Title
Safe Online Decision-Making for Non-Stationary Systems

Author
Ding, Yuhao

Publication Date
2023

Peer reviewed|Thesis/dissertation

Safe Online Decision-Making for Non-Stationary Systems

by

Yuhao Ding

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering - Industrial Engineering and Operations Research

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Associate Professor Javad Lavaei, Chair
Assistant Professor Paul Grigas
Professor Murat Arcak

Spring 2023

Safe Online Decision-Making for Non-Stationary Systems

Abstract

Safe Online Decision-Making for Non-Stationary Systems

by

Yuhao Ding

Doctor of Philosophy in Engineering - Industrial Engineering and Operations Research

University of California, Berkeley

Associate Professor Javad Lavaei, Chair

Despite several progresses of control-theoretic techniques in the past decade, these methods still struggle to bridge the widening gap between theory and reality, which is exacerbated by the increasing complexity, uncertainty, and safety requirements. Consequently, the creation of online control algorithms for safety-critical applications in non-stationary environments could pave the way for a new chapter in modern control theory, substantially enhancing the reliability of intelligent systems as they function in dynamic, uncertain, and potentially hostile conditions subject to physical and computational limitations. Safe non-stationary decision-making not only encompasses the core challenges of traditional decision-making but also presents new hurdles, such as *(i)* fast adaptation under the non-stationary environments, *(ii)* global optimality convergence of the non-convex optimization, *(iii)* continual balancing of objective and constraints. The above challenges go beyond current capabilities in computation and theory and manifest in various aspects of practical and theoretical interests, from sample complexity and non-convergence issues to computational tractability and enforcement of safety constraints for real-time control. This thesis aims to pioneer system operation at the nexus of reinforcement learning, online learning, statistical learning, and nonlinear optimization. The design of provably efficient and safe online decision-making algorithms that exploit prediction and prior knowledge while grappling with the effects of dynamic feedback and non-stationary environment will push the frontiers of computational verification and synthesis of control policies for safety-critical systems.

To overcome these challenges and realize the full potential of online decision-making approaches for adaptability and performance gains, this thesis aims to extend the foundational knowledge in systems and control and broaden our understanding of performance limits and engineering trade-offs when the system must operate outside of the assumptions of known models and needs to adapt to its environment in real-time. In particular, we develop a new mathematical foundation and a set of computational tools for the design of safe online decision-making algorithms that can be deployed in environments that undergo changes. Along this line, we

will address the following objectives: *(i)* escaping spurious local minimum trajectories in online time-varying non-convex optimization, *(ii)* provably efficient primal-dual reinforcement learning for CMDPs with non-stationary objectives and constraints, *(iii)* non-stationary risk-sensitive reinforcement learning with near-optimal dynamic regret, adaptive detection, and separation design.

To my parents and wife, for their unconditional love and support

Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my deepest gratitude and appreciation to all those who have supported and contributed to the successful completion of this thesis. This thesis would not have been possible without their guidance, encouragement, and unwavering belief in my potential.

First and foremost, I would like to extend my sincere thanks to my advisor, Prof. Javad Lavaei, for his invaluable mentorship, patience, and expertise throughout my research. His dedication to my growth as a researcher and his constant support have been instrumental in shaping my academic and professional path. I am greatly indebted to Javad for his endless help and support throughout my PhD, for helping me become an independent researcher, and for giving me the courage to tackle hard problems. Beyond his support in my academic growth, Javad has also provided me with substantial assistance in my personal life, such as increasing my salary during the COVID-19 pandemic so that I could live and work more securely. I will always be grateful to him for having faith in me, providing me with many research opportunities, and making my PhD journey enriching and meaningful.

I would also like to acknowledge the contributions of my thesis committee members, Prof. Murat Arack and Prof. Paul Grigas. Their insightful feedback and constructive criticism have been essential in refining my research and ensuring its quality. I feel fortunate to have had the opportunity to collaborate with Murat on the problem of time-varying optimization, which leads to a series of publications. I am also truly grateful to Paul for his support and guidance during my PhD. I gained extensive knowledge on optimization and experienced immense benefits from his class "Modern Optimization for Statistical Learning", which has become my favorite.

I am happy to have worked in a brilliant research group and a department with students and post-docs of outstanding proficiency and personality. Throughout my PhD, I was fortunate to receive significant advice and assistance from Prof. Ming Jin, who is now a faculty member at Virginia Tech. My gratitude extends to my other co-authors: Somayeh Sojoudi, Salar Fattahi, Cedric Josz, Junzi Zhang, Reza Mohammadi-Ghazi, Han Feng, Donghao Ying, Yunkai Zhang and Alec Koppel.

I consider myself incredibly lucky to have been in such a friendly environment at UC Berkeley. A special shoutout goes to my IEOR cohort Yoon Lee and Mahan Tajrobehkar for co-founding the "Stronger Group" where we worked out together and supported each other. My sincere gratitude also goes to the peers from my home IEOR department: Richard Zhang, SangWoo Park, Igor Molybog, Yingjie Bi, Heyuan Liu, Hansheng Jiang, Tianyi Lin, Renyuan Xu, Junyu Cao, Haoyang Cao, Anran Hu, Mahbod Olfat, Georgios Patsakis, Pedro Hespanhol, Marie (Pelagie) Elimbi, Julie Mulvaney-Kemp, Yoon Lee, Mahan Tajrobehkar, Haixiang Zhang, Baturalp Yalcin, Jihun Kim, Ying Chen, Eli Brock, Hyunin Lee, Haoting Zhang, Jingxu Xu, Jiaming Wang, and Mo Liu.

I would like to extend my heartfelt gratitude to my friends Yunbo Liu and Yiling You who are the first two friends I made at Berkeley. Our friendship has thrived for five years, and it will continue to flourish. My personal well-being during my PhD was hugely indebted

to many other amazing friends, including Fan Zhang, Can Huang, Haotian Gu, Sizhu Lu and He Yin.

I must also express my profound appreciation for my family, whose unwavering love, support, and encouragement have been a constant source of strength throughout my academic journey. To my parents, Yixin Ding and Zhiqun Huang, thank you for your endless patience and faith in my abilities. To my wife, Danlin Xiao, your understanding and support have been indispensable in helping me persevere through the challenges of this journey.

Chapter 1

Introduction

This dissertation focuses on developing computational tools and analytical assurance for modern *safety-critical* systems in *non-stationary* environments. The employment of intelligent autonomous systems has seen a marked increase across various domains, including robotics, communication, transportation, and power systems. With the opportunities also come challenges to the classical control paradigm, increasingly confronted by the widening gap between theory and reality due to the unprecedented uncertainty and complexity of real-world systems. On the other hand, the potential of online decision-making techniques such as online optimization and reinforcement learning (RL) to enhance system performance and adaptivity has been observed in many applications over the past few years. However, there is an equally-vast array of real-world applications for which the existing online decision-making techniques are not yet applicable or are too risky to employ. Those applications often require the *nonconvex optimization*, *safety assurance*, and the underlying environment may undergo changes and be *nonstationary*. While these aspects have been tackled separately in the literature to some limited extent, there remains a substantial gap when these issues arise simultaneously, imposing challenges for the deployment of concurrent methods in real-world systems. For example, in autonomous driving [103], it is essential to guarantee the safety, such as collision avoidance and traffic rules, while handling time-varying conditions related to weather and traffic; similarly, in most safety-critical human-computer interaction applications, e.g., automated medical care, human behavior changes over time.

In the following sections of this chapter, we first provide a general introductory overview of the problems that are considered in this dissertation, as well as discuss the potential challenges we may encounter in addressing them. Following that, we provide a brief summary of our contributions. We then identify the pertinent publications referenced throughout the dissertation. We conclude this chapter by presenting the basic notations that are used throughout the dissertation.

1.1 Safe Decision-Making under Non-Stationarity

This dissertation is devoted to solving safe decision-making problems under non-stationarity in the form of

$$\min_{x_t \in \mathbb{R}^n} \quad f(x_t; \theta_t) \tag{1.1a}$$

$$\text{subject to} \quad x_t \in \mathcal{X}(\theta_t) \tag{1.1b}$$

for $t = 1, 2, 3, \ldots$, where

- $\theta_t \in \mathbb{R}^m$ is the time-varying exogenous vector that (directly or indirectly) captures the non-stationary parameters of the problem. For instance, it may capture the electricity demands in a power system which may vary over time during a day and exhibit strong seasonality. It can also encapsulate specific non-stationary parameters of a dynamical system, such as the moving targets or obstacles in the motion-planning problem of a robotic system.

- $x_t \in \mathbb{R}^n$ is the targeted multivariate decision variable at the time t. For instance, it may capture the amount of generations for different generators in a power system, or it may indicate an optimal control policy for a dynamical system. Due to the non-stationarity of the parameters θ_t, the decision variable should be optimized at each time t to adapt to the varying systems.

- $f(x_t; \theta_t)$ is the objective function with respect to x_t and parameterized by θ_t. For example, it may correspond to the operational cost of a power system, or it may be the accumulated costs that a reinforcement learning agent aims to minimize.

- $\mathcal{X}(\theta_t)$ is the feasible set of the decision-making problems parameterized by θ_t, i.e., the set of all feasible values that can be attributed to the decision variable x_t. The feasible set $\mathcal{X}(\theta_t)$ is usually explicitly characterized by a set of inequality or equality constraints that are parameterized by θ_t. For instance, it may correspond to the constraints to match the power generation and electricity demand in a power system, or it may capture certain safety requirements such as the collision-avoidance and risk control for a reinforcement learning agent.

As will be shown later in the dissertation, many real-world problems can be cast as instances of the problem (1.1). Our goal is to provide computational tools and analytical assurance for safe decision-making problems in non-stationary environments under safety constraints. While it is mentioned in the seminal book by Sutton and Barto [116] that "nonstationarity is the case most commonly encountered in reinforcement learning," the field is still in its infancy. As will be delineated later, there are some unique technical challenges in this setting:

- **Non-stationarity**: Non-stationarity presents a substantial challenge for many real-world safety-critical applications [97] since we may not have a full knowledge of how the parameter vector θ_t may vary over the time. For instance, the exact real-time electricity demand in the future is not available at the time of scheduling the power generation. Similarly, the true model of a dynamical system is rarely known in practice and may vary over time due to the varying operating conditions. Indeed, the inference of the non-stationary parameter θ_t based on a limited number of noisy observations/samples of $\{\theta_i\}_{i=1}^{t-1}$ is the key challenge for solving the problem (1.1). Since the data associated with $\{\theta_i\}_i$ is not independent and identically distributed (i.i.d.), a traditional estimator based on i.i.d. assumption is no longer effective. Furthermore, without a precise estimation of θ_t, solving the problem (1.1) accurately for each time t is impossible. Thus, the performance guarantee on solving the problem (1.1) will also depend on the underlying non-stationarity structure of the parameter $\{\theta_i\}_i$.

- **Non-convexity**: Nonconvexity is inherent in many real-world problems such as training of deep neural networks [82], the optimal power flow problem [79], and reinforcement learning [4]. If the objective function $f(\cdot;\theta_t)$ in the problem (1.1) is non-convex, it may possess multiple local/global solutions, any of which may be recovered and returned as a candidate solution using the local search algorithms. From the classical complexity theory, this nonconvexity is perceived to be the main contributor to the intractability of these problems. Although there has been recently shown that simple local search methods, such as gradient-based algorithms, have a superb performance in solving nonconvex optimization problems. However, these results are all for time-invariant optimization problems for which the landscape is time-invariant. In contrast, many real-world problems should be solved sequentially over time with time-varying data. Therefore, it is essential to study the effect of the temporal variation on the landscape of time-varying nonconvex optimization problems.

- **Continual balancing among constraints and objective**: As mentioned before, the problems that are considered in this dissertation are motivated by safety-critical applications. Most works in safe decision-making have resorted to a primal-dual formulation/algorithm. The dual variables play the role of balancing the constraints and objective (similar to penalty coefficients); while the optimal dual solution is fixed in a stationary environment, the optimal dual solution is time-varying in a nonstationary environment, introducing a technical challenge in algorithm design and analysis.

1.2 Summary of Contributions

In view of the above fundamental challenges, nonstationary safe decision-making encapsulates various *open problems* that, if addressed, will substantially extend the foundational knowledge in systems & control and broaden our understanding of performance limits and engineering trade-offs when the system must operate outside of the assumptions of known models and

needs to adapt to its environment in real-time. In this section, we will briefly summarize the contributions of the dissertation.

- **Chapter 2**
  A major limitation of online algorithms that track the optimizers of time-varying nonconvex optimization problems is that they focus on a specific local minimum trajectory, which may lead to poor spurious local solutions. In Chapter 2, we show that the natural temporal variation may help simple online tracking methods find and track time-varying global minima. To this end, we investigate the properties of a time-varying projected gradient flow system with inertia, which can be regarded as the continuous-time limit of (1) the optimality conditions for a discretized sequential optimization problem with a proximal regularization and (2) the online tracking scheme. We introduce the notion of the dominant trajectory and show that the inherent temporal variation could reshape the landscape of the Lagrange functional and help a proximal algorithm escape the spurious local minimum trajectories if the global minimum trajectory is dominant. For a problem with twice continuously differentiable objective function and constraints, sufficient conditions are derived to guarantee that no matter how a local search method is initialized, it will track a time-varying global solution after some time. The results are illustrated on a benchmark example with many local minima.

- **Chapter 3**
  We consider primal-dual-based reinforcement learning (RL) in episodic constrained Markov decision processes (CMDPs) with non-stationary objectives and constraints, which plays a central role in ensuring the safety of RL in time-varying environments. In this problem, the reward/utility functions and the state transition functions are both allowed to vary arbitrarily over time as long as their cumulative variations do not exceed certain known variation budgets. Designing safe RL algorithms in time-varying environments is particularly challenging because of the need to integrate the constraint violation reduction, safe exploration, and adaptation to the non-stationarity. To this end, we identify two alternative conditions on the time-varying constraints under which we can guarantee the safety in the long run. We also propose the Periodically Restarted Optimistic Primal-Dual Proximal Policy Optimization (PROPD-PPO) algorithm that can coordinate with both two conditions. Furthermore, a dynamic regret bound and a constraint violation bound are established for the proposed algorithm in both the linear kernel CMDP function approximation setting and the tabular CMDP setting under two alternative conditions. This chapter provides the first provably efficient algorithm for non-stationary CMDPs with safe exploration.

- **Chapter 4**
  We study risk-sensitive reinforcement learning (RL) based on an entropic risk measure in episodic non-stationary Markov decision processes (MDPs). Both the reward functions and the state transition kernels are unknown and allowed to vary arbitrarily over time with a budget on their cumulative variations. When this variation budget is known a

prior, we propose two restart-based algorithms, namely Restart-RSMB and Restart-RSQ, and establish their dynamic regrets. Based on these results, we further present a meta-algorithm that does not require any prior knowledge of the variation budget and can adaptively detect the non-stationarity on the exponential value functions. A dynamic regret lower bound is then established for non-stationary risk-sensitive RL to certify the near-optimality of the proposed algorithms. Our results also show that the risk control and the handling of the non-stationarity can be separately designed in the algorithm if the variation budget is known a prior, while the non-stationary detection mechanism in the adaptive algorithm depends on the risk parameter. This work offers the first non-asymptotic theoretical analyses for the non-stationary risk-sensitive RL in the literature.

1.3 Related Publications

- **Chapter 2**
  Main paper:
  - Yuhao Ding, Javad Lavaei and Murat Arcak. “Time-variation in online nonconvex optimization enables escaping from spurious local minima”, IEEE Transactions on Automatic Control, vol. 68, no. 1, pp. 156-171, Jan. 2023.

  Related paper:
  - Yuhao Ding, Javad Lavaei and Murat Arcak. “Escaping Spurious Local Minimum Trajectories in Online Time-varying Nonconvex Optimization", 2021 American Control Conference (ACC), New Orleans, LA, USA, 2021, pp. 454-461.
  - Salar Fattahi; Cedric Josz; Yuhao Ding; Reza Mohammadi; Javad Lavaei; Somayeh Sojoudi. "On the Absence of Spurious Local Trajectories in Time-Varying Nonconvex Optimization", in IEEE Transactions on Automatic Control, vol. 68, no. 1, pp. 80-95, Jan. 2023.
- **Chapter 3**
  Main paper:
  - Yuhao Ding and Javad Lavaei. “Provably Efficient Primal-Dual Method for CMDPs with Non-stationary Objectives and Constraints‘”, Proceedings of the AAAI Conference on Artificial Intelligence. 2023

  Related paper:
  - Donghao Ying, Yuhao Ding and Javad Lavaei. “A dual approach to constrained markov decision processes with entropy regularization", Proceedings of The 25th International Conference on Artificial Intelligence and Statistics, PMLR 151:1887-1909, 2022.

  - Vanshaj Khattar, Yuhao Ding, Bilgehan Sel, Javad Lavaei and Ming Jin. "A CMDP-within-online framework for Meta-Safe Reinforcement Learning", International Conference on Learning Representations, 2023.
  - Donghao Ying, Mengzi Guo, Yuhao Ding, Javad Lavaei and Zunjun Shen. "Policy-based Primal-Dual Methods for Convex Constrained Markov Decision Processes", Proceedings of the AAAI Conference on Artificial Intelligence. 2023

- **Chapter 4**
  Main paper:

  - Yuhao Ding, Ming Jin and Javad Lavaei. "Non-stationary Risk-sensitive Reinforcement Learning: Near-optimal Dynamic Regret, Adaptive Detection, and Separation Design", Proceedings of the AAAI Conference on Artificial Intelligence. 2023

1.4 Notations

Scalars, vectors, matrices, and sets: For a vector x, we use x^T to denote the transpose of x, and use x_i or $(x)_i$ to denote the i-th entry of x. For vectors x and y, we use $x \geq y$ to denote an entry-wise inequality. We use the standard notations that $\|x\|_1 = \sum_i |x_i|$, $\|x\|_2 = \sqrt{\sum_i x_i^2}$, and $\|x\|_\infty = \max_i |x_i|$. For simplicity, we will also use the notation $\|\cdot\|$ to represent the Euclidean norm $\|\cdot\|_2$. We denote $\text{Proj}_{[a,b]}(x)$ as the projection of x onto the interval $[a,b]$ and $(x)_+$ as the maximum between x and 0. Let I_n denote the $n \times n$ identity matrix. For a matrix A, we use A_{ij} to denote its (i,j)-th entry and use $\lambda_{\min}(A)$ to denote its minimum eigenvalue. We use $\|v\|_A$ to denote the norm induced by a positive definite matrix A for vector v, i.e., $\|v\|_A = \sqrt{v^\top A v}$. Let $|\mathcal{S}|$ denote the cardinality of set $\mathcal{S}$. Let $\mathbb{R}$ represent the set of real numbers. The interior of the interval $\bar{I}_{t,2}$ is denoted by $\text{int}(\bar{I}_{t,2})$. The symbol $\mathcal{B}_r(h(t)) = \{x \in \mathbb{R}^n : \|x - h(t)\| \leq r\}$ denotes the region centered around a trajectory $h(t)$ with radius r at time t. We use the shorthand notation $[n]$ for the set $\{1, 2, \ldots, n\}$.

Functions: We denote the solution of $\dot{x} = f(x,t)$ starting from x_0 at the initial time t_0 with $x(t, t_0, x_0)$ or the short-hand notation $x(t)$ if the initial condition (t_0, x_0) is clear from the context. When applying a scalar function to a vector x, e.g. $\log x$, the operation is understood as entry-wise. For a function $f(x)$, let $\nabla_x f(x)$ denote its gradient with respect to x, and we may omit x in the subscript when it is clear from the context. Let $\arg\min f(x)$ (resp. $\arg\max f(x)$) denote any arbitrary global minimum (resp. global maximum) of $f(x)$. Given a variable x, the notation $a = \mathcal{O}(b(x))$ means that $a \leq C \cdot b(x)$ for some constant $C > 0$ that is independent of x. Similarly, $a = \widetilde{\mathcal{O}}(b(x))$ indicates that the previous inequality may also depend on the function $\log(x)$, where $C > 0$ is again independent of x. In addition, the notation $a = \Omega(b(x))$ means that $a \geq C \cdot b(x)$ for some constant $C > 0$ that is independent of x.

Probability: When the variable s follows the distribution ρ, we write it as $s \sim \rho$. Let $\mathbb{E}[\cdot]$ and $\mathbb{E}[\cdot \mid \cdot]$ denote the expectation and conditional expectation of a random variable, respectively.

Chapter 2

Time-varying non-convex optimization

Consider the following equality-constrained time-varying optimization problem:

$$\begin{aligned} \min_{x(t)\in\mathbb{R}^n} \quad & f(x(t),t) \\ \text{s.t.} \quad & g(x(t),t) = 0 \end{aligned} \tag{2.1}$$

where $t \geq 0$ denotes the time and $x(t)$ is the optimization variable that depends on t. Moreover, the objective function $f : \mathbb{R}^n \times [0,\infty) \to \mathbb{R}$ and the constraint function $g(x,t) = (g_1(x,t),\ldots,g_m(x,t))$ with $g_k : \mathbb{R}^n \times [0,\infty) \to \mathbb{R}$ for $k = 1,...,m$ are assumed to be twice continuously differentiable in state x and continuously differentiable in time t. For each time t, the function $f(x,t)$ could potentially be nonconvex in x with many local minima and the function $g(x,t)$ could also potentially be nonlinear in x, leading to a nonconvex feasible set. The objective is to solve the above problem online under the assumption that at any given time t the function $f(x,t')$ and $g(x,t')$ are known for all $t' \leq t$ while no knowledge about $f(x,t')$ or $g(x,t')$ may be available for any $t' > t$. Therefore, the problem (2.1) cannot be minimized off-line and should be solved sequentially. Another issue is that the optimization problem at each time instance could be highly complex due to NP-hardness, which is an impediment to finding its global minima. This chapter aims to investigate under what conditions simple local search algorithms can solve the above online optimization problem to almost global optimality after some finite time. More precisely, the goal is to devise an algorithm that can track a global solution of (2.1) as a function of time t with some error at the initial time and a diminishing error after some time.

If $f(x,t)$ and $g(x,t)$ do not change over time, the problem reduces to a classic (time-invariant) optimization problem. It is known that simple local search methods, such as stochastic gradient descent (SGD) [63], may be able to find a global minimum of such time-invariant problems (under certain conditions) for almost all initializations due to the randomness embedded in SGD [68, 54, 75]. The objective of this chapter is to significantly extend the above result from a single optimization problem to infinitely-many problems parametrized by time t. In other words, it is desirable to investigate the following question: **Can the temporal variation in the landscape of time-varying nonconvex**

optimization problems enable online local search methods to find and track global trajectories? To answer this question, we study a first-order time-varying ordinary differential equation (ODE), which is the counterpart of the classic projected gradient flow system for time-invariant optimization problems [119] and serves as a continuous-time limit of the discrete online tracking method for (2.1) with the proximal regularization. This ODE is given as

$$\dot{x}(t) = -\frac{1}{\alpha}\mathcal{P}(x(t),t)\nabla_x f(x(t),t) - \mathcal{Q}(x(t),t)g'(x(t),t) \tag{P-ODE}$$

where $\alpha > 0$ is a constant parameter named **inertia** due to a **proximal regularization**, $g'(z,t) = \frac{\partial g(z,t)}{\partial t}$, $\mathcal{P}(x(t),t)$ and $\mathcal{Q}(x(t),t)$ are matrices related to the Jacobian of $g(x,t)$ that will be derived in detail later. A system of the form (P-ODE) is called a **time-varying projected gradient system with inertia** α. The behavior of the solutions of this system initialized at different points depends on the value of α. In the unconstrained case, this ODE reduces to the **time-varying gradient system with inertia** α given as

$$\dot{x}(t) = -\frac{1}{\alpha}\nabla_x f(x,t) \tag{ODE}$$

In what follows, we offer a motivating example without constraints (to simplify the visualization) before stating the goals of this chapter.

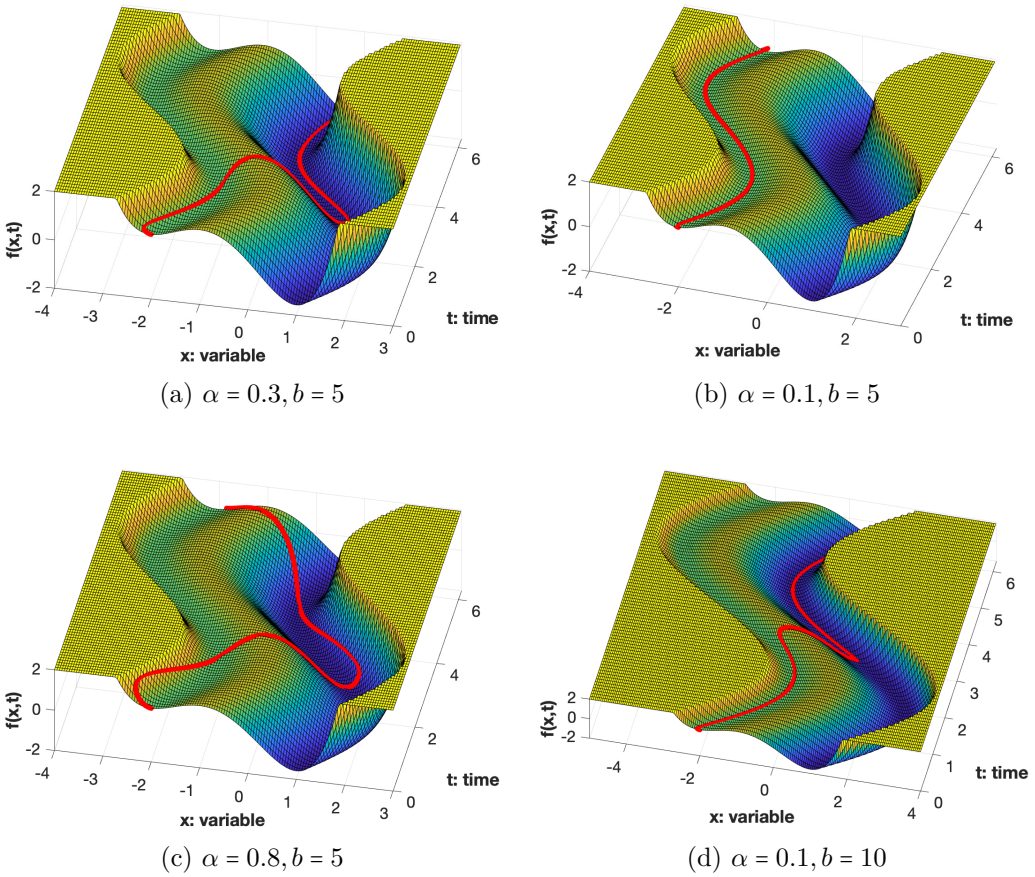
Motivating example

Example 1. Consider $f(x,t) := \bar{f}(x - b\sin(t))$, where

$$\bar{f}(y) := \frac{1}{4}y^4 + \frac{2}{3}y^3 - \frac{1}{2}y^2 - 2y$$

This time-varying objective has a spurious (non-global) local minimum trajectory at $-2 + b\sin(t)$, a local maximum trajectory at $-1 + b\sin(t)$, and a global minimum trajectory at $1 + b\sin(t)$. In Figure 2.1, we show a bifurcation phenomenon numerically. The red lines are the solutions of (P-ODE) with the initial point -2. In the case with $\alpha = 0.3$ and $b = 5$, the solution of (P-ODE) winds up in the region of attraction of the global minimum trajectory. However, for the case with $\alpha = 0.1$ and $b = 5$, the solution of (P-ODE) remains in the region of attraction of the spurious local minimum trajectory. In the case with $\alpha = 0.8$ and $b = 5$, the solution of (P-ODE) fails to track any local minimum trajectory. In the case with $\alpha = 0.1$ and $b = 10$, the solution of (P-ODE) winds up in the region of attraction of the global minimum trajectory.

Two observations can be made here. First, jumping from a local minimum trajectory to a better trajectory tends to occur with the help of a relatively large inertia when the local minimum trajectory changes the direction abruptly and there happens to exist a better local minimum trajectory in the direction of the inertia. Second, when the inertia α is relatively small, the solution of (P-ODE) tends to track a local (or global) minimum trajectory closely and converges to that trajectory quickly.



(a) $\alpha = 0.3, b = 5$ (b) $\alpha = 0.1, b = 5$

(c) $\alpha = 0.8, b = 5$ (d) $\alpha = 0.1, b = 10$

Figure 2.1: Illustration of Example 1 (in order to increase visibility, the objective function values are rescaled). Jumping from a spurious local minimum trajectory to a global minimum trajectory occurs in Figure 2.1a and 2.1d when the inertia α and the change (controlled by the parameter b) of local minimum trajectory are appropriate.

Example 2. Consider the time-varying optimal power flow (OPF) problem, as the most fundamental problem for the operation of electric power grids that aims to match supply with demand while satisfying network and physical constraints. Let $f(x,t)$ be the function to be minimized at time t, which is the sum of the total energy cost and a penalty term taking care of all the inequality constraints of the problem. Let $g(x,t) = 0$ describe the time-varying demand constraint. Assume that the load data corresponds to the California data for August 2019. As discussed in [93], this time-varying OPF has 16 local minima at t=0 and many more for some values of $t > 0$. However, if (ODE) is run from any of these local minima, the 16 trajectories will all converge to the globally optimal trajectory, as shown in Figure 2.2. This observation has been made in [93] for a discrete-time version of the problem, but it also holds true for the continuous-time (ODE) model.

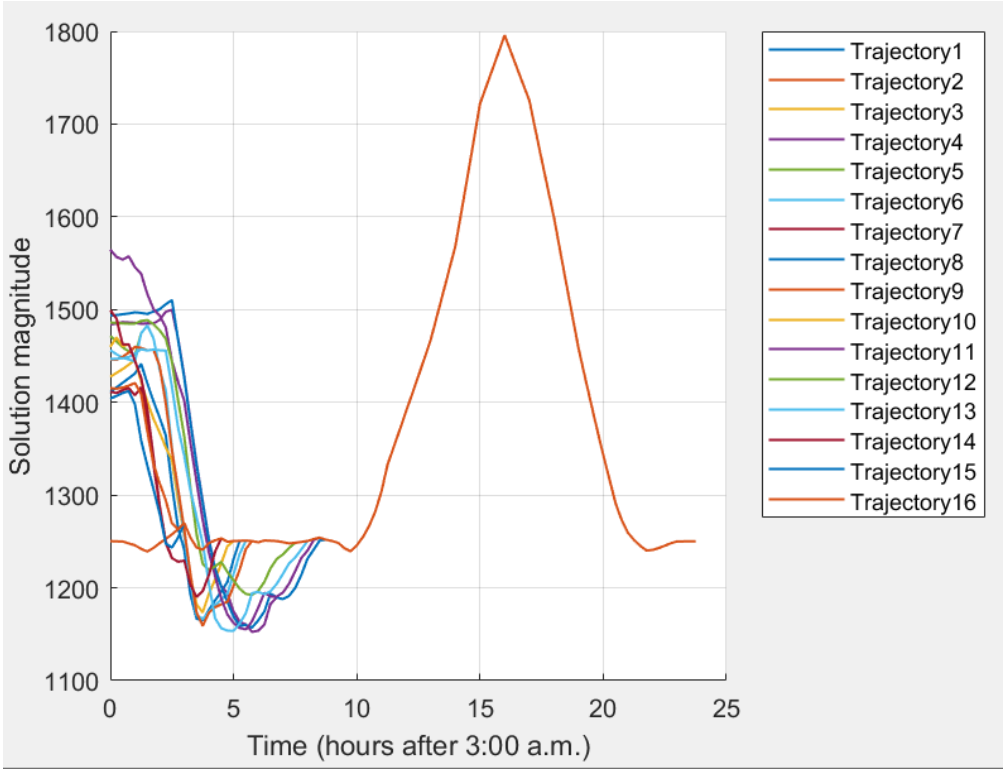


Figure 2.2: $|x(t)|$ (magnitude of the solution of (ODE)).

Our contributions

To mathematically study the observations made in Example 1 and Example 2 for a general time-varying nonconvex optimization problem with equality constraints, we focus on the aforementioned time-varying projected gradient flow system with inertia α as a continuous-time limit of an online updating scheme for (2.1). We first introduce a time-varying Lagrange functional to unify the analysis of unconstrained problems and equality-constrained problems, and make the key assumption that the time-varying Lagrange functional is locally one-point strongly convex around each local minimum trajectory. This assumption is justified by the second-order sufficient optimality conditions. A key property of (P-ODE) is that its solution will remain in the time-varying feasible region if the initial point is feasible for (2.1), which allows us to use the Lyapunov technique without worrying about the feasibility of the solution. Then, we show that the time-varying projected gradient flow system with inertia α is a continuous-time limit of the Karush–Kuhn–Tucker (KKT) optimality conditions for a discretized sequential optimization problem with a proximal regularization. The existence and uniqueness of the solution for such ODE is proven.

As a main result of this work, it is proven that the natural temporal variation of the time-varying optimization problem encourages the exploration of the state space and re-shaping the

landscape of the objective function (in the unconstrained case) or the Langrange functional (in the constrained case) by making it one-point strongly convex over a large region during some time interval. We introduce the notion of the dominant trajectory and show that if a given spurious local minimum trajectory is dominated by the global minimum trajectory, then the temporal variation of the time-varying optimization would trigger escaping the spurious local minimum trajectory for free. We develop two sufficient conditions under which the ODE solution will jump from a certain local minimum trajectory to a more desirable local minimum trajectory. We then derive sufficient conditions on the inertia α to guarantee that the solution of (P-ODE) can track a global minimum trajectory. To illustrate how the time variation nature of an online optimization problem promotes escaping a spurious minimum trajectory, we offer a case study with many shallow minimum trajectories.

2.1 Related Work

Online time-varying optimization problems: Time-varying optimization problems of the form (2.1) arise in the real-time optimal power flow problem [120, 62] for which the power loads and renewable generations are time-varying and operational decisions should be made every 5 minutes, as well as in the real-time estimation of the state of a nonlinear dynamic system [101]. Other examples include model predictive control [16], time-varying compressive sensing [104, 9] and online economic optimization [72, 130]. There are many researches on the design of efficient online algorithms for tracking the optimizers of time-varying convex optimization problems [111, 44, 12, 110]. With respect to time-varying nonconvex optimization problems, the work [57] presents a comprehensive theory on the structure and singularity of the KKT trajectories for time-varying optimization problems. On the algorithm side, [120] provides regret-type results in the case where the constraints are lifted to the objective function via penalty functions. [121] develops a running regularized primal-dual gradient algorithm to track a KKT trajectory, and offers asymptotic bounds on the tracking error. [89] obtains an ODE to approximate the KKT trajectory and derives an algorithm based on a predictor-corrector method to track the ODE solution. Recently, [42] proposed the question of whether the natural temporal variation in a time-varying nonconvex optimization problem could help a local tracking method escape spurious local minimum trajectories. It developed a differential equation to characterize this phenomenon (which is the basis of the current work), but it lacked mathematical conditions to guarantee this desirable behavior. The paper [93] also studies this phenomenon in the context of power systems and verifies on real data for California that the natural load variation enables escaping local minima of the optimal power flow problem. The current work significantly generalizes the results of [42] and [93] by mathematically studying when such an escaping is possible.

Local search methods for global optimization: Nonconvexity is inherent in many real-world problems: the classical compressive sensing and matrix completion/sensing [39, 22, 23], training of deep neural networks [82], the optimal power flow problem [79], and others. From the classical complexity theory, this nonconvexity is perceived to be the main

contributor to the intractability of these problems. However, it has been recently shown that simple local search methods, such as gradient-based algorithms, have a superb performance in solving nonconvex optimization problems. For example, [81] shows that the gradient descent with a random initialization could avoid the saddle points almost surely, and [68] and [54] prove that a perturbed gradient descent and SGD could escape the saddle points efficiently. Furthermore, it has been shown that nearly-isotropic classes of problems in matrix completion/sensing [15, 53, 131], robust principle component analysis [43, 71], and dictionary recovery [115] have benign landscape, implying that they are free of spurious local minima. The work [75] proves that SGD could help escape sharp local minima of a loss function by taking the alternative view that SGD works on a convolved (thus smoothed) version of the loss function. However, these results are all for time-invariant optimization problems for which the landscape is time-invariant. In contrast, many real-world problems should be solved sequentially over time with time-varying data. Therefore, it is essential to study the effect of the temporal variation on the landscape of time-varying nonconvex optimization problems.

Continuous-time interpretation of discrete numerical algorithms: Many iterative numerical optimization algorithms for time-invariant optimization problems can be interpreted as a discretization of a continuous-time process. Then, several new insights have been obtained due to the known results for continuous-time dynamical systems [74, 59]. Perhaps, the simplest and oldest example is the gradient flow system for the gradient descent algorithm with an infinitesimally small step size. The recent papers [114, 76, 125] study accelerated gradient methods for convex optimization problems from a continuous-time perspective. In addition, the continuous-time limit of the gradient descent is also employed to analyze various non-convex optimization problems, such as deep linear neural networks [106] and matrix regression [58]. It is natural to analyze the continuous-time limit of an online algorithm for tracking a KKT trajectory of time-varying optimization problem [111, 121, 89, 42].

2.2 Preliminaries and Problem Formulation

Time-varying optimization with equality constraints

The first-order KKT conditions for the time-varying optimization (2.1) are as follows:

$$0 = \nabla_x f(x(t), t) + \mathcal{J}_g(x(t), t)^\intercal \lambda(t) \tag{2.2a}$$

$$0 = g(x(t), t) \tag{2.2b}$$

where $\mathcal{J}_g(z,t) \coloneqq \frac{\partial g(z,t)}{\partial z}$ denotes the Jacobian of $g(\cdot,\cdot)$ with respect to the first argument and $\lambda(t) \in \mathbb{R}^m$ is a Lagrange multiplier associated with the equality constraint. We first make some assumptions below.

Assumption 1. *$f : \mathbb{R}^n \times [0,\infty) \to \mathbb{R}$ is twice continuously differentiable in $x \in \mathbb{R}^n$ and continuously differentiable in $t \geq 0$. $g_k : \mathbb{R}^n \times [0,\infty) \to \mathbb{R}$ is twice continuously differentiable in $x \in \mathbb{R}^n$ and twice continuously differentiable in $t \geq 0$ for $k = 1, \ldots, m$. Moreover, at any given*

time t, $f(x,t)$ is uniformly bounded from below over the set $\{x \in \mathbb{R}^n : g(x,t) = 0\}$, meaning that there exists a constant M such that $f(x,t) \geq M$ for all $x \in \{x \in \mathbb{R}^n : g(x,t) = 0\}$ and $t \geq 0$.

Assumption 2. *The feasible set at t defined as*

$$\mathcal{M}(t) \coloneqq \{x \in \mathbb{R}^n : g(x,t) = 0\}$$

is nonempty for all $t \geq 0$.

Assumption 3. *For all $t \geq 0$ and $x \in \mathcal{M}(t)$, the matrix $\mathcal{J}_g(x,t)$ has full row-rank.*

Remark 1. *Although Assumption 3 is somewhat stronger than the Linear independence constraint qualification [13], it is necessary for our following analysis because with different values of α and different initial points, the solution of* (P-ODE) *may land anywhere in the feasible region. Furthermore, Sard's theorem [105] ensures that if the constraint function $g(\cdot,t)$ is sufficiently smooth, then the set of values of $g(\cdot,t)$, denoted as $\mathcal{S}(t)$, for which $\mathcal{J}_g(x,t)$ is not full row-rank has measure 0. Thus, Assumption 3 is satisfied if $0 \notin \mathcal{S}(t)$ where $\mathcal{S}(t)$ is only a set with measure 0. Finally, if the inertia parameter α is fixed and the initial point of* (P-ODE) *is a local solution, then the work [42] provides a sophisticated proof for the existence and uniqueness of the solution for a special class of* (P-ODE) *under a minor assumption that the Jacobian has full-row rank only at the discrete local trajectories (which is defined in the paragraph after equation* (2.10) *in our work). However, to be able to study the solution of* (P-ODE) *for all $\alpha > 0$ and any initial feasible point and keep the focus of the paper on studying the escaping behavior, we made Assumption 3.*

Under Assumption 3, the matrix $\mathcal{J}_g(x(t),t)\mathcal{J}_g(x(t),t)^\top$ is invertible and therefore $\lambda(t)$ in (2.2a) can be written as

$$\lambda(t) = -(\mathcal{J}_g(x(t),t)\mathcal{J}_g(x(t),t)^\top)^{-1}\mathcal{J}_g(x(t),t)\nabla_x f(x(t),t) \tag{2.3}$$

Since $\lambda(t)$ is written as a function of $x(t)$ in (2.3), we also denote it as $\lambda(x(t),t)$. Now, (2.2a) can be written as

$$0 = \Big[I_n - \mathcal{J}_g(x(t),t)^\top(\mathcal{J}_g(x(t),t)\mathcal{J}_g(x(t),t)^\top)^{-1}\mathcal{J}_g(x(t),t)\Big]\nabla_x f(x(t),t) \tag{2.4}$$

where I_n is the identity matrix in $\mathbb{R}^{n\times n}$. For the sake of readability, we introduce the symbolic notation

$$\mathcal{P}(x(t),t) \coloneqq I_n - \mathcal{J}_g(x(t),t)^\top(\mathcal{J}_g(x(t),t)\mathcal{J}_g(x(t),t)^\top)^{-1}\mathcal{J}_g(x(t),t)$$

which is the orthogonal projection operation onto T_x^t, where T_x^t denotes the tangent plane of $g(x(t),t)$ at the point $x(t)$ and the time t. It is convenient and conventional to introduce the time-varying Lagrange functional

$$L(x,\lambda,t) = f(x,t) + \lambda g(x,t) \tag{2.5}$$

In terms of this functional, (2.4) can be written as

$$0 = \nabla_x L(x, \lambda, t) \tag{2.6}$$

where λ is given in (2.3). Here, $\nabla_x L(x, \lambda, t)$ means first taking the partial gradient with respect to the first argument and then using the formula (2.3) for λ. Since the solution is time-varying, we define the notion of the local (or global) minimum trajectory below.

Definition 1. *A continuous trajectory $h : I_t \to \mathbb{R}^n$, where $I_t \subseteq [0, \infty)$, is said to be a **local (or global) minimum trajectory** of the time-varying optimization (2.1) if each point of $h(t)$ is a local (or global) minimum of the time varying optimization (2.1) for every $t \in I_t$.*

In this chapter, we focus on the case when the local minimum trajectories will not cross, bifurcate or disappear by assuming the following uniform regularity condition.

Assumption 4. *For each local minimum trajectory $h(t)$, its domain I_t is $[0, \infty)$ and $h(t)$ satisfies the second-order sufficient optimality conditions uniformly, meaning that $\nabla^2_{xx} L(h(t), \lambda, t)$ is positive definite on $T^t_{h(t)} = \{y : \mathcal{J}_g(h(t), t)^\top y = 0\}$ for all $t \in [0, \infty)$.*

Lemma 1. *Under Assumptions 1-4, each local minimum trajectory $h(t)$ is differentiable and isolated, and therefore it can not bifurcate or merge with other local minimum trajectories.*

After freezing the time t in (2.1) at a particular value, one may use local search methods, like Rosen's gradient projection method [102], to minimize $f(x, t)$ over the feasible region $\mathcal{M}(t)$. If the initial point is feasible and close enough to a local solution and the step size is small enough, the algorithm will converge to the local minimum. This leads to the notion of region of attraction defined by resorting to the continuous-time model of Rosen's gradient projection method [119] (for which the step size is not important anymore).

Definition 2. *The **region of attraction** of a local minimum point $h(t)$ of $f(\cdot, t)$ in the feasible set $\mathcal{M}(t)$ at a given time t is defined as:*

$$RA^{\mathcal{M}(t)}(h(t)) = \Big\{ x_0 \in \mathcal{M}(t) \;\Big|\; \lim_{\tilde{t} \to \infty} \tilde{x}(\tilde{t}) = h(t) \quad \text{where}$$
$$\frac{d\tilde{x}(\tilde{t})}{d\tilde{t}} = -\mathcal{P}(\tilde{x}(\tilde{t}), t) \nabla_x f(\tilde{x}(\tilde{t}), t) \quad \text{and} \quad \tilde{x}(0) = x_0 \Big\}.$$

In the unconstrained case, the notion of the locally one-point strong convexity can be defined as follows:

Definition 3. *Consider arbitrary positive scalars c and r. The function $f(x, t)$ is said to be **locally (c, r)-one-point strongly convex** around the local minimum trajectory $h(t)$ if*

$$\nabla_x f(e + h(t), t)^\top e \geq c \|e\|^2, \quad \forall e \in D, \quad \forall t \in [0, \infty) \tag{2.7}$$

where $D = \{e \in \mathbb{R}^n : \|e\| \leq r\}$. The region $D = \{e \in \mathbb{R}^n : \|e\| \leq r\}$ is called the region of locally (c, r)-one-point strong convexity around $h(t)$.

This definition resembles the (locally) strong convexity condition for the function $f(x,t)$, but it is only expressed around the point $h(t)$. This restriction to a single point constitutes the definition of one-point strong convexity and it does not imply that the function is convex. The following result paves the way for the generalization of the notion of the locally one-point strong convexity from the unconstrained case to the equality constrained case.

Lemma 2. *Consider an arbitrary local minimum trajectory $h(t)$ satisfying Assumption 4, there exist positive constants $\hat{r}$ and $\hat{c}$ such that*

$$e(t)^\top \nabla_x L(e(t)+h(t), \lambda(e(t)+h(t),t), t) \geq \hat{c}\,\|e(t)\|^2$$

for all $e(t) \in \{e + h(t) \in \mathcal{M}(t) : \|e\| \leq \hat{r}\}$.

Definition 4. *Consider arbitrary positive scalars c and r. The Lagrange function $L(x,\lambda,t)$ with λ given in (2.3) is said to be **locally (c,r)-one-point strongly convex** with respect to x around the local minimum trajectory $h(t)$ in the feasible set $\mathcal{M}(t)$ if:*

$$e^\top \nabla_x L(e+h(t), \lambda(e+h(t),t), t) \geq c\,\|e\|^2 \tag{2.8}$$

for all $e \in D^{\mathcal{M}(t)}$ and $t \in [0,\infty)$, where $D^{\mathcal{M}(t)} = \{e + h(t) \in \mathcal{M}(t) : \|e\| \leq r\}$. The region $D^{\mathcal{M}(t)} = \{e + h(t) \in \mathcal{M}(t) : \|e\| \leq r\}$ is called the region of locally (c,r)-one-point strong convexity of the Lagrange function $L(x,\lambda,t)$ around $h(t)$ in the feasible set $\mathcal{M}(t)$.

Remark 2. *The Lagrange function $L(x,\lambda,t)$ with λ given in (2.3) being locally (c,r)-one-point strongly convex with respect to x around $h(t)$ is equivalent to the vector field $\mathcal{P}(x,t)\nabla_x f(x(t),t)$ being **locally (c,r)-one-point strongly monotone** with respect to x around $h(t)$.*

Derivation of time-varying projected gradient flow system

In practice, one can only hope to sequentially solve the time-varying optimization problem (2.1) at some discrete time instances $0 = \tau_0 < \tau_1 < \tau_2 < \tau_3 < \dots$ as follows:

$$\min_{x\in\mathbb{R}^n} \quad f(x,\tau_i), \quad \text{s.t.} \quad g(x,\tau_i) = 0, \quad i = 1,2,\dots \tag{2.9}$$

In many real-world applications, it is neither practical nor realistic to have solutions that abruptly change over time. To meet this requirement, we impose a soft constraint to the objective function by penalizing the deviation of its solution from the one obtained in the previous time step. This leads to the following sequence of optimization problems with **proximal regularization** (except for the initial optimization problem):

$$\begin{aligned}
&\min_{x\in\mathbb{R}^n} \quad f(x,\tau_0), && \text{(2.10a)}\\
&\text{s.t.} \quad g(x,\tau_0) = 0,\\
&\min_{x\in\mathbb{R}^n} \quad f(x,\tau_i) + \frac{\alpha}{2(\tau_i-\tau_{i-1})}\,\|x - x^*_{i-1}\|^2, && \text{(2.10b)}\\
&\text{s.t.} \quad g(x,\tau_i) = 0, \quad i = 1,2,\dots
\end{aligned}$$

where x_{i-1}^* denotes an arbitrary local minimum of the modified optimization problem (2.10) obtained using a local search method at time iteration $i-1$. A local optimal solution sequence $x_0^*, x_1^*, x_2^*, \ldots$ is said to be a **discrete local trajectory** of the sequential regularized optimization (2.10). The parameter α is called inertia because it acts as a resistance to changes x at time step τ_i with respect to x at the previous time step τ_{i-1}. Note that α could be time-varying (and adaptively changing) in the analysis of this chapter, but we restrict our attention to a fixed regularization term to simplify the presentation.

Under Assumption 3, all solutions x^* of (2.10b) must satisfy the KKT conditions:

$$0 = \nabla_x f(x_i^*, \tau_i) + \alpha \frac{x_i^* - x_{i-1}^*}{\tau_i - \tau_{i-1}} + \mathcal{J}_g(x_i, \tau_i)^\top \bar{\lambda}_i, \tag{2.11a}$$

$$0 = g(x_i, \tau_i), \tag{2.11b}$$

where $\bar{\lambda}_i$'s are the Lagrange multipliers for the sequence of optimization problems with proximal regularization in (2.10). Similar to [89], we can write the right-hand side of the constraint (2.11b) as:

$$\frac{g(x_i, \tau_i) - g(x_i, \tau_{i-1}) + g(x_i, \tau_{i-1}) - g(x_{i-1}, \tau_{i-1})}{\tau_i - \tau_{i-1}} \tag{2.12}$$

Since the function $f(x,t)$ and $g(x,t)$ are nonconvex in general, the problem (2.10) may not have a unique solution x_i^*. In order to cope with this issue, we study the continuous-time limit of (2.11) as the time step $\tau_{i+1} - \tau_i$ diminishes to zero. This yields the following time-varying ordinary differential equations:

$$0 = \nabla_x f(x(t), t) + \alpha \dot{x}(t) + \mathcal{J}_g(x(t), t)^\top \bar{\lambda}(t), \tag{2.13a}$$

$$0 = \mathcal{J}_g(x(t), t)\dot{x}(t) + g'(x(t), t), \tag{2.13b}$$

where $g' = \frac{\partial g(x,t)}{\partial t}$ denotes the partial derivative of g with respect to t. Since $\mathcal{J}_g(x(t),t)\mathcal{J}_g(x(t),t)^\top$ is invertible, we have

$$\begin{aligned} 0 =& (\mathcal{J}_g(x(t),t)\mathcal{J}_g(x(t),t)^\top)^{-1}\mathcal{J}_g(x(t),t)\nabla_x f(x(t),t) \\ & - \alpha(\mathcal{J}_g(x(t),t)\mathcal{J}_g(x(t),t)^\top)^{-1} g'(x(t),t) + \bar{\lambda}(t). \end{aligned} \tag{2.14}$$

Therefore, $\bar{\lambda}(t)$ can be written as a function of x, t and α:

$$\begin{aligned} \bar{\lambda}(t) =& -(\mathcal{J}_g(x(t),t)\mathcal{J}_g(x(t),t)^\top)^{-1}\mathcal{J}_g(x(t),t)\nabla_x f(x(t),t) \\ & + \alpha(\mathcal{J}_g(x(t),t)\mathcal{J}_g(x(t),t)^\top)^{-1} g'(x(t),t) \\ =& \lambda(x(t),t) + \alpha(\mathcal{J}_g(x,t)\mathcal{J}_g(x,t)^\top)^{-1} g'(x,t) \end{aligned} \tag{2.15}$$

We alternatively denote $\bar{\lambda}(t)$ as $\bar{\lambda}(x(t), t, \alpha)$. When $\alpha = 0$, we have $\bar{\lambda}(x(t), t, \alpha) = \lambda(x(t), t)$ and the differential equation (2.13) reduces to the algebraic equation (2.2), which is indeed

the first-order KKT condition for the unregularized time-varying optimization (2.1). When $\alpha > 0$, substituting $\bar{\lambda}(x(t), t, \alpha)$ into (2.13a) yields the following time-varying ODE:

$$\dot{x}(t) = -\frac{1}{\alpha}\mathcal{P}(x(t),t)\nabla_x f(x(t),t) - \mathcal{Q}(x(t),t)g'(x(t),t), \tag{P-ODE}$$

where $\mathcal{Q}(x(t),t) = \mathcal{J}_g(x(t),t)^\top(\mathcal{J}_g(x(t),t)\mathcal{J}_g(x(t),t)^\top)^{-1}$. In terms of the Lagrange functional, (P-ODE) can be written as

$$\dot{x} = -\frac{1}{\alpha}\nabla_x L(x,\bar{\lambda},t) = -\frac{1}{\alpha}\nabla_x L(x,\lambda,t) - \mathcal{Q}(x,t)g'(x,t). \tag{2.16}$$

Here, $\nabla_x L(x,\bar{\lambda},t)$ means first taking the partial gradient with respect to the first argument and then using the formula (2.15) for $\bar{\lambda}$. It can be shown that if the initial point of (P-ODE) is in the feasible set $M(t_0)$, the solution of (P-ODE) will stay in the feasible set $M(t)$.

Lemma 3. *Suppose that the solution $x(t,t_0,x_0)$ of* (P-ODE) *is defined in $[t_0,\infty)$ with the initial point x_0. If $x_0 \in \mathcal{M}(t_0)$, then the solution $x(t,t_0,x_0)$ belongs to $\mathcal{M}(t)$ for all $t \geq t_0$.*

Therefore, as long as the initial point of (P-ODE) is in the feasible set $M(t_0)$, the above lemma guarantees that we can analyze the stability of (P-ODE) using the standard Lyapunov's theorem without worrying about the feasibility of the solution. When $\alpha > 0$, we will show that for any initial point x_0, (P-ODE) has a unique solution defined for all $t \in I_t \subseteq [0,\infty)$ if there exists a local minimum trajectory $h(t)$ such that the solutions of (P-ODE) lie in a compact set around $h(t)$ [1].

Theorem 1 (Existence and uniqueness)**.** *Under Assumptions 1-4 and given any initial point $x_0 \in \mathcal{M}(t_0)$, suppose that there exists a local minimum trajectory $h(t)$ with the property that $x(t) - h(t)$ lies entirely in D for all $t \in I_t \subseteq [0,\infty)$ where D is a compact subset of $\mathbb{R}^n$ containing $x_0 - h(t_0)$ tand $x(t)$ denotes the solution of* (P-ODE) *with the initial point x_0. Then,* (P-ODE) *has a unique solution starting from x_0 that is defined for all $t \geq 0$.*

In online optimization, it is sometimes desirable to predict the solution at a future time (namely, τ_i) only based on the information at the current time (namely, τ_{i-1}). This can be achieved by implementing the forward Euler method to obtain a numerical approximation to the solution of (P-ODE):

$$\bar{x}_i^* = \bar{x}_{i-1}^* - (\tau_i - \tau_{i-1})\Big(\frac{1}{\alpha}\mathcal{P}(\bar{x}_{i-1}^*,\tau_{i-1})\nabla_x f(\bar{x}_{i-1}^*,\tau_{i-1} + \mathcal{Q}(\bar{x}_{i-1}^*,\tau_{i-1})g'(\bar{x}_{i-1}^*,\tau_{i-1})\Big) \tag{2.17}$$

(note that $\bar{x}_0^*, \bar{x}_1^*, \bar{x}_2^*, \ldots$ show the approximate solutions). The following theorem explains the reason behind studying the continuous-time problem (P-ODE) in the remainder of this chapter.

