes the stability of the linear system and the compactness of the feasible region Θ_t . The following lemma establishes a formal result.

LEMMA 8. Suppose that Assumption 3 holds. The policy gradient objective function has bounded gradients, i.e., $\|\nabla_t l(\theta)\|_F \leq G$ for any $0 \leq t \leq T-1$. Furthermore, G is polynomial in the model parameters $(m, n, T, \bar{\sigma}_Q, \bar{\sigma}_R, \bar{\sigma}_\Theta, \bar{\sigma}_X, \bar{\sigma}_W, \|B\|_2)$.

Sequential decomposition inequalities are more complicated to verify. However, the structure of the LQR problem helps to construct the relationship between the difference in gradients and the difference in optimal Q-value functions. To see this, define $\Pi_{[j_1:j_2]} := (A + B\theta_{j_2})(A + B\theta_{j_2-1}) \dots (A + B\theta_{j_1+1})(A + B\theta_{j_1})$ for $j_1 \leq j_2$. Recall the gradient formulation in Proposition 2, for any $1 \leq t < k \leq T$:

$$\begin{aligned} & \nabla_t l(\theta_{[0:k-1]}, \theta_k, \theta_{[k+1:T-1]}^*) - \nabla_t l(\theta_{[0:k-1]}, \theta_k^*, \theta_{[k+1:T-1]}^*) \\ &= 2B^\top (P_{t+1}(\theta_{[0:k-1]}, \theta_k, \theta_{[k+1:T-1]}^*) - P_{t+1}(\theta_{[0:k-1]}, \theta_k^*, \theta_{[k+1:T-1]}^*)) (A + B\theta_t) \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} [s_t s_t^\top] \\ &= 2B^\top (A + B\theta_{t+1})^\top (P_{t+2}(\theta_{[0:k-1]}, \theta_k, \theta_{[k+1:T-1]}^*) \\ &\quad - P_{t+2}(\theta_{[0:k-1]}, \theta_k^*, \theta_{[k+1:T-1]}^*)) (A + B\theta_{t+1})(A + B\theta_t) \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} [s_t s_t^\top] \\ &\quad \dots \\ &= 2B^\top \Pi_{[t+1:k-1]}^\top \left(P_k(\theta_{[0:k-1]}, \theta_k, \theta_{[k+1:T-1]}^*) - P_k(\theta_{[0:k-1]}, \theta_k^*, \theta_{[k+1:T-1]}^*) \right) \Pi_{[t:k-1]} \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} [s_t s_t^\top]. \end{aligned}$$

The second equation uses the update (12). Next, we proceed to the difference in optimal Q-value functions. Utilizing the explicit expression of the Q-value function in Proposition 2, we conclude that

$$\begin{aligned} & \mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} \left[Q_k^*(s_k, \pi_k(s_k|\theta_k)) \right] - \mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} \left[Q_k^*(s_k, \pi_k(s_k|\theta_k^*)) \right] \\ &= \mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} \left[s_k^\top (P_k(\theta_{[0:k-1]}, \theta_k, \theta_{[k+1:T-1]}^*) - P_k(\theta_{[0:k-1]}, \theta_k^*, \theta_{[k+1:T-1]}^*)) s_k \right] \\ &= \text{Tr} \left((P_k(\theta_{[0:k-1]}, \theta_k, \theta_{[k+1:T-1]}^*) - P_k(\theta_{[0:k-1]}, \theta_k^*, \theta_{[k+1:T-1]}^*)) \mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} [s_k s_k^\top] \right). \end{aligned}$$

Both the difference in gradients of $l(\theta)$ and the difference in expected optimal Q-value functions have the component $P_k(\theta_{[0:k-1]}, \theta_k, \theta_{[k+1:T-1]}^*) - P_k(\theta_{[0:k-1]}, \theta_k^*, \theta_{[k+1:T-1]}^*)$. Leveraging this, we can prove the *Sequential Decomposition* condition under some mild assumptions.

LEMMA 9. Suppose that Assumption 3 holds. The Sequential Decomposition condition holds with $M_g > 0$. Furthermore, M_g is polynomial in the model parameters $(m, n, T, \bar{\sigma}_Q, \bar{\sigma}_R, \bar{\sigma}_\Theta, \bar{\sigma}_X, \bar{\sigma}_W, \|B\|_2)$.

With the *Differentiability* condition (Lemma 7), *KL Condition of Optimal Q-value Function* (Lemma 7), *Bounded Gradient* condition (Lemma 8), and *Sequential Decomposition* condition (Lemma 9), we are ready to demonstrate the KL condition of the policy gradient optimization problem.

THEOREM 3. Consider the LQR problem. Suppose that Assumption 3 holds. The policy gradient objective function $l(\theta)$ satisfies the KL condition on Θ :

$$l(\theta) - l(\theta^*) \leq \frac{1}{2\mu_l} \min_{g \in \partial \delta_\Theta(\theta)} \sum_{t=0}^{T-1} \|\nabla_t l(\theta) + g_t\|_F^2, \quad \forall \theta \in \Theta,$$

where $g = (g_0, \dots, g_{T-1})$. In addition, the reciprocal of KL constant μ_l is polynomial in the model parameters $(m, n, T, \bar{\sigma}_Q, \bar{\sigma}_R, \underline{\sigma}_R^{-1}, \bar{\sigma}_\Theta, \bar{\sigma}_X, \underline{\sigma}_X^{-1}, \bar{\sigma}_W, \|B\|_2)$.

Proof of Theorem 3 Plugging Lemma 7, 8, and 9 into Theorem 1 can get the results. \square

Following the proof of Hambly et al. (2021, Theorem 3.3), we can establish a linear convergence rate of exact policy gradient methods. It's worth noting that the KL constant in Theorem 3 is different from that in Hambly et al. (2021, Lemma 3.6). They fully explored the special structure of the LQR problem and used an important fact that expected Q-value functions induced by *any* policy π_θ is quadratic. To illustrate the general applicability of our unified framework, we instead utilize the KL condition of expected *optimal* Q-value functions, as this is a less restrictive condition shared with other operations problems. Interestingly, through our framework, one can easily extend the same results to the setting with general strongly convex cost functions, linear transition kernels, and linear policy classes.

6. Inventory Models

This section demonstrates how to validate the assumptions in Theorem 1 to establish the KL condition for the policy gradient optimization problem of the multi-period inventory system with Markov-modulated demands, where unsatisfied demands are backlogged. One can extend the result to the lost sales model and derive a sample complexity with the same order as the objectives in the two settings, which exhibit the same landscape but with a constant difference.

Many works in inventory control assume that random demands are independent across time. However, this assumption is often unrealistic in the real world, and the underlying demand process might be correlated, i.e., economic conditions and seasons affect the random demands of different periods. To capture correlations, some literature models the underlying demand process by an exogenous discrete-time, discrete-state Markov chain (Song and Zipkin 1993). We briefly state the problem formulation and validate the KL condition of the policy gradient optimization problem using a state-dependent base-stock policy class.

6.1. Problem Formulation

Consider a MDP framework with $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, C, T, \rho)$. At the beginning of period t , the decision-maker observes state $s_t = (x_t, i_t)$ and determines the order quantity $a_t \geq 0$, where x_t represents the current inventory level and i_t is the state of the world. The inventory level x_1 follows an initial distribution ρ . The replenishment immediately raises the inventory level to $y_t = x_t + a_t \geq x_t$. Subsequently, the decision-maker observes a random demand D_t whose distribution depends on the current state of the world i_t . Finally, the inventory level at the beginning of the next period

follows the linear transition kernel $x_{t+1} = y_t - D_t = x_t + a_t - D_t$, where the negative inventory level represents backlogged demands.

Suppose the exogenous Markov chain has a finite state space \mathcal{I} . In state $i_t \in \mathcal{I}$, the random demand D_t follows a cumulative distribution function $P_D(\cdot|i_t)$. State i moves to the next state j with probability $p(j|i) \in [0, 1]$ and $\sum_{j \in \mathcal{I}} p(j|i) = 1$ for any $i \in \mathcal{I}$. As a finite-state time-homogeneous Markov Chain has at least one stationary distribution, let us pick $\nu \in \mathbb{R}^{|\mathcal{I}|}$ as one of its stationary distributions. Assume that initial state i_1 follows the stationary distribution ν , e.g., $i_t \sim \nu$ for any $t \in [T]$.

When the on-hand inventory level exceeds the realized demand D_t , it incurs a holding cost of $h_t \geq 0$ per unit. Otherwise, insufficient inventory causes a backlogging cost of $b_t \geq 0$ per unit. Let L_t denote the expected holding and backlogging cost for period t , which is a convex function of the order-up-to level y_t :

$$L_t(y_t|i_t) = \mathbb{E}_{D_t \sim P_D(\cdot|i_t)} \left[h_t(y_t - D_t)^+ + b_t(y_t - D_t)^- \right].$$

For simplicity, we omit the ordering cost. Our analysis can easily accommodate positive linear ordering costs, which can be subsumed in holding and backlogging costs. The per-period cost in the MDP framework is $C_t(s_t, a_t) = L_t(x_t + a_t|i_t)$. We aim to identify an ordering policy that minimizes the total expected cost over T periods:

$$\min_{\{a_t\}_{t=1}^T} \mathbb{E} \left[\sum_{t=1}^T L_t(x_t + a_t|i_t) \middle| x_1 \sim \rho, i_1 \sim \nu \right], \quad (14)$$

where $x_{t+1} = x_t + a_t - D_t$, $D_t \sim P_D(\cdot|i_t)$, and $i_{t+1} \sim p(\cdot|i_t)$. One can use dynamic programming to reformulate (14). Let V_t^* denote the cost-to-go function, which starts at the beginning of period t . It satisfies the Bellman optimality equation (2):

$$\begin{aligned} V_t^*(x_t, i_t) &= \min_{a_t \geq 0} \left\{ L_t(x_t + a_t|i_t) + \mathbb{E}_{i_{t+1} \sim p(\cdot|i_t), D_t \sim P_D(\cdot|i_t)} \left[V_{t+1}^*(x_t + a_t - D_t, i_{t+1}) \right] \right\} \\ &= \min_{y_t \geq x_t} \left\{ L_t(y_t|i_t) + \mathbb{E}_{i_{t+1} \sim p(\cdot|i_t), D_t \sim P_D(\cdot|i_t)} \left[V_{t+1}^*(y_t - D_t, i_{t+1}) \right] \right\}, \end{aligned}$$

with $V_{T+1}^*(\cdot) = 0$. Let us define

$$f_t(y_t|i_t) := L_t(y_t|i_t) + \mathbb{E}_{i_{t+1} \sim p(\cdot|i_t), D_t \sim P_D(\cdot|i_t)} \left[V_{t+1}^*(y_t - D_t, i_{t+1}) \right]. \quad (15)$$

Song and Zipkin (1993) demonstrated the convexity of $V_t^*(x_t, i_t)$ with respect to x_t and the convexity of $f_t(y_t|i_t)$ with respect to y_t for any $i_t \in \mathcal{I}$ by mathematical induction. Given the convexity of $f_t(y_t|i_t)$, they proved that a state-dependent base-stock policy is optimal.

Thus, in the following, we focus on the state-dependent base-stock policy class. Specifically, for each state of the exogenous Markov Chain and period t , we define the corresponding base-stock

level $\theta_{t,i} \in \mathbb{R}$. Let $\theta_t \in \mathbb{R}^{|\mathcal{I}|}$ to denote the vector of all the base-stock levels in period t with $\theta_{t,i}$ as its i -th component. Given the current inventory level x_t and the state of the world i_t , the decision-maker orders $\pi_t(x_t, i_t | \theta_t) = (\theta_{t,i_t} - x_t)^+$. For the optimal base-stock levels θ_t^* , its i -th component $\theta_{t,i}^*$ is a minimizer of $f_t(\cdot | i)$.

Given the state-dependent base-stock policy class, we can express the policy gradient objective function of all the base-stock levels $\theta = (\theta_1, \dots, \theta_T) \in \mathbb{R}^{T \times |\mathcal{I}|}$ by

$$l(\theta) = \mathbb{E} \left[\sum_{t=1}^T L_t(x_t \vee \theta_{t,i_t} | i_t) \middle| x_1 \sim \rho, i_1 \sim \nu \right],$$

with $x_{t+1} = x_t \vee \theta_{t,i_t} - D_t$, $D_t \sim P_D(\cdot | i_t)$, and $i_{t+1} \sim p(\cdot | i_t)$. Here \vee denotes a component-wise max operator. In the remaining part of this section, we focus on the landscape of $l(\theta)$ over a compact convex set $\Theta = \Theta_1 \times \dots \times \Theta_T$ with $\Theta_t = \{\theta_t : \theta_{t,i} \in [0, B], \forall i \in \mathcal{I}\}$. This is reasonable in practice since one can treat $B \geq 0$ as the capacity of the warehouse. The policy class is $\Pi_\Theta := \{\pi(x, i, t | \theta) = x \vee \theta_{t,i} : \theta_{t,i} \in [0, B]\}$.

6.2. Nonconvex Landscape

In this subsection, we demonstrate the KL condition of the policy gradient objective function $l(\theta)$ by verifying the conditions in Theorem 1. To establish the KL condition, we rely on the following assumptions.

ASSUMPTION 4. *Assume that all the following assumptions hold.*

1. *The initial inventory level x_1 is independent of the exogenous Markov chain. In addition, x_1 is independent of the random demands.*
2. *The initial cumulative distribution function of x_1 is L_ρ -Lipschitz continuous. Furthermore, the cumulative distribution function $P_D(\cdot | i)$ is L_D -Lipschitz continuous for all $i \in \mathcal{I}$.*
3. *There exists a positive constant α_D such that $P_D(B | i) \leq 1 - \alpha_D$ for any $i \in \mathcal{I}$.*
4. *The probability density functions of random demands at each period are uniformly bounded below by $\mu_D > 0$ on $[0, B]$.*

Assumption 4.1 is standard in the setting of Markov modulated demands (Song and Zipkin 1993, Chen and Song 2001). It allows the correlation of random demands across different periods through the exogenous Markov Chain, which is less restrictive than the standard independent assumption in the literature (Huh and Rusmevichientong 2014, Cheung and Simchi-Levi 2019). Assumption 4.2 ensures the *differentiability* condition, which is valid for many commonly used distributions, e.g., uniform, exponential, and Erlang distributions. Assumption 4.3 is crucial for demonstrating the *KL condition of expected optimal Q-value functions* as it excludes suboptimal stationary points

(see the proof of Lemma 10). Assumption 4.4 leads to strongly convex costs $L_t(\cdot|i_t)$ and holds for numerous distributions, such as uniform, exponential, and Erland distributions.

For any policy $\pi_\theta \in \Pi_\Theta$, the expected Q-value functions satisfy the Bellman equation (1):

$$\begin{aligned} & \mathbb{E}_{(x_t, i_t) \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^{\pi_\theta}(x_t, i_t, \pi_t(x_t, i_t|\theta_t)) \right] \\ &= \mathbb{E}_{(x_t, i_t) \sim \rho_t(\cdot|\pi_\theta)} \left[L_t(x_t \vee \theta_{t, i_t} | i_t) + \mathbb{E}_{i_{t+1} \sim p(\cdot|i_t), D_t \sim P_D(\cdot|i_t)} \left[V_{t+1}^{\pi_\theta}(x_t \vee \theta_{t, i_t} - D_t, i_{t+1}) \right] \right]. \end{aligned} \quad (16)$$

We express the expected Q-value function by $\mathbb{E}_{(x_t, i_t) \sim \rho_t(\cdot|\pi_\theta)} [h(x_t \vee \theta_{t, i_t} | i_t)]$, where

$$h(y_{t, i_t} | i_t) = L_t(y | i_t) + \mathbb{E}_{i_{t+1} \sim p(\cdot|i_t), D_t \sim P_D(\cdot|i_t)} [V_{t+1}^{\pi_\theta}(y - D_t, i_{t+1})].$$

Clearly, $h(\cdot|i_t)$ is continuously differentiable by backward mathematical induction. The cumulative distribution function $\rho_t(\cdot|\pi_\theta)$ is continuous. Thus, the expected Q-value function is also continuously differentiable. Furthermore, replacing π_θ with π^* in (16) gives the expression of expected optimal Q-value functions:

$$F_t(\theta_t) := \mathbb{E}_{(x_t, i_t) \sim \rho_t(\cdot|\pi_\theta)} [Q_t^*(x_t, i_t, \pi_t(x_t, i_t|\theta_t))] = \mathbb{E}_{(x_t, i_t) \sim \rho_t(\cdot|\pi_\theta)} [f_t(x_t \vee \theta_{t, i_t} | i_t)]. \quad (17)$$

It's not straightforward to verify the KL condition of $F_t(\theta_t)$, and we present a proof sketch to understand why this condition holds. From (17), the suboptimality gap of $F_t(\cdot)$ can be upper bounded using the suboptimality gap of $f_t(\cdot|i_t)$. Since $f_t(\cdot|i_t)$ is strongly convex, its suboptimality gap is dominated by its gradient norm. Applying (17) again, we establish the connection between the gradients of $F_t(\cdot)$ and $f_t(\cdot|i_t)$ and prove the KL condition of $F_t(\theta_t)$. For more rigorous proof, we refer readers to Appendix D.

LEMMA 10. *Suppose that Assumption 4 holds. The expected Q-value function is continuously differentiable on Θ_t . Furthermore, the expected optimal Q-value function $F_t(\theta_t)$ satisfies the KL condition on Θ_t with KL constant $\mu_Q = \mu_D \alpha_D^2 \min_{i \in \mathcal{I}} \{\nu_i\}$.*

Similar to previous applications, verifying remaining conditions requires a gradient expression of the policy gradient objective function. Let $V_t^{\pi_\theta}$ denote the value function, which starts at the beginning of period t and follows policy π_θ . It satisfies the Bellman equation (1):

$$V_t^{\pi_\theta}(x_t, i_t) = L_t(x_t \vee \theta_{t, i_t} | i_t) + \mathbb{E}_{i_{t+1} \sim p(\cdot|i_t), D_t \sim P_D(\cdot|i_t)} [V_{t+1}^{\pi_\theta}(x_t \vee \theta_{t, i_t} - D_t, i_{t+1})],$$

with $V_{T+1}^{\pi_\theta}(\cdot) = 0$. The subsequent proposition presents the result in a recursive form.

PROPOSITION 3 (Policy Gradient Expression). *For any $\theta \in \Theta$ and $t \in [T]$, the partial derivatives of value functions satisfy the following recursive form:*

$$\nabla_x V_t^{\pi_\theta}(x_t, i_t) = \mathbf{1}(x_t \geq \theta_{t, i_t}) \times \left(L'_t(x_t | i_t) + \sum_{i_{t+1} \in \mathcal{I}} p(i_{t+1} | i_t) \mathbb{E}_{D_t \sim P_D(\cdot|i_t)} [\nabla_x V_{t+1}^{\pi_\theta}(x_t - D_t, i_{t+1})] \right), \quad (18)$$

where $\nabla_x V_{T+1}^{\pi_\theta}(\cdot, \cdot) = 0$ and $\nabla_x V_t^{\pi_\theta}(x_t, i_t)$ represent the partial derivative of x_t . Additionally, the policy gradient objective function $l(\theta)$ has the following gradient form for any $t \in [T]$ and $i \in \mathcal{I}$:

$$\begin{aligned} \frac{\partial}{\partial \theta_{t,i}} l(\theta) = & \mathbb{E}_{(x_t, i_t) \sim \rho_t(\cdot | \pi_\theta)} \left[\mathbf{1}(i_t = i, \theta_{t,i} \geq x_t) \right. \\ & \times \left(L'_t(\theta_{t,i} | i) + \sum_{i_{t+1} \in \mathcal{I}} p(i_{t+1} | i) \mathbb{E}_{D_t \sim P_D(\cdot | i)} \left[\nabla_x V_{t+1}^{\pi_\theta}(\theta_{t,i} - D_t, i_{t+1}) \right] \right) \Big]. \end{aligned}$$

Given the gradient formulation in Proposition 3, one can easily construct a stochastic policy gradient estimator using samples from trajectories and verify that Assumptions 1 and 2 holds with all parameters scaling polynomially in the planning horizon. Due to the page limits, we omit the proof. Leveraging Proposition 3, the following two lemmas verify the *bounded gradient* condition and the *sequential decomposition inequality*.

LEMMA 11. Suppose that Assumption 4 holds. The policy gradient objective function $l(\theta)$ has bounded gradients for any $\theta \in \Theta$ and $t \in [T]$:

$$\|\nabla_{\theta_t} l(\theta)\|_2 \leq \max_{t \in \{1, \dots, T\}} \{\max\{h_t, b_t\}\} T.$$

LEMMA 12. Suppose that Assumption 4 holds. The sequential decomposition inequality holds for any $\theta \in \Theta$ and $1 \leq t < k \leq T$, i.e.,

$$\left\| \nabla_{\theta_t} l(\theta_{[1:k-1]}, \theta_k, \theta_{[k+1:T]}^*) - \nabla_{\theta_t} l(\theta_{[1:k-1]}, \theta_k^*, \theta_{[k+1:T]}^*) \right\| \leq \frac{L_D}{\alpha_D} (F_k(\theta_k) - F_k(\theta_k^*)).$$

Plugging Lemmas 10, 11, and 12, we have the main result of this section.

THEOREM 4. Suppose that Assumption 4 holds. The policy gradient objective function $l(\theta)$ of the multi-period inventory system with Markov-modulated demand satisfies the KL condition on Θ with KL constant

$$\mu_l = \frac{\mu_D^3 \alpha_D^8 \min_{i \in \mathcal{I}} \{v_i\}^3}{e L_D^2 \max_{t \in \{1, \dots, T\}} \{\max\{h_t, b_t\}\}^2 T^4}.$$

More specifically, we have

$$l(\theta) - l(\theta^*) \leq \frac{1}{2\mu_l} \min_{g \in \partial \delta_\Theta(\theta)} \|\nabla l(\theta) + g\|_2^2, \quad \forall \theta \in \Theta.$$

REMARK 4. The KL constant in Theorem 4 has a dependence on $\min_{i \in \mathcal{I}} \{v_i\}$. Since $\sum_{i \in \mathcal{I}} \nu_i = 1$, the smallest probability is at most the order of $1/|\mathcal{I}|$.

Leveraging the KL condition, we establish the global convergence of stochastic policy gradient methods by Lemma 1. The sample complexity required for achieving an ϵ -optimal state-dependent base-stock policy is $\tilde{\mathcal{O}}(\epsilon^{-1} \text{poly}(T))$. This is the first sample complexity result for the multi-period

inventory system with Markov-modulated demands in the literature. As a byproduct, we improve the sample complexity for stochastic gradient methods to solve the inventory system with independent demands, which is a special case with no exogenous Markov chain ($|Z| = 1$). Our sample complexity admits a polynomial dependence on the time horizon, representing a significant improvement compared to the exponential dependence in [Huh and Rusmevichientong \(2014\)](#) for a biased stochastic gradient method. We remark that [Huh and Rusmevichientong \(2014\)](#) assumes the convexity of cost-go-functions, whereas Theorem 1 assumes the KL Condition of expected optimal Q-value functions (which is a generalization of strong convexity). While the analysis in [Huh and Rusmevichientong \(2014\)](#) can be extended to the case with strongly convex cost-to-go functions, it remains unclear whether the exponential dependence on T can be improved.

