First-Order Regret in Reinforcement Learning with Linear Function Approximation: A Robust Estimation Approach

Andrew Wagenmaker* Yifang Chen† Max Simchowitz‡ Simon S. Du§
Kevin Jamieson¶

June 18, 2022

Abstract

Obtaining first-order regret bounds—regret bounds scaling not as the worst-case but with some measure of the performance of the optimal policy on a given instance—is a core question in sequential decision-making. While such bounds exist in many settings, they have proven elusive in reinforcement learning with large state spaces. In this work we address this gap, and show that it is possible to obtain regret scaling as $\widetilde{\mathcal{O}}(\sqrt{d^3H^3 \cdot V_1^\star \cdot K} + d^{3.5}H^3 \log K)$ in reinforcement learning with large state spaces, namely the linear MDP setting. Here $V_1^\star$ is the value of the optimal policy and K is the number of episodes. We demonstrate that existing techniques based on least squares estimation are insufficient to obtain this result, and instead develop a novel robust self-normalized concentration bound based on the robust Catoni mean estimator, which may be of independent interest.

1 Introduction

A central question in reinforcement learning (RL) is understanding precisely how long an agent must interact with its environment before learning to behave near-optimally. One popular way to measure this duration of interaction is by studying the *regret* $\mathcal{R}_K$, or cumulative suboptimality, of online reinforcement algorithms that explore an unknown environment across K episodes of interaction. Typical regret guarantees scale as $\mathcal{R}_K \le \mathcal{O}(\sqrt{\mathrm{poly}(d, H) \cdot K})$, where d measures the "size" of the environment and H the horizon length of each episode.

In many cases, however, regret bounds scaling at least as large as $\Omega(\sqrt{K})$ may be deeply unsatisfactory. Consider, for example, an environment where the agent receives rewards only at very hard-to-reach states; that is, states which can only be visited with some small probability $p \ll 1$. In this case, the maximal cumulative