[1] In Theorems 3 and 4, the compactness assumption is included in the definition of the dominant trajectory. In Theorem 5, checking the compactness assumption can be carried out via the Lyapunov's method without solving the differential equation due to the one-point strong convexity condition around $h(t)$.

Theorem 2 (Convergence)**.** *Under Assumptions 1-4 and given a local minimum x_0^* of* (2.10a)*, as the time difference $\Delta_\tau = \tau_{i+1} - \tau_i$ approaches zero, any sequence of discrete local trajectories (x_k^Δ) converges to the* (P-ODE) *in the sense that for all fixed $T > 0$:*

$$\lim_{\Delta_\tau \to 0} \max_{0 \le k \le \frac{T}{\Delta\tau}} \left\| x_k^\Delta - x(\tau_k, \tau_0, x_0^*) \right\| = 0 \tag{2.18}$$

and any sequence of $(\bar{x}_k^\Delta)$ updated by (2.17) *converges to the* (P-ODE) *in the sense that for all fixed $T > 0$:*

$$\lim_{\Delta_\tau \to 0} \max_{0 \le k \le \frac{T}{\Delta\tau}} \left\| \bar{x}_k^\Delta - x(\tau_k, \tau_0, x_0^*) \right\| = 0 \tag{2.19}$$

Theorem 2 guarantees that the solution of (P-ODE) is a reasonable approximation in the sense that it is the continuous-time limit of both the solution of the sequential regularized optimization problem (2.10) and the solution of the online updating scheme (2.17). For this reason, we only study the continuous-time problem (P-ODE) in the remainder of this chapter.

Jumping, tracking and escaping

In this chapter, the objective is to study the case where there are at least two local minimum trajectories of the online time-varying optimization problem. Consider two local minimum trajectories $h_1(t)$ and $h_2(t)$. We provide the definitions of jumping, tracking and escaping below.

Definition 5. *It is said that the solution of* (P-ODE) ***(v,u)-jumps*** *from $h_1(t)$ to $h_2(t)$ over the time interval $[t_1, t_2]$ if there exist $u > 0$ and $v > 0$ such that*

$$\mathcal{B}_v(h_1(t_1)) \cap \mathcal{M}(t_1) \subseteq RA^{\mathcal{M}(t_1)}(h_1(t_1)) \tag{2.20a}$$

$$\mathcal{B}_u(h_2(t_2)) \cap \mathcal{M}(t_2) \subseteq RA^{\mathcal{M}(t_2)}(h_2(t_2)) \tag{2.20b}$$

$$\forall x_1 \in \mathcal{B}_v(h_1(t_1)) \cap \mathcal{M}(t_1) \Longrightarrow x(t_2, t_1, x_1) \in \mathcal{B}_u(h_2(t_2)) \cap \mathcal{M}(t_2) \tag{2.20c}$$

Definition 6. *Given $x_0 \in \mathcal{M}(t_0)$, it is said that $x(t, t_0, x_0)$* ***u-tracks*** *$h_2(t)$ if there exist a finite time $T > 0$ and a constant $u > 0$ such that*

$$x(t, t_0, x_0) \in \mathcal{B}_u(h_2(t)) \cap \mathcal{M}(t), \qquad \forall t \ge T \tag{2.21a}$$

$$\mathcal{B}_u(h_2(t)) \cap \mathcal{M}(t) \subseteq RA^{\mathcal{M}(t)}(h_2(t)), \qquad \forall t \ge T \tag{2.21b}$$

In this chapter, the objective is to study the scenario where a solution $x(t, t_0, x_0)$ tracking a poor solution $h_1(t)$ at the beginning ends up tracking a better solution $h_2(t)$ after some time. This needs the notion of "escaping" which is a combination of jumping and tracking.

Definition 7. *It is said that the solution of* (ODE) ***(v,u)-escapes*** *from $h_1(t)$ to $h_2(t)$ if there exist $T > 0$, $u > 0$ and $v > 0$ such that*

$$\mathcal{B}_v(h_1(t_0)) \cap \mathcal{M}(t_0) \subseteq RA^{\mathcal{M}(t_0)}(h_1(t_0)) \tag{2.22a}$$

$$\mathcal{B}_u(h_2(t)) \cap \mathcal{M}(t) \subseteq RA^{\mathcal{M}(t)}(h_2(t)), \ \ \forall t \ge T \tag{2.22b}$$

$$\forall x_0 \in \mathcal{B}_v(h_1(t_0)) \cap \mathcal{M}(t_0) \Longrightarrow x(t, t_0, x_0) \in \mathcal{B}_u(h_2(t)) \cap \mathcal{M}(t), \ \ \forall t \ge T \tag{2.22c}$$

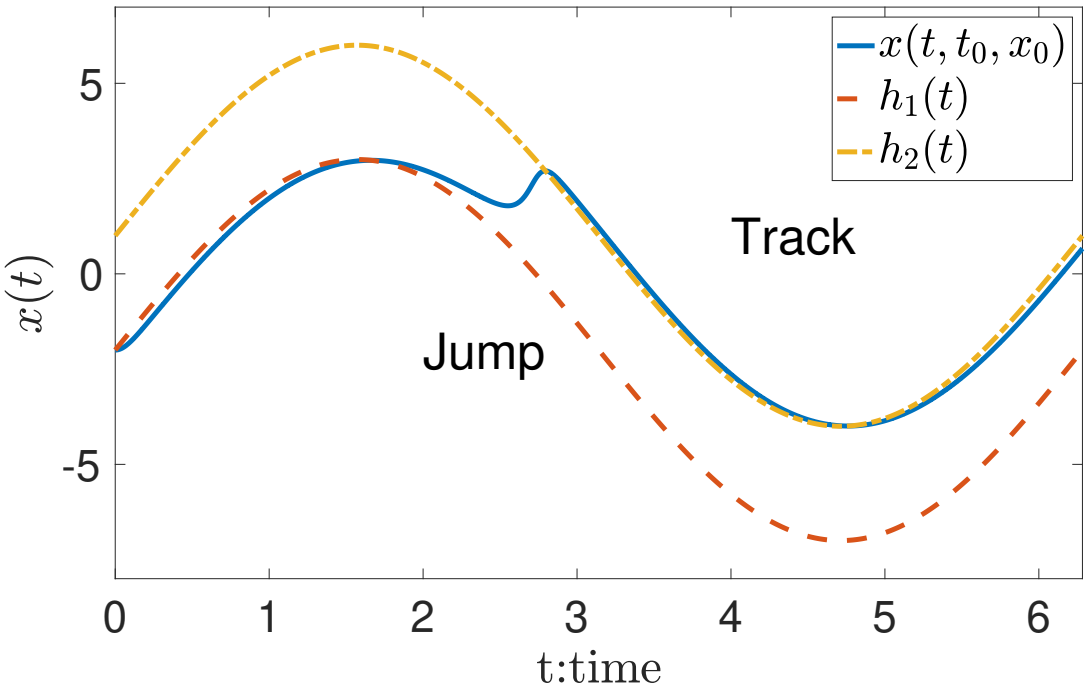


Figure 2.3: Illustration of jumping and tracking.

Figure 2.3 illustrates the definitions of jumping and tracking for Example 1 with $\alpha = 0.3$ and $b = 5$. The objective of this chapter is to study when the solution of (P-ODE) started at a poor local minimum at the initial time jumps to and tracks a better (or global) minimum of the problem after some time. In other words, it is desirable to investigate the escaping property from $h_1(t)$ and $h_2(t)$.

2.3 Change of variables

Given two isolated local minimum trajectories $h_1(t)$, $h_2(t)$. One may use the change of variables $x(t, t_0, x_0) = e(t, t_0, e_0) + h_2(t)$ to transform (P-ODE) into the form

$$\dot{e}(t) = -\frac{1}{\alpha}\mathcal{P}(e(t) + h_2(t), t)\nabla_x f(e(t) + h_2(t), t) - \mathcal{Q}(e(t) + h_2(t), t)g'(e(t) + h_2(t), t) - \dot{h}_2(t) \tag{2.23a}$$

$$= -\frac{1}{\alpha}\nabla_x\Big(L\big(e(t) + h_2(t), \bar{\lambda}(e(t) + h_2(t), t, \alpha), t\big) + \alpha\dot{h}_2(t)^\top e(t)\Big) \tag{2.23b}$$

We use $e(t, t_0, e_0)$ to denote the solution of this differential equation starting at time $t = t_0$ with the initial point $e_0 = x_0 - h_2(t_0)$ and use $-\frac{1}{\alpha}U(e(t), t, \alpha)$ to denote the right-hand side of (2.23).

Note that $h_1(t)$ and $h_2(t)$ are local solutions of (2.1) and as long as (2.1) is time-varying, these functions cannot satisfy (P-ODE) in general. We denote $\mathcal{M}^h(t) \coloneqq \{e \in \mathbb{R}^n : g(e + h(t), t) = 0\}$.

Unconstrained optimization landscape after a change of variables

In this subsection, we study the unconstrained case to enable a better visualization of the optimization landscape. In the unconstrained case, (2.23) is reduced to

$$\dot{e}(t) = -\frac{1}{\alpha}\nabla_x f(e(t) + h_2(t), t) - \dot{h}_2(t). \tag{2.24}$$

Inertia encouraging the exploration

The first term $\nabla_x f(e + h_2(t), t)$ in (2.24) can be understood as a time-varying gradient term that encourages the solution of (2.24) to track $h_2(t)$, while the second term $\dot{h}_2(t)$ represents the inertia from this trajectory. In particular, if $\dot{h}_2(t)$ points toward outside of the region of attraction of $h_2(t)$ during some time interval, the term $\dot{h}_2(t)$ acts as an **exploration** term that encourages the solution of (ODE) to leave the region of attraction of $h_2(t)$. The parameter α balances the roles of the gradient and the inertia.

In the extreme case where α goes to infinity, $e(t)$ converges to $-h_2(t)$ and $x(t)$ approaches a constant trajectory determined by the initial point x_0; when α is sufficiently small, the time-varying gradient term dominates the inertia term and the solution of (ODE) would track $h_2(t)$ closely. With an appropriate proximal regularization α that keeps the balance between the time-varying gradient term and the inertia term, the solution of (ODE) could temporarily track a local minimum trajectory with the potential of exploring other local minimum trajectories.

Inertia creating a one-point strongly convex landscape

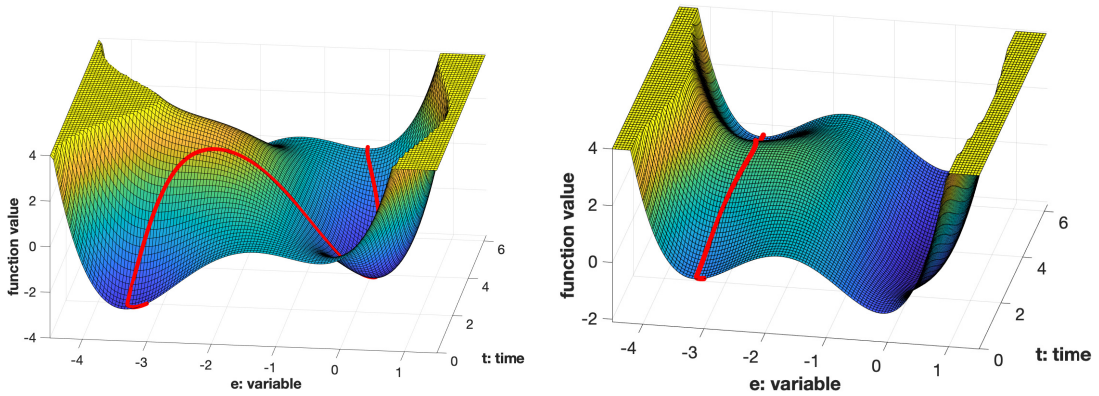
The differential equation (2.24) can be written as

$$\dot{e}(t) = -\frac{1}{\alpha}\nabla_e\Big(f(e(t) + h_2(t), t) + \alpha\dot{h}_2(t)^\top e(t)\Big) \tag{2.25}$$

This can be regarded as a time-varying gradient flow system of the original objective function $f(e + h_2(t), t)$ plus a time-varying perturbation $\alpha\dot{h}_2(t)^\top e$. During some time interval $[t_1, t_2]$, the time-varying perturbation $\alpha\dot{h}_2(t)^\top e$ may enable the time-varying objective function $f(e+h_2(t), t)+\alpha\dot{h}_2(t)^\top e$ over a neighborhood of $h_1(t)$ to become one-point strongly convexified with respect to $h_2(t)$. Under such circumstances, the time-varying perturbation $\alpha\dot{h}_2(t)^\top e$ prompts the solution of (2.25) starting in a neighborhood of $h_1(t)$ to move towards a neighborhood of $h_2(t)$. Before analyzing this phenomenon, we illustrate the concept in an example.

Consider again Example 1 and recall that $\bar{f}(x)$ has 2 local minima at $x = -2$ and $x = 1$. By taking $b = 5$, $h_1(t) = -2 + 5\sin(t)$ and $h_2(t) = 1 + 5\sin(t)$, the differential equation (2.25)

can be expressed as $\dot{e}(t) = -\frac{1}{\alpha}\nabla_e\Big(\bar{f}(1+e(t)) + 5\alpha\cos(t)e(t)\Big)$. The landscape of the new time-varying function $\bar{f}(1+e) + 5\alpha\cos(t)e$ with the variable e is shown for two cases $\alpha = 0.3$ and $\alpha = 0.1$ in Figure 2.4. The red curves are the solutions of (2.25) starting from $e = -3$. One can observe that when $\alpha = 0.3$, the new landscape becomes one-point strongly convex around $h_2(t)$ over the whole region for some time interval, which provides (2.25) with the opportunity of escaping from the region around $h_1(t)$ to the region around $h_2(t)$. However, when $\alpha = 0.1$, there are always two locally one-point strongly convex regions around $h_1(t)$ and $h_2(t)$ and, therefore, (2.25) fails to escape the region around $h_1(t)$.

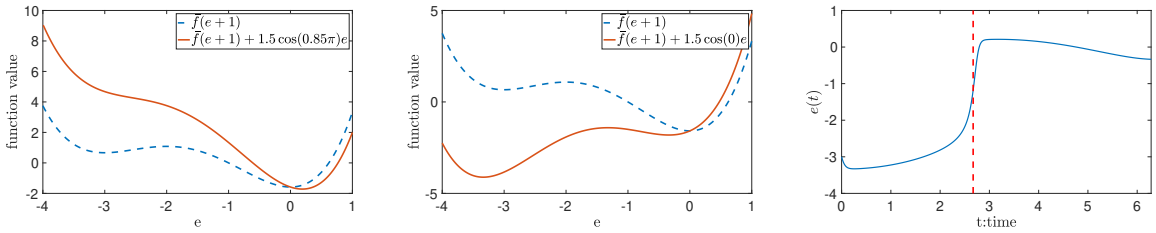


(a) $\bar{f}(1+e) + 1.5\cos(t)e$ (b) $\bar{f}(1+e) + 0.5\cos(t)e$

Figure 2.4: Illustration of time-varying landscape after change of variables for Example 1.

To further inspect the case $\alpha = 0.3$, observe in Figure 2.5a that the landscape of the objective function $\bar{f}(1+e) + 1.5\cos(0.85\pi)e$ shows that the region around the spurious local minimum trajectory $h_1(t)$ is one-point strongly convexified with respect to $h_2(t)$ at time $t = 0.85\pi$. This is consistent with the fact that the solution of $\dot{e} = -\frac{1}{0.3}\nabla_x\bar{f}(1+e) - 5\cos(t)$ starting from $e = -3$ jumps to the neighborhood of 0 around time $t = 0.85\pi$, as demonstrated in Figure 2.5c. Furthermore, if the time interval $[t_1, t_2]$ is large enough to allow transitioning from a neighborhood of $h_1(t)$ to a neighborhood of $h_2(t)$, then the solution of (2.25) would move to the neighborhood of $h_2(t)$. In contrast, the region around $1 + b\sin(t)$ is never one-point strongly convexified with respect to $-2 + b\sin(t)$, as shown in Figure 2.5b.

From the right-hand side of (2.25), it can be inferred that if the gradient of $f(\cdot, t)$ is relatively small around some local minimum trajectory, then its landscape is easier to be re-shaped by the time-varying linear perturbation $\alpha\dot{h}_2(t)^\top e$. The local minimum trajectory in a neighborhood with small gradients usually corresponds to a shallow minimum trajectory in which the trajectory has a relatively flat landscape and a relatively small region of attraction. Thus, the one-point strong convexication introduced by the time-varying perturbation could help escape the shallow minimum trajectories.



(a) $\bar{f}(1+e)+1.5\cos(0.85\pi)e$ (b) $\bar{f}(1+e)+1.5\cos(0)e$ (c) solution of $\dot{e}=-\frac{1}{0.3}\nabla_x\bar{f}(1+e)-5\cos(t)$ starting from $e_0=-3$

Figure 2.5: Illustration of one-point strong convexification for Example 1.

Table 2.1: A unified view for unconstrained and equality-constrained problems

	Unconstrained problem	Equality-constrained problem
First-order optimality condition(FOC)	$0=\nabla_x f(x,t)$	$0=\nabla_x L(x,\lambda,t)$
ODE (continuous time limit of FOC for regularized problem)	$\dot{x}=-\frac{1}{\alpha}\nabla_x f(x,t)$	$\dot{x}=-\frac{1}{\alpha}\nabla_x L(x,\bar{\lambda},t)$
Change of variables: $x=h+e$	$\dot{e}=-\frac{1}{\alpha}\nabla_e f(e+h,t)-\dot{h}$	$\dot{e}=-\frac{1}{\alpha}\nabla_e L(e+h,\bar{\lambda},t)-\dot{h}$
Key assumption: one-point strong convexity	$e^\top\nabla_e f(e+h,t)\geq c\left\|e\right\|^2$	$e^\top\nabla_e L(e+h,\lambda,t)\geq c\left\|e\right\|^2$
Reshaping of the landscape: one-point strong convexification	$e^\top\left(\nabla_e f(e+h,t)+\alpha\dot{h}\right)\geq w\left\|e\right\|^2$	$e^\top\left(\nabla_e L(e+h,\bar{\lambda},t)+\alpha\dot{h}\right)\geq w\left\|e\right\|^2$

Dominant trajectory

In this subsection, we will formalize the intuitions discussed in Section 2.3. We first define the notion of the shallow local minimum trajectory.

Definition 8. *Consider a positive number α and assume that $\dot{h}_1(t)$ is L-Lipschitz continuous. It is said that the local minimum trajectory $h_1(t)$ is α-**shallow** during the time period $[t_0,t_0+\delta]$ if*

$$\epsilon>E(\alpha)+L\delta \text{ and } r\leq\frac{1}{2}\delta(\epsilon-E(\alpha)-L\delta),$$

where $\epsilon=\sup_{t\in[t_0,t_0+\delta]}\left\|\dot{h}_1(t)\right\|$, $r=\sup_{t\in[t_0,t_0+\delta]}\sup_{x(t)\in RA^{\mathcal{M}(t)}(h_1(t))}\left\|x(t)-h_1(t)\right\|$, $E(\alpha)=\sup_{t\in[t_0,t_0+\delta]}\sup_{x(t)\in RA^{\mathcal{M}(t)}(h_1(t))}\left\|\frac{1}{\alpha}\nabla_x L(x,\bar{\lambda},t)\right\|$, $\frac{1}{\alpha}\nabla_x L(x,\bar{\lambda},t)$ is defined in (2.16).

In other words, a local minimum trajectory is shallow if it has a large time variation but a small region of attraction. We next show that whenever a local minimum trajectory $h_1(t)$ is

shallow during some time interval, the solution of (P-ODE) starting anywhere in the region of attraction of $h_1(t)$ will leave its region of attraction at some time.

Lemma 4. *If the local minimum trajectory $h_1(t)$ is α-shallow during $[t_0, t_0+\delta]$, then for any $x(t_0) \in RA^{\mathcal{M}(t_0)}(h_1(t_0))$, then there exists a time $t \in [t_0, t_0+\delta]$ such that $x(t) \notin RA^{\mathcal{M}(t)}(h_1(t))$.*

On the one hand, Lemma 4 shows that any shallow local minimum trajectory is unstable in the sense that the time-variation in the minimum trajectory will force the solution of (P-ODE) to leave its region of attraction. If the shallow local minimum trajectory happens to be a non-global local solution, then the solution of (P-ODE), acting as a tracking algorithm, will help avoid the bad local solutions for free. On the other hand, Lemma 4 does not specify where the solution of (P-ODE) will end up after leaving the region of attraction of a shallow local minimum trajectory. Simulations (such as those provided in Sections 2.3 and 2.5) suggest that, with some appropriate α, the solution of (P-ODE) may move towards a nearby local minimum trajectory that has an enlarged region of one-point strong convexity. This leads to the following definition of the region of the domination and the dominant local minimum trajectory.

Definition 9. *Given two local minimum trajectories $h_1(t)$ and $h_2(t)$, suppose that the time-varying Lagrange function $L(x, \lambda, t)$ with λ given in (2.3) is locally (c_2, r_2)-one-point strongly convex with respect to x around $h_2(t)$ in the region $\mathcal{M}^{h_2}(t) \cap \mathcal{B}_{r_2}(0)$. A set D_{v,ρ,r_2} is said to be **the region of domination** for $h_2(t)$ with respect to $h_1(t)$ if it satisfies the following properties:*

- *D_{v,ρ,r_2} is a compact subset such that*

$$e_1 \in D_{v,\rho,r_2} \Rightarrow e(t, t_1, e_1) \in D_{v,\rho,r_2}, \forall t \in [t_1, t_2] \tag{2.26}$$

  where $e(t, t_1, e_1)$ is the solution of (2.23) staring from the feasible initial point $e_1 \in \mathcal{M}^{h_2}(t_1)$ at the initial time t_1.

- *$D_{v,\rho,r_2} \supseteq D'_v \cup \mathcal{B}_\rho(0)$ where*

$$D'_v = \{e_1 \in \mathbb{R}^n : e_1 + h_2(t_1) \in \mathcal{M}(t_1) \cap \mathcal{B}_v(h_1(t_1)) \subseteq RA^{\mathcal{M}(t_1)}(h_1(t_1))\}, \tag{2.27}$$

$$\rho \geq \sup_{t\in[t_1,t_2]} \sup_{\substack{\bar{e}(t):\|\bar{e}(t)\|<r_2,\\ 0=U(\bar{e}(t),t,\alpha)}} \|\bar{e}(t)\|. \tag{2.28}$$

The condition (2.26) is a set invariance property, which requires that the solution of (2.23) starting from an initial point in D_{v,ρ,r_2} stays in D_{v,ρ,r_2} during the time period $[t_1, t_2]$. For the visualization of D_{v,ρ,r_2}, $\mathcal{B}_\rho$ and D'_v in Definition 9, we consider again Example 1. In Fig 2.6, the red curve corresponds to the landscape of the function $\bar{f}(1+e) + 1.5\cos(0.85\pi)e$, $e = 0$ corresponds to $h_2(t)$ and $e = -3$ corresponds to $h_1(t)$. $\mathcal{B}_\rho$ is a region around $h_2(t)$ containing all zeros of $0 = U(\cdot, t, \alpha)$ during a time period around 0.85π and D'_v is a neighborhood around $h_1(t)$. In this example, the region of domination for $h_2(t)$ with respect to $h_1(t)$ is $D_{v,\rho,r_2} = [-4, 1]$ which contains $\mathcal{B}_\rho$ and D'_v if $h_1(t)$ if it also satisfies (2.26).

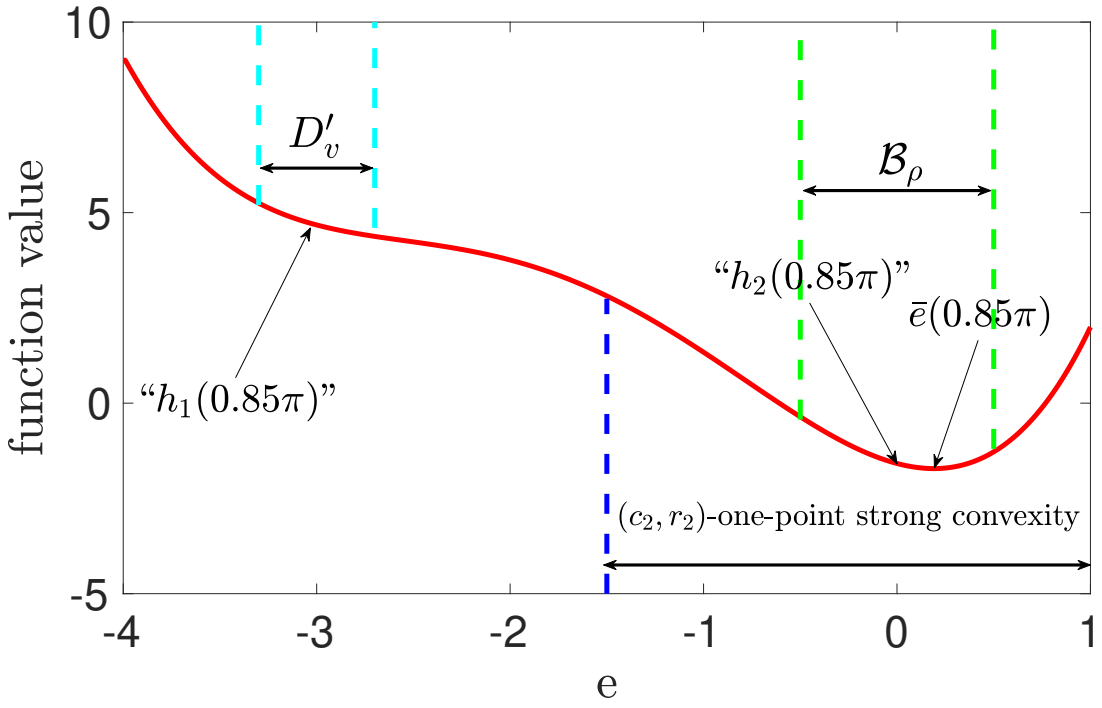


Figure 2.6: Illustration of Definition 9: the region of domination.

Definition 10. *It is said that $h_2(t)$ is a (α, w)-**dominant trajectory** with respect to $h_1(t)$ during the time period $[t_1, t_2]$ over the region D_{v,ρ,r_2} if the time variation of $h_2(t)$ makes the time-varying function $U(e(t), t, \alpha)$ become one-point strongly monotone over D_{v,ρ,r_2}, i.e.,*

$$\begin{aligned} &U(e(t), t, \alpha)^\top \Big(e(t) - \bar{e}(t)\Big) \geq w \left\| e(t) - \bar{e}(t) \right\|^2, \\ &\forall e(t) \in D_{v,\rho,r_2} \cap \mathcal{M}(t), t \in [t_1, t_2], \end{aligned} \tag{2.29}$$

where $w > 0$ is a constant and $\bar{e}(t)$ is defined in (2.28).

Note that $h_2(t)$ being a dominant trajectory with respect to $h_1(t)$ is equivalent to the statement that the inertia of $h_2(t)$ creates a strongly convex landscape over D_{v,ρ,r_2}, as discussed in Section 2.3.

Remark 3. *The intuition behind Definition 10 is that if the time variation in the time-varying optimization could make the landscape after the change of variables become one-point strongly convex with respect to $h_2(t)$ in a neighborhood including both $h_1(t)$ and $h_2(t)$, then the minimum trajectory $h_2(t)$ is dominant (with respect to $h_1(t)$).*

The role of temporal variations of the constraints

From the perspective of the landscape of the Lagrange functional, (2.23b) can be regarded as a time-varying gradient flow system of the Lagrange functional $L\big(e(t)+h_2(t), \bar{\lambda}(e(t)+h_2(t), t, \alpha), t\big)$ t(the partial gradient is taken with respect to the first argument of L) plus a linear time-varying perturbation $\alpha \dot{h}_2^g(t)^\intercal e(t)$. Besides the linear time-varying perturbation $\alpha \dot{h}_2^g(t)^\intercal e(t)$ induced by the inertia of the minimum trajectory similar to the unconstrained case, the constraints' temporal variation $g'(\cdot, t)$ plays the role of shifting the Lagrange multiplier from λ in (2.3) to $\bar{\lambda}$ in (2.15), which results in a nonlinear time-varying perturbation of the landscape of the Lagrange functional.

From the perspective of the perturbed gradient, the constraints' temporal variation $g'(\cdot, t)$ perturbs the projected gradient $\mathcal{P}(\cdot, t)\nabla_x f(\cdot, t)$ in an orthogonal direction $\mathcal{Q}(\cdot, t)g'(\cdot, t)$ to drive the trajectory of (2.23a) towards satisfying the time-varying constraints.

Lemma 5. *At any given time t, the vector $\mathcal{P}(x, t)\nabla_x f(x, t)$ is orthogonal to the vector $\mathcal{Q}(x, t)g'(x, t)$.*

Therefore, in the equality-constrained problem, the time-varying projected gradient flow system after a change of variables in (2.23a) can be regarded as a composition of a time-varying projected term $\mathcal{P}(e+h_2(t), t)\nabla_x f(e+h_2(t), t)$, a time-varying constraint-driven term $\mathcal{Q}(e+h_2(t), t)g'(e+h_2(t), t)$ and an inertia term $\dot{h}_2(t)$ due to the time variation of the local minimum trajectory.

A unified view for unconstrained and equality-constrained problems

By introducing the Lagrange functional in (2.5) and (2.16), we can unify the analysis of how the temporal variation and the proximal regularization help reshape the optimization landscape and potentially make the landscape become one-point strongly convex over a larger region, for both unconstrained and equality constrained problems. This unified view is illustrated in Table 2.1.

2.4 Main results

In this section, we study the jumping, tracking and escaping properties for the time-varying nonconvex optimization.

Jumping

The following theorem shows that the solution of (P-ODE) could jump to the dominant trajectory as long as the time-interval of such domination is large enough.

Theorem 3 (Sufficient conditions for jumping from $h_1(t)$ to $h_2(t)$)**.** *Suppose that the local minimum trajectory $h_2(t)$ is a (α, w)-dominant trajectory with respect to $h_1(t)$ during $[t_1, t_2]$*

over the region D_{v,ρ,r_2}. *Let* $e_1 \in D'_v$ *be the initial point of* (2.23), *and consider* $\bar{e}(t)$ *defined in* (2.28). *Assume that* $U(e,t,\alpha)$ *is non-singular for all* $t \in [t_1,t_2]$ *and* $e \in D_{v,\rho,r_2}$ *and there exists a constant* $\theta \in (0,1)$ *such that*

$$t_2 - t_1 \geq \max\left\{ \frac{\alpha\rho}{(r_2-\rho)\theta w}, \frac{\alpha \ln\left(\frac{\|e_1-\bar{e}(t_1)\|}{r_2-\rho}\right)}{(1-\theta)w} \right\}. \tag{2.30}$$

Then, the solution of (P-ODE) *will* (v,r_2)*-jump from* $h_1(t)$ *to* $h_2(t)$ *over the time interval* $[t_1,t_2]$.

We also offer an approach based on the time-averaged dynamics over a small time interval and name it "small interval averaging"[2]. This technique guarantees that the solution of the time-varying differential equation (or system) will converge to a residual set of the origin of (2.24), provided that: (i) there is a time interval $[t_1,t_2]$ such that the temporal variation makes the averaged objective function during this interval locally one-point strongly convex around $h_2(t)$ not only just over a neighborhood of $h_2(t)$ but also over a neighborhood of $h_1(t)$, (ii) the original time-varying system is not too distant from the time-invariant averaged system, (iii) $[t_1,t_2]$ is large enough to allow the transition of points from a neighborhood of $h_1(t)$ to a neighborhood of $h_2(t)$. Therefore, the time interval $[t_1,t_2]$ and the time-averaged dynamics over this time interval serve as a certificate for jumping from $h_1(t)$ to $h_2(t)$. In what follows, we introduce the notion of averaging a time-varying function over a time interval $[t_1,t_2]$.

Definition 11. *A function* $U_{av}(e,\alpha)$ *is said to be the* ***average function*** *of* $U(e,t,\alpha)$ *over the time interval* $[t_1,t_2]$ *if*

$$U_{av}(e,\alpha) = \frac{1}{t_2-t_1}\int_{t_1}^{t_2} U(e,\tau,\alpha)d\tau$$

The averaged system of (2.23) over the time interval $[t_1,t_2]$ can be written as

$$\dot{e} = -\frac{1}{\alpha}U_{\text{av}}(e,\alpha) \tag{2.31}$$

Then, (2.23) can be regarded as a time-invariant system (2.31) with the time-varying perturbation term $p(e(t),t,\alpha) = -\frac{1}{\alpha}(U(e(t),t,\alpha) - U_{\text{av}}(e(t),\alpha))$. For the averaged system, we can define the on-average region of domination D_{v,ρ,r_2} for $h_2(t)$ with respect to $h_1(t)$ similarly as Definition 9 by replacing (2.28) with

$$\rho \geq \sup_{\bar{e}:\|\bar{e}\|<r_2, 0=U_{\text{av}}(\bar{e},\alpha)} \|\bar{e}\|. \tag{2.32}$$

[2]Our averaging approach distinguishes from classic averaging methods [59, 74, 122, 3] and the partial averaging method [99] in the sense that: (1) it is averaged over a small time interval instead of the entire time horizon, and (2) there is no two-time-scale behavior because there is no parameter in (2.24) that can be taken sufficiently small.

The corresponding on-average (α, w)-dominant trajectory with respect to $h_1(t)$ during $[t_1, t_2]$ over the region D_{v,ρ,r_2} can also be defined similarly as Definition 10 by replacing (2.29) with

$$U_{\text{av}}(e,\alpha)^{\top}(e-\bar{e}) \geq w \left\|e-\bar{e}\right\|^2, \ \forall e \in D_{v,\rho,r_2} \cup \left(\cup_{[t_1,t_2]}\mathcal{M}(t)\right) \tag{2.33}$$

where $\bar{e}$ is defined in (2.32).

Theorem 4 (Sufficient conditions for jumping from $h_1(t)$ to $h_2(t)$ using averaging)**.** *Suppose that the local minimum trajectory $h_2(t)$ is a on-average (α, w)-dominant trajectory with respect to $h_1(t)$ during $[t_1, t_2]$ over the region D_{v,ρ,r_2}. Assume that the following conditions are satisfied:*

1. *There exist some time-varying scalar functions $\delta_1(\alpha, t)$ and $\delta_2(\alpha, t)$ such that*

$$\|p(e(t), t, \alpha)\| \leq \delta_1(\alpha, t) \left\|e-\bar{e}\right\| + \delta_2(\alpha, t), \tag{2.34}$$

   for all $t \in [t_1, t_2]$, and there exist some positive constants $\eta_1(\alpha)$ and $\eta_2(\alpha)$ such that

$$\int_{t_1}^{t} \delta_1(\alpha, \tau) d\tau \leq \eta_1(\alpha)(t-t_1) + \eta_2(\alpha). \tag{2.35}$$

2. *The inequality*

$$\beta_2(\alpha) \left\|e_1-\bar{e}\right\| e^{-\beta_1(\alpha)(t_2-t_1)} + \beta_2(\alpha) \int_{t_1}^{t_2} e^{-\beta_1(\alpha)(t_2-\tau)} \delta_2(\alpha, \tau) d\tau \leq r_2 - \rho, \forall e_1 \in D_v' \tag{2.36}$$

   holds, where $\beta_1(\alpha) = \frac{w}{\alpha} - \eta_1(\alpha) > 0$ and $\beta_2(\alpha) = e^{\eta_2(\alpha)} \geq 1$.

Then, the solution of (P-ODE) *will (v, r_2)-jump from $h_1(t)$ to $h_2(t)$ over the time interval $[t_1, t_2]$.*

Remark 4. *If the global minimum trajectory is the dominant trajectory with respect to the spurious local minimum trajectories, then Theorems 3 and 4 guarantee that the solution of* (P-ODE) *will jump to the neighborhood of the global minimum trajectory.*

Remark 5. *The condition in Theorem 3 and Condition 2 in Theorem 4 mean that $[t_1, t_2]$ needs to be large enough to allow the transition of points from a neighborhood of $h_1(t)$ to a neighborhood of $h_2(t)$. Condition 1 in Theorem 4 means that the original time-varying system should not be too distant from the time-invariant averaged system.*

Remark 6. *To make the one-point strong monotonicity conditions* (2.29) *and* (2.33) *hold, the inertia parameter α cannot be too small.*

Remark 7. *The locally one-point strongly convex parameter w in* (2.29) *and* (2.33) *determines the convergence rate during $[t_1, t_2]$, which is reflected in* (2.30) *and* (2.36)*.*

Remark 8. *In Theorem 4, to ensure that the time-invariant partial interval averaged system is a reasonable approximation of the time-varying system, the time interval $[t_1, t_2]$ should not be very large. On the other hand, to guarantee that the solution of* (2.23) *has enough time to jump, the time interval $[t_1, t_2]$ should not be very small. This trade-off is reflected in* (2.36)*.*

Tracking

In this subsection, we study the tracking property of the local minimum trajectory $h_2(t)$. First, notice that if $h_2(t)$ is not constant, the right-hand side of (P-ODE) is nonzero while the left-hand side is zero. Therefore, $h_2(t)$ is not a solution of (P-ODE) in general. This is because the solution of (P-ODE) approximates the continuous limit of a discrete local trajectory of the sequential regularized optimization problem (2.10). However, to preserve the optimality of the solution with regards to the original time-varying optimization problem without any proximal regularization, it is required to guarantee that the solution of (P-ODE) is close to $h_2(t)$.

If the solution of (2.23) can be shown to be in a small residual set around 0 on the time-varying manifold $\mathcal{M}(t)$, then it is guaranteed that $x(t, t_0, x_0)$ tracks its nearby local minimum trajectory. Notice that (2.23) can be regarded as a time-varying perturbation of the system

$$\dot{e} = -\frac{1}{\alpha}\mathcal{P}(e + h_2(t), t)\nabla_x f(e + h_2(t), t), \quad \forall t \geq t_0 \tag{2.37}$$

Since $h_2(t)$ is a local minimum trajectory, it is obvious that $e(t) \equiv 0$ is an equilibrium point of (2.37). In addition, if the time-varying Lagrange function $L(x, \lambda, t)$ with λ given in (2.3) is locally one-point strongly convex with respect to x around $h_2(t)$ in the time-varying feasible set $\mathcal{M}(t)$, after noticing the fact that the solution of (2.23) will remain in $\mathcal{M}^{h_2}(t)$ if the initial point $e_0 \in \mathcal{M}^{h_2}(t_0)$ from Lemma 3, one would expect that the solution of (2.23) stays in a small residual set of $e = 0$ if the perturbation $\mathcal{Q}(e(t) + h_2(t), t)g'(e(t) + h_2(t), t) + \dot{h}_2(t)$ is relatively small. The perturbation $\mathcal{Q}(e(t) + h_2(t), t)g'(e(t) + h_2(t), t) + \dot{h}_2(t)$ being small is equivalent to α being small. The next theorem shows that every local minimum trajectory can be tracked for a relatively small α.

Theorem 5 (Sufficient condition for tracking)**.** *Assume that the time-varying Lagrange function $L(x, \lambda, t)$ with λ given in* (2.3) *is locally (c_2, r_2)-one-point strongly convex with respect to x around $h_2(t)$. Given $\gamma(t)$ such that $\|\dot{h}_2(t)\| \leq \gamma(t)$, suppose that there exist time-varying scalar functions $\delta_1(t)$ and $\delta_2(t)$ such that the perturbed gradient due to the time-variation of constraints satisfies the inequality*

$$\|\mathcal{Q}(e(t) + h_2(t), t)g'(e(t) + h_2(t), t)\| \leq \delta_1(t)\|e\| + \delta_2(t), \tag{2.38}$$

and there exist some positive constants η_1 and η_2 such that

$$\int_{t_1}^{t} \delta_1(\tau)d\tau \leq \eta_1(t - t_1) + \eta_2. \tag{2.39}$$

If $\sup_{t \geq t_1}(\delta_2(t) + \gamma(t))$ is bounded and the following conditions hold

$$\|x_0 - h_2(0)\| \leq \frac{r_2}{e^{\eta_2}}, \tag{2.40a}$$

$$\alpha \leq \frac{c_2 r_2}{e^{\eta_2}\sup_{t \geq t_1}(\delta_2(t) + \gamma(t)) + \eta_1 r_2}, \tag{2.40b}$$

then the solution $x(t, t_0, x_0)$ *will* r_2*-track* $h_2(t)$*. More specifically, we have*

$$\|x(t,t_0,x_0) - h_2(t)\| \leq e^{\eta_2} \|e_1\| e^{-(\frac{c_2}{\alpha}-\eta_1)(t-t_1)} + e^{\eta_2} \int_{t_1}^{t} e^{-(\frac{c_2}{\alpha}-\eta_1)(t-\tau)} (\delta_2(t) + \gamma(t)) d\tau \leq r_2. \tag{2.41}$$

Remark 9. *The inequality* (2.41) *implies that the smaller the regularization parameter* α *is, the smaller the tracking error* $x(t, t_0, x_0) - h_2(t)$ *is and the faster* $x(t, t_0, x_0)$ *converges to the neighbourhood of* $h_2(t)$*.*

Remark 10. *In the case that the local minimum trajectory* $h_2(t)$ *is a constant, the upper bound on* α *simply becomes* $\alpha < \infty$*. This implies that if* $h_2(t)$ *is constant, then it will be perfectly tracked with any regularization parameter and can not be escaped by tuning the regularization parameter.*

Remark 11. *In the unconstrained case or the case with the time-invariant constraints,* $\delta_1(t)$ *and* $\delta_2(t)$ *in* (2.38) *simply become zero. Then, the tracking conditions in* (2.40) *become* $\|x_0 - h_2(0)\| \leq r_2$ *and* $\alpha \leq \frac{c_2 r_2}{\sup_{t \geq t_0} \gamma(t)}$*, and the tracking error bound in* (2.41) *becomes*

$$\|e(t)\| \leq \|e_1\| e^{-\frac{c_2}{\alpha}(t-t_1)} + \int_{t_1}^{t} e^{-\frac{c_2}{\alpha}(t-\tau)} \gamma(t) d\tau \leq \frac{\alpha \sup_{t \geq t_1} \gamma(t)}{c_2}$$

Remark 12. *After the solution of* (P-ODE) *has escaped the spurious local trajectories and started tracking the globally minimum trajectory, one may use the state-of-the-art tracking methods in [121] and [130] to improve the tracking of the globally minimum trajectory.*

Escaping

Combining the results of jumping and tracking immediately yields a sufficient condition on escaping from one local minimum trajectory to a more desirable local (or global) minimum trajectory. The proof is omitted for brevity.

Theorem 6 (Sufficient conditions for escaping from $h_1(t)$ to $h_2(t)$)**.** *Given two local minimum trajectories* $h_1(t)$ *and* $h_2(t)$*, suppose that the Lagrange function* $L(x, \lambda, t)$ *with* λ *given in* (2.3) *is locally* (c_2, r_2)*-one-point strongly convex with respect to* x *around* $h_2(t)$ *in the time-varying feasible set* $\{e \in \mathbb{R}^n : e + h_2(t) \in \mathcal{M}(t), \|e\| \leq r_2\}$ *and let* $\mathcal{B}_v(h_1(t_1)) \subseteq RA^{\mathcal{M}(t_1)}(h_1(t_1))$*. Under the conditions of Theorem 3 or 4, if* (2.38)-(2.40) *hold, then the solution of* (P-ODE) *will* (v, r_2)*-escape from* $h_1(t)$ *to* $h_2(t)$ *after* $t \geq t_2$*.*

Discussions

Adaptive inertia: To leverage the potential of the time-varying perturbation $\alpha \mathcal{Q}(e(t) + h_2(t), t) g'(e(t) + h_2(t), t) + \alpha \dot{h}_2(t)$ in re-shaping the landscape of the Langrange function or the objective function to become locally one-point strongly convex in x over a large region,

the regularization parameter α should be selected relatively large. On the other hand, to ensure that the solution of (2.23) and (2.25) will end up tracking a desirable local (or global) minimum trajectory, Theorem 5 prescribes small values for α. In practice, especially when the time-varying objective function has many spurious shallow minimum trajectories, this suggests using a relatively large regularization parameter α at the beginning of the time horizon to escape spurious shallow minimum trajectories and then switching to a relative small regularization parameter α for reducing the ultimate tracking error bound.

Sequential jumping: When the time-varying optimization problem has many local minimum trajectories, the solution of (P-ODE) or (ODE) may sequentially jump from one local minimum trajectory to a better local minimum trajectory. To illustrate this concept, consider the local minimum trajectories $h_1(t), h_2(t), ..., h_m(t)$, where $h_m(t)$ is a global trajectory. Assume that there exists a sequence of time intervals $[t_1^i, t_2^i]$ for $i = 1, 2, \ldots, m-1$ such that the conditions of Theorem 3 or 4 are satisfied for $h_i(t)$ and $h_{i+1}(t)$ during each time interval. Then, by sequentially deploying Theorem 3 or 4, it can be concluded that the solution of (P-ODE) or (ODE) will jump from $h_1(t)$ to $h_m(t)$ after $t \geq t_2^m$. Furthermore, if $h_m(t)$ can be tracked with the given α, the solution of (P-ODE) or (ODE) will escape from $h_1(t)$ to $h_m(t)$ after $t \geq t_2^m$.

2.5 Numerical Examples

Example 3. Consider the non-convex function

$$\bar{f}(x) = 0.5e + 20e^{-d} - 20e^{-\sqrt{0.5(x_1^2+x_2^2)+d^2}} - 0.5e^{(0.5(\cos(2\pi x_1)+\cos(2\pi x_2)))}.$$

This function has a global minimum at $(0, 0)$ with the optimal value 0 and many spurious local minima. Its landscape is shown in Figure 2.7. When $d = 0$, this function is called the Ackley function [2], which is a benchmark function for global optimization algorithms. To make this function twice continuously differentiable, we choose $d = 0.01$.

Consider the time-varying objective function $f(x, t) = \bar{f}(x - z(t))$ and the time-varying constraint $g(x, t) = (x_1 - z_1(t)) - 1/2(x_2 - z_2(t))^2 = 0$, where $z(t) = [24\sin(t), \cos(t)]^\top$. This constrained time-varying optimization problem has the global minimum trajectory $[0, 0]^\top + z(t)$ and many spurious local minimum trajectories. Two local minimum trajectories are $h_1(t) = [1.92, 1.96]^\top + z(t)$ and $h_2(t) = [0, 0]^\top + z(t)$. It can be shown that $L(x, \lambda, t)$ is locally $(20, 0.5)$-one-point strongly convex with respect to $h_2(t)$.

We take $D_{v,\rho,r_2} = D_{0.04,0.01,1} = [-0.1, 2] \times [-0.1, 2]$ in Definition 10. The condition in (2.26) can be verified by checking the signs of the derivatives of $e_1(t)$ and $e_2(t)$ along the dynamics (2.23) on the boundary points of $D_{0.04,0.01,1} \cap \mathcal{M}^{h_2}(t)$. Furthermore, (2.33) is satisfied for $w = 1$. Thus, $h_2(t)$ is a $(0.2, 1)$-dominant trajectory with respect to $h_1(t)$ during $[0, \frac{\pi}{8}]$ over the region $D_{0.04,0.01,1}$.

Regarding Theorem 3, if we select $\theta = 0.2$, the inequality (2.36) is satisfied for $\alpha = 0.2$ and $t_2 - t_1 = \pi/8$. Thus, the solution of (P-ODE) will $(0.04, 0.5)$-jump from $h_1(t)$ to $h_2(t)$.

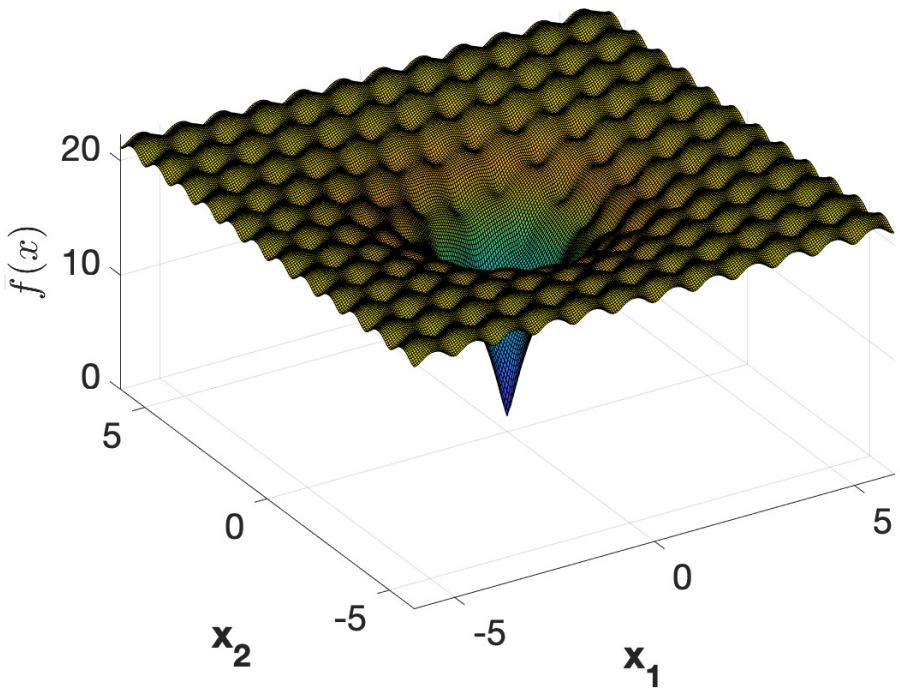


Figure 2.7: Illustration of Example 3.

Regarding Theorem 5, δ_1 and δ_2 in the inequality (2.38) can be taken as 0 and $24\sqrt{2}\cos(t) + \sqrt{2}\sin(t)$, respectively. Then the inequality (2.40b) reduces to $\alpha \leq \frac{10}{\sqrt{2(24^2+1)}} \approx 0.29$, which is satisfied by $\alpha = 0.2$. Thus, the solution of (P-ODE) will 0.5-track $h_2(t)$. Putting the above findings together, we can conclude that the solution of (2.23) will $(0.04, 0.5)$-escape from $h_1(t)$ to $h_2(t)$.

In addition, by choosing the inertia parameter $\alpha = 0.2$, the simulation shows that for 1000 runs of random initialization with $x_2(0) - z(0) \in [-5, 5]$ and $x_1(0)$ determined by the equality constraint, all solutions of the corresponding (P-ODE) will sequentially jump over the local minimum trajectories and end up tracking the global trajectory after $t \geq 5\pi$.

Appendix

2.A Omitted proofs of Section 2.2

Proof of Lemma 1

Proof. Under Assumptions 1-4, one can apply the inverse function theorem to (2.2) (see [113, Theorem 4.4, Example 4.7]) to conclude that for every $h(\bar{t})$ and $\bar{t}$, there exist an open set $\mathcal{S}_{h(\bar{t})}$ containing $h(\bar{t})$ and an open set $\mathcal{S}_{\bar{t}}$ containing $\bar{t}$ such that there exist a unique differentiable function $x(t)$ in $\mathcal{S}_{h(\bar{t})}$ for all $t \in \mathcal{S}_{\bar{t}}$ where $x(t)$ is the isolated local minimizer of the time-varying optimization problem (2.1). Because of this uniqueness property and the continuity of the local minimum trajectory $h(t)$, $x(t)$ must coincide with $h(t)$ for all $t \in \mathcal{S}_{\bar{t}}$. Then, because the above property holds uniformly for every $t \in [0, \infty)$, $h(t)$ must be a differentiable isolated minimum trajectory. □

Proof of Lemma 2

Proof. Due to the second-order sufficient conditions for the equality constrained minimization problem, $\nabla^2_{xx} L(h(t), \lambda(h(t), t), t)$ is positive definite on $T^t_{h(t)}$ for all $t \in [0, \infty)$, meaning that for every nonzero vector $y \in T^t_{h(t)}$, there exists a positive constant $\bar{c}$ such that $y \nabla^2_{xx} L(h(t), \lambda, t) y > \bar{c} \|y\|^2$. Since $\mathcal{P}(h(t), t)$ is the orthogonal projection matrix onto the tangent plane $T^t_{h(t)}$, we have $y \nabla^2_{xx} L(h(t), \lambda(h(t), t), t) \mathcal{P}(h(t), t) y > \bar{c} \|y\|^2$ for all $y \in T^t_{h(t)}$ and $y \neq 0$, and $y \nabla^2_{xx} L(h(t), \lambda(h(t), t), t)\ \mathcal{P}(h(t), t) y = 0$ for all $y \notin T^t_{h(t)}$. Taking the first-order Taylor expansion of $\nabla_x L(x, \lambda(x, t), t)$ with respect to x around $h(t)$ and using the following result from [84, Corollary 1]:

$$\frac{\partial}{\partial x} \nabla_x L(x, \lambda(x, t), t)\big|_{x=h(t)} = \nabla^2_{xx} L(h(t), \lambda(h(t), t), t) \mathcal{P}(h(t), t),$$

it yields that

$$\begin{aligned} e(t)^\top \nabla_x L(e(t) + h(t), \lambda, t) &= e(t)^\top \nabla_x L(h(t), \lambda, t) + e(t)^\top \nabla^2_{xx} L(h(t), \lambda, t) \mathcal{P}(h(t), t) e(t) + o(e(t)^3) \\ &= e(t)^\top \nabla^2_{xx} L(h(t), \lambda, t) \mathcal{P}(h(t), t) e(t) + o(e(t)^3) \end{aligned}$$

From Lemma 6 in the online report [36], we know that $\nabla^2_{xx} L(x, \lambda, t) \mathcal{P}(x, t)$ is continuous in x and t. In addition, $g(x, t)$ is also continuous in x and t. As a result, there exist positive

constants $\hat{r}$ and $\hat{c}$ such that

$$e(t)^\top \nabla_x L(e(t) + h(t), \lambda, t) \geq \hat{c} \left\| e(t) \right\|^2$$

for all $e(t) \in \{e + h(t) \in \mathcal{M}(t) : \|e\| \leq \hat{r}\}$ □

Proof of Lemma 3

Proof. On examining the evolution of $g(x(t), t)$ along the flow of the system (P-ODE), we obtain

$$\dot{g}(x(t), t) = \mathcal{J}_g(x(t), t)\dot{x}(t) + g'(x(t), t) = 0$$

Hence, $g(x(t_0), t_0) = g(x(t, t_0, x_0), t)$ for all $t \geq t_0$. □

Proof of Theorem 1

Proof. Since $h(t)$ is differentiable by Lemma 1, we can use the change of variables $e(t) = x(t) - h(t)$ to rewrite (P-ODE) as:

$$\dot{e}(t) = -\frac{1}{\alpha}\mathcal{P}(e(t) + h(t), t)\nabla_x f(e(t) + h(t), t) - \mathcal{Q}(e(t) + h(t), t)g'(e(t) + h(t), t) - \dot{h}(t) \tag{2.42}$$

In light of the conditions in Theorem 1, the solution of (2.42) stays in a compact set. Then, by Lemma 3 and [74, Theorem 3.3], the equation (2.42) has a unique solution. Thus, (P-ODE) must also have a unique solution. □

Proof of Theorem 2

Proof. The first part follows from Theorem 2 in [42]. For the second part, a direct application of the classical results on convergence of the forward Euler method [66] immediately shows that the solution of (P-ODE) starting at a local minimum of (2.10a) is the continuous limit of the discrete local trajectory of the sequential regularized optimization (2.10). □

2.B Omitted proofs of Section 2.3

Proof of Lemma 4

Proof. Let $b(t_0)$ be the unit vector $-\frac{\dot{h}_1(t_0)}{\|\dot{h}_1(t_0)\|}$. One can write

$$-\dot{h}_1(t)^\top b(t_0) \geq -\dot{h}_1(t_0)^\top b(t_0) - L|t - t_0| \geq \epsilon - L\delta := \epsilon'$$

For any $t \in [t_0, t_0 + \delta]$ and $e(t) \in RA^{\mathcal{M}(t)}(h_1(t))$, we have

$$\begin{aligned}(\dot{x}(t) - \dot{h}_1(t))^\top b(t_0) =& -\frac{1}{\alpha}\nabla_x L(x, \bar{\lambda}, t)^\top b(t_0) - \dot{h}_1(t)^\top b(t_0) \\ \geq& \epsilon' - \left\| \frac{1}{\alpha}\nabla_x L(x, \bar{\lambda}, t) \right\| \geq \epsilon' - E\end{aligned}$$

Hence,

$$\begin{aligned}r \geq& \|x(t_0 + \delta) - h_1(t_0 + \delta)\| \\ \geq& (x(t_0 + \delta) - h_1(t_0 + \delta))^\top b(t_0) \\ \geq& (x(t_0) - h_1(t_0))^\top b(t_0) + \int_{t_0}^{t_0+\delta} (\epsilon' - E) dt \\ \geq& -r + (\epsilon' - E)\delta\end{aligned}$$

The above contradiction completes the proof. □

Proof of Lemma 5

Proof. Recall that $\mathcal{P}(x,t)$ is the orthogonal projection matrix on the tangent plane of $g(x(t),t)$ at the point $x(t)$ after the freezing time t. Thus, we have $\mathcal{P}(x,t)\nabla_x f(x,t) \in T_x^t$. For the vector $\mathcal{Q}(x,t)g'(x,t)$, it can be shown that

$$\mathcal{P}(x,t)\mathcal{Q}(x,t)g'(x,t) = 0$$

This implies that the orthogonal projection of the vector $\mathcal{Q}(x,t)g'(x,t)$ onto the tangent plane T_x^t is 0. Thus, $\mathcal{Q}(x,t)g'(x,t)$ must be orthogonal to T_x^t. □

2.C Omitted proofs of Section 2.4

Proof of Theorem 3

Proof. First, notice that if $U(e,t,\alpha)$ is uniformly non-singular for all $t \in [t_1, t_2]$ and $e \in D_{v,\rho,r_2}$, then $\bar{e}(t)$ defined in (2.28) is continuously differentiable for $t \in [t_1, t_2]$. Then, notice that every solution of (2.23) with an initial point in $D_{v,\rho,r_2} \cap \mathcal{M}(t_1)$ will remain in D_{v,ρ,r_2}. It follows from Theorem 1 that (2.23) has a unique solution defined for all $t \in [t_1, t_2]$ whenever $e_1 \in D_{v,\rho,r_2} \cap \mathcal{M}(t_1)$.