7. Stochastic Cash Balance Problem

The cash balance problem originally refers to a cost minimization problem when a firm has to decide how much cash to hold to meet the transaction requirements over a finite planning horizon. It can also model inventory management with rented equipment ([Whisler 1967](#), [Chen and Simchi-Levi 2009](#)). In this section, we will briefly describe the problem formulation under the inventory setting and validate the KL condition of the objectives for policy gradient methods using a two-sided base-stock policy class.

7.1. Problem Formulation

The problem formulation of the stochastic cash balance problem is similar to the multi-period inventory system. We use the same notations in Section 6. Unlike the class inventory system where the decision-maker can only raise the inventory level, the stochastic cash balance problem allows the decision-maker to reduce the inventory level. Let $a_t = y_t - x_t$ denote the ordering ($a_t > 0$) or return ($a_t < 0$) quantity. The transaction cost is a piecewise linear function:

$$c(y, x) = \begin{cases} k(y - x), & y \geq x, \\ q(x - y), & y < x, \end{cases}$$

with $k + q \geq 0$. The assumption for $k + q \geq 0$ implies that the unit refund can not exceed the unit ordering cost. Therefore, the transaction cost function is jointly convex in (x, y) .

Additionally, the random demand D_t can be positive or negative. Negative demand means the decision-maker receives more returns from customers than their purchase. For simplicity, we assume that demands among different periods are independent and identically distributed. One can easily extend our results to the setting when random demands are not identically distributed. Let F_D represent the cumulative distribution function of random demands. All other settings are the same as the inventory model in Section 6.

For convenience, we recall some useful notations. Let L_t denote the expected cost for the period t as a function of the inventory level y_t after the ordering and return decisions:

$$L_t(y_t) = \mathbb{E}_{D_t} \left[h_t(y_t - D_t)^+ + b_t(y_t - D_t)^- \right].$$

Then the per-period cost in the MDP framework is $C_t(s_t, a_t) = c(s_t + a_t, s_t) + L_t(s_t + a_t)$. The objective of the decision-maker is to minimize the total expected costs over the finite horizon T :

$$\min_{\{a_t\}_{t=1}^T} \mathbb{E} \left[\sum_{t=1}^T (c(s_t + a_t, s_t) + L_t(s_t + a_t)) \middle| s_1 \sim \rho \right], \quad (19)$$

with $s_{t+1} = s_t + a_t - D_t$. Like the classic stochastic inventory system, one can present the optimization problem (19) by a dynamic program. Let V_t^* be the cost-to-go function which starts at the beginning of period t , it satisfies the Bellman equation (2):

$$V_t^*(s_t) = \min_{y_t} \left\{ c(y_t, s_t) + L_t(y_t) + \mathbb{E}_{D_t} [V_{t+1}^*(y_t - D_t)] \right\},$$

with $V_{T+1}^*(\cdot) = 0$. By mathematical induction, we can easily show the convexity of cost-to-go functions since the transaction cost function is jointly convex in (x, y) and $L_t(y_t)$ is a convex function. Define $f_t(x) = L_t(x) + \mathbb{E}_{D_t} [V_{t+1}^*(x - D_t)]$, we can rewrite the dynamic programming recursion:

$$V_t^*(s_t) = \min \left\{ \min_{y_t \geq s_t} \{k(y_t - s_t) + f_t(y_t)\}, \min_{y_t \leq s_t} \{q(s_t - y_t) + f_t(y_t)\} \right\}.$$

Whisler (1967) and Eppen and Fama (1969) studied the stochastic cash balance problem and proved the optimality of the two-sided base-stock policy. That is, at period t there exists two parameters $\underline{\theta}_t$ and $\bar{\theta}_t$ with $\underline{\theta}_t \leq \bar{\theta}_t$, such that the optimal inventory level $y_t(s_t)$ satisfies:

$$y_t(s_t) = \begin{cases} \underline{\theta}_t, & s_t \leq \underline{\theta}_t, \\ s_t, & \underline{\theta}_t < s_t < \bar{\theta}_t, \\ \bar{\theta}_t, & s_t \geq \bar{\theta}_t. \end{cases}$$

Based on the convexity of the cost-to-go functions, if we let

$$\begin{cases} \underline{f}_t(y_t) = ky_t + f_t(y_t) = ky_t + L_t(y_t) + \mathbb{E}_{D_t} [V_{t+1}^*(y_t - D_t)], \\ \bar{f}_t(y_t) = -qy_t + f_t(y_t) = -qy_t + L_t(y_t) + \mathbb{E}_{D_t} [V_{t+1}^*(y_t - D_t)], \end{cases}$$

then both $\underline{f}_t(y_t)$ and $\bar{f}_t(y_t)$ are convex functions with minimizers $\underline{\theta}_t^*$ and $\bar{\theta}_t^*$. It is well known that the optimal parameters for the two-sided base-stock policy are $\underline{\theta}_t^*$ and $\bar{\theta}_t^*$ (Whisler 1967, Eppen and Fama 1969).

Let $\theta_t = (\underline{\theta}_t, \bar{\theta}_t)$ denote the parameters at period t . Then we can rewrite the two-sided base-stock policy by $a_t = \pi_t(s_t | \theta_t) = (s_t \vee \underline{\theta}_t) \wedge \bar{\theta}_t - s_t$. Thus the policy gradient objective function is

$$l(\theta) = \mathbb{E} \left[\sum_{t=1}^T \left(c((s_t \vee \underline{\theta}_t) \wedge \bar{\theta}_t, s_t) + L_t((s_t \vee \underline{\theta}_t) \wedge \bar{\theta}_t) \right) \middle| s_1 \sim \rho \right],$$

with $s_{t+1} = (s_t \vee \underline{\theta}_t) \wedge \bar{\theta}_t - D_t$. In the remaining part of this section, we only analyze the nonconvex landscape of $l(\theta)$ over a convex set $\Theta = \Theta_1 \times \dots \times \Theta_T$ with $\Theta_t = \{(\underline{\theta}_t, \bar{\theta}_t) : \underline{B} \leq \underline{\theta}_t \leq \bar{\theta}_t \leq \bar{B}\}$. This is reasonable in practice since one can treat \bar{B} as the capacity of the warehouse, and the lower bound \underline{B} ensures the firms will never hold too many backlogged demands.

7.2. Nonconvex Landscape

Like Section 6, we establish the KL condition for the policy gradient objective function by verifying the conditions in Theorem 1. Before showing the main results, we make some assumptions.

ASSUMPTION 5. *Assume that all the following assumptions hold.*

1. *The initial state s_1 and random demands D_t in different periods t are independent of each other.*
2. *The initial distribution ρ is L_ρ -Lipschitz continuous. Furthermore, the cumulative distribution function F_D of random demands is L_D -Lipschitz continuous.*
3. *There exists a positive constant α_D such that $\mathbb{P}(D_t \geq \bar{B}) \geq \alpha_D$ and $\mathbb{P}(D_t \leq \underline{B}) \geq \alpha_D$ for any $t \in [T]$.*
4. *The probability density function of the random demand D_t is bounded below by $\mu_D > 0$ on $[\underline{B}, \bar{B}]$.*

Assumption 5 is very similar to Assumption 4 except that Assumption 5.3 further requires $\mathbb{P}(D_t \leq \underline{B}) \geq \alpha_D$, which plays the same role for excluding suboptimal stationary points.

For any policy $\pi_\theta \in \Pi_\Theta$, the expected Q-value functions satisfy the Bellman equation (1):

$$\begin{aligned} & \mathbb{E}_{s_t \sim \rho_t(\cdot | \pi_\theta)} \left[Q_t^{\pi_\theta}(s_t, \pi_t(s_t | \theta_t)) \right] \\ &= \mathbb{E}_{s_t \sim \rho_t(\cdot | \pi_\theta)} \left[c((s_t \vee \underline{\theta}_t) \wedge \bar{\theta}_t, s_t) + L_t((s_t \vee \underline{\theta}_t) \wedge \bar{\theta}_t) + \mathbb{E}_{D_t} \left[V_{t+1}^{\pi_\theta}((s_t \vee \underline{\theta}_t) \wedge \bar{\theta}_t - D_t) \right] \right]. \end{aligned}$$

Same as Section 6, the continuous differentiability of the expected Q-value function comes from the continuity of the cumulative distribution function of s_t , which holds under Assumption 5.2. Recall the definition $f_t(x) = L_t(x) + \mathbb{E}_{D_t} [V_{t+1}^*(x - D_t)]$. We get the expression of expected optimal Q-value functions by replacing π_θ with π^* :

$$F_t(\theta_t) := \mathbb{E}_{s_t \sim \rho_t(\cdot | \pi_\theta)} \left[Q_t^*(s_t, \pi_t(s_t | \theta_t)) \right] = \mathbb{E}_{s_t \sim \rho_t(\cdot | \pi_\theta)} \left[c((s_t \vee \underline{\theta}_t) \wedge \bar{\theta}_t, s_t) + f_t((s_t \vee \underline{\theta}_t) \wedge \bar{\theta}_t) \right].$$

The structure of the expected optimal Q-value function is close to that for the inventory system in Section 6. The following lemma establishes the *KL condition for the expected optimal Q-value function* by a similar proof.

LEMMA 13. *Suppose that Assumption 5 holds. The expected Q-value function is continuously differentiable on Θ_t for any $t \in [T]$. In addition, the expected optimal Q-value function $F_t(\theta_t)$ satisfies the KL condition on Θ_t with KL constant $\mu_D \alpha_D^2$ for any $t \in [T]$.*

Again, we need an explicit expression of the gradient $\nabla l(\theta)$ to validate the remaining conditions. Let $V_t^{\pi_\theta}$ be the value function that starts at the beginning of period t . It satisfies the Bellman equation (1):

$$\begin{aligned} V_t^{\pi_\theta}(s_t) &= c(\pi_t(s_t|\theta_t), s_t) + L_t(\pi_t(s_t|\theta_t)) + \mathbb{E}_{D_t} \left[V_{t+1}^{\pi_\theta}(\pi_t(s_t|\theta_t) - D_t) \right] \\ &= c((s_t \vee \underline{\theta}_t) \wedge \bar{\theta}_t, s_t) + L_t((s_t \vee \underline{\theta}_t) \wedge \bar{\theta}_t) + \mathbb{E}_{D_t} \left[V_{t+1}^{\pi_\theta}((s_t \vee \underline{\theta}_t) \wedge \bar{\theta}_t - D_t) \right], \end{aligned}$$

with $V_{T+1}^{\pi_\theta}(\cdot) = 0$. The subsequent proposition presents a gradient formulation for $l(\theta)$.

PROPOSITION 4 (Policy Gradient Expression). *For any $\theta \in \Theta$ and $t \in [T]$, the derivatives of value functions satisfy the following recursive form:*

$$(V_t^{\pi_\theta})'(s_t) = k\mathbf{1}(s_t \leq \underline{\theta}_t) - q\mathbf{1}(s_t \geq \bar{\theta}_t) + \mathbf{1}(\underline{\theta}_t < s_t < \bar{\theta}_t) \times \left(L_t'(s_t) + \mathbb{E}_{D_t} [(V_{t+1}^{\pi_\theta})'(s_t - D_t)] \right) \quad (20)$$

with $(V_{T+1}^{\pi_\theta})(\cdot) = 0$. Additionally, the policy gradient objective function $l(\theta)$ has the following gradient form for any $t \in [T]$:

$$\begin{cases} \frac{\partial}{\partial \underline{\theta}_t} l(\theta) = \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[\mathbf{1}(\underline{\theta}_t \geq s_t) \times \left(k + L_t'(\underline{\theta}_t) + \mathbb{E}_{D_t} [(V_{t+1}^{\pi_\theta})'(\underline{\theta}_t - D_t)] \right) \right], \\ \frac{\partial}{\partial \bar{\theta}_t} l(\theta) = \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[\mathbf{1}(\bar{\theta}_t \leq s_t) \times \left(-q + L_t'(\bar{\theta}_t) + \mathbb{E}_{D_t} [(V_{t+1}^{\pi_\theta})'(\bar{\theta}_t - D_t)] \right) \right]. \end{cases}$$

With the explicit expression of the gradient, we can verify the *bounded Gradient* condition and the *sequential Decomposition Inequality* by the following two lemmas.

LEMMA 14. *Suppose that Assumption 5 holds. It follows that the policy gradient objective function $l(\theta)$ has bounded gradients for any $\theta \in \Theta$ and $t \in [T]$:*

$$\|\nabla_{\theta_t} l(\theta)\|_2 \leq 2(k + |q| + \max_{t \in \{1, \dots, T\}} \{\max\{h_t, b_t\}\})T.$$

LEMMA 15. *Suppose that Assumption 5 holds. For any $\theta \in \Theta$ and $1 \leq t < k \leq T$, sequential decomposition inequalities hold, i.e.,*

$$\begin{aligned} & \left\| \nabla_{\theta_t} l(\theta_{[1:k-1]}, \theta_k, \theta_{[k+1:T]}^*) - \nabla_{\theta_t} l(\theta_{[1:k-1]}, \theta_k^*, \theta_{[k+1:T]}^*) \right\|_2 \\ & \leq \frac{L_D}{\alpha_D} \left(\mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^*(s_t, \pi_t(s_t|\theta_t)) \right] - \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^*(s_t, \pi_t(s_t|\theta_t^*)) \right] \right). \end{aligned}$$

Equipped with *KL condition of Optimal Q-value Function* (Lemma 13), *Bounded Gradient* condition (Lemma 14), and *Sequential Decomposition* inequality (Lemma 15), we can demonstrate the KL condition of the policy gradient objective function $l(\theta)$ by applying Theorem 1.

THEOREM 5. *Suppose that Assumption 5 holds. The policy gradient objective function $l(\theta)$ of the stochastic cash balance problem satisfies the KL condition on Θ with KL constant*

$$\mu_l = \frac{\mu_D^3 \alpha_D^8}{16eL_D^2 (k + |q| + \max_{t \in \{1, \dots, T\}} \{\max\{h_t, b_t\}\})^2 T^4}.$$

Proof of Theorem 5 Plugging Lemmas 14, 13, and 15 into Theorem 1 can get the result. \square

Similar to the inventory system in Section 6, we establish an $\tilde{O}(\epsilon^{-1}\text{poly}(T))$ sample complexity of stochastic policy gradient methods converging to globally optimal policies by Lemma 1. To the best of our knowledge, this is the first sample complexity result for data-driven methods solving the stochastic cash balance problem. Additionally, one can check that KL condition holds for the stochastic cash balance problem with Markov-modulated demands as well.

8. Conclusion

This work provides a framework with several easily verified conditions to establish the KL condition for policy gradient optimization problems of finite-horizon MDPs with general state and action spaces. Despite nonconvexity, the KL condition guarantees a linear convergence rate for exact policy gradient methods and an $\tilde{O}(\epsilon^{-1})$ sample complexity for stochastic policy gradient methods. Our framework covers a broad range of control and operations models, including entropy-regularized tabular MDPs, LQR problems, multi-period inventory systems with Markov-modulated demands, and stochastic cash balance problems. Furthermore, we establish the first sample complexity solving the stochastic cash balance problem and multi-period inventory system with Markov-modulated demand allowing backorders, and an extension to the lost sales model. The complexity has a polynomial instead of an exponential dependence on the planning horizon.

Our work opens up several directions for future research. Firstly, in many of the applications discussed, one can further explore the structural properties to build a more precise characterization of the KL constant and reduce the dependence on T . Secondly, our results build upon the KL condition of expected optimal Q-value functions, requiring strongly convex per-period costs. It remains interesting to further generalize the results to general convex per-period costs for applications like inventory models. Adding regularization could be one potential solution, yet it might deteriorate the dependence on accuracy ϵ . Finally, it is interesting to explore the applicability of our developed framework to other applications.

Acknowledgments

Yifan Hu is supported by the Swiss National Science Foundation under NCCR Automation, grant agreement 51NF40_180545.

References

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *The Journal of Machine Learning Research*, 22(1):4431–4506, 2021. (Cited on pages 1, 4, and 16.)

-
- Mihai Anitescu. Degenerate nonlinear programming with a quadratic growth condition. *SIAM Journal on Optimization*, 10(4):1116–1135, 2000. (Cited on page 4.)
- Hedy Attouch and Jérôme Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116:5–16, 2009. (Cited on page 4.)
- Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013. (Cited on pages 4, 10, 12, and 39.)
- Richard Bellman. On the theory of dynamic programming. *Proceedings of the national Academy of Sciences*, 38(8):716–719, 1952. (Cited on page 9.)
- Aharon Ben-Tal and Dick Den Hertog. Hidden conic quadratic representation of some nonconvex quadratic optimization problems. *Mathematical Programming*, 143:1–29, 2014. (Cited on page 3.)
- Aharon Ben-Tal and Marc Teboulle. Hidden convexity in some nonconvex quadratically constrained quadratic programming. *Mathematical Programming*, 72(1):51–63, 1996. (Cited on page 3.)
- Aharon Ben-Tal, Dick Den Hertog, and Monique Laurent. Hidden convexity in partially separable optimization. *Available at SSRN 1865208*, 2011. (Cited on page 3.)
- D.P. Bertsekas. *Dynamic Programming and Optimal Control*. Number v. 1 in Athena scientific optimization and computation series. Athena Scientific, 1995. ISBN 9781886529120. (Cited on pages 2, 20, 21, and 48.)
- Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *Operations Research*, 2024. (Cited on pages 1, 3, 4, 5, 13, 16, 17, 19, and 48.)
- Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007. (Cited on pages 3 and 4.)
- Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4):2563–2578, 2022. (Cited on page 4.)
- Fangruo Chen and Jing-Sheng Song. Optimal policies for multiechelon inventory problems with markov-modulated demand. *Operations Research*, 49(2):226–234, 2001. (Cited on page 25.)
- Xin Chen and Xiangyu Gao. Stochastic optimization with decisions truncated by positively dependent random variables. *Operations Research*, 67(5):1321–1327, 2019. (Cited on page 3.)
- Xin Chen and David Simchi-Levi. A new approach for the stochastic cash balance problem with fixed costs. *Probability in the Engineering and Informational Sciences*, 23(4):545–562, 2009. (Cited on page 28.)
- Xin Chen, Xiangyu Gao, and Zhan Pang. Preservation of structural properties in optimization with decisions truncated by random variables and its applications. *Operations Research*, 66(2):340–357, 2018. (Cited on page 3.)

- Xin Chen, Niao He, Yifan Hu, and Zikun Ye. Efficient algorithms for a class of stochastic hidden convex optimization and its applications in network revenue management. *Operations Research*, 2024. (Cited on pages 3 and 51.)
- Yiwei Chen and Cong Shi. Network revenue management with online inverse batch gradient descent method. *Production and Operations Management*, 32(7):2123–2137, 2023. (Cited on page 3.)
- Wang Chi Cheung and David Simchi-Levi. Sampling-based approximation schemes for capacitated stochastic inventory control models. *Mathematics of Operations Research*, 44(2):668–692, 2019. (Cited on pages 5 and 25.)
- Gary D Eppen and Eugene F Fama. Cash balance and simple dynamic portfolio problems with proportional costs. *International Economic Review*, 10(2):119–133, 1969. (Cited on page 29.)
- Ilyas Fatkhullin, Anas Barakat, Anastasia Kireeva, and Niao He. Stochastic policy gradient methods: Improved sample complexity for fisher-non-degenerate policies. In *International Conference on Machine Learning*, pages 9827–9869. PMLR, 2023a. (Cited on page 4.)
- Ilyas Fatkhullin, Niao He, and Yifan Hu. Stochastic optimization under hidden convexity. *arXiv preprint arXiv:2401.00108*, 2023b. (Cited on page 3.)
- Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International conference on machine learning*, pages 1467–1476. PMLR, 2018. (Cited on pages 5 and 19.)
- Qi Feng and J George Shanthikumar. Supply and demand functions in inventory models. *Operations Research*, 66(1):77–91, 2018. (Cited on page 3.)
- Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *International Conference on Machine Learning*, pages 2160–2169. PMLR, 2019. (Cited on page 16.)
- Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. (Cited on page 11.)
- Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU press, 2013. (Cited on page 49.)
- Xiao-Yue Gong and David Simchi-Levi. Bandits atop reinforcement learning: Tackling online inventory models with cyclic demands. *Management Science*, 2023. (Cited on page 6.)
- Ben Hambly, Renyuan Xu, and Huining Yang. Policy gradient methods for the noisy linear quadratic regulator over a finite horizon. *SIAM Journal on Control and Optimization*, 59(5):3359–3391, 2021. (Cited on pages 3, 5, 19, 20, 21, and 23.)
- Yinbin Han, Meisam Razaviyayn, and Renyuan Xu. Policy gradient converges to the globally optimal policy for nearly linear-quadratic regulators. *arXiv preprint arXiv:2303.08431*, 2023. (Cited on page 5.)
- Yifan Hu, Jie Wang, Xin Chen, and Niao He. Multi-level monte-carlo gradient methods for stochastic optimization with biased oracles. *arXiv preprint arXiv:2408.11084*, 2024. (Cited on page 3.)

-
- Woonghee Tim Huh and Paat Rusmevichientong. Online sequential optimization with biased gradients: theory and applications to censored demand. *INFORMS Journal on Computing*, 26(1):150–159, 2014. (Cited on pages 6, 25, and 28.)
- Jemin Hwangbo, Joonho Lee, Alexey Dosovitskiy, Dario Bellicoso, Vassilios Tsounis, Vladlen Koltun, and Marco Hutter. Learning agile and dynamic motor skills for legged robots. *Science Robotics*, 4(26): eaau5872, 2019. (Cited on page 1.)
- Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002. (Cited on pages 14 and 44.)
- Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 795–811. Springer, 2016. (Cited on pages 3, 4, 5, 10, 11, 14, and 38.)
- Sara Klein, Simon Weissmann, and Leif Döring. Beyond stationarity: Convergence analysis of stochastic softmax policy gradient methods. *arXiv preprint arXiv:2310.02671*, 2023. (Cited on pages 4 and 16.)
- Anton J Kleywegt, Alexander Shapiro, and Tito Homem-de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on optimization*, 12(2):479–502, 2002. (Cited on page 5.)
- Sumit Kunnumkal and Huseyin Topaloglu. Using stochastic approximation methods to compute optimal base-stock levels in inventory control problems. *Operations Research*, 56(3):646–664, 2008. (Cited on page 6.)
- Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. In *Annales de l’institut Fourier*, volume 48, pages 769–783, 1998. (Cited on pages 2, 3, and 4.)
- Guanghui Lan. Policy optimization over general state and action spaces. *arXiv preprint arXiv:2211.16715*, 2022. (Cited on page 4.)
- Guanghui Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, 198(1):1059–1106, 2023. (Cited on pages 4 and 16.)
- Retsef Levi, Robin O Roundy, and David B Shmoys. Provably near-optimal sampling-based policies for stochastic inventory control models. *Mathematics of Operations Research*, 32(4):821–839, 2007. (Cited on page 5.)
- Eitan Levin, Joe Kileel, and Nicolas Boumal. The effect of smooth parametrizations on nonconvex optimization landscapes. *Mathematical Programming*, pages 1–49, 2024. (Cited on page 3.)
- Feng-Yi Liao, Lijun Ding, and Yang Zheng. Error bounds, pl condition, and quadratic growth for weakly convex functions, and linear convergences of proximal point methods. In *6th Annual Learning for Dynamics & Control Conference*, pages 993–1005. PMLR, 2024. (Cited on page 4.)

