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

Xin Chen
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
xin.chen@isye.gatech.edu

Yifan Hu
Department of Statistics
Rutgers University
yifan.hu@rutgers.edu

Minda Zhao
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
mindazhao@gatech.edu

Abstract

Policy gradient methods are widely used in reinforcement learning. Yet, the non-convexity of policy optimization imposes significant challenges in understanding the global convergence of policy gradient methods. For a class of finite-horizon MDPs with general state and action spaces, we develop a framework that provides a set of easily verifiable assumptions to ensure the Polyak-Łojasiewicz-Kurdyka (PŁK) condition of the policy optimization. Leveraging the PŁK 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 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 $\tilde{\mathcal{O}}(\epsilon^{-1})$ 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.

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 RL problems is policy gradient methods, which search over a parameterized policy space by applying first-order methods on the total expected cost of a Markov Decision Process (MDP). Despite wide applicability, the understanding of the global convergence and non-asymptotic convergence behavior of policy gradient methods remains limited due to the nonconvexity of the policy gradient optimization problem (Agarwal et al., 2021; Bhandari & Russo, 2024).

To address this gap, this paper seeks to understand the nonconvex landscape of the policy gradient optimization problem and establish non-asymptotic convergence rates for policy gradient methods. Unlike the tabular setting, solving MDPs with general state and action spaces presents significant challenges, as it is generally impossible to enumerate all the states and actions. To resolve this issue, we focus on a subset of MDPs where the optimal policy can be characterized by finite-dimensional parameters, as observed in various control and operations models. We aim to establish the Polyak-Łojasiewicz-Kurdyka (PŁK) condition (Polyak et al., 1963; Lojasiewicz, 1963; Kurdyka, 1998) of the

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

policy gradient optimization problem for such MDPs. Informally, the PŁK 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 & Wright, 1999) is globally optimal. Policy gradient methods are designed to find these stationary points and thus converge globally on nonconvex MDP problems.

A key challenge in establishing the PŁK condition across various problems lies in the need for case-specific analysis. The lack of a unified understanding of the structural properties that ensure the PŁK condition makes it hard to generalize the analysis to other problems. For this purpose, we introduce a framework with several easily verifiable assumptions to establish the PŁK condition of the policy gradient optimization problem in finite-horizon MDPs with general state and action spaces. Specifically, we demonstrate that the policy gradient optimization problem satisfies the PŁK condition when (i) the policy objective has bounded gradients, (ii) expected optimal Q-value functions satisfy the PŁK condition, and (iii) sequential decomposition inequalities hold. Roughly speaking, sequential decomposition inequalities state that the difference in partial gradients of expected optimal Q-value functions, when evaluated under an arbitrary policy versus the optimal policy, is upper-bounded by the corresponding difference in function values.

To illustrate the practical relevance of our framework, we validate that a variety of control and operations models, along with their corresponding optimal policy classes, satisfy the proposed assumptions (i)-(iii), thereby satisfying the PŁK condition. These models include (1) entropy-regularized tabular MDPs with stochastic policies, (2) Linear Quadratic Regulator (LQR) problems with affine policies, (3) multi-period inventory systems with Markov-modulated demands employing state-dependent base-stock policies, and (4) stochastic cash balance problems with two-sided base-stock policies. Notably, models (1)-(3) are commonly seen in dynamic programming textbooks (Bertsekas, 1995; Puterman, 2014). All of these models exhibit (hidden) convexity in the dynamic programming recursions, a structural property that is crucial for ensuring the PŁK condition holds.

Leveraging the PŁK condition, we establish a linear convergence rate for exact policy gradient methods to achieve an ϵ-optimal policy for entropy-regularized tabular MDPs and LQR problems, which aligns with existing results (Bhandari & 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. All these complexities exhibit a polynomial dependence on the time horizon, improving over the exponential dependence established by the framework of Huh & Rusmevichientong (2014). Such an improvement is mainly due to a newly established technical Lemma, which could be of independent interest.

Our contribution advances the state-of-the-art results of policy gradient methods for solving MDPs with general state and action spaces in several key aspects. First, our work focuses on finite-horizon MDPs with the discounted factor $\gamma = 1$. One cannot directly apply existing results of infinite-horizon discounted MDPs (Ju & Lan, 2022; Bhandari & Russo, 2024) by treating finite-horizon MDPs as a special case due to the explicit dependence on $1/(1-\gamma)$ in complexity bounds. While Bhandari & Russo (2024, Theorem 2) can extend to the setting of finite-horizon MDPs, its applicability relies on stronger structural conditions, e.g., gradient dominance across all Q-value functions and closure under policy improvement, which fail in inventory models. In contrast, our analysis applies to a more extensive class of control and operations problems.