We take $V(e(t),t) = \frac{1}{2}\|e(t) - \bar{e}(t)\|^2$ as the Lyapunov function for the system (2.23). Because of Lemma 3, any solution of (2.23) stating in $\mathcal{M}(t_1)$ will remain in $\mathcal{M}(t)$ for all $t \geq t_1$. Therefore, the derivative of $V(e(t),t)$ along the trajectories of (2.23) in $\mathcal{M}(t)$ can be

expressed as

$$\begin{aligned}\dot{V} =&(e(t)-\bar{e}(t))^\top\Big(-\frac{1}{\alpha}U(e(t),t,\alpha)\Big)-(e(t)-\bar{e}(t))^\top\dot{\bar{e}}(t),\ \forall e(t)\in D_{v,\rho,r_2}\cap\mathcal{M}^{h_2}(t)\\ \leq&-\frac{w}{\alpha}\left\|e(t)-\bar{e}(t)\right\|^2+\left\|\dot{\bar{e}}(t)\right\|\left\|e(t)-\bar{e}(t)\right\|,\ \forall e(t)\in D_{v,\rho,r_2}\cap\mathcal{M}^{h_2}(t)\\ \leq&-(1-\theta)\frac{w}{\alpha}\left\|e(t)-\bar{e}(t)\right\|^2-\theta\frac{w}{\alpha}\left\|e(t)-\bar{e}(t)\right\|^2+\delta\left\|e(t)-\bar{e}(t)\right\|,\ \forall e(t)\in D_{v,\rho,r_2}\cap\mathcal{M}^{h_2}(t)\\ \leq&-(1-\theta)\frac{w}{\alpha}\left\|e(t)-\bar{e}(t)\right\|^2,\ \forall e(t)\in\{e(t)\in D_{v,\rho,r_2}\cap\mathcal{M}^{h_2}(t):\left\|e(t)-\bar{e}(t)\right\|\geq\frac{\alpha\delta}{\theta w}\}\end{aligned}\tag{2.43}$$

where $\delta\coloneqq\sup_{t\in[t_1,t_2]}\left\|\dot{\bar{e}}(t)\right\|$. By taking $e_1\in D'_v\cap\mathcal{M}(t_1)$, since D_{v,ρ,r_2} satisfies the condition (2.26), the solution of (2.23) starting from e_1 will stay in D_{v,ρ,r_2}. Thus, the bound in (2.43) is valid. To ensure that the trajectory of (2.23) enters the time-varying set $\mathcal{B}_{r_2-\rho}(\bar{e}(t))$, it is sufficient to have $\frac{\alpha\delta}{\theta w}\leq r_2-\rho$ or $\alpha\leq\frac{(r_2-\rho)\theta w}{\delta}$. Since $\delta=\sup_{t\in[t_1,t_2]}\left\|\dot{\bar{e}}(t)\right\|\geq\frac{\rho}{t_2-t_1}$. We can further bound α as $\alpha\leq\frac{(r_2-\rho)\theta w(t_2-t_1)}{\rho}$ which is equivalent to $t_2-t_1\geq\frac{\alpha\rho}{(r_2-\rho)\theta w}$.

Now, it is desirable to show that if the time interval $[t_1,t_2]$ is large enough, the solution of (2.23a) will enter the time-varying set $\mathcal{B}_{r_2-\rho}(\bar{e}(t))$ with an exponential convergence rate. Since $\dot{V}(\cdot,\cdot)$ is negative in $\Gamma(t)\coloneqq\{e\in D_{v,\rho,r_2}\cap\mathcal{M}^{h_2}(t):\left\|e-\bar{e}(t)\right\|\geq\frac{\alpha\delta}{\theta w}\}$ and because of (2.26), a trajectory starting from $\Gamma(t_1)$ must stay in D_{v,ρ,r_2} and move in a direction of decreasing $V(e,t)$. The function $V(e,t)$ will continue decreasing until the trajectory enters the set $\{e\in D_{v,\rho,r_2}\cap\mathcal{M}^{h_2}(t):\left\|e-\bar{e}(t)\right\|\leq\frac{\alpha\delta}{\theta w}\}$ or until time t_2. Let us show that the trajectory enters $\mathcal{B}_{r_2-\rho}(\bar{e}(t))$ before t_2 if $t_2-t_1>\frac{\alpha}{w(1-\theta)}\ln(\frac{\left\|e_1-\bar{e}(t_1)\right\|}{r_2-\rho})$. Since $V(e(t),t)=\frac{1}{2}\left\|e(t)-\bar{e}(t)\right\|^2$, (2.43) can be written as

$$\dot{V}(e(t),t)\leq-(1-\theta)\frac{2w}{\alpha}V(e(t),t),\ \forall e\in\Big\{e\in D_{v,\rho,r_2}\cap\mathcal{M}^{h_2}(t):\left\|e(t)-\bar{e}(t)\right\|\geq\frac{\alpha\delta}{\theta w}\}\Big\},$$

By the comparison lemma[74, Lemma 3.4],

$$V(e(t),t)\leq\exp\{-(1-\theta)\frac{2w}{\alpha}(t-t_1)\}V(e_1,t_1)$$

Hence,

$$\left\|e(t)-\bar{e}(t)\right\|\leq\exp\{-(1-\theta)\frac{w}{\alpha}(t-t_1)\}\left\|e_1-\bar{e}(t_1)\right\|.$$

The inequality $\left\|e(t_2)-\bar{e}(t_2)\right\|\leq r_2-\rho$ holds if $t_2-t_1\geq\frac{\alpha}{w(1-\theta)}\ln(\frac{\left\|e_1-\bar{e}(t_1)\right\|}{r_2-\rho})$. □

Proof of Theorem 4

Proof. Due to the space restriction, we move the proof to the online report [36]. □

Proof of Theorem 5

Proof. Consider $V(e) = \frac{1}{2}\|e\|^2 : \mathcal{B}_{r_2}(0) \to \mathbb{R}$ as the Lyapunov function for the system (2.23). Because of Lemma 3, any solution of (2.23) stating in $\mathcal{M}(t_1)$ will remain in $\mathcal{M}(t)$ for all $t \geq t_1$. The derivative of $V(e)$ along the trajectories of (2.23) can be obtained as

$$\begin{aligned}\dot{V} &= e(t)^\top \Big(-\frac{1}{\alpha}\mathcal{P}(e(t)+h_2(t),t)\nabla_x f(e(t)+h_2(t),t) \\ &\quad - \mathcal{Q}(e(t)+h_2(t),t)g'(t)(e(t)+h_2(t),t) - \dot{h}_2^g(t)\Big), \\ &\leq -\frac{c}{\alpha}\|e(t)\|^2 + \delta_1(t)\|e(t)\|^2 + (\delta_2(t)+\gamma(t))\|e(t)\|\end{aligned}$$

Since $V(e) = \frac{1}{2}\|e\|^2$, one can derive an upper bound on $\dot{V}$ as

$$\dot{V} \leq -\Big[\frac{2c}{\alpha} - 2\delta_1(t)\Big]V + (\delta_2(t)+\gamma(t))\sqrt{2V}$$

Using the same proof procedure as in Theorem 4 of the online report [36] and by taking $\beta_1(\alpha) = \frac{c}{\alpha} - \eta_1 > 0$ and $\beta_2 = e^{\eta_2} \geq 1$, it can be shown that

$$\|e(t)\| \leq \beta_2\|e_1\| e^{-\beta_1(\alpha)(t-t_1)} + \beta_2 \int_{t_1}^{t} e^{-\beta_1(\alpha)(t-\tau)}(\delta_2(t)+\gamma(t))d\tau \tag{2.44}$$

To make the bound in (2.44) valid, we must ensure that $e(t) \in \mathcal{B}_{r_2}(0)$ for all $t \geq t_1$. Note that

$$\begin{aligned}\|e(t)\| &\leq \beta_2\|e_1\| e^{-\beta_1(\alpha)(t-t_1)} + \frac{\beta_2}{\beta_1(\alpha)}(1 - e^{-\beta_1(\alpha)(t-\tau)}) \sup_{t\geq t_0}(\delta_2(t)+\gamma(t)) \\ &\leq \max\Big\{\beta_2\|e_1\|, \frac{\beta_2}{\beta_1(\alpha)}\sup_{t\geq t_0}(\delta_2(t)+\gamma(t))\Big\}\end{aligned}$$

It can be verified that the condition $e(t) \in \mathcal{B}_{r_2}(0)$ will be satisfied if (2.40) holds. Furthermore, by $e(t) \in \mathcal{B}_{r_2}(0)$ and Theorem 1, there must exist a unique solution for (P-ODE) for all $t \geq t_1$. □

Chapter 3

Non-Stationary Constrained MDPs

Safe reinforcement learning (RL) studies how an agent learns to maximize its expected total reward by interacting with an unknown environment over time while dealing with restrictions/constraints arising from real-world problems [6, 40, 52]. A standard approach for modeling the safe RL is based on Constrained Markov Decision Processes (CMDPs) [5], where one seeks to maximize the expected total reward under a safety-related constraint on the expected total utility.

While classical safe RL and CMDPs assume that an agent interacts with a time-invariant (stationary) environment, both the reward/utility functions and transition kernels can be time-varying for many real-world safety-critical applications. For example, in autonomous driving [103] or power grid control [37], it is essential to guarantee safety, such as collision-avoidance and contingency, while handling time-varying conditions related to traffic and load demands. Similarly, in most safety-critical human-computer interaction applications, e.g., automated medical care, human behavior changes over time. In such scenarios, if the automated system is not adapted to take such changes into account, then the system could quickly violate the safety constraint and incur a severe loss [26, 92]. Despite the importance of non-stationary safe RL problems, the literature lacks provably efficient algorithms and theoretical results.

In this work, we formulate a general non-stationary safe exploration problem as an episodic CMDP in which the transition model is unknown and non-stationary, the reward/utility feedback after each episode is bandit and non-stationary, and the variation budget is known. The goal is to design an algorithm that can perform a non-stationary safe exploration, that is, to adaptively explore the unknown and time-varying environment and learn to satisfy time-varying constraints in the long run.

The safe exploration in non-stationary CMDPs is more challenging since the utilities and dynamics are time-varying and unknown a priori. Thus, it is difficult/impossible to guarantee a small/zero constraint violation without knowing how CMDPs will change. Previous constraint violation analyses [34, 83] strongly rely on the conditions of having the same transition dynamics and rewards over all episodes, which are not applicable to non-stationary CMDPs. In view of the aforementioned challenges, we propose a new primal-dual method

and develop novel techniques to decouple the optimality gap and the constraint violation.

3.1 Related Work

Non-stationary RL. Non-stationary RL has been mostly studied in the unconstrained setting [67, 7, 95, 38, 86, 134, 123, 46, 132, 27, 124]. Our work is related to policy-based methods for non-stationary RL since the optimal solution of CMDP is usually a stochastic policy [5] and thus a policy-based method is preferred. When the variation budget is known a prior, [46] propose the first policy-based method for non-stationary RL, but they assume stationary transitions and adversarial full-information rewards in the tabular setting. [132] extends the above results to a more general setting where both the transitions and rewards can vary over episodes. To eliminate the assumption of having prior knowledge on variation budgets, [124] recently outline that an adaptive restart approach can be used to convert any upper-confidence-bound-type stationary RL algorithm to a dynamic-regret-minimizing algorithm. Beyond the non-stationary unconstrained RL, [100] consider the online CMDPs where the reward is adversarial but the transition model is fixed and the constraints are stochastic over episodes. In summary, the above papers only consider the non-stationarity in the objective and may not work for the more general safe RL problems where there is also time-varying constraints.

CMDP. The study of RL algorithms for CMDPs has received considerable attention due to the safety requirement [5, 98, 129, 40, 52, 56, 55]. Our work is closely related to Lagrangian-based CMDP algorithms with optimistic policy evaluations [41, 112, 34, 83, 100]. In particular, [41, 112] leverage upper confidence bound (UCB) bonus on fixed reward/utility and transition probability to propose sample efficient algorithms for tabular CMDPs. [34] generalize the above results to the linear kernel CMDPs. Under some mild conditions and additional computation cost, [83] propose two algorithms to learn policies with a zero or bounded constraint violation for CMDPs. Beyond the stationary CMDP, [100] consider the online CMDPs where only the rewards in objective can vary over episodes. In contrast, our work focuses on a more general and realistic safe RL setting where the dynamics and rewards/utilities can all change over episodes, and thus we significantly extend the existing results.

3.2 Problem formulation

Model. In this chapter, we study safe RL in non-stationary environments via episodic CMDPs with adversarial bandit-information reward/utility feedback and unknown adversarial transition kernels. At each episode m, a CMDP is defined by the state space $\mathcal{S}$, the action space $\mathcal{A}$, the fixed length of each episode H, a collection of transition probability measure $\{\mathbb{P}_h^m\}_{h=1}^H$, a collection of reward functions $\{r_h^m\}_{h=1}^H$, a collection of utility functions $\{g_h^m\}_{h=1}^H$ and the constraint offset b_m. We assume that $\mathcal{S}$ is a measurable space with a possibly infinite

number of elements, and that $\mathcal{A}$ is a finite set. In addition, we assume $r_h^m : \mathcal{S} \times \mathcal{A} \to [0,1]$ and $g_h^m : \mathcal{S} \times \mathcal{A} \to [0,1]$ are deterministic reward and utility functions. Our analysis readily generalizes to the setting where the reward/utility functions are random. In this chapter, we focus on a bandit setting where the agent only observes the values of reward and utility functions, $r_h^m(x_h^m, a_h^m)$ and $g_h^m(x_h^m, a_h^m)$ at the visited state-action pair (x_h^m, a_h^m). To avoid triviality, we take $b_m \in (0, H]$ and assume that it is known to the agent.

Let the policy space $\Delta(\mathcal{A}|\mathcal{S}, H)$ be $\{\{\pi_h(\cdot|\cdot)\}_{h=1}^H : \pi_h(\cdot|s) \in \Delta(\mathcal{A}), \forall x \in \mathcal{S}, h \in [H]\}$, where $\Delta(\mathcal{A})$ denotes a probability simplex over the action space. Let $\pi^m \in \Delta(\mathcal{A}|\mathcal{S}, H)$ be a policy taken by the agent at episode m, where $\pi_h^m(\cdot|x_h^m) : \mathcal{S} \to \mathcal{A}$ is the action that the agent takes at state x_h^m. For simplicity, we assume the initial state x_1^m to be fixed as x_1 in different episodes. The episode terminates at state x_H^m in which no control action is needed and both reward and utility functions are equal to zero.

Given a policy $\pi \in \Delta(\mathcal{A}|\mathcal{S}, H)$ and the episode m, the value function $V_{r,h}^{\pi,m}$ associated with the reward function r at step h in episode m is the expected value of the total reward, $V_{r,h}^{\pi,m}(x) = \mathbb{E}_{\pi,\mathbb{P}^m}\left[\sum_{i=h}^H r_i^m(x_i, a_i) | x_h = x\right]$, for all $x \in \mathcal{S}$ and $h \in [H]$, where the expectation $\mathbb{E}_{\pi,\mathbb{P}^m}$ is taken over the random state-action sequence $\{(x_i^m, a_i^m)\}_{i=h}^H$, the action a_h^m follows the policy $\pi_h^m(\cdot|x_h^m)$, and the next state x_{h+1} follows the transition dynamics $\mathbb{P}_h^m(\cdot|x_h^m, a_h^m)$.

The action-value function is defined as $Q_{r,h}^{\pi,m}(x,a) = \mathbb{E}_{\pi,\mathbb{P}^m}\left[\sum_{i=h}^H r_i^m(x_i^m, a_i^m) | x_h^m = x, a_h^m = a\right]$, for all $x \in \mathcal{S}, a \in \mathcal{A}$ and $h \in [H]$. Similarly, we define the value function $V_{g,h}^{\pi,m} : \mathcal{S} \to \mathbb{R}$ and the action-value function $Q_{g,h}^{\pi,m} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ associated with the utility function g. For brevity, we use the symbol $\diamond$ to denote r or g. and take the shorthand $\mathbb{P}_h^m V_{\diamond,h}^{\pi,m}(x,a) \coloneqq \mathbb{E}_{x' \sim \mathbb{P}_h^m(\cdot|x,a)}\left[V_{\diamond,h+1}^{\pi,m}(x')\right]$. The Bellman equation associated with a policy π is given by

$$Q_{\diamond,h}^{\pi,m}(x,a) = (\diamond_h^m + \mathbb{P}_h^m V_{\diamond,h+1}^{\pi,m})(x,a), \tag{3.1a}$$

$$V_{\diamond,h}^{\pi,m}(x) = \left\langle Q_{\diamond,h}^{\pi,m}(x,\cdot), \pi_h(\cdot|x)\right\rangle_{\mathcal{A}}, \tag{3.1b}$$

for all $(x,a) \in \mathcal{S} \times \mathcal{A}$, where $\langle\cdot,\cdot\rangle_{\mathcal{A}}$ denotes the inner product over $\mathcal{A}$ and we will omit the subscript $\mathcal{A}$ in the sequel when it is clear from the context.

Constrained MDP. In constrained MDPs, the agent aims to approximate the optimal non-stationary policy by interacting with the environment. In each episode m, the agent aims to maximize the expected total reward while satisfying the constraints on the expected total utility

$$\max_{\pi \in \Delta(\mathcal{A}|\mathcal{S},H)} V_{r,1}^{\pi,m} \text{ subject to } V_{g,1}^{\pi,m} \geq b_m \tag{3.2}$$

for all $m = 1, 2, \ldots$, where the reward/utility functions and the transition kernels are potentially different across the episodes. The associated Lagrangian of problem (3.2) is given by

$$\mathcal{L}^m(\pi, \mu) \coloneqq V_{r,1}^{\pi,m} + \mu\left(V_{g,1}^{\pi,m} - b_m\right) \tag{3.3}$$

where the policy π is the primal variable and $\mu \geq 0$ is the dual variable. We can reformulate the constrained optimization problem (3.2) as the saddle-point problem

$$\max_{\pi \in \Delta(\mathcal{A}|\mathcal{S},H)} \min_{\mu \geq 0} \mathcal{L}^m(\pi, \mu).$$

Let $\mathcal{D}^m(Y) \coloneqq \text{maximize } {}_\pi\mathcal{L}^m(\pi,\mu)$ be the dual function, $\mu^{\star,m} \coloneqq \operatorname{argmin}_{\mu\geq 0}\mathcal{D}^m(\mu)$ be an optimal dual variable and $\pi^{\star,m}$ be a globally optimal solution of (3.2) at episode m. Unlike the unconstrained MDP, the optimal solution of CMDP is usually a stochastic policy and the best deterministic policy can lose as much as the difference between the respective values of the best and the worst policies [5]. As a consequence, RL methods that implicitly rely on the existence of a deterministic optimal policy (e.g., Q learning) may not be suitable for this type of problem. This further inspires the study of randomized policies and take on a policy gradient approach for non-stationary CMDP.

Performance metrics. Suppose that the agent executes policy π^m in episode m. We now define the dynamic regret and the constraint violation in the long run as:

$$\mathrm{DR}(M) \coloneqq \sum_{m=1}^{M}\left(V_{r,1}^{\pi^{\star,m},m} - V_{r,1}^{\pi^m,m}\right), \tag{3.4}$$

$$\mathrm{CV}(M) \coloneqq \left[\sum_{m=1}^{M}\left(b_m - V_{g,1}^{\pi^m,m}\right)\right]_+. \tag{3.5}$$

There are two main reasons for considering the constraint violation in the long run. Firstly, in many applications such as supply chain and energy systems, the requirements of balancing the time-varying and unknown demands with the supply are formulated as some time-varying constraints. As long as the supply and the demand can be balanced in the long run, the policy is considered safe. Secondly, since the utility function g_h^m is unknown *a priori* and time-varying, the constraint $V_{g,1}^{\pi,m} \geq b_m$ may not be satisfied in every episode m. Rather, the agent strives to satisfy the constraints in the long run. In other words, the agent aims to ensure the long-term constraint $\sum_{m=1}^{M}(V_{g,1}^{\pi,m} - b_m) \geq 0$ over some given period of episodes M.

Linear function approximation We focus on a class of CMDPs, where transition kernels and reward/utility functions are linear in feature maps.

Assumption 5 (Linear Kernel CMDP)**.** *For every $m \in [M]$, the CMDP$(\mathcal{S},\mathcal{A},H,\mathbb{P}^m,r^m,g^m)$ satisfies the following conditions: (1) there exist a kernel feature map $\psi : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to \mathbb{R}^{d_1}$ and a vector $\theta_h^m \in \mathbb{R}^{d_1}$ with $\left\|\theta_h^m\right\|_2 \leq \sqrt{d_1}$ such that*

$$\mathbb{P}_h^m(x' \mid x,a) = \langle\psi(x,a,x'),\theta_h^m\rangle$$

for all $(x,a,x') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$ and $h \in [H]$; (2) there exist a feature map $\varphi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^{d_2}$ and vectors $\theta_{r,h}^m, \theta_{g,h}^m \in \mathbb{R}^{d_2}$ such that

$$r_h^m(x,a) = \left\langle\varphi(x,a),\theta_{r,h}^m\right\rangle \text{ and } g_h^m(x,a) = \left\langle\varphi(x,a),\theta_{g,h}^m\right\rangle$$

for all $(x,a) \in \mathcal{S}\times\mathcal{A}$ and $h \in [H]$, where $\max\left(\left\|\theta_{r,h}^m\right\|_2, \left\|\theta_{g,h}^m\right\|_2\right) \leq \sqrt{d_2}$; (3) for every function $V : \mathcal{S} \to [0,H]$, $\left\|\int_{\mathcal{S}}\psi(x,a,x')V(x')dx'\right\| \leq \sqrt{d_1}H$ for all $(x,a) \in \mathcal{S}\times\mathcal{A}$ and $\max(d_1,d_2) \leq d$.

This assumption adapts the definition of linear kernel MDP [8, 21, 133] to CMDP and has also been used in [34] for stationary constrained MDP problems. Some examples of linear

kernel MDPs include tabular MDPs [133], feature embedded transition models [126], and linear combinations of base models [90]. The linear kernal MDP defined in Assumption 5 is different from linear MDP [127, 70] since they define transition dynamics using the different feature maps, although both of them encapsulate the tabular MDP as the special case. They are not comparable since one cannot be implied by the other [133].

Variation budget. Note that $\mathbb{P}_h^m$ and r_h^m, g_h^m are determined by the unknown measures $\{\theta_h^m\}_{h\in[H],m\in[M]}$ and the latent vectors $\{\theta_{\diamond,h}^m\}_{h\in[H],m\in[M]}$ for $\diamond = r$ or g which can vary across the indexes $(m,h) \in [M]\times[H]$ in general. We measure the non-stationarity of the CMDP in terms of its variation in $\theta_h^m, \theta_{r,h}^m$ and $\theta_{g,h}^m$:

$$B_{\mathbb{P}} \coloneqq \sum_{m=2}^{M}\sum_{h=1}^{H}\left\|\theta_h^m - \theta_h^{m-1}\right\|_2, \tag{3.6a}$$

$$B_{\diamond} \coloneqq \sum_{m=2}^{M}\sum_{h=1}^{H}\left\|\theta_{\diamond,h}^m - \theta_{\diamond,h}^{m-1}\right\|_2, \text{ for } \diamond = r \text{ or } g, \tag{3.6b}$$

and denote $B_\Delta = B_{\mathbb{P}} + B_r + B_g$. Note that our definition of variation only imposes restrictions on the summation of non-stationarity across two different episodes, and it does not put any restriction on the difference between two consecutive steps in the same episode. In addition to the variations defined above, we introduce the total variation in the optimal policies of adjacent episodes:

$$B_\star \coloneqq \sum_{m=2}^{M}\sum_{h=1}^{H}\max_{x\in\mathcal{S}}\left\|\pi_h^{\star,m}(\cdot\mid x) - \pi_h^{\star,m-1}(\cdot\mid x)\right\|_1. \tag{3.7}$$

The notion of $B_\star$ is also used for online convex optimization with a dynamic regret criterion [14, 60, 61, 24] and for policy-based methods in non-stationary unconstrained MDPs [46, 132]. It is worth noting that the variations $(B_{\mathbb{P}}, B_\diamond)$ and $B_\star$ do not imply each other.

A special but important example of the non-stationarity is the system with piece-wise constant dynamics and rewards/utilities where the number of switches is S. In this case, all variation budgets $(B_{\mathbb{P}}, B_\diamond)$ and $B_\star$ can be upper bounded by $\mathcal{O}(SH)$. As one of the first works to investigate the non-stationary CMDP, we assume that we have access to quantities B_Δ and $B_\star$ or some upper bounds on them via an oracle.

3.3 Assumptions on Time-Varying Constraints

In this chapter, we consider two scenarios for the non-stationary CMDPs, each requiring some specific knowledge to enable safe exploration under the non-stationarity.

The first scenario assumes the knowledge of local variation budgets of constraints. We first define local variation budgets of constraints. To adapt the non-stationarity, the restart estimation of the value function is used, which breaks the M episodes into $\lceil\frac{M}{L}\rceil$ epochs. For every $\mathcal{E}\in\left[\lceil\frac{M}{L}\rceil\right]$, define $B_{g,\mathcal{E}}$ and $B_{\mathbb{P},\mathcal{E}}$ to be the local variation budgets of the utility function and transitions within epoch $\mathcal{E}$. By definition, we have $\sum_{\mathcal{E}=1}^{\lceil\frac{M}{L}\rceil} B_{g,\mathcal{E}} \le B_g$ and $\sum_{\mathcal{E}=1}^{\lceil\frac{M}{L}\rceil} B_{\mathbb{P},\mathcal{E}} \le B_{\mathbb{P}}$.

Assumption 6 (Local variation budgets of constraints)**.** *We have access to the local variation budget $B_{g,\mathcal{E}}$ and $B_{\mathbb{P},\mathcal{E}}$ for every $\mathcal{E} \in \left[\lceil \frac{M}{L} \rceil\right]$, and also the constrained optimization problems given in* (3.2) *are uniformly feasible.*

The second scenario extends the strict feasibility (also known as Slater condition) for problem (3.2) to non-stationary constrained optimization problems.

Assumption 7 (Uniformly strict feasibility)**.** *We have access to a sequence of constraint thresholds $\{b_m\}_{m=1}^M$ and a constant γ such that the constrained optimization problems in* (3.2) *are $\gamma-$uniformly strictly feasible, i.e., there exist $\gamma > 0$ and $\bar{\pi}^m \in \Delta(\mathcal{A} \mid \mathcal{S}, H)$ such that $V_{g,1}^{\bar{\pi}^m,m}(x_1) \geq b_m + \gamma$ for all $m = 1, \ldots, M$.*

Under this assumption, one can establish the strong duality and the boundedness of the optimal dual variable.

Lemma 6 (Lemma 1 in [34])**.** *Under Assumption 7, it holds that $V_{r,1}^{\pi^{\star,m},m}(x_1) = \mathcal{D}^m(\mu^{\star,m})$ and $0 \leq \mu^{\star,m} \leq H/\gamma$ for all $m = 1, \ldots, M$.*

Remark 13. *We require either Assumption 6 or Assumption* (7)*, and both of them need not hold simultaneously. Assumption 6 requires the local variation budgets of constraints, but does not enforce every instance problem* (3.2) *to be strictly feasible. It is suitable for the case with a forecasting oracle for the constraints. For example, in supply chain or energy systems, the supply is desired to match the time-varying and unknown demands where a forecasting oracle for the demands is usually available. In addition, it is also suitable for the case with only non-stationary rewards such as collision avoidance in a maze with a moving target. On the other hand, Assumption 7 needs the knowledge of strict feasible constraint thresholds, but does not require the local variation budgets of constraints. It is suitable for the case with a relatively large feasibility threshold γ.*

3.4 Safe Exploration under The Non-Stationarity

In Algorithm 1, we develop a new efficient method named Periodically Restarted Optimistic Primal-Dual Proximal Policy Optimization (PROPD-PPO) algorithm. In each episode, our algorithm consists of three main stages: periodically restarted policy improvement, dual update, and periodically restarted policy evaluation. We first present the high-level idea behind our method.

High-Level Idea

Safe exploration in non-stationary CMDPs is more challenging in that we need to reduce the constraint violation even when the constraints vary over the episodes. To overcome this issue, we develop our method based on some assumed knowledge on the constraints. Under Assumption 6, since the optimal dual variables may not be well-bounded, we need to add a dual regularization to stabilize the dual updates and fully utilize the convexity of the dual

Algorithm 1 Periodically Restarted Optimistic Primal-Dual Proximal Policy Optimization

1: **Inputs:** Time horizon M, restart period W, L, $\{Q^0_{r,h}, Q^0_{g,h}\}_{h=1}^H$ and $V^0_{g,1}$ being zero functions, initial policy $\{\pi^0_h\}_{h\in[H]}$ being uniform distributions on $\mathcal{A}$, initial dual variable $\mu^0 = 0$, dual regularization parameter ξ, learning rates $\alpha, \eta > 0$, χ.
2: **for** $m = 1, \ldots, M$ **do**
3: Set the initial state $x^m_1 = x_1$, $\ell^m_\pi = (\lceil \frac{m}{L} \rceil - 1)L + 1$, $\ell^m_Q = (\lceil \frac{m}{W} \rceil - 1)W + 1$.
4: **if** $m = \ell^m_\pi$ **then**
5: Set $\{Q^{m-1}_{r,h}, Q^{m-1}_{g,h}\}_{h=1}^H$ as zero functions and set $\{\pi^{m-1}_h\}_{h=1}^H$ as uniform distributions on $\mathcal{A}$.
6: **end if**
7: **for** $h = 1, 2, \ldots, H$ **do**
8: Update the policy $\pi^m_h(\cdot \,|\, \cdot) \propto \pi^{m-1}_h(\cdot \,|\, \cdot) \exp\left(\alpha \left(Q^{m-1}_{r,h} + \mu^{m-1} Q^{m-1}_{g,h}\right)(\cdot, \cdot)\right)$.
9: Take an action $a^m_h \sim \pi^m_h(\cdot \,|\, x^m_h)$ and receive reward/utility $r_h(x^m_h, a^m_h), g_h(x^m_h, a^m_h)$.
10: Observe the next state x^m_{h+1}.
11: **end for**
12: Update the dual variable by $\mu^m = \mathrm{Proj}_{[0,\chi]}\left(\mu^{m-1} + \eta\left(b_m - V^{m-1}_{g,1}(x_1) - \xi\mu^{m-1}\right)\right)$.
13: Estimate $\{Q^m_{r,h}, Q^m_{g,h}\}_{h=1}^H$ and $V^m_{g,1}$ via $\mathrm{LSTD}\Big(\{x^\tau_h, a^\tau_h, r^\tau_h(x^\tau_h, a^\tau_h), g^\tau_h(x^\tau_h, a^\tau_h)\}^{H,m}_{h=1,\tau=\ell^m_Q}\Big)$.
14: **end for**

function. In addition, the knowledge of local variation of the constraints is needed to obtain an optimistic estimator of constraint functions, so that a large dual variable cannot amplify the estimation error of the constraint functions. This is different from the dual update that has been used in Lagrangian-based stationary CMDPs under the strict feasible condition [33, 34, 128, 41, 83, 100]. On the other hand, under Assumption 7, the optimal dual variables can be bounded by Lemma 6. Then, the dual regularization and an optimistic estimator for the constraint functions are not necessary. Thus, a standard dual update will be enough.

Periodically Restarted Policy Improvement

One way to update the policy π^m is to solve the Lagrangian-based policy optimization problem $\max_{\pi\in\Delta(\mathcal{A}|\mathcal{S},H)} \mathcal{L}^m_\xi(\pi, \mu^{m-1})$, where $\mathcal{L}^m_\xi(\pi, \mu^{m-1})$ is defined in (3.9) and the dual variable μ^{m-1} is from episode $m-1$. Motivated by the policy improvement step in NPG [73], TRPO [108], and PPO [107], we perform a simple policy update in the online mirror descent fashion by

$$\arg\max_{\pi\in\Delta(\mathcal{A}|\mathcal{S},H)} \sum_{h=1}^H \left\langle \left(Q^{m-1}_{r,h} + \mu^{m-1} Q^{m-1}_{g,h}\right)(x_h, \cdot), \pi_h - \pi^{m-1}_h \right\rangle - \frac{1}{\alpha} \sum_{h=1}^H D\left(\pi_h(\cdot|x_h) \,|\, \pi^{m-1}_h(\cdot|x_h)\right). \quad (3.8)$$

Since the above update is separable over H steps, we can update the policy π^m as line 8 in Algorithm 1, leading to a closed-form solution for each step $h \in [H]$. Furthermore, in order to guarantee the policy to be exploratory enough in new environments, our policy improvement step also features a periodic restart mechanism, which resets its policy to a uniform distribution over the action space $\mathcal{A}$ every L episodes.

Remark 14. *Although policy improvement step* (3.8) *has been used in stationary CMDPs [34], our method differs in the sense that we remove the requirement to mix the policy with a uniform policy at every iteration. This is due to a technical improvement in the analysis by replacing the "pushback property of KL-divergence lemma" (Lemma 14 in [34]) with the "one-step descent lemma" for the KL-regularized optimization.*

Dual Update

We first define the modified Lagrangian of (3.3) to be

$$\mathcal{L}_\xi^m(\pi,\mu) \coloneqq V_{r,1}^{\pi,m} + \mu\left(V_{g,1}^{\pi,m} - b_m\right) + \frac{\xi}{2}\|\mu\|_2^2 \tag{3.9}$$

where $\xi \geq 0$ is the dual regularization parameter to be determined later. Since the value function $V_{g,1}^{\pi,m}$ is unknown, in order to infer the constraint violation for the dual update, we estimate $V_{g,1}^{\pi^m,m}(x_1)$ via an optimistic policy evaluation. We update the Lagrange multiplier μ by moving μ^m to the direction of minimizing the estimated Lagrangian $\mathcal{L}(\pi,\mu)$:

$$\widetilde{\mathcal{L}}_\xi^m(\pi,\mu) \coloneqq V_{r,1}^m + \mu\left(V_{g,1}^m - b_m\right) + \frac{\xi}{2}\|\mu\|_2^2. \tag{3.10}$$

over $\mu \geq 0$ in line 14 of Algorithm 1, where $\eta > 0$ is a stepsize and $\text{Prof}_{[0,\chi]}$ is a projection onto $[0,\chi]$ with an upper bound χ on μ^m. The choices of the parameters χ and ξ depend on the assumption:

$$\begin{aligned}
&\xi > 0,\ \chi = \infty,\ \text{under Assumption 6,}\\
&\xi = 0,\ \chi = \frac{2H}{\gamma}, \text{under Assumption 7.}
\end{aligned}$$

Under Assumption 6, since the strictly feasibility may not hold for all episodes (corresponding to $\gamma = 0$), we may not have a finite upper bound on the dual variable μ. Thus, a dual regularization with $\xi > 0$ is needed to stabilize the dual updates under the non-stationarity. The value of ξ depends on the number of episodes M and the variation budgets $B_{\mathbb{P}}, B_g$. On the other hand, under Assumption 7, we choose $\chi = \frac{2H}{\gamma} \geq 2\mu^{*,m}$ similarly as [34, 41], so that the projection interval $[0,\chi]$ includes all optimal dual variables $\{\mu^{*,m}\}_{m=1}^M$ in light of Lemma 6.

Periodically Restarted Optimistic Policy Evaluation

To evaluate the policy under the unknown nonstationarity, we take the Least-Squares Temporal Difference (LSTD) [20, 80] with UCB to properly handle the exploration-exploitation trade-off and apply the restart strategy to adapt to the unknown nonstationarity. In particular, we apply the restart strategy and evaluate the policy π^m only based on the previous historical

trajectories from the episode ℓ_Q^m to the episode m instead of the all previous historical trajectories. The method is standard and summarized in Appendix.

After obtaining the estimates of $\mathbb{P}_h^m V_{\diamond,h+1}^m$ and $\diamond_h^m(\cdot,\cdot)$ for $\diamond = r$ or g, we update the estimated action-value function $\left\{Q_{\diamond,h}^m\right\}_{h=1}^H$ iteratively and add UCB bonus terms $\Gamma_h^m(\cdot,\cdot)$, $\Gamma_{\diamond,h}^m(\cdot,\cdot) : \mathcal{S}\times\mathcal{A}\to\mathbb{R}^+$ so that

$$\Omega_{1,\diamond} \coloneqq (\varphi^m)^\top u_{\diamond,h}^m + \Gamma_h^m \quad \text{and} \quad \Omega_{2,\diamond} \coloneqq \left(\phi_{\diamond,h}^m\right)^\top w_{\diamond,h}^m + \Gamma_{\diamond,h}^m$$

all become upper bounds on $\mathbb{P}_h^m V_{\diamond,h+1}^m$ and $\diamond_h^m(\cdot,\cdot)$ (up to some errors due to the non-stationarity). Here, the weights $u_{\diamond,h}^m, w_{\diamond,h}^m$ and the bonus terms $\Gamma_h^m, \Gamma_{\diamond,h}^m$ are defined in Appendix. Moreover,

$$\begin{aligned} Q_{r,h}^m(\cdot,\cdot) &= \min\left(H-h+1, \Omega_{1,r}(\cdot,\cdot) + \Omega_{2,r}(\cdot,\cdot)\right)_+, \\ Q_{g,h}^m(\cdot,\cdot) &= \min\left(H-h+1, \Omega_{1,g}(\cdot,\cdot) + \Omega_{2,g}(\cdot,\cdot) + LV\right)_+ \end{aligned}$$

where $LV > 0$ depends on the local variation budgets of the constraint $B_{\mathbb{P},\mathcal{E}}$, $B_{g,\mathcal{E}}$ under Assumption 6, $LV = 0$ under Assumption 7, and $(x)_+$ denotes the maximum between x and 0. The reason for introducing a positive LV term under Assumption 6 is to guarantee that the model prediction error in $Q_{g,h}^m$ is non-positive when the dual variable μ is very large.

3.5 Main Results

We now present the dynamic regret and the constraint violation bounds for Algorithm 1 under the two alternative assumptions introduced in Section 3.3. The choices of the algorithm parameters will depend on the assumption used for the analysis. When both assumptions are satisfied, one can check which one yields a tighter bound, and this depends on the value of the strict feasibility threshold γ (and the values of H, M if in the tabular CMDP setting).

Linear Kernal CMDP

We first present the results for linear Kernal CMDP under each of Assumptions 6 and 7.

Theorem 7 (Linear Kernal CMDP + Assumption 6)**.** *Let Assumptions 5 and 6 hold. Given $p \in (0,1)$, we set $\alpha = H^{-1}M^{-\frac{1}{2}}(\sqrt{d}B_\Delta + B_\star)^{\frac{1}{3}}$, $L = M^{\frac{3}{4}}(\sqrt{d}B_\Delta + B_\star)^{-\frac{2}{3}}$, $\eta = M^{-\frac{1}{2}}$, $\xi = 2H(\sqrt{d}B_\Delta + B_\star)^{\frac{1}{3}}M^{-\frac{1}{2}}$, $W = d^{-\frac{1}{4}}H^{-1}M^{\frac{1}{2}}B_\Delta^{-\frac{1}{2}}$, in Algorithm 1 and set set $\beta = C_1\sqrt{dH^2\log(dW/p)}$, $LV = B_{\mathbb{P},\mathcal{E}}H^2d_1\sqrt{d_1W} + B_{g,\mathcal{E}}\sqrt{d_2W}$. Then, with probability $1-p$, the dynamic regret and the constraint violation satisfy*

$$\begin{aligned} \mathrm{DR}(M) &\le \widetilde{\mathcal{O}}\left(d^{\frac{9}{8}}H^{\frac{5}{2}}M^{\frac{3}{4}}(\sqrt{d}B_\Delta + B_*)^{\frac{1}{3}}\right), \\ \mathrm{CV}(M) &\le \widetilde{\mathcal{O}}\left(d^{\frac{9}{8}}H^{\frac{5}{2}}M^{\frac{3}{4}}(\sqrt{d}B_\Delta + B_*)^{\frac{1}{3}}\right). \end{aligned}$$

Theorem 8 (Linear Kernal CMDP + Assumption 7)**.** *Let Assumptions 5 and 7 hold. Given* $p \in (0,1)$*, we set* $\alpha = \gamma H^{-\frac{3}{2}} M^{-\frac{1}{3}} (\sqrt{d} B_\Delta + B_\star)^{\frac{1}{3}}$*,* $L = M^{\frac{2}{3}} (\sqrt{d} B_\Delta + B_\star)^{-\frac{2}{3}}$*,* $\eta = M^{-\frac{1}{2}}$*,* $\xi = 0$*,* $W = d^{-\frac{1}{4}} H^{-1} M^{\frac{1}{2}} B_\Delta^{-\frac{1}{2}}$ *in Algorithm 1 and set set* $\beta = C_1 \sqrt{dH^2 \log(dW/p)}$*,* $LV = 0$*. Then, with probability* $1-p$*, the dynamic regret and the constraint violation satisfy*

$$\mathrm{DR}(M) \le \widetilde{\mathcal{O}}\left(\gamma^{-1} d^{\frac{9}{8}} H^{\frac{5}{2}} M^{\frac{3}{4}} (\sqrt{d} B_\Delta + B_*)^{\frac{1}{3}}\right),$$
$$\mathrm{CV}(M) \le \widetilde{\mathcal{O}}\left(\gamma^{-1} d^{\frac{9}{8}} H^{\frac{5}{2}} M^{\frac{3}{4}} (\sqrt{d} B_\Delta + B_*)^{\frac{1}{3}}\right).$$

The proofs for Theorems 7 and 8 can be found in Appendix. Our dynamic regret bounds in Theorems 7 and 8 have the optimal dependence on the total number of episodes M. This matches the existing bounds in the general non-stationary linear kernel MDP setting without any constraints [132, 134, 123]. The dependence on the variation budgets $(B_\Delta, B_\star)$ also matches the existing bound in policy-based method for the non-stationary linear kernel MDP setting [132]. Regarding the long-term safe exploration, we provide the first finite-time constraint violation result in the non-stationary CMDP setting.

In the linear kernel CMDP setting, the same dynamic regret and constraint violation bounds are obtained under either of Assumptions 6 and 7, except that the dynamic regret and constraint violation under Assumption 7 also depend on the strict feasibility threshold γ. When γ is small, i.e., there exist some episodes for which the CMDP problem (3.2) does not have a large enough strict feasibility threshold, the dynamic regret and constraint violation bounds in Theorem 8 may be large.

Tabular CMDP

A special case of the linear kernel CMDP in Assumption 5 is the tabular CMDP with $|\mathcal{S}| < \infty$ and $|\mathcal{A}| < \infty$. In the tabular case, improved results can be obtained by incorporating Algorithm 1 with a variant of the optimistic policy evaluation method. We refer the reads to Appendix for such procedures and state the result below:

Theorem 9 (Tabular CMDP + Assumption 6)**.** *Let Assumption 6 hold and consider a tabular CMDP. Given* $p \in (0,1)$ *and* $\rho \in [\frac{1}{3}, \frac{1}{2}]$*, we set* $\alpha = H^{-\frac{1}{3}} M^{-\rho} (B_\Delta + B_\star)^{\frac{1}{3}}$*,* $L = H^{-\frac{1}{3}} M^{\frac{1+\rho}{2}} (B_\Delta + B_\star)^{-\frac{2}{3}}$*,* $\eta = H^{-\frac{1}{3}} M^{-\frac{1}{2}}$*,* $\xi = 2H^{\frac{5}{3}} (B_\Delta + B_\star)^{\frac{1}{3}} M^{-\rho}$*,* $W = H^{\frac{2}{3}} |\mathcal{S}|^{\frac{2}{3}} |\mathcal{A}|^{\frac{1}{3}} \left(\frac{M}{B_\Delta}\right)^{\frac{2}{3}}$ *in Algorithm 1 and* $\beta = C_4 H \sqrt{|\mathcal{S}| \log(|\mathcal{S}||\mathcal{A}|W/p)}$*,* $LV = B_{\mathbb{P},\mathcal{E}} H + B_{g,\mathcal{E}}$*. Then, with probability* $1-p$*, the dynamic regret and the constraint violation satisfy*

$$\mathrm{DR}(M) \le \widetilde{\mathcal{O}}\left(|\mathcal{S}|^{\frac{2}{3}} |\mathcal{A}|^{\frac{1}{3}} H^{\frac{5}{3}} M^{\frac{1+\rho}{2}} (B_\Delta + B_*)^{\frac{1}{3}}\right),$$
$$\mathrm{CV}(M) \le \widetilde{\mathcal{O}}\left(|\mathcal{S}|^{\frac{2}{3}} |\mathcal{A}|^{\frac{1}{3}} H^{\frac{5}{3}} M^{\frac{2-\rho}{2}} (B_\Delta + B_*)^{\frac{1}{3}}\right).$$

Theorem 10 (Tabular CMDP + Assumption 7)**.** *Let Assumption 7 hold and consider a tabular CMDP. Given* $p \in (0,1)$*, we set* $\alpha = \gamma H^{-\frac{3}{2}} M^{-\frac{1}{3}} (B_\Delta + B_\star)^{\frac{1}{3}}$*,* $L = M^{\frac{2}{3}} (B_\Delta + B_\star)^{-\frac{2}{3}}$*,*

$\eta = M^{-\frac{1}{2}}$, $\xi = 0$, $W = |\mathcal{S}|^{\frac{2}{3}}|\mathcal{A}|^{\frac{1}{3}}\left(\frac{M}{B_\Delta}\right)^{\frac{2}{3}}$ in Algorithm 1 and $\beta = C_4 H\sqrt{|\mathcal{S}|\log(|\mathcal{S}||\mathcal{A}|W/p)}$, $LV = 0$. Then, with probability $1-p$, the dynamic regret and the constraint violation satisfy

$$\mathrm{DR}(M) \leq \widetilde{\mathcal{O}}\left(\gamma^{-1}|\mathcal{S}|^{\frac{2}{3}}|\mathcal{A}|^{\frac{1}{3}}H^{\frac{5}{2}}M^{\frac{2}{3}}(B_\Delta + B_\star)^{\frac{1}{3}}\right),$$
$$\mathrm{CV}(M) \leq \widetilde{\mathcal{O}}\left(\gamma^{-1}|\mathcal{S}|^{\frac{2}{3}}|\mathcal{A}|^{\frac{1}{3}}H^{\frac{5}{2}}M^{\frac{2}{3}}(B_\Delta + B_\star)^{\frac{1}{3}}\right).$$

The proofs for Theorems 9 and 10 can be found in Appendix. For the tabular CMDP under Assumption 6, there is a trade-off for the dependence on the total number of episodes M between the dynamic regret and the constraint violation. This trade-off is controlled by the primal update parameter α and the dual regularization parameter ξ. Such trade-off does not appear in the linear kernel CMDP setting because the dynamic regret and constraint violation in the linear kernel CMDP are bottlenecked by the error in the non-stationary policy evaluation.

The dynamic regret and constraint violation bounds in Theorem 10 have an improved dependence on the total number of episodes M compared to Theorems 7 and 8. This improvement is due to the improved result of the policy evaluation step in the tabular setting. The dependence on M in Theorem 10 is also better than that of Theorem 9. This is due to a sharper analysis for the constraint violation under Assumption 7 based on [11, Proposition 3.60]. However, the dynamic regret and constraint violation bounds in Theorem 10 have a worse dependence on the horizon H and are also dependent on the feasibility threshold γ compared to Theorem 9. In addition, the dependence of the dynamic regret on M and $(B_\Delta, B_\star)$ matches the existing bound in the non-stationary tabular MDP setting without any constraints [86].

3.6 Summary

Our results are summarized in Table 3.6.1 and our method is the first provably efficient algorithm for non-stationary CMDPs with safe exploration.

Setting	**Dynamic regret**	**Constraint violation**
Tabular+ A6	$\widetilde{\mathcal{O}}\left(\lvert\mathcal{S}\rvert^{\frac{2}{3}}\lvert\mathcal{A}\rvert^{\frac{1}{3}}H^{\frac{5}{3}}M^{\frac{1+\rho}{2}}(B_\Delta+B_*)^{\frac{1}{3}}\right)$	$\widetilde{\mathcal{O}}\left(\lvert\mathcal{S}\rvert^{\frac{2}{3}}\lvert\mathcal{A}\rvert^{\frac{1}{3}}H^{\frac{5}{3}}M^{\frac{2-\rho}{2}}(B_\Delta+B_*)^{\frac{1}{3}}\right)$
Tabular+ A7	$\widetilde{\mathcal{O}}\left(\gamma^{-1}\lvert\mathcal{S}\rvert^{\frac{2}{3}}\lvert\mathcal{A}\rvert^{\frac{1}{3}}H^{\frac{5}{2}}M^{\frac{2}{3}}(B_\Delta+B_\star)^{\frac{1}{3}}\right)$	$\widetilde{\mathcal{O}}\left(\gamma^{-1}\lvert\mathcal{S}\rvert^{\frac{2}{3}}\lvert\mathcal{A}\rvert^{\frac{1}{3}}H^{\frac{5}{2}}M^{\frac{2}{3}}(B_\Delta+B_\star)^{\frac{1}{3}}\right)$
Linear kernel+A6	$\widetilde{\mathcal{O}}\left(d^{\frac{9}{8}}H^{\frac{5}{2}}M^{\frac{3}{4}}(\sqrt{d}B_\Delta+B_*)^{\frac{1}{3}}\right)$	$\widetilde{\mathcal{O}}\left(d^{\frac{9}{8}}H^{\frac{5}{2}}M^{\frac{3}{4}}(\sqrt{d}B_\Delta+B_*)^{\frac{1}{3}}\right)$
Linear kernel+A7	$\widetilde{\mathcal{O}}\left(\gamma^{-1}d^{\frac{9}{8}}H^{\frac{5}{2}}M^{\frac{3}{4}}(\sqrt{d}B_\Delta+B_*)^{\frac{1}{3}}\right)$	$\widetilde{\mathcal{O}}\left(\gamma^{-1}d^{\frac{9}{8}}H^{\frac{5}{2}}M^{\frac{3}{4}}(\sqrt{d}B_\Delta+B_*)^{\frac{1}{3}}\right)$

Table 3.6.1: We summarize the dynamic regrets and constraint violations obtained in this chapter for tabular and linear kernel CMDPs under different assumptions. Here, A6 and A7 represent the assumption 6 and assumption 7 respectively, γ is the strict feasibility threshold of the constraints and is defined in Assumption 7, H is the horizon of each episode, M is the total number of episodes, d is the dimension of the feature mapping, $|\mathcal{S}|$ and $|\mathcal{A}|$ are the cardinalities of the state and action spaces, and B_Δ, B_* are the variation budgets defined in (3.6) and (3.7). There is a trade-off controlled by $\rho \in [\frac{1}{3}, \frac{1}{2}]$ between the dynamic regret and constraint violation for the tabular CMDP under Assumption 6.

Appendix

Due to the space limit, we provide the proof for results related to the linear Kernel MDP setting here ane the proofs for the tabular MDP setting is similar and can be found in [35].