- Stanislaw Lojasiewicz. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 117(87-89):2, 1963. (Cited on pages 2 and 3.)
- Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993. (Cited on page 4.)
- Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International conference on machine learning*, pages 6820–6829. PMLR, 2020. (Cited on page 4.)
- Sentao Miao and Yining Wang. Network revenue management with nonparametric demand learning: \sqrt{T} -regret and polynomial dimension dependency. *Available at SSRN 3948140*, 2021. (Cited on page 3.)
- B Sh Mordukhovich. Maximum principle in the problem of time optimal response with nonsmooth constraints. *Journal of Applied Mathematics and Mechanics*, 40(6):960–969, 1976. (Cited on page 10.)
- Ion Necoara, Yu Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175:69–107, 2019. (Cited on page 4.)
- Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009. (Cited on page 11.)
- Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999. (Cited on pages 2 and 41.)
- Boris Teodorovich Polyak et al. Gradient methods for minimizing functionals. *Zhurnal vychislitel’noi matematiki i matematicheskoi fiziki*, 3(4):643–653, 1963. (Cited on page 3.)
- Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014. (Cited on pages 2, 7, and 9.)
- Hanzhang Qin, David Simchi-Levi, and Li Wang. Data-driven approximation schemes for joint pricing and inventory control models. *Management Science*, 68(9):6591–6609, 2022. (Cited on page 5.)
- Hanzhang Qin, David Simchi-Levi, and Ruihao Zhu. Sailing through the dark: Provably sample-efficient inventory control. *Available at SSRN 4652347*, 2023. (Cited on page 6.)
- R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009. (Cited on page 10.)
- David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International conference on machine learning*, pages 387–395. PMLR, 2014. (Cited on pages 14, 43, 44, 54, and 61.)
- David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016. (Cited on page 1.)
- Jing-Sheng Song and Paul Zipkin. Inventory control in a fluctuating demand environment. *Operations Research*, 41(2):351–370, 1993. (Cited on pages 23, 24, and 25.)

-
- Ronald J Stern and Henry Wolkowicz. Indefinite trust region subproblems and nonsymmetric eigenvalue perturbations. *SIAM Journal on Optimization*, 5(2):286–313, 1995. (Cited on page 3.)
- Yue Sun and Maryam Fazel. Learning optimal controllers by policy gradient: Global optimality via convex parameterization. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 4576–4581. IEEE, 2021. (Cited on page 3.)
- Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999. (Cited on pages 9, 17, and 46.)
- William D Whisler. A stochastic inventory model for rented equipment. *Management Science*, 13(9):640–647, 1967. (Cited on pages 28 and 29.)
- Lin Xiao. On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282):1–36, 2022. (Cited on page 4.)
- Yaqi Xie, Will Ma, and Linwei Xin. Vc theory for inventory policies. *arXiv preprint arXiv:2404.11509*, 2024. (Cited on page 6.)
- Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. *Advances in Neural Information Processing Systems*, 33:4572–4583, 2020. (Cited on page 3.)
- Kairen Zhang, Xiangyu Gao, Zhanyue Wang, and Sean Zhou. Sampling-based approximation for serial multi-echelon inventory system. *Available at SSRN 3859856*, 2022. (Cited on page 6.)

Online Appendices

Appendix A: Omitted Proofs in Section 3

A.1. Strong Convexity, Gradient Dominance, and KL condition

DEFINITION 4 (GRADIENT DOMINANCE). Consider a convex and compact set $\mathcal{X} \subseteq \mathbb{R}^d$ and a differentiable function f . Suppose that f^* is the optimal objective value $f^* := \min_{x \in \mathcal{X}} f(x)$. The function f is said to be (α, μ) -gradient dominated over \mathcal{X} if there exists constants $\alpha > 0$ and $\mu \geq 0$ such that

$$f(x) - f^* \leq \max_{x' \in \mathcal{X}} \left\{ \alpha \langle \nabla f(x), x - x' \rangle - \frac{\mu}{2} \|x - x'\|_2^2 \right\}, \quad \forall x \in \mathcal{X}. \quad (21)$$

One important property is that the $(\alpha, 0)$ -gradient dominance (for $\mu > 0$) and (α, μ) -gradient dominance are generalizations of convexity and μ -strong convexity, respectively. In particular, if f is convex (or μ -strongly convex), then f is $(1, 0)$ -gradient dominated (or $(1, \mu)$ -gradient dominated) by definition.

LEMMA 16 (Karimi et al. (2016)). Consider a convex and compact set $\mathcal{X} \subseteq \mathbb{R}^n$. If a function $f : \mathcal{X} \rightarrow \mathbb{R}$ is (α, μ) -gradient dominated over \mathcal{X} , then f satisfies the KL condition with constant μ/α^2 over \mathcal{X} .

REMARK 5. Karimi et al. (2016, Appendix G) analyzed the equivalence between the proximal-PL condition and the KL condition. Since the proximal-PL condition differs from the (α, μ) -gradient dominance slightly, one can apply the same proof to prove Lemma 16.

COROLLARY 1. Consider a convex and compact set $\mathcal{X} \subseteq \mathbb{R}^n$. If the function $f : \mathcal{X} \rightarrow \mathbb{R}$ is μ -strongly convex over \mathcal{X} , then f satisfies the KL condition over \mathcal{X} with constant μ .

Proof of Corollary 1 From the μ -strong convexity of f , we have

$$f(x) - f^* \leq \langle \nabla f(x), x - x^* \rangle - \frac{\mu}{2} \|x - x^*\|_2^2 \leq \max_{x' \in \mathcal{X}} \left\{ \langle \nabla f(x), x - x' \rangle - \frac{\mu}{2} \|x - x'\|_2^2 \right\}$$

Here the last inequality uses the fact that $x^* \in \mathcal{X}$. Then applying Lemma 16 completes the proof.

□

A.2. No Suboptimal Stationary Points

Proof of Proposition 1 Suppose $\bar{x} \in \mathcal{X}$ satisfies the first-order necessary optimality condition. We have:

$$\langle \nabla f(\bar{x}), \bar{x} - x \rangle \leq 0, \quad \forall x \in \mathcal{X}.$$

Recall that the subdifferential of $\delta_{\mathcal{X}}(\bar{x})$ is the normal cone of \mathcal{X} at \bar{x} , i.e.,

$$\partial \delta_{\mathcal{X}}(\bar{x}) = \{g \mid \langle g, x - \bar{x} \rangle \leq 0, \forall x \in \mathcal{X}\}.$$

Therefore, we have $-\nabla f(\bar{x}) \in \partial\delta_{\mathcal{X}}(\bar{x})$, which implies

$$f(\bar{x}) - f^* \leq \frac{1}{2\mu} \min_{g \in \partial\delta_{\mathcal{X}}(\bar{x})} \|\nabla f(\bar{x}) + g\|_2^2 \stackrel{(a)}{\leq} \frac{1}{2\mu} \|\nabla f(\bar{x}) - \nabla f(\bar{x})\|_2^2 = 0, \quad \forall \bar{x} \in \mathcal{X}.$$

Here inequality (a) holds because $-\nabla f(\bar{x}) \in \partial\delta_{\mathcal{X}}(\bar{x})$. Since $f(\bar{x}) \geq f^*$ by definition, it follows that $f(\bar{x}) = f^*$ and \bar{x} is a global optimal point. \square

A.3. Convergence Rate under the KL Condition

Proof of Lemma 1 The non-asymptotic convergence result of projected gradient descent can be found in Attouch et al. (2013). As for the projected stochastic gradient descent, the proof follows the framework in Attouch et al. (2013) with an extension to the stochastic setting. From the optimality condition of $x_{k+1} = \arg \min_{x \in \mathcal{X}} \|x - \gamma_k \nabla \hat{f}(x_k)\|_2^2$, we have

$$\|x_{k+1} - x_k\|_2^2 + \langle x_{k+1} - x_k, \gamma_k \nabla \hat{f}(x_k) \rangle \leq 0. \quad (22)$$

From the smoothness of f , we conclude that

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\leq \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|_2^2 \\ &= \langle \nabla \hat{f}(x_k), x_{k+1} - x_k \rangle + \langle \nabla f(x_k) - \nabla \hat{f}(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|_2^2 \\ &\stackrel{(a)}{\leq} \langle \nabla f(x_k) - \nabla \hat{f}(x_k), x_{k+1} - x_k \rangle - \frac{L}{2} \|x_{k+1} - x_k\|_2^2 \\ &\stackrel{(b)}{\leq} \frac{1}{L} \|\nabla f(x_k) - \nabla \hat{f}(x_k)\|_2^2 + \frac{L}{4} \|x_{k+1} - x_k\|_2^2 - \frac{L}{2} \|x_{k+1} - x_k\|_2^2 \\ &= \frac{1}{L} \|\nabla f(x_k) - \nabla \hat{f}(x_k)\|_2^2 - \frac{L}{4} \|x_{k+1} - x_k\|_2^2. \end{aligned} \quad (23)$$

Here inequality (a) uses (22) and $\gamma_k = \frac{1}{L}$, and (b) uses the inequality $2\langle a, b \rangle \leq \|a\|_2^2 + \|b\|_2^2$. From the optimality condition of the optimization problem that defines the projection operator,

$$x_k - \gamma_k \nabla \hat{f}(x_k) - x_{k+1} \in \partial\delta_{\mathcal{X}}(x_{k+1}).$$

Since the subdifferentials of $\delta_{\mathcal{X}}$ is a normal cone, we have

$$\frac{x_k - x_{k+1}}{\gamma_k} - \nabla \hat{f}(x_k) \in \partial\delta_{\mathcal{X}}(x_{k+1}).$$

Thus, we get

$$\begin{aligned} \min_{g \in \partial\delta_{\mathcal{X}}(x_{k+1})} \|\nabla f(x_{k+1}) + g\|_2 &\leq \left\| \frac{x_k - x_{k+1}}{\gamma_k} + \nabla f(x_{k+1}) - \nabla \hat{f}(x_k) \right\|_2 \\ &\stackrel{(a)}{\leq} \left\| \frac{x_k - x_{k+1}}{\gamma_k} \right\|_2 + \|\nabla f(x_{k+1}) - \nabla f(x_k)\|_2 + \|\nabla f(x_k) - \nabla \hat{f}(x_k)\|_2 \\ &\stackrel{(b)}{\leq} 2L \|x_{k+1} - x_k\|_2 + \|\nabla f(x_k) - \nabla \hat{f}(x_k)\|_2. \end{aligned}$$

Inequality (a) uses the triangle inequality, and inequality (b) uses the smoothness of f and $\gamma_k = \frac{1}{L}$. By the KL Condition, we have

$$\begin{aligned}
f(x_{k+1}) - f^* &\leq \frac{1}{2\mu} \min_{g \in \partial \delta_{\mathcal{X}}(x_{k+1})} \|\nabla f(x_{k+1}) + g\|_2^2 \\
&\leq \frac{1}{2\mu} \left(2L\|x_{k+1} - x_k\|_2 + \|\nabla f(x_k) - \nabla \hat{f}(x_k)\|_2 \right)^2 \\
&\stackrel{(a)}{\leq} \frac{1}{2\mu} \left(8L^2\|x_{k+1} - x_k\|_2^2 + 2\|\nabla f(x_k) - \nabla \hat{f}(x_k)\|_2^2 \right) \\
&\stackrel{(b)}{\leq} \frac{16L}{\mu} (f(x_k) - f(x_{k+1})) + \frac{17}{\mu} \|\nabla f(x_k) - \nabla \hat{f}(x_k)\|_2^2.
\end{aligned}$$

Here (a) uses the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, and inequality (b) uses (23). Notice that $x_k = x_k(\xi_{[k-1]})$ is a function of the $\xi_{[k-1]} = (\xi_1, \dots, \xi_{k-1})$. For simplicity, we will use $\mathbb{E}[f(x_{k+1})]$ and $\mathbb{E}[f(x_k)]$ to denote $\mathbb{E}_{\xi_{[k]}}[f(x_{k+1})]$ and $\mathbb{E}_{\xi_{[k-1]}}[f(x_k)]$ respectively. Then, taking the expectation on both sides and using the assumption that $\mathbb{E}_{\xi_k} \|\nabla \hat{f}(x_k) - \nabla f(x_k)\|_2^2 \leq \sigma^2/N$, we have

$$\mathbb{E}[f(x_{k+1})] - f^* \leq \frac{16L}{\mu} \left(\mathbb{E}[f(x_k)] - \mathbb{E}[f(x_{k+1})] \right) + \frac{17\sigma^2}{\mu N}.$$

Rearranging the terms, we have

$$\mathbb{E}[f(x_{k+1})] - f^* \leq \left(1 - \frac{\mu}{16L + \mu} \right) \left(\mathbb{E}[f(x_k)] - f^* \right) + \frac{17\sigma^2}{(16L + \mu)N}.$$

Taking the telescoping sum, we have

$$\begin{aligned}
\mathbb{E}[f(x_k)] - f^* &\leq \left(1 - \frac{\mu}{16L + \mu} \right)^k \left(\mathbb{E}[f(x_0)] - f^* \right) + \frac{17\sigma^2}{(16L + \mu)N} \sum_{l=0}^{k-1} \left(1 - \frac{\mu}{16L + \mu} \right)^l \\
&\leq \left(1 - \frac{\mu}{16L + \mu} \right)^k \left(\mathbb{E}[f(x_0)] - f^* \right) + \frac{17\sigma^2}{\mu N}.
\end{aligned}$$

Here the last inequality holds because $\sum_{l=0}^{k-1} (1-q)^l \leq 1/q$ for any $q \in (0, 1)$. \square

A.4. KL Condition in Policy Gradient Formulation

We first prove a technical lemma (Lemma 2), which is useful for our main results in Theorem 1.

Proof of Lemma 2 Let us define $u_t^2 := \sum_{l=t}^T X_l^2$, $v_t^2 := \sum_{l=t}^T Y_l^2$. Without loss of generality, we assume that $v_t > 0$ for any $t \in [T]$. Otherwise, there exists $t \in [T]$ such that $Y_l = 0$ for any $t \leq l \leq T$, which implies $X_l = Y_l = 0$ for any $t \leq l \leq T$ by (9). Therefore, discarding these terms will not affect the final result.

Furthermore, we define $f_t := u_t^2/v_t^2$ and a series $\{\delta_t\}_{t=1}^T$ where

$$\delta_t := (1 + 2Y_t M_g + M_g^2 \delta_{t+1} v_{t+1}^2) \delta_{t+1},$$

and $\delta_T = 1$. We use backward mathematical induction to show that $f_t \leq \delta_t$ for any $t \in [T]$.

Induction Base: Since $f_T = u_T^2/v_T^2 = X_T^2/Y_T^2 = 1$, we have $f_T = \delta_T = 1$.

Induction Step: Suppose that $f_{t+1} \leq \delta_{t+1}$ holds. We have

$$\begin{aligned}
 f_t &= \frac{u_t^2}{v_t^2} = \frac{u_{t+1}^2 + X_t^2}{v_{t+1}^2 + Y_t^2} \stackrel{(a)}{\leq} \frac{u_{t+1}^2 + (M_g u_{t+1}^2 + Y_t)^2}{v_{t+1}^2 + Y_t^2} = \frac{u_{t+1}^2 + Y_t^2 + M_g^2 u_{t+1}^4 + 2Y_t M_g u_{t+1}^2}{v_{t+1}^2 + Y_t^2} \\
 &\stackrel{(b)}{=} \frac{(1 + 2Y_t M_g + M_g^2 f_{t+1} v_{t+1}^2) f_{t+1} v_{t+1}^2 + Y_t^2}{v_{t+1}^2 + Y_t^2} \\
 &\stackrel{(c)}{\leq} (1 + 2Y_t M_g + M_g^2 \delta_{t+1} v_{t+1}^2) \delta_{t+1} \\
 &= \delta_t.
 \end{aligned}$$

Here inequality (a) uses (9), equation (b) comes from the definition $f_{t+1} = u_{t+1}^2/v_{t+1}^2$, and inequality (c) utilizes the induction assumption $f_{t+1} \leq \delta_{t+1}$ and the fact that δ_t is non-increasing in t with $\delta_t \geq \delta_T = 1$. Therefore, by mathematical induction, we can conclude that $f_t \leq \delta_t$ for any $t \in [T]$. By definition, we have the following inequalities,

$$\delta_t = \prod_{l=t}^{T-1} (1 + 2Y_l M_g + M_g^2 \delta_{l+1} v_{l+1}^2) \stackrel{(a)}{\leq} \prod_{l=t}^{T-1} (1 + 2Y_l M_g + M_g^2 \delta_{t+1} v_t^2), \quad \forall t \in [T-1].$$

Inequality (a) holds because δ_t and v_t^2 are non-increasing in t . Consider the following optimization problem:

$$\begin{aligned}
 &\max_{Y_t, \dots, Y_{T-1}} \quad \prod_{l=t}^{T-1} (1 + 2Y_l M_g + M_g^2 \delta_{t+1} v_t^2), \\
 &\text{s.t.} \quad \sum_{l=t}^{T-1} Y_l^2 \leq v_t^2, \\
 &\quad Y_l \geq 0, \quad \forall t \leq l \leq T-1.
 \end{aligned}$$

Since all the terms are non-negative, the optimal solution satisfies the equation $\sum_{l=t}^{T-1} Y_l^2 = v_t^2$. Otherwise, we can increase one of Y_l for a larger objective. By the KKT condition (Nocedal and Wright 1999), the optimal solution has an explicit form $(Y_l^*)^2 = \frac{1}{T-t} v_t^2$ for any $t \leq l \leq T-1$. This implies

$$\delta_t \leq (1 + 2M_g \frac{1}{\sqrt{T-t}} v_t + M_g^2 \delta_{t+1} v_t^2)^{T-t}, \quad \forall t \in [T-1]. \quad (24)$$

Next, we consider the following two cases:

1. Suppose that there exists $\tilde{t} \in [T-1]$ such that $2M_g v_{\tilde{t}} \frac{1}{\sqrt{T-\tilde{t}}} + M_g^2 v_{\tilde{t}}^2 e > \frac{1}{T-\tilde{t}}$. Since $v_t \geq 0$ for all $t \in [T-1]$, we can conclude that

$$v_{\tilde{t}} > \frac{1}{\sqrt{T-\tilde{t}} M_g} \times \frac{\sqrt{e+1}-1}{e}.$$

Since v_t is non-increasing and $v_1 \geq v_{\tilde{t}}$, it holds that

$$f_1 = \frac{u_1^2}{v_1^2} \leq \frac{T G^2}{v_{\tilde{t}}^2} \leq 4e M_g^2 G^2 T^2.$$

The first inequality holds as $u_1^2 = \sum_{t=1}^T X_t^2 \leq \sum_{t=1}^T G^2 = T G^2$.

2. Suppose that for any $t \in [T-1]$, we have $2M_g v_t \frac{1}{\sqrt{T-t}} + M_g^2 v_t^2 e \leq \frac{1}{T-t}$. We use backward mathematical induction to show $\delta_t \leq e$ for all $t \in [T]$.

Induction Base: We have $\delta_T = 1 < e$.

Induction Step: Suppose that $\delta_{t+1} \leq e$ holds. From (24), we have

$$\delta_t \leq (1 + 2M_g \frac{1}{\sqrt{T-t}} v_t + M_g^2 \delta_{t+1} v_t^2)^{T-t} \leq (1 + 2M_g \frac{1}{\sqrt{T-t}} v_t + M_g^2 v_t^2 e)^{T-t} \leq (1 + \frac{1}{T-t})^{T-t} \leq e.$$

Therefore, we have $\delta_t \leq e, \forall t \in [T]$ by mathematical induction, which implies $f_1 \leq \delta_1 \leq e$.

Combining these two cases, we conclude that $f_1 \leq \max\{e, 4eM_g^2 G^2 T^2\}$. Since $f_1 = u_1^2/v_1^2$, we have $\sum_{t=1}^T X_t^2 \leq \max\{e, 4eM_g^2 G^2 T^2\} \sum_{t=1}^T Y_t^2$. This completes the proof. \square

The bound in Lemma 2 depends crucially on M_g , G , and T . This dependence could influence the KL constant in Theorem 1. In what follows, we demonstrate the tightness of the dependence on M_g , G , and T in Lemma 2. Without loss of generality, we consider the case when $M_g > 1$, $G > 1$, and $M_g G \geq 4$.

We first construct two sequences $\{X_t\}_{t=1}^T$ and $\{Y_t\}_{t=1}^T$ such that f_1 is order of $M_g^2 G^2 T^2$ up to some logarithmic factors. In particular, let us define $Z := \log_2(M_g G)$ and

$$X_t = \begin{cases} \frac{2^{Z+1-t}}{M_g} & t < \lfloor Z \rfloor, \\ \frac{Z}{tM_g} & t \geq \lfloor Z \rfloor, \end{cases} \quad Y_t = \begin{cases} 0 & t < T, \\ \frac{Z}{tM_g} & t = T. \end{cases}$$

With the sequences $\{X_t\}_{t=1}^T$ and $\{Y_t\}_{t=1}^T$, we first verify the inequality (9). Since $Y_t = 0$ for any $t < T$, the inequality (9) simplifies to $X_t \leq M_g \sum_{k=t+1}^T X_k^2$. Introducing a transformation $\bar{X}_t = M_g X_t$, it is sufficient to check that if the sequence $\{\bar{X}_t\}_{t=1}^T$ satisfies

$$\bar{X}_t \leq \sum_{k=t+1}^T \bar{X}_k^2, \quad (25)$$

with $0 \leq \bar{X}_t \leq M_g G$ for any $t \in [T]$. To validate (25), it is sufficient to verify that $\bar{X}_t \leq \bar{X}_{t+1} + \bar{X}_{t+1}^2$ by mathematical induction. We consider the following three cases:

1. For any $t \geq \lfloor Z \rfloor$, we have

$$\bar{X}_{t+1} + \bar{X}_{t+1}^2 - \bar{X}_t = Z \left(\frac{1}{t+1} - \frac{1}{t} + \frac{Z}{(t+1)^2} \right) = Z \frac{Zt - t - 1}{t(t+1)^2} \stackrel{(a)}{\geq} Z \frac{t-1}{t(t+1)^2} \stackrel{(b)}{\geq} 0.$$

Here inequality (a) holds because $Z = \log_2(M_g G) \geq 2$, and inequality (b) follows since $t \geq 1$.