Second, we provide a stronger characterization of the landscape for the policy gradient optimization by demonstrating the PŁK condition. Bhandari & Russo (2024, Theorem 3) establish 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. Yet, it remains unclear if asymptotic convergence translates to an exponential or a polynomial dependence on the planning horizon. In fact, they "leave the study of a gradient dominance condition for finite-horizon problems as future work". Our work bridges this gap by establishing the PŁK condition (which is equivalent to the gradient dominance condition under some mild conditions (Karimi et al., 2016, Appendix G)) of the policy gradient optimization. The established PŁK condition enables us to achieve a non-asymptotic convergence rate for policy gradient methods, yielding a sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-1})$ with polynomial dependence on the time horizon.

Finally, our work complements the understanding of the PŁK condition by identifying a class of optimization problems satisfying the PŁK condition. Many optimization papers impose the PŁK

condition as an important assumption (Attouch et al., 2013; Fatkhullin et al., 2022; Lewis & Tian, 2025), but very few examples satisfy the PŁK condition (Li & Pong, 2018). As we highlight in our analysis, the (hidden) convexity within the dynamic programming recursion is critical to establishing the PŁK condition. This observation suggests the possibility of extending our framework to a broader class of dynamic programming problems with convex recursive structures.

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 $\mathcal{S} = \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_T$, 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 $P : \mathcal{S} \times \mathcal{A} \times [T] \to \mathcal{S}$, 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 $C : \mathcal{S} \times \mathcal{A} \times [T] \to \mathbb{R}$, 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 $P_t(\cdot|s, a) \coloneqq P(\cdot|s, a, t)$, and $C_t(s, a) \coloneqq C(s, a, t)$ 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 $\pi : \mathcal{S} \times [T] \to \mathcal{A}$ 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 $\pi_t(s) \coloneqq \pi(s, t)$ 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 $(s_1, \ldots, s_T)$, where s_1 is the initial state and $s_{t+1} \sim P_t(\cdot|s_t, \pi_t(s_t))$ for all $t = 1, \ldots, T-1$. A policy π^* is optimal if it minimizes the total expected cost $J(\pi)$ with the initial distribution ρ:

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

Given a policy π, we define $\rho_t(\cdot|\pi)$ as the cumulative distribution function of s_t incurred by π starting with the initial distribution ρ. Furthermore, we define the value function

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

which represents the total expected cost at time t starting with the initial state s and policy π. In the same manner, we define the function

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

as the action-value (or Q-value) function. We use V_t^* and Q_t^* to denote the value function and the Q-value function corresponding to the optimal policy π^*, respectively.

Note that general policy optimization falls into functional optimization as we search over the function class Π, which is computationally intractable. To avoid the computational issue, it is common to parameterize the policy through finite-dimensional parameters $\theta = (\theta_1, \ldots, \theta_T)$ (Sutton et al., 1999). At time t, the policy $\pi_t(\cdot|\theta_t)$ is parameterized by θ_t, which belongs to a convex and compact set $\Theta_t \subseteq \mathbb{R}^d$. The parameter space Θ, defined as the Cartesian product $\Theta_1 \times \cdots \times \Theta_T$, forms a convex and compact feasible region of θ. In such a case, the parameterized policy class is $\Pi_\Theta = \{\pi(s, t|\theta) : \mathcal{S} \times [T] \times \Theta \to \mathcal{A}\} \subseteq \Pi$. 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 $l(\theta) \coloneqq J(\pi_\theta)$, called the policy gradient objective function. Let θ^* denote one of the minimizers of $\min_{\theta\in\Theta} l(\theta)$ and π_{θ^*} as

the corresponding policy. We define a policy π_θ to be ϵ-optimal if θ is an ϵ-optimal solution of $l(\theta)$. We denote the gradient $\nabla l(\theta) = (\nabla_{\theta_1} l(\theta), \ldots, \nabla_{\theta_T} l(\theta))$. Policy gradient methods apply first-order algorithms to minimize the total expected cost $l(\theta)$ over Θ by the iteration $\theta^{k+1} \leftarrow \text{Proj}_\Theta(\theta^k - \eta_k \nabla \hat{l}(\theta^k))$, where η_k represents the stepsize and $\nabla \hat{l}(\theta^k)$ represents the stochastic gradient estimator. Standard results for first-order methods in nonconvex settings typically guarantee convergence to first-order stationary points (Ghadimi & Lan, 2013). In the next section, we identify structural properties under which $l(\theta)$ satisfies the PŁK condition, thereby enabling policy gradient methods to converge to globally optimal policies despite nonconvexity.

3 Landscape Characterization

To characterize the nonconvex landscape of constrained smooth optimization problems, we utilize a specific form of the PŁK condition (Karimi et al., 2016, Appendix G).

Definition 1 (PŁK 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 $f^* \coloneqq \min_{x \in \mathcal{X}} f(x)$. The function f satisfies the PŁK condition on $\mathcal{X}$ if there exists $\mu > 0$ such that*

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

where $\delta_{\mathcal{X}}(x)$ represents the indicator function for $\mathcal{X}$, i.e., $\delta_{\mathcal{X}}(x) = 0$ when $x \in \mathcal{X}$ and $+\infty$ otherwise. Here, μ refers to the PŁK constant.