3.A Policy Evaluation Algorithm

Policy Evaluation Algorithm for Linear Kernel MDP Setting

For episode m and each step $h \in [H]$, we estimate $\mathbb{P}_h^m V_{r,h+1}^m$ in the Bellman equation (3.1) by ${\phi_{r,h}^m}^\top w_{r,h}^m$, where $w_{r,h}^m$ is updated by the minimizer of the regularized least-squares problem over w,

$$\sum_{\tau=\ell_Q^m}^{m-1} \left(V_{r,h+1}^\tau(x_{h+1}^\tau) - \phi_{r,h}^\tau(x_h^\tau, a_h^\tau)^\top w\right)^2 + \lambda \|w\|^2 \tag{3.11}$$

where

$$\phi_{r,h}^\tau(\cdot,\cdot) \coloneqq \textstyle\int_{\mathcal{S}} \psi(\cdot,\cdot,x') V_{r,h+1}^\tau(x') dx' \tag{3.12a}$$

$$V_{r,h+1}^\tau(\cdot) = \left\langle Q_{r,h+1}^\tau(\cdot,\cdot), \pi_{h+1^\tau(\cdot|\cdot)} \right\rangle_{\mathcal{A}} \tag{3.12b}$$

for all $h \in [H-1]$ and $V_{r,H+1}^\tau = 0$, and $\lambda > 0$ is the regularization parameter.

Similarly, we estimate $\mathbb{P}_h^m V_{g,h+1}^m$ by $(\phi_{g,h}^m)^\intercal w_{g,h}^m$. We display the least-squares solution in lines 3-5 of Algorithm 2 where the symbol $\diamond$ denotes r or g. In addition, since we consider the bandit reward/utility feedback in the linear function approximation setting, we also need to estimate $r_h^m(\cdot,\cdot)$ by $\left(\varphi^m(\cdot,\cdot)\right)^\top u_{r,h}^m$, where $u_{r,h}^m$ is updated by the minimizer of another regularized least-squares problem,

$$\sum_{\tau=\ell_Q^m}^{m-1} \left(r_h^\tau(x_h^\tau, a_h^\tau) - \left(\varphi^\tau(x_h^\tau, a_h^\tau)\right)^\top u\right)^2 + \lambda \|u\|_2^2 \tag{3.13}$$

where λ is the regularization parameter. Similarly, we estimate $g_h^m(\cdot,\cdot)$ by $\left(\varphi^m(\cdot,\cdot)\right)^\top u_{g,h}^m$. The least-squares solutions lead to lines 8-9 of Algorithm 2.

Algorithm 2 Least-Squares Temporal Difference with UCB exploration (LSTD)

Inputs: $\{x_h^\tau, a_h^\tau, r_h^\tau(x_h^\tau, a_h^\tau), g_h^\tau(x_h^\tau, a_h^\tau)\}_{h=1,\tau=\ell_Q^m}^{H,m}$, regularization parameter λ, UCB parameter β, local variation budgets $B_{\mathbb{P},\mathcal{E}}, B_{g,\mathcal{E}}$.
for $h = H, H-1, \ldots, 1$ **do**
  $\Lambda_{\diamond,h}^m = \sum_{\tau=\ell_Q^m}^{m-1} \phi_{\diamond,h}^\tau\left(x_h^\tau, a_h^\tau\right) \phi_{\diamond,h}^\tau\left(x_h^\tau, a_h^\tau\right)^\top + \lambda I.$
  $w_{\diamond,h}^m = \left(\Lambda_{\diamond,h}^m\right)^{-1} \sum_{\tau=\ell_Q^m}^{m-1} \phi_{\diamond,h}^\tau\left(x_h^\tau, a_h^\tau\right) V_{\diamond,h+1}^\tau\left(x_{h+1}^\tau\right).$
  $\phi_{\diamond,h}^m(\cdot,\cdot) = \int_{\mathcal{S}} \psi(\cdot,\cdot,x') V_{\diamond,h+1}^m(x')\, dx'.$
  $\Gamma_{\diamond,h}^m(\cdot,\cdot) = \beta\left(\phi_{\diamond,h}^m(\cdot,\cdot)^\top \left(\Lambda_{\diamond,h}^m\right)^{-1} \phi_{\diamond,h}^m(\cdot,\cdot)\right)^{1/2}.$
  $LV = \begin{cases} B_{\mathbb{P},\mathcal{E}} H^2 d_1 \sqrt{d_1 W} + B_{g,\mathcal{E}} \sqrt{d_2 W}, \text{Under Assumption 6,} \\ 0, \text{Under Assumption 7.} \end{cases}$
  $\Lambda_h^m = \sum_{\tau=\ell_Q^m}^{m-1} \varphi\left(x_h^\tau, a_h^\tau\right) \varphi\left(x_h^\tau, a_h^\tau\right)^\top + \lambda I\ .$
  $u_{\diamond,h}^m = \left(\Lambda_h^m\right)^{-1} \sum_{\tau=\ell_Q^m}^{m-1} \varphi\left(x_h^\tau, a_h^\tau\right) \diamond_h^m\left(x_h^\tau, a_h^\tau\right).$
  $\Gamma_h^m(\cdot,\cdot) = \beta\left(\varphi(\cdot,\cdot)^\top \left(\Lambda_h^m\right)^{-1} \varphi(\cdot,\cdot)\right)^{1/2}.$
  $Q_{r,h}^m(\cdot,\cdot) = \min\left(H-h+1, \varphi(\cdot,\cdot)^\top u_{r,h}^m + \phi_{r,h}^m(\cdot,\cdot)^\top w_{r,h}^m + \left(\Gamma_h^m + \Gamma_{r,h}^m\right)(\cdot,\cdot)\right)_+,$
  $Q_{g,h}^m(\cdot,\cdot) = \min\left(H-h+1, \varphi(\cdot,\cdot)^\top u_{g,h}^m + \phi_{g,h}^m(\cdot,\cdot)^\top w_{g,h}^m + \left(\Gamma_h^m + \Gamma_{g,h}^m\right)(\cdot,\cdot) + LV\right)_+.$
  $V_{\diamond,h}^m(\cdot) = \left\langle Q_{\diamond,h}^m(\cdot,\cdot), \pi_h^m(\cdot \mid \cdot)\right\rangle_{\mathcal{A}}^+.$
end for
Output: $\{Q_{r,h}^m, Q_{g,h}^m\}_{h=1}^H$ and $V_{g,1}^m$.

3.B Proof for Linear Kernel CMDP Case under Assumption 6

Proof of Dynamic Regret Bound in Theorem 7

Our analysis for the dynamic regret begins with the decomposition of the regret given in (3.4):

Lemma 7 (Dynamic regret decomposition)**.** *The dynamic regret in* (3.4) *can be expanded as*

$$\begin{aligned}
\mathrm{DR}(M) = &\sum_{m=1}^{M}\sum_{h=1}^{H} \mathbb{E}_{\pi^\star,\ell_\pi^m,\mathbb{P}^{\ell_\pi^m}}\left[\left\langle Q_{r,h}^m(x_h,\cdot), \pi_h^{\star,m}(\cdot \mid x_h) - \pi_h^m(\cdot \mid x_h)\right\rangle\right] \\
&+ \sum_{m=1}^{M}\sum_{h=1}^{H} \left(\mathbb{E}_{\pi^\star,m,\mathbb{P}^m} - \mathbb{E}_{\pi^\star,\ell_\pi^m,\mathbb{P}^{\ell_\pi^m}}\right)\left[\left\langle Q_{r,h}^m(x_h,\cdot), \pi_h^{\star,m}(\cdot \mid x_h) - \pi_h^m(\cdot \mid x_h)\right\rangle\right] \\
&+ \sum_{m=1}^{M}\sum_{h=1}^{H} \mathbb{E}_{\pi^\star,m,\mathbb{P}^m}\left[\iota_{r,h}^m(x_h,a_h)\right] - \sum_{m=1}^{M}\sum_{h=1}^{H} \iota_{r,h}^m\left(x_h^m, a_h^m\right) + S_{r,H,2}^M
\end{aligned}$$

where $\left\{S^m_{r,h,k}\right\}_{(m,h,k)\in[M]\times[H]\times[2]}$ *is a martingale.*

Proof. We have

$$\mathrm{DR}(M) = \underbrace{\sum_{m=1}^{M}\left(V^{\pi^{\star,m},m}_{r,1}(x_1) - V^m_{r,1}(x_1)\right)}_{\text{(R.I)}} + \underbrace{\sum_{m=1}^{M}\left(V^m_{r,1}(x_1) - V^{\pi^m,m}_{r,1}(x_1)\right)}_{\text{(R.II)}}, \tag{3.14}$$

where the policy $\pi^{\star,m}$ is the best policy in hindsight for the constrained optimization problem (3.2) at episode m; the policy π^m is the policy updated in line 10 of Algorithm 1; $V^{\pi^{\star,m},m}_{r,1}, V^{\pi^m,m}_{r,1}(x_1)$ are the value functions corresponding to the policies $\pi^{\star,m}$ and π^m, and the value function $V^m_{r,1}(x_1)$ is estimated from an optimistic policy evaluation by Algorithm 2. To bound the total regret (3.14), we need to analyze the two terms (R.I) and (R.II) separately.

To analyze the first term (R.I), we define the model prediction error for the reward at episode m as

$$\iota^m_{r,h} \coloneqq r^m_h + \mathbb{P}^m_h V^m_{r,h+1} - Q^m_{r,h} \tag{3.15}$$

for all $(m,h)\in[M]\times[H]$, which describes the error in the Bellman equation (3.1) using $V^m_{r,h+1}$ instead of $V^{\pi^m,m}_{r,h+1}$ and using the sample estimation of $\mathbb{P}^m_h$. With this notation, we expand the term (R.I) in (3.14) into

$$\sum_{m=1}^{M}\sum_{h=1}^{H}\mathbb{E}_{\pi^{\star,m},\mathbb{P}^m}\left[\left\langle Q^m_{r,h}(x_h,\cdot), \pi^{\star,m}_h(\cdot\mid x_h) - \pi^m_h(\cdot\mid x_h)\right\rangle\right] + \sum_{m=1}^{M}\sum_{h=1}^{H}\mathbb{E}_{\pi^{\star,m},\mathbb{P}^m}\left[\iota^m_{r,h}(x_h,a_h)\right] \tag{3.16}$$

where the first double sum is linear in terms of the policy difference and the second one describes the total model prediction errors. The above expansion is proved in Lemma 20. We can further decompose the first term in (3.16) as

$$\begin{aligned}
&\sum_{m=1}^{M}\sum_{h=1}^{H}\mathbb{E}_{\pi^{\star,m},\mathbb{P}^m}\left[\left\langle Q^m_{r,h}(x_h,\cdot), \pi^{\star,m}_h(\cdot\mid x_h) - \pi^m_h(\cdot\mid x_h)\right\rangle\right]\\
=&\sum_{m=1}^{M}\sum_{h=1}^{H}\mathbb{E}_{\pi^{\star},\ell^m_\pi,\mathbb{P}^{\ell^m_\pi}}\left[\left\langle Q^m_{r,h}(x_h,\cdot), \pi^{\star,m}_h(\cdot\mid x_h) - \pi^m_h(\cdot\mid x_h)\right\rangle\right]\\
&+\sum_{m=1}^{M}\sum_{h=1}^{H}\left(\mathbb{E}_{\pi^{\star,m},\mathbb{P}^m} - \mathbb{E}_{\pi^{\star},\ell^m_\pi,\mathbb{P}^{\ell^m_\pi}}\right)\cdot\left[\left\langle Q^m_{r,h}(x_h,\cdot), \pi^{\star,m}_h(\cdot\mid x_h) - \pi^m_h(\cdot\mid x_h)\right\rangle\right].
\end{aligned} \tag{3.17}$$

To analyze the second term (R.II) in (3.14), we will first introduce some notations. for every $(m,h)\in[M]\times[H]$, we define $\mathcal{F}^m_{h,1}$ as a $\sigma-$algebra generated by state-action sequences, reward and utility functions,

$$\left\{(x^\tau_i, a^\tau_i)\right\}_{(\tau,i)\in[m-1]\times[H]}\bigcup\left\{(x^m_i,a^m_i)\right\}_{i\in[H]}.$$

Similarly, we define $\mathcal{F}_{h,2}^m$ as an $\sigma-$algebra generated by

$$\{(x_i^\tau, a_i^\tau)\}_{(\tau,i)\in[m-1]\times[H]} \bigcup \{(x_i^m, a_i^m)\}_{i\in[H]} \bigcup \{x_{h+1}^m\}.$$

Here, x_{H+1}^m is a null state for every $m \in [M]$. A filtration is a sequence of $\sigma-$algebras $\{\mathcal{F}_{h,k}^m\}_{(k,h,m)\in[K]\times[H]\times[2]}$ in terms of the time index

$$t(m,h,k) \coloneqq 2(m-1)H + 2(h-1) + k \tag{3.18}$$

such that $F_{h,k}^m \subset F_{h',k'}^{m'}$ for every $t(m,h,k) \le t(m',h',k')$. The estimated reward/utility value functions $V_{r,h}^m, V_{g,h}^m$ and the associated Q-functions $Q_{r,h}^m, Q_{g,h}^m$ are $\mathcal{F}_{1,1}^m$ measurable since they are obtained from previous $m-1$ historical trajectories. With these notations, we can expand the term (R.II) in (3.14) into

$$-\sum_{m=1}^{M}\sum_{h=1}^{H} \iota_{r,h}^m\left(x_h^m, a_h^m\right) + S_{r,H,2}^M \tag{3.19}$$

where $\left\{S_{r,h,k}^m\right\}_{(m,h,k)\in[M]\times[H]\times[2]}$ is a martingale adapted to the filtration $\left\{\mathcal{F}_{h,k}^m\right\}_{(m,h,k)\in[M]\times[H]\times[2]}$ in terms of the time index t. We define $S_{r,H,2}^M$ and prove (3.19) in Lemma 15. □

In the following proofs, we use the shorthand notation $\left\langle Q_{r,h}^{m-1} + \mu^{m-1}Q_{g,h}^{m-1}, \pi_h\right\rangle$ for $\left\langle \left(Q_{r,h}^{m-1} + \mu^{m-1}Q_{g,h}^{m-1}\right)(x_h,\cdot), \pi_h(\cdot \mid x_h)\right\rangle$ and the shorthand notation $D\left(\pi_h \mid \pi_h^{m-1}\right)$ for $D\left(\pi_h(\cdot \mid x_h) \mid \pi_h^{m-1}(\cdot \mid x_h)\right)$ if dependence on the state-action sequence $\{x_h, a_h\}_{h=1}^H$ is clear from the context.

Lemma 8 (Primal step for dynamic regret)**.** *Let Assumption 5 hold. For the primal update rule in line 10 of Algorithm 1, we have*

$$\begin{aligned}\sum_{h=1}^{H}\left\langle Q_{r,h}^{m-1}, \pi_h^{\star,m-1} - \pi_h^{m-1}\right\rangle \le &-\mu^{m-1}\sum_{h=1}^{H}\left\langle Q_{g,h}^{m-1}, \pi_h^{\star,m-1} - \pi_h^{m-1}\right\rangle + \frac{\alpha(1+\mu^{m-1})^2H^2}{2} \\ &+ \frac{1}{\alpha}\sum_{h=1}^{H}\left[D\left(\pi_h^{\star,m-1} \mid \pi_h^{m-1}\right) - D\left(\pi_h^{\star,m-1} \mid \pi_h^{m}\right)\right]\end{aligned} \tag{3.20}$$

Proof. This result follows immediately from the "one-step descent" lemma in Lemma 19 and the fact $Q_{r,h}^{m-1} + \mu^{m-1}Q_{g,h}^{m-1} \in [0, (1+\mu^{m-1})H]$.

□

Lemma 9 (Bound for the first term in (3.17))**.** *Let Assumption 5 hold. Then*

$$\begin{aligned}&\sum_{m=1}^{M}\sum_{h=1}^{H}\mathbb{E}_{\pi^\star,\ell_\pi^m,\mathbb{P}^{\ell_\pi^m}}\left[\left\langle Q_{r,h}^m, \pi_h^{\star,m} - \pi_h^m(\cdot \mid x_h)\right\rangle\right] \\ \le &-\sum_{m=1}^{M}\mu^m\sum_{h=1}^{H}\mathbb{E}_{\pi^\star,\ell_\pi^m,\mathbb{P}^{\ell_\pi^m}}\left[\left\langle Q_{g,h}^m, \pi_h^{\star,m} - \pi_h^m\right\rangle\right] + \alpha H^2\sum_{m=1}^{M}(1+|\mu^m|^2) + \frac{1}{\alpha}HML^{-1}\log|\mathcal{A}| + H^2LB_\star\end{aligned}$$

Proof. By further decomposing the first term in (3.17), we obtain

$$\begin{aligned}
&\sum_{m=1}^{M}\sum_{h=1}^{H}\mathbb{E}_{\pi^\star,\ell_\pi^m,\mathbb{P}^{\ell_\pi^m}}\left[\left\langle Q_{r,h}^m,\pi_h^{\star,m}-\pi_h^m(\cdot\mid x_h)\right\rangle\right]\\
&=\sum_{m=1}^{M}\sum_{h=1}^{H}\mathbb{E}_{\pi^\star,\ell_\pi^m,\mathbb{P}^{\ell_\pi^m}}\left[\left\langle Q_{r,h}^m,\pi_h^{\star,\ell_\pi^m}-\pi_h^m(\cdot\mid x_h)\right\rangle\right]+\sum_{m=1}^{M}\sum_{h=1}^{H}\mathbb{E}_{\pi^\star,\ell_\pi^m,\mathbb{P}^{\ell_\pi^m}}\left[\left\langle Q_{r,h}^m,\pi_h^{\star,m}-\pi_h^{\star,\ell_\pi^m}(\cdot\mid x_h)\right\rangle\right]
\end{aligned}\tag{3.21}$$

where $\pi_h^{\star,\ell_\pi^m}$ is the optimal policy at episode ℓ_π^m.

For the first term in (3.21), we have

$$\begin{aligned}
&\sum_{m=1}^{M}\sum_{h=1}^{H}\mathbb{E}_{\pi^\star,\ell_\pi^m,\mathbb{P}^{\ell_\pi^m}}\left[\left\langle Q_{r,h}^m,\pi_h^{\star,\ell_\pi^m}-\pi_h^m(\cdot\mid x_h)\right\rangle\right]\\
&\le-\sum_{m=1}^{M}\mu^m\sum_{h=1}^{H}\mathbb{E}_{\pi^\star,\ell_\pi^m,\mathbb{P}^{\ell_\pi^m}}\left[\left\langle Q_{g,h}^m,\pi_h^{\star,m}-\pi_h^m\right\rangle\right]+\alpha H^2\sum_{m=1}^{M}(1+|\mu^m|^2)\\
&\quad+\frac{1}{\alpha}\sum_{h=1}^{H}\sum_{\mathcal{E}=1}^{\lceil\frac{M}{L}\rceil}\mathbb{E}_{\pi^\star,(\mathcal{E}-1)L,\mathbb{P}^{(\mathcal{E}-1)L}}\left[\sum_{m=(\mathcal{E}-1)L}^{\mathcal{E}L}D\left(\pi_h^{\star,m}\mid\pi_h^m\right)-D\left(\pi_h^{\star,m}\mid\pi_h^{m+1}\right)\right]\\
&\le-\sum_{m=1}^{M}\mu^m\sum_{h=1}^{H}\mathbb{E}_{\pi^\star,\ell_\pi^m,\mathbb{P}^{\ell_\pi^m}}\left[\left\langle Q_{g,h}^m,\pi_h^{\star,m}-\pi_h^m\right\rangle\right]+\alpha H^2\sum_{m=1}^{M}(1+|\mu^m|^2)\\
&\quad+\frac{1}{\alpha}\sum_{h=1}^{H}\sum_{\mathcal{E}=1}^{\lceil\frac{M}{L}\rceil}\mathbb{E}_{\pi^\star,(\mathcal{E}-1)L,\mathbb{P}^{(\mathcal{E}-1)L}}\left[D\left(\pi_h^{\star,(\mathcal{E}-1)L}\mid\pi_h^{(\mathcal{E}-1)L}\right)\right]\\
&\le-\sum_{m=1}^{M}\mu^m\sum_{h=1}^{H}\mathbb{E}_{\pi^\star,\ell_\pi^m,\mathbb{P}^{\ell_\pi^m}}\left[\left\langle Q_{g,h}^m,\pi_h^{\star,m}-\pi_h^m\right\rangle\right]+\alpha H^2\sum_{m=1}^{M}(1+|\mu^m|^2)+\frac{1}{\alpha}HML^{-1}\log|\mathcal{A}|
\end{aligned}$$

where the first inequality follows from Lemma 8 and the fact that $(1+|\mu^m|)^2\le 2+2|\mu^m|^2$, the second inequality results from the telescoping, and the last inequality is due to

$$D\left(\pi_h^{\star,(\mathcal{E}-1)L}\mid\pi_h^{(\mathcal{E}-1)L}\right)=\sum_{a\in\mathcal{A}}\pi_h^{\star,(\mathcal{E}-1)L}\cdot\log\left(|\mathcal{A}|\cdot\pi_h^{\star,(\mathcal{E}-1)L}\right)\le\log|\mathcal{A}|.$$

For the second term in (3.21), it holds that

$$\begin{aligned}
&\sum_{m=1}^{M}\sum_{h=1}^{H}\mathbb{E}_{\pi^\star,\ell_\pi^m,\mathbb{P}^{\ell_\pi^m}}\left[\left\langle Q_{r,h}^m,\pi_h^{\star,m}(\cdot\mid x_h)-\pi_h^{\star,\ell_\pi^m}(\cdot\mid x_h)\right\rangle\right]\\
\leq&\sum_{m=1}^{M}\sum_{h=1}^{H}\mathbb{E}_{\pi^\star,\ell_\pi^m,\mathbb{P}^{\ell_\pi^m}}\left[H\left\|\pi_h^{\star,m}(\cdot\mid x_h)-\pi_h^{\star,\ell_\pi^m}(\cdot\mid x_h)\right\|_1\right]\\
\leq&\sum_{m=1}^{M}\sum_{h=1}^{H}H\max_{x_h\in\mathcal{S}}\left\|\pi_h^{\star,m}(\cdot\mid x_h)-\pi_h^{\star,\ell_\pi^m}(\cdot\mid x_h)\right\|_1\\
\leq&\sum_{\mathcal{E}=1}^{\lceil\frac{M}{L}\rceil}\sum_{h=1}^{H}\sum_{m=(\mathcal{E}-1)L+1}^{\mathcal{E}L}HB_{\star,\mathcal{E}}\\
\leq&H^2LB_\star
\end{aligned}$$

where the first policy holds by Holder's inequality and the fact that $\left\|Q_h^m(s,\cdot)\right\|_\infty\leq H$, the third step is due to the definition of $B_{\star,\mathcal{E}}=\sum_{m=(\mathcal{E}-1)L+1}^{\mathcal{E}L}\sum_{h=1}^{H}\left\|\pi_h^{\star,m}-\pi_h^{\star,m-1}\right\|_\infty$ and the last inequality follows from the definition of $B_\star=\sum_{m=1}^{M}\sum_{h=1}^{H}\left\|\pi_h^{\star,m}-\pi_h^{\star,m-1}\right\|_\infty$. This completes the proof. □

Lemma 10 (Bound for the second term in (3.17))**.** *Let Assumption 5 hold. Then*

$$\begin{aligned}
&\sum_{m=1}^{M}\sum_{h=1}^{H}\left(\mathbb{E}_{\pi^\star,m,\mathbb{P}^m}-\mathbb{E}_{\pi^\star,\ell_\pi^m,\mathbb{P}^{\ell_\pi^m}}\right)\cdot\left[\left\langle Q_{r,h}^m(x_h,\cdot),\pi_h^{\star,m}(\cdot\mid x_h)-\pi_h^m(\cdot\mid x_h)\right\rangle\right]\\
&\leq 2H^2L\left(\sqrt{d_1}B_{\mathbb{P}}+B_\star\right).
\end{aligned}$$

Proof. We denote by $\mathbb{1}(x_h)$ the indicator function for state x_h. It holds that

$$\begin{aligned}
&\sum_{m=1}^{M}\sum_{h=1}^{H}\left(\mathbb{E}_{\pi^\star,m,\mathbb{P}^m}-\mathbb{E}_{\pi^\star,\ell_\pi^m,\mathbb{P}^{\ell_\pi^m}}\right)\cdot\left[\left\langle Q_{r,h}^m(x_h,\cdot),\pi_h^{\star,m}(\cdot\mid x_h)-\pi_h^m(\cdot\mid x_h)\right\rangle\right]\\
\leq&\sum_{m=1}^{M}\sum_{h=1}^{H}\left(\mathbb{E}_{\pi^\star,m,\mathbb{P}^m}-\mathbb{E}_{\pi^\star,\ell_\pi^m,\mathbb{P}^{\ell_\pi^m}}\right)\left[2H\mathbb{1}(x_h)\right]\\
=&2H\sum_{\mathcal{E}=1}^{\lceil\frac{M}{L}\rceil}\sum_{m=(\mathcal{E}-1)L+1}^{\mathcal{E}L}\sum_{h=1}^{H}\sum_{j=(\mathcal{E}-1)L+2}^{m}\left(\mathbb{E}_{\pi^\star,j,\mathbb{P}^j}-\mathbb{E}_{\pi^\star,j-1,\mathbb{P}^{j-1}}\right)\left[\mathbb{1}(x_h)\right]\\
\leq&2HL\sum_{\mathcal{E}=1}^{\lceil\frac{M}{L}\rceil}\sum_{h=1}^{H}\sum_{j=(\mathcal{E}-1)L+1}^{\mathcal{E}L}\left(\mathbb{E}_{\pi^\star,j,\mathbb{P}^j}-\mathbb{E}_{\pi^\star,j-1,\mathbb{P}^{j-1}}\right)\left[\mathbb{1}(x_h)\right]\\
\leq&2H^2L\left(\sqrt{d_1}B_{\mathbb{P}}+B_\star\right)
\end{aligned}$$

where the first step follows from $\left|\left\langle Q_{r,h}^m(x_h,\cdot),\pi_h^{\star,m}(\cdot\mid x_h)-\pi_h^m(\cdot\mid x_h)\right\rangle\right|\leq 2H\mathbb{1}(x_h)$, the third steps holds by telescoping, and the last step follows from Lemma 21. This completes the proof. □

Lemma 11 (Dual step for dynamic regret)**.** *It holds that*

$$-\sum_{m=1}^{M}\mu^m\left(V_{g,1}^{\pi^{\star,m},m}(x_1)-V_{g,1}^{m}(x_1)\right)\le\eta H^2(M+1)+\sum_{m=1}^{M+1}(\eta\xi^2-\xi)|\mu^{m-1}|^2.$$

Proof. By the dual update in line 14 in Algorithm 1 and $\chi=\infty$, we have

$$\begin{aligned}
0&\le\left(\mu^{m+1}\right)^2\\
&=\sum_{m=1}^{M+1}\left((\mu^m)^2-\left(\mu^{m-1}\right)^2\right)\\
&=\sum_{m=1}^{M+1}\left(\left(\mu^{m-1}+\eta\left(b^m-\xi\mu^{m-1}-V_{g,1}^{m-1}(x_1)\right)\right)\right)^2-\left(\mu^{m-1}\right)^2\\
&\le\sum_{m=1}^{M+1}2\eta\mu^{m-1}\left(V_{g,1}^{\pi^{\star,m-1}}(x_1)-\xi\mu^{m-1}-V_{g,1}^{m-1}(x_1)\right)+\eta^2\left(b^m-\xi\mu^{m-1}-V_{g,1}^{m-1}(x_1)\right)^2
\end{aligned}$$

where we use the feasibility of $\pi^{\star,m-1}$ in the last inequality. Since $\mu^0=0$ and $\left|b^m-V_{g,1}^{m-1}(x_1)\right|\le H$, the above inequality implies that

$$\begin{aligned}
&-\sum_{m=1}^{M}\mu^{m-1}\left(V_{g,1}^{\pi^{\star,m},m}(x_1)-V_{g,1}^{m}(x_1)\right)\\
\le&\sum_{m=1}^{M+1}\frac{\eta}{2}\left(b^m-\xi\mu^{m-1}-V_{g,1}^{m-1}(x_1)\right)^2-\sum_{m=1}^{M+1}\xi|\mu^{m-1}|^2 &(3.22)\\
\le&\sum_{m=1}^{M+1}\eta\left(b^m-V_{g,1}^{m-1}(x_1)\right)^2+\sum_{m=1}^{M+1}(\eta\xi^2-\xi)|\mu^{m-1}|^2\\
\le&\eta H^2(M+1)+\sum_{m=1}^{M+1}(\eta\xi^2-\xi)|\mu^{m-1}|^2. &(3.23)
\end{aligned}$$

This completes the proof. □

Lemma 12 (Model prediction error bound for dynamic regret)**.** *Let Assumption 5 and 6 hold. Fix $p\in(0,1)$ and let $\mathcal{E}$ be the epoch that the episode m belongs to. If we set $\lambda=1$, $LV=B_{\mathbb{P},\mathcal{E}}H^2d_1\sqrt{d_1W}+B_{g,\mathcal{E}}\sqrt{d_2W}$, and*

$$\begin{aligned}
\Gamma_h^m(\cdot,\cdot)&=\beta\left(\varphi(\cdot,\cdot)^\top\left(\Lambda_h^m\right)^{-1}\varphi(\cdot,\cdot)\right)^{1/2},\\
\Gamma_{r,h}^m&=\beta\left((\phi_{r,h}^m)^\top(\Lambda_{r,h}^m)^{-1}\phi_{r,h}^m\right)^{1/2},\\
\Gamma_{g,h}^m&=\beta\left((\phi_{g,h}^m)^\top(\Lambda_{g,h}^m)^{-1}\phi_{g,h}^m\right)^{1/2}
\end{aligned}$$

with $\beta = C_1\sqrt{dH^2\log(dW/p)}$ *in Algorithm 2 , then with probability at least* $1-p/2$ *it holds that*

$$\begin{aligned}
&\sum_{m=1}^{M}\sum_{h=1}^{H}\left(\mathbb{E}_{\pi^{\star,m},\mathbb{P}^m}\left[\iota_{r,h}^m(x_h,a_h)+\mu^m\iota_{g,h}^m(x_h,a_h)\right]-\iota_{r,h}^m(x_h^m,a_h^m)\right)\\
\leq& C_2dH^2MW^{-\frac{1}{2}}\sqrt{\log\left(dH^2W+1\right)\log\left(\frac{dW}{p}\right)}+B_{\mathbb{P}}H^3d_1W\sqrt{d_1W}+B_rHW\sqrt{d_2W}
\end{aligned}$$

where C_1 *and* C_2 *are absolute constants.*

Proof. By Lemma 17, for every $(m,h)\in[M]\times[H]$ and $(x,a)\in\mathcal{S}\times\mathcal{A}$, the following inequality holds with probability at least 1-p/2:

$$\begin{aligned}
-2\left(\Gamma_h^m+\Gamma_{r,h}^m\right)(x,a)-B_{\mathbb{P},\mathcal{E}}H^2d_1\sqrt{d_1W}-B_{r,\mathcal{E}}\sqrt{d_2W}\\
\leq\iota_{r,h}^m(x,a)\leq B_{\mathbb{P},\mathcal{E}}H^2d_1\sqrt{d_1W}+B_{r,\mathcal{E}}\sqrt{d_2W}.
\end{aligned}$$

By the definition of $\iota_{r,h}^m(x,a)$, we have $\left|\iota_{r,h}^m(x,a)\right|\leq 2H$. Hence, it holds with probability at least $1-p/2$ that

$$\begin{aligned}
&\mathbb{E}_{\pi^{\star,m},\mathbb{P}^m}\left[\iota_{r,h}^m(x_h,a_h)\right]-\iota_{r,h}^m(x,a)\\
\leq&2\min\left(H,\left(\Gamma_h^m+\Gamma_{r,h}^m\right)(x,a)+B_{\mathbb{P},\mathcal{E}}H^2d_1\sqrt{d_1W}+B_{r,\mathcal{E}}\sqrt{d_2W}\right)
\end{aligned}$$

for every $(m,h)\in[M]\times[H]$ and $(x,a)\in\mathcal{S}\times\mathcal{A}$, where $\Gamma_h^m(\cdot,\cdot)=\beta\left(\varphi(\cdot,\cdot)^\top\left(\Lambda_h^m\right)^{-1}\varphi(\cdot,\cdot)\right)^{1/2}$ and $\Gamma_{r,h}^m(\cdot,\cdot)=\beta\left(\phi_{r,h}^m(\cdot,\cdot)^\top\left(\Lambda_{r,h}^m\right)^{-1}\phi_{r,h}^m(\cdot,\cdot)\right)^{1/2}$. Therefore, we have

$$\begin{aligned}
&\sum_{m=1}^{M}\sum_{h=1}^{H}\left(\mathbb{E}_{\pi^{\star,m},\mathbb{P}^m}\left[\iota_{r,h}^m(x_h,a_h)\mid x_1\right]-\iota_{r,h}^m(x_h^m,a_h^m)\right)\\
\leq&2\sum_{m=1}^{M}\sum_{h=1}^{H}\min\left(H,\left(\Gamma_h^m+\Gamma_{r,h}^m\right)(x_h^m,a_h^m)+B_{\mathbb{P},\mathcal{E}}H^2d_1\sqrt{d_1W}+B_{r,\mathcal{E}}\sqrt{d_2W}\right)\\
\leq&2\sum_{m=1}^{M}\sum_{h=1}^{H}\min\left(H,\left(\Gamma_h^m+\Gamma_{r,h}^m\right)(x_h^m,a_h^m)\right)+2\sum_{\mathcal{E}=1}^{\lceil\frac{M}{W}\rceil}\left(B_{\mathbb{P},\mathcal{E}}H^3d_1W\sqrt{d_1W}+B_{r,\mathcal{E}}HW\sqrt{d_2W}\right)\\
\leq&2\sum_{m=1}^{M}\sum_{h=1}^{H}\min\left(H,\left(\Gamma_h^m+\Gamma_{r,h}^m\right)(x_h^m,a_h^m)\right)+2B_{\mathbb{P}}H^3d_1W\sqrt{d_1W}+2B_rHW\sqrt{d_2W}
\end{aligned}$$

where the last inequality follows from the definition of the variation budgets $B_{\mathbb{P}}\coloneqq\sum_{\mathcal{E}=1}^{\lceil\frac{M}{W}\rceil}B_{\mathbb{P},\mathcal{E}}$

and $B_r \coloneqq \sum_{\mathcal{E}=1}^{\lceil \frac{M}{W} \rceil} B_{r,\mathcal{E}}$. It results from the Cauchy-Schwartz inequality that

$$\begin{aligned}
&\sum_{m=1}^{M} \sum_{h=1}^{H} \min\left(H, \left(\Gamma_h^m + \Gamma_{r,h}^m\right)\left(x_h^m, a_h^m\right)\right) \\
&= \sum_{\mathcal{E}=1}^{\lceil \frac{M}{W} \rceil} \sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W} \sum_{h=1}^{H} \min\left(H, \left(\Gamma_h^m + \Gamma_{r,h}^m\right)\left(x_h^m, a_h^m\right)\right) \\
&\leq \beta \sum_{\mathcal{E}=1}^{\lceil \frac{M}{W} \rceil} \sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W} \sum_{h=1}^{H} \min\Bigg(H/\beta, \left(\varphi\left(x_h^m, a_h^m\right)^\top \left(\Lambda_h^m\right)^{-1} \varphi\left(x_h^m, a_h^m\right)\right)^{1/2} \\
&+ \left(\phi_{r,h}^m\left(x_h^m, a_h^m\right)^\top \left(\Lambda_{r,h}^m\right)^{-1} \phi_{r,h}^m\left(x_h^m, a_h^m\right)\right)^{1/2}\Bigg)
\end{aligned} \tag{3.24}$$

Since we take $\beta = C_1\sqrt{dH^2 \log(dW/p)}$ with $C_1 > 1$, we have $H/\beta \leq 1$. It remains to apply Lemma 18. First, for every $h \in [H]$ it holds that

$$\sum_{m=1}^{M} \phi_{r,h}^m\left(x_h^m, a_h^m\right)^\top \left(\Lambda_{r,h}^m\right)^{-1} \phi_{r,h}^m\left(x_h^m, a_h^m\right) \leq 2 \log\left(\frac{\det\left(\Lambda_{r,h}^{M+1}\right)}{\det\left(\Lambda_{r,h}^1\right)}\right).$$

Due to $\left\|\phi_{r,h}^m\right\| \leq \sqrt{d}H$ in Assumption 5 and $\Lambda_{r,h}^1 = \lambda I$ in Algorithm 2, it is clear that for every $h \in [H]$,

$$\Lambda_{r,h}^{M+1} = \sum_{m=1}^{M} \phi_{r,h}^m\left(x_h^m, a_h^m\right) \phi_{r,h}^m\left(x_h^m, a_h^m\right)^\top + \lambda I \leq \left(dH^2 M + \lambda\right) I.$$

Thus,

$$\log\left(\frac{\det\left(\Lambda_{r,h}^{K+1}\right)}{\det\left(\Lambda_{r,h}^1\right)}\right) \leq \log\left(\frac{\det\left(\left(dH^2 M + \lambda\right) I\right)}{\det(\lambda I)}\right) \leq d \log\left(\frac{dH^2 K + \lambda}{\lambda}\right).$$

Therefore,

$$\sum_{m=1}^{M} \phi_{r,h}^m\left(x_h^m, a_h^m\right)^\top \left(\Lambda_{r,h}^m\right)^{-1} \phi_{r,h}^m\left(x_h^m, a_h^m\right) \leq 2d \log\left(\frac{dH^2 K + \lambda}{\lambda}\right). \tag{3.25}$$

Similarly, one can show that

$$\sum_{m=1}^{M} \varphi\left(x_h^m, a_h^m\right)^\top \left(\Lambda_h^m\right)^{-1} \varphi\left(x_h^m, a_h^m\right) \leq 2d \log\left(\frac{dK + \lambda}{\lambda}\right). \tag{3.26}$$

Applying the Cauchy-Schwartz inequality and the inequalities (3.25) and (3.26) to (3.24) leads to

$$\begin{aligned}
&\sum_{\mathcal{E}=1}^{\lceil \frac{M}{W} \rceil} \sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W} \sum_{h=1}^{H} \min\left(H, \left(\Gamma_h^m + \Gamma_{r,h}^m\right)(x_h^m, a_h^m)\right) \\
&\leq \beta \sum_{\mathcal{E}=1}^{\lceil \frac{M}{W} \rceil} \sum_{h=1}^{H} \min\Bigg(W, \sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W} \left(\varphi(x_h^m, a_h^m)^\top \left(\Lambda_h^m\right)^{-1} \varphi(x_h^m, a_h^m)\right)^{1/2} \\
&+ \left(\phi_{r,h}^m(x_h^m, a_h^m)^\top \left(\Lambda_{r,h}^m\right)^{-1} \phi_{r,h}^m(x_h^m, a_h^m)\right)^{1/2}\Bigg) \\
&\leq \beta \sum_{\mathcal{E}=1}^{\lceil \frac{M}{W} \rceil} \sum_{h=1}^{H} \Bigg(\left(W \sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W} \varphi(x_h^m, a_h^m)^\top \left(\Lambda_h^m\right)^{-1} \varphi(x_h^m, a_h^m)\right)^{1/2} + \\
&\left(W \sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W} \phi_{r,h}^m(x_h^m, a_h^m)^\top \left(\Lambda_{r,h}^m\right)^{-1} \phi_{r,h}^m(x_h^m, a_h^m)\right)^{1/2}\Bigg) \\
&\leq \beta M W^{-\frac{1}{2}} H \left(\left(2d \log\left(\frac{dW+\lambda}{\lambda}\right)\right)^{1/2} + \left(2d \log\left(\frac{dH^2W+\lambda}{\lambda}\right)\right)^{1/2}\right).
\end{aligned}$$

Therefore, by setting $\lambda = 1$, we have

$$\begin{aligned}
&\sum_{m=1}^{M} \sum_{h=1}^{H} \left(\mathbb{E}_{\pi^\star, m, \mathbb{P}^m}\left[\iota_{r,h}^m(x_h, a_h) \mid x_1\right] - \iota_{r,h}^m(x_h^m, a_h^m)\right) \\
\leq & C_2 d H^2 M W^{-\frac{1}{2}} \sqrt{\log\left(dH^2W+1\right) \log\left(\frac{dW}{p}\right)} + 2B_{\mathbb{P}} H^3 d_1 W \sqrt{d_1 W} + 2B_r H W \sqrt{d_2 W}
\end{aligned}$$

where C_2 is some constant. In addition, by Lemma 17, for every $(m,h) \in [M] \times [H]$ and $(x,a) \in \mathcal{S} \times \mathcal{A}$, the following inequality holds with probability at least $1 - p/2$:

$$-2\left(\Gamma_h^m + \Gamma_{r,h}^m\right)(x,a) - 2B_{\mathbb{P},\mathcal{E}} H^2 d_1 \sqrt{d_1 W} - 2B_{r,\mathcal{E}} \sqrt{d_2 W} \leq \iota_{r,h}^m(x,a) \leq 0.$$

Thus, it holds that

$$\sum_{m=1}^{M} \sum_{h=1}^{H} \mathbb{E}_{\pi^\star, m, \mathbb{P}^m}\left[\mu^m \iota_{g,h}^m(x_h, a_h)\right] \leq 0.$$

Finally, by combining the above two inequalities, we obtain the desired result. □

Lemma 13 (Martingale bound for dynamic regret)**.** *Fix $p \in (0,1)$. In Algorithm 1, it holds with probability at least $1 - p/2$ that*

$$\left|S_{r,H,2}^M\right| \leq 4\sqrt{H^2 T \log\left(\frac{4}{p}\right)} \tag{3.27}$$

where $T = HM$.

Proof. In the expansion of the term (R.III) in (3.19) and Lemma 15, we introduce the following martingale:

$$S^M_{r,H,2} = \sum_{m=1}^{M}\sum_{h=1}^{H}\left(D^m_{r,h,1} + D^m_{r,h,2}\right)$$

where

$$D^m_{r,h,1} = \left(\mathcal{I}^m_h\left(Q^m_{r,h} - Q^{\pi^m,m}_{r,h}\right)\right)(x^m_h) - \left(Q^m_{r,h} - Q^{\pi^m,m}_{r,h}\right)(x^m_h, a^m_h),$$
$$D^m_{r,h,2} = \left(\mathbb{P}^m_h V^m_{r,h+1} - \mathbb{P}^m_h V^{\pi^m,m}_{r,h+1}\right)(x^m_h, a^m_h) - \left(V^m_{r,h+1} - V^{\pi^m,m}_{r,h+1}\right)(x^m_{h+1})$$

and $\left(\mathcal{I}^m_h f\right)(x) \coloneqq \left\langle f(x,\cdot), \pi^m_h(\cdot \mid x)\right\rangle$. Due to the truncation in line 10 of Algorithm 2, we know that $Q^m_{r,h}, Q^{\pi^m,m}_{r,h}, V^m_{r,h+1}, V^{\pi^m,m}_{r,h+1} \in [0,H]$. This shows that $\left|D^m_{r,h,1}\right| \le 2H, \left|D^m_{r,h,2}\right| \le 2H$ for all $(m,h) \in [M]\times[H]$. The Azuma-Hoeffding inequality yields that,

$$P\left(\left|S^M_{r,H,2}\right| \ge s\right) \le 2\exp\left(\frac{-s^2}{16H^2T}\right).$$

For $p \in (0,1)$, if we set $s = 4H\sqrt{T\log(4/p)}$, then the inequality (3.27) holds with probability at least $1 - p/2$. □

Proof of dynamic regret in Theorem 7

By combining Lemmas 9 and 10, we can conclude that

$$\begin{aligned}
&\mathrm{DR}(M)\\
\le&\frac{1}{\alpha}HML^{-1}\log|\mathcal{A}| + \alpha H^2\sum_{m=1}^{M}(1+|\mu^m|^2) + H^2LB_\star + 2H^2L\left(\sqrt{d_1}B_{\mathbb{P}} + B_\star\right)\\
&- \sum_{m=1}^{M}\mu^m\sum_{h=1}^{H}\mathbb{E}_{\pi^\star,m,\mathbb{P}^m}\left[\left\langle Q^m_{g,h}(x_h,\cdot), \pi^{\star,m}_h(\cdot\mid x_h) - \pi^m_h(\cdot\mid x_h)\right\rangle\right] + \sum_{m=1}^{M}\sum_{h=1}^{H}\mathbb{E}_{\pi^\star,m,\mathbb{P}^m}\left[\iota^m_{r,h}(x_h,a_h)\right]\\
&- \sum_{m=1}^{M}\sum_{h=1}^{H}\iota^m_{r,h}(x^m_h,a^m_h) + S^M_{r,H,2}.
\end{aligned}$$

Then, by Lemma 20, the above inequality further implies that

$$\begin{aligned}
&\mathrm{DR}(M)\\
\le&\frac{1}{\alpha}HML^{-1}\log|\mathcal{A}| + \alpha H^2\sum_{m=1}^{M}(1+|\mu^m|^2) + H^2LB_\star + 2H^2L\left(\sqrt{d_1}B_{\mathbb{P}} + B_\star\right)\\
&- \sum_{m=1}^{M}\mu^m\left(V^{\pi^\star,m,m}_{g,1}(x_1) - V^m_{g,1}(x_1)\right) + \sum_{m=1}^{M}\sum_{h=1}^{H}\mathbb{E}_{\pi^\star,m,\mathbb{P}^m}\left[\iota^m_{r,h}(x_h,a_h) + \mu^m\iota^m_{g,h}(x_h,a_h)\right]\\
&- \sum_{m=1}^{M}\sum_{h=1}^{H}\iota^m_{r,h}(x^m_h,a^m_h) + S^M_{r,H,2}.
\end{aligned} \tag{3.28}$$

Due to the dual update in Lemma 11, we obtain

$$\begin{aligned}\mathrm{DR}(M) \leq & \frac{1}{\alpha} HML^{-1} \log|\mathcal{A}| + \alpha H^2 M + H^2 L B_\star + 2H^2 L\left(\sqrt{d_1} B_{\mathbb{P}} + B_\star\right) \\ & + \eta H^2 (M+1) + \sum_{m=1}^{M+1} (\alpha H^2 + \eta\xi^2 - \xi)\|\mu^m\|^2 \\ & + \sum_{m=1}^{M} \sum_{h=1}^{H} \mathbb{E}_{\pi^{\star,m},\mathbb{P}^m}\left[\iota_{r,h}^m(x_h,a_h) + \mu^m \iota_{g,h}^m(x_h,a_h)\right] \\ & - \sum_{m=1}^{M} \sum_{h=1}^{H} \iota_{r,h}^m(x_h^m,a_h^m) + S_{r,H,2}^M.\end{aligned} \tag{3.29}$$

Then, by controlling the model prediction error in Lemma 12 and the martingale bound in Lemma 13, we have

$$\begin{aligned}\mathrm{DR}(M) \leq & \frac{1}{\alpha} HML^{-1} \log|\mathcal{A}| + \alpha H^2 M + H^2 L\left(2\sqrt{d_1} B_{\mathbb{P}} + 3B_\star\right) + \eta H^2 (M+1) \\ & + \sum_{m=1}^{M+1} (\alpha H^2 + \eta\xi^2 - \xi)\|\mu^m\|^2 + 4\sqrt{H^2 T \log\left(\frac{4}{p}\right)} \\ & + C_2 d H^2 M W^{-\frac{1}{2}} \sqrt{\log\left(dH^2 W + 1\right) \log\left(\frac{dW}{p}\right)} \\ & + B_{\mathbb{P}} H^3 d_1 W \sqrt{d_1 W} + B_r H W \sqrt{d_2 W}\end{aligned}$$

with probability at least $1-p$. Finally, by setting $\alpha = H^{-1}M^{-\frac{1}{2}}(\sqrt{d}B_\Delta + B_\star)^{\frac{1}{3}}$, $L = M^{\frac{3}{4}}(\sqrt{d}B_\Delta + B_\star)^{-\frac{2}{3}}$, $\eta = M^{-\frac{1}{2}}$, $\xi = 2H(\sqrt{d}B_\Delta + B_\star)^{\frac{1}{3}} M^{-\frac{1}{2}}$, $W = d^{-\frac{1}{4}} H^{-1} M^{\frac{1}{2}} B_\Delta^{-\frac{1}{2}}$, it holds that

$$\text{D-Regret(M)} \leq \widetilde{\mathcal{O}}\left(d^{\frac{9}{8}} H^{\frac{5}{2}} M^{\frac{3}{4}} (\sqrt{d}B_\Delta + B_*)^{\frac{1}{3}}\right)$$

with probability at least $1-p$. This completes the proof.

3.C Proof for Linear Kernel CMDP Case under Assumption 7

Model Prediction Error

Lemma 14 (Model prediction error bound for dynamic regret under uniform Slater condition)**.** *Let Assumption 5 and 7 hold. Fix $p \in (0,1)$ and let $\mathcal{E}$ be the epoch that the episode m belongs to. If we set $\lambda = 1$, $LV = 0$ and*

$$\begin{aligned}\Gamma_h^m(\cdot,\cdot) &= \beta\left(\varphi(\cdot,\cdot)^\top (\Lambda_h^m)^{-1} \varphi(\cdot,\cdot)\right)^{1/2}, \\ \Gamma_{r,h}^m &= \beta\left((\phi_{r,h}^m)^\top (\Lambda_{r,h}^m)^{-1} \phi_{r,h}^m\right)^{1/2}, \\ \Gamma_{g,h}^m &= \beta\left((\phi_{g,h}^m)^\top (\Lambda_{g,h}^m)^{-1} \phi_{g,h}^m\right)^{1/2},\end{aligned}$$

with $\beta = C_1\sqrt{dH^2\log(dW/p)}$ *in Algorithm 2, then with probability at least* $1-p/2$ *it holds that*

$$\begin{aligned}
&\sum_{m=1}^{M}\sum_{h=1}^{H}\left(\mathbb{E}_{\pi^\star,m,\mathbb{P}^m}\left[\iota_{r,h}^m(x_h,a_h)+\mu^m\iota_{g,h}^m(x_h,a_h)\right]-\iota_{r,h}^m(x_h^m,a_h^m)\right)\\
\leq& C_2dH^2MW^{-\frac{1}{2}}\sqrt{\log\left(dH^2W+1\right)\log\left(\frac{dW}{p}\right)}\\
&+(2+\chi)B_{\mathbb{P}}H^3d_1W\sqrt{d_1W}+(2B_r+\chi B_g)HW\sqrt{d_2W}
\end{aligned}$$

where C_1 *and* C_2 *are absolute constants and* $\mu^m\leq\chi$.

Proof. By Lemma 16, for every $(m,h)\in[M]\times[H]$ and $(x,a)\in\mathcal{S}\times\mathcal{A}$, the following inequality holds with probability at least 1-p/2:

$$\begin{aligned}
&-2\left(\Gamma_h^m+\Gamma_{\diamond,h}^m\right)(x,a)-B_{\mathbb{P},\mathcal{E}}H^2d_1\sqrt{d_1W}-B_{\diamond,\mathcal{E}}\sqrt{d_2W}\\
&\leq\iota_{\diamond,h}^m(x,a)\leq B_{\mathbb{P},\mathcal{E}}H^2d_1\sqrt{d_1W}+B_{\diamond,\mathcal{E}}\sqrt{d_2W}.
\end{aligned}$$

for $\diamond=r$ or g. The rest of the proof is similar to Lemma 12 and is thus omitted. □

Proof of Dynamic Regret in Theorem 8

From equation (3.29), we have

$$\begin{aligned}
\mathrm{DR}(M)\leq&\frac{1}{\alpha}HML^{-1}\log|\mathcal{A}|+\alpha H^2\sum_{m=1}^{M}(1+|\mu^m|^2)\\
&+H^2L\left(2\sqrt{d_1}B_{\mathbb{P}}+3B_\star\right)+\eta H^2(M+1)+\sum_{m=1}^{M+1}(\eta\xi^2-\xi)|\mu^{m-1}|^2\\
&+\sum_{m=1}^{M}\sum_{h=1}^{H}\mathbb{E}_{\pi^\star,m,\mathbb{P}^m}\left[\iota_{r,h}^m(x_h,a_h)+\mu^m\iota_{g,h}^m(x_h,a_h)\right]-\sum_{m=1}^{M}\sum_{h=1}^{H}\iota_{r,h}^m(x_h^m,a_h^m)+S_{r,H,2}^M\\
\leq&\frac{1}{\alpha}HML^{-1}\log|\mathcal{A}|+\alpha H^2\sum_{m=1}^{M}(1+\chi^2)+H^2L\left(2\sqrt{d_1}B_{\mathbb{P}}+3B_\star\right)\\
&+\eta H^2(M+1)+\sum_{m=1}^{M+1}(\eta\xi^2-\xi)\chi^2\\
&+\sum_{m=1}^{M}\sum_{h=1}^{H}\mathbb{E}_{\pi^\star,m,\mathbb{P}^m}\left[\iota_{r,h}^m(x_h,a_h)+\mu^m\iota_{g,h}^m(x_h,a_h)\right]-\sum_{m=1}^{M}\sum_{h=1}^{H}\iota_{r,h}^m(x_h^m,a_h^m)+S_{r,H,2}^M
\end{aligned}\tag{3.30}$$

where the second inequality follows from the fact $\mu^m \le \chi$ for all $m \in [M]$ under the uniform Slater condition. Then, by Lemma 14, it holds that

$$\begin{aligned}\mathrm{DR}(M) \le & \frac{1}{\alpha} HML^{-1} \log|\mathcal{A}| + \alpha H^2 \sum_{m=1}^{M} (1+\chi^2) + H^2 L \left(2\sqrt{d_1} B_{\mathbb{P}} + 3B_\star\right) \\ & + \eta H^2 (M+1) + \sum_{m=1}^{M+1} (\eta\xi^2 - \xi)\chi^2 \\ & + C_2 d H^2 M W^{-\frac{1}{2}} \sqrt{\log\left(dH^2W+1\right) \log\left(\frac{dW}{p}\right)} \\ & + (2+\chi) B_{\mathbb{P}} H^3 d_1 W \sqrt{d_1 W} + (2B_r + \chi B_g) HW \sqrt{d_2 W}.\end{aligned} \tag{3.31}$$

Furthermore, by substituting the parameters $\alpha = \gamma H^{-\frac{3}{2}} M^{-\frac{1}{3}} (\sqrt{d} B_\Delta + B_\star)^{\frac{1}{3}}$, $L = M^{\frac{2}{3}} (\sqrt{d} B_\Delta + B_\star)^{-\frac{2}{3}}$, $\eta = M^{-\frac{1}{2}}$, $\xi = 0$, $W = d^{-\frac{1}{4}} H^{-1} M^{\frac{1}{2}} B_\Delta^{-\frac{1}{2}}$, we obtain

$$\mathrm{DR}(M) \le \widetilde{\mathcal{O}}\left(\gamma^{-1} d^{\frac{9}{8}} H^{\frac{5}{2}} M^{\frac{3}{4}} (\sqrt{d} B_\Delta + B_*)^{\frac{1}{3}}\right). \tag{3.32}$$

This completes the proof.