2. For $t = \lfloor Z \rfloor - 1$, we have

$$\bar{X}_{t+1} + \bar{X}_{t+1}^2 - \bar{X}_t = \frac{Z}{\lfloor Z \rfloor} + \left(\frac{Z}{\lfloor Z \rfloor} \right)^2 - 2^{Z-\lfloor Z \rfloor} \geq 1 + 1 - 2 = 0.$$

The inequality applies $0 \leq Z - \lfloor Z \rfloor \leq 1$.

3. For any $t < \lfloor Z \rfloor - 1$, we obtain

$$\bar{X}_{t+1} + \bar{X}_{t+1}^2 - \bar{X}_t \stackrel{(a)}{\geq} \bar{X}_{t+1}^2 - \bar{X}_t = (2^{Z-t})^2 - 2^{Z+1-t} \stackrel{(b)}{\geq} 0.$$

Here inequality (a) holds as $X_t \geq 0$ for any $t \in [T]$, and inequality (b) applies $Z \geq \lfloor Z \rfloor > t + 1$.

Therefore, the sequences $\{X_t\}_{t=1}^T$ and $\{Y_t\}_{t=1}^T$ satisfy (9). Additionally, we have

$$f_1 = \frac{u_1^2}{v_1^2} \geq \frac{X_1^2}{Y_T^2} = \frac{M_g^2 G^2 T^2}{Z^2} = \frac{M_g^2 G^2 T^2}{\log_2(M_g G)^2}.$$

This example establishes a lower bound of f_1 , indicating that the dependence on M_g , G , and T is tight up to some logarithmic factors and the bound in Lemma 2 is sharp. Next, we prove the main result (Theorem 1).

Proof of Theorem 1 The proof mainly follows the proof sketch in Section 3.3. For readers' convenience, we divide the proof into several parts.

Step 1: Bounded Gradient Mismatch Inequality. For any $g_t \in \partial \delta_{\Theta_t}(\theta_t)$, we have

$$\begin{aligned} & \left\| \nabla_{\theta_t} \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^{\pi_{\theta^*}}(s_t, \pi_t(s_t|\theta_t)) \right] + g_t \right\|_2 - \left\| \nabla_{\theta_t} \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^{\pi_\theta}(s_t, \pi_t(s_t|\theta_t)) \right] + g_t \right\|_2 \\ & \leq \left\| \nabla_{\theta_t} \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^{\pi_{\theta^*}}(s_t, \pi_t(s_t|\theta_t)) \right] + g_t - \nabla_{\theta_t} \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^{\pi_\theta}(s_t, \pi_t(s_t|\theta_t)) \right] - g_t \right\|_2 \\ & \stackrel{(a)}{=} \left\| \nabla_{\theta_t} l(\theta_{[1:t]}, \theta_{[t+1:T]}^*) - \nabla_{\theta_t} l(\theta_{[1:t]}, \theta_{[t+1:T]}) \right\|_2 \\ & \leq \sum_{k=t+1}^T \left\| \nabla_{\theta_t} l(\theta_1, \dots, \theta_k, \theta_{k+1}^*, \dots, \theta_T^*) - \nabla_{\theta_t} l(\theta_1, \dots, \theta_{k-1}, \theta_k^*, \dots, \theta_T^*) \right\|_2 \\ & \stackrel{(b)}{\leq} \sum_{k=t+1}^T M_g \left(\mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} \left[Q_k^{\pi_{\theta^*}}(s_k, \pi_k(s_k|\theta_k)) \right] - \mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} \left[Q_k^{\pi_{\theta^*}}(s_k, \pi_k(s_k|\theta_k^*)) \right] \right) \\ & \stackrel{(c)}{\leq} \sum_{k=t+1}^T \frac{M_g}{2\mu_Q} \min_{g_k \in \partial \delta_{\Theta_k}(\theta_k)} \left\| \nabla_{\theta_k} \mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} \left[Q_k^{\pi_{\theta^*}}(s_k, \pi_k(s_k|\theta_k)) \right] + g_k \right\|_2^2. \end{aligned} \tag{26}$$

Here (a) comes from the deterministic policy gradient theorem (Silver et al. 2014), (b) relies on *sequential decomposition inequalities*, and (c) uses the *KL condition of expected optimal Q-value functions*. Define

$$\begin{aligned} X_t &:= \min_{g_t \in \partial \delta_{\Theta_t}(\theta_t)} \left\| \nabla_{\theta_t} \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^{\pi_{\theta^*}}(s_t, \pi_t(s_t|\theta_t)) \right] + g_t \right\|_2, \\ Y_t &:= \min_{g_t \in \partial \delta_{\Theta_t}(\theta_t)} \left\| \nabla_{\theta_t} \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^{\pi_\theta}(s_t, \pi_t(s_t|\theta_t)) \right] + g_t \right\|_2. \end{aligned}$$

We have

$$\begin{aligned}
& X_t - Y_t \\
& \stackrel{(a)}{\leq} \max_{g_t \in \partial \delta_{\Theta_t}(\theta_t)} \left\{ \left\| \nabla_{\theta_t} \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^{\pi_{\theta^*}}(s_t, \pi_t(s_t|\theta_t)) \right] + g_t \right\|_2 - \left\| \nabla_{\theta_t} \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^{\pi_\theta}(s_t, \pi_t(s_t|\theta_t)) \right] + g_t \right\|_2 \right\} \\
& \stackrel{(b)}{\leq} \frac{M_g}{2\mu_Q} \sum_{k=t+1}^T \min_{g_k \in \partial \delta_{\Theta_k}(\theta_k)} \left\| \nabla_{\theta_k} \mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} \left[Q_k^{\pi_{\theta^*}}(s_k, \pi_k(s_k|\theta_k)) \right] + g_k \right\|_2^2 \\
& = \frac{M_g}{2\mu_Q} \sum_{k=t+1}^T X_k^2.
\end{aligned}$$

Inequality (a) utilizes the fact that $\min_{x \in \mathcal{X}} f(x) - \min_{x \in \mathcal{X}} h(x) \leq \max_{x \in \mathcal{X}} \{f(x) - h(x)\}$. Inequality (b) comes from (26). Similarly, we have $Y_t - X_t \leq \frac{M_g}{2\mu_Q} \sum_{k=t+1}^T X_k^2$. Given the fact that $X_T = Y_T$ and $X_t, Y_t \leq G$ for all $t = 1, \dots, T$, applying Lemma 2, we have that the bounded gradient mismatch condition holds:

$$\sum_{t=1}^T X_t^2 \leq \frac{eM_g^2 G^2 T^2}{\mu_Q^2} \sum_{t=1}^T Y_t^2.$$

Step 2: KL Condition of $l(\theta)$. Although Kakade and Langford (2002) only derived the Performance Difference Lemma for infinite-horizon discounted MDPs, one can use the same trick to get a similar result for finite-horizon MDPs. Therefore, we have

$$\begin{aligned}
l(\theta) - l(\theta^*) &= \sum_{t=1}^T \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^{\pi_{\theta^*}}(s_t, \pi_t(s_t|\theta_t)) - V_t^{\pi_{\theta^*}}(s_t) \right] \\
&= \sum_{t=1}^T \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^{\pi_{\theta^*}}(s_t, \pi_t(s_t|\theta_t)) - Q_t^{\pi_{\theta^*}}(s_t, \pi_t(s_t|\theta_t^*)) \right] \\
&\stackrel{(a)}{\leq} \sum_{t=1}^T \frac{1}{2\mu_Q} \min_{g_t \in \partial \delta_{\Theta_t}(\theta_t)} \left\| \nabla_{\theta_t} \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^{\pi_{\theta^*}}(s_t, \pi_t(s_t|\theta_t)) \right] + g_t \right\|_2^2 \\
&\stackrel{(b)}{\leq} \frac{eM_g^2 G^2 T^2}{2\mu_Q^3} \sum_{t=1}^T \min_{g_t \in \partial \delta_{\Theta_t}(\theta_t)} \left\| \nabla_{\theta_t} \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^{\pi_\theta}(s_t, \pi_t(s_t|\theta_t)) \right] + g_t \right\|_2^2.
\end{aligned}$$

Inequality (a) utilizes the *KL condition of optimal Q-value functions*, and inequality (b) comes from the bounded gradient mismatch inequality. Employing the deterministic policy gradient theorem in Silver et al. (2014), we have

$$\nabla_{\theta_t} l(\theta) = \nabla_{\theta_t} \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^{\pi_\theta}(s_t, \pi_t(s_t|\theta_t)) \right]. \quad (27)$$

Therefore, we obtain

$$\begin{aligned}
l(\theta) - l(\theta^*) &\leq \frac{eM_g^2 G^2 T^2}{2\mu_Q^3} \sum_{t=1}^T \min_{g_t \in \partial \delta_{\Theta_t}(\theta_t)} \left\| \nabla_{\theta_t} \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^{\pi_\theta}(s_t, \pi_t(s_t|\theta_t)) \right] + g_t \right\|_2^2 \\
&\stackrel{(a)}{=} \frac{eM_g^2 G^2 T^2}{2\mu_Q^3} \min_{g \in \partial \delta_{\Theta}(\theta)} \left\| \nabla l(\theta) + g \right\|_2^2.
\end{aligned}$$

The equality (a) comes from (27) and the fact that

$$\nabla l(\theta) = \begin{bmatrix} \nabla_{\theta_1} \mathbb{E}_{s_1 \sim \rho} \left[Q_1^{\pi_\theta}(s_1, \pi_1(s_1|\theta_1)) \right] \\ \nabla_{\theta_2} \mathbb{E}_{s_2 \sim \rho_2(\cdot|\pi_\theta)} \left[Q_2^{\pi_\theta}(s_2, \pi_2(s_2|\theta_2)) \right] \\ \vdots \\ \nabla_{\theta_T} \mathbb{E}_{s_T \sim \rho_T(\cdot|\pi_\theta)} \left[Q_T^{\pi_\theta}(s_T, \pi_T(s_T|\theta_T)) \right] \end{bmatrix}, \quad g = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_T \end{bmatrix}.$$

Therefore, $l(\theta)$ satisfies the KL condition. This completes the proof. \square

A.5. KL Constant under Weaker Assumptions

In section 3.3, we provide a standard approach that can establish a slightly weaker condition than *sequential decomposition inequalities* in Theorem 1:

$$\begin{aligned} & \left\| \nabla_{\theta_t} l(\theta_1, \dots, \theta_k, \theta_{k+1}^*, \dots, \theta_T^*) - \nabla_{\theta_t} l(\theta_1, \dots, \theta_{k-1}, \theta_k^*, \dots, \theta_T^*) \right\|_2 \\ & \leq M_g \sqrt{\mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} \left[Q_k^{\pi_{\theta^*}}(s_k, \pi_k(s_k|\theta_k)) \right] - \mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} \left[Q_k^{\pi_{\theta^*}}(s_k, \pi_k(s_k|\theta_k^*)) \right]}. \end{aligned} \quad (28)$$

In what follows, we show that if relying on the weaker condition (28), our analysis leads to a suboptimal characterization of the KL constant, resulting in an exponential dependence on T .

We proceed with the same proof sketch in Theorem 1. In step 1, we aim to establish the *bounded gradient mismatch* inequality (8). Using the definition of X_t and Y_t and the weaker condition (28), we have

$$\begin{aligned} X_t - Y_t & \stackrel{(a)}{\leq} \sum_{k=t+1}^T \max_{g_k \in \partial \delta_{\Theta_k}(\theta_k)} \left\| \nabla_{\theta_t} l(\theta_1, \dots, \theta_{k-1}, \theta_k, \theta_{k+1}^*, \dots, \theta_T^*) - \nabla_{\theta_t} l(\theta_1, \dots, \theta_{k-1}, \theta_k^*, \theta_{k+1}^*, \dots, \theta_T^*) \right\|_2 \\ & \stackrel{(b)}{\leq} M_g \sum_{k=t+1}^T \sqrt{\mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} \left[Q_k^{\pi_{\theta^*}}(s_k, \pi_k(s_k|\theta_k)) \right] - \mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} \left[Q_k^{\pi_{\theta^*}}(s_k, \pi_k(s_k|\theta_k^*)) \right]} \\ & \stackrel{(c)}{\leq} \frac{M_g}{\sqrt{2\mu_Q}} \sum_{k=t+1}^T \min_{g_k \in \partial \delta_{\Theta_k}(\theta_k)} \left\| \nabla_{\theta_k} \mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} \left[Q_k^{\pi_{\theta^*}}(s_k, \pi_k(s_k|\theta_k)) \right] + g_k \right\|_2 \\ & = \frac{M_g}{\sqrt{2\mu_Q}} \sum_{k=t+1}^T X_k. \end{aligned}$$

Inequality (a) uses the first three steps in (26). (b) comes from the assumption (28). (c) uses the *KL condition of optimal expected Q-value functions*. To proceed, we establish a hard instance when $M_g > 1$ and $G > 1$.

LEMMA 17. Assume that the nonnegative sequences $\{X_t\}_{t=1}^T$ and $\{Y_t\}_{t=1}^T$ satisfy

$$|X_t - Y_t| \leq M_g \sum_{k=t+1}^T X_k, \quad X_T = Y_T, \quad X_t, Y_t \leq G, \quad \forall t \in [T], \quad (29)$$

with constants $M_g > 1$ and $G > 1$. Then the best M with $\sum_{t=1}^T X_t^2 \leq M \sum_{t=1}^T Y_t^2$ for all sequences $\{X_t\}_{t=1}^T$ and $\{Y_t\}_{t=1}^T$ satisfying (29) cannot be smaller than $M_g^{2(T-1)}$.

Proof of Lemma 17 The analysis is through construction of a hard instance. Let $X_t = GM_g^{1-t}$ for any $t \in [T]$. Set $Y_T = GM_g^{1-T}$ and $Y_t = 0$ for $1 \leq t < T$. The two sequences satisfy (29) because

$$|X_t - Y_t| = |X_t| = GM_g^{1-t} = M_g \times GM_g^{-t} = M_g X_{t+1} \leq M_g \sum_{k=t+1}^T X_k, \quad \forall t < T,$$

and $|X_T - Y_T| = 0$. Therefore, we have

$$\frac{\sum_{t=1}^T X_t^2}{\sum_{t=1}^T Y_t^2} = \frac{\sum_{t=1}^T X_t^2}{Y_T^2} \geq \frac{X_1^2}{Y_T^2} = M_g^{2(T-1)}.$$

This concludes the proof. \square

Lemma 17 implies that the constant of the *bounded gradient mismatch* inequality (8) depends at least exponentially on T . Following the same steps in the proof of Theorem 1, the KL constant admits an exponential dependence on T . It remains clear whether the dependence can be improved under (28) through alternative analysis. To remove the exponential dependence, we apply the stronger *sequential decomposition inequalities* in Theorem 1 that hold in various applications.

Appendix B: Omitted Proofs in Section 4

B.1. Bounded Gradient

Proof of Lemma 4 From the Policy Gradient Theorem (Sutton et al. 1999, Theorem 1), we have

$$\begin{aligned} \nabla_{\theta_t(s_t, i)} l(\theta) &= \nabla_{\theta_t(s_t, i)} \mathbb{E}_{s_t \sim \rho_t(\cdot | \pi_\theta)} \left[Q_t^{\pi_\theta}(s_t, \pi_t(s_t | \theta_t)) \right] \\ &= \rho_t(s_t | \pi_\theta) \left(\underbrace{-\frac{\lambda}{n\theta_t(s_t, i)}}_{(I)} + \underbrace{C_t(s_t, i)}_{(II)} + \sum_{s_{t+1} \in \mathcal{S}} P_t(s_{t+1} | s_t, i) \underbrace{V_{t+1}^{\pi_\theta}(s_{t+1})}_{(III)} \right). \end{aligned}$$

Utilizing the assumption that $\theta_t(s_t, i) \geq \underline{p}$, we have $|(I)| \leq \lambda/(n\underline{p})$. The second term admits $|(II)| \leq \bar{C}$. Lastly, we have the recursive form for (III) using the Bellman equation (1):

$$V_t^{\pi_\theta}(s_t) = C_t^r(s_t, \pi_t(s_t | \theta_t)) + \sum_{s_{t+1}} P_t(s_{t+1} | s_t, \pi_t(s_t | \theta_t)) V_{t+1}^{\pi_\theta}(s_{t+1}).$$

By definition, we have

$$\begin{aligned} \left| C_t^r(s_t, \pi_t(s_t | \theta_t)) \right| &= \left| C_t(s_t, \pi_t(s_t | \theta_t)) + \lambda \mathcal{R}(\pi_t(s_t | \theta_t)) \right| \\ &\leq \sum_{i \in \mathcal{N}} \theta_t(s_t, i) |C_t(s_t, i)| + \left| \lambda \mathcal{R}(\pi_t(s_t | \theta_t)) \right| \\ &\leq \bar{C} + \lambda \log(1/(n\underline{p})). \end{aligned}$$

The last inequality uses the assumption that $\pi_t(a_t | s_t) \geq \underline{p}$. Therefore, by mathematical induction, we have

$$|V_t^{\pi_\theta}(s_t)| \leq (T-t)\bar{C} + \lambda(T-t) \log(1/(n\underline{p})), \quad \forall s_t \in \mathcal{S}.$$

Combining all the results, we have

$$\begin{aligned}
\|\nabla_{\theta_t} l(\theta)\|_F &\leq \sum_{s_t \in \mathcal{S}, i \in \mathcal{N}} |\nabla_{\theta_t(s_t, i)} l(\theta)| \\
&\leq \sum_{s_t \in \mathcal{S}, i \in \mathcal{N}} \left| \rho_t(s_t | \pi_\theta) \left(-\frac{\lambda}{n\theta_t(s_t, i)} + C_t(s_t, i) + \sum_{s_{t+1} \in \mathcal{S}} P_t(s_{t+1} | s_t, i) V_{t+1}^{\pi_\theta}(s_{t+1}) \right) \right| \\
&\leq \sum_{s_t \in \mathcal{S}, i \in \mathcal{N}} \rho_t(s_t | \pi_\theta) \left(\frac{\lambda}{n\underline{p}} + \bar{C} + \sum_{s_{t+1}} P_t(s_{t+1} | s_t, i) \left((T-t)\bar{C} + \lambda(T-t) \log\left(\frac{1}{n\underline{p}}\right) \right) \right) \\
&\leq T\bar{C} + \frac{\lambda}{n\underline{p}} + \lambda T \log\left(\frac{1}{n\underline{p}}\right).
\end{aligned}$$

This completes the proof. \square

B.2. Sequential Decomposition Inequality

Proof of Lemma 5 Let $\theta_\alpha = (\theta_1, \dots, \theta_{k-1}, \theta_k, \theta_{k+1}^*, \dots, \theta_T^*)$ and $\theta_\beta = (\theta_1, \dots, \theta_{k-1}, \theta_k^*, \theta_{k+1}^*, \dots, \theta_T^*)$, then we have the following inequalities:

$$\begin{aligned}
&\left\| \nabla_{\theta_t} l(\theta_1, \dots, \theta_{k-1}, \theta_k, \theta_{k+1}^*, \dots, \theta_T^*) - \nabla_{\theta_t} l(\theta_1, \dots, \theta_{k-1}, \theta_k^*, \theta_{k+1}^*, \dots, \theta_T^*) \right\|_F \\
&\leq \sum_{s_t \in \mathcal{S}, i \in \mathcal{N}} \left| \nabla_{\theta_t(s_t, i)} l(\theta_1, \theta_{k-1}, \dots, \theta_k, \theta_{k+1}^*, \dots, \theta_T^*) - \nabla_{\theta_t(s_t, i)} l(\theta_1, \dots, \theta_{k-1}, \theta_k^*, \theta_{k+1}^*, \dots, \theta_T^*) \right| \\
&= \sum_{s_t \in \mathcal{S}, i \in \mathcal{N}} \left| \rho_t(s_t | \pi_\theta) \sum_{s_{t+1} \in \mathcal{S}} P_t(s_{t+1} | s_t, i) \left(Q_{t+1}^{\pi_\alpha}(s_{t+1}, \pi_{t+1}(s_{t+1} | \theta_{t+1})) - Q_{t+1}^{\pi_\beta}(s_{t+1}, \pi_{t+1}(s_{t+1} | \theta_{t+1})) \right) \right|.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
&Q_{t+1}^{\pi_\alpha}(s_{t+1}, \pi_{t+1}(s_{t+1} | \theta_{t+1})) - Q_{t+1}^{\pi_\beta}(s_{t+1}, \pi_{t+1}(s_{t+1} | \theta_{t+1})) \\
&= \sum_{i_{t+1} \in \mathcal{N}} \theta_{t+1}(s_{t+1}, i_{t+1}) \sum_{s_{t+2} \in \mathcal{S}} P_t(s_{t+2} | s_{t+1}, i_{t+1}) \\
&\quad \left(Q_{t+2}^{\pi_\alpha}(s_{t+2}, \pi_{t+2}(s_{t+2} | \theta_{t+2})) - Q_{t+2}^{\pi_\alpha}(s_{t+2}, \pi_{t+2}(s_{t+2} | \theta_{t+2})) \right). \tag{30}
\end{aligned}$$

Applying (30) recursively, we conclude that

$$\begin{aligned}
&\left\| \nabla_{\theta_t} l(\theta_1, \dots, \theta_{k-1}, \theta_k, \theta_{k+1}^*, \dots, \theta_T^*) - \nabla_{\theta_t} l(\theta_1, \dots, \theta_{k-1}, \theta_k^*, \theta_{k+1}^*, \dots, \theta_T^*) \right\|_F \\
&\leq \sum_{s_t \in \mathcal{S}, i \in \mathcal{N}} \left| \rho_t(s_t | \pi_\theta) \sum_{s_{t+1} \in \mathcal{S}} P_t(s_{t+1} | s_t, i) \sum_{i_{t+1} \in \mathcal{N}} \theta_{t+1}(s_{t+1}, i_{t+1}) \sum_{s_{t+2} \in \mathcal{S}} P_t(s_{t+2} | s_{t+1}, i_{t+1}) \dots \right. \\
&\quad \left. \sum_{i_{k-1} \in \mathcal{N}} \theta_{k-1}(s_{k-1}, i_{k-1}) \sum_{s_k \in \mathcal{S}} P_{k-1}(s_k | s_{k-1}, i_{k-1}) \left(Q_k^*(s_k, \pi_k(s_k | \theta_k)) - Q_k^*(s_k, \pi_k(s_k | \theta_k^*)) \right) \right|.
\end{aligned}$$