The subdifferential of $\delta_{\mathcal{X}}(x)$ is the normal cone of $\mathcal{X}$ at x. Notably, the PŁK condition reduces to the classical PŁ condition when $\mathcal{X} = \mathbb{R}^n$. Like the PŁ condition, the PŁK condition excludes suboptimal stationary points, offering a convergence guarantee for first-order methods (Karimi et al., 2016).

Leveraging the PŁK condition, one can establish the linear convergence rate (Attouch et al., 2013) and $\tilde{\mathcal{O}}(\epsilon^{-1})$ sample complexity of first-order methods. However, verifying the PŁK condition for the policy gradient optimization is challenging. To address this difficulty, we develop a general framework to validate the PŁK condition for a class of MDPs in the following theorem.

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 is differentiable on Θ_t with the 2-norm of its gradient upper bounded by $G > 0$.*
2. ***(PŁK Condition of Expected Optimal Q-value Functions)** For any $t \in [T]$, the expected optimal Q-value function satisfies the PŁK condition with a PŁK constant $\mu_Q > 0$.*
3. ***(Sequential Decomposition Inequality)** For any $\theta \in \Theta$ and $1 \leq t < k \leq T$, there exists $M_g > 0$:*

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

Then $l(\theta)$ satisfies the PŁK condition on Θ with the PŁK constant $\mu_l = \mu_Q^3 / e M_g^2 G^2 T^2$.

4 Conclusion

This work provides a framework with easily verifiable conditions to establish the PŁK condition for policy gradient optimization of finite-horizon MDPs with general state and action spaces. Despite nonconvexity, the PŁK condition guarantees a linear convergence rate for exact policy gradient methods and an $\tilde{\mathcal{O}}(\epsilon^{-1})$ sample complexity for stochastic policy gradient methods. Our model is applicable to various control and operations models, including entropy-regularized tabular MDPs, LQR problems, multi-period inventory systems with Markov-modulated demands, and stochastic cash balance problems. Notably, we establish the first sample complexity results for solving the stochastic cash balance problem and the multi-period inventory system with Markov-modulated demand, allowing backorders, and an extension to the lost sales model. The complexity admits a polynomial dependence on the planning horizon. Our understanding of the structures of operations models contributes to both *optimization* and *reinforcement learning*.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *The Journal of Machine Learning Research*, 22 (1):4431–4506, 2021.

Attouch, H., Bolte, J., and Svaiter, B. F. 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.

Bertsekas, D. *Dynamic Programming and Optimal Control*. Number v. 1 in Athena scientific optimization and computation series. Athena Scientific, 1995. ISBN 9781886529120.

Bhandari, J. and Russo, D. Global optimality guarantees for policy gradient methods. *Operations Research*, 72(5):1906–1927, 2024.

Fatkhullin, I., Etesami, J., He, N., and Kiyavash, N. Sharp analysis of stochastic optimization under global kurdyka-lojasiewicz inequality. *Advances in Neural Information Processing Systems*, 35: 15836–15848, 2022.

Ghadimi, S. and Lan, G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Hambly, B., Xu, R., and Yang, H. Policy gradient methods for the noisy linear quadratic regulator over a finite horizon. *SIAM Journal on Control and Optimization*, 59(5):3359–3391, 2021.

Huh, W. T. and Rusmevichientong, P. Online sequential optimization with biased gradients: theory and applications to censored demand. *INFORMS Journal on Computing*, 26(1):150–159, 2014.

Hwangbo, J., Lee, J., Dosovitskiy, A., Bellicoso, D., Tsounis, V., Koltun, V., and Hutter, M. Learning agile and dynamic motor skills for legged robots. *Science Robotics*, 4(26):eaau5872, 2019.

Ju, C. and Lan, G. Policy optimization over general state and action spaces. *arXiv preprint arXiv:2211.16715*, 2022.

Karimi, H., Nutini, J., and Schmidt, M. 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*, pp. 795–811. Springer, 2016.

Kurdyka, K. On gradients of functions definable in o-minimal structures. In *Annales de l'institut Fourier*, volume 48, pp. 769–783, 1998.

Lewis, A. and Tian, T. The complexity of first-order optimization methods from a metric perspective. *Mathematical Programming*, 212(1):49–78, 2025.

Li, G. and Pong, T. K. Calculus of the exponent of kurdyka–łojasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of computational mathematics*, 18(5): 1199–1232, 2018.

Lojasiewicz, S. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 117(87-89):2, 1963.

Nocedal, J. and Wright, S. J. *Numerical optimization*. Springer, 1999.

Polyak, B. T. et al. Gradient methods for minimizing functionals. *Zhurnal vychislitel'noi matematiki i matematicheskoi fiziki*, 3(4):643–653, 1963.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.