Proof of Constraint Violation in Theorem 8

By the dual update in line 14 in Algorithm 1 and $\xi = 0$, for any $\mu \in [0, \chi]$ we have

$$\begin{aligned}\left|\mu^{m+1} - \mu\right|^2 &= \left|\mathrm{Proj}_{[0,\chi]}\left(\mu^m + \eta\left(b_m - V_{g,1}^m(x_1)\right)\right) - \mathrm{Proj}_{[0,\chi]}(\mu)\right|^2 \\ &\le \left|\mu^m + \eta\left(b_m - V_{g,1}^m(x_1)\right) - \mu\right|^2 \\ &\le \left(\mu^m - \mu\right)^2 + 2\eta\left(b_m - V_{g,1}^m(x_1)\right)\left(\mu^m - \mu\right) + \eta^2 H^2\end{aligned}$$

where we apply the non-expansiveness of projection in the first inequality and $\left|b_m - V_{g,1}^m(x_1)\right| \le H$ for the last inequality. By summing the above inequality from $m = 1$ to $m = M$, we have

$$0 \le \left|\mu^{M+1} - \mu\right|^2 = \left|\mu^1 - \mu\right|^2 + 2\eta \sum_{m=1}^{M} \left(b_m - V_{g,1}^m(x_1)\right)\left(\mu^m - \mu\right) + \eta^2 H^2 M$$

which implies that

$$\begin{aligned}\sum_{m=1}^{M} \left(b_m - V_{g,1}^m(x_1)\right)\left(\mu - \mu^m\right) &\le \frac{1}{2\eta}\left|\mu^1 - \mu\right|^2 + \frac{\eta}{2} H^2 M \\ &\le \frac{1}{2\eta}\mu^2 + \frac{\eta}{2} H^2 M.\end{aligned} \tag{3.33}$$

In addition, from equation (3.28), we obtain

$$\begin{aligned}
&\sum_{m=1}^{M}\left(V_{r,1}^{\pi^{\star,m},m}(x_1)-V_{r,1}^{m}(x_1)\right)+\sum_{m=1}^{M}\mu^m\left(b_m-V_{g,1}^{m}(x_1)\right)\\
\leq&\frac{1}{\alpha}HML^{-1}\log|\mathcal{A}|+\alpha H^2\sum_{m=1}^{M}(1+|\mu^m|^2)+H^2L\left(2\sqrt{d_1}B_{\mathbb{P}}+3B_\star\right) \qquad (3.34)\\
&+\sum_{m=1}^{M}\sum_{h=1}^{H}\mathbb{E}_{\pi^{\star,m},\mathbb{P}^m}\left[\iota_{r,h}^{m}(x_h,a_h)+\mu^m\iota_{g,h}^{m}(x_h,a_h)\right]-\sum_{m=1}^{M}\sum_{h=1}^{H}\iota_{r,h}^{m}(x_h^m,a_h^m)+S_{r,H,2}^{M}\\
\leq&\frac{1}{\alpha}HML^{-1}\log|\mathcal{A}|+\alpha H^2\sum_{m=1}^{M}(1+\chi^2)+H^2L\left(2\sqrt{d_1}B_{\mathbb{P}}+3B_\star\right)\\
&+C_2dH^2MW^{-\frac{1}{2}}\sqrt{\log\left(dH^2W+1\right)\log\left(\frac{dW}{p}\right)}+(2+\chi)B_{\mathbb{P}}H^3d_1W\sqrt{d_1W}\\
&+(2B_r+\chi B_g)HW\sqrt{d_2W}\\
\leq&\widetilde{\mathcal{O}}\left(\gamma^{-1}d^{\frac{9}{8}}H^{\frac{5}{2}}M^{\frac{3}{4}}(\sqrt{d}B_\Delta+B_*)^{\frac{1}{3}}\right),
\end{aligned}$$

where the second inequality follow from Lemma 14 and the last inequality follows by substituting the parameters $\alpha=\gamma H^{-\frac{3}{2}}M^{-\frac{1}{3}}(\sqrt{d}B_\Delta+B_\star)^{\frac{1}{3}}$, $L=M^{\frac{2}{3}}(\sqrt{d}B_\Delta+B_\star)^{-\frac{2}{3}}$, $\eta=M^{-\frac{1}{2}}$, $\xi=0$, $W=d^{-\frac{1}{4}}H^{-1}M^{\frac{1}{2}}B_\Delta^{-\frac{1}{2}}$. Then, by combining the above inequality with (3.33) and setting $\mu=\chi$, it holds that

$$\begin{aligned}
&\sum_{m=1}^{M}\left(V_{r,1}^{\pi^{\star,m},m}(x_1)-V_{r,1}^{m}(x_1)\right)+\sum_{m=1}^{M}\chi\left(b_m-V_{g,1}^{m}(x_1)\right)\\
\leq&\widetilde{\mathcal{O}}\left(\gamma^{-1}d^{\frac{9}{8}}H^{\frac{5}{2}}M^{\frac{3}{4}}(\sqrt{d}B_\Delta+B_*)^{\frac{1}{3}}+M^{\frac{1}{2}}\chi^2\right).
\end{aligned}$$

Finally, by Corollary 22, we obtain

$$\left[\sum_{m=1}^{M}b_m-V_{g,1}^{\pi^m,m}(x_1)\right]_+\leq\widetilde{\mathcal{O}}\left(\gamma^{-1}d^{\frac{9}{8}}H^{\frac{5}{2}}M^{\frac{3}{4}}(\sqrt{d}B_\Delta+B_*)^{\frac{1}{3}}\right).$$

This completes the proof.

3.D Auxiliary Lemmas

Model Prediction Error

We first show that the prediction error in the value function can be expanded as the summation of the model prediction error and a martingale.

Lemma 15 (Value prediction error expansion, Lemma 26 in [35])**.** *It holds that*

$$\sum_{m=1}^{M}\left(V_{r,1}^{m}(x_1)-V_{r,1}^{\pi^m,m}(x_1)\right)=-\sum_{m=1}^{M}\sum_{h=1}^{H}\iota_{r,h}^{m}(x_h^m,a_h^m)+S_{r,H,2}^{M}.$$

Lemma 16 (Lemma 30 in [35]). *Let Assumption 5 hold. Fix $p \in (0,1)$ and let $\mathcal{E}$ be the epoch that the episode m belongs to. If we set $\lambda = 1$, $LV = 0$ and*

$$\begin{aligned}
\Gamma_h^m(\cdot,\cdot) &= \beta\left(\varphi(\cdot,\cdot)^\top \left(\Lambda_h^m\right)^{-1} \varphi(\cdot,\cdot)\right)^{1/2}, \\
\Gamma_{r,h}^m &= \beta\left((\phi_{r,h}^m)^\top (\Lambda_{r,h}^m)^{-1} \phi_{r,h}^m\right)^{1/2}, \\
\Gamma_{g,h}^m &= \beta\left((\phi_{g,h}^m)^\top (\Lambda_{g,h}^m)^{-1} \phi_{g,h}^m\right)^{1/2},
\end{aligned}$$

with $\beta = C_1\sqrt{dH^2\log(dW/p)}$ in Algorithm 2, then it holds that

$$\begin{aligned}
&-2\left(\Gamma_h^m + \Gamma_{\diamond,h}^m\right)(x,a) - B_{\mathbb{P},\mathcal{E}} H^2 d_1 \sqrt{d_1 W} - B_{\diamond,\mathcal{E}}\sqrt{d_2 W} \\
&\le \iota_{\diamond,h}^m(x,a) \le B_{\mathbb{P},\mathcal{E}} H^2 d_1 \sqrt{d_1 W} + B_{\diamond,\mathcal{E}}\sqrt{d_2 W}
\end{aligned}$$

with probability at least $1 - p/2$ for every $(m,h) \in [M] \times [H]$ and $(x,a) \in \mathcal{S} \times \mathcal{A}$, where the symbol $\diamond$ is equal to r or g.

Lemma 17 (Lemma 31 in [35]). *Let Assumptions 5 and 6 hold. Fix $p \in (0,1)$ and let $\mathcal{E}$ be the epoch that the episode m belongs to. If we set $\lambda = 1$, $LV = B_{\mathbb{P},\mathcal{E}} H^2 d_1 \sqrt{d_1 W} + B_{g,\mathcal{E}}\sqrt{d_2 W}$, and*

$$\begin{aligned}
\Gamma_h^m(\cdot,\cdot) &= \beta\left(\varphi(\cdot,\cdot)^\top \left(\Lambda_h^m\right)^{-1} \varphi(\cdot,\cdot)\right)^{1/2}, \\
\Gamma_{r,h}^m &= \beta\left((\phi_{r,h}^m)^\top (\Lambda_{r,h}^m)^{-1} \phi_{r,h}^m\right)^{1/2}, \\
\Gamma_{g,h}^m &= \beta\left((\phi_{g,h}^m)^\top (\Lambda_{g,h}^m)^{-1} \phi_{g,h}^m\right)^{1/2}
\end{aligned}$$

with $\beta = C_1\sqrt{dH^2\log(dW/p)}$ in Algorithm 2, then it holds that

$$\begin{aligned}
&-2\left(\Gamma_h^m + \Gamma_{r,h}^m\right)(x,a) - B_{\mathbb{P},\mathcal{E}} H^2 d_1 \sqrt{d_1 W} - B_{r,\mathcal{E}}\sqrt{d_2 W} \\
&\le \iota_{r,h}^m(x,a) \le B_{\mathbb{P},\mathcal{E}} H^2 d_1 \sqrt{d_1 W} + B_{r,\mathcal{E}}\sqrt{d_2 W} \\
&-2\left(\Gamma_h^m + \Gamma_{g,h}^m\right)(x,a) - 2B_{\mathbb{P},\mathcal{E}} H^2 d_1 \sqrt{d_1 W} - 2B_{g,\mathcal{E}}\sqrt{d_2 W} \\
&\le \iota_{g,h}^m(x,a) \le 0
\end{aligned}$$

with probability at least $1 - p/2$ for every $(m,h) \in [M] \times [H]$ and $(x,a) \in \mathcal{S} \times \mathcal{A}$.

Lemma 18 (Elliptical Potential Lemma, Lemma D.2 in [70] or [21]). *Let $\{\phi_t\}_{t=1}^\infty$ be a sequence of functions in $\mathbb{R}^d$ and $\Lambda_0 \in \mathbb{R}^{d\times d}$ be a positive definite matrix. Let $\Lambda_t = \Lambda_0 + \sum_{i=1}^{t-1} \phi_i \phi_i^\top$. Assume that $\|\phi_t\|_2 \le 1$ and $\lambda_{\min}(\Lambda_0) \ge 1$. For every $t \ge 1$, it holds that*

$$\log\left(\frac{\det(\Lambda_{t+1})}{\det(\Lambda_1)}\right) \le \sum_{i=1}^{t} \phi_i^\top \Lambda_i^{-1} \phi_i \le 2\log\left(\frac{\det(\Lambda_{t+1})}{\det(\Lambda_1)}\right).$$

Performance Difference Lemmas

Lemma 19 (One-step descent lemma, Lemma 3.3 in [21]). *For every two distributions $\pi^\star$ and π supported on $\mathcal{A}$, state $s \in \mathcal{S}$ and function $Q : \mathcal{S} \times \mathcal{A} \to [0, H]$, it holds that for a distribution π' supported on $\mathcal{A}$ with $\pi'(\cdot) \propto \pi(\cdot) \cdot \exp\{\alpha Q(s, \cdot)\}$ we have*

$$\langle Q(s,\cdot), \pi^\star(\cdot) - \pi(\cdot)\rangle \le \frac{1}{2}\alpha H^2 + \frac{1}{\alpha}\left[D(\pi^\star(\cdot) \mid \pi(\cdot)) - D(\pi^\star(\cdot) \mid \pi'(\cdot))\right].$$

We then introduce a variation of the performance difference lemma with the model prediction error.

Lemma 20 (Performance difference lemma with model prediction error, Lemma 26 in [35]). *For $\diamond = r$ or g, it holds that*

$$\begin{aligned} &V_{\diamond,1}^{\pi^\star,m}(x_1) - V_{\diamond,1}^{m}(x_1) \\ &= \sum_{h=1}^{H} \mathbb{E}_{\pi^\star,m,\mathbb{P}^m}\left[\left\langle Q_{\diamond,h}^{m}(x_h,\cdot), \pi_h^{\star,m}(\cdot \mid x_h) - \pi_h^{m}(\cdot \mid x_h)\right\rangle\right] + \sum_{h=1}^{H} \mathbb{E}_{\pi^\star,m,\mathbb{P}^m}\left[\iota_{\diamond,h}^{m}(x_h,a_h)\right]. \end{aligned}$$

Smoothness Property for The Visitation Measure

We recall the operator $\left(\widetilde{\mathcal{I}}_h^{\star,m} f\right)(x) = \left\langle f(x,\cdot), \pi_h^{\star,m}(\cdot \mid x)\right\rangle$ and note that $\left(\widetilde{\mathcal{I}}_h^{\star,m}\mathbb{P}_h^m\right)(x' \mid x) = \sum_{a\in\mathcal{A}} \mathbb{P}_h^m(x' \mid x, a)\pi_h(a \mid x)$ is the transition kernel in step h under policy π at the episode m. We fix $h \in [H]$. Under policies $\{\pi_h^m\}_{h=1}^H$, the distribution of x_h conditional on x_1 is given by

$$\widetilde{\mathcal{I}}_1^{\star,m}\mathbb{P}_1^m\widetilde{\mathcal{I}}_2^{\star,m}\mathbb{P}_2^m \cdots \widetilde{\mathcal{I}}_{h-1}^{\star,m}\mathbb{P}_{h-1}^m(x_h \mid x_1) \coloneqq \sum_{x_2,\ldots,x_{h-1}} \prod_{i\in[h-1]} \left(\widetilde{\mathcal{I}}_i^{\star,m}\mathbb{P}_i^m\right)(x_{i+1} \mid x_i).$$

We have the following smoothness property for the visitation measure $\widetilde{\mathcal{I}}_1^{\star,m}\mathbb{P}_1^m\widetilde{\mathcal{I}}_2^{\star,m}\mathbb{P}_2^m \cdots \widetilde{\mathcal{I}}_{h-1}^{\star,m}\mathbb{P}_{h-1}^m(x_h \mid x_1)$.

Lemma 21 (Lemma 41 in [35]). *Under Assumption 5, it holds that*

$$\sum_{m=2}^{M}\sum_{h=1}^{H}\left(\mathbb{E}_{\pi^\star,m,\mathbb{P}^m} - \mathbb{E}_{\pi^\star,m-1,\mathbb{P}^{m-1}}\right)[\mathbb{1}(x_h)] \le H\left(\sqrt{d_1}B_{\mathbb{P}} + B_\star\right),$$

where $\mathbb{1}(x_h)$ denotes the indicator function for the state x_h and $d_1 > 0$ is a constant defined in Assumption 5.

Constraints Violation under Uniform Slater Condition

Lemma 22 (Constraint Violation under Uniform Slater Condition, Lemma 47 in [35]). *Let the uniform Slater condition hold and $\mu^{\star,m} \in \Lambda^{\star,m}$. Let $\bar{C}^\star \ge 2\max_{m\in[M]}\mu^{\star,m}$. Assume that $\{\pi^m\}_{m=1}^M$ satisfies*

$$\sum_{m=1}^{M} V_{r,1}^{\pi^\star,m,m}(x_1) - V_{r,1}^{\pi^m,m}(x_1) + \bar{C}^\star\sum_{m=1}^{M}\left(b_m - V_{g,1}^{\pi^m,m}(x_1)\right) \le \delta.$$

Then,

$$\sum_{m=1}^{M} \left(b_m - V_{g,1}^{\pi^m, m}\left(x_1\right) \right) \leq \frac{2\delta}{\bar{C}^\star}.$$

Chapter 4

Non-Stationary Risk-Sensitive RL

Risk-sensitive RL considers problems in which the objective takes into account risks that arise during the learning process, in contrast to the typical expected accumulated reward objective. Effective management of the variability of the return in RL is essential in various applications in finance [88], autonomous driving [52] and human behavior modeling [94].

While classical risk-sensitive RL assumes that an agent interacts with a time-invariant (stationary) environment, both the reward functions and the transition kernels can be time-varying for many risk-sensitive applications. For example, in finance [88], the federal reserve adjusts the interest rate or the balance sheet in a non-stationary way and the market participants should adjust their trading policies accordingly. In the medical treatments [85], the patient's health condition and the sensitivity of the patient's internal body organs to the medicine vary over time. This non-stationarity should be accounted for to minimize the risk of any potential side effects of the treatment. A similar requirement holds for the power grid control [37] where the power grid contingency needs to be prepared with the time-varying electricity loads.

Despite the importance and ubiquity of non-stationary risk-sensitive RL problems, the literature lacks provably efficient algorithms and theoretical results. In this work, we study risk-sensitive RL with an entropic risk measure [65] under episodic Markov decision processes with unknown and time-varying reward functions and state transition kernels.

The non-stationary RL problem with an entropic risk measure has the following technical challenges. (1) Due to the non-stationarity of the model, any estimation error of the expectation operator may be tremendously amplified in the value function when the risk parameter β is small. (2) In addition, the exponential Bellman equation (see Equation (4.2)) used in our risk-sensitive analysis associates the instantaneous reward and value function of the next step in a multiplicative way [47]. However, this multiplicative feature of the exponential Bellman equation will also involve the policy evaluation errors due to the non-stationary drifting as multiplicative terms, which makes it more difficult to gauge the bounds than the risk-neural non-stationary setting in which all policy evaluation errors are in an additive way. (3) Furthermore, the non-linearity of the objective function (see Equation (4.1a)) makes it difficult to obtain an unbiased estimation of the value function, which is

needed in the design of a non-stationary detection mechanism in risk-neutral non-stationary RL [124]. (4) It is unclear whether the risk control and the handling of the non-stationarity can be separately designed when achieving the optimal dynamic regret. To address these difficulties, we develop a novel analysis to carefully quantify the effect of the non-stationarity in risk-sensitive RL. Our main theoretical contributions are as follows

- When the variation budget is known a prior, we propose two provably efficient restart algorithms, namely Restart-RSMB and Restart-RSQ, and establish their dynamic regrets. The stationary version of the model-based method Restart-RSMB is also the first model-based risk-sensitive algorithm in the stationary setting in the literature.
- When the variation budget is unknown (parameter-free), we propose a meta-algorithm that adaptively detects the non-stationarity of the exponential value functions. The proposed adaptive algorithms, namely Adaptive-RSMB and Adaptive-RSQ, can achieve the (almost) same dynamic regret as the algorithms requiring the knowledge of the variation budget.
- We establish a lower bound result for non-stationary RL with entropic risk measure that certifies the near-optimality of our upper bounds.
- Our results also show that the risk control and the handling of the non-stationarity can be separately designed if the variation budget is known a prior, while the non-stationary detection mechanism in the adaptive algorithms depends on the risk parameter.

4.1 Related Work

Many risk-sensitive objectives have been investigated in the literature and applied to RL, such as the entropic risk measure, Markowitz mean-variance model, Value-at-Risk (VaR), and Conditional Value at Risk (CVaR) [91, 28, 30, 77, 31, 117, 118, 65]. Our work is closely related to the entropic risk measure. Following the seminal paper [65], this line of work includes [10, 17, 19, 18, 25, 29, 32, 49, 51, 64, 96, 50, 109, 48, 45, 47]. In particular, when transitions are unknown and simulators of the environment are unavailable, the first non-asymptotic regret guarantees are established under the tabular setting in [48] and the function approximation setting in [45]. Then, a simple transformation of the risk-sensitive Bellman equations is proposed in [47], which leads to improved regret upper bounds. However, the above papers all assume that the environment is stationary, and therefore their results may quickly collapse in a non-stationary environment.

4.2 Problem formulation

Episodic MDP and Risk-Sensitive Objective

In this chapter, we study risk-sensitive RL in non-stationary environments via episodic MDPs with adversarial bandit-information reward feedback and unknown adversarial transition

dynamics. At each episode m, an episodic MDP is defined by the finite state space $\mathcal{S}$, the finite action space $\mathcal{A}$, a collection of transition probability measure $\{\mathcal{P}_h^m\}_{h=1}^H$ specifying the transition probability $\mathcal{P}_h^m(s' \mid s, a)$ from state s to the next state s' under action $a \in \mathcal{A}$, a collection of reward functions $\{r_h^m\}_{h=1}^H$ where $r_h^m : \mathcal{S} \times \mathcal{A} \to [0,1]$, and $H > 0$ as the length of episodes. In this chapter, we focus on a bandit setting where the agent only observes the values of reward functions, i.e., $r_h^m(s_h^m, a_h^m)$ at the visited state-action pair (s_h^m, a_h^m). We also assume that reward functions are deterministic to streamline the presentation, while our analysis readily generalizes to the setting where reward functions are random.

For simplicity, we assume the initial state s_1^m to be fixed as s_1 in different episodes. We use the convention that the episode terminates when a state s_{H+1} at step $H+1$ is reached, at which the agent does not take any further action and receives no reward.

A policy $\pi^m = \left\{\pi_h^m\right\}_{h\in[H]}$ of an agent is a sequence of functions $\pi_h^m : \mathcal{S} \to \mathcal{A}$, where $\pi_h^m(s)$ is the action that the agent takes in state s at step h at episode m. For each $h \in [H]$ and $m \in [M]$, we define the value function $V_h^{\pi,m} : \mathcal{S} \to \mathbb{R}$ of a policy π as the expected value of the cumulative rewards the agent receives under a risk measure of exponential utility by executing π starting from an arbitrary state at step h. Specifically, we have

$$V_h^{\pi,m}(s) \coloneqq \frac{1}{\beta} \log\left\{\mathbb{E}_{\pi,\mathcal{P}^m}\left[\exp\left(\beta \sum_{i=h}^{H} r_i^m\left(s_i, a_i\right)\right) \mid s_h = s\right]\right\}$$

where the expectation $\mathbb{E}_{\pi,\mathcal{P}^m}$ is taken over the random state-action sequence $\{(x_i^m, a_i^m)\}_{i=h}^H$, the action a_i^m follows the policy $\pi_i^m(\cdot \mid x_i^m)$, and the next state x_{i+1} follows the transition dynamics $\mathcal{P}_i^m(\cdot \mid x_i^m, a_i^m)$. Here $\beta \neq 0$ is the risk parameter of the exponential utility: $\beta > 0$ corresponds to a risk-seeking value function, $\beta < 0$ corresponds to a risk-averse value function, and as $\beta \to 0$ the agent tends to be risk-neutral and we recover the classical value function $V_h^{\pi,m}(s) = \mathbb{E}_{\pi,\mathcal{P}^m}\left[\sum_{t=1}^H r_h^m\left(s_t, a_t\right) \mid s_0 = s\right]$ in standard RL.

We further define the action-value function $Q_h^{\pi,m} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, for each $h \in [H]$ and $m \in [M]$, which gives the expected value of the risk measured by the exponential utility when the agent starts from an arbitrary state-action pair and follows the policy π afterwards; that is,

$$\begin{aligned} Q_h^{\pi,m} \coloneqq & \frac{1}{\beta} \log\left\{\exp\left(\beta \cdot r_h^m(s,a)\right) \mathbb{E}\left[\exp\left(\beta \sum_{i=h}^{H} r_i^m\left(s_t, a_t\right)\right) \middle| s_h = s, a_h = a\right]\right\} \\ = & r_h^m(s,a) + \frac{1}{\beta} \log\left\{\mathbb{E}\left[\exp\left(\beta \sum_{i=h+1}^{H} r_i^m\left(s_t, a_t\right)\right) \middle| s_h = s, a_h = a\right]\right\} \end{aligned}$$

for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Under some mild regularity conditions [10], for each episode m, there always exists an optimal policy, denoted as $\pi^{*,m}$, that yields the optimal value $V_h^{\pi^{*,m},m}(s) \coloneqq \sup_\pi V_h^{\pi,m}(s)$ for all $(h,s) \in [H] \times \mathcal{S}$. For convenience, we denote $V_h^{\pi^{*,m},m}(s)$ as $V_h^{*,m}(s)$ when it is clear from the context.

Exponential Bellman Equation

For all $(s, a, h, m) \in \mathcal{S} \times \mathcal{A} \times [H] \times [M]$, the Bellman equation associated with π is given by

$$Q_h^{\pi,m}(s,a) = r_h^m(s,a) + \frac{1}{\beta} \log\left\{\mathbb{E}_{s' \sim P_h^m(\cdot|s,a)}\left[e^{\beta \cdot V_{h+1}^{\pi,m}(s')}\right]\right\}, \tag{4.1a}$$

$$V_h^{\pi,m}(s) = Q_h^{\pi,m}(s, \pi(s)), \quad V_{H+1}^{\pi,m}(s) = 0. \tag{4.1b}$$

In Equation (4.1), it can be seen that the action value $Q_h^{\pi,m}$ of step h is a non-linear function of the value function $V_{h+1}^{\pi,m}$ of the later step. Based on Equation (4.1), for $h \in [H]$ and $m \in [M]$, the Bellman optimality equation is given by

$$Q_h^{*,m}(s,a) = r_h^m(s,a) + \frac{1}{\beta} \log\left\{\mathbb{E}_{s' \sim P_h^m(\cdot|s,a)}\left[e^{\beta \cdot V_{h+1}^{*,m}(s')}\right]\right\},$$

$$V_h^{*,m}(s) = \max_{a \in \mathcal{A}} Q_h^{*,m}(s,a), \quad V_{H+1}^{*,m}(s) = 0.$$

It has been recently shown in [47] that under the risk-sensitive measurement, it is easier to analyze a simple transformation of the Bellman equation (by taking exponential on both sides of (4.1)), which is called *exponential Bellman equation*: for every policy π and tuple (s, a, h, m), we have

$$e^{\beta \cdot Q_h^{\pi,m}(s,a)} = \mathbb{E}_{s' \sim P_h^m(\cdot|s,a)}\left[e^{\beta\left(r_h^m(s,a) + V_{h+1}^{\pi,m}(s')\right)}\right]. \tag{4.2}$$

When $\pi = \pi^{*,m}$, we obtain the corresponding optimality equation

$$e^{\beta \cdot Q_h^{*,m}(s,a)} = \mathbb{E}_{s' \sim P_h^m(\cdot|s,a)}\left[e^{\beta\left(r_h^m(s,a) + V_{h+1}^{*,m}(s')\right)}\right]. \tag{4.3}$$

Note that Equation (4.2) associates the current and future cumulative utilities $\left(Q_h^{\pi,m}$ and $V_{h+1}^{\pi,m}\right)$ in a multiplicative way, rather than in an additive way as in the standard Bellman equations (4.1).

Non-stationarity and Variation Budget

In this work, we focus on a non-stationary environment where the transition function P_h^m and reward functions r_h^m can vary over the episodes. We measure the non-stationarity of the MDP over an interval $\mathcal{I}$ in terms of its variation in the reward functions and transition kernels:

$$B_{r,\mathcal{I}} \coloneqq \sum_{m \in \mathcal{I}} \sum_{h=1}^{H} \sup_{s,a} \left|r_h^m(s,a) - r_h^{m+1}(s,a)\right|,$$

$$B_{\mathcal{P},\mathcal{I}} \coloneqq \sum_{m \in \mathcal{I}} \sum_{h=1}^{H} \sup_{s,a} \left\|\mathcal{P}_h^m(\cdot \mid s,a) - \mathcal{P}_h^{m+1}(\cdot \mid s,a)\right\|_1.$$

Note that our definition of variation only imposes restrictions on the summation of non-stationarity across different episodes, and does not put any restriction on the difference between two steps in the same episode. We further let $B_r \coloneqq B_{r,[1,M]}$, $B_p \coloneqq B_{p,[1,M]}$, and $B \coloneqq B_r + B_p$, and assume $B > 0$.

Performance Metrics

Since both the reward and the transition dynamics vary over the episodes and are revealed only after a policy is decided, the agent aims to ensure the long-term optimality guarantee over some given period of episodes M. Suppose that the agent executes policy π^m in episode m. We now define the dynamic regret as the difference between the total reward value of policy $\{\pi^{\star,m}\}_{m=1}^{M}$ and that of the agent's policy π^m over M episodes:

$$\text{D-Regret}(M) := \sum_{m=1}^{M} \left(V_1^{*,m} - V_1^{\pi^m,m}\right).$$

4.3 Restart Algorithms with The Knowledge of Variation Budget

Periodically Restarted Risk-Sensitive Model-Based Method

We first present the Periodically Restarted Risk-sensitive Model-based method (Restart-RSMB) in Algorithm 3. It consists of two main stages: estimation of value function (line 7-13) with the periodical restart (line 5) and the policy execution (line 15).

To estimate the value function under the unknown non-stationarity, we take the optimistic value evaluation to properly handle the exploration-exploitation trade-off and apply the restart strategy to adapt to the unknown non-stationarity. In particular, we reset the visitation counters $N_h^m(s,a,s')$ and $N_h^m(x,a)$ to zero every W episodes (line 5). Then, the reward and transition dynamics are estimated using only the data from the episode $\ell^m = \left(\lceil \frac{m}{W} \rceil - 1\right)W + 1$ to the episode m by

$$\widehat{\mathcal{P}}_h^m\left(s' \mid s,a\right) = \frac{N_h^m(s,a,s') + \frac{\lambda}{|\mathcal{S}|}}{N_h^m(s,a)+\lambda}, \text{ for all } (s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}, \tag{4.4a}$$

$$\widehat{r}_h^m(s,a) = \frac{\sum_{\tau=\ell^m}^{m-1} 1\{(s,a)=(s_h^\tau,a_h^\tau)\} r_h^\tau(s_h^\tau,a_h^\tau)}{N_h^m(s,a)+\lambda}, \text{ for all } (s,a) \in \mathcal{S}\times\mathcal{A}, \tag{4.4b}$$

which are used to compute the estimated cumulative rewards at step h (line 9). To encourage a sufficient exploration in the uncertain environment, Algorithm 3 applies the counter-based Upper Confidence Bound (UCB). Under the entropic risk measure, this bonus term takes the form

$$\begin{cases} C_1\left(\left(e^{\beta(H-h+1)}-1\right)+e^{\beta(H-h+1)}\beta\right)\sqrt{\frac{|\mathcal{S}|\log(6WH|\mathcal{S}||\mathcal{A}|/p)}{N_h^m(s,a)+1}}, \text{ if } \beta>0, \\ C_1\left(\left(1-e^{\beta(H-h+1)}\right)-\beta\right)\sqrt{\frac{|\mathcal{S}|\log(6WH|\mathcal{S}||\mathcal{A}|/p)}{N_h^m(s,a)+1}}, \text{ if } \beta<0, \end{cases} \tag{4.5}$$

for some constant $C_1 > 1$. Bonus terms of the form (4.5) are called "doubly decaying bonus" since they shrink deterministically and exponentially across the horizon steps due to the term $e^{\beta(H-h+1)}$, apart from decreasing in the visit count. We refer the reader to [45] for more discussion.

Algorithm 3 Periodically Restarted Risk-sensitive Model-based RL (Restart-RSMB)

1: **Inputs:** Time horizon M, restart period W;
2: **for** $m = 1, \ldots, M$ **do**
3: Set the initial state $x_1^m = x_1$ and $\ell^m = (\lceil \frac{m}{W} \rceil - 1)W + 1$;
4: **if** $m = \ell^m$ **then**
5: $Q_h^m(s,a), V_h^m(s) \leftarrow H - h + 1$ if $\beta > 0$, $Q_h^m(s,a), V_h^m(s) \leftarrow 0$ if $\beta < 0$, $N_h^m(s,a) \leftarrow 0, N_h^m(s,a,s') \leftarrow 0$ for all $(s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$;
6: **end if**
7: **for** $h = H, \ldots, 1$ **do**
8: **for** $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
9: $w_h^m(s,a) = \sum_{s'} \hat{\mathcal{P}}_h^m(s' \mid s,a) \left[e^{\beta[\hat{r}_h^m(s,a) + V_{h+1}^m(s')]} \right]$ where $\hat{\mathcal{P}}_h^m$, $\hat{r}_h^m$ are defined in (4.4);
10: $G_h^m(s,a) \leftarrow \begin{cases} \min\left\{ e^{\beta(H-h+1)}, w_h^m(s,a) + \Gamma_h^m(s,a) \right\}, \text{if } \beta > 0; \\ \max\left\{ e^{\beta(H-h+1)}, w_h^m(s,a) - \Gamma_h^m(s,a) \right\}, \text{if } \beta < 0; \end{cases}$ where Γ_h^m is defined in (4.5);
11: $V_h^m(s) \leftarrow \max_{a' \in \mathcal{A}} \frac{1}{\beta} \log G_h^m(s,a')$;
12: **end for**
13: **end for**
14: **for** $h = 1, 2, \ldots, H$ **do**
15: Take an action $a_h^m \leftarrow \arg\max_{a' \in \mathcal{A}} \frac{1}{\beta} \log\{G_h^m(s_h^m, a')\}$, and observe $r_h(s_h^m, a_h^m)$ and s_{h+1}^m;
16: $N_h^m(s_h^m, a_h^m) \leftarrow N_h^m(s_h^m, a_h^m) + 1$; $N_h^m(s_h^m, a_h^m, s_{h+1}^m) \leftarrow N_h^m(s_h^m, a_h^m, s_{h+1}^m,) + 1$;
17: **end for**
18: **end for**

Periodically Restarted Risk-Sensitive Q-Learning

Next, we introduce Periodically Restarted Risk-sensitive Q-learning (Restart-RSQ) in Algorithm 4, which is model-free and inspired by RSQ2 in [47]. Similar to Algorithm 3, we use the optimistic value evaluation to handle the exploration-exploitation trade-off and apply the restart strategy to adapt to the unknown non-stationarity. In particular, we re-initialize the value functions $Q_h^m(s,a), V_h^m(s)$ and reset the visitation counter $N_h^m(x,a)$ to zero every W episodes (line 5). The algorithm then updates the exponential Q values using the Q-learning style update (line 11-12) for the state action pair that just visited (line 8). The learning rate α_t is defined as $\frac{H+1}{H+t}$, which is motivated by [69] and ensures that only the last $\mathcal{O}(\frac{1}{H})$ fraction of samples in each epoch is given non-negligible weights when used to estimate the optimistic Q-values under the non-stationarity. Algorithm 4 also applies the UCB by incorporating a "doubly decaying bonus" term that takes the form

$$\Gamma_{h,t}^m(s_h^m, a_h^m) \leftarrow C_2 \left| e^{\beta(H-h+1)} - 1 \right| \sqrt{\frac{|\mathcal{S}| \log(MH|\mathcal{S}||\mathcal{A}|/\delta)}{t}} \tag{4.6}$$

for some constant $C_2 > 1$.

Algorithm 4 Periodically Restarted Risk-sensitive Q-learning (Restart-RSQ)

Inputs: Time horizon M, restart period W;
for $m = 1, \ldots, M$ **do**
  Set the initial state $x_1^m = x_1$ and $\ell^m = (\lceil \frac{m}{W} \rceil - 1)W + 1$;
  if $m = \ell^m$ **then**
    $Q_h^m(s,a), V_h^m(s) \leftarrow H - h + 1$ if $\beta > 0$, $Q_h^m(s,a), V_h^m(s) \leftarrow 0$ if $\beta < 0$, $N_h^m(s,a) \leftarrow 0$ for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$;
  end if
  for $h = 1, 2, \ldots, H$ **do**
    Take an action $a_h^m \leftarrow \arg\max_{a' \in \mathcal{A}} \frac{1}{\beta} \log\{G_h^m\left(s_h^m, a'\right)\}$, and observe $r_h^m(s_h^m, a_h^m)$ and s_{h+1}^m;
    $N_h^m(s_h^m, a_h^m) \leftarrow N_h^m(s_h^m, a_h^m) + 1$; $t \leftarrow N_h^m(s_h^m, a_h^m)$;
    Set $\alpha_t = \frac{H+1}{H+t}$ and define $\Gamma_{h,t}^m(s_h^m, a_h^m)$ as in (4.6);
    $w_h^m(s_h^m, a_h^m) = (1 - \alpha_t) \cdot G_h(s_h^m, a_h^m) + \alpha_t \cdot \left[e^{\beta[r_h^m(s_h^m, a_h^m) + V_{h+1}^m(s')]}\right]$;
    $G_h^m(s_h^m, a_h^m) \leftarrow \begin{cases} \min\left\{e^{\beta(H-h+1)}, w_h^m(s_h^m, a_h^m) + \alpha_t \Gamma_{h,t}^m(s_h^m, a_h^m)\right\}, & \text{if } \beta > 0; \\ \max\left\{e^{\beta(H-h+1)}, w_h^m(s_h^m, a_h^m) - \alpha_t \Gamma_{h,t}^m(s_h^m, a_h^m)\right\}, & \text{if } \beta < 0; \end{cases}$
    $V_h^m(s_h^m) \leftarrow \max_{a' \in \mathcal{A}} \frac{1}{\beta} \log G_h^m\left(s_h^m, a'\right)$;
  end for
end for

Theoretical Results and Discussions

We now present our main theoretical results for Algorithms 3 and 4.

Theorem 11. *For every $\delta \in (0,1]$, with probability at least $1 - \delta$ there exists a universal constant $c_1 > 0$ (used in Algorithm 3) such that the dynamic regret of Algorithm 3 with $W = M^{\frac{2}{3}} B^{-\frac{2}{3}} |\mathcal{S}|^{\frac{2}{3}} |\mathcal{A}|^{\frac{1}{3}}$ is bounded by*

$$\text{D-Regret(M)} \leq \widetilde{\mathcal{O}}\left(e^{|\beta|H} |\mathcal{S}|^{\frac{2}{3}} |\mathcal{A}|^{\frac{1}{3}} H^2 M^{\frac{2}{3}} B^{\frac{1}{3}}\right).$$

Theorem 12. *For every $\delta \in (0,1]$, with probability at least $1 - \delta$ there exists a universal constant $c_2 > 0$ (used in Algorithm 4) such that the dynamic regret of Algorithm 4 with $W = M^{\frac{2}{3}} H^{-\frac{3}{4}} B^{-\frac{2}{3}} |\mathcal{S}|^{\frac{2}{3}} |\mathcal{A}|^{\frac{1}{3}}$ is bounded by*

$$\text{D-Regret(M)} \leq \widetilde{\mathcal{O}}\left(e^{|\beta|H} |\mathcal{S}|^{\frac{1}{3}} |\mathcal{A}|^{\frac{1}{3}} H^{\frac{9}{4}} M^{\frac{2}{3}} B^{\frac{1}{3}}\right).$$

The proofs of the two theorems are provided in Appendices 4.A and 4.B, respectively. Note that the above results generalize those in the literature of risk-neutral non-stationary RL. In particular, when $\beta \to 0$, we recover the regret bounds with the same dependence on M and B for the restart model-based RL [38] and restart Q-learning [87].

Algorithm 5 Risk-sensitive MALG with Stationary Tests and Restarts (Adaptive-ALG)

1: **Inputs:** ALG and its associated $\rho(\cdot)$, $\hat{n} = \log_2 M + 1$, $\hat{\rho}(m) = 6\hat{n}\log(\frac{M}{\delta})\rho(m)$;
2: **for** $n = 0, 1, \ldots,$ **do**
3: Set $m_n \leftarrow m$ and run MALG-Initialization (Algorithm 6) for the block $[m_n, m_n + 2^n - 1]$;
4: **while** $m < m_n + 2^n$ **do**
5: Identify the unique active instance covering the episode m and denote it as alg;
6: Construct the optimistic estimator g_m for the active instance alg;
7: Follow alg's decision π_m, receive estimated value $R_m = e^{\beta \sum_{h=1}^{H} r_h^m}$, and update alg;
8: Set $U_m = \begin{cases} \min_{\tau \in [m_n, m]} g_\tau, & \text{if } \beta > 0, \\ \max_{\tau \in [m_n, m]} g_\tau, & \text{if } \beta < 0; \end{cases}$
9: Perform **Test1** and **Test2**; Increment $t \leftarrow t + 1$;
10: **If** either test returns *fail*, **then** restart from Line 2.
11: **end while**
12: **end for**
13: **Test1**: Return *fail* if $m = alg.e$ for some order-k alg and
$\begin{cases} \frac{1}{2^k} \sum_{\tau=alg.s}^{alg.e} R_\tau - U_t \geq 9\hat{\rho}(2^k), & \text{if } \beta > 0, \\ U_t - \frac{1}{2^k} \sum_{\tau=alg.s}^{alg.e} R_\tau \geq 9\hat{\rho}(2^k), & \text{if } \beta < 0; \end{cases}$
14: **Test2**: Return *fail* if $\begin{cases} \frac{1}{m-m_n+1} \sum_{\tau=m_n}^{m} (g_\tau - R_\tau) \geq 3\hat{\rho}(m - m_n + 1), & \text{if } \beta > 0, \\ \frac{1}{m-m_n+1} \sum_{\tau=m_n}^{m} (R_\tau - g_\tau) \geq 3\hat{\rho}(m - m_n + 1), & \text{if } \beta < 0, \end{cases}$

4.4 Adaptive Algorithm without The Knowledge of Variation Budget

In Theorems 11 and 12, we need to set the restart period to $W = \mathcal{O}(B^{-\frac{2}{3}} M^{\frac{2}{3}})$, which clearly requires the variation budget B in advance. To overcome this limitation, we propose a meta-algorithm that adaptively detects the non-stationarity without the knowledge of B, while still achieving the similar dynamic regret as in Theorems 11 and 12. In particular, we generalize the black-box approach [124] to the risk-sensitive RL setting and design a non-stationarity detection based on the exponential Bellman equations (4.2).

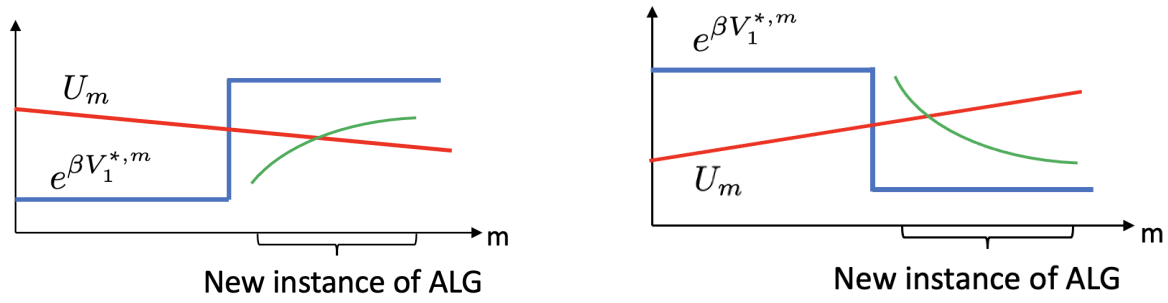
Risk-Sensitive Non-Stationary Detection

We first sketch the high-level idea of the black-box reduction approach for risk-sensitive non-stationary RL with $\beta > 0$. Note that the dynamic regret can be bounded and decomposed

as follows:

$$\text{D-Regret}(M) \leq \underbrace{\frac{1}{\beta}\sum_{m=1}^{M}\left(e^{\beta V_1^{*,m}} - e^{\beta V_1^{m}}\right)}_{\mathbf{R1}} + \underbrace{\frac{1}{\beta}\sum_{m=1}^{M}\left(e^{\beta V_1^{m}} - e^{\beta V_1^{\pi^m,m}}\right)}_{\mathbf{R2}} \tag{4.7}$$

where V_1^m is an UCB-based optimistic estimator of the value function as constructed in Algorithms 3 and 4. In a stationary environment with $\beta > 0$, the base algorithms, such as Algorithms 3 and 4 without the restart mechanism (that is, $W = M$), ensure that **R1** is simply non-positive and **R2** is bounded by $\widetilde{\mathcal{O}}(M^{\frac{1}{2}})$. However, in a non-stationary environment, both terms can be substantially larger. Thus, if we can detect the event that either of the two terms is abnormally larger than the promised bound for a stationary environment, we learn that the environment has changed substantially and should restart the base algorithm. This detection can be easily performed for **R2** since both $e^{\beta V_1^m}$ and $e^{\beta V_1^{\pi^m,m}}$ are observable [1], but not for **R1** since $V_1^{*,m}$ is unknown. To address this issue, we fully utilize the fact that $e^{\beta V_1^m}$ is a UCB-based optimistic estimator to facilitate non-stationary detection.



(a) $\beta > 0$ (b) $\beta < 0$

Figure 4.4.1: An illustration of the risk-sensitive non-stationarity detection. The green curves represent the learner's average performance in new ALG. Since both U_m and learner's average performance depend on the risk-sensitive parameter β in a non-linear way. The non-stationarity detection relies on the choice of β and thus the risk control and the handling of the non-stationarity can not be separately designed.

We illustrate the idea of non-stationary detection for risk-sensitive RL in Figure 4.4.1. Here, the value of $V_1^{*,m}$ drastically increases which results to an increase in $e^{\beta V_1^{*,m}}$ for $\beta > 0$ and an decrease in $e^{\beta V_1^{*,m}}$ for $\beta < 0$. If we start running another instance of base algorithm after this environment change, then its performance will gradually approach due to its regret guarantee in a stationary environment. Since the optimistic estimators should always be an upper bound of the learner's average performance in a stationary environment for

[1] More precisely, $\sum_{m=1}^{M} e^{\beta V_1^{\pi^m,m}}$ can be estimated from $\sum_{m=1}^{M} e^{\beta \sum_{h=1}^{H} r_h^m}$ using the Azuma's inequality.

$\beta > 0$ or a lower bound of the learner's average performance in a stationary environment for $\beta < 0$, if, at some point, we find that the new instance of the base algorithm significantly outperformances/underperformances (depending on the value of β) this quantity, we can infer that the environment has changed.

Multi-Scale ALG (MALG) and Non-Stationarity Tests

To detect the non-stationarity at different scales, we schedule and run instances of the base algorithm ALG in a randomized and multi-scale manner. In particular, Adaptive-ALG runs MALG in a sequence of blocks with doubling lengths. Within each block, Adaptive-ALG first initializes a MALG schedule (Algorithm 6 in Appendix 4.C), and then interacts the unique active instance at each episode with the environment (lines 5-7 in Algorithm 5). At the end of each episode, Adaptive-ALG performs two non-stationarity tests (line 10 in Algorithm 5), and if either of them returns *fail*, the restart is triggered. We now describe these three parts in detail below.

MALG-initialization. MALG is run for an interval of length 2^n (unless it is terminated by the non-stationarity detection), which is called a *block*. During the initialization, MALG partions the block equally into 2^{n-k} sub-intervals of length 2^k for $k = 0, 1, \ldots, n$, and an instance of based algorithm (denoted by ALG) is scheduled for each of these sub-intervals with probability $\frac{\rho(2^n)}{\rho(2^k)}$, where ρ is a non-increasing function associated with the bound on **R2** for ALG in a stationary environment (see Appendix 4.C). We refer to these instances of length 2^k as order-k instances.

MALG-interaction. After the initialization, MALG starts interacting with the environment as follows. In each episode m, the unique instance alg that covers this episode with the shortest length is considered as active, while all others are regarded as inactive. MALG follows the decision of the active instance alg and updates it after receiving the feedback from the environment. All inactive instances do not make any decisions or updates, that is, they are paused but may be resumed at some future episode. We refer the read to Appendix 4.C for an illustrative example for MALG procedure.

Non-stationarity detection For $\beta > 0$, two non-stationarity tests are performed for the two terms in the decomposition (4.7). In particular, **Test1** prevents **R1** from growing too large by testing if there is some order$-k$ instance's interval during which the learner's average performance $\frac{1}{2^k}\sum_{\tau=alg.s}^{alg.e} R_\tau$ is larger than the promised optimistic estimator $U_m = \min_{\tau\in[m_n,m]} g_\tau$ (for a stationary environment) by a certain amount. On the other hand, **Test2** prevents **R2** from growing too large by directly testing if its average is large than the promised regret bound. The two non-stationarity tests for $\beta < 0$ are similar but with $\frac{1}{2^k}\sum_{\tau=alg.s}^{alg.e} R_\tau$ and U_m exchanged in **TEST1**, as well as with g_τ and R_τ exchanged in **TEST2**.

Theoretical Results and Discussions

For simplicity, we denote the revised Algorithms 3 and 4 without the restart mechanism (that is, $W = M$) as RSMB and RSQ, respectively. We now present our main theoretical result for

Algorithm 5 when the base algorithms are RSMB and RSQ, respectively.

Theorem 13. *For every $\delta \in (0,1]$, with probability at least $1-\delta$ it holds for Algorithm 5 that*

$$\text{D-Regret(M)} \leq \begin{cases} \widetilde{\mathcal{O}}\left(e^{|\beta|H}|S|^{\frac{2}{3}}|A|^{\frac{1}{3}}H^2M^{\frac{2}{3}}B^{\frac{1}{3}}\right), \textit{if ALG is RSMB,} \\ \widetilde{\mathcal{O}}\left(e^{|\beta|H}|S|^{\frac{2}{3}}|A|^{\frac{1}{3}}H^{\frac{5}{3}}M^{\frac{2}{3}}B^{\frac{1}{3}}\right), \textit{if ALG is RSQ.} \end{cases}$$

The above results show that the dynamic regret bound of the adaptive Algorithm 5 (almost) matches that of the restart Algorithms 3-4 that require the knowledge of the variation budget. The proof of Theorem 13 relies on the results in Theorems 11-4 and is provided in Appendix 4.C.

4.5 Lower Bound

We now present a lower bound on the dynamic regret which complements the upper bounds in Theorems 11, 12 and 13.

Theorem 14. *For sufficiently large M, there exists an instance of non-stationary MDP with H horizons, state space $\mathcal{S}$, action space $\mathcal{A}$ and variation budget B such that*

$$\text{D-Regret(M)} \geq \Omega\left(\frac{e^{\frac{2|\beta|H}{3}}-1}{|\beta|}|\mathcal{S}|^{\frac{1}{3}}|\mathcal{A}|^{\frac{1}{3}}M^{\frac{2}{3}}B^{\frac{1}{3}}\right).$$

Theorem 14 shows that the exponential dependence on $|\beta|$ and H in Theorems 11, 12 and 13 is essentially indispensable and that the results in Theorems 11, 12 and 13 are nearly optimal in their dependence on $|\mathcal{A}|, M$ and B. When $\beta \to 0$, we recover the existing lower bound for the non-stationary risk-neutral episodic MDP problems [87].

The proof is given in Appendix 4.C. In the proof, the hard instance we construct is a non-stationary MDP with piecewise constant dynamics on each segment of the horizon, and its dynamics experience an abrupt change at the beginning of each new segment. In each segment, we construct a $|\mathcal{S}||\mathcal{A}|$-arm bandit model with Bernoulli reward for each arm. This bandit model can be seen as a special case of our episodic MDP problem, and then we show the expected regret, in terms of the logarithmic-exponential objective, that any bandit algorithm has to incur.

4.6 Summary

Our main theoretical contributions are summarized in Table 4.6.1.

Algorithm	**D-Regret**	**Parameter-free**	**Model-free**	**Separation**
Restart-RSMB	$\widetilde{\mathcal{O}}\left(e^{\lvert\beta\rvert H}\lvert\mathcal{S}\rvert^{\frac{2}{3}}\lvert\mathcal{A}\rvert^{\frac{1}{3}}H^2M^{\frac{2}{3}}B^{\frac{1}{3}}\right)$	✗	✗	✓
Restart-RSQ	$\widetilde{\mathcal{O}}\left(e^{\lvert\beta\rvert H}\lvert\mathcal{S}\rvert^{\frac{1}{3}}\lvert\mathcal{A}\rvert^{\frac{1}{3}}H^{\frac{9}{4}}M^{\frac{2}{3}}B^{\frac{1}{3}}\right)$	✗	✓	✓
Adaptive-RSMB	$\widetilde{\mathcal{O}}\left(e^{\lvert\beta\rvert H}\lvert\mathcal{S}\rvert^{\frac{2}{3}}\lvert\mathcal{A}\rvert^{\frac{1}{3}}H^2M^{\frac{2}{3}}B^{\frac{1}{3}}\right)$	✓	✗	✗
Adaptive-RSQ	$\widetilde{\mathcal{O}}\left(e^{\lvert\beta\rvert H}\lvert\mathcal{S}\rvert^{\frac{1}{3}}\lvert\mathcal{A}\rvert^{\frac{1}{3}}H^{\frac{5}{3}}M^{\frac{2}{3}}B^{\frac{1}{3}}\right)$	✓	✓	✗
Lower bound	$\Omega\left(\frac{e^{\frac{2\lvert\beta\rvert H}{3}}-1}{\lvert\beta\rvert}\lvert\mathcal{S}\rvert^{\frac{1}{3}}\lvert\mathcal{A}\rvert^{\frac{1}{3}}M^{\frac{2}{3}}B^{\frac{1}{3}}\right)$	N/A	N/A	N/A

Table 4.6.1: We summarize the dynamic regrets and lower bound obtained in this paper. Here, β is the risk parameter, H is the horizon of each episode, M is the total number of episodes, B is the total variation measurement, and $|\mathcal{S}|$ and $|\mathcal{A}|$ are the cardinalities of the state and action spaces.

Appendix

4.A Proof of Theorem 11

Preliminaries

First, we set some notations and definitions. Define $\iota := \log(6H|\mathcal{S}||\mathcal{A}|W/p)$ for a given $p \in (0,1]$. We adopt the shorthand notations $\mathbb{1}_h^m(s,a) := \mathbb{1}\left\{\left(s_h^m, a_h^m\right) = (s,a)\right\}$ and $r_h^m := r_h\left(s_h^m, a_h^m\right)$ for $(m,h) \in [M] \times [H]$. The epoch is defined as an interval that starts at the first episode after a restart and ends at the first time when the restart is triggered. In Algorithm 3, the restart mechanism divides M episodes into $\lceil \frac{M}{W} \rceil$ epochs.

For every $(m,h) \in [M] \times [H]$, and $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, we define two visitation counters $N_h^m(s,a,s')$ and $N_h^m(x,a)$ at step h in episode m as follows:

$$\begin{aligned} N_h^m(s,a,s') &= \sum_{\tau=\ell^m}^{m-1} 1\left\{(s,a,s') = \left(s_h^\tau, a_h^\tau, s_{h+1}^\tau\right)\right\}, \\ N_h^m(s,a) &= \sum_{\tau=\ell^m}^{m-1} 1\left\{(s,a) = \left(s_h^\tau, a_h^\tau\right)\right\}. \end{aligned} \tag{4.8a}$$

This allows us to estimate the transition kernel $\mathcal{P}_h^m$ and reward function r^m for episode m using only the data from the episode $\ell^m = (\lceil \frac{m}{W} \rceil - 1)W + 1$ to the episode m by

$$\widehat{\mathcal{P}}_h^m(s' \mid s,a) = \frac{N_h^m(s,a,s') + \frac{\lambda}{|\mathcal{S}|}}{N_h^m(s,a) + \lambda}, \text{ for all } (s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \tag{4.9a}$$

$$\widehat{r}_h^m(s,a) = \frac{1}{N_h^m(s,a)+\lambda} \sum_{\tau=\ell^m}^{m-1} 1\left\{(s,a) = \left(s_h^\tau, a_h^\tau\right)\right\} r_h^\tau\left(s_h^\tau, a_h^\tau\right), \text{ for all } (s,a) \in \mathcal{S} \times \mathcal{A}, \tag{4.9b}$$

where $\lambda > 0$ is the regularization parameter. We denote by V_h^m, G_h^m, Γ_h^m the values of V_h, G_h, Γ_h after the updates in step h of episode m, respectively. We also set $Q_h^m = \frac{1}{\beta} \log\left\{G_h^m\right\}$.