Since $Q_k^*(s_k, \pi_k(s_k|\theta_k)) \geq Q_k^*(s_k, \pi_k(s_k|\theta_k^*))$, the absolute function can be removed. Then we have

$$\begin{aligned}
& \left\| \nabla_{\theta_t} l(\theta_1, \dots, \theta_{k-1}, \theta_k, \theta_{k+1}^*, \dots, \theta_T^*) - \nabla_{\theta_t} l(\theta_1, \dots, \theta_{k-1}, \theta_k^*, \theta_{k+1}^*, \dots, \theta_T^*) \right\|_F \\
& \leq \sum_{s_t \in \mathcal{S}, i \in \mathcal{N}} \rho_t(s_t|\pi_\theta) \sum_{s_{t+1} \in \mathcal{S}} P_t(s_{t+1}|s_t, i) \sum_{i_{t+1} \in \mathcal{N}} \theta_{t+1}(s_{t+1}, i_{t+1}) \dots \left(Q_k^*(s_k, \pi_k(s_k)) - Q_k^*(s_k, \pi_k^*(s_k)) \right) \\
& \stackrel{(a)}{\leq} \frac{1}{\underline{p}} \sum_{s_t \in \mathcal{S}} \rho_t(s_t|\pi_\theta) \sum_{i \in \mathcal{N}} \theta_t(s_t, i) \sum_{s_{t+1} \in \mathcal{S}} P_t(s_{t+1}|s_t, i) \sum_{i_{t+1} \in \mathcal{N}} \theta_{t+1}(s_{t+1}, i_{t+1}) \sum_{s_{t+2} \in \mathcal{S}} P_t(s_{t+2}|s_{t+1}, i_{t+1}) \dots \\
& \quad \sum_{i_{k-1} \in \mathcal{N}} \theta_{k-1}(s_{k-1}, i_{k-1}) \sum_{s_k \in \mathcal{S}} P_{k-1}(s_k|s_{k-1}, i_{k-1}) \left(Q_k^*(s_k, \pi_k(s_k|\theta_k)) - Q_k^*(s_k, \pi_k(s_k|\theta_k^*)) \right) \\
& = \frac{1}{\underline{p}} \left(\mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} \left[Q_k^*(s_k, \pi_k(s_k|\theta_k)) \right] - \mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} \left[Q_k^*(s_k, \pi_k^*(s_k|\theta_k)) \right] \right).
\end{aligned}$$

Inequality (a) uses the assumption that $\theta_t(s_t, i) \geq \underline{p}$ for any $i \in \mathcal{N}$. This completes the proof. \square

Appendix C: Omitted Proofs in Section 5

C.1. KL Condition of Optimal Q-value Function

Proof of Lemma 7 From (1), the Q-value function has the following expression:

$$V_t^{\pi_\theta}(s_t) = Q_t^{\pi_\theta}(s_t, \pi_t(s_t|\theta_t)) = \underbrace{s_t^\top Q_t s_t + s_t^\top \theta_t^\top R_t \theta_t s_t}_{(I)} + \underbrace{\mathbb{E}_{w_t} \left[V_{t+1}^{\pi_\theta}((A + B\theta_t)s_t + w_t) \right]}_{(II)}.$$

Suppose that $V_{t+1}^{\pi_\theta}(s_{t+1})$ is continuously differentiable in s_{t+1} . Term (I) is a quadratic function of θ_t and, therefore, is continuously differentiable. Term (II) is continuously differentiable since the composition of a continuously differentiable function and a linear function is continuously differentiable. From the induction base $V_T^{\pi_\theta}(s_T) = s_T^\top Q_T s_T$ that is continuously differentiable, we can prove that the Q-value function is continuously differentiable by mathematical induction, which implies the continuous differentiability of the expected Q-value function.

If we plug π^* into (II), we get the explicit expression of the optimal Q-value function. Bertsekas (1995) demonstrated the convexity of V_t^* by mathematical induction. Therefore, its composition with linear function $As_t + B(\theta_t s_t) + w_t$ is still convex, which implies the convexity of the term (II). From Assumption 3 and Lemma 6, we know that matrices R_t and $\mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)}[s_t s_t^\top]$ are positive definite. Taking the expectation on the term (I) gives a $2\underline{\sigma}_X \underline{\sigma}_R$ -strongly convex function in θ_t (Bhandari and Russo 2024). Combining these, we conclude that the expected optimal Q-value function is $2\underline{\sigma}_X \underline{\sigma}_R$ -strongly convex. Leveraging Corollary 1, we establish the KL condition of the expected optimal Q-value function with KL constant $2\underline{\sigma}_X \underline{\sigma}_R$ for all $t = 0, \dots, T-1$. \square

C.2. Bounded Gradient

Proof of Lemma 8 For any $\theta \in \Theta$, we derive the following inequality using (12):

$$\begin{aligned}
\|P_t\|_2 & \leq \|Q_t\|_2 + \|\theta_t R_t \theta_t\|_2 + \|(A + B\theta)^\top P_{t+1} (A + B\theta)\|_2 \\
& \stackrel{(a)}{\leq} \bar{\sigma}_Q + \|\theta_t\|_2^2 \|R_t\|_2 + \|A + B\theta\|_2^2 \|P_{t+1}\|_2 \\
& \stackrel{(b)}{\leq} \bar{\sigma}_Q + \bar{\sigma}_\Theta^2 \bar{\sigma}_R + \|P_{t+1}\|_2.
\end{aligned}$$

Here inequality (a) uses the assumption that $\sigma_{\max}(Q_t) \leq \bar{\sigma}_Q$ for all $0 \leq t \leq T$, and inequality (b) uses the assumption that $\|\theta_t\|_2 \leq \bar{\sigma}_\Theta$ and $\|A + B\theta_t\|_2 \leq 1$ for all $0 \leq t \leq T-1$. Since $P_T = Q_T$, we conclude that

$$\|P_t\|_2 \leq (T-t+1)\bar{\sigma}_Q + (T-t)\bar{\sigma}_\Theta^2\bar{\sigma}_R, \quad \forall t=0, \dots, T. \quad (31)$$

Next, we have the following inequality from (13) for all $t=0, \dots, T-1$:

$$\begin{aligned} \|E_t\|_2 &= \|R_t\theta_t + B^\top P_{t+1}(A + B\theta_t)\|_2 \\ &\leq \|R_t\theta_t\|_2 + \|B^\top P_{t+1}(A + B\theta_t)\|_2 \\ &\stackrel{(a)}{\leq} \bar{\sigma}_\Theta\bar{\sigma}_R + \|B\|_2\|P_{t+1}\|_2\|A + B\theta_t\|_2 \\ &\stackrel{(b)}{\leq} \bar{\sigma}_\Theta\bar{\sigma}_R + (T-t+1)\bar{\sigma}_Q\|B\|_2 + (T-t)\bar{\sigma}_\Theta^2\bar{\sigma}_R\|B\|_2, \end{aligned} \quad (32)$$

where inequality (a) comes from $\sigma_{\max}(R_t) \leq \bar{\sigma}_R$ for any $0 \leq t \leq T-1$ and $\|\theta_t\|_2 \leq \bar{\sigma}_\Theta$ for any $\theta_t \in \Theta_t, 0 \leq t \leq T-1$. Inequality (b) uses (31) and $\|A + B\theta_t\|_2 \leq 1$ for any $\theta_t \in \Theta_t, 0 \leq t \leq T-1$. Recall the linear dynamic function $s_{t+1} = As_t + Ba_t + w_t$, we have the following result:

$$\mathbb{E}[s_{t+1}s_{t+1}^\top] = (A + B\theta_t)\mathbb{E}[s_t s_t^\top](A + B\theta_t)^\top + \mathbb{E}[w_t w_t^\top].$$

Therefore, we can derive the following inequality for all $t=0, \dots, T-1$:

$$\begin{aligned} \|\mathbb{E}[s_{t+1}s_{t+1}^\top]\|_2 &\leq \|A + B\theta_t\|_2^2 \|\mathbb{E}[s_t s_t^\top]\|_2 + \|\mathbb{E}[w_t w_t^\top]\|_2 \\ &\stackrel{(a)}{\leq} \|\mathbb{E}[s_t s_t^\top]\|_2 + \|\mathbb{E}[w_t w_t^\top]\|_2 \\ &\stackrel{(b)}{\leq} \|\mathbb{E}[s_t s_t^\top]\|_2 + \bar{\sigma}_W. \end{aligned}$$

Here inequality (a) uses $\|A + B\theta_t\|_2 \leq 1$ for any $\theta_t \in \Theta_t, 0 \leq t \leq T-1$ and inequality (b) comes from $\sigma_{\max}(\mathbb{E}[w_t w_t^\top]) \leq \bar{\sigma}_W$. Taking the telescoping sum, for any $t=0, \dots, T-1$, we have

$$\begin{aligned} \|\mathbb{E}[s_{t+1}s_{t+1}^\top]\|_2 &\leq \sigma_{\max}(\mathbb{E}[s_0 s_0^\top]) + (t+1)\bar{\sigma}_W \\ &\leq \bar{\sigma}_X + (t+1)\bar{\sigma}_W. \end{aligned} \quad (33)$$

Thus, combining (32) and (33), we conclude that

$$\begin{aligned} \|\nabla_t l(\theta)\|_F &\stackrel{(a)}{\leq} \sqrt{\min\{m, n\}} \|\nabla_t l(\theta)\|_2 \\ &\leq 2\sqrt{\min\{m, n\}} \|E_t\|_2 \|\mathbb{E}[s_t s_t^\top]\|_2 \\ &\leq 2\sqrt{\min\{m, n\}} (\bar{\sigma}_\Theta\bar{\sigma}_R + (T-t+1)\bar{\sigma}_Q\|B\|_2 + (T-t)\bar{\sigma}_\Theta^2\bar{\sigma}_R\|B\|_2) (\bar{\sigma}_X + t\bar{\sigma}_W), \end{aligned}$$

where (a) uses $\|A\|_F \leq \sqrt{r}\|A\|_2$ with $r = \text{rank}(A)$ (Golub and Van Loan 2013). The right-hand side of the inequality is polynomial in the model parameters $(m, n, T, \bar{\sigma}_Q, \bar{\sigma}_R, \bar{\sigma}_\Theta, \bar{\sigma}_X, \bar{\sigma}_W, \|B\|_2)$. \square

C.3. Sequential Decomposition Inequality

Proof of Lemma 9 First define

$$\Pi_{[j_1:j_2]} := (A + B\theta_{j_2})(A + B\theta_{j_2-1}) \dots (A + B\theta_{j_1+1})(A + B\theta_{j_1}),$$

for $j_1 \leq j_2$. Therefore, we have $\|\Pi_{[j_1:j_2]}\|_2 \leq 1$ for any $0 \leq j_1 \leq j_2 \leq T-1$ since $\|A + B\theta_t\|_2 \leq 1$ for any $\theta_t \in \Theta_t, 0 \leq t \leq T-1$. Recall the gradient formulation in Proposition 2, we can derive

$$\begin{aligned} & \nabla_t l(\theta_{[0:k-1]}, \theta_k, \theta_{[k+1:T-1]}^*) - \nabla_t l(\theta_{[0:k-1]}, \theta_k^*, \theta_{[k+1:T-1]}^*) \\ &= 2B^\top (P_{t+1}(\theta_{[0:k-1]}, \theta_k, \theta_{[k+1:T-1]}^*) - P_{t+1}(\theta_{[0:k-1]}, \theta_k^*, \theta_{[k+1:T-1]}^*)) (A + B\theta_t) \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} [s_t s_t^\top] \\ &\stackrel{(a)}{=} 2B^\top (A + B\theta_{t+1})^\top (P_{t+2}(\theta_{[0:k-1]}, \theta_k, \theta_{[k+1:T-1]}^*) \\ &\quad - P_{t+2}(\theta_{[0:k-1]}, \theta_k^*, \theta_{[k+1:T-1]}^*)) (A + B\theta_{t+1})(A + B\theta_t) \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} [s_t s_t^\top] \\ &\quad \dots \\ &= 2B^\top \Pi_{[t+1:k-1]}^\top \left(P_k(\theta_{[0:k-1]}, \theta_k, \theta_{[k+1:T-1]}^*) - P_k(\theta_{[0:k-1]}, \theta_k^*, \theta_{[k+1:T-1]}^*) \right) \Pi_{[t:k-1]} \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} [s_t s_t^\top]. \end{aligned}$$

Equation (a) uses the update (12). Utilizing the explicit expression of the Q-value function in Proposition 2, we conclude that

$$\begin{aligned} & \mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} \left[Q_k^*(s_k, \pi_k(s_k|\theta_k)) \right] - \mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} \left[Q_k^*(s_k, \pi_k(s_k|\theta_k^*)) \right] \\ &= \mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} \left[s_k^\top (P_k(\theta_{[0:k-1]}, \theta_k, \theta_{[k+1:T-1]}^*) - P_k(\theta_{[0:k-1]}, \theta_k^*, \theta_{[k+1:T-1]}^*)) s_k \right] \\ &= \text{Tr} \left((P_k(\theta_{[0:k-1]}, \theta_k, \theta_{[k+1:T-1]}^*) - P_k(\theta_{[0:k-1]}, \theta_k^*, \theta_{[k+1:T-1]}^*)) \mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} [s_k s_k^\top] \right). \end{aligned}$$

Therefore, we have

$$\begin{aligned} & \left\| \nabla_t l(\theta_{[0:k-1]}, \theta_k, \theta_{[k+1:T-1]}^*) - \nabla_t l(\theta_{[0:k-1]}, \theta_k^*, \theta_{[k+1:T-1]}^*) \right\|_F^2 \\ &= 4 \left\| B^\top \Pi_{[t+1:k-1]}^\top \left(P_k(\theta_{[0:k-1]}, \theta_k, \theta_{[k+1:T-1]}^*) - P_k(\theta_{[0:k-1]}, \theta_k^*, \theta_{[k+1:T-1]}^*) \right) \Pi_{[t:k-1]} \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} [s_t s_t^\top] \right\|_F^2 \\ &\stackrel{(a)}{\leq} \frac{4 \|B\|_2^2 \left\| \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} [s_t s_t^\top] \right\|_2^2}{\underline{\sigma}_X^2} \left\| (P_k(\theta_{[0:k-1]}, \theta_k, \theta_{[k+1:T-1]}^*) - P_k(\theta_{[0:k-1]}, \theta_k^*, \theta_{[k+1:T-1]}^*)) \mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} [s_k s_k^\top] \right\|_F^2 \\ &\stackrel{(b)}{\leq} \frac{4 \|B\|_2^2 \left\| \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} [s_t s_t^\top] \right\|_2^2}{\underline{\sigma}_X^2} \left\| \mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} \left[Q_k^*(s_k, \pi_k(s_k|\theta_k)) \right] - \mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} \left[Q_k^*(s_k, \pi_k(s_k|\theta_k^*)) \right] \right\|_F^2. \end{aligned}$$

Inequality (a) comes from $\|\Pi_{[j_1:j_2]}\|_2 \leq 1$ for any $0 \leq j_1 \leq j_2 \leq T-1$ and uses the assumption $\underline{\sigma}_X = \min_{t=0, \dots, T} \sigma_{\min}(\mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} [s_t s_t^\top])$. Inequality (b) utilizes $\|A\|_F^2 = \text{Tr}(A^\top A) \leq \text{Tr}(A)^2$ when A is positive semi-definite. Since θ_k^* is the optimal solution, we have

$$\mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} \left[Q_k^*(s_k, \pi_k(s_k|\theta_k)) \right] \geq \mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} \left[Q_k^*(s_k, \pi_k(s_k|\theta_k^*)) \right], \quad \forall \theta_k \in \Theta_k.$$

Therefore, we conclude that

$$\begin{aligned} & \left\| \nabla_t l(\theta_{[0:k-1]}, \theta_k, \theta_{[k+1:T-1]}^*) - \nabla_t l(\theta_{[0:k-1]}, \theta_k^*, \theta_{[k+1:T-1]}^*) \right\|_F \\ &\leq \frac{2 \|B\|_2 \left\| \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} [s_t s_t^\top] \right\|_2}{\underline{\sigma}_X} \left(\mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} \left[Q_k^*(s_k, \pi_k(s_k|\theta_k)) \right] - \mathbb{E}_{s_k \sim \rho_k(\cdot|\pi_\theta)} \left[Q_k^*(s_k, \pi_k(s_k|\theta_k^*)) \right] \right). \end{aligned}$$

From the proof of Lemma 8, we know that $\|\mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)}[s_t s_t^\top]\|_2$ is polynomial in model parameters. This concludes the proof. \square

Appendix D: Omitted Proofs in Section 6

In Appendix D and E, we frequently use the following argument. Let $f : \mathcal{X} \rightarrow \mathbb{R}$ be a continuously differentiable function. Assume that ξ is a random variable whose cumulative distribution function \mathbb{P} is Lipschitz continuous. Consider a function $F(x) := \mathbb{E}_{\xi \sim \mathbb{P}}[f(x \wedge \xi)]$. Chen et al. (2024) proved that:

1. $\frac{\partial}{\partial x} f(x \wedge \xi) \stackrel{\text{a.s.}}{=} \mathbf{1}(x \leq \xi) \times f'(x \wedge \xi) = \mathbf{1}(x \leq \xi) \times f'(x)$.
2. $F'(x) = \frac{\partial}{\partial x} \mathbb{E}_{\xi \sim \mathbb{P}}[f(x \wedge \xi)] = \mathbb{E}_{\xi \sim \mathbb{P}}[\frac{\partial}{\partial x} f(x \wedge \xi)] = \mathbb{E}_{\xi \sim \mathbb{P}}[\mathbf{1}(x \leq \xi) \times f'(x)] = \mathbb{P}(x \leq \xi) \times f'(x)$.

Similar arguments hold when $F(x) := \mathbb{E}_{\xi \sim \mathbb{P}}[f(x \vee \xi)]$. In the following sections, we directly use the results.

D.1. KL Condition of Expected Optimal Q-value Functions

Proof of Lemma 10 We use three parts to complete the proof. First, we demonstrate the relationships between suboptimality gaps $F_t(\theta_t) - F_t(\theta_t^*)$ and $f_t(\theta_{t,i}|i) - f_t(\theta_{t,i}^*|i)$. Next, we show the connections of their gradients. Finally, we prove the KL property of F_t .

Step 1: Relationship between suboptimality gaps. Applying the Bellman equation (2), it holds that

$$\begin{aligned} & \mathbb{E}_{(x_t, i_t) \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^*(x_t, i_t, \pi_t(x_t, i_t|\theta_t)) \right] \\ &= \mathbb{E}_{(x_t, i_t) \sim \rho_t(\cdot|\pi_\theta)} \left[L_t(x_t \vee \theta_{t,i_t}|i_t) + \mathbb{E}_{i_{t+1} \sim p(\cdot|i_t), D_t \sim P_D(\cdot|i_t)} [V_{t+1}^*(x_t \vee \theta_{t,i_t} - D_t, i_{t+1})] \right]. \end{aligned}$$

Recalling the definition (15), we express expected optimal Q-value functions as

$$F_t(\theta_t) = \mathbb{E}_{(x_t, i_t) \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^*(x_t, i_t, \pi_t(x_t, i_t|\theta_t)) \right] = \mathbb{E}_{(x_t, i_t) \sim \rho_t(\cdot|\pi_\theta)} \left[f_t(x_t \vee \theta_{t,i_t}|i_t) \right].$$

By the law of total expectation, we rewrite the suboptimality gap:

$$\begin{aligned} F_t(\theta_t) - F_t(\theta_t^*) &= \mathbb{E}_{(x_t, i_t) \sim \rho_t(\cdot|\pi_\theta)} \left[f_t(x_t \vee \theta_{t,i_t}|i_t) - f_t(x_t \vee \theta_{t,i_t}^*|i_t) \right] \\ &= \mathbb{E}_{i_t \sim \nu} \left[\mathbb{E}_{x_t} \left[f_t(x_t \vee \theta_{t,i_t}|i_t) - f_t(x_t \vee \theta_{t,i_t}^*|i_t) \middle| i_t \right] \right] \\ &= \sum_{i \in \mathcal{I}} \nu_i \times \mathbb{E}_{x_t} \left[f_t(x_t \vee \theta_{t,i}|i) - f_t(x_t \vee \theta_{t,i}^*|i) \middle| i = i \right]. \end{aligned}$$

Without loss of generality, we assume that $\theta_{t,i} \leq \theta_{t,i}^*$. For any random variable ξ and its corresponding cumulative distribution function $P(\xi)$, we have

$$\begin{aligned}
& \mathbb{E}_{\xi \sim P(\xi)} \left[f_t(\xi \vee \theta_{t,i}|i) - f_t(\xi \vee \theta_{t,i}^*|i) \right] \\
&= \int_{-\infty}^{\theta_{t,i}} \left(f_t(\theta_{t,i}|i) - f_t(\theta_{t,i}^*|i) \right) dP(\xi) + \int_{\theta_{t,i}}^{\theta_{t,i}^*} \left(f_t(\xi|i) - f_t(\theta_{t,i}^*|i) \right) dP(\xi) \\
&\stackrel{(a)}{\leq} \int_{-\infty}^{\theta_{t,i}} \left(f_t(\theta_{t,i}|i) - f_t(\theta_{t,i}^*|i) \right) dP(\xi) + \int_{\theta_{t,i}}^{\theta_{t,i}^*} \left(f_t(\theta_{t,i}|i) - f_t(\theta_{t,i}^*|i) \right) dP(\xi) \\
&\stackrel{(b)}{\leq} f_t(\theta_{t,i}|i) - f_t(\theta_{t,i}^*|i).
\end{aligned}$$

Inequality (a) holds as $\theta_{t,i}^*$ is a minimizer of $f_t(\cdot|i)$, which implies that $f_t(\cdot|i)$ is non-increasing on the interval $[\theta_{t,i}, \theta_{t,i}^*]$ due to its convexity. Inequality (b) comes from the fact that $f_t(\theta_{t,i}|i) - f_t(\theta_{t,i}^*|i) \geq 0$. The same result hold when $\theta_{t,i} > \theta_{t,i}^*$. Therefore, we have

$$\begin{aligned}
F_t(\theta_t) - F_t(\theta_t^*) &= \sum_{i \in \mathcal{I}} \nu_i \times \mathbb{E}_{x_t} \left[f_t(x_t \vee \theta_{t,i}|i) - f_t(x_t \vee \theta_{t,i}^*|i) \middle| i_t = i \right] \\
&\leq \sum_{i \in \mathcal{I}} \nu_i \times \left[f_t(\theta_{t,i}|i) - f_t(\theta_{t,i}^*|i) \right].
\end{aligned}$$

Step 2: Relationship between gradients. By definition, we calculate the partial gradient of F_t :

$$\begin{aligned}
\nabla_{\theta_{t,i}} F_t(\theta_t) &= \nabla_{\theta_{t,i}} \mathbb{E}_{(x_t, i_t) \sim \rho_t(\cdot|\pi_\theta)} \left[f_t(x_t \vee \theta_{t,i}|i_t) \right] \\
&\stackrel{(a)}{=} \mathbb{E}_{(x_t, i_t) \sim \rho_t(\cdot|\pi_\theta)} \left[\mathbf{1}(i_t = i) \times \mathbf{1}(\theta_{t,i} \geq x_t) \times f'_t(\theta_{t,i}|i) \right] \\
&= \mathbb{P}(i_t = i, \theta_{t,i} \geq x_t) \times f'_t(\theta_{t,i}|i).
\end{aligned}$$

Here equation (a) utilizes the chain rule. Next, we analyze the property of $\mathbb{P}(i_t = i, \theta_{t,i} \geq x_t)$:

$$\begin{aligned}
\mathbb{P}(i_t = i, \theta_{t,i} \geq x_t) &\stackrel{(a)}{=} \mathbb{E}_{i_{t-1}} \left[\mathbb{P}(i_t = i, \theta_{t,i} \geq x_t | i_{t-1}) \right] \\
&\stackrel{(b)}{=} \mathbb{E}_{i_{t-1}} \left[\mathbb{P}(i_t = i | i_{t-1}) \times \mathbb{P}(\theta_{t,i} \geq x_t | i_{t-1}) \right] \\
&= \mathbb{E}_{i_{t-1}} \left[p(i | i_{t-1}) \times \mathbb{P}(\theta_{t,i} \geq x_t | i_{t-1}) \right].
\end{aligned}$$

Equation (a) uses the law of total expectation. Equation (b) holds because x_t and i_t are independent conditioned on i_{t-1} . For simplicity, we use $\rho_{t-1}^x(\cdot|i_{t-1})$ to denote the CDF of x_{t-1} conditioned on i_{t-1} . From the transition kernel $x_t = x_{t-1} \vee \theta_{t-1, i_{t-1}} - D_{t-1}$, we derive the following inequalities:

$$\begin{aligned}
\mathbb{P}(\theta_{t,i} \geq x_t | i_{t-1}) &= \mathbb{P}(\theta_{t,i} \geq x_{t-1} \vee \theta_{t-1, i_{t-1}} - D_{t-1} | i_{t-1}) \\
&= \mathbb{P}(\theta_{t,i} \geq x_{t-1} - D_{t-1}, \theta_{t,i} \geq \theta_{t-1, i_{t-1}} - D_{t-1} | i_{t-1}) \\
&= \int_0^{\theta_{t-1, i_{t-1}} - \theta_{t,i}} \mathbb{P}(\theta_{t,i} \geq x_{t-1} - D, \theta_{t,i} \geq \theta_{t-1, i_{t-1}} - D | i_{t-1}) dP_D(D | i_{t-1}) \\
&\quad + \int_{\theta_{t-1, i_{t-1}} - \theta_{t,i}}^\infty \mathbb{P}(\theta_{t,i} \geq x_{t-1} - D, \theta_{t,i} \geq \theta_{t-1, i_{t-1}} - D | i_{t-1}) dP_D(D | i_{t-1}) \\
&= \int_{\theta_{t-1, i_{t-1}} - \theta_{t,i}}^\infty \rho_{t-1}^x(D + \theta_{t,i} | i_{t-1}) dP_D(D | i_{t-1}).
\end{aligned}$$

The last inequality holds as $\mathbf{1}(\theta_{t,i} \geq \theta_{t-1,i_{t-1}} - D) = 0$ for any $D \in [0, \theta_{t-1,i_{t-1}} - \theta_{t,i})$. Since $\rho_{t-1}^x(\cdot|i_{t-1})$ is non-decreasing and non-negative, we have

$$\int_{\theta_{t-1,i_{t-1}} - \theta_{t,i}}^{\infty} \rho_{t-1}^x(D + \theta_{t,i}|i_{t-1}) dP_D(D|i_{t-1}) \geq \int_{B - \theta_{t,i}}^{\infty} \rho_{t-1}^x(D + \theta_{t,i}|i_{t-1}) dP_D(D|i_{t-1}).$$

From the transition kernel, we know that $x_{t+1} = x_t \vee \theta_{t,i_t} - D_t$. Since $x_1 \in (-\infty, B]$, $\theta_{t,i} \in [0, B]$, and $D_t \in [0, +\infty)$, we have $x_t \leq B$ for any $t \in [T]$. Therefore, $\rho_{t-1}^x(B|i_{t-1}) = 1$, which implies that

$$\mathbb{P}(\theta_{t,i} \geq x_t|i_{t-1}) \geq \int_{B - \theta_{t,i}}^{\infty} \rho_{t-1}^x(D + \theta_{t,i}|i_{t-1}) dP_D(D|i_{t-1}) \geq \int_{B - \theta_{t,i}}^{\infty} dP_D(D|i_{t-1}).$$

Therefore, we have

$$\mathbb{P}(\theta_{t,i} \geq x_t|i_{t-1}) \geq \int_{B - \theta_{t,i}}^{\infty} dP_D(D|i_{t-1}) \stackrel{(a)}{\geq} 1 - P_D(B|i_{t-1}) \geq \alpha_D > 0.$$

Inequality (a) holds because $\theta_{t,i} \geq 0$. Thus, we have

$$\mathbb{P}(i_t = i, \theta_{t,i} \geq x_t) = \mathbb{E}_{i_{t-1}}[p(i|i_{t-1}) \times \mathbb{P}(\theta_{t,i} \geq x_t|i_{t-1})] \geq \alpha_D \mathbb{E}_{i_{t-1}}[p(i|i_{t-1})] = \alpha_D \nu_i.$$

Step 3: KL condition of F_t . First from step 1, we have

$$F_t(\theta_t) - F_t(\theta_t^*) \leq \sum_{i \in \mathcal{I}} \nu_i \times [f_t(\theta_{t,i}|i) - f_t(\theta_{t,i}^*|i)].$$

Based on Assumption 4.4, we know that the per-period cost $L_t(\cdot|i)$ exhibits μ_D -strong convexity. Combining this with the convexity of the cost-to-go function $V_{t+1}^*(\cdot, i)$, we conclude that $f_t(\cdot|i)$ also possesses μ_D -strong convexity for any $i \in \mathcal{I}$. By Corollary 1, we know that strong convexity implies KL condition. Therefore,

$$\begin{aligned} F_t(\theta_t) - F_t(\theta_t^*) &\stackrel{(a)}{\leq} \sum_{i \in \mathcal{I}} \frac{\nu_i}{2\mu_D} \times \min_{g_{t,i} \in \partial \delta_{[0,B]}(\theta_{t,i})} |f'_t(\theta_{t,i}|i) + g_{t,i}|^2 \\ &\stackrel{(b)}{=} \sum_{i \in \mathcal{I}} \frac{\nu_i}{2\mu_D} \min_{g_{t,i} \in \partial \delta_{[0,B]}(\theta_{t,i})} \left| \frac{\nabla_{\theta_{t,i}} F_t(\theta_t)}{\mathbb{P}(i_t = i, \theta_{t,i} \geq x_t)} + g_{t,i} \right|^2 \\ &\stackrel{(c)}{=} \sum_{i \in \mathcal{I}} \frac{\nu_i}{2\mu_D \mathbb{P}(i_t = i, \theta_{t,i} \geq x_t)^2} \min_{g_{t,i} \in \partial \delta_{[0,B]}(\theta_{t,i})} |\nabla_{\theta_{t,i}} F_t(\theta_t) + g_{t,i}|^2 \\ &\stackrel{(d)}{\leq} \sum_{i \in \mathcal{I}} \frac{1}{2\mu_D \nu_i \alpha_D^2} \min_{g_{t,i} \in \partial \delta_{[0,B]}(\theta_{t,i})} |\nabla_{\theta_{t,i}} F_t(\theta_t) + g_{t,i}|^2 \\ &\stackrel{(e)}{\leq} \frac{1}{2\mu_D \alpha_D^2 \min_{i \in \mathcal{I}} \{\nu_i\}} \min_{g_t \in \partial \delta_{\Theta_t}(\theta_t)} \|\nabla F_t(\theta_t) + g_t\|_2^2. \end{aligned}$$

Here inequality (a) utilizes the KL condition of $f_t(\cdot|i)$. Equation (b) uses the relationship between different gradients in step 2. Equation (c) holds because $\partial \delta_{[0,B]}(\theta_{t,i})$ is a cone. Inequality (d) holds because $\mathbb{P}(i_t = i, \theta_{t,i} \geq x_t) \geq \alpha_D \nu_i$. Equation (e) is true since $\nu_i^{-1} \leq \min_{i \in \mathcal{I}} \{\nu_i\}^{-1}$. This completes the proof. \square

D.2. Gradient Formulation

Proof of Proposition 3 We first prove the recursive form for partial derivatives of value functions. From the Bellman equation (1), we have:

$$\begin{aligned}\nabla_x V_t^{\pi_\theta}(x_t, i_t) &= \frac{\partial}{\partial x_t} Q_t^{\pi_\theta}(x_t, i_t, \pi_t(x_t, i_t | \theta_t)) \\ &= \frac{\partial}{\partial x_t} \left(L_t(x_t \vee \theta_{t,i_t} | i_t) + \sum_{i_{t+1} \in \mathcal{I}} p(i_{t+1} | i_t) \mathbb{E}_{D_t \sim P_D(\cdot | i_t)} [V_{t+1}^{\pi_\theta}(x_t \vee \theta_{t,i_t} - D_t, i_{t+1})] \right) \\ &= \mathbf{1}(x_t \geq \theta_{t,i_t}) \times \left(L'_t(x_t | i_t) + \sum_{i_{t+1} \in \mathcal{I}} p(i_{t+1} | i_t) \mathbb{E}_{D_t \sim P_D(\cdot | i_t)} [\nabla_x V_{t+1}^{\pi_\theta}(x_t - D_t, i_{t+1})] \right),\end{aligned}$$

where $\nabla_x V_{T+1}^{\pi_\theta}(\cdot, \cdot) = 0$. The last equation uses the chain rule. Then for the policy gradient objective function $l(\theta)$, we calculate its partial derivative by:

$$\begin{aligned}\frac{\partial}{\partial \theta_{t,i}} l(\theta) &\stackrel{(a)}{=} \mathbb{E}_{(x_t, i_t) \sim \rho_t(\cdot | \pi_\theta)} \left[\frac{\partial}{\partial \theta_{t,i}} Q_t^{\pi_\theta}(x_t, i_t, \pi_t(x_t, i_t | \theta_t)) \right] \\ &\stackrel{(b)}{=} \mathbb{E}_{(x_t, i_t) \sim \rho_t(\cdot | \pi_\theta)} \left[\frac{\partial}{\partial \theta_{t,i}} \pi_t(x_t, i_t | \theta_t) \times \frac{\partial}{\partial a_t} Q_t^{\pi_\theta}(x_t, i_t, a_t) \Big|_{a_t = \pi_t(x_t, i_t | \theta_t)} \right] \\ &\stackrel{(c)}{=} \mathbb{E}_{(x_t, i_t) \sim \rho_t(\cdot | \pi_\theta)} \left[\mathbf{1}(i_t = i, \theta_{t,i} \geq x_t) \right. \\ &\quad \left. \times \left(L'_t(\theta_{t,i} | i) + \sum_{i_{t+1} \in \mathcal{I}} p(i_{t+1} | i) \mathbb{E}_{D_t \sim P_D(\cdot | i)} [\nabla_x V_{t+1}^{\pi_\theta}(\theta_{t,i} - D_t, i_{t+1})] \right) \right].\end{aligned}$$

Here equation (a) utilizes the Deterministic Policy Gradient Theorem (Silver et al. 2014). Equation (b) applies the chain rule. Equation (c) uses the explicit expression of $\pi_t(x_t, i_t | \theta_t) = (\theta_{t,i_t} - x_t)^+$ and the Bellman equation (16). This concludes the proof. \square

D.3. Bounded Gradient

Proof of Lemma 11 From Proposition 3, we bound the partial derivative as follows:

$$\begin{aligned}\left| \frac{\partial}{\partial \theta_{t,i}} l(\theta) \right| &= \left| \mathbb{P}(i_t = i, \theta_{t,i} \geq x_t) \times \left(L'_{t,i}(\theta_{t,i}) + \sum_{i_{t+1} \in \mathcal{I}} p(i_{t+1} | i) \mathbb{E}_{D_t \sim P_D(\cdot | i)} [\nabla_x V_{t+1}^{\pi_\theta}(\theta_{t,i} - D_t, i_{t+1})] \right) \right| \\ &\leq \mathbb{P}(i_t = i) \times \left| \left(L'_{t,i}(\theta_{t,i}) + \sum_{i_{t+1} \in \mathcal{I}} p(i_{t+1} | i) \mathbb{E}_{D_t \sim P_D(\cdot | i)} [\nabla_x V_{t+1}^{\pi_\theta}(\theta_{t,i} - D_t, i_{t+1})] \right) \right| \\ &\stackrel{(a)}{\leq} \nu_i \left(\left| L'_{t,i}(\theta_{t,i}) \right| + \sum_{i_{t+1} \in \mathcal{I}} p(i_{t+1} | i) \left| \mathbb{E}_{D_t \sim P_D(\cdot | i)} [\nabla_x V_{t+1}^{\pi_\theta}(\theta_{t,i} - D_t, i_{t+1})] \right| \right) \\ &\stackrel{(b)}{\leq} \nu_i \left(\max_{t \in \{1, \dots, T\}} \{ \max\{h_t, b_t\} \} + \sum_{i_{t+1} \in \mathcal{I}} p(i_{t+1} | i) \left| \mathbb{E}_{D_t \sim P_D(\cdot | i)} [\nabla_x V_{t+1}^{\pi_\theta}(\theta_{t,i} - D_t, i_{t+1})] \right| \right).\end{aligned}$$

Here inequality (a) employs the triangle inequality and inequality (b) holds because $|L'_{t,i}(\theta_{t,i})| \leq \max_{t \in \{1, \dots, T\}} \{\max\{h_t, b_t\}\}$ for any $\theta \in \Theta$, $t \in [T]$ and $i \in \mathcal{I}$. From (18), we have

$$\begin{aligned} |\nabla_x V_t^{\pi\theta}(x_t, i_t)| &= \left| \mathbf{1}(x_t \geq \theta_{t,i}) \times \left(L'_{t,i}(x_t) + \sum_{i_{t+1} \in \mathcal{I}} p(i_{t+1}|i_t) \mathbb{E}_{D_t \sim P_D(\cdot|i_t)} [\nabla_x V_{t+1}^{\pi\theta}(x_t - D_t, i_{t+1})] \right) \right| \\ &\leq |L'_{t,i}(x_t)| + \sum_{i_{t+1} \in \mathcal{I}} p(i_{t+1}|i_t) \left| \mathbb{E}_{D_t \sim P_D(\cdot|i_t)} [\nabla_x V_{t+1}^{\pi\theta}(x_t - D_t, i_{t+1})] \right| \\ &\leq \max_{t \in \{1, \dots, T\}} \{\max\{h_t, b_t\}\} + \sum_{i_{t+1} \in \mathcal{I}} p(i_{t+1}|i_t) \left| \mathbb{E}_{D_t \sim P_D(\cdot|i_t)} [\nabla_x V_{t+1}^{\pi\theta}(x_t - D_t, i_{t+1})] \right|. \end{aligned} \quad (34)$$

We use mathematical induction to prove $|\nabla_x V_t^{\pi\theta}(x_t, i_t)| \leq (T - t + 1) \max_{t \in \{1, \dots, T\}} \{\max\{h_t, b_t\}\}$ for any $\theta \in \Theta$, $t \in [T]$, $x_t \in (-\infty, B]$, and $i_t \in \mathcal{I}$.

Induction Base: As $\nabla_x V_{T+1}^{\pi\theta}(\cdot, \cdot) = 0$, it's obvious that

$$|\nabla_x V_T^{\pi\theta}(x_t, i_t)| \leq \max_{t \in \{1, \dots, T\}} \{\max\{h_t, b_t\}\}.$$

Induction Step: Suppose we have $|\nabla_x V_{t+1}^{\pi\theta}(x_{t+1}, i_{t+1})| \leq (T - t) \max_{t \in \{1, \dots, T\}} \{\max\{h_t, b_t\}\}$ for any $\theta \in \Theta$, $t \in [T]$, $x_{t+1} \in (-\infty, B]$, and $i_{t+1} \in \mathcal{I}$, applying (34) recursively, we have

$$\begin{aligned} |\nabla_x V_t^{\pi\theta}(x_t, i_t)| &\leq \max_{t \in \{1, \dots, T\}} \{\max\{h_t, b_t\}\} + \sum_{i_{t+1} \in \mathcal{I}} p(i_{t+1}|i_t) \left| \mathbb{E}_{D_t \sim P_D(\cdot|i_t)} [\nabla_x V_{t+1}^{\pi\theta}(x_t - D_t, i_{t+1})] \right| \\ &\leq \max_{t \in \{1, \dots, T\}} \{\max\{h_t, b_t\}\} + \sum_{i_{t+1} \in \mathcal{I}} p(i_{t+1}|i_t) (T - t) \max_{t \in \{1, \dots, T\}} \{\max\{h_t, b_t\}\} \\ &= (T - t + 1) \max_{t \in \{1, \dots, T\}} \{\max\{h_t, b_t\}\}. \end{aligned}$$

By mathematical induction, we have $|\nabla_x V_t^{\pi\theta}(x_t, i_t)| \leq (T - t + 1) \max_{t \in \{1, \dots, T\}} \{\max\{h_t, b_t\}\}$ for any $\theta \in \Theta$, $t \in [T]$, $x_t \in (-\infty, B]$, and $i_t \in \mathcal{I}$. Therefore, we conclude that

$$\begin{aligned} \left| \frac{\partial}{\partial \theta_{t,i}} l(\theta) \right| &\leq \nu_i \left(\max_{t \in \{1, \dots, T\}} \{\max\{h_t, b_t\}\} + \sum_{i_{t+1} \in \mathcal{I}} p(i_{t+1}|i) \left| \mathbb{E}_{D_t \sim P_D(\cdot|i)} [\nabla_x V_{t+1}^{\pi\theta}(\theta_{t,i} - D_t, i_{t+1})] \right| \right) \\ &\leq \nu_i (T - t + 1) \max_{t \in \{1, \dots, T\}} \{\max\{h_t, b_t\}\} \\ &\leq \nu_i T \max_{t \in \{1, \dots, T\}} \{\max\{h_t, b_t\}\} \end{aligned}$$

Thus, we obtain

$$\|\nabla_{\theta_t} l(\theta)\|_2 \leq \|\nabla_{\theta_t} l(\theta)\|_1 \leq \sum_{i \in \mathcal{I}} \nu_i T \max_{t \in \{1, \dots, T\}} \{\max\{h_t, b_t\}\} = T \max_{t \in \{1, \dots, T\}} \{\max\{h_t, b_t\}\}.$$

This completes the proof. \square

D.4. Sequential Decomposition Inequality

Proof of Lemma 12 For simplicity, we define $\theta_\alpha = (\theta_{[1:k]}, \theta_{[k+1:T]}^*)$ and $\theta_\beta = (\theta_{[1:k-1]}, \theta_{[k:T]}^*)$. Furthermore, we denote π_α and π_β as the policies deploying parameters θ_α and θ_β , respectively. Similar to the previous discussion, let π_θ denote the policy using parameters $\theta = (\theta_1, \dots, \theta_T)$. Then

$$\|\nabla_{\theta_t} l(\theta_\alpha) - \nabla_{\theta_t} l(\theta_\beta)\|_2 \leq \|\nabla_{\theta_t} l(\theta_\alpha) - \nabla_{\theta_t} l(\theta_\beta)\|_1 = \sum_{i \in \mathcal{I}} \left| \frac{\partial}{\partial \theta_{t,i}} l(\theta_\alpha) - \frac{\partial}{\partial \theta_{t,i}} l(\theta_\beta) \right|.$$

For any $i \in \mathcal{I}$, we can derive the following inequalities by Proposition 4:

$$\begin{aligned} & \left| \frac{\partial}{\partial \theta_{t,i}} l(\theta_\alpha) - \frac{\partial}{\partial \theta_{t,i}} l(\theta_\beta) \right| \\ &= \left| \mathbb{E}_{(x_t, i_t) \sim \rho_t(\cdot | \pi_\theta)} \left[\mathbf{1}(i_t = i, \theta_{t,i} \geq x_t) \right. \right. \\ & \quad \times \sum_{i_{t+1} \in \mathcal{I}} p(i_{t+1} | i) \left(\mathbb{E}_{D_t \sim P_D(\cdot | i)} [\nabla_x V_{t+1}^{\pi_\alpha}(\theta_{t,i} - D_t, i_{t+1})] - \mathbb{E}_{D_t \sim P_D(\cdot | i)} [\nabla_x V_{t+1}^{\pi_\beta}(\theta_{t,i} - D_t, i_{t+1})] \right) \left. \right] \Big| \\ &\leq \nu_i \sum_{i_{t+1} \in \mathcal{I}} p(i_{t+1} | i) \left| \mathbb{E}_{D_t \sim P_D(\cdot | i)} [\nabla_x V_{t+1}^{\pi_\alpha}(\theta_{t,i} - D_t, i_{t+1})] - \mathbb{E}_{D_t \sim P_D(\cdot | i)} [\nabla_x V_{t+1}^{\pi_\beta}(\theta_{t,i} - D_t, i_{t+1})] \right| \\ &\stackrel{(a)}{\leq} \nu_i \sum_{i_{t+1} \in \mathcal{I}} p(i_{t+1} | i) \sum_{i_{t+2} \in \mathcal{I}} p(i_{t+2} | i_{t+1}) \\ & \quad \times \left| \mathbb{E}_{D_{[t:t+1]}} [\nabla_x V_{t+2}^{\pi_\alpha}(\theta_{t,i} - D_{[t:t+1]}, i_{t+2})] - \mathbb{E}_{D_{[t:t+1]}} [\nabla_x V_{t+2}^{\pi_\beta}(\theta_{t,i} - D_{[t:t+1]}, i_{t+2})] \right| \\ &\leq \nu_i \sum_{i_{t+1} \in \mathcal{I}} p(i_{t+1} | i) \sum_{i_{t+2} \in \mathcal{I}} p(i_{t+2} | i_{t+1}) \sum_{i_{t+3} \in \mathcal{I}} p(i_{t+3} | i_{t+2}) \\ & \quad \times \left| \mathbb{E}_{D_{[t:t+2]}} [\nabla_x V_{t+3}^{\pi_\alpha}(\theta_{t,i} - D_{[t:t+2]}, i_{t+3})] - \mathbb{E}_{D_{[t:t+2]}} [\nabla_x V_{t+3}^{\pi_\beta}(\theta_{t,i} - D_{[t:t+2]}, i_{t+3})] \right| \\ &\quad \dots \\ &\stackrel{(b)}{\leq} \nu_i \sum_{i_{t+1} \in \mathcal{I}} p(i_{t+1} | i) \sum_{i_{t+2} \in \mathcal{I}} p(i_{t+2} | i_{t+1}) \cdots \sum_{i_k \in \mathcal{I}} p(i_k | i_{k-1}) \\ & \quad \times \left| \mathbb{E}_{D_{[t:k-1]}} [\nabla_x V_k^{\pi_\alpha}(\theta_{t,i} - D_{[t:k-1]}, i_k)] - \mathbb{E}_{D_{[t:k-1]}} [\nabla_x V_k^{\pi_\beta}(\theta_{t,i} - D_{[t:k-1]}, i_k)] \right|. \end{aligned}$$