Let us fix a pair $(s,a) \in \mathcal{S} \times \mathcal{A}$. Recall from Algorithm 3 that

$$w_h^m(s,a) = \sum_{s'} \hat{\mathcal{P}}_h^m(s' \mid s,a)\left[e^{\beta\left[\hat{r}_h^m(s,a) + V_{h+1}^m(s')\right]}\right].$$

We define

$$q_{h,1}^{m,+}(s,a) := \begin{cases} w_h^m(s,a) + \Gamma_h^m(s,a), & \text{if } \beta > 0 \\ w_h^m(s,a) - \Gamma_h^m(s,a), & \text{if } \beta < 0 \end{cases}$$

$$q_{h,1}^{m}(s,a) := \begin{cases} \min\left\{q_{h,1}^{m,+}(s,a), e^{\beta(H-h+1)}\right\}, & \text{if } \beta > 0 \\ \max\left\{q_{h,1}^{m,+}(s,a), e^{\beta(H-h+1)}\right\}, & \text{if } \beta < 0 \end{cases}$$

and

$$q_{h,2}^{m}(s,a) := \mathbb{E}_{s' \sim \mathcal{P}_h^m(\cdot|s,a)}\left[e^{\beta\left[r_h^m(s,a) + V_{h+1}^m(s')\right]}\right], \tag{4.10}$$

as well as the following for a policy π,

$$q_{h,3}^{m,\pi}(s,a) := \mathbb{E}_{s' \sim \mathcal{P}_h^m(\cdot|s,a)}\left[e^{\beta\left[r_h^m(s,a) + V_{h+1}^{\pi,m}(s')\right]}\right] \tag{4.11}$$

Model Prediction Errors

Lemma 23. *Define* $\overline{\mathcal{V}}_{h+1} := \left\{\bar{V}_{h+1} : \mathcal{S} \to \mathbb{R} \mid \forall s \in \mathcal{S}, \bar{V}_{h+1}(s) \in [0, H-h]\right\}$. *For any* $p \in (0,1]$, *with probability* $1 - p/2$, *we have*

$$\begin{aligned} &\left|\sum_{s' \in \mathcal{S}}\left(\widehat{\mathcal{P}}_h^m\left(s' \mid s,a\right) e^{\beta\left[r_h^m(s,a) + \bar{V}(s')\right]} - \mathcal{P}_h^m\left(s' \mid s,a\right) e^{\beta\left[r_h^m(s,a) + \bar{V}(s')\right]}\right)\right| \\ \leq &\Gamma_h^m + \left|e^{\beta(H-h+1)} - 1\right| B_{\mathcal{P},\mathcal{E}} \end{aligned}$$

for every $(s,a,m,h) \in \mathcal{S} \times \mathcal{A} \times [M] \times [H]$ *and* $\bar{V} \in \overline{\mathcal{V}}_{h+1}$, *where* Γ_h^m *is defined in* (4.5).

Proof. For the ease of notation, we denote $\sum_{s' \in \mathcal{S}} \mathcal{P}_h^m\left(s' \mid s,a\right) e^{\beta\left[r_h^m(s,a) + \bar{V}(s')\right]}$ as $\left(\mathcal{P}_h^m e^{\beta\left[r_h^m + \bar{V}\right]}\right)(s,a)$. Then, for every $\bar{V} \in \mathcal{V}_{h+1}$, we consider the difference between $\sum_{s' \in \mathcal{S}} \widehat{\mathcal{P}}_h^m\left(s' \mid \cdot,\cdot\right) e^{\beta\left[r_h^m(s,a) + \bar{V}(s')\right]}$

and $\sum_{s'\in\mathcal{S}} \mathcal{P}_h^m(s' \mid \cdot,\cdot)\, e^{\beta[r_h^m(s,a)+\bar{V}(s')]}$ as follows:

$$
\begin{aligned}
&(N_h^m(s,a)+\lambda)\left|\sum_{s'\in\mathcal{S}}\left(\widehat{\mathcal{P}}_h^m(s' \mid s,a)\, e^{\beta[r_h^m(s,a)+\bar{V}(s')]} - \mathcal{P}_h^m(s' \mid s,a)\, e^{\beta[r_h^m(s,a)+\bar{V}(s')]}\right)\right| && (4.12)\\
&= \left|\sum_{s'\in\mathcal{S}}\left(N_h^m(s,a,s') + \frac{\lambda}{|\mathcal{S}|}\right) e^{\beta[r_h^m(s,a)+\bar{V}(s')]} - (N_h^m(s,a)+\lambda)\left(\mathcal{P}_h^m e^{\beta[r_h^m+\bar{V}]}\right)(s,a)\right| \\
&\le \left|\sum_{s'\in\mathcal{S}} N_h^m(s,a,s')\, e^{\beta[r_h^m(s,a)+\bar{V}(s')]} - N_h^m(s,a)\left(\mathcal{P}_h^m e^{\beta[r_h^m+\bar{V}]}\right)(s,a)\right| \\
&\quad + \lambda\left|\frac{1}{|\mathcal{S}|}\sum_{s'\in\mathcal{S}} e^{\beta[r_h^m(s,a)+\bar{V}(s')]} - \mathcal{P}_h^m e^{\beta[r_h^m+\bar{V}]}(s,a)\right| \\
&= \left|\sum_{\tau=\ell_Q^m}^{m-1} \mathbb{1}\{(s,a)=(s_h^\tau,a_h^\tau)\}\left(e^{\beta[r_h^m(s_h^\tau,a_h^\tau)+\bar{V}(s_{h+1}^\tau)]} - \left(\mathcal{P}_h^m e^{\beta[r_h^m+\bar{V}]}\right)(s,a)\right)\right| \\
&\quad + \lambda\left|\frac{1}{|\mathcal{S}|}\sum_{s'\in\mathcal{S}} e^{\beta[r_h^m(s,a)+\bar{V}(s')]} - \mathcal{P}_h^m e^{\beta[r_h^m+\bar{V}]}(s,a)\right| \\
&= \left|\sum_{\tau=\ell_Q^m}^{m-1} \mathbb{1}\{(s,a)=(s_h^\tau,a_h^\tau)\}\, e^{\beta r_h^m(s_h^\tau,a_h^\tau)}\left(e^{\beta\bar{V}(s_{h+1}^\tau)} - \left(\mathcal{P}_h^m e^{\beta\bar{V}}\right)(s,a)\right)\right| \\
&\quad + \lambda\left|\frac{1}{|\mathcal{S}|}\sum_{s'\in\mathcal{S}} e^{\beta[r_h^m(s,a)+\bar{V}(s')]} - \mathcal{P}_h^m e^{\beta[r_h^m+\bar{V}]}(s,a)\right| \\
&\le \left|\sum_{\tau=\ell_Q^m}^{m-1} \mathbb{1}\{(s,a)=(s_h^\tau,a_h^\tau)\}\, e^{\beta r_h^m(s_h^\tau,a_h^\tau)}\left(e^{\beta\bar{V}(s_{h+1}^\tau)} - \left(\mathcal{P}_h^\tau e^{\beta\bar{V}}\right)(s,a)\right)\right| && (4.13)\\
&\quad + \left|\sum_{\tau=\ell_Q^m}^{m-1} \mathbb{1}\{(s,a)=(s_h^\tau,a_h^\tau)\}\, e^{\beta r_h^m(s_h^\tau,a_h^\tau)}\left(\left(\mathcal{P}_h^\tau e^{\beta\bar{V}}\right)(s,a) - \left(\mathcal{P}_h^m e^{\beta\bar{V}}\right)(s,a)\right)\right| && (4.14)\\
&\quad + \lambda\left|\frac{1}{|\mathcal{S}|}\sum_{s'\in\mathcal{S}} e^{\beta[r_h^m(s,a)+\bar{V}(s')]} - \mathcal{P}_h^m e^{\beta[r_h^m+\bar{V}]}(s,a)\right| && (4.15)
\end{aligned}
$$

for every $(m,h)\in[M]\times[H]$ and $(s,a)\in\mathcal{S}\times\mathcal{A}$.

To analyze the term in (4.13), we let $\eta_h^\tau := e^{\beta[r_h^m(s_h^\tau,a_h^\tau)+\bar{V}(s_{h+1}^\tau)]} - \left(\mathcal{P}_h^\tau e^{\beta[r_h^m+\bar{V}]}\right)(s_h^\tau,a_h^\tau)$. Conditioning on the filtration $\mathcal{F}_{h,1}^m$, the term η_h^τ is a zero-mean and $|e^{\beta(H-h+1)}-1|$-sub-Gaussian random variable. By Lemma 44, we use $Y=\lambda I$ and $X_\tau = \mathbb{1}\left\{(s,a)=\left(s_h^\tau,a_h^\tau\right)\right\}$ and thus with

probability at least $1-\delta$ it holds for every $m \in [M]$ that

$$
\begin{aligned}
&\left(N_h^m(s,a)+\lambda\right)^{-1/2}\left|\sum_{\tau=\ell_Q^m}^{m-1}\mathbb{1}\left\{(s,a)=(s_h^\tau,a_h^\tau)\right\}\left(e^{\beta[r_h^m(s_h^\tau,a_h^\tau)+\bar{V}](s_{h+1}^\tau)}-\left(\mathcal{P}_h^\tau e^{\beta[r_h^m+\bar{V}]}\right)(s_h^\tau,a_h^\tau)\right)\right| \\
&\leq\sqrt{\frac{\left(e^{\beta(H-h+1)}-1\right)^2}{2}\log\left(\frac{\left(N_h^m(s,a)+\lambda\right)^{1/2}\lambda^{-1/2}}{\delta}\right)} \\
&\leq\sqrt{\frac{\left(e^{\beta(H-h+1)}-1\right)^2}{2}\log\left(\frac{W}{\delta}\right)}
\end{aligned}
$$

where W is the restart period.

For the term in (4.14), by the definition of $B_{\mathcal{P},\mathcal{E}}$ and N_h^m, we have

$$
\begin{aligned}
&\left|\sum_{\tau=\ell_Q^m}^{m-1}\mathbb{1}\left\{(s,a)=(s_h^\tau,a_h^\tau)\right\}\left(\left(\mathcal{P}_h^\tau e^{\beta[r_h^m+\bar{V}]}\right)(s,a)-\left(\mathcal{P}_h^m e^{\beta[r_h^m+\bar{V}]}\right)(s,a)\right)\right| \\
&=\left|\sum_{\tau=\ell_Q^m}^{m-1}\mathbb{1}\left\{(s,a)=(s_h^\tau,a_h^\tau)\right\}\left(\left(\mathcal{P}_h^\tau\left(e^{\beta[r_h^m+\bar{V}]}-1\right)\right)(s,a)-\left(\mathcal{P}_h^m\left(e^{\beta[r_h^m+\bar{V}]}-1\right)\right)(s,a)\right)\right| \\
&\leq\left|\sum_{\tau=\ell_Q^m}^{m-1}\mathbb{1}\left\{(s,a)=(s_h^\tau,a_h^\tau)\right\}\right|\left|e^{\beta(H-h+1)}-1\right|B_{\mathcal{P},\mathcal{E}} \\
&\leq\left(N_h^m(s,a)+\lambda\right)\left|e^{\beta(H-h+1)}-1\right|B_{\mathcal{P},\mathcal{E}}.
\end{aligned}
$$

where the first equality is due to $\mathcal{P}_h^m 1=\mathcal{P}_h^\tau 1$ for all $\tau\in[\ell^m,m-1]$. For the term in (4.15), we have

$$
\begin{aligned}
\lambda\left|\frac{1}{|\mathcal{S}|}\sum_{s'\in\mathcal{S}}e^{\beta[r_h^m(s,a)+\bar{V}(s')]}-\mathcal{P}_h^m e^{\beta[r_h^m+\bar{V}]}(s,a)\right| &\leq\frac{\lambda}{|\mathcal{S}|}\sum_{s'\in\mathcal{S}}\left|e^{\beta[r_h^m(s,a)+\bar{V}(s')]}-\mathcal{P}_h^m e^{\beta[r_h^m+\bar{V}]}(s,a)\right| \\
&\leq\lambda\left|e^{\beta(H-h+1)}-1\right|.
\end{aligned}
$$

By returning to (4.12) and setting $\lambda=1$, with probability at least $1-\delta$ it holds that

$$
\begin{aligned}
&\left|\sum_{s'\in\mathcal{S}}\left(\widehat{\mathcal{P}}_h^m\left(s'\mid s,a\right)e^{\beta\left[r_h^m(s,a)+\bar{V}(s')\right]}-\mathcal{P}_h^m\left(s'\mid s,a\right)e^{\beta\left[r_h^m(s,a)+\bar{V}(s')\right]}\right)\right| \\
&\leq\left(N_h^m(s,a)+\lambda\right)^{-\frac{1}{2}}\left|e^{\beta(H-h+1)}-1\right|\sqrt{\frac{1}{2}\left(\log\left(\frac{W}{\delta}\right)\right)}+\left|e^{\beta(H-h+1)}-1\right|B_{\mathcal{P},\mathcal{E}}+\left|e^{\beta(H-h+1)}-1\right| \\
&\leq C_1\left(N_h^m(s,a)+\lambda\right)^{-\frac{1}{2}}\left|e^{\beta(H-h+1)}-1\right|\sqrt{\left(\log\left(\frac{W}{\delta}\right)\right)}+\left|e^{\beta(H-h+1)}-1\right|B_{\mathcal{P},\mathcal{E}}
\end{aligned}
$$

for all $m\in[M]$ and for some constant $C_1>1$.

Furthermore, let $d(V, V') = \max_{s\in\mathcal{S}} |V(s) - V'(s)|$ be a distance on $\mathcal{V}_{h+1}$. For every ϵ, an ϵ-covering $\mathcal{V}^{\epsilon}_{h+1}$ of $\mathcal{V}_{h+1}$ with respect to distance $d(\cdot,\cdot)$ satisfies $\left|\mathcal{V}^{\epsilon}_{h+1}\right| \le \left(\frac{1}{\epsilon}\right)^{|\mathcal{S}|}$. Then, for every $V \in \mathcal{V}_{h+1}$, there exists $V' \in \mathcal{V}^{\epsilon}_{h+1}$ such that $\max_{s\in\mathcal{S}} |V(s) - V'(s)| \le \epsilon$, which further implies that

$$\max_{s,a,s'} \left| e^{\beta\left[r^m_h(s,a)+V(s')\right]} - e^{\beta\left[r^m_h(s,a)+V'(s')\right]} \right| \le g_h(\beta)\epsilon,$$

where

$$g_h(\beta) = \begin{cases} e^{\beta(H-h+1)}\beta, & \text{if } \beta > 0, \\ -\beta, & \text{if } \beta < 0. \end{cases} \tag{4.16}$$

Thus, by the triangle inequality and (4.12), we have

$$\begin{aligned}
&\left| \sum_{s'\in\mathcal{S}} \left(\widehat{\mathcal{P}}^m_h\left(s' \mid s,a\right) e^{\beta\left[r^m_h(s,a)+V(s')\right]} - \mathcal{P}^m_h\left(s' \mid s,a\right) e^{\beta\left[r^m_h(s,a)+V(s')\right]} \right) \right| \\
\le& \left| \sum_{s'\in\mathcal{S}} \left(\widehat{\mathcal{P}}^m_h\left(s' \mid s,a\right) e^{\beta\left[r^m_h(s,a)+V'(s')\right]} - \mathcal{P}^m_h\left(s' \mid s,a\right) e^{\beta\left[r^m_h(s,a)+V'(s')\right]} \right) \right| + 2g_h(\beta)\beta\epsilon \\
\le& C_1 \left(N^m_h(s,a)+\lambda\right)^{-1/2} \left| e^{\beta(H-h+1)} - 1 \right| \sqrt{\log\left(\frac{W}{\delta}\right)} + \left| e^{\beta(H-h+1)} - 1 \right| B_{\mathcal{P},\mathcal{E}} + 2g_h(\beta)\epsilon.
\end{aligned}$$

Then, by choosing $\delta = (p/2)/\left(\left|\mathcal{V}^{\epsilon}_{h+1}\right| H\,|\mathcal{S}|\,|\mathcal{A}|\right)$, $\epsilon = \frac{1}{4\sqrt{W}}$, and taking a union bound over $V \in \mathcal{V}^{\epsilon}_{h+1}$ and $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, it holds with probability at least $1 - p/2$ that

$$\begin{aligned}
&\sup_{V\in\mathcal{V}_{h+1}} \left\{ \left| \sum_{s'\in\mathcal{S}} \left(\widehat{\mathcal{P}}^m_h\left(s' \mid s,a\right) e^{\beta\left[r^m_h(s,a)+V(s')\right]} - \mathcal{P}^m_h\left(s' \mid s,a\right) e^{\beta\left[r^m_h(s,a)+V(s')\right]} \right) \right| \right\} \\
\le& C_1 \left(N^m_h(s,a)+\lambda\right)^{-1/2} \left| e^{\beta(H-h+1)} - 1 \right| \sqrt{\left(\log\left(\frac{6W\left|\mathcal{V}^{\epsilon}_{h+1}\right| H\,|\mathcal{S}|\,|\mathcal{A}|}{p} \right) \right)} + \left| e^{\beta(H-h+1)} - 1 \right| B_{\mathcal{P},\mathcal{E}} \\
&+ 2g_h(\beta)\epsilon \\
\le& C_1 \left(N^m_h(s,a)+\lambda\right)^{-1/2} \left| e^{\beta(H-h+1)} - 1 \right| \sqrt{|\mathcal{S}| \left(\log\left(\frac{6WH\,|\mathcal{S}|\,|\mathcal{A}|}{p} \right) \right)} + \left| e^{\beta(H-h+1)} - 1 \right| B_{\mathcal{P},\mathcal{E}} \\
&+ g_h(\beta) W^{-1/2} \\
\le& \left(C_1 \left| e^{\beta(H-h+1)} - 1 \right| + g_h(\beta) \right) \left(N^m_h(s,a)+\lambda\right)^{-1/2} \sqrt{|\mathcal{S}| \left(\log\left(\frac{6WH\,|\mathcal{S}|\,|\mathcal{A}|}{p} \right) \right)} + \left| e^{\beta(H-h+1)} - 1 \right| B_{\mathcal{P},\mathcal{E}} \\
\le& C_1 \left(\left| e^{\beta(H-h+1)} - 1 \right| + g_h(\beta) \right) \left(N^m_h(s,a)+\lambda\right)^{-1/2} \sqrt{|\mathcal{S}| \left(\log\left(\frac{6WH\,|\mathcal{S}|\,|\mathcal{A}|}{p} \right) \right)} + \left| e^{\beta(H-h+1)} - 1 \right| B_{\mathcal{P},\mathcal{E}}
\end{aligned}$$

for every $(s,a,m,h) \in \mathcal{S}\times\mathcal{A}\times[M]\times[H]$. By our choice of Γ_h^m, with probability at least $1-p/2$ it holds that

$$\begin{aligned}&\left|\sum_{s'\in\mathcal{S}}\left(\widehat{\mathcal{P}}_h^m\left(s'\mid s,a\right)e^{\beta\left[r_h^m(s,a)+\bar{V}(s')\right]}-\mathcal{P}_h^m\left(s'\mid s,a\right)e^{\beta\left[r_h^m(s,a)+\bar{V}(s')\right]}\right)\right|\\&\leq\Gamma_h^m+\left|e^{\beta(H-h+1)}-1\right|B_{\mathcal{P},\mathcal{E}}\end{aligned}$$

for every $(s,a,m,h)\in\mathcal{S}\times\mathcal{A}\times[M]\times[H]$. □

Lemma 24. *For every $(s,a,m,h)\in\mathcal{S}\times\mathcal{A}\times[M]\times[H]$ and $\bar{V}\in\overline{\mathcal{V}}_{h+1}$, we have*

$$\left|\sum_{s'\in\mathcal{S}}\left(\widehat{\mathcal{P}}_h^m\left(s'\mid s,a\right)e^{\beta\left[\hat{r}_h^m(s,a)+\bar{V}(s')\right]}-\widehat{\mathcal{P}}_h^m\left(s'\mid s,a\right)e^{\beta\left[r_h^m(s,a)+\bar{V}(s')\right]}\right)\right|\leq\Gamma_h^m+g_h(\beta)B_{r,\mathcal{E}}$$

where $g_h(\beta)$ is defined in (4.16).

Proof. Since

$$|e^{\beta x}-e^{\beta y}|\leq\begin{cases}\beta e^{\beta u}|x-y|, & \text{if } \beta>0,\\ -\beta|x-y|, & \text{if } \beta<0\end{cases}$$

for every $0\leq x\leq u$ and $0\leq y\leq u$ where $u>0$ is some constant, it holds that

$$\begin{aligned}&\left|\sum_{s'\in\mathcal{S}}\left(\widehat{\mathcal{P}}_h^m\left(s'\mid s,a\right)e^{\beta\left[\hat{r}_h^m(s,a)+\bar{V}(s')\right]}-\widehat{\mathcal{P}}_h^m\left(s'\mid s,a\right)e^{\beta\left[r_h^m(s,a)+\bar{V}(s')\right]}\right)\right|\\&\leq g_h(\beta)\left|\hat{r}_h^m(s,a)-r_h^m(s,a)\right|.\end{aligned}\tag{4.17}$$

Furthermore, by our estimation $\hat{r}_h^m(x,a)$, we have

$$\begin{aligned}&\left|\hat{r}_h^m(x,a)-r_h^m(x,a)\right|\\&=\left|\hat{r}_h^m(x,a)-r_h^m(x,a)\right|\\&=\left(n_h^m(x,a)+\lambda\right)^{-1}\left|\sum_{\tau=\ell^m}^{m-1}1\left\{(x,a)=(x_h^\tau,a_h^\tau)\right\}\left(r_h^\tau(x_h^\tau,a_h^\tau)-r_h^m(x,a)\right)-\lambda r_h^m(x,a)\right|\\&\leq B_{r,\mathcal{E}}+\left(n_h^m(x,a)+\lambda\right)^{-1}\left|\lambda r_h^m(x,a)\right|\\&\leq B_{r,\mathcal{E}}+\left(n_h^m(x,a)+\lambda\right)^{-1}\lambda\\&\leq B_{r,\mathcal{E}}+\left(n_h^m(x,a)+\lambda\right)^{-1/2}\lambda\end{aligned}$$

By substituting the above inequality into (4.17) and setting $\lambda=1$, we obtain the desired results. □

Lemma 25. *For every $p\in(0,1]$, with probability $1-p/2$, we have*

$$\begin{aligned}&\left|\sum_{s'\in\mathcal{S}}\left(\widehat{\mathcal{P}}_h^m\left(s'\mid s,a\right)e^{\beta\left[\hat{r}_h^m(s,a)+\bar{V}(s')\right]}-\mathcal{P}_h^m\left(s'\mid s,a\right)e^{\beta\left[r_h^m(s,a)+\bar{V}(s')\right]}\right)\right|\\&\leq 2\Gamma_h^m+\left|e^{\beta(H-h+1)}-1\right|B_{\mathcal{P},\mathcal{E}}+g_h(\beta)B_{r,\mathcal{E}}\end{aligned}$$

where $g_h(\beta)$ is defined in (4.16)*, for every $(s,a,m,h)\in\mathcal{S}\times\mathcal{A}\times[M]\times[H]$ and $\bar{V}\in\overline{\mathcal{V}}_{h+1}$.*

Proof. The proof follows from Lemma 23, Lemma 24 and Cauchy-Schwartz inequality. □

Value Difference Bounds

Lemma 26. *Recall the definition of Γ_h^m from Algorithm 3. For all $(m,h,s,a) \in [M]\times[H]\times\mathcal{S}\times\mathcal{A}$, the following statement holds with probability at least $1-p/2$:*

- *If $\beta > 0$:*

$$\begin{aligned} -\left|e^{\beta(H-h+1)}-1\right| B_{\mathcal{P},\mathcal{E}} - g_h(\beta) B_{r,\mathcal{E}} \le & \left(q_{h,1}^m - q_{h,2}^m\right)(s,a) \\ \le & 4\Gamma_h^m + \left|e^{\beta(H-h+1)}-1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta) B_{r,\mathcal{E}}. \end{aligned}$$

- *If $\beta < 0$:*

$$\begin{aligned} -\left|e^{\beta(H-h+1)}-1\right| B_{\mathcal{P},\mathcal{E}} - g_h(\beta) B_{r,\mathcal{E}} \le & \left(q_{h,2}^m - q_{h,1}^m\right)(s,a) \\ \le & 4\Gamma_h^m + \left|e^{\beta(H-h+1)}-1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta) B_{r,\mathcal{E}}. \end{aligned}$$

(Note that $g_h(\beta)$ is defined in (4.16)).

Proof. We focus on the case of $\beta > 0$ since the proof for $\beta < 0$ is similar. We first fix a tuple $(m,h,s,a) \in [M]\times[H]\times\mathcal{S}\times\mathcal{A}$. By the definitions of $q_{h,1}^{m,+}$ and $q_{h,2}^m$, one can compute

$$\begin{aligned} & \left|\left(q_{h,1}^{m,+} - 2\Gamma_h^m - q_{h,2}^m\right)(s,a)\right| \\ &= \left|\left(w_h^m - q_{h,2}^m\right)(s,a)\right| \\ &= \left|\sum_{s'\in\mathcal{S}} \left(\widehat{\mathcal{P}}_h^m\left(s' \mid s,a\right) e^{\beta\left[\hat{r}_h^m(s,a)+\bar{V}(s')\right]} - \mathcal{P}_h^m\left(s' \mid s,a\right) e^{\beta\left[r_h^m(s,a)+\bar{V}(s')\right]}\right)\right| \\ &\le 2\Gamma_h^m + \left|e^{\beta(H-h+1)}-1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta) B_{r,\mathcal{E}} \end{aligned}$$

where the last step holds by Lemma 23. Then, we have

$$\begin{aligned} -\left|e^{\beta(H-h+1)}-1\right| B_{\mathcal{P},\mathcal{E}} - g_h(\beta) B_{r,\mathcal{E}} \le & \left(q_{h,1}^{m,+} - q_{h,2}^m\right)(s,a) \\ \le & 4\Gamma_h^m + \left|e^{\beta(H-h+1)}-1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta) B_{r,\mathcal{E}}. \end{aligned}$$

Furthermore, if $q_{h,1}^{m,+} \le e^{\beta(H-h+1)}$, one can write

$$q_{h,1}^{m,+} - q_{h,2}^m = q_{h,1}^m - q_{h,2}^m \ge -\left|e^{\beta(H-h+1)}-1\right| B_{\mathcal{P},\mathcal{E}} - g_h(\beta) B_{r,\mathcal{E}}.$$

If $q_{h,1}^{m,+} \ge e^{\beta(H-h+1)}$, we have $q_{h,1}^{m,+} - q_{h,2}^m = e^{\beta(H-h+1)} - q_{h,2}^m \ge 0$. In addition, since $q_{h,1}^{m,+} \ge q_{h,1}^m$, it holds that $q_{h,1}^m - q_{h,2}^m \le q_{h,1}^{m,+} - q_{h,2}^m$. This completes the proof.

□

Lemma 27. *On the event of Lemma 26, for all $(m,h,s,a) \in [M]\times[H]\times\mathcal{S}\times\mathcal{A}$ and every policy π:*

- *If* $\beta > 0$*:*

$$e^{\beta \cdot Q_h^m(s,a)} - e^{\beta \cdot Q_h^{\pi,m}(s,a)} \geq -(H-h+1)\left[\left|e^{\beta(H-h+1)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta) B_{r,\mathcal{E}}\right].$$

- *If* $\beta < 0$*:*

$$e^{\beta \cdot Q_h^m(s,a)} - e^{\beta \cdot Q_h^{\pi,m}(s,a)} \leq (H-h+1)\left[\left|e^{\beta(H-h+1)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta) B_{r,\mathcal{E}}\right].$$

Proof. We focus on the case of $\beta > 0$ since the proof for $\beta < 0$ is similar. For the purpose of the proof, we set $Q_{H+1}^{\pi,m}(s,a) = Q_{H+1}^{*,m}(s,a) = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. We fix a tuple $(m,s,a) \in [M] \times \mathcal{S} \times \mathcal{A}$ and use strong induction on h. The base case for $h = H+1$ is satisfied since $e^{\beta \cdot Q_{H+1}^m(s,a)} = e^{\beta \cdot Q_{H+1}^{\pi,m}(s,a)} = 1$ for all $m \in [M]$ by definition. Now, we fix an index $h \in [H]$ and assume that

$$e^{\beta \cdot Q_{h+1}^m(s,a)} - e^{\beta \cdot Q_{h+1}^{\pi,m}(s,a)} \geq -(H-h)\left[\left|e^{\beta(H-h)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta) B_{r,\mathcal{E}}\right].$$

Moreover, by the induction assumption, we have

$$\begin{aligned} e^{\beta \cdot V_{h+1}^m(s)} &= \max_{a' \in \mathcal{A}} e^{\beta \cdot Q_{h+1}^m(s,a')} \\ &\geq \max_{a' \in \mathcal{A}} e^{\beta \cdot Q_{h+1}^{\pi,m}(s,a')} - (H-h)\left[\left|e^{\beta(H-h)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta) B_{r,\mathcal{E}}\right] \\ &\geq e^{\beta \cdot V_{h+1}^{\pi,m}(s)} - (H-h)\left[\left|e^{\beta(H-h)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta) B_{r,\mathcal{E}}\right]. \end{aligned} \tag{4.18}$$

By the definitions of $q_{h,2}^m$ and $q_{h,3}^{m,\pi}$, it follows from (4.18) that

$$q_{h,2}^m - q_{h,3}^{m,\pi} \geq -(H-h)\left[\left|e^{\beta(H-h+1)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta) B_{r,\mathcal{E}}\right].$$

In addition, on the event of Lemma 26, we also have

$$q_{h,1}^m - q_{h,2}^m \geq -\left[\left|e^{\beta(H-h+1)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta) B_{r,\mathcal{E}}\right].$$

Therefore, it follows that

$$\begin{aligned} \left(e^{\beta \cdot Q_h^m} - e^{\beta \cdot Q_h^{\pi,m}}\right)(s,a) &= \left(q_{h,1}^m - q_{h,3}^{m,\pi}\right)(s,a) \\ &= \left(q_{h,1}^m - q_{h,2}^m\right)(s,a) + \left(q_{h,2}^m - q_{h,3}^{m,\pi}\right)(s,a) \\ &\geq -(H-h+1)\left[\left|e^{\beta(H-h+1)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta) B_{r,\mathcal{E}}\right] \end{aligned}$$

which completes the induction. □

Lemma 28. *For all* $(m,h,s) \in [M] \times [H] \times \mathcal{S}$*, policy* π *and* $\delta \in (0,1]$*, with probability at least* $1 - \delta/2$*:*

- *If* $\beta > 0$*:*

$$e^{\beta \cdot V_h^m(s,a)} - e^{\beta \cdot V_h^{\pi,m}(s,a)} \geq -(H-h+1)\left[\left|e^{\beta(H-h+1)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta) B_{r,\mathcal{E}}\right].$$

- *If* $\beta < 0$*:*

$$e^{\beta \cdot V_h^m(s,a)} - e^{\beta \cdot V_h^{\pi,m}(s,a)} \leq (H-h+1)\left[\left|e^{\beta(H-h+1)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta) B_{r,\mathcal{E}}\right].$$

Proof. The result follows from Lemma 27 and Equation (4.18).

Proof of Theorem 11

We first consider $\beta > 0$. For $h \in [H]$, we define

$$\delta_h^m := e^{\beta V_h^m(s_h^m)} - e^{\beta V_h^{\pi^m,m}(s_h^m)}, \tag{4.19a}$$

$$\begin{aligned}\zeta_{h+1}^m &:= q_{h,2}^m - q_{h,3}^m - e^{\beta r_h^m(s_h^m,a_h^m)}\delta_{h+1}^m \\ &= \left[P_h^m\left(e^{\beta\left[r_h^m(s_h^m,a_h^m)+V_{h+1}^m(s')\right]} - e^{\beta\left[r_h^m(s_h^m,a_h^m)+V_{h+1}^{\pi^m,m}(s')\right]}\right)\right](s_h^m,a_h^m) - e^{\beta r_h^m(s_h^m,a_h^m)}\delta_{h+1}^m,\end{aligned} \tag{4.19b}$$

where $\left[P_h^m f\right](s,a) := \mathbb{E}_{s'\sim P_h^m(|\cdot|s,a)}\left[f(s')\right]$ for every $f : \mathcal{S} \to \mathbb{R}$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$. Then, for every $(m,h) \in [M]\times[H]$, we have

$$\begin{aligned}\delta_h^m &\overset{(i)}{=} \left(e^{\beta\cdot Q_h^m} - e^{\beta\cdot Q_h^{\pi^m,m}}\right)(s_h^m,a_h^m) \\ &\overset{(ii)}{=} q_{h,1}^m(s_h^m,a_h^m) - q_{h,2}^m(s_h^m,a_h^m) + q_{h,2}^m(s_h^m,a_h^m) - q_{h,3}^m(s_h^m,a_h^m) \\ &\overset{(iii)}{\le} 4\Gamma_h^m + \left|e^{\beta(H-h+1)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta)B_{r,\mathcal{E}} + q_{h,2}^m(s_h^m,a_h^m) - q_{h,3}^m(s_h^m,a_h^m) \\ &= 4\Gamma_h^m + \left|e^{\beta(H-h+1)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta)B_{r,\mathcal{E}} + e^{\beta\cdot r_h^m(s_h^m,a_h^m)}\delta_{h+1}^m + \zeta_{h+1}^m.\end{aligned} \tag{4.20}$$

In the above equation, step (i) holds by the construction of Algorithm 3 and the definition of $V_h^{\pi^m}$ in Equation (4.1b); step (ii) holds by Equations (4.10) and (4.11); step (iii) holds on the event of Lemma 26; the last step follows from the definition of δ_h^m and ζ_h^m in Equations 4.19a and 4.19b.

Using the fact that $V_{H+1}^m(s) = V_{H+1}^{\pi^m}(s) = 0$, we can expand the recursion in Equation (4.20) to obtain

$$\begin{aligned}\delta_1^m &\le \sum_{h\in[H]} e^{\beta\sum_{i=1}^{h-1} r_i^m}\zeta_{h+1}^m + \sum_{h\in[H]} e^{\beta\sum_{i=1}^{h-1} r_i^m}\left(4\Gamma_h^m + \left|e^{\beta(H-h+1)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta)B_{r,\mathcal{E}}\right) \\ &\le \sum_{h\in[H]} e^{\beta\sum_{i=1}^{h-1} r_i^m}\zeta_{h+1}^m + \sum_{h\in[H]} e^{\beta(h-1)}\left(4\Gamma_h^m + \left|e^{\beta(H-h+1)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta)B_{r,\mathcal{E}}\right).\end{aligned}$$

where the last step follows from $r_h^m(\cdot,\cdot) \in [0,1]$. Summing the above display over $m \in [M]$ gives

$$\begin{aligned}&\sum_{m\in[M]} \delta_1^m \\ &\le \sum_{m\in[M]}\sum_{h\in[H]} e^{\beta\sum_{i=1}^{h-1} r_i^m}\zeta_{h+1}^m + \sum_{m\in[M]}\sum_{h\in[H]} e^{\beta(h-1)}\left(4\Gamma_h^m + \left|e^{\beta(H-h+1)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta)B_{r,\mathcal{E}}\right) \\ &= \sum_{\mathcal{E}=1}^{\lceil\frac{M}{W}\rceil}\sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W}\sum_{h\in[H]}\left(e^{\beta\sum_{i=1}^{h-1} r_i^m}\zeta_{h+1}^m + e^{\beta(h-1)}\left(4\Gamma_h^m + \left|e^{\beta(H-h+1)} - 1\right| B_{\mathcal{P},\mathcal{E}} + g_h(\beta)B_{r,\mathcal{E}}\right)\right) \\ &= \sum_{\mathcal{E}=1}^{\lceil\frac{M}{W}\rceil}\sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W}\sum_{h\in[H]}\left(e^{\beta\sum_{i=1}^{h-1} r_i^m}\zeta_{h+1}^m + 4e^{\beta(h-1)}\Gamma_h^m\right) + WH\left(\left|e^{\beta H} - 1\right| B_{\mathcal{P}} + g_1(\beta)B_r\right).\end{aligned} \tag{4.21}$$

We aim to control the terms in (4.21). Since $\left\{e^{\beta\sum_{i=1}^{h-1} r_i^m}\zeta_{h+1}^m\right\}$ is a martingale difference sequence satisfying $\left|e^{\beta\sum_{i=1}^{h-1} r_i^m}\zeta_{h+1}^m\right| \le 2|e^{\beta H}-1|$ for all $(m,h) \in [M]\times[H]$, by the Azuma-Hoeffding inequality, we have:

$$\mathcal{P}\left(\sum_{m\in[M]}\sum_{h\in[H]} e^{\beta\sum_{i=1}^{h-1} r_i^m}\zeta_{h+1}^m \ge t\right) \le \exp\left(-\frac{t^2}{8HM\left(e^{\beta H}-1\right)^2}\right), \quad \forall t>0.$$

Hence, with probability $1-\delta/2$, it holds that

$$\sum_{k\in[K]}\sum_{h\in[H]} e^{\beta(h-1)}\zeta_{h+1}^m \le \left(e^{\beta H}-1\right)\sqrt{2HM\log(2/\delta)} \le 2\left(e^{\beta H}-1\right)\sqrt{2HM\iota}, \tag{4.22}$$

where $\iota = \log(6H|\mathcal{S}||\mathcal{A}|W/\delta)$. Furthermore, recall the definition of Γ_h^m, we can derive

$$\begin{aligned}
&\sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W}\sum_{h\in[H]} e^{\beta(h-1)}\Gamma_h^m \\
&\le \sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W}\sum_{h\in[H]} \left(C_1\left|e^{\beta(H-h+1)}-1\right| + g_h(\beta)\right)\sqrt{|\mathcal{S}|\iota}\sqrt{\frac{1}{N_h^m\left(s_h^m,a_h^m\right)+1}} \\
&\le \left(C_1\left|e^{\beta H}-1\right| + g_1(\beta)\right)\sqrt{|\mathcal{S}|\iota}\sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W}\sum_{h\in[H]}\sqrt{\frac{1}{N_h^m\left(s_h^m,a_h^m\right)+1}} \\
&\overset{(i)}{\le} \left(C_1\left|e^{\beta H}-1\right| + g_1(\beta)\right)\sqrt{|\mathcal{S}|\iota}\sum_{h\in[H]}\sqrt{W}\sqrt{\sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W}\frac{1}{N_h^m\left(s_h^m,a_h^m\right)+1}} \\
&\le \left(C_1\left|e^{\beta H}-1\right| + g_1(\beta)\right)\sqrt{|\mathcal{S}|\iota}\sqrt{2H^2|S||A|W\iota}
\end{aligned}$$

where step (i) follows the Cauchy-Schwarz inequality and the last step holds by the pigeonhole principle. Thus, it holds that

$$\sum_{m\in[M]}\sum_{h\in[H]} e^{\beta(h-1)}\Gamma_h^m \le \left(C_1\left|e^{\beta H}-1\right| + e^{\beta H}|\beta|\right)\sqrt{2H^2|S|^2|A|\iota^2}\frac{M}{\sqrt{W}}. \tag{4.23}$$

Substituting (4.22) and (4.23) into (4.21) yields that

$$\begin{aligned}
\sum_{m\in[M]}\delta_1^m \le& 2\left|e^{\beta H}-1\right|\sqrt{2HM\iota} + \left(C_1\left|e^{\beta H}-1\right| + g_1(\beta)\right)\sqrt{2H^2|S|^2|A|\iota^2}\frac{M}{\sqrt{W}} \\
&+ WH\left(\left|e^{\beta H}-1\right|B_{\mathcal{P}} + g_1(\beta)B_r\right)
\end{aligned} \tag{4.24}$$

For $\beta > 0$, we have that $g_1(\beta) = e^{\beta H}\beta$ and the dynamic regret can be decomposed based on Lemma 43:

$$
\begin{aligned}
&\text{D-Regret}(M)\\
\leq& \sum_{m\in[M]} \frac{1}{\beta}\left[e^{\beta\cdot V_1^{*,m}(s_1^m)} - e^{\beta\cdot V_1^{m}(s_1^m)}\right] + \sum_{m\in[M]} \frac{1}{\beta}\left[e^{\beta\cdot V_1^{m}(s_1^m)} - e^{\beta\cdot V_1^{\pi^m,m}(s_1^m)}\right]\\
\leq& \frac{1}{\beta}\sum_{\mathcal{E}=1}^{\lceil \frac{M}{W}\rceil}\sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W} H\left(\left|e^{\beta H}-1\right| B_{\mathcal{P},\mathcal{E}} + g_1(\beta)B_{r,\mathcal{E}}\right) + \sum_{m\in[M]} \frac{1}{\beta}\left[e^{\beta\cdot V_1^{m}(s_1^m)} - e^{\beta\cdot V_1^{\pi^m,m}(s_1^m)}\right]\\
\leq& \frac{1}{\beta} WH\left(\left|e^{\beta H}-1\right| B_{\mathcal{P}} + g_1(\beta)B_r\right) + \frac{1}{\beta}\sum_{m\in[M]}\delta_1^m\\
\leq& \frac{1}{\beta}\left(2\left(e^{\beta H}-1\right)\sqrt{2HM\iota} + \left(C_1\left(e^{\beta H}-1\right) + e^{\beta H}\beta\right)\sqrt{2H^2|S|^2|A|\iota^2}\frac{M}{\sqrt{W}}\right.\\
&\left. + WH\left(\left(e^{\beta H}-1\right)B_{\mathcal{P}} + e^{\beta H}\beta B_r\right)\right)\\
\leq& 2e^{\beta H}H\sqrt{2HM\iota} + e^{\beta H}\left(C_1 H + 1\right)\sqrt{2H^2|S|^2|A|\iota^2}\frac{M}{\sqrt{W}} + WHe^{\beta H}\left(HB_{\mathcal{P}} + B_r\right)\\
\leq& 2e^{\beta H}H\sqrt{2HM\iota} + (C_1+1)e^{\beta H}H\sqrt{2H^2|S|^2|A|\iota^2}\frac{M}{\sqrt{W}} + WH^2e^{\beta H}\left(B_{\mathcal{P}} + B_r\right) \qquad (4.25)
\end{aligned}
$$

where the second inequality follows from Lemma 28, the third inequality holds because of the definition of $B_{\mathcal{P}}$, B_r and δ_1^m, the forth inequality is due to (4.24), and the fifth inequality follows from $e^{\beta H} - 1 \leq \beta H e^{\beta H}$ for $\beta > 0$.

Finally, by setting $W = M^{\frac{2}{3}}\left(B_{\mathcal{P}} + B_r\right)^{-\frac{2}{3}}|S|^{\frac{2}{3}}|A|^{\frac{1}{3}}$, we conclude that

$$
\text{D-Regret(M)} \leq \widetilde{\mathcal{O}}\left(e^{\beta H}|S|^{\frac{2}{3}}|A|^{\frac{1}{3}}H^2M^{\frac{2}{3}}\left(B_{\mathcal{P}} + B_r\right)^{\frac{1}{3}}\right).
$$

The proof of $\beta < 0$ follows a similar procedure and is therefore omitted.

4.B Proof of Theorem 12

Preliminaries

We first lay out some additional notations to facilitate our proof. Let N_h^m, G_h^m, V_h^m be N_h, G_h, V_h at the beginning of episode m, before t is updated. We also set $Q_h^m := \frac{1}{\beta}G_h^m$. Let $\widehat{P}_h^m(\cdot \mid s,a)$ denote the delta function centered at s_{h+1}^m for all $(m,h,s,a) \in [M]\times[H]\times\mathcal{S}\times\mathcal{A}$. This means that $\mathbb{E}_{s'\sim\widehat{P}_h^m(\cdot|s,a)}\left[f\left(s'\right)\right] = f\left(s_{h+1}^m\right)$ for every $f : \mathcal{S} \to \mathbb{R}$. Denote by $n_h^m := N_h^m\left(s_h^m, a_h^m\right)$. Recall from Algorithm 4 that the learning rate is defined as

$$
\alpha_t := \frac{H+1}{H+t}
$$

for $t \in \mathbb{Z}$. We also define

$$\alpha_t^0 := \prod_{j=1}^{t} (1-\alpha_j), \quad \alpha_t^i := \alpha_i \prod_{j=i+1}^{t} (1-\alpha_j) \tag{4.26}$$

for integers $i, t \geq 1$. We set $\alpha_t^0 = 1$ and $\sum_{i\in[t]} \alpha_t^i = 0$ if $t = 0$, and $\alpha_t^i = \alpha_i$ if $t < i+1$.

The epoch is defined as an interval that starts at the first episode after a restart and ends at the first time when the restart is triggered. In Algorithm 4, the restart mechanism divides M episodes into $\lceil \frac{M}{W} \rceil$ epochs.

Define the shorthand notation $\iota := \log(|\mathcal{S}||\mathcal{A}|MH/\delta)$ for $\delta \in (0,1]$. We fix a tuple $(m,h,s,a) \in [M]\times[H]\times\mathcal{S}\times\mathcal{A}$ with $m_i^{\mathcal{E}} \leq M$ being the episode in which (s,a) is visited the i-th time at step h in epoch $\mathcal{E}$. Let us define

$$q_{h,1}^{m,+}(s,a) := \alpha_t^0 e^{\beta(H-h+1)} + \begin{cases} \sum_{i\in[t]} \alpha_t^i \left[e^{\beta\left[r_h^{m_i^{\mathcal{E}}}(s,a) + V_{h+1}^{m_i^{\mathcal{E}}}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]} + \Gamma_{h,i} \right], & \text{if } \beta > 0, \\ \sum_{i\in[t]} \alpha_t^i \left[e^{\beta\left[r_h^{m_i^{\mathcal{E}}}(s,a) + V_{h+1}^{m_i^{\mathcal{E}}}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]} - \Gamma_{h,i} \right], & \text{if } \beta < 0, \end{cases}$$

$$q_{h,1}^{m}(s,a) := \begin{cases} \min\left\{ q_{h,1}^{m,+}(s,a), e^{\beta(H-h+1)} \right\}, & \text{if } \beta > 0, \\ \max\left\{ q_{h,1}^{m,+}(s,a), e^{\beta(H-h+1)} \right\}, & \text{if } \beta < 0, \end{cases}$$

and

$$q_{h,2}^{m,\circ}(s,a) := \alpha_t^0 e^{\beta(H-h+1)} + \sum_{i\in[t]} \alpha_t^i \left[e^{\beta\left[r_h^{m_i^{\mathcal{E}}}(s,a) + V_{h+1}^{*,m_i^{\mathcal{E}}}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]} \right]$$

$$q_{h,2}^{m,+}(s,a) := \alpha_t^0 e^{\beta(H-h+1)} + \begin{cases} \sum_{i\in[t]} \alpha_t^i \left[e^{\beta\left[r_h^{m_i^{\mathcal{E}}}(s,a) + V_{h+1}^{*,m_i^{\mathcal{E}}}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]} + \Gamma_{h,i} \right], & \text{if } \beta > 0 \\ \sum_{i\in[t]} \alpha_t^i \left[e^{\beta\left[r_h^{m_i^{\mathcal{E}}}(s,a) + V_{h+1}^{*,m_i^{\mathcal{E}}}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]} - \Gamma_{h,i} \right], & \text{if } \beta < 0 \end{cases}$$

$$q_{h,2}^{m}(s,a) := \begin{cases} \min\left\{ q_{h,2}^{m,+}(s,a), e^{\beta(H-h+1)} \right\}, & \text{if } \beta > 0 \\ \max\left\{ q_{h,2}^{m,+}(s,a), e^{\beta(H-h+1)} \right\}, & \text{if } \beta < 0 \end{cases}$$

and

$$q_{h,3}^{m}(s,a) := \alpha_t^0 e^{\beta\cdot Q_h^{*,m}(s,a)} + \sum_{i\in[t]} \alpha_t^i \left[\mathbb{E}_{s'\sim P_h^m(\cdot|s,a)} e^{\beta\left[r_h^m(s,a) + V_{h+1}^{*,m}(s')\right]} \right].$$

By the definition of $q_{h,2}^{m,\circ}$, $q_{h,2}^{m,+}$ and $q_{h,2}^{m}$, it can be seen that $q_{h,2}^{m,\circ} \leq q_{h,2}^{m}$ if $\beta > 0$, and $q_{h,2}^{m,\circ} \geq q_{h,2}^{m}$ if $\beta < 0$. In addition, by definition, we have $\left(e^{\beta\cdot Q_h^m} - e^{\beta\cdot Q_h^{*,m}}\right)(s,a) = \left(q_{h,1}^m - q_{h,3}^m\right)(s,a)$.

Value Difference Bounds

Lemma 29. *For every triple (s,a,h) and episodes m_1, m_2 in the epoch $\mathcal{E}$, it holds that $\left|V_h^{*,m_1}(s) - V_h^{*,m_2}(s)\right| \le B_{r,\mathcal{E}} + HB_{\mathcal{P},\mathcal{E}}$.*

Proof. Let $a_1 = \arg\max_a Q_h^{*,m_1}(s,a)$ and $a_2 = \arg\max_a Q_h^{*,m_2}(s,a)$, it holds that

$$\begin{aligned} V_h^{*,m_1}(s) = Q_h^{*,m_1}(s,a_1) \ge Q_h^{*,m_1}(s,a_2) \ge & Q_h^{*,m_2}(s,a_2) - B_{r,\mathcal{E}} - HB_{\mathcal{P},\mathcal{E}} \\ = & V_h^{*,m_2}(s) - B_{r,\mathcal{E}} - HB_{\mathcal{P},\mathcal{E}} \end{aligned}$$

where the second inequality follows from [87, Lemma 1]. Similarly, we have

$$V_h^{*,m_2}(s) \ge V_h^{*,m_1}(s) - B_{r,\mathcal{E}} - HB_{\mathcal{P},\mathcal{E}}.$$

This completes the proof. □

Lemma 30. *For every $(m,h,s,a) \in [M]\times[H]\times\mathcal{S}\times\mathcal{A}$ and $m_1,\dots,m_t < m$ with $t = N_h^m(s,a)$, we have*

$$\begin{aligned} &\left|\sum_{i\in[t]} \alpha_t^i \left[e^{\beta\left[r_h^{m_i^{\mathcal{E}}}(s,a) + V_{h+1}^{*,m_i^{\mathcal{E}}}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]} - \mathbb{E}_{s'\sim P_h^m(\cdot|s,a)}\left[e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m}(s')\right]}\right]\right]\right| \\ \le & \Gamma_{h,t} + 2g_h(\beta)B_{r,\mathcal{E}} + \left(g_h(\beta)H + \left|e^{\beta(H-h+1)} - 1\right|\right) B_{\mathcal{P},\mathcal{E}} \end{aligned}$$

with probability at least $1-\delta$, and

$$\sum_{i\in[t]} \alpha_t^i \Gamma_{h,i} \in \left[\Gamma_{h,t}, 2\Gamma_{h,t}\right],$$

where $\Gamma_{h,t}$ is defined in (4.6).

Proof. For every $(m,h,s,a) \in [M]\times[H]\times\mathcal{S}\times\mathcal{A}$, we have the following decomposition:

$$\begin{aligned} & e^{\beta\left[r_h^{m_i^{\mathcal{E}}}(s,a) + V_{h+1}^{*,m_i^{\mathcal{E}}}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]} - \mathbb{E}_{s'\sim P_h^m(\cdot|s,a)}\left[e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m}(s')\right]}\right] \\ = & e^{\beta\left[r_h^{m_i^{\mathcal{E}}}(s,a) + V_{h+1}^{*,m_i^{\mathcal{E}}}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]} - e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m_i^{\mathcal{E}}}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]} && \text{(4.27a)} \\ & + e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m_i^{\mathcal{E}}}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]} - e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]} && \text{(4.27b)} \\ & + e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]} - \mathbb{E}_{s'\sim P_h^{m_i^{\mathcal{E}}}(\cdot|s,a)}\left[e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m}(s')\right]}\right] && \text{(4.27c)} \\ & + \mathbb{E}_{s'\sim P_h^{m_i^{\mathcal{E}}}(\cdot|s,a)}\left[e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m}(s')\right]}\right] - \mathbb{E}_{s'\sim P_h^m(\cdot|s,a)}\left[e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m}(s')\right]}\right]. && \text{(4.27d)} \end{aligned}$$

For the terms in (4.27a), it holds that

$$\begin{aligned} \left|e^{\beta\left[r_h^{m_i^{\mathcal{E}}}(s,a) + V_{h+1}^{*,m_i^{\mathcal{E}}}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]} - e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m_i^{\mathcal{E}}}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]}\right| \le & g_h(\beta)\left|r_h^{m_i^{\mathcal{E}}}(s,a) - r_h^m(s,a)\right| \\ \le & g_h(\beta)B_{r,\epsilon}, \end{aligned} \tag{4.28}$$

where the first inequality follows from the Lipschitz continuity of $e^{\beta x}$ with respect to x and the second inequality is due to the definition of the local variation budget $B_{r,\epsilon}$.

For the terms in (4.27b), it holds that

$$\begin{aligned}\left|e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m_i^{\mathcal{E}}}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]}-e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]}\right| &\leq g_h(\beta)\left|V_{h+1}^{*,m_i^{\mathcal{E}}}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)-V_{h+1}^{*,m}(s_{h+1}^{m_i^{\mathcal{E}}})\right| \\ &\leq g_h(\beta)\left(B_{r,\mathcal{E}}+HB_{\mathcal{P},\mathcal{E}}\right) \end{aligned} \tag{4.29}$$

where the second inequality follows from Lemma 29.

For the terms in (4.27d), we have

$$\begin{aligned}&\left|\mathbb{E}_{s'\sim P_h^{m_i^{\mathcal{E}}}(\cdot|s,a)}\left[e^{\beta[r_h^m(s,a)+V_{h+1}^{*,m}(s')]}\right]-\mathbb{E}_{s'\sim P_h^m(\cdot|s,a)}\left[e^{\beta[r_h^m(s,a)+V_{h+1}^{*,m}(s')]}\right]\right| \\ &=\left|\mathbb{E}_{s'\sim P_h^{m_i^{\mathcal{E}}}(\cdot|s,a)}\left[e^{\beta[r_h^m(s,a)+V_{h+1}^{*,m}(s')]}-1\right]-\mathbb{E}_{s'\sim P_h^m(\cdot|s,a)}\left[e^{\beta[r_h^m(s,a)+V_{h+1}^{*,m}(s')]}-1\right]\right| \\ &\leq\left|e^{\beta(H-h+1)}-1\right|B_{\mathcal{P},\mathcal{E}} \end{aligned} \tag{4.30}$$

where the first step follows from $\mathcal{P}_h^m 1(s,a)=\mathcal{P}_h^\tau 1(s,a)$ for all $\tau\in[\ell^m,m-1]$ and the last step holds by the definition of $B_{\mathcal{P},\mathcal{E}}$.