Here inequality (a) applies (18) and utilizes $\mathbf{1}(\theta_{t,i} - D_t \geq \theta_{t+1, i_{t+1}}) \leq 1$. Inequality (b) holds by applying (18) recursively. Recall the definition of $\theta_\alpha = (\theta_{[1:k]}, \theta_{[k+1:T]}^*)$, $\theta_\beta = (\theta_{[1:k-1]}, \theta_{[k:T]}^*)$, and $f_t(\cdot | i)$. For any sample path (i, i_{t+1}, \dots, i_k) , we have

$$\begin{cases} \nabla_x V_k^{\pi_\alpha}(\theta_{t,i} - D_{[t:k-1]}, i_k) = \mathbf{1}(\theta_{t,i} - D_{[t:k-1]} \geq \theta_{k, i_k}) \times f'_k(\theta_{t,i} - D_{[t:k-1]} | i_k), \\ \nabla_x V_k^{\pi_\beta}(\theta_{t,i} - D_{[t:k-1]}, i_k) = \mathbf{1}(\theta_{t,i} - D_{[t:k-1]} \geq \theta_{k, i_k}^*) \times f'_k(\theta_{t,i} - D_{[t:k-1]} | i_k). \end{cases}$$

Without loss of generality, we assume that $\theta_{k,i_k} \leq \theta_{k,i_k}^*$. Then for any sample path (i, i_{t+1}, \dots, i_k) , $D_t \sim F_D(\cdot|i)$, and $D_j \sim F_D(\cdot|i_j), \forall t+1 \leq j \leq k-1$, we have

$$\begin{aligned} & \left| \mathbb{E}_{D_{[t:k-1]}} [\nabla_x V_k^{\pi\alpha}(\theta_{t,i} - D_{[t:k-1]}, i_k)] - \mathbb{E}_{D_{[t:k-1]}} [\nabla_x V_k^{\pi\beta}(\theta_{t,i} - D_{[t:k-1]}, i_k)] \right| \\ & \leq \mathbb{E}_{D_{[t:k-1]}} \left[\left| \mathbf{1}(\theta_{k,i_k} \leq \theta_{t,i} - D_{[t:k-1]} \leq \theta_{k,i_k}^*) \times f'_k(\theta_{t,i} - D_{[t:k-1]}|i_k) \right| \right] \\ & \stackrel{(a)}{=} \int_{\theta_{k,i_k}}^{\theta_{k,i_k}^*} -f'_k(x|i_k)\psi(x)dx, \end{aligned}$$

where ψ is the probability density function of $x := \theta_{t,i} - D_{[t:k-1]}$. Equation (a) holds because $f_k(\cdot|i_k)$ is convex, therefore $f'_k(\cdot|i_k) \leq 0$ within the interval $[\theta_{k,i_k}, \theta_{k,i_k}^*]$. From Assumption 4.2, the cumulative distribution function of the random demand D_t is L_D -Lipschitz continuous. Then the probability density function of D_t is upper bounded by L_D . Suppose $\psi_{D_1}(\cdot)$ and $\psi_{D_2}(\cdot)$ are probability density functions of D_1 and D_2 , we have the following inequalities:

$$\psi_{D_1+D_2}(\omega) = \int_0^\omega \psi_{D_1}(\nu)\psi_{D_2}(\omega-\nu)d\nu \leq L_D \int_0^\omega \psi_{D_1}(\nu)d\nu \leq L_D,$$

which implies that the probability density function of cumulative demands is upper bounded by L_D and thus $\psi(\cdot) \leq L_D$. Hence,

$$\int_{\theta_{k,i_k}}^{\theta_{k,i_k}^*} -f'_k(x|i_k)\psi(x)dx \leq L_D \int_{\theta_{k,i_k}}^{\theta_{k,i_k}^*} -f'_k(x|i_k)dx = L_D \left(f_k(\theta_{k,i_k}|i_k) - f_k(\theta_{k,i_k}^*|i_k) \right).$$

The same result holds when $\theta_k > \theta_k^*$ following a similar derivation. Therefore,

$$\begin{aligned} \left\| \nabla_{\theta_t} l(\theta_\alpha) - \nabla_{\theta_t} l(\theta_\beta) \right\|_2 & \leq \sum_{i \in \mathcal{I}} \left| \frac{\partial}{\partial \theta_{t,i}} l(\theta_\alpha) - \frac{\partial}{\partial \theta_{t,i}} l(\theta_\beta) \right| \\ & \leq L_D \sum_{i \in \mathcal{I}} \nu_i \sum_{i_{t+1} \in \mathcal{I}} p(i_{t+1}|i) \cdots \sum_{i_k \in \mathcal{I}} p(i_k|i_{k-1}) \left(f_k(\theta_{k,i_k}|i_k) - f_k(\theta_{k,i_k}^*|i_k) \right) \\ & = L_D \sum_{i \in \mathcal{I}} \nu_i \left(f_k(\theta_{k,i}|i) - f_k(\theta_{k,i}^*|i) \right). \end{aligned}$$

The last equation is true as ν is a stationary distribution of the exogenous Markov chain. Recalling the definition of $F_k(\theta_k)$, we have

$$F_k(\theta_k) - F_k(\theta_k^*) = \sum_{i \in \mathcal{I}} \nu_i \times \mathbb{E}_{x_k} \left[f_k(x_k \vee \theta_{k,i}|i) - f_k(x_k \vee \theta_{k,i}^*|i) \middle| i_k = i \right].$$

Without loss of generality, we assume that $\theta_{k,i} \leq \theta_{k,i}^*$. For any random variable ξ and its corresponding cumulative distribution function $P(\xi)$, we have

$$\begin{aligned} & \mathbb{E}_{\xi \sim P(\xi)} \left[f_k(\xi \vee \theta_{k,i}|i) - f_k(\xi \vee \theta_{k,i}^*|i) \right] \\ & = \int_{-\infty}^{\theta_{k,i}} \left(f_k(\theta_{k,i}|i) - f_k(\theta_{k,i}^*|i) \right) dP(\xi) + \int_{\theta_{k,i}}^{\theta_{k,i}^*} \left(f_k(\xi|i) - f_k(\theta_{k,i}^*|i) \right) dP(\xi) \\ & \geq \int_{-\infty}^{\theta_{k,i}} \left(f_k(\theta_{k,i}|i) - f_k(\theta_{k,i}^*|i) \right) dP(\xi). \end{aligned}$$

The last inequality holds as $f_k(\xi|i) \geq f_k(\theta_{k,i}|i)$ for any $\xi \in [\theta_{k,i}, \theta_{k,i}^*]$. Therefore, we have

$$\begin{aligned} F_k(\theta_k) - F_k(\theta_k^*) &\geq \sum_{i \in \mathcal{I}} \nu_i \int_{-\infty}^{\theta_{k,i}} dP(x_k|i_k=i) \left(f_t(\theta_{k,i}|i) - f_t(\theta_{k,i}^*|i) \right) \\ &= \sum_{i \in \mathcal{I}} \mathbb{P}(x_k \leq \theta_{k,i}, i_k=i) \left(f_k(\theta_{k,i}|i) - f_k(\theta_{k,i}^*|i) \right). \end{aligned}$$

Similar results hold when $\theta_{k,i} > \theta_{k,i}^*$. Following the same procedure in the proof of Lemma 10 step 2, we have $\mathbb{P}(x_k \leq \theta_{k,i}, i_k=i) \geq \alpha_D \nu_i$. Thus, we conclude that

$$\begin{aligned} \|\nabla_{\theta_t} l(\theta_\alpha) - \nabla_{\theta_t} l(\theta_\beta)\|_2 &\leq L_D \sum_{i \in \mathcal{I}} \nu_i \left(f_k(\theta_{k,i}|i) - f_k(\theta_{k,i}^*|i) \right) \\ &\leq \frac{L_D}{\alpha_D} \sum_{i \in \mathcal{I}} \mathbb{P}(x_k \leq \theta_{k,i}, i_k=i) \left(f_k(\theta_{k,i}|i) - f_k(\theta_{k,i}^*|i) \right) \\ &\leq \frac{L_D}{\alpha_D} (F_k(\theta_k) - F_k(\theta_k^*)). \end{aligned}$$

This completes the proof. \square

Appendix E: Omitted Proofs in Section 7

The proof for the stochastic cash balance problem in Section 7 shares some similarities with the inventory system in Section 6 yet the per-period decision is a two-dimensional vector. We demonstrate the full proof for completeness.

E.1. KL Condition of Optimal Q-value Function

Proof of Lemma 13 We divide the proof into three parts. First, we demonstrate the relationship between three suboptimality gaps $F_t(\theta_t) - F_t(\theta_t^*)$, $\underline{f}_t(\underline{\theta}_t) - \underline{f}_t(\underline{\theta}_t^*)$, and $\bar{f}_t(\bar{\theta}_t) - \bar{f}_t(\bar{\theta}_t^*)$. Next, we show how their gradients relate to each other. Finally, we prove the KL property of F_t .

Step 1: Relationship between suboptimality gaps. In the feasible region Θ_t , we have $\underline{\theta}_t \leq \bar{\theta}_t$. For function f_t , the following equations hold:

$$f_t((s_t \vee \underline{\theta}_t) \wedge \bar{\theta}_t) = f_t(s_t \vee \underline{\theta}_t) + f_t(s_t \wedge \bar{\theta}_t) - f_t(s_t). \quad (35)$$

It further holds that

$$\begin{aligned} &F_t(\theta_t) - F_t(\theta_t^*) \\ &= \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[Q_t^*(s_t, \pi_t(s_t|\theta_t)) - Q_t^*(s_t, \pi_t(s_t|\theta_t^*)) \right] \\ &= \mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[c((s_t \vee \underline{\theta}_t) \wedge \bar{\theta}_t, s_t) + f_t((s_t \vee \underline{\theta}_t) \wedge \bar{\theta}_t) - c((s_t \vee \underline{\theta}_t^*) \wedge \bar{\theta}_t^*, s_t) - f_t((s_t \vee \underline{\theta}_t^*) \wedge \bar{\theta}_t^*) \right] \\ &= \underbrace{\mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[c(s_t \vee \underline{\theta}_t, s_t) + f_t(s_t \vee \underline{\theta}_t) - c(s_t \vee \underline{\theta}_t^*, s_t) - f_t(s_t \vee \underline{\theta}_t^*) \right]}_{(I)} \\ &\quad + \underbrace{\mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[c(s_t \wedge \bar{\theta}_t, s_t) + f_t(s_t \wedge \bar{\theta}_t) - c(s_t \wedge \bar{\theta}_t^*, s_t) - f_t(s_t \wedge \bar{\theta}_t^*) \right]}_{(II)}, \end{aligned}$$

where the last equation comes from (35). We analyze the first term (I). Without loss of generality, we assume that $\underline{\theta}_t \leq \underline{\theta}_t^*$. With the expression of c , it holds that

$$\begin{aligned} \text{(I)} &= \int_{-\infty}^{\underline{\theta}_t} (\underline{f}_t(\underline{\theta}_t) - ks_t) d\rho_t(s_t|\pi_\theta) + \int_{\underline{\theta}_t}^{+\infty} f_t(s_t) d\rho_t(s_t|\pi_\theta) \\ &\quad - \int_{-\infty}^{\underline{\theta}_t^*} (\underline{f}_t(\underline{\theta}_t^*) - ks_t) d\rho_t(s_t|\pi_\theta) - \int_{\underline{\theta}_t^*}^{+\infty} f_t(s_t) d\rho_t(s_t|\pi_\theta) \\ &= \int_{-\infty}^{\underline{\theta}_t} (\underline{f}_t(\underline{\theta}_t) - \underline{f}_t(\underline{\theta}_t^*)) d\rho_t(s_t|\pi_\theta) + \int_{\underline{\theta}_t}^{\underline{\theta}_t^*} (f_t(s_t) + ks_t - \underline{f}_t(\underline{\theta}_t^*)) d\rho_t(s_t|\pi_\theta). \end{aligned}$$

For the right-hand-side, we have the following inequalities:

$$\begin{aligned} \int_{\underline{\theta}_t}^{\underline{\theta}_t^*} (f_t(s_t) + ks_t - \underline{f}_t(\underline{\theta}_t^*)) d\rho_t(s_t|\pi_\theta) &\stackrel{(a)}{=} \int_{\underline{\theta}_t}^{\underline{\theta}_t^*} (\underline{f}_t(s_t) - \underline{f}_t(\underline{\theta}_t^*)) d\rho_t(s_t|\pi_\theta) \\ &\stackrel{(b)}{\leq} \int_{\underline{\theta}_t}^{\underline{\theta}_t^*} (\underline{f}_t(\underline{\theta}_t) - \underline{f}_t(\underline{\theta}_t^*)) d\rho_t(s_t|\pi_\theta), \end{aligned}$$

where Equation (a) uses the definition of \underline{f}_t and f_t , and inequality (b) holds because \underline{f}_t is non-increasing on the interval $[\underline{\theta}_t, \underline{\theta}_t^*]$. Therefore, we conclude that $\text{(I)} \leq \underline{f}_t(\underline{\theta}_t) - \underline{f}_t(\underline{\theta}_t^*)$. The same result holds when $\underline{\theta}_t > \underline{\theta}_t^*$.

For the second term (II), we apply the same technique. Without loss of generality, we assume that $\bar{\theta}_t \leq \bar{\theta}_t^*$:

$$\begin{aligned} \text{(II)} &= \int_{-\infty}^{\bar{\theta}_t} f_t(s_t) d\rho_t(s_t|\pi_\theta) + \int_{\bar{\theta}_t}^{+\infty} (\bar{f}_t(\bar{\theta}_t) + qs_t) d\rho_t(s_t|\pi_\theta) \\ &\quad - \int_{-\infty}^{\bar{\theta}_t^*} f_t(s_t) d\rho_t(s_t|\pi_\theta) - \int_{\bar{\theta}_t^*}^{+\infty} (\bar{f}_t(\bar{\theta}_t^*) + qs_t) d\rho_t(s_t|\pi_\theta) \\ &= \int_{\bar{\theta}_t}^{\bar{\theta}_t^*} (\bar{f}_t(\bar{\theta}_t) + qs_t - f_t(s_t)) d\rho_t(s_t|\pi_\theta) + \int_{\bar{\theta}_t^*}^{+\infty} (\bar{f}_t(\bar{\theta}_t) - \bar{f}_t(\bar{\theta}_t^*)) d\rho_t(s_t|\pi_\theta). \end{aligned}$$

Similarly, it holds that

$$\begin{aligned} \int_{\bar{\theta}_t}^{\bar{\theta}_t^*} (\bar{f}_t(\bar{\theta}_t) + qs_t - f_t(s_t)) d\rho_t(s_t|\pi_\theta) &\stackrel{(a)}{=} \int_{\bar{\theta}_t}^{\bar{\theta}_t^*} (\bar{f}_t(s_t) - \bar{f}_t(\bar{\theta}_t^*)) d\rho_t(s_t|\pi_\theta) \\ &\stackrel{(b)}{\leq} \int_{\bar{\theta}_t}^{\bar{\theta}_t^*} (\bar{f}_t(\bar{\theta}_t) - \bar{f}_t(\bar{\theta}_t^*)) d\rho_t(s_t|\pi_\theta), \end{aligned}$$

where Equation (a) uses the definition of \bar{f}_t and f_t , and Inequality (b) holds because \bar{f}_t is non-increasing on the interval $[\bar{\theta}_t, \bar{\theta}_t^*]$. We thus have $\text{(II)} \leq \bar{f}_t(\bar{\theta}_t) - \bar{f}_t(\bar{\theta}_t^*)$. The same result holds when $\bar{\theta}_t > \bar{\theta}_t^*$. Combining all the results, we conclude that

$$F_t(\theta_t) - F_t(\theta_t^*) \leq \text{(I)} + \text{(II)} \leq \underline{f}_t(\underline{\theta}_t) - \underline{f}_t(\underline{\theta}_t^*) + \bar{f}_t(\bar{\theta}_t) - \bar{f}_t(\bar{\theta}_t^*)$$

Step 2: Relationship between gradients. By definition, we calculate the gradient of F_t . We first show the partial derivative for $\underline{\theta}$ using the definition of F_t , f_t , \underline{f}_t , and \bar{f}_t .

$$\begin{aligned}
\nabla_{\underline{\theta}_t} F_t(\theta_t) &= \nabla_{\underline{\theta}_t} \mathbb{E}_{s_t \sim \rho_t(\cdot | \pi_\theta)} \left[Q_t^*(s_t, \pi_t(s_t | \theta_t)) \right] \\
&= \nabla_{\underline{\theta}_t} \left[\int_{\underline{B}} (\underline{f}_t(\underline{\theta}_t) - k s_t) d\rho_t(s_t | \pi_\theta) + \int_{\underline{\theta}_t}^{\bar{\theta}_t} f_t(s_t) d\rho_t(s_t | \pi_\theta) + \int_{\bar{\theta}_t}^{\bar{B}} (\bar{f}_t(\bar{\theta}_t) + q s_t) d\rho_t(s_t | \pi_\theta) \right] \\
&\stackrel{(a)}{=} \underline{f}_t(\underline{\theta}_t) - k \underline{\theta}_t + \int_{\underline{B}} \underline{f}_t'(\underline{\theta}_t) d\rho_t(s_t | \pi_\theta) - f_t(\underline{\theta}_t) \\
&= \mathbb{P}(s_t \leq \underline{\theta}_t) \underline{f}_t'(\underline{\theta}_t).
\end{aligned}$$

Equation (a) uses the Leibniz rule. Similarly, we derive $\nabla_{\bar{\theta}_t} F_t(\theta_t) = \mathbb{P}(s_t \geq \bar{\theta}_t) \bar{f}_t'(\bar{\theta}_t)$.

Step 3: KL Condition of F_t . By Assumption 5.4, the per-period holding or backlogging cost is μ_D -strongly convex over $[\underline{B}, \bar{B}]$. By the convexity of cost-to-go functions, we have that \underline{f}_t and \bar{f}_t are both μ_D -strongly convex over $[\underline{B}, \bar{B}]$. Therefore, we can derive

$$\begin{aligned}
F_t(\theta_t) - F_t(\theta_t^*) &\leq \underline{f}_t(\underline{\theta}_t) - \underline{f}_t(\underline{\theta}_t^*) + \bar{f}_t(\bar{\theta}_t) - \bar{f}_t(\bar{\theta}_t^*) \\
&\stackrel{(a)}{\leq} \underline{f}_t'(\underline{\theta}_t)(\underline{\theta}_t - \underline{\theta}_t^*) - \frac{\mu_D}{2} \|\underline{\theta}_t - \underline{\theta}_t^*\|_2^2 + \bar{f}_t'(\bar{\theta}_t)(\bar{\theta}_t - \bar{\theta}_t^*) - \frac{\mu_D}{2} \|\bar{\theta}_t - \bar{\theta}_t^*\|_2^2 \\
&= \frac{\nabla_{\underline{\theta}_t} F_t(\theta_t)}{\mathbb{P}(s_t \leq \underline{\theta}_t)} (\underline{\theta}_t - \underline{\theta}_t^*) - \frac{\mu_D}{2} \|\underline{\theta}_t - \underline{\theta}_t^*\|_2^2 + \frac{\nabla_{\bar{\theta}_t} F_t(\theta_t)}{\mathbb{P}(s_t \geq \bar{\theta}_t)} (\bar{\theta}_t - \bar{\theta}_t^*) - \frac{\mu_D}{2} \|\bar{\theta}_t - \bar{\theta}_t^*\|_2^2 \\
&\stackrel{(b)}{\leq} \alpha_D^{-1} \left\langle \nabla_{\theta_t} F_t(\theta_t), \begin{bmatrix} \underline{\theta}_t \\ \bar{\theta}_t \end{bmatrix} - \begin{bmatrix} \underline{\theta}_t^* \\ \bar{\theta}_t^* \end{bmatrix} \right\rangle - \frac{\mu_D}{2} \left\| \begin{bmatrix} \underline{\theta}_t \\ \bar{\theta}_t \end{bmatrix} - \begin{bmatrix} \underline{\theta}_t^* \\ \bar{\theta}_t^* \end{bmatrix} \right\|_2^2 \\
&\stackrel{(c)}{\leq} \max_{\theta_t' \in \Theta_t} \left\{ \alpha_D^{-1} \langle \nabla_{\theta_t} F_t(\theta_t), \theta_t - \theta_t' \rangle - \frac{\mu_D}{2} \|\theta_t - \theta_t'\|_2^2 \right\}.
\end{aligned}$$

Inequality (a) uses the strong convexity of \underline{f}_t and \bar{f}_t . The equality holds by the explicit expression derived in Step 2. Inequality (b) holds because $\mathbb{P}(s_t \leq \underline{\theta}_t) \geq \alpha_D > 0$ and $\mathbb{P}(s_t \geq \bar{\theta}_t) \geq \alpha_D > 0$. Inequality (c) utilizes the fact that $\theta^* \in \Theta_t$. Therefore, $F_t(\theta_t)$ satisfies the (α_D^{-1}, μ_D) gradient dominance condition, and thus the KL condition with constant $\mu_D \alpha_D^2$ by Lemma 16. This completes the proof. \square

E.2. Gradient Formulation

Proof of Proposition 4 By the Bellman equation (1), we derive the recursive form of $(V_t^{\pi_\theta})'(s_t)$ for any $t \in [T]$:

$$\begin{aligned}
(V_t^{\pi_\theta})'(s_t) &= \frac{\partial}{\partial s_t} Q_t^{\pi_\theta}(s_t, \pi_t(s_t | \theta_t)) \\
&= \frac{\partial}{\partial s_t} \left(c((s_t \vee \underline{\theta}_t) \wedge \bar{\theta}_t, s_t) + L_t((s_t \vee \underline{\theta}_t) \wedge \bar{\theta}_t) + \mathbb{E}_{D_t} \left[(V_{t+1}^{\pi_\theta})'((s_t \vee \underline{\theta}_t) \wedge \bar{\theta}_t - D_t) \right] \right) \\
&= -k \mathbf{1}(s_t \leq \underline{\theta}_t) + q \mathbf{1}(s_t \geq \bar{\theta}_t) + \mathbf{1}(\underline{\theta}_t < s_t < \bar{\theta}_t) \times \left(L_t'(s_t) + \mathbb{E}_{D_t} \left[(V_{t+1}^{\pi_\theta})'(s_t - D_t) \right] \right)
\end{aligned}$$

with $(V_{T+1}^{\pi_\theta})'(\cdot) = 0$. For the policy gradient objective function $l(\theta)$, we calculate the partial derivative

$$\begin{aligned}
\frac{\partial}{\partial \underline{\theta}_t} l(\theta) &\stackrel{(a)}{=} \mathbb{E}_{s_t \sim \rho_t(\cdot | \pi_\theta)} \left[\frac{\partial}{\partial \underline{\theta}_t} Q_t^{\pi_\theta}(s_t, \pi_t(s_t | \theta_t)) \right] \\
&\stackrel{(b)}{=} \mathbb{E}_{s_t \sim \rho_t(\cdot | \pi_\theta)} \left[\frac{\partial}{\partial \underline{\theta}_t} \pi_t(s_t | \theta_t) \times \frac{\partial}{\partial a_t} Q_t^{\pi_\theta}(s_t, a_t) \Big|_{a_t = \pi_t(s_t | \theta_t)} \right] \\
&\stackrel{(c)}{=} \mathbb{E}_{s_t \sim \rho_t(\cdot | \pi_\theta)} \left[\mathbf{1}(\underline{\theta}_t \geq s_t) \times \frac{\partial}{\partial a_t} \left(C_t(s_t, a_t) + \mathbb{E}_{D_t} [V_{t+1}^{\pi_\theta}(s_t + a_t - D_t)] \right) \Big|_{a_t = \pi_t(s_t | \theta_t)} \right] \\
&\stackrel{(d)}{=} \mathbb{E}_{s_t \sim \rho_t(\cdot | \pi_\theta)} \left[\mathbf{1}(\underline{\theta}_t \geq s_t) \times \left(k + L'_t(\underline{\theta}_t) + \mathbb{E}_{D_t} [(V_{t+1}^{\pi_\theta})'(\underline{\theta}_t - D_t)] \right) \right].
\end{aligned}$$

Equation (a) utilizes the Deterministic Policy Gradient Theorem (Silver et al. 2014). Equation (b) applies the chain rule. Equation (c) uses the Bellman equation (1). Lastly, equation (d) holds because $\mathbf{1}(\underline{\theta}_t \geq s_t)$. Similarly, we can calculate the partial derivative

$$\frac{\partial}{\partial \bar{\theta}_t} l(\theta) = \mathbb{E}_{s_t \sim \rho_t(\cdot | \pi_\theta)} \left[\mathbf{1}(\bar{\theta}_t \leq s_t) \times \left(-q + L'_t(\bar{\theta}_t) + \mathbb{E}_{D_t} [(V_{t+1}^{\pi_\theta})'(\bar{\theta}_t - D_t)] \right) \right].$$

This concludes the proof. \square

E.3. Bounded Gradient

Proof of Lemma 14 Following Proposition 4, we can bound the partial derivative as follows:

$$\begin{aligned}
\left| \frac{\partial}{\partial \underline{\theta}_t} l(\theta) \right| &= \left| \mathbb{E}_{s_t \sim \rho_t(\cdot | \pi_\theta)} \left[\mathbf{1}(\underline{\theta}_t \geq s_t) \times \left(k + L'_t(\underline{\theta}_t) + \mathbb{E}_{D_t} [(V_{t+1}^{\pi_\theta})'(\underline{\theta}_t - D_t)] \right) \right] \right| \\
&\stackrel{(a)}{\leq} k + |L'_t(\underline{\theta}_t)| + |\mathbb{E}_{D_t} [(V_{t+1}^{\pi_\theta})'(\underline{\theta}_t - D_t)]| \\
&\stackrel{(b)}{\leq} k + \max_{t \in \{1, \dots, T\}} \{ \max\{h_t, b_t\} \} + |\mathbb{E}_{D_t} [(V_{t+1}^{\pi_\theta})'(\underline{\theta}_t - D_t)]|.
\end{aligned}$$

Inequality (a) employs the triangle inequality and utilizes the fact that $\mathbf{1}(\underline{\theta}_t \geq s_t) \leq 1$. Inequality (b) holds because $|L'_t(\underline{\theta}_t)| \leq \max_{t \in \{1, \dots, T\}} \{ \max\{h_t, b_t\} \}$ for any θ_t and $t \in [T]$. From (20), we have

$$\begin{aligned}
|(V_t^{\pi_\theta})'(s_t)| &= \left| -k \mathbf{1}(s_t \leq \underline{\theta}_t) + q \mathbf{1}(s_t \geq \bar{\theta}_t) + \mathbf{1}(\underline{\theta}_t < s_t < \bar{\theta}_t) \times \left(L'_t(s_t) + \mathbb{E}_{D_t} [(V_{t+1}^{\pi_\theta})'(s_t - D_t)] \right) \right| \\
&\leq k + |q| + |L'_t(s_t)| + |\mathbb{E}_{D_t} [(V_{t+1}^{\pi_\theta})'(\underline{\theta}_t - D_t)]| \\
&\leq k + |q| + \max_{t \in \{1, \dots, T\}} \{ \max\{h_t, b_t\} \} + |\mathbb{E}_{D_t} [(V_{t+1}^{\pi_\theta})'(\underline{\theta}_t - D_t)]|.
\end{aligned} \tag{36}$$

Applying (36) recursively, we derive

$$\begin{aligned}
\left| \frac{\partial}{\partial \underline{\theta}_t} l(\theta) \right| &\leq k + \max_{t \in \{1, \dots, T\}} \{ \max\{h_t, b_t\} \} + |\mathbb{E}_{D_t} [(V_{t+1}^{\pi_\theta})'(\underline{\theta}_t - D_t)]| \\
&\leq (k + |q| + \max_{t \in \{1, \dots, T\}} \{ \max\{h_t, b_t\} \}) T.
\end{aligned}$$

Similarly, we have

$$\left| \frac{\partial}{\partial \bar{\theta}_t} l(\theta) \right| \leq (k + |q| + \max_{t \in \{1, \dots, T\}} \{ \max\{h_t, b_t\} \}) T.$$

Thus, we obtain

$$\|\nabla_{\theta_t} l(\theta)\|_2 \leq \|\nabla_{\theta_t} l(\theta)\|_1 \leq 2(k + |q| + \max_{t \in \{1, \dots, T\}} \{\max\{h_t, b_t\}\})T.$$

This completes the proof. \square

E.4. Sequential Decomposition Inequality

Proof of Lemma 15 For simplicity, we define $\theta_\alpha = (\theta_{[1:k]}, \theta_{[k+1:T]}^*)$ and $\theta_\beta = (\theta_{[1:k-1]}, \theta_{[k:T]}^*)$. Furthermore, we denote π_α and π_β as the policies deploying parameters θ_α and θ_β , respectively. Let π_θ denote the policy using parameters $\theta = (\theta_1, \dots, \theta_T)$. Then

$$\|\nabla_{\theta_t} l(\theta_\alpha) - \nabla_{\theta_t} l(\theta_\beta)\|_2 \leq \|\nabla_{\theta_t} l(\theta_\alpha) - \nabla_{\theta_t} l(\theta_\beta)\|_1 = \underbrace{\left| \frac{\partial}{\partial \underline{\theta}_t} l(\theta_\alpha) - \frac{\partial}{\partial \underline{\theta}_t} l(\theta_\beta) \right|}_{(I)} + \underbrace{\left| \frac{\partial}{\partial \bar{\theta}_t} l(\theta_\alpha) - \frac{\partial}{\partial \bar{\theta}_t} l(\theta_\beta) \right|}_{(II)}.$$

For the first part (I), we can derive the following inequality by Proposition 4,

$$\begin{aligned} (I) &= \left| \mathbb{E}_{s_t \sim \rho_t(\cdot | \pi_\theta)} \left[\mathbf{1}(\underline{\theta}_t \geq s_t) \times \left(\mathbb{E}_{D_t} [(V_{t+1}^{\pi_\alpha})'(\underline{\theta}_t - D_t)] - \mathbb{E}_{D_t} [(V_{t+1}^{\pi_\beta})'(\underline{\theta}_t - D_t)] \right) \right] \right| \\ &\leq \left| \mathbb{E}_{D_t} [(V_{t+1}^{\pi_\alpha})'(\underline{\theta}_t - D_t)] - \mathbb{E}_{D_t} [(V_{t+1}^{\pi_\beta})'(\underline{\theta}_t - D_t)] \right|. \end{aligned}$$

Applying (20) recursively, we have

$$\begin{aligned} (I) &\leq \left| \mathbb{E}_{D_t} [(V_{t+1}^{\pi_\alpha})'(\underline{\theta}_t - D_t)] - \mathbb{E}_{D_t} [(V_{t+1}^{\pi_\beta})'(\underline{\theta}_t - D_t)] \right| \\ &= \left| \mathbb{E}_{D_t} \left[k \mathbf{1}(\underline{\theta}_t - D_t \leq \underline{\theta}_{t+1}) - q \mathbf{1}(\underline{\theta}_t - D_t \geq \bar{\theta}_{t+1}) \right. \right. \\ &\quad \left. \left. + \mathbf{1}(\underline{\theta}_{t+1} < \underline{\theta}_t - D_t < \bar{\theta}_{t+1}) \times \left(L'_{t+1}(\underline{\theta}_t - D_t) + \mathbb{E}_{D_{t+1}} [(V_{t+2}^{\pi_\alpha})'(\underline{\theta}_t - D_{[t:t+1]})] \right) \right] \right. \\ &\quad \left. - \mathbb{E}_{D_t} \left[k \mathbf{1}(\underline{\theta}_t - D_t \leq \underline{\theta}_{t+1}) - q \mathbf{1}(\underline{\theta}_t - D_t \geq \bar{\theta}_{t+1}) \right. \right. \\ &\quad \left. \left. + \mathbf{1}(\underline{\theta}_{t+1} < \underline{\theta}_t - D_t < \bar{\theta}_{t+1}) \times \left(L'_{t+1}(\underline{\theta}_t - D_t) + \mathbb{E}_{D_{t+1}} [(V_{t+2}^{\pi_\beta})'(\underline{\theta}_t - D_{[t:t+1]})] \right) \right] \right| \\ &\leq \mathbb{E}_{D_{[t:t+1]}} \left[\left| (V_{t+2}^{\pi_\alpha})'(\underline{\theta}_t - D_{[t:t+1]}) - (V_{t+2}^{\pi_\beta})'(\underline{\theta}_t - D_{[t:t+1]}) \right| \right] \\ &\dots \\ &\leq \mathbb{E}_{D_{[t:k-1]}} \left[\left| (V_k^{\pi_\alpha})'(\underline{\theta}_t - D_{[t:k-1]}) - (V_k^{\pi_\beta})'(\underline{\theta}_t - D_{[t:k-1]}) \right| \right]. \end{aligned} \tag{37}$$

Therefore, we can derive the following inequality using (20):

$$\begin{aligned} (V_k^{\pi_\alpha})'(\underline{\theta}_t - D_{[t:k-1]}) &= -k \mathbf{1}(\underline{\theta}_t - D_{[t:k-1]} \leq \underline{\theta}_k) + q \mathbf{1}(\underline{\theta}_t - D_{[t:k-1]} \geq \bar{\theta}_k) \\ &\quad + \mathbf{1}(\underline{\theta}_k < \underline{\theta}_t - D_{[t:k-1]} < \bar{\theta}_k) \times \left(L'_k(\underline{\theta}_t - D_{[t:k-1]}) + \mathbb{E}_{D_k} [(V_{k+1}^{\pi_\alpha})'(\underline{\theta}_t - D_{[t:k]})] \right) \\ &\stackrel{(a)}{=} -k \mathbf{1}(\underline{\theta}_t - D_{[t:k-1]} \leq \underline{\theta}_k) + q \mathbf{1}(\underline{\theta}_t - D_{[t:k-1]} \geq \bar{\theta}_k) \\ &\quad + \mathbf{1}(\underline{\theta}_k < \underline{\theta}_t - D_{[t:k-1]} < \bar{\theta}_k) \times \left(L'_k(\underline{\theta}_t - D_{[t:k-1]}) + \mathbb{E}_{D_k} [(V_{k+1}^*)'(\underline{\theta}_t - D_{[t:k]})] \right) \\ &\stackrel{(b)}{=} -k \mathbf{1}(\underline{\theta}_t - D_{[t:k-1]} \leq \underline{\theta}_k) + q \mathbf{1}(\underline{\theta}_t - D_{[t:k-1]} \geq \bar{\theta}_k) \\ &\quad + \mathbf{1}(\underline{\theta}_k < \underline{\theta}_t - D_{[t:k-1]} < \bar{\theta}_k) \times (f'_k(\underline{\theta}_t - D_{[t:k-1]})). \end{aligned}$$

Here Equation (a) holds because π_α uses optimal $\theta_{[k+1:T]}^*$ starting from period $k+1$, and Equation (b) comes from the definition of f_k . Again, by definitions of \underline{f}_k and \bar{f}_k , we have

$$\begin{aligned} (V_k^{\pi_\alpha})'(\underline{\theta}_t - D_{[t:k-1]}) &= -k\mathbf{1}(\underline{\theta}_t - D_{[t:k-1]} \leq \underline{\theta}_k) + q\mathbf{1}(\underline{\theta}_t - D_{[t:k-1]} \geq \bar{\theta}_k) \\ &\quad + \mathbf{1}(\underline{\theta}_k < \underline{\theta}_t - D_{[t:k-1]} < \bar{\theta}_k) \times (f'_k(\underline{\theta}_t - D_{[t:k-1]})) \\ &= f'_k(\underline{\theta}_t - D_{[t:k-1]}) - \mathbf{1}(\underline{\theta}_t - D_{[t:k-1]} \leq \underline{\theta}_k) \times \underline{f}'_k(\underline{\theta}_t - D_{[t:k-1]}) \\ &\quad - \mathbf{1}(\underline{\theta}_t - D_{[t:k-1]} \geq \bar{\theta}_k) \times \bar{f}'_k(\underline{\theta}_t - D_{[t:k-1]}). \end{aligned} \quad (38)$$

Similarly, we can derive that

$$\begin{aligned} (V_k^{\pi_\beta})'(\underline{\theta}_t - D_{[t:k-1]}) &= f'_k(\underline{\theta}_t - D_{[t:k-1]}) - \mathbf{1}(\underline{\theta}_t - D_{[t:k-1]} \leq \underline{\theta}_k^*) \times \underline{f}'_k(\underline{\theta}_t - D_{[t:k-1]}) \\ &\quad - \mathbf{1}(\underline{\theta}_t - D_{[t:k-1]} \geq \bar{\theta}_k^*) \times \bar{f}'_k(\underline{\theta}_t - D_{[t:k-1]}). \end{aligned} \quad (39)$$

Plugging the results of (38) and (39) into (37), we have

$$\begin{aligned} (I) &\leq \mathbb{E}_{D_{[t:k-1]}} \left[\left| (V_k^{\pi_\alpha})'(\underline{\theta}_t - D_{[t:k-1]}) - (V_k^{\pi_\beta})'(\underline{\theta}_t - D_{[t:k-1]}) \right| \right] \\ &= \mathbb{E}_{D_{[t:k-1]}} \left[\left| \underline{f}'_k(\underline{\theta}_t - D_{[t:k-1]}) \times (\mathbf{1}(\underline{\theta}_t - D_{[t:k-1]} \leq \underline{\theta}_k) - \mathbf{1}(\underline{\theta}_t - D_{[t:k-1]} \leq \underline{\theta}_k^*)) \right. \right. \\ &\quad \left. \left. + \bar{f}'_k(\underline{\theta}_t - D_{[t:k-1]}) \times (\mathbf{1}(\underline{\theta}_t - D_{[t:k-1]} \geq \bar{\theta}_k) - \mathbf{1}(\underline{\theta}_t - D_{[t:k-1]} \geq \bar{\theta}_k^*)) \right| \right] \\ &\leq \underbrace{\mathbb{E}_{D_{[t:k-1]}} \left[\left| \underline{f}'_k(\underline{\theta}_t - D_{[t:k-1]}) \times (\mathbf{1}(\underline{\theta}_t - D_{[t:k-1]} \leq \underline{\theta}_k) - \mathbf{1}(\underline{\theta}_t - D_{[t:k-1]} \leq \underline{\theta}_k^*)) \right| \right]}_{(III)} \\ &\quad + \underbrace{\mathbb{E}_{D_{[t:k-1]}} \left[\left| \bar{f}'_k(\underline{\theta}_t - D_{[t:k-1]}) \times (\mathbf{1}(\underline{\theta}_t - D_{[t:k-1]} \geq \bar{\theta}_k) - \mathbf{1}(\underline{\theta}_t - D_{[t:k-1]} \geq \bar{\theta}_k^*)) \right| \right]}_{(IV)}. \end{aligned}$$

For part (III), without loss of generality, we assume $\underline{\theta}_k \leq \underline{\theta}_k^*$. Then

$$(III) = \mathbb{E}_{D_{[t:k-1]}} \left[\left| \underline{f}'_k(\underline{\theta}_t - D_{[t:k-1]}) \times \mathbf{1}(\underline{\theta}_k < \underline{\theta}_t - D_{[t:k-1]} \leq \underline{\theta}_k^*) \right| \right] \stackrel{(a)}{=} \int_{\underline{\theta}_k}^{\underline{\theta}_k^*} -\underline{f}'_k(x) \psi(x) dx,$$

where ψ is the probability density function of $\underline{\theta}_t - D_{[t:k-1]}$. Equation (a) holds because $\underline{f}_k(\cdot)$ is convex, $\underline{\theta}_k^*$ is its minimizer, and it uses the variable change $x = \underline{\theta}_t - D_{[t:k-1]}$. From Assumption 5.4, the cumulative distribution function of the random demand D_t is L_D -Lipschitz continuous. Then the probability density function of D_t is upper bounded by L_D . Using a similar derivation in Section 6, the probability density function of cumulative demands is upper bounded by L_D and thus $\psi(\cdot) \leq L_D$. Hence,

$$(III) \leq L_D \int_{\underline{\theta}_k}^{\underline{\theta}_k^*} -\underline{f}'_k(x) dx = L_D \left(\underline{f}_k(\underline{\theta}_k) - \underline{f}_k(\underline{\theta}_k^*) \right).$$

The same result holds when $\underline{\theta}_k > \underline{\theta}_k^*$. As for part (IV), we can derive a similar bound

$$(IV) \leq L_D \left(\bar{f}_k(\bar{\theta}_k) - \bar{f}_k(\bar{\theta}_k^*) \right).$$

As a result, we have

$$(I) \leq (III) + (IV) \leq L_D \left(\underline{f}_k(\underline{\theta}_k) - \underline{f}_k(\underline{\theta}_k^*) \right) + L_D \left(\bar{f}_k(\bar{\theta}_k) - \bar{f}_k(\bar{\theta}_k^*) \right).$$

For the second part (II), we can derive a similar bound using the same technique.

$$(II) \leq L_D \left(\underline{f}_k(\underline{\theta}_k) - \underline{f}_k(\underline{\theta}_k^*) \right) + L_D \left(\bar{f}_k(\bar{\theta}_k) - \bar{f}_k(\bar{\theta}_k^*) \right).$$

This implies that

$$\|\nabla_{\theta_t} l(\theta_\alpha) - \nabla_{\theta_t} l(\theta_\beta)\|_2 \leq (I) + (II) \leq 2L_D \left(\underline{f}_k(\underline{\theta}_k) - \underline{f}_k(\underline{\theta}_k^*) + \bar{f}_k(\bar{\theta}_k) - \bar{f}_k(\bar{\theta}_k^*) \right). \quad (40)$$

Recalling the definition of $F_k(\theta_k)$, we have

$$\begin{aligned} F_k(\theta_k) - F_k(\theta_k^*) &= \underbrace{\mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[c(s_t \vee \underline{\theta}_t, s_t) + f_t(s_t \vee \underline{\theta}_t) - c(s_t \vee \underline{\theta}_t^*, s_t) - f_t(s_t \vee \underline{\theta}_t^*) \right]}_{(I)} \\ &\quad + \underbrace{\mathbb{E}_{s_t \sim \rho_t(\cdot|\pi_\theta)} \left[c(s_t \wedge \bar{\theta}_t, s_t) + f_t(s_t \wedge \bar{\theta}_t) - c(s_t \wedge \bar{\theta}_t^*, s_t) - f_t(s_t \wedge \bar{\theta}_t^*) \right]}_{(II)}. \end{aligned} \quad (41)$$

Without loss of generality, we assume that $\underline{\theta}_t \leq \underline{\theta}_t^*$. Then we have

$$\begin{aligned} (I) &= \int_{-\infty}^{\underline{\theta}_t} (\underline{f}_t(\underline{\theta}_t) - \underline{f}_t(\underline{\theta}_t^*)) d\rho_t(s_t|\pi_\theta) + \int_{\underline{\theta}_t}^{\underline{\theta}_t^*} (f_t(s_t) + ks_t - \underline{f}_t(\underline{\theta}_t^*)) d\rho_t(s_t|\pi_\theta) \\ &\geq \int_{-\infty}^{\underline{\theta}_t} (\underline{f}_t(\underline{\theta}_t) - \underline{f}_t(\underline{\theta}_t^*)) d\rho_t(s_t|\pi_\theta). \end{aligned}$$

The last inequality holds as $\underline{f}_t(s_t) \geq \underline{f}_t(\underline{\theta}_t^*)$ for any $s_t \in [\underline{\theta}_t, \underline{\theta}_t^*]$. Therefore, we have

$$(I) \geq \mathbb{P}(s_t \leq \underline{\theta}_t) (\underline{f}_t(\underline{\theta}_t) - \underline{f}_t(\underline{\theta}_t^*)) \geq \alpha_D (\underline{f}_t(\underline{\theta}_t) - \underline{f}_t(\underline{\theta}_t^*)). \quad (42)$$

Similar results hold when $\theta_{k,i} > \theta_{k,i}^*$. For term (II), we apply the same technique and derive that

$$(II) \geq \alpha_D (\bar{f}_t(\bar{\theta}_t) - \bar{f}_t(\bar{\theta}_t^*)). \quad (43)$$

Combining (40), (41), (42), and (43), we conclude that

$$\begin{aligned} \|\nabla_{\theta_t} l(\theta_\alpha) - \nabla_{\theta_t} l(\theta_\beta)\|_2 &\leq 2L_D \left(\underline{f}_k(\underline{\theta}_k) - \underline{f}_k(\underline{\theta}_k^*) + \bar{f}_k(\bar{\theta}_k) - \bar{f}_k(\bar{\theta}_k^*) \right) \\ &\leq \frac{L_D}{\alpha_D} (F_k(\theta_k) - F_k(\theta_k^*)). \end{aligned}$$

This concludes the proof. \square