We now analyze the terms in (4.27c). For every $(m,h,s,a)\in[M]\times[H]\times\mathcal{S}\times\mathcal{A}$, we define

$$\psi(i,m,h,s,a)\coloneqq e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]}-\mathbb{E}_{s'\sim P_h^{m_i^{\mathcal{E}}}(\cdot|s,a)}\left[e^{\beta[r_h^m(s,a)+V_{h+1}^{*,m}(s')]}\right].$$

For a fix tuple $(m,h,s,a)\in[M]\times[H]\times\mathcal{S}\times\mathcal{A}$, $\{\psi(i,m,h,s,a)\}_{i\in[t]}$ with $t=N_h^m(s,a)$ is a martingale difference sequence. By the Azuma-Hoeffding inequality, with probability at least $1-\delta/(HM|\mathcal{S}||\mathcal{A}|)$, it holds that

$$\left|\sum_{i\in[t]}\alpha_t^i\cdot\psi(i,m,h,s,a)\right|\leq\frac{C_2}{2}\left|e^{\beta(H-h+1)}-1\right|\sqrt{\iota\sum_{i\in[t]}(\alpha_t^i)^2}\leq C_2\left|e^{\beta(H-h+1)}-1\right|\sqrt{\frac{H\iota}{t}}$$

where $C_2>0$ is some universal constant, the first step holds since $r_h(s,a)+V_{h+1}^*(s')\in[0,H-h+1]$ for $s'\in\mathcal{S}$, and the last step follows from the second property in Lemma 45. Then, applying the union bound over $(m,h,s,a)\in[M]\times[H]\times\mathcal{S}\times\mathcal{A}$, we have that the following holds for all $(m,h,s,a)\in[M]\times[H]\times\mathcal{S}\times\mathcal{A}$ with probability at least $1-\delta$:

$$\left|\sum_{i\in[t]}\alpha_t^i\cdot\psi(i,m,h,s,a)\right|\leq C_2\left|e^{\beta(H-h+1)}-1\right|\sqrt{\frac{H\iota}{t}}, \tag{4.31}$$

where $t=N_h^m(s,a)$.

Finally, by combining Equations (4.28)-(4.31) and noticing that $\sum_{i\in[t]}\alpha_t^i = 1$ from the forth property in Lemma 45, we have

$$\begin{aligned}
&\left|\sum_{i\in[t]}\alpha_t^i\left[e^{\beta\left[r_h^{m_i^{\mathcal{E}}}(s,a)+V_{h+1}^{*,m_i^{\mathcal{E}}}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]}-\mathbb{E}_{s'\sim P_h^m(\cdot|s,a)}\left[e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m}(s')\right]}\right]\right]\right| \\
\leq & C_2\left|e^{\beta(H-h+1)}-1\right|\sqrt{\frac{H\iota}{t}}+2g_h(\beta)B_{r,\mathcal{E}}+\left(g_h(\beta)H+\left|e^{\beta(H-h+1)}-1\right|\right)B_{\mathcal{P},\mathcal{E}}
\end{aligned}$$

For bounds on $\sum_{i\in[t]}\alpha_t^i\Gamma_{h,i}$, we recall the definition of $\{\Gamma_{h,t}\}$ in (4.6) and compute

$$\begin{aligned}
\sum_{i\in[t]}\alpha_t^i\Gamma_{h,i} &= C_2\left|e^{\beta(H-h+1)}-1\right|\sum_{i\in[t]}\alpha_t^i\sqrt{\frac{H\iota}{i}} \\
&\in\left[C_2\left|e^{\beta(H-h+1)}-1\right|\sqrt{\frac{H\iota}{t}},2C_2\left|e^{\beta(H-h+1)}-1\right|\sqrt{\frac{H\iota}{t}}\right]
\end{aligned}$$

where the last step holds by the first property in Lemma 45. □

Lemma 31. *For all (m,h,s,a) and $\delta\in(0,1]$, the following statements hold with probability at least $1-\delta$:*

- *If $\beta>0$:*

$$\begin{aligned}
&-2e^{\beta(H-h+1)}\beta B_{r,\mathcal{E}}-\left(e^{\beta(H-h+1)}\beta H+\left(e^{\beta(H-h+1)}-1\right)\right)B_{\mathcal{P},\mathcal{E}}\leq q_{h,2}^m(s,a)-q_{h,3}^m(s,a) \\
&\leq\alpha_t^0\left(e^{\beta(H-h+1)}-1\right)+2\sum_{i\in[t]}\alpha_t^i\Gamma_{h,i}+2e^{\beta(H-h+1)}\beta B_{r,\mathcal{E}}+\left(e^{\beta(H-h+1)}\beta H+\left(e^{\beta(H-h+1)}-1\right)\right)B_{\mathcal{P},\mathcal{E}}.
\end{aligned}$$

- *If $\beta<0$:*

$$\begin{aligned}
&2\beta B_{r,\mathcal{E}}-\left(-\beta H+\left(1-e^{\beta(H-h+1)}\right)\right)B_{\mathcal{P},\mathcal{E}}\leq q_{h,3}^m(s,a)-q_{h,2}^m(s,a) \\
&\leq\alpha_t^0\left(1-e^{\beta(H-h+1)}\right)+2\sum_{i\in[t]}\alpha_t^i\Gamma_{h,i}-2\beta B_{r,\mathcal{E}}+\left(-\beta H+\left(1-e^{\beta(H-h+1)}\right)\right)B_{\mathcal{P},\mathcal{E}}.
\end{aligned}$$

Proof. We focus on the case where $\beta>0$ and the case for $\beta<0$ can be proved similarly. By the definition of $q_{h,2}^{m,+}$ and $q_{h,3}^m$, it holds that

$$\begin{aligned}
q_{h,2}^{m,+}-q_{h,3}^m =&\alpha_t^0\left(e^{\beta(H-h+1)}-e^{\beta Q_h^{*,m}(s,a)}\right) \\
&+\sum_{i\in[t]}\alpha_t^i\left[e^{\beta\left[r_h^{m_i^{\mathcal{E}}}(s,a)+V_{h+1}^{*,m_i^{\mathcal{E}}}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)\right]}+\Gamma_{h,i}-\mathbb{E}_{s'\sim P_h^m(\cdot|s,a)}e^{\beta\left[r_h^m(s,a)+V_{h+1}^{*,m}(s')\right]}\right].
\end{aligned}$$

Due to $e^{\beta(H-h+1)}\geq e^{\beta Q_h^{*,m}(s,a)}\geq 1$ and Lemma 30, we have

$$q_{h,2}^{m,+}-q_{h,3}^m\geq -2g_h(\beta)B_{r,\mathcal{E}}-\left(g_h(\beta)H+\left|e^{\beta(H-h+1)}-1\right|\right)B_{\mathcal{P},\mathcal{E}}$$

and

$$q_{h,2}^{m,+} - q_{h,3}^{m} \leq \alpha_t^0 \left(e^{\beta(H-h+1)} - 1\right) + 2 \sum_{i \in [t]} \alpha_t^i \Gamma_{h,i} + 2 g_h(\beta) B_{r,\mathcal{E}} + \left(g_h(\beta) H + \left|e^{\beta(H-h+1)} - 1\right|\right) B_{\mathcal{P},\mathcal{E}}.$$

Furthermore, if $q_{h,2}^{m,+} \leq e^{\beta(H-h+1)}$, then we have $q_{h,2}^{m} = q_{h,2}^{m,+}$. On the other hand, if $q_{h,2}^{m,+} \geq e^{\beta(H-h+1)}$, then $q_{h,2}^{m} = e^{\beta(H-h+1)} \leq q_{h,2}^{m,+}$. Thus, it holds that $0 \leq q_{h,2}^{m} - q_{h,3}^{m} \leq q_{h,2}^{m,+} - q_{h,3}^{m}$. This completes the proof. □

The next two lemmas compare the iterate $e^{\beta \cdot Q_h^m}$ (and $e^{\beta \cdot V_h^m}$) with the optimal exponential value function $e^{\beta \cdot Q_h^{*,m}}$ (and $e^{\beta \cdot V_h^{*,m}}$).

Lemma 32. *For all (m, h, s, a) and $\delta \in (0, 1]$, it holds with probability at least $1 - \delta$:*

- *If $\beta > 0$:*

$$\begin{aligned}&(e^{\beta Q_h^m} - e^{\beta Q_h^{*,m}})(s,a) \\ \geq &-(H-h+1)\left(2e^{\beta(H-h+1)}\beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h+1)}\beta H + \left(e^{\beta(H-h+1)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}}\right).\end{aligned}$$

- *If $\beta < 0$:*

$$(e^{\beta Q_h^m} - e^{\beta Q_h^{*,m}})(s,a) \leq (H-h+1)\left(-2\beta B_{r,\mathcal{E}} + \left(-\beta H + \left(1 - e^{\beta(H-h+1)}\right)\right) B_{\mathcal{P},\mathcal{E}}\right).$$

Proof. We focus only on the case where $\beta > 0$ since the proof for $\beta < 0$ is similar. For the purpose of the proof, we set $Q_{H+1}^m(s,a) = Q_{H+1}^*(s,a) = 0$ for all $(m, s, a) \in [M] \times \mathcal{S} \times \mathcal{A}$. We fix a pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ and use strong induction on m and h. Without loss of generality, we assume that there exists a pair (m,h) such that $(s,a) = \left(s_h^m, a_h^m\right)$ (that is, (s,a) has been visited at some point in Algorithm 4), since otherwise $e^{\beta \cdot Q_h^m(s,a)} = e^{\beta(H-h+1)} \geq e^{\beta \cdot Q_h^*(s,a)}$ for all $(m,h) \in [M] \times [H]$ and we are done.

The base case for $m = 1$ and $h = H+1$ is satisfied since $e^{\beta \cdot Q_{H+1}^{m'}(s,a)} = e^{\beta \cdot Q_{H+1}^{*,m}(s,a)}$ for $m' \in [M]$ by definition. We fix a pair $(m,h) \in [M] \times [H]$ and assume that

$$e^{\beta \cdot Q_{h+1}^{m_i^{\mathcal{E}}}(s,a)} - e^{\beta \cdot Q_{h+1}^{*,m_i^{\mathcal{E}}}(s,a)} \geq -(H-h)\left(2e^{\beta(H-h)}\beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h)}\beta H + \left(e^{\beta(H-h)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}}\right)$$

for each $m_1^{\mathcal{E}}, \ldots, m_t^{\mathcal{E}}$ (here $t = N_h^m(s,a)$). We have for $i \in [t]$ that

$$\begin{aligned}e^{\beta \cdot V_{h+1}^{m_i^{\mathcal{E}}}(s)} =& \max_{a' \in \mathcal{A}} e^{\beta \cdot Q_{h+1}^{m_i^{\mathcal{E}}}(s,a')} - (H-h)\left(2e^{\beta(H-h)}\beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h)}\beta H + \left(e^{\beta(H-h)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}}\right) \\ \geq& \max_{a' \in \mathcal{A}} e^{\beta \cdot Q_{h+1}^{*,m_i^{\mathcal{E}}}(s,a')} - (H-h)\left(2e^{\beta(H-h)}\beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h)}\beta H + \left(e^{\beta(H-h)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}}\right) \\ =& e^{\beta \cdot V_{h+1}^{*,m_i^{\mathcal{E}}}(s)} - (H-h)\left(2e^{\beta(H-h)}\beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h)}\beta H + \left(e^{\beta(H-h)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}}\right) \qquad (4.32)\end{aligned}$$

where the first equality holds by the update procedure in Algorithm 4. Then, it holds that

$$\begin{aligned}
(q^{m,+}_{h,1} - q^{m}_{h,2})(s,a) &\geq (q^{m,+}_{h,1} - q^{m,+}_{h,2})(s,a) \\
&\geq \sum_{i\in[t]} \alpha^i_t \left[e^{\beta\left[r_h^{m^{\mathcal{E}}_i}(s,a) + V_{h+1}^{m^{\mathcal{E}}_i}\left(s_{h+1}^{m^{\mathcal{E}}_i}\right)\right]} - e^{\beta\left[r_h^{m^{\mathcal{E}}_i}(s,a) + V_{h+1}^{*,m^{\mathcal{E}}_i}\left(s_{h+1}^{m^{\mathcal{E}}_i}\right)\right]} \right] \\
&= \sum_{i\in[t]} \alpha^i_t e^{\beta r_h^{m^{\mathcal{E}}_i}(s,a)} \left[e^{\beta\left[V_{h+1}^{m^{\mathcal{E}}_i}\left(s_{h+1}^{m^{\mathcal{E}}_i}\right)\right]} - e^{\beta\left[V_{h+1}^{*,m^{\mathcal{E}}_i}\left(s_{h+1}^{m^{\mathcal{E}}_i}\right)\right]} \right] \\
&\geq -(H-h) \sum_{i\in[t]} \alpha^i_t e^{\beta} \left(2e^{\beta(H-h)}\beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h)}\beta H + \left(e^{\beta(H-h)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}}\right) \\
&\geq -(H-h) \sum_{i\in[t]} \alpha^i_t \left(2e^{\beta(H-h+1)}\beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h+1)}\beta H + \left(e^{\beta(H-h+1)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}}\right) \\
&\geq -(H-h) \left(2e^{\beta(H-h+1)}\beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h+1)}\beta H + \left(e^{\beta(H-h+1)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}}\right)
\end{aligned}$$

where the first inequality follows from the definitions of $q^{m,+}_{h,1}$, $q^{m,+}_{h,2}$, the second inequality holds by the induction hypothesis, the third inequality follows from $e^{\beta} > 1$ for $\beta > 0$, and the last inequality holds by $\sum_{i\in[t]} \alpha^i_t \leq 1$ from Lemma 45. Furthermore, when $q^{m}_{h,1} = e^{\beta(H-h+1)} \leq q^{m,+}_{h,1}$, we have $q^{m}_{h,1} - q^{m}_{h,2} \geq 0$ since $q^{m}_{h,2} \leq e^{\beta(H-h+1)}$ by definition. Thus, we can conclude that

$$(q^{m}_{h,1} - q^{m}_{h,2})(s,a) \geq -(H-h)\left(2e^{\beta(H-h+1)}\beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h+1)}\beta H + \left(e^{\beta(H-h+1)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}}\right) \quad (4.33)$$

In addition, from Lemma 31 , we also have

$$(q^{m}_{h,2} - q^{m}_{h,3})(s,a) \geq -2e^{\beta(H-h+1)}\beta B_{r,\mathcal{E}} - \left(e^{\beta(H-h+1)}\beta H + \left(e^{\beta(H-h+1)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}} \quad (4.34)$$

Finally, by combining (4.33) and (4.34), we obtain

$$\begin{aligned}
&(e^{\beta Q^{m}_h} - e^{\beta Q^{*,m}_h})(s,a) \\
=&(q^{m}_{h,1} - q^{m}_{h,2})(s,a) + (q^{m}_{h,2} - q^{m}_{h,3})(s,a) \\
\geq& -(H-h+1)\left(2e^{\beta(H-h+1)}\beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h+1)}\beta H + \left(e^{\beta(H-h+1)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}}\right).
\end{aligned}$$

The induction is completed. □

Lemma 33. *For all (m,h,s,a) and $\delta \in (0,1]$, it holds with probability at least $1-\delta$:*

- *If $\beta > 0$:*

$$\begin{aligned}
&(e^{\beta Q^{m}_h} - e^{\beta Q^{*,m}_h})(s,a) \\
\leq& \sum_{i\in[t]} \alpha^i_t e^{\beta r_h^{m^{\mathcal{E}}_i}(s,a)} \left[e^{\beta\left[V_{h+1}^{m^{\mathcal{E}}_i}\left(s_{h+1}^{m^{\mathcal{E}}_i}\right)\right]} - e^{\beta\left[V_{h+1}^{*,m^{\mathcal{E}}_i}\left(s_{h+1}^{m^{\mathcal{E}}_i}\right)\right]} \right] + 3\sum_{i\in[t]} \alpha^i_t \Gamma_{h,i} \\
&+ \alpha^0_t \left(e^{\beta(H-h+1)} - 1\right) + 2e^{\beta(H-h+1)}\beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h+1)}\beta H + \left(e^{\beta(H-h+1)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}}.
\end{aligned}$$

- *If $\beta < 0$:*

$$
\begin{aligned}
&(e^{\beta Q_h^m} - e^{\beta Q_h^{*,m}})(s,a) \\
&\geq \sum_{i\in[t]} \alpha_t^i e^{\beta r_h^{m_i^\mathcal{E}}(s,a)} \left[e^{\beta\left[V_{h+1}^{*,m_i^\mathcal{E}}\left(s_{h+1}^{m_i^\mathcal{E}}\right)\right]} - e^{\beta\left[V_{h+1}^{m_i^\mathcal{E}}\left(s_{h+1}^{m_i^\mathcal{E}}\right)\right]} \right] - 3\sum_{i\in[t]} \alpha_t^i \Gamma_{h,i} \\
&\quad - \alpha_t^0 \left(1 - e^{\beta(H-h+1)}\right) + 2\beta B_{r,\mathcal{E}} - \left(-\beta H + \left(1 - e^{\beta(H-h+1)}\right)\right) B_{\mathcal{P},\mathcal{E}}.
\end{aligned}
$$

Proof. We focus on the case where $\beta > 0$ since the case for $\beta < 0$ can be proved similarly. By the definition of $q_{h,1}^m$ and $q_{h,2}^m$, we have

$$
\begin{aligned}
\left(q_{h,1}^m - q_{h,2}^m\right)(s,a) &\leq \left(q_{h,1}^{m,+} - q_{h,2}^{m,\circ}\right)(s,a) \\
&\leq \sum_{i\in[t]} \alpha_t^i \left[e^{\beta\left[r_h^{m_i^\mathcal{E}}(s,a) + V_{h+1}^{m_i^\mathcal{E}}\left(s_{h+1}^{m_i^\mathcal{E}}\right)\right]} - e^{\beta\left[r_h^{m_i^\mathcal{E}}(s,a) + V_{h+1}^{*,m_i^\mathcal{E}}\left(s_{h+1}^{m_i^\mathcal{E}}\right)\right]} \right] + \sum_{i\in[t]} \alpha_t^i \Gamma_{h,i} \\
&= \sum_{i\in[t]} \alpha_t^i e^{\beta r_h^{m_i^\mathcal{E}}(s,a)} \left[e^{\beta\left[V_{h+1}^{m_i^\mathcal{E}}\left(s_{h+1}^{m_i^\mathcal{E}}\right)\right]} - e^{\beta\left[V_{h+1}^{*,m_i^\mathcal{E}}\left(s_{h+1}^{m_i^\mathcal{E}}\right)\right]} \right] + \sum_{i\in[t]} \alpha_t^i \Gamma_{h,i}
\end{aligned}
$$

where the first inequality follows from $q_{h,1}^m \leq q_{h,1}^{m,+}$ and $q_{h,2}^m \geq q_{h,2}^{m,\circ}$, and the second inequality holds by the definition of $q_{h,1}^{m,+}$ and $q_{h,2}^{m,\circ}$. Then, by Lemma 31, we obtain

$$
\begin{aligned}
&(e^{\beta Q_h^m} - e^{\beta Q_h^{*,m}})(s,a) \\
=&(q_{h,1}^m - q_{h,2}^m)(s,a) + (q_{h,2}^m - q_{h,3}^m)(s,a) \\
\leq &\sum_{i\in[t]} \alpha_t^i e^{\beta r_h^{m_i^\mathcal{E}}(s,a)} \left[e^{\beta\left[V_{h+1}^{m_i^\mathcal{E}}\left(s_{h+1}^{m_i^\mathcal{E}}\right)\right]} - e^{\beta\left[V_{h+1}^{*,m_i^\mathcal{E}}\left(s_{h+1}^{m_i^\mathcal{E}}\right)\right]} \right] + 3\sum_{i\in[t]} \alpha_t^i \Gamma_{h,i} \\
&+ \alpha_t^0 \left(e^{\beta(H-h+1)} - 1\right) + 2e^{\beta(H-h+1)}\beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h+1)}\beta H + \left(e^{\beta(H-h+1)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}}.
\end{aligned}
$$

This completes the proof. □

Proof of Theorem 12

For now, we consider the case for $\beta > 0$. We define the following quantities to ease the notations for the proof:

$$
\begin{aligned}
\delta_h^m &:= e^{\beta \cdot V_h^m(s_h^m)} - e^{\beta \cdot V_h^{\pi^m}(s_h^m)}, \\
\phi_h^m &:= e^{\beta \cdot V_h^m(s_h^m)} - e^{\beta \cdot V_h^{*,m}(s_h^m)}, \\
\xi_{h+1}^m &:= \left[\left(\mathcal{P}_h^m - \widehat{\mathcal{P}}_h^m\right)\left(e^{\beta \cdot V_{h+1}^{*,m}} - e^{\beta \cdot V_{h+1}^{\pi^m}}\right)\right](s_h^m, a_h^m)
\end{aligned}
$$

For each fixed $(m,h) \in [M] \times [H]$, we let $t = N_h^m\left(s_h^m, a_h^m\right)$. Then, it holds that

$$
\begin{aligned}
\delta_h^m &\overset{(i)}{=} e^{\beta \cdot Q_h^m\left(s_h^m,a_h^m\right)} - e^{\beta \cdot Q_h^{\pi^m,m}\left(s_h^m,a_h^m\right)} \\
&= \left[e^{\beta \cdot Q_h^m\left(s_h^m,a_h^m\right)} - e^{\beta \cdot Q_h^{*,m}\left(s_h^m,a_h^m\right)}\right] + \left[e^{\beta \cdot Q_h^{*,m}\left(s_h^m,a_h^m\right)} - e^{\beta \cdot Q_h^{\pi^m}\left(s_h^m,a_h^m\right)}\right] \\
&\overset{(ii)}{=} \left[e^{\beta \cdot Q_h^m\left(s_h^m,a_h^m\right)} - e^{\beta \cdot Q_h^{*,m}\left(s_h^m,a_h^m\right)}\right] + e^{\beta \cdot r_h^m\left(s_h^m,a_h^m\right)}\left[\mathcal{P}_h^m\left(e^{\beta \cdot V_{h+1}^{*,m}} - e^{\beta \cdot V_{h+1}^{\pi^m,m}}\right)\right]\left(s_h^m, a_h^m\right) \\
&\overset{(iii)}{\leq} \left[e^{\beta \cdot Q_h^m\left(s_h^m,a_h^m\right)} - e^{\beta \cdot Q_h^{*,m}\left(s_h^m,a_h^m\right)}\right] + e^{\beta}\left[\mathcal{P}_h^m\left(e^{\beta \cdot V_{h+1}^{*,m}} - e^{\beta \cdot V_{h+1}^{\pi^m,m}}\right)\right]\left(s_h^m, a_h^m\right) \\
&= \left[e^{\beta \cdot Q_h^m\left(s_h^m,a_h^m\right)} - e^{\beta \cdot Q_h^{*,m}\left(s_h^m,a_h^m\right)}\right] + e^{\beta}\left(\delta_{h+1}^m - \phi_{h+1}^m + \xi_{h+1}^m\right) \\
&\overset{(iv)}{\leq} \alpha_t^0\left(e^{\beta(H-h+1)} - 1\right) + 3\sum_{i \in [t]} \alpha_t^i \Gamma_{h,i} + \sum_{i \in [t]} \alpha_t^i \cdot e^{\beta \cdot r_h^{m_i^{\mathcal{E}}}\left(s_h^m,a_h^m\right)}\left[e^{\beta \cdot V_{h+1}^{m_i^{\mathcal{E}}}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)} - e^{\beta \cdot V_{h+1}^{*}\left(s_{h+1}^{m_i^{\mathcal{E}}}\right)}\right] \\
&\quad + 2e^{\beta(H-h+1)}\beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h+1)}\beta H + \left(e^{\beta(H-h+1)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}} \\
&\quad + e^{\beta}\left(\delta_{h+1}^m - \phi_{h+1}^m + \xi_{h+1}^m\right) \\
&= \alpha_t^0\left(e^{\beta(H-h+1)} - 1\right) + \sum_{i \in [t]} \alpha_t^i \cdot e^{\beta \cdot r_h^{m_i^{\mathcal{E}}}\left(s_h^m,a_h^m\right)} \phi_{h+1}^{m_i^{\mathcal{E}}} + e^{\beta}\left(\delta_{h+1}^m - \phi_{h+1}^m + \xi_{h+1}^m\right) \qquad (4.35) \\
&\quad + 3\sum_{i \in [t]} \alpha_t^i \Gamma_{h,i} + 2e^{\beta(H-h+1)}\beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h+1)}\beta H + \left(e^{\beta(H-h+1)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}} \qquad (4.36)
\end{aligned}
$$

where step (i) holds since $V_h^m\left(s_h^m\right) = \max_{a' \in \mathcal{A}} Q_h^m\left(s_h^m, a'\right) = Q_h^m\left(s_h^m, a_h^m\right)$ and $V_h^{\pi^m,m}\left(s_h^m\right) = Q_h^{\pi^m,m}\left(s_h^m, \pi_h^m\left(s_h^m\right)\right) = Q_h^{\pi^m,m}\left(s_h^m, a_h^m\right)$; step (ii) holds by the exponential Bellman equation (4.2); step (iii) holds since $V_{h+1}^{*,m} \geq V_{h+1}^{\pi^m,m}$ implies $e^{\beta \cdot V_{h+1}^{*,m}} \geq e^{\beta \cdot V_{h+1}^{\pi^m,m}}$ given that $\beta > 0$; step (iv) holds on the event of Lemma 33.

We bound each term in (4.35) and (4.36) one by one. First, we have

$$
\begin{aligned}
\sum_{m \in [M]} \alpha_{n_h^m}^0\left(e^{\beta(H-h+1)} - 1\right) &= \left(e^{\beta(H-h+1)} - 1\right) \sum_{m \in [M]} \mathbb{1}\left\{n_h^m = 0\right\} \\
&\leq \left(e^{\beta(H-h+1)} - 1\right)|\mathcal{S}||\mathcal{A}|.
\end{aligned}
$$

To bound the second term in (4.35), we first define

$$
\hat{\phi}_{h+1}^{m_i^{\mathcal{E}}\left(s_h^m,a_h^m\right)} \coloneqq \phi_{h+1}^{m_i^{\mathcal{E}}\left(s_h^m,a_h^m\right)} + (H-h)\left(2e^{\beta(H-h)}\beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h)}\beta H + \left(e^{\beta(H-h)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}}\right)
$$

which is non-negative from Lemma 32 and (4.32):

$$
\begin{aligned}
\sum_{m \in [M]}\left(\sum_{i \in [t]} \alpha_t^i \cdot e^{\beta \cdot r_h^{m_i^{\mathcal{E}}}\left(s_h^m,a_h^m\right)} \phi_{h+1}^{m_i^{\mathcal{E}}}\right) &= \sum_{m \in [M]}\left(\sum_{i \in [n_h^m]} \alpha_{n_h^m}^i \cdot e^{\beta \cdot r_h^{m_i^{\mathcal{E}}}\left(s_h^m,a_h^m\right)} \phi_{h+1}^{m_i^{\mathcal{E}}\left(s_h^m,a_h^m\right)}\right) \\
&\leq e^{\beta} \sum_{m \in [M]}\left(\sum_{i \in [n_h^m]} \alpha_{n_h^m}^i \hat{\phi}_{h+1}^{m_i^{\mathcal{E}}\left(s_h^m,a_h^m\right)}\right)
\end{aligned}
$$

where $m_i^{\mathcal{E}}\left(s_h^m, a_h^m\right)$ denotes the episode in which $\left(s_h^m, a_h^m\right)$ was taken at step h for the i-th time in the epoch $\mathcal{E}$. We re-group the above summation by changing the order of the summation. For every $\hat{m}^{\mathcal{E}}$ in the epoch $\mathcal{E}$, the term $\phi_{h+1}^{\hat{m}^{\mathcal{E}}}$ appears in the summand with $m > \hat{m}^{\mathcal{E}}$ if and only if $\left(s_h^m, a_h^m\right) = \left(s_h^{m'}, a_h^{m'}\right)$ and the episode m is in the epoch $\mathcal{E}$. Since the inverse of the mapping $i \to m_i^{\mathcal{E}}(s_h^m, a_h^m)$ is $\hat{m}^{\mathcal{E}} \to n_h^{\hat{m}^{\mathcal{E}}}$, we can continue the above display as

$$
\begin{aligned}
e^{\beta} \sum_{m \in [M]} \left(\sum_{i \in \left[n_h^m\right]} \alpha_{n_h^m}^i \hat{\phi}_{h+1}^{m_i^{\mathcal{E}}\left(s_h^m a_h^m\right)} \right) &\leq e^{\beta} \sum_{\mathcal{E}=1}^{\lceil \frac{M}{W} \rceil} \sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W} \left(\sum_{i \in \left[n_h^m\right]} \alpha_{n_h^m}^i \hat{\phi}_{h+1}^{m_i^{\mathcal{E}}\left(s_h^m a_h^m\right)} \right) \\
&\leq e^{\beta} \sum_{\mathcal{E}=1}^{\lceil \frac{M}{W} \rceil} \sum_{m'=(\mathcal{E}-1)W}^{\mathcal{E}W} \hat{\phi}_{h+1}^{m'} \left(\sum_{t \geq n_h^{m'}+1} \alpha_t^{n_h^{m'}} \right) \\
&\leq e^{\beta} \left(1 + \frac{1}{H}\right) \sum_{\mathcal{E}=1}^{\lceil \frac{M}{W} \rceil} \sum_{m'=(\mathcal{E}-1)W}^{\mathcal{E}W} \hat{\phi}_{h+1}^{m'}
\end{aligned}
$$

where the last step follows the third property in Lemma 45. Collecting the above results and substituting them into (4.35)-(4.36), we have

$$
\begin{aligned}
\sum_{m \in [M]} \delta_h^m \leq & \left(e^{\beta(H-h+1)} - 1\right) |\mathcal{S}||\mathcal{A}| + \left(1 + \frac{1}{H}\right) e^{\beta} \sum_{\mathcal{E}=1}^{\lceil \frac{M}{W} \rceil} \sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W} \hat{\phi}_{h+1}^m \\
& + \sum_{\mathcal{E}=1}^{\lceil \frac{M}{W} \rceil} \sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W} e^{\beta} \left(\delta_{h+1}^m - \phi_{h+1}^m + \xi_{h+1}^m\right) + 3 \sum_{\mathcal{E}=1}^{\lceil \frac{M}{W} \rceil} \sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W} \sum_{i \in [t]} \alpha_t^i \Gamma_{h,i} \\
& + \sum_{\mathcal{E}=1}^{\lceil \frac{M}{W} \rceil} \sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W} \left(2e^{\beta(H-h+1)} \beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h+1)} \beta H + \left(e^{\beta(H-h+1)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}}\right) \\
\leq & \left(e^{\beta(H-h+1)} - 1\right) |\mathcal{S}||\mathcal{A}| + \left(1 + \frac{1}{H}\right) \sum_{\mathcal{E}=1}^{\lceil \frac{M}{W} \rceil} \sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W} e^{\beta} \delta_{h+1}^m \\
& + \sum_{\mathcal{E}=1}^{\lceil \frac{M}{W} \rceil} \sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W} \left(3 \sum_{i \in [t]} \alpha_t^i \Gamma_{h,i} + e^{\beta} \xi_{h+1}^m\right) \\
& + 3(H-h) \sum_{\mathcal{E}=1}^{\lceil \frac{M}{W} \rceil} \sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W} \left(2e^{\beta(H-h+1)} \beta B_{r,\mathcal{E}} + \left(e^{\beta(H-h+1)} \beta H + \left(e^{\beta(H-h+1)} - 1\right)\right) B_{\mathcal{P},\mathcal{E}}\right) \\
\leq & \left(e^{\beta(H-h+1)} - 1\right) |\mathcal{S}||\mathcal{A}| + \left(1 + \frac{1}{H}\right) \sum_{m \in [M]} e^{\beta} \delta_{h+1}^m \\
& + 3 \sum_{\mathcal{E}=1}^{\lceil \frac{M}{W} \rceil} \sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W} \sum_{i \in [t]} \alpha_t^i \Gamma_{h,i} + \sum_{m \in [M]} e^{\beta} \xi_{h+1}^m \\
& + 3(H-h) \left(2e^{\beta(H-h+1)} \beta W B_r + \left(e^{\beta(H-h+1)} \beta H + \left(e^{\beta(H-h+1)} - 1\right)\right) W B_{\mathcal{P}}\right)
\end{aligned}
$$

where the second step holds since $\delta_{h+1}^m \geq \phi_{h+1}^m$ (due to the fact that $\beta > 0$ and $V_{h+1}^{*,m} \geq V_{h+1}^{\pi^m,m}$) and the definition of $\hat{\phi}_{h+1}^m$; the last step follows from the definition of B_r and $B_{\mathcal{P}}$. Now, we unroll the quantity $\sum_{m\in[M]} \delta_h^m$ recursively in the form of Equation (36), and get

$$
\begin{aligned}
&\sum_{m\in[M]} \delta_1^m \qquad (4.37)\\
&\leq \sum_{h\in[H]} \left[\left(1+\frac{1}{H}\right)e^{\beta}\right]^{h-1} \Bigg[\left(e^{\beta(H-h+1)}-1\right)|\mathcal{S}||\mathcal{A}| + 3\sum_{\mathcal{E}=1}^{\lceil \frac{M}{W} \rceil} \sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W} \sum_{i\in[t]} \alpha_t^i \Gamma_{h,i} + \sum_{m\in[M]} \left(e^{\beta}\xi_{h+1}^m\right)\\
&\quad +3(H-h)\left(2e^{\beta(H-h+1)}\beta W B_r + \left(e^{\beta(H-h+1)}\beta H + \left(e^{\beta(H-h+1)}-1\right)\right) W B_{\mathcal{P}}\right)\Bigg]\\
&\leq \sum_{h\in[H]} \left(1+\frac{1}{H}\right)^{h-1} \Bigg[\left(e^{\beta H}-1\right)|\mathcal{S}||\mathcal{A}| + 3\sum_{\mathcal{E}=1}^{\lceil \frac{M}{W} \rceil} \sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W} e^{\beta(h-1)} \sum_{i\in[t]} \alpha_t^i \Gamma_{h,i} + \sum_{m\in[M]} e^{\beta h}\xi_{h+1}^m\\
&\quad +3(H-h)\left(2e^{\beta H}\beta W B_r + \left(e^{\beta H}\beta H + \left(e^{\beta H}-1\right)\right) W B_{\mathcal{P}}\right)\Bigg]\\
&\leq e\left[\left(e^{\beta H}-1\right)H|\mathcal{S}||\mathcal{A}| + 3e\sum_{\mathcal{E}=1}^{\lceil \frac{M}{W} \rceil} \sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W} \sum_{h\in[H]} e^{\beta(h-1)} \sum_{i\in[t]} \alpha_t^i \Gamma_{h,i}\right] + \sum_{h\in[H]} \sum_{m\in[M]} \left(1+\frac{1}{H}\right)^{h-1} e^{\beta h}\xi_{h+1}^m\\
&\quad + 3eH^2\left(2e^{\beta H}\beta W B_r + \left(e^{\beta H}\beta H + \left(e^{\beta H}-1\right)\right) W B_{\mathcal{P}}\right)
\end{aligned}
$$

where the first step uses the fact that $\delta_{H+1}^m = 0$ for $m \in [M]$; the last step holds since $(1+1/H)^h \leq (1+1/H)^H \leq e$ for all $h \in [H]$. Furthermore, the definition of $\Gamma_{h,i}$ and Lemma 45 imply that

$$\sum_{i\in[t]} \alpha_t^i \Gamma_{h,i} \leq C_2\left(e^{\beta(H-h+1)-1}\right)\sqrt{\frac{H\iota}{t}}.$$

for some constant $C_2 > 0$. By the pigeonhole principle, for any $h \in [H]$ we have

$$
\begin{aligned}
\sum_{\mathcal{E}=1}^{\lceil \frac{M}{W} \rceil} \sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W} \sum_{h\in[H]} e^{\beta(h-1)} \sum_{i\in[t]} \alpha_t^i \Gamma_{h,i} &\leq C_2\left(e^{\beta H}-1\right)\sum_{\mathcal{E}=1}^{\lceil \frac{M}{W} \rceil} \sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W} \sqrt{\frac{H\iota}{n_h^m}}\\
&\leq C_2\left(e^{\beta H}-1\right)\sum_{\mathcal{E}=1}^{\lceil \frac{M}{W} \rceil} \sqrt{W}\sqrt{\sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W} \frac{H\iota}{n_h^m}}\\
&\leq C_2\left(e^{\beta H}-1\right) M\sqrt{H|\mathcal{S}||\mathcal{A}|\iota/W} \qquad (4.38)
\end{aligned}
$$

where the second step follows from the Cauchy-Schwarz inequality, the third step holds since $\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} N_h^m(s,a) = W$ and the right-hand side of the second step is maximized when $N_h^m(s,a) = W/(|\mathcal{S}||\mathcal{A}|)$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. Finally, the Azuma-Hoeffding inequality and the fact that $\left|\left(1+\frac{1}{H}\right)^{h-1} e^{\beta h}\xi_{h+1}^m\right| \leq e\left(e^{\beta H}-1\right)$ for $h \in [H]$ together imply that with probability at

least $1-\delta$, we have

$$\left|\sum_{h\in[H]}\sum_{m\in[M]}\left(1+\frac{1}{H}\right)^{h-1}e^{\beta h}\xi_{h+1}^{m}\right|\leq C_3\left(e^{\beta H}-1\right)\sqrt{HM\iota} \tag{4.39}$$

for some constant $C_3>0$. Plugging Equations (4.38) and (4.39) into (4.37), we have

$$\begin{aligned}\sum_{m\in[M]}\delta_1^m \leq&\mathcal{O}\Big(\left(e^{\beta H}-1\right)M\sqrt{H|\mathcal{S}||\mathcal{A}|\iota/W}+\left(e^{\beta H}-1\right)\sqrt{HM\iota}\\ &+H^2\left(2e^{\beta H}\beta WB_r+\left(e^{\beta H}\beta H+\left(e^{\beta H}-1\right)\right)WB_{\mathcal{P}}\right)\Big)\end{aligned} \tag{4.40}$$

when M is large enough. Invoking Lemma 43 yields that

$$\begin{aligned}&\text{D-Regret}(M)\\ \leq&\sum_{m\in[M]}\frac{1}{\beta}\left[e^{\beta\cdot V_1^{*,m}(s_1^m)}-e^{\beta\cdot V_1^{m}(s_1^m)}\right]+\sum_{m\in[M]}\frac{1}{\beta}\left[e^{\beta\cdot V_1^{m}(s_1^m)}-e^{\beta\cdot V_1^{\pi^m,m}(s_1^m)}\right]\\ \leq&\frac{1}{\beta}\sum_{\mathcal{E}=1}^{\lceil\frac{M}{W}\rceil}\sum_{m=(\mathcal{E}-1)W}^{\mathcal{E}W}H\left(2e^{\beta H}\beta B_{r,\mathcal{E}}+\left(e^{\beta H}\beta H+\left(e^{\beta H}-1\right)\right)B_{\mathcal{P},\mathcal{E}}\right)\\ &+\sum_{m\in[M]}\frac{1}{\beta}\left[e^{\beta\cdot V_1^{m}(s_1^m)}-e^{\beta\cdot V_1^{\pi^m,m}(s_1^m)}\right]\\ \leq&\frac{1}{\beta}WH\left(2e^{\beta H}\beta B_r+\left(e^{\beta H}\beta H+\left(e^{\beta H}-1\right)\right)B_{\mathcal{P}}\right)+\frac{1}{\beta}\sum_{m\in[M]}\delta_1^m\\ \leq&\frac{1}{\beta}WH\left(2e^{\beta H}\beta B_r+\left(e^{\beta H}\beta H+\left(e^{\beta H}-1\right)\right)B_{\mathcal{P}}\right)\\ &+\frac{1}{\beta}\mathcal{O}\Big(\left(e^{\beta H}-1\right)M\sqrt{H|\mathcal{S}||\mathcal{A}|\iota/W}+\left(e^{\beta H}-1\right)\sqrt{HM\iota}\\ &+\quad H^2\left(2e^{\beta H}\beta WB_r+\left(e^{\beta H}\beta H+\left(e^{\beta H}-1\right)\right)WB_{\mathcal{P}}\right)\Big)\\ \leq&\mathcal{O}\left(e^{\beta H}HM\sqrt{H|\mathcal{S}||\mathcal{A}|\iota/W}+e^{\beta H}H\sqrt{HM\iota}+H^2e^{\beta H}W\left(B_r+HB_{\mathcal{P}}\right)\right) &(4.41)\\ \leq&\widetilde{\mathcal{O}}\left(e^{\beta H}M\sqrt{H^3|\mathcal{S}||\mathcal{A}|/W}+e^{\beta H}\sqrt{H^3M}+H^3e^{\beta H}W\left(B_r+B_{\mathcal{P}}\right)\right) &(4.42)\end{aligned}$$

where the second step holds by (4.32), the third inequality holds because of the definition of $B_{\mathcal{P}}$, B_r and δ_1^m, the forth inequality is due to (4.40), and the fifth inequality follows from $e^{\beta H}-1\leq\beta He^{\beta H}$ for $\beta>0$. Finally, by setting $W=M^{\frac{2}{3}}H^{-\frac{3}{4}}\left(B_{\mathcal{P}}+B_r\right)^{-\frac{2}{3}}|S|^{\frac{1}{3}}|A|^{\frac{1}{3}}$, we conclude that

$$\text{D-Regret(M)}\leq\widetilde{\mathcal{O}}\left(e^{\beta H}|S|^{\frac{1}{3}}|A|^{\frac{1}{3}}H^{\frac{9}{4}}M^{\frac{2}{3}}\left(B_{\mathcal{P}}+B_r\right)^{\frac{1}{3}}\right).$$

The proof is similar for the case of $\beta < 0$, and one only needs to exchange the role of V_h^m, $V_h^{\pi^m,m}$ and $V_h^{*,m}$ in the definitions of $\delta_h^m, \phi_h^m, \xi_h^m$:

$$\delta_h^m := e^{\beta \cdot V_h^{\pi^m}}(s_h^m) - e^{\beta \cdot V_h^m}(s_h^m),$$
$$\phi_h^m := e^{\beta \cdot V_h^{*,m}}(s_h^m) - e^{\beta \cdot V_h^m}(s_h^m),$$
$$\xi_{h+1}^m := \left[\left(\mathcal{P}_h^m - \widehat{\mathcal{P}}_h^m\right)\left(e^{\beta \cdot V_{h+1}^{\pi^m}} - e^{\beta \cdot V_{h+1}^{*,m}}\right)\right](s_h^m, a_h^m)$$

to derive the counterparts of (4.35) and (4.36), and complete the remaining analysis.

4.C Proof of Theorem 13

Multi-Scale ALG Initialization

Algorithm 6 Multi-scale ALG Initialization (MALG-initialization)

1: **Inputs:** ALG and its associated $\rho(\cdot)$, n;
2: **for** $\tau = 0, \ldots, 2^n - 1$ **do**
3: **for** k=n,n-1,...,0 **do**
4: If τ is a multiple of 2^k, with probability $\frac{\rho(2^n)}{\rho(2^k)}$, schedule a new instance alg of ALG that starts at $alg.s = \tau + 1$ and ends at $alg.e = \tau + 2^k$
5: **end for**
6: **end for**

An Illustrative Example

For better illustration, we give an example with $n = 4$. This example has also been shown in [124] and we present here for completeness. By Algorithm 6, one possible realization of the MALG initialization is shown in Figure 4.C.1 with one order-4 instance (red), zero order-3 instance, two order-2 instances (green), two order-1 instances (purple) and five order-0 instances (blue). The bolder part of the segment indicates the period of time when the instances are active, while the thinner part indicates the inactive period. At any point of time, the active instance is always the one with the shortest length. The dashed arrow marked with ① indicates that ALG is executed as of the two sides of the arrow are concatenated. On the other hand, the two blue instances on the two sides of the dashed line marked with ② are two different order-0 instances, so the second one should start from scratch even though they are consecutive.

Preliminaries

Similar to [124], our approach takes a base algorithm that tackles the risk-sensitive RL problem when the environment is (near-)stationary, and turns it into another algorithm that

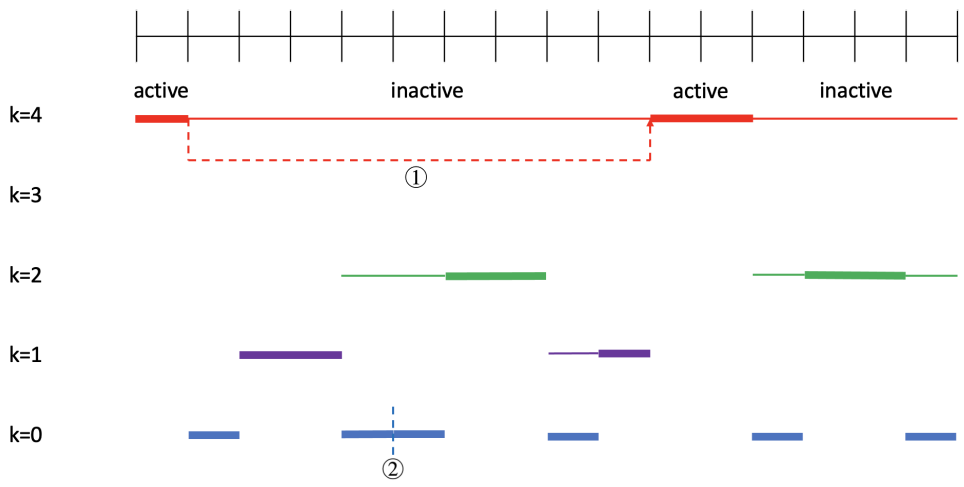


Figure 4.C.1: An illustrate example of MALG with $n = 4$.

can deal with non-stationary environments. The base algorithm is assumed to satisfy the following requirement:

Assumption 8. *ALG outputs an auxiliary quantity $e^{\beta V_1^m(s_1)} \in [0, e^{\beta H}]$ at the beginning of each round m. There exist a non-stationarity measure Δ and a non-increasing function $\rho : [M] \to \mathbb{R}$ such that running ALG satisfies the following: for all $m \in [M]$, as long as $\Delta_{[1,m]} \le \rho(m)$, without knowing $\Delta_{[1,m]}$ ALG ensures with probability at least $1 - \frac{\delta}{M}$: if $\beta > 0$, it holds that*

$$e^{\beta V_1^m(s_1)} \ge \min_{\tau \in [1,m]} e^{\beta V_1^{*,\tau}(s_1)} - \Delta_{[1,m]} \quad \text{and} \quad \frac{1}{m}\sum_{\tau=1}^{m}\left(e^{\beta V_1^\tau(s_1)} - e^{\beta \sum_{h=1}^H r_h^\tau}\right) \le \rho(m) + \Delta_{[1,m]},$$

and if $\beta < 0$, it holds that

$$\max_{\tau \in [1,m]} e^{\beta V_1^{*,\tau}(s_1)} \ge e^{\beta V_1^m(s_1)} - \Delta_{[1,m]} \quad \text{and} \quad \frac{1}{m}\sum_{\tau=1}^{m}\left(e^{\beta \sum_{h=1}^H r_h^\tau} - e^{\beta V_1^\tau(s_1)}\right) \le \rho(m) + \Delta_{[1,m]},$$

Furthermore, we assume that $\rho(m) \ge \frac{1}{\sqrt{m}}$ and $C(m) = m\rho(m)$ is a non-decreasing function.

Under Assumption 8, the multi-scale nature of MALG allows the learner's regret to also enjoy a multi-scale structure, as shown in the next lemma:

Lemma 34. *Let $\widehat{n} = \log_2 M + 1$ and $\widehat{\rho}(m) = 6\widehat{n}\log(M/\delta)\rho(m)$. MALG with input $n \le \log_2 M$ guarantees the following: for every instance alg that MALG maintains and every $m \in [alg.s, alg.e]$, as long as $\Delta_{[alg.s\,,t]} \le \rho(m')$ where $m' = m - alg.s + 1$, we have with probability at least $1 - \frac{\delta}{M}$: if $\beta > 0$, it holds that*

$$g_m \ge \min_{\tau \in [alg.s, m]} e^{\beta V_1^{*,\tau}(s_1)} - \Delta_{[alg.s\,,t]}, \quad \frac{1}{m'}\sum_{\tau=alg.s}^{m}\left(g_\tau - e^{\beta \sum_{h=1}^H r_h^\tau}\right) \le \widehat{\rho}(m') + \widehat{n}\Delta_{[alg.s,m]},$$

and if $\beta < 0$, it holds that

$$\max_{\tau\in[alg.s,m]} e^{\beta V_1^{*,\tau}(s_1)} \geq g_m - \Delta_{[alg.s\ ,t]}, \quad \frac{1}{m'}\sum_{\tau=alg.s}^{m}\left(e^{\beta\sum_{h=1}^{H} r_h^\tau} - g_\tau\right) \leq \widehat{\rho}(m') + \widehat{n}\Delta_{[alg.s,m]},$$

where g_m is the UCB-based optimistic estimator $e^{\beta V_1^m(s_1)}$ for the unique active instance alg at the episode m, and the number of instances started within $[alg.s, m]$ is upper bounded by $6\widehat{n}\log(M/\delta)\frac{C(m')}{C(1)}$.

Proof. The proof is similar to that of Lemma 3 in [124] with the standard value functions replaced by the exponential value functions and is thus omitted. □

Lemma 34 states that even if there are multiple instances interleaving in a complicated way, the regret for a specific interval is still almost the same as running ALG alone on this interval, due to the carefully chosen probability $\frac{\rho(2^n)}{\rho(2^k)}$ in Algorithm 6. Built on Lemma 34, the regret on a single block $[m_n, E_n]$, where E_n is either $m_n + 2^n - 1$ or something smaller in the case where a restart is triggered, is bounded in the following lemma:

Lemma 35. *For Algorithm 5 with ALG satisfying Assumption 8 and on every block $\mathcal{J} = [m_n, E_n]$ where $E_n \leq m_n + 2^n - 1$, it holds with high probability that:*

$$\begin{cases}\sum_{\tau\in\mathcal{J}}\left(e^{\beta V_1^{*,\tau}(s_1)} - R_\tau\right) \leq \widetilde{\mathcal{O}}\left(\sum_{i=1}^{\ell} C\left(|\mathcal{I}_i'|\right) + \sum_{m=0}^{n}\frac{\rho(2^m)}{\rho(2^n)}C\left(2^m\right)\right), & \text{if } \beta > 0,\\ \sum_{\tau\in\mathcal{J}}\left(R_\tau - e^{\beta V_1^{*,\tau}(s_1)}\right) \leq \widetilde{\mathcal{O}}\left(\sum_{i=1}^{\ell} C\left(|\mathcal{I}_i'|\right) + \sum_{m=0}^{n}\frac{\rho(2^m)}{\rho(2^n)}C\left(2^m\right)\right), & \text{if } \beta < 0,\end{cases}$$

where $\left\{\mathcal{I}_1', \ldots, \mathcal{I}_\ell'\right\}$ is any partition of $\mathcal{J}$ such that $\Delta_{\mathcal{I}_i'} \leq \rho\left(|\mathcal{I}_i'|\right)$ for all i.

Proof. The proof is similar to that of Lemma 4 in [124] with the standard value functions replaced by the exponential value functions and is thus omitted. □

Built on the dynamic regret over a block, we can further bound the dynamic regret over a single-epoch. The epoch is defined as an interval that starts at the first episode after a restart and ends at the first time when the restart is triggered.

Lemma 36. *Assume that $C(m)$ takes the form of $C(m) = c_1 m^{\frac{1}{2}}$ for some constant c_1. Then, for Algorithm 5 with ALG satisfying Assumption 8 and on every epoch $\mathcal{E}$, it holds with high probability that:*

$$\begin{cases}\sum_{\tau\in\mathcal{E}}\left(e^{\beta V_1^{*,\tau}(s_1)} - R_\tau\right) \leq \widetilde{\mathcal{O}}\left(c_1^{\frac{2}{3}}\Delta_{\mathcal{E}}^{\frac{1}{3}}|\mathcal{E}|^{\frac{2}{3}} + c_1|\mathcal{E}|^{\frac{1}{2}}\right), & \text{if } \beta > 0,\\ \sum_{\tau\in\mathcal{E}}\left(R_\tau - e^{\beta V_1^{*,\tau}(s_1)}\right) \leq \widetilde{\mathcal{O}}\left(c_1^{\frac{2}{3}}\Delta_{\mathcal{E}}^{\frac{1}{3}}|\mathcal{E}|^{\frac{2}{3}} + c_1|\mathcal{E}|^{\frac{1}{2}}\right), & \text{if } \beta < 0,\end{cases}$$

Proof. The proof is similar to that of Lemma 22 in [124] with the standard value functions replaced by the exponential value functions and is thus omitted. □

Finally, we have the following bound on the number of epoch:

Lemma 37 (Lemma 24 in [124]). *Assume that $C(m)$ takes the form of $C(m) = c_1 m^{\frac{1}{2}}$ for some constant c_1. Then, with high probability, the number of epoch is upper-bounded by $1 + 2(c_1^{-\frac{1}{3}} \Delta^{\frac{2}{3}} M^{\frac{1}{3}})$.*

Proof of Theorem 13

We first focus on the case for $\beta > 0$. Let $\mathcal{E}_1, \ldots, \mathcal{E}_N$ be epochs in $[1, M]$. If Assumption 8 holds, by Lemma 36, the dynamic regret of the exponential value functions over M episodes is upper-bounded by

$$\begin{aligned}\sum_{m=1}^{M} \left(e^{\beta V_1^{*,m}(s_1)} - R_m \right) \leq & \widetilde{\mathcal{O}} \left(\sum_{i=1}^{N} \left(c_1^{\frac{2}{3}} \Delta_{\mathcal{E}_i}^{\frac{1}{3}} |\mathcal{E}_i|^{\frac{2}{3}} + c_1 |\mathcal{E}_i|^{\frac{1}{2}} \right) \right) \\ \leq & \widetilde{\mathcal{O}} \left(c_1^{\frac{2}{3}} \Delta^{\frac{1}{3}} M^{\frac{2}{3}} + c_1 N^{\frac{1}{2}} M^{\frac{1}{2}} \right) \\ \leq & \widetilde{\mathcal{O}} \left(c_1^{\frac{2}{3}} \Delta^{\frac{1}{3}} M^{\frac{2}{3}} \right). \end{aligned} \tag{4.43}$$

where the second inequality follows from Hölder's inequality and the facts that $\sum_{i=1}^{N} \Delta_{\mathcal{E}_i} \leq \Delta$ and $\sum_{i=1}^{N} |\mathcal{E}_i| \leq M$, the last step holds by the bound on N from Lemma 37.

Now, it remains to show that the base algorithms RSVI and RSQ satisfy Assumption 8 and provide the concrete form of $\Delta(m)$, $\rho(m)$, c_1 and c_2.

- RSVI as the base algorithm: it has been shown in Lemma 28 and (4.24) in the proof of Theorem 11 that RSVI satisfies Assumption 8 with the following choices:

  $$\begin{aligned} \Delta(m) &= H \left(\left| e^{\beta H} - 1 \right| B_{\mathcal{P},m} + g_1(\beta) B_{r,m} \right), \\ \rho(m) &= \mathcal{O} \left(\left(\left| e^{\beta H} - 1 \right| + g_1(\beta) \right) \sqrt{H^2 |S|^2 |A| \iota^2 / m} \right), \\ c_1 &= \left(\left| e^{\beta H} - 1 \right| + g_1(\beta) \right) \sqrt{H^2 |S|^2 |A| \iota^2}. \end{aligned}$$

  Then, by plugging in the form of Δ and c_1 in (4.43), and using $e^{\beta H} - 1 \leq \beta H e^{\beta H}$ for $\beta > 0$, we have

  $$\sum_{m=1}^{M} \left(e^{\beta V_1^{*,m}(s_1)} - R_m \right) \leq \widetilde{\mathcal{O}} \left(\beta e^{\beta H} H^2 |\mathcal{S}|^{\frac{2}{3}} |\mathcal{A}|^{\frac{1}{3}} B^{\frac{1}{3}} M^{\frac{2}{3}} \right).$$

  Invoking the above inequality with Lemma 43 and applying Azuma's inequality to bound $\sum_{m=1}^{M} (R_m - e^{\beta V_1^{\pi^m,m}})$ yield that:

  $$\text{D-Regret(M)} \leq \widetilde{\mathcal{O}} \left(e^{\beta H} H^2 |\mathcal{S}|^{\frac{2}{3}} |\mathcal{A}|^{\frac{1}{3}} B^{\frac{1}{3}} M^{\frac{2}{3}} \right).$$

- RSQ as the base algorithm: it has also been shown in Lemma 32 and (4.40) in the proof of Theorem 12 that RSQ satisfies Assumption 8 with the following choices:

$$\Delta(m) = H\left(2g_1(\beta)B_{r,m} + \left(g_1(\beta)H + \left|e^{\beta H} - 1\right|\right)B_{\mathcal{P},m}\right)$$
$$\rho(m) = \mathcal{O}\left(\left|e^{\beta H} - 1\right|\sqrt{H|\mathcal{S}||\mathcal{A}|\iota/m}\right),$$
$$c_1 = \mathcal{O}\left(\left|e^{\beta H} - 1\right|\sqrt{H|\mathcal{S}||\mathcal{A}|\iota}\right).$$

  Then, by plugging in the form of Δ and c_1 in (4.43), and using $e^{\beta H} - 1 \leq \beta H e^{\beta H}$ for $\beta > 0$, we have

$$\sum_{m=1}^{M}\left(e^{\beta V_1^{*,m}(s_1)} - R_m\right) \leq \widetilde{\mathcal{O}}\left(\beta e^{\beta H} H^{\frac{5}{3}}|\mathcal{S}|^{\frac{1}{3}}|\mathcal{A}|^{\frac{1}{3}}B^{\frac{1}{3}}M^{\frac{2}{3}}\right).$$

  Invoking the above inequality with Lemma 43 and applying Azuma's inequality to bound $\sum_{m=1}^{M}(R_m - e^{\beta V_1^{\pi^m,m}})$ yield that:

$$\text{D-Regret(M)} \leq \widetilde{\mathcal{O}}\left(e^{\beta H} H^{\frac{5}{3}}|\mathcal{S}|^{\frac{1}{3}}|\mathcal{A}|^{\frac{1}{3}}B^{\frac{1}{3}}M^{\frac{2}{3}}\right).$$

For the case of $\beta < 0$, note that from Lemma 43, the dynamic regret can be bounded and decomposed as follows:

$$\text{D-Regret}(M) \leq \frac{e^{-\beta H}}{(-\beta)}\sum_{m\in[M]}\left[e^{\beta\cdot V_1^m(s_1^m)} - e^{\beta\cdot V_1^{*,m}(s_1^m)}\right] + \frac{e^{-\beta H}}{(-\beta)}\sum_{m\in[M]}\left[e^{\beta\cdot V_1^{\pi^m,m}(s_1^m)} - e^{\beta\cdot V_1^m(s_1^m)}\right].$$

Then, following a procedure similar to the one used for the case $\beta > 0$ and noticing that $g_1(\beta)H = -\beta H \geq 1 - e^{\beta H}$ for $\beta < 0$, we obtain the desired result.

4.D Proof of Theorem 14

Case $\beta > 0$

Consider a stochastic k-arm and M horizons bandit environment ν, where the reward for pulling arm $j \in \{1, 2, \ldots, k\}$ is given by the scaled Bernoulli random variable $Ber(p_j)$

$$X_j = \begin{cases} H, & \text{with probability } p_j, \\ 0, & \text{with probability } 1 - p_j \end{cases}$$

where $H \geq 1$ specifies the range of the reward. We let the arm i be the unique optimal arm and all the other $k-1$ arms have the same p_j, that is, $p_1 = p_2 = \cdots = p_{i-1} = p_{i+1} = \cdots = p_k = p$ and $p_i = p + \Delta$ for some constants $p > 0$ and $\Delta > 0$. Define X_j^m to be the outcome of arm j (if pulled) in round m, and Y^m to be the outcome of arm actually pulled in round m.

Lemma 38. *For the Bernoulli bandit ν described above, if $p = e^{-\beta H}$, $\Delta \le e^{-\beta H}$ and $H \ge \frac{\log 2}{\beta}$, then for every policy π, the regret with the entropic risk measure in ν satisfies*

$$\begin{aligned}\text{Regret}(M) &\coloneqq \sum_{m=1}^{M} \frac{1}{\beta} \left(\log\left[\mathbb{E}[\exp\left(\beta X_1^m\right)]\right] - \log\left[\mathbb{E}[\exp\left(\beta Y^m\right)]\right]\right) \\ &\ge \sum_{j \in [k]/\ i} \mathbb{E}\left[T_j(M)\right] \frac{\Delta(e^{\beta H} - 1)}{4\beta}\end{aligned}$$

Proof. By the definition of $\text{Regret}(M)$, we have

$$\begin{aligned}\text{Regret(M)} &= \sum_{m=1}^{M} \frac{1}{\beta} \left(\log\left[\mathbb{E}[\exp\left(\beta X_1^m\right)]\right] - \log\left[\mathbb{E}[\exp\left(\beta Y^m\right)]\right]\right) \\ &= \sum_{j \in [k]/\ i} \frac{T_i(M)}{\beta} \left(\log\left[\mathbb{E}[\exp\left(\beta X_1\right)]\right] - \log\left[\mathbb{E}[\exp\left(\beta X_i\right)]\right]\right) \qquad (4.44)\end{aligned}$$

where the last step holds because of the independence among $\{X_1^m\}_{m=1}^M$ and the independence among $\{Y^m\}_{m=1}^M$. Taking the expectation over M on both sides of (4.44), we have

$$\begin{aligned}\mathbb{E}\left[\text{Regret(M)}\right] &= \sum_{j \in [k]/\ i} \frac{\mathbb{E}\left[T_i(M)\right]}{\beta} \left(\log\left[\mathbb{E}[\exp\left(\beta X_i\right)]\right] - \log\left[\mathbb{E}[\exp\left(\beta X_j\right)]\right]\right) \\ &= \sum_{j \in [k]/\ i} \frac{\mathbb{E}\left[T_j(M)\right]}{\beta} \log\left(\frac{(p+\Delta)e^{\beta H} + (1-p-\Delta)}{pe^{\beta H} + (1-p)}\right) \\ &= \sum_{j \in [k]/\ i} \frac{\mathbb{E}\left[T_j(M)\right]}{\beta} \log\left(1 + \frac{\Delta(e^{\beta H} - 1)}{pe^{\beta H} + (1-p)}\right) \\ &= \sum_{j \in [k]/\ i} \frac{\mathbb{E}\left[T_j(M)\right]}{\beta} \log\left(1 + \frac{\Delta(e^{\beta H} - 1)}{2 - e^{-\beta H}}\right) \\ &\ge \sum_{j \in [k]/\ i} \frac{\mathbb{E}\left[T_j(M)\right]}{\beta} \log\left(1 + \frac{\Delta(e^{\beta H} - 1)}{2}\right) \\ &\ge \sum_{j \in [k]/\ i} \mathbb{E}\left[T_j(M)\right] \frac{\Delta(e^{\beta H} - 1)}{4\beta}\end{aligned}$$

where the forth equality holds since $p = e^{-\beta H}$, the first inequality follows from $e^{\beta H} \ge 2$, and the second inequality holds since $\Delta \le e^{-\beta H}$ and $\log(1+x) \ge \frac{x}{2}$ for $x \in [0,1]$. □

Lemma 39. *Let $k > 1$. For every policy π and sufficiently large M and H, there exists a k-arm bandit instance such that*

$$\mathbb{E}_{\vec{p}}\left[\text{Regret(M)}\right] > \frac{e^{\beta H/2} - 1}{\beta} \frac{\sqrt{Mk}}{64e}.$$

Proof. Fix a policy π. Let $\Delta \in [0, e^{-\beta H}]$ be some constant to be chosen later. We start with a Bernoulli bandit where the reward of each arm is a scaled Bernoulli random variable $\mathrm{Ber}(p_i)$ with $\vec{p} \coloneqq (p_1, \ldots, p_k) = (\Delta + p, p, \ldots, p)$. This environment and the policy π give rise to the probability measure $\mathbb{P}_{\vec{p}}$ on the canonical bandit model (Section 4.6 in [78]) induced by the M-round interconnection of π and ν. Expectation under $\mathbb{P}_{\vec{p}}$ will be denoted as $\mathbb{E}_{\vec{p}}$. To choose the second environment, let

$$i = \arg\min_{j>1} \mathbb{E}_{\vec{p}}\left[T_j(M)\right].$$

Since $\sum_{j=1}^{k} \mathbb{E}_{\vec{p}}\left[T_j(M)\right] = M$, it holds that

$$\mathbb{E}_{\vec{p}}\left[T_i(M)\right] \leq \frac{M}{k-1} \tag{4.45}$$

The second bandit is also a Bernoulli bandit where the reward of each arm is a scaled Bernoulli random variable $\mathrm{Ber}(p'_i)$ with $\vec{p}' \coloneqq (p'_1, \ldots, p'_k) = (\Delta + p, p, \ldots, 2\Delta + p, p, \ldots, p)$, where specifically $p'_i = 2\Delta + p$. Therefore, $p_j = p'_j$ except at index i and the optimal arm in $\nu_{\vec{p}}$ is the first arm, while in $\nu_{\vec{p}'}$ arm i is optimal. Then, Lemma 38 and a simple calculation lead to

$$\begin{aligned}
\mathbb{E}_{\vec{p}}\left[\mathrm{Regret(M)}\right] &\geq \mathbb{P}_{\vec{p}}(T_1(M) \leq \frac{M}{2}) \frac{M\Delta(e^{\beta H}-1)}{8\beta}, \\
\mathbb{E}_{\vec{p}'}\left[\mathrm{Regret(M)}\right] &> \mathbb{P}_{\vec{p}'}(T_1(M) > \frac{M}{2}) \frac{M\Delta(e^{\beta H}-1)}{8\beta}.
\end{aligned}$$

Then, applying the Bretagnolle-Huber inequality in Lemma 46 leads to

$$\begin{aligned}
&\mathbb{E}_{\vec{p}}\left[\mathrm{Regret(M)}\right] + \mathbb{E}_{\vec{p}'}\left[\mathrm{Regret(M)}\right] \\
>&\frac{M\Delta(e^{\beta H}-1)}{8\beta}\left(\mathbb{P}_{\vec{p}}(T_1(M) \leq \frac{M}{2}) + \mathbb{P}_{\vec{p}'}(T_1(M) > \frac{M}{2})\right) \\
\geq&\frac{M\Delta(e^{\beta H}-1)}{8\beta} \exp\left(-D_{\mathrm{KL}}(\mathbb{P}_{\vec{p}} \mid \mathbb{P}_{\vec{p}'})\right)
\end{aligned}$$

It remains to upper-bound $D_{\mathrm{KL}}(\mathbb{P}_{\vec{p}} \mid \mathbb{P}_{\vec{p}'})$. For this, we use Lemma 48:

$$\begin{aligned}
D_{\mathrm{KL}}\left(\mathbb{P}_{\nu} \mid \mathbb{P}_{\nu'}\right) =& \mathbb{E}_{\mathbb{P}_{\vec{p}}}\left[T_i(M)\right] \mathrm{D}_{\mathrm{KL}}\left(\mathrm{Ber}(p_i) \mid \mathrm{Ber}(p'_i)\right) \\
=& \mathbb{E}_{\mathbb{P}_{\vec{p}}}\left[T_i(M)\right] \mathrm{D}_{\mathrm{KL}}\left(p \mid 2\Delta + p\right)) \\
\leq& \mathbb{E}_{\mathbb{P}_{\vec{p}}}\left[T_i(M)\right] \cdot \frac{4\Delta^2}{(2\Delta+p)(1-2\Delta-p)} \\
\leq& \frac{M}{k-1} \cdot \frac{4\Delta^2}{(2\Delta+p)(1-2\Delta-p)} \\
\leq& \frac{16M\Delta^2}{kp} \\
\leq& \frac{16e^{\beta H}M\Delta^2}{k}
\end{aligned} \tag{4.46}$$

where the first inequality follows from Lemma 47, the second inequality holds by (4.45), the third step follows from $1-2\Delta-p\geq\frac{1}{2}$ and $k\geq 3$, and the last step holds by $p=e^{-\beta H}$.

Substituting this into the previous expression, we find that

$$\begin{aligned}\mathbb{E}_{\vec{p}}\left[\text{Regret(M)}\right]+\mathbb{E}_{\vec{p}'}\left[\text{Regret(M)}\right]&>\frac{M\Delta(e^{\beta H}-1)}{8\beta}\exp\left(-\frac{16e^{\beta H}M\Delta^2}{k}\right)\\&>\frac{e^{\beta H/2}-1}{\beta}\frac{\sqrt{Mk}}{32e}\end{aligned}$$

where the second inequality holds by choosing $\Delta=\sqrt{k/(16Me^{\beta H})}\leq e^{-\beta H}$ with M sufficiently large. This result is completed by using $2\max(a,b)\geq a+b$. □

Lemma 40. *For every policy π and sufficiently large M and H, there exists a MDP instance with horizon H, $S\geq 3$ states and A actions such that*

$$\mathbb{E}\left[\text{Regret(M)}\right]>\frac{e^{\beta H/2}-1}{\beta}\frac{\sqrt{MSA}}{64e}.$$

Proof. Note that the M-round k-arm bandit model described in Lemma 39 is a special case of an M-episode $(H+2)$-horizon MDP with S states and $\frac{S-1}{2}$ actions where $S\geq 3$ is odd. Let s_1 be the initial state, and all other states be absorbing regardless of actions taken. At the initial state s_1, we may choose to take action a_1, $a_2,\ldots,a_{\frac{S-1}{2}}$. If a_j is taken at state s_1, then we transition to state $s_{1+2(j-1)+1}$ with probability p_j and to state $s_{1+2(j-1)+2}$ with probability $1-p_j$. The reward function satisfies $r_h(s_{1+2(j-1)+1},a)=1$, $r_h(s_{1+2(j-1)+2},a)=0$ and $r_h(s_1,a)=0$ for all $h\in[H+2]$, $a\in\mathcal{A}$ and $j=1,\ldots,\frac{S-1}{2}$. □

Based on Lemma 40, let us now incorporate the non-stationarity of the MDP and derive a lower bound for the dynamic regret D-Regret(M). We will construct the non-stationary environment as a switching-MDP. For each segment of length M_0, the environment is held constant, and the regret lower bound for each segment is $\mathcal{O}\left(\frac{e^{\beta H/2}-1}{\beta}\sqrt{SAM_0}\right)$. At the beginning of each new segment, we uniformly sample a new action at random at the state s_1 from the action space $\mathcal{A}$ to be the optimal action at the state s_1 for the new segment. In this case, the learning algorithm cannot use the information it learned during its previous interactions with the environment, even if it knows the switching structure of the environment. Therefore, the algorithm needs to learn a new (static) MDP in each segment, which leads to a dynamic regret lower bound of

$$\mathcal{O}\left(\frac{e^{\beta H/2}-1}{\beta}L\sqrt{SAM_0}\right)=\mathcal{O}\left(\frac{e^{\beta H/2}-1}{\beta}\sqrt{SAML}\right),$$

where L is the number of segments. Every time that the optimal action at the state s_1 varies, it will cause a variation of magnitude $2\Delta=\sqrt{SA/(4M_0e^{\beta H})}$ in the transition kernel. The constraint of the overall variation budget requires that

$$2\Delta L=\sqrt{\frac{SA}{4M_0e^{\beta H}}}L=\sqrt{\frac{SAL^3}{4Me^{\beta H}}}\leq B,$$

which in turn requires $L \le 4^{\frac{1}{3}} B^{\frac{2}{3}} M^{\frac{1}{3}} e^{\frac{\beta H}{3}} S^{-\frac{1}{3}} A^{-\frac{1}{3}}$. Finally, by assigning the largest possible value to L subject to the variation budget, we obtain a dynamic regret lower bound of

$$\mathcal{O}\left(\frac{e^{\frac{2\beta H}{3}} - 1}{\beta} S^{\frac{1}{3}} A^{\frac{1}{3}} B^{\frac{1}{3}} M^{\frac{2}{3}}\right).$$

This completes the proof of Theorem 14 for the case $\beta > 0$.

Case $\beta < 0$

The proof of the base $\beta < 0$ is similar to that of the case $\beta > 0$. For $\beta < 0$, consider a stochastic k-arm and M horizons bandit environment ν, where the reward for pulling arm $j \in \{1, 2, \ldots, k\}$ is given by the scaled Bernoulli random variable $Ber(1 - p_j)$

$$X_j = \begin{cases} 0, & \text{with probability } p_j, \\ H, & \text{with probability } 1 - p_j \end{cases}$$

where $H \ge 1$ specifies the range of the reward. We let the arm i be the unique optimal arm and all the other $k - 1$ arms have the same p_j, that is, $p_1 = p_2 = \cdots = p_{i-1} = p_{i+1} = \cdots = p_k = p$ and $p_i = p + \Delta$ for some constants $p > 0$ and $\Delta < 0$. Define X_j^m to be the outcome of arm j (if pulled) in round m, and Y^m to be the outcome of arm actually pulled in round m.

Lemma 41. *For the Bernoulli bandit ν described above, if $p = e^{\beta H}$ and $\Delta \ge -e^{\beta H}$, then for every policy π, the regret with the entropic risk measure in ν satisfies*

$$\begin{aligned} \text{Regret}(M) &\coloneqq \sum_{m=1}^{M} \frac{1}{\beta} \left(\log\left[\mathbb{E}[\exp(\beta X_1^m)]\right] - \log\left[\mathbb{E}[\exp(\beta Y^m)]\right]\right) \\ &\ge \sum_{j \in [k]/\ i} \mathbb{E}\left[T_j(M)\right] \frac{\Delta(e^{-\beta H} - 1)}{2\beta} \end{aligned}$$

Proof. Taking the expectation over M on both sides of (4.44), we have

$$\begin{aligned} \mathbb{E}\left[\text{Regret(M)}\right] &= \sum_{j \in [k]/\ i} \frac{\mathbb{E}\left[T_i(M)\right]}{\beta} \left(\log\left[\mathbb{E}[\exp(\beta X_i)]\right] - \log\left[\mathbb{E}[\exp(\beta X_j)]\right]\right) \\ &= \sum_{j \in [k]/\ i} \frac{\mathbb{E}\left[T_j(M)\right]}{\beta} \log\left(\frac{(1 - p - \Delta)e^{\beta H} + (p + \Delta)}{(1 - p)e^{\beta H} + p}\right) \\ &= \sum_{j \in [k]/\ i} \frac{\mathbb{E}\left[T_j(M)\right]}{\beta} \log\left(1 + \frac{\Delta(1 - e^{\beta H})}{(1 - p)e^{\beta H} + p}\right) \\ &\ge \sum_{j \in [k]/\ i} \frac{\mathbb{E}\left[T_j(M)\right]}{\beta} \log\left(1 + \frac{\Delta(1 - e^{\beta H})}{2e^{\beta H}}\right) \\ &\ge \sum_{j \in [k]/\ i} \mathbb{E}\left[T_j(M)\right] \frac{\Delta(e^{-\beta H} - 1)}{2\beta} \end{aligned}$$

where the first inequality holds since $p = e^{\beta H}$, the second inequality holds since $\Delta \le e^{-\beta H}$ and $\log(1+x) \le x$ for $x > -1$. □

Lemma 42. *Let $k > 1$. For every policy π and sufficiently large M and H, there exists a k-arm bandit instance such that*

$$\mathbb{E}_{\vec{p}}\left[\text{Regret(M)}\right] > \frac{e^{-\beta H/2}-1}{-\beta}\frac{\sqrt{Mk}}{64e}.$$

Proof. The proof is similar to that of Lemma 39 by replacing Lemma 38 with Lemma 41, replacing (4.46) by

$$D_{\text{KL}}\left(\mathbb{P}_\nu \mid \mathbb{P}_{\nu'}\right) = \mathbb{E}_{\mathbb{P}_{\vec{p}}}\left[T_i(M)\right] \text{D}_{\text{KL}}\left(\text{Ber}(1-p_i) \mid \text{Ber}(1-p_i')\right)$$

and by choosing $\Delta = -\sqrt{k/(16Me^{-\beta H})} \ge -e^{\beta H}$. □

The rest of the proof is similar to that for the case $\beta > 0$ and is thus omitted.

4.E Auxiliary lemmas

Lemma 43. *For $\beta > 0$, the dynamic regret is bounded by*

$$\text{D-Regret}(M) \le \sum_{m\in[M]} \frac{1}{\beta}\left[e^{\beta\cdot V_1^*(s_1^m)} - e^{\beta\cdot V_1^m(s_1^m)}\right] + \sum_{m\in[M]} \frac{1}{\beta}\left[e^{\beta\cdot V_1^m(s_1^m)} - e^{\beta\cdot V_1^{\pi^m,m}(s_1^m)}\right],$$

and for $\beta < 0$, the dynamic regret is bounded by

$$\text{D-Regret}(M) \le \sum_{m\in[M]} \frac{e^{-\beta H}}{(-\beta)}\left[e^{\beta\cdot V_1^m(s_1^m)} - e^{\beta\cdot V_1^{*,m}(s_1^m)}\right] + \sum_{m\in[M]} \frac{e^{-\beta H}}{(-\beta)}\left[e^{\beta\cdot V_1^{\pi^m,m}(s_1^m)} - e^{\beta\cdot V_1^m(s_1^m)}\right].$$

Proof. For $\beta > 0$, we have

$$\begin{aligned}
&\text{D-Regret}(M)\\
&= \sum_{m\in[M]} \left(V_1^{*,m} - V_1^{\pi^m,m}\right)(s_1^m)\\
&= \sum_{m\in[M]} \left(V_1^{*,m} - V_1^m\right)(s_1^m) + \sum_{m\in[M]} \left(V_1^m - V_1^{\pi^m}\right)(s_1^m)\\
&= \sum_{m\in[M]} \left[\frac{1}{\beta}\log\left\{e^{\beta\cdot V_1^{*,m}(s_1^m)}\right\} - \frac{1}{\beta}\log\left\{e^{\beta\cdot V_1^m(s_1^m)}\right\}\right] + \sum_{m\in[M]} \left[\frac{1}{\beta}\log\left\{e^{\beta\cdot V_1^m(s_1^m)}\right\} - \frac{1}{\beta}\log\left\{e^{\beta\cdot V_1^{\pi^m}(s_1^m)}\right\}\right]\\
&\le \sum_{m\in[M]} \frac{1}{\beta}\left[e^{\beta\cdot V_1^{*,m}(s_1^m)} - e^{\beta\cdot V_1^m(s_1^m)}\right] + \sum_{m\in[M]} \frac{1}{\beta}\left[e^{\beta\cdot V_1^m(s_1^m)} - e^{\beta\cdot V_1^{\pi^m,m}(s_1^m)}\right]
\end{aligned}$$

where the last step holds by the 1-Lipschitzness of the function $f(x) = \log x$ for $x \ge 1$.

For $\beta < 0$, we similarly have

$$
\begin{aligned}
&\text{D-Regret}(M) \\
&= \sum_{m\in[M]} \left(V_1^{*,m} - V_1^{\pi^m,m}\right)(s_1^m) \\
&= \sum_{m\in[M]} \left(V_1^{*,m} - V_1^{m}\right)(s_1^m) + \sum_{m\in[M]} \left(V_1^{m} - V_1^{\pi^m}\right)(s_1^m) \\
&= \sum_{m\in[M]} \left[\frac{1}{-\beta}\log\left\{e^{\beta\cdot V_1^m(s_1^m)}\right\} - \frac{1}{-\beta}\log\left\{e^{\beta\cdot V_1^{*,m}(s_1^m)}\right\}\right] \\
&\quad + \sum_{m\in[M]} \left[\frac{1}{-\beta}\log\left\{e^{\beta\cdot V_1^{\pi^m,m}(s_1^m)}\right\} - \frac{1}{-\beta}\log\left\{e^{\beta\cdot V_1^{m}(s_1^m)}\right\}\right] \\
&\le \sum_{m\in[M]} \frac{e^{-\beta H}}{(-\beta)}\left[e^{\beta\cdot V_1^m(s_1^m)} - e^{\beta\cdot V_1^{*,m}(s_1^m)}\right] + \sum_{m\in[M]} \frac{e^{-\beta H}}{(-\beta)}\left[e^{\beta\cdot V_1^{\pi^m,m}(s_1^m)} - e^{\beta\cdot V_1^{m}(s_1^m)}\right]
\end{aligned}
$$

where the last step holds by the $(e^{-\beta H})$-Lipschitzness of the function $f(x) = \log x$ for $x \ge e^{\beta H}$. □

Lemma 44 (Theorem 1 in [1]). *Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration and $\{\eta_t\}_{t=1}^{\infty}$ be a $\mathbb{R}$-valued stochastic process such that η_t is $\mathcal{F}_t$-measurable for every $t \ge 0$. Assume that for every $t \ge 0$, conditioning on $\mathcal{F}_t, \eta_t$ is a zero-mean and σ-subGaussian random variable with the variance proxy $\sigma^2 > 0$, i.e., $\mathbb{E}\left[e^{\lambda\eta_t} \mid \mathcal{F}_t\right] \le e^{\lambda^2\sigma^2/2}$ for every $\lambda \in \mathbb{R}$. Let $\{X_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process such that X_t is $\mathcal{F}_t$-measurable for every $t \ge 0$. Let $Y \in \mathbb{R}^{d\times d}$ be a deterministic and positive-definite matrix. For every $t \ge 0$, we define*

$$\bar{Y}_t := Y + \sum_{\tau=1}^{t} X_\tau X_\tau^\top \text{ and } S_t = \sum_{\tau=1}^{t} \eta_\tau X_\tau.$$

Then, for every fixed $\delta \in (0,1)$, it holds with probability at least $1-\delta$ that

$$\|S_t\|^2_{(\bar{Y}_t)^{-1}} \le 2\sigma^2 \log\left(\frac{\det\left(\bar{Y}_t\right)^{1/2}\det(Y)^{-1/2}}{\delta}\right)$$

for every $t \ge 0$.

Lemma 45 (Fact 1 in [45]). *The following properties hold for α_t^i defined in (4.26):*

1. *$\frac{1}{\sqrt{t}} \le \sum_{i\in[t]} \frac{\alpha_t^i}{\sqrt{i}} \le \frac{2}{\sqrt{t}}$ for every integer $t \ge 1$.*
2. *$\max_{i\in[t]} \alpha_t^i \le \frac{2H}{t}$ and $\sum_{i\in[t]} \left(\alpha_t^i\right)^2 \le \frac{2H}{t}$ for every integer $t \ge 1$.*
3. *$\sum_{t=i}^{\infty} \alpha_t^i = 1 + \frac{1}{H}$ for every integer $i \ge 1$.*
4. *$\sum_{i\in[t]} \alpha_t^i = 1$ and $\alpha_t^0 = 0$ for every integer $t \ge 1$, and $\sum_{i\in[t]} \alpha_t^i = 0$ and $\alpha_t^0 = 1$ for $t = 0$.*

Lemma 46 (Lemma 14.2 in [78]). *Let P, Q be probability measures on the same measurable space $(\Omega, \mathcal{F})$, and let $A \in \mathcal{F}$ be an arbitrary event. Then,*

$$P(A) + Q(A^c) \geq \frac{1}{2} \exp\left(-D_{KL}(P \mid Q)\right),$$

where D_{KL} denotes the KL divergence and $A^c = \Omega / A$ is the complement of A.

Lemma 47 (Lemma 14 in [48]). *Let $p, p' \in (0,1)$ be such that $p > p'$. We have*

$$D_{\mathrm{KL}}\left(\mathrm{Ber}\left(p'\right) \| \mathrm{Ber}(p)\right) \leq \frac{\left(p - p'\right)^2}{p(1-p)}.$$

Lemma 48 (Divergence decomposition, Lemma 15.1 in [78]). *Let $\nu = (P_1, \ldots, P_k)$ be the reward distributions associated with one k-armed bandit, and let $\nu' = \left(P'_1, \ldots, P'_k\right)$ be the reward distributions associated with another k-armed bandit. Fix some policy π and let $\mathbb{P}_\nu = \mathbb{P}_{\nu\pi}$ and $\mathbb{P}_{\nu'} = \mathbb{P}_{\nu'\pi}$ be the probability measures on the canonical bandit model (Section 4.6 in [78]) induced by the M-round interconnection of π and ν (respectively, π and ν'). Then,*

$$\mathrm{D}_{\mathrm{KL}}\left(\mathbb{P}_\nu, \mathbb{P}_{\nu'}\right) = \sum_{i=1}^{k} \mathbb{E}_\nu\left[T_i(M)\right] \mathrm{D}_{\mathrm{KL}}\left(P_i, P'_i\right)$$

Chapter 5

Conclusions

This thesis aims to develop a set of computational tools with a rigorous analysis for the design of safe online decision-making algorithms in the presence of non-stationarity. In each of the three chapters, we take into account the underlying non-stationarity and the safety requirements of real-world problems. In what follows, we briefly summarize our contributions and future directions.

In Chapter 2, we study the landscape of time-varying nonconvex optimization problems. The objective is to understand when simple local search algorithms can find (and track) time-varying global solutions of the problem over time. We introduce a time-varying projected gradient flow system with controllable inertia as a continuous-time limit of the optimality conditions for discretized sequential optimization problems with proximal regularization and online updating scheme. Via a change of variables, the time-varying projected gradient flow system is regarded as a composition of a time-varying projected gradient term, a time-varying constraint-driven term and an inertia term due to the time variation of the local minimum trajectory. We show that the time-varying perturbation term due to the inertia encourages the exploration of the state space and reshapes the landscape by potentially making it one-point strongly convex over a large region during some time interval. We introduce the notions of jumping and escaping, and use them to develop sufficient conditions under which the time-varying solution escapes from a poor local trajectory to a better (or global) minimum trajectory over a finite time interval. We illustrate in a benchmark example with many shallow minimum trajectories that the natural time variation of the problem enables escaping spurious local minima over time. Avenues for future work include the characterization of the class of problems in which all spurious local minimum trajectories are shallow compared with the global minimum trajectory.

In Chapter 3, we formulate a general non-stationary safe RL problem as a non-stationary episodic CMDP. To solve this problem, we identify two alternative conditions on the time-varying constraints under which we can guarantee the safety in the long run. We also develop a new algorithm named PROPD-PPO, which consists of three main mechanisms: periodic-restart-based policy improvement, dual update with dual regularization, and periodic-restart-based optimistic policy evaluation. We establish the dynamic regret bound and constraint

violation bounds for the proposed algorithm in both the linear kernel CMDP function approximation setting and the tabular CMDP setting under two alternative assumptions. This paper provides the first provably efficient algorithm for non-stationary CMDPs with safe exploration. An interesting future direction is to relax the assumption on the prior knowledge of the variation budgets and generalize the non-stationarity detection mechanism to our CMDP setting.

In Chapter 4, we provide strong theoretical analyses for the non-stationary risk-sensitive RL problem, which is motivated by various risk-sensitive applications. We propose two restart based algorithms that require the knowledge of the variation budget, as well as a black-box approach to turn a certain risk-sensitive RL algorithm in a (near-)stationary environment into another algorithm in a non-stationary environment without requiring the knowledge of the variation budge. The dynamic regret bounds of these algorithms are obtained and a lower bound is established to verify the near-optimality of the proposed upper bounds. Our results also reveal the condition under which the risk control and the handling of the non-stationarity can be separately designed in the algorithm. One important future direction lies in extending our results to other notions of risk, such as the general coherent risk measures. Furthermore, it is useful to study how to adjust the risk sensitivity parameter adaptively in a non-stationary environment.

Bibliography

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. "Improved algorithms for linear stochastic bandits". In: *Advances in neural information processing systems* 24 (2011), pp. 2312–2320.

[2] David H. Ackley. *A Connectionist Machine for Genetic Hillclimbing*. Norwell, MA, USA: Kluwer Academic Publishers, 1987. ISBN: 0-89838-236-X.

[3] Dirk Aeyels and Joan Peuteman. "On exponential stability of nonlinear time-varying differential equations". In: *Automatica* 35.6 (1999), pp. 1091–1100.

[4] Alekh Agarwal et al. "On the theory of policy gradient methods: Optimality, approximation, and distribution shift". In: *arXiv preprint arXiv:1908.00261* (2019).

[5] Eitan Altman. *Constrained Markov decision processes*. Vol. 7. CRC Press, 1999.

[6] Dario Amodei et al. "Concrete problems in AI safety". In: *arXiv preprint arXiv:1606.06565* (2016).

[7] Peter Auer, Pratik Gajane, and Ronald Ortner. "Adaptively tracking the best bandit arm with an unknown number of distribution changes". In: *Conference on Learning Theory*. PMLR. 2019, pp. 138–158.

[8] Alex Ayoub et al. "Model-based reinforcement learning with value-targeted regression". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 463–474.

[9] Aurele Balavoine, Christopher J Rozell, and Justin Romberg. "Discrete and Continuous-time Soft-Thresholding with Dynamic Inputs". In: *arXiv preprint arXiv:1405.1361* (2014).

[10] Nicole Bäuerle and Ulrich Rieder. "More risk-sensitive Markov decision processes". In: *Mathematics of Operations Research* 39.1 (2014), pp. 105–120.

[11] Amir Beck. *First-order methods in optimization*. SIAM, 2017.

[12] Andrey Bernstein, Emiliano Dall'Anese, and Andrea Simonetto. "Online primal-dual methods with measurement feedback for time-varying convex optimization". In: *IEEE Transactions on Signal Processing* 67.8 (2019), pp. 1978–1991.

[13] Dimitri P Bertsekas. *Nonlinear programming*. Athena Scientific, 2016.

[14] Omar Besbes, Yonatan Gur, and Assaf Zeevi. "Non-stationary stochastic optimization". In: *Operations Research* 63.5 (2015), pp. 1227–1244.

[15] Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. "Global optimality of local search for low rank matrix recovery". In: *Advances in Neural Information Processing Systems*. 2016, pp. 3873–3881.

[16] Thomas Binder et al. "Introduction to model based optimization of chemical processes on moving horizons". In: *Online optimization of large scale systems*. Springer, 2001, pp. 295–339.

[17] Vivek S Borkar. "A sensitivity formula for risk-sensitive cost and the actor–critic algorithm". In: *Systems & Control Letters* 44.5 (2001), pp. 339–346.

[18] Vivek S Borkar. "Q-learning for risk-sensitive control". In: *Mathematics of operations research* 27.2 (2002), pp. 294–311.

[19] Vivek S Borkar and Sean P Meyn. "Risk-sensitive optimal control for Markov decision processes with monotone cost". In: *Mathematics of Operations Research* 27.1 (2002), pp. 192–209.

[20] Steven J Bradtke and Andrew G Barto. "Linear least-squares algorithms for temporal difference learning". In: *Machine learning* 22.1 (1996), pp. 33–57.

[21] Qi Cai et al. "Provably efficient exploration in policy optimization". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 1283–1294.

[22] Emmanuel J Candes and Yaniv Plan. "Matrix completion with noise". In: *Proceedings of the IEEE* 98.6 (2010), pp. 925–936.

[23] Emmanuel J Candès and Benjamin Recht. "Exact matrix completion via convex optimization". In: *Foundations of Computational Mathematics* 9.6 (2009), p. 717.

[24] Xuanyu Cao and KJ Ray Liu. "Online convex optimization with time-varying constraints and bandit feedback". In: *IEEE Transactions on automatic control* 64.7 (2018), pp. 2665–2680.

[25] Rolando Cavazos-Cadena and Emmanuel Fernández-Gaucherand. "The vanishing discount approach in Markov chains with risk-sensitive criteria". In: *IEEE Transactions on Automatic Control* 45.10 (2000), pp. 1800–1816.

[26] Yash Chandak et al. "Optimizing for the future in non-stationary MDPs". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 1414–1425.

[27] Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. "Reinforcement learning for non-stationary Markov decision processes: The blessing of (more) optimism". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 1843–1854.

[28] Yinlam Chow and Mohammad Ghavamzadeh. "Algorithms for CVaR optimization in MDPs". In: *Advances in neural information processing systems* 27 (2014).

[29] Stefano P Coraluppi and Steven I Marcus. "Risk-sensitive and minimax control of discrete-time, finite-state Markov decision processes". In: *Automatica* 35.2 (1999), pp. 301–309.

[30] Erick Delage and Shie Mannor. "Percentile optimization for Markov decision processes with parameter uncertainty". In: *Operations research* 58.1 (2010), pp. 203–213.

[31] Dotan Di Castro, Aviv Tamar, and Shie Mannor. "Policy gradients with variance related risk criteria". In: *arXiv preprint arXiv:1206.6404* (2012).

[32] Giovanni B Di Masi and Lukasz Stettner. "Risk-sensitive control of discrete-time Markov processes with infinite horizon". In: *SIAM Journal on Control and Optimization* 38.1 (1999), pp. 61–78.

[33] Dongsheng Ding et al. "Natural Policy Gradient Primal-Dual Method for Constrained Markov Decision Processes." In: *NeurIPS*. 2020.

[34] Dongsheng Ding et al. "Provably efficient safe exploration via primal-dual policy optimization". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 3304–3312.

[35] Yuhao Ding and Javad Lavaei. "Provably efficient primal-dual reinforcement learning for CMDPs with non-stationary objectives and constraints". In: *AAAI Conference on Artificial Intelligence* (2023).

[36] Yuhao Ding, Javad Lavaei, and Murat Arcak. "Time-variation in online nonconvex optimization enables escaping from spurious local minima". In: (2020). URL: `https://lavaei.ieor.berkeley.edu/Online_opt_2020_2.pdf`.

[37] Yuhao Ding, Javad Lavaei, and Murat Arcak. "Time-variation in online nonconvex optimization enables escaping from spurious local minima". In: *IEEE Transactions on Automatic Control* (2021).

[38] Omar Darwiche Domingues et al. "A kernel-based approach to non-stationary reinforcement learning in metric spaces". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 3538–3546.

[39] David L Donoho. "For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution". In: *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* 59.6 (2006), pp. 797–829.

[40] Gabriel Dulac-Arnold, Daniel Mankowitz, and Todd Hester. "Challenges of real-world reinforcement learning". In: *arXiv preprint arXiv:1904.12901* (2019).

[41] Yonathan Efroni, Shie Mannor, and Matteo Pirotta. "Exploration-exploitation in constrained MDPs". In: *arXiv preprint arXiv:2003.02189* (2020).

[42] S Fattahi et al. "Absence of spurious local trajectories in time-varying optimization". In: *arXiv preprint arXiv:1905.09937* (2019).

[43] Salar Fattahi and Somayeh Sojoudi. "Exact guarantees on the absence of spurious local minima for non-negative robust principal component analysis". In: *arXiv preprint arXiv:1812.11466* (2018).

[44] Mahyar Fazlyab et al. "Self-triggered time-varying convex optimization". In: *2016 IEEE 55th Conference on Decision and Control (CDC)*. IEEE. 2016, pp. 3090–3097.

[45] Yingjie Fei, Zhuoran Yang, and Zhaoran Wang. "Risk-sensitive reinforcement learning with function approximation: A debiasing approach". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 3198–3207.

[46] Yingjie Fei et al. "Dynamic regret of policy optimization in non-stationary environments". In: *arXiv preprint arXiv:2007.00148* (2020).

[47] Yingjie Fei et al. "Exponential Bellman Equation and Improved Regret Bounds for Risk-Sensitive Reinforcement Learning". In: *Advances in Neural Information Processing Systems* 34 (2021).

[48] Yingjie Fei et al. "Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret". In: *arXiv preprint arXiv:2006.13827* (2020).

[49] Emmanuel Fernández-Gaucherand and Steven I Marcus. "Risk-sensitive optimal control of hidden Markov models: Structural results". In: *IEEE Transactions on Automatic Control* 42.10 (1997), pp. 1418–1422.

[50] Wendell H Fleming and William M McEneaney. "Risk sensitive optimal control and differential games". In: *Stochastic theory and adaptive control*. Springer, 1992, pp. 185–197.

[51] Wendell H Fleming and William M McEneaney. "Risk-sensitive control on an infinite time horizon". In: *SIAM Journal on Control and Optimization* 33.6 (1995), pp. 1881–1915.

[52] Javier Garcıa and Fernando Fernández. "A comprehensive survey on safe reinforcement learning". In: *Journal of Machine Learning Research* 16.1 (2015), pp. 1437–1480.

[53] Rong Ge, Jason D Lee, and Tengyu Ma. "Matrix completion has no spurious local minimum". In: *Advances in Neural Information Processing Systems*. 2016, pp. 2973–2981.

[54] Rong Ge et al. "Escaping from saddle points—online stochastic gradient for tensor decomposition". In: *Conference on Learning Theory*. 2015, pp. 797–842.

[55] Shangding Gu et al. "A Review of Safe Reinforcement Learning: Methods, Theory and Applications". In: *arXiv preprint arXiv:2205.10330* (2022).

[56] Shangding Gu et al. "Multi-agent constrained policy optimisation". In: *arXiv preprint arXiv:2110.02793* (2021).

[57] Jürgen Guddat, F Guerra Vazquez, and Hubertus Th Jongen. *Parametric optimization: singularities, pathfollowing and jumps*. Springer, 1990.

[58] Suriya Gunasekar et al. "Implicit regularization in matrix factorization". In: *Advances in Neural Information Processing Systems*. 2017, pp. 6151–6159.

[59] Jack K Hale. *Ordinary differential equations*. Wiley-Inter-science, 1980.

[60] Eric Hall and Rebecca Willett. "Dynamical models and tracking regret in online convex programming". In: *International Conference on Machine Learning*. PMLR. 2013, pp. 579–587.

[61] Eric C Hall and Rebecca M Willett. "Online convex optimization in dynamic environments". In: *IEEE Journal of Selected Topics in Signal Processing* 9.4 (2015), pp. 647–662.

[62] Adrian Hauswirth et al. "Time-varying projected dynamical systems with applications to feedback optimization of power systems". In: *2018 IEEE Conference on Decision and Control (CDC)*. IEEE. 2018, pp. 3258–3263.

[63] Elad Hazan et al. "Introduction to online convex optimization". In: *Foundations and Trends® in Optimization* 2.3-4 (2016), pp. 157–325.

[64] Daniel Hernández-Hernández and Steven I Marcus. "Risk sensitive control of Markov processes in countable state space". In: *Systems & control letters* 29.3 (1996), pp. 147–155.

[65] Ronald A Howard and James E Matheson. "Risk-sensitive Markov decision processes". In: *Management science* 18.7 (1972), pp. 356–369.

[66] Arieh Iserles. *A first course in the numerical analysis of differential equations*. 44. Cambridge University Press, 2009.

[67] Thomas Jaksch, Ronald Ortner, and Peter Auer. "Near-optimal Regret Bounds for Reinforcement Learning." In: *Journal of Machine Learning Research* 11.4 (2010).

[68] Chi Jin et al. "How to escape saddle points efficiently". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 1724–1732.

[69] Chi Jin et al. "Is Q-learning provably efficient?" In: *Advances in neural information processing systems* 31 (2018).

[70] Chi Jin et al. "Provably efficient reinforcement learning with linear function approximation". In: *Conference on Learning Theory*. PMLR. 2020, pp. 2137–2143.

[71] Cedric Josz et al. "A theory on the absence of spurious solutions for nonconvex and nonsmooth optimization". In: *Advances in Neural Information Processing Systems*. 2018, pp. 2441–2449.

[72] JV Kadam et al. "Towards integrated dynamic real-time optimization and control of industrial processes". In: *Proceedings foundations of computer-aided process operations (FOCAPO2003)* (2003), pp. 593–596.

[73] Sham M Kakade. "A natural policy gradient". In: *Advances in neural information processing systems* 14 (2001).

[74] Hassan K Khalil. *Nonlinear systems*. Upper Saddle River, 2002.

[75] Robert Kleinberg, Yuanzhi Li, and Yang Yuan. "An alternative view: When does SGD escape local minima?" In: *arXiv preprint arXiv:1802.06175* (2018).

[76] Walid Krichene, Alexandre Bayen, and Peter L Bartlett. "Accelerated mirror descent in continuous and discrete time". In: *Advances in Neural Information Processing Systems*. 2015, pp. 2845–2853.

[77] Prashanth La and Mohammad Ghavamzadeh. "Actor-critic algorithms for risk-sensitive MDPs". In: *Advances in neural information processing systems* 26 (2013).

[78] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[79] Javad Lavaei and Steven H Low. "Zero duality gap in optimal power flow problem". In: *IEEE Transactions on Power Systems* 27.1 (2012), pp. 92–107.

[80] Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. "Finite-sample analysis of LSTD". In: *ICML-27th International Conference on Machine Learning*. 2010, pp. 615–622.

[81] Jason D Lee et al. "First-order methods almost always avoid saddle points". In: *arXiv preprint arXiv:1710.07406* (2017).

[82] Hao Li et al. "Visualizing the loss landscape of neural nets". In: *Advances in Neural Information Processing Systems*. 2018, pp. 6389–6399.

[83] Tao Liu et al. "Learning Policies with Zero or Bounded Constraint Violation for Constrained MDPs". In: *arXiv preprint arXiv:2106.02684* (2021).

[84] David G Luenberger. "The gradient projection method along geodesics". In: *Management Science* 18.11 (1972), pp. 620–631.

[85] Chiara Dalla Man et al. "The UVA/PADOVA type 1 diabetes simulator: new features". In: *Journal of diabetes science and technology* 8.1 (2014), pp. 26–34.

[86] Weichao Mao et al. "Model-Free Non-Stationary RL: Near-Optimal Regret and Applications in Multi-Agent RL and Inventory Control". In: *https://arxiv.org/abs/2010.03161* (2020).

[87] Weichao Mao et al. "Model-Free Non-Stationary RL: Near-Optimal Regret and Applications in Multi-Agent RL and Inventory Control". In: *arXiv preprint arXiv:2010.03161* (2020).

[88] Harry M Markowitz. "Portfolio selection". In: *Portfolio selection*. Yale university press, 1968.

[89] Olivier Massicot and Jakub Marecek. "On-line Non-Convex Constrained Optimization". In: *arXiv preprint arXiv:1909.07492* (2019).

[90] Aditya Modi et al. "Sample complexity of reinforcement learning using linearly combined model ensembles". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 2010–2020.

[91] John Moody and Matthew Saffell. "Learning to trade via direct reinforcement". In: *IEEE transactions on neural Networks* 12.4 (2001), pp. 875–889.

[92] Brett L Moore et al. "Reinforcement learning for closed-loop propofol anesthesia: a study in human volunteers". In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 655–696.

[93] Julie Mulvaney-Kemp, Salar Fattahi, and Javad Lavaei. "Smoothing property of load variation promotes finding global solutions of time-varying optimal power flow". In: *https://lavaei.ieor.berkeley.edu/DOPF_2020_2.pdf* (2020).

[94] Yael Niv et al. "Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain". In: *Journal of Neuroscience* 32.2 (2012), pp. 551–562.

[95] Ronald Ortner, Pratik Gajane, and Peter Auer. "Variational regret bounds for reinforcement learning". In: *Uncertainty in Artificial Intelligence*. PMLR. 2020, pp. 81–90.

[96] Takayuki Osogami. "Robustness and risk-sensitivity in Markov decision processes". In: *Advances in Neural Information Processing Systems* 25 (2012).

[97] Sindhu Padakandla. "A survey of reinforcement learning algorithms for dynamically varying environments". In: *ACM Computing Surveys (CSUR)* 54.6 (2021), pp. 1–25.

[98] Santiago Paternain et al. "Safe policies for reinforcement learning via primal-dual methods". In: *arXiv preprint arXiv:1911.09101* (2019).

[99] Joan Peuteman and Dirk Aeyels. "Exponential stability of nonlinear time-varying differential equations and partial averaging". In: *Mathematics of Control, Signals and Systems* 15.1 (2002), pp. 42–70.

[100] Shuang Qiu et al. "Upper confidence primal-dual reinforcement learning for CMDP with adversarial loss". In: *arXiv preprint arXiv:2003.00660* (2020).

[101] Christopher V Rao, James B Rawlings, and David Q Mayne. "Constrained state estimation for nonlinear discrete-time systems: Stability and moving horizon approximations". In: *IEEE Transactions on Automatic Control* 48.2 (2003), pp. 246–258.

[102] JB Rosen. "The gradient projection method for nonlinear programming. Part II. Nonlinear constraints". In: *Journal of the Society for Industrial and Applied Mathematics* 9.4 (1961), pp. 514–532.

[103] Ahmad EL Sallab et al. "Deep reinforcement learning framework for autonomous driving". In: *Electronic Imaging* 2017.19 (2017), pp. 70–76.

[104] M Salman Asif and Justin Romberg. "Sparse Recovery of Streaming Signals Using L1-Homotopy". In: *arXiv preprint arXiv:1306.3331* (2013).

[105] Arthur Sard. "The measure of the critical values of differentiable maps". In: *Bulletin of the American Mathematical Society* 48.12 (1942), pp. 883–890.

[106] Andrew M Saxe, James L McClelland, and Surya Ganguli. "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks". In: *arXiv preprint arXiv:1312.6120* (2013).

[107] John Schulman et al. "Proximal policy optimization algorithms". In: *arXiv preprint arXiv:1707.06347* (2017).

[108] John Schulman et al. "Trust region policy optimization". In: *International conference on machine learning*. PMLR. 2015, pp. 1889–1897.

[109] Yun Shen, Wilhelm Stannat, and Klaus Obermayer. "Risk-sensitive Markov control processes". In: *SIAM Journal on Control and Optimization* 51.5 (2013), pp. 3652–3672.

[110] Andrea Simonetto. "Time-varying convex optimization via time-varying averaged operators". In: *arXiv preprint arXiv:1704.07338* (2017).

[111] Andrea Simonetto et al. "A class of prediction-correction methods for time-varying convex optimization". In: *IEEE Transactions on Signal Processing* 64.17 (2016), pp. 4576–4591.

[112] Rahul Singh, Abhishek Gupta, and Ness B Shroff. "Learning in Markov decision processes under constraints". In: *arXiv preprint arXiv:2002.12435* (2020).

[113] Georg Still. "Lectures on parametric optimization: An introduction". In: *Optimization Online* (2018).

[114] Weijie Su, Stephen Boyd, and Emmanuel Candes. "A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights". In: *Advances in Neural Information Processing Systems*. 2014, pp. 2510–2518.

[115] Ju Sun, Qing Qu, and John Wright. "Complete dictionary recovery over the sphere I: Overview and the geometric picture". In: *IEEE Transactions on Information Theory* 63.2 (2016), pp. 853–884.

[116] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[117] Aviv Tamar, Yonatan Glassner, and Shie Mannor. "Optimizing the CVaR via sampling". In: *Twenty-Ninth AAAI Conference on Artificial Intelligence*. 2015.

[118] Aviv Tamar et al. "Policy gradient for coherent risk measures". In: *Advances in neural information processing systems* 28 (2015).

[119] KUNIO Tanabe. "An algorithm for constrained maximization in nonlinear programming". In: *J. Oper. Res. Soc. Jpn* 17 (1974), pp. 184–201.

[120] Yujie Tang, Krishnamurthy Dvijotham, and Steven Low. "Real-time optimal power flow". In: *IEEE Transactions on Smart Grid* 8.6 (2017), pp. 2963–2973.

[121] Yujie Tang et al. "Running Primal-Dual Gradient Method for Time-Varying Nonconvex Problems". In: *arXiv preprint arXiv:1812.00613* (2018).

[122] Andrew R Teel, Joan Peuteman, and Dirk Aeyels. "Semi-global practical asymptotic stability and averaging". In: *Systems & Control Letters* 37.5 (1999), pp. 329–334.

[123] Ahmed Touati and Pascal Vincent. "Efficient learning in non-stationary linear Markov decision processes". In: *arXiv preprint arXiv:2010.12870* (2020).

[124] Chen-Yu Wei and Haipeng Luo. "Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach". In: *Conference on Learning Theory*. PMLR. 2021, pp. 4300–4354.

[125] Andre Wibisono, Ashia C Wilson, and Michael I Jordan. "A variational perspective on accelerated methods in optimization". In: *Proceedings of the National Academy of Sciences* 113.47 (2016), E7351–E7358.

[126] Lin Yang and Mengdi Wang. "Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 10746–10756.

[127] Lin Yang and Mengdi Wang. "Sample-optimal parametric Q-learning using linearly additive features". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 6995–7004.

[128] Donghao Ying, Yuhao Ding, and Javad Lavaei. "A dual approach to constrained markov decision processes with entropy regularization". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 1887–1909.

[129] Ming Yu et al. "Convergent policy optimization for safe reinforcement learning". In: *Advances in Neural Information Processing Systems* 32 (2019), pp. 3127–3139.

[130] Victor M Zavala et al. "On-line economic optimization of energy systems using weather forecast information". In: *Journal of Process Control* 19.10 (2009), pp. 1725–1736.

[131] Richard Y Zhang, Somayeh Sojoudi, and Javad Lavaei. "Sharp restricted isometry bounds for the inexistence of spurious local minima in nonconvex matrix recovery". In: *Journal of Machine Learning Research* (2019).

[132] Han Zhong, Zhuoran Yang, and Zhaoran Wang Csaba Szepesvári. "Optimistic Policy Optimization is Provably Efficient in Non-stationary MDPs". In: *arXiv preprint arXiv:2110.08984* (2021).

[133] Dongruo Zhou, Jiafan He, and Quanquan Gu. "Provably efficient reinforcement learning for discounted MDPs with feature mapping". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 12793–12802.

[134] Huozhi Zhou et al. "Nonstationary reinforcement learning with linear function approximation". In: *arXiv preprint arXiv:2010.04244* (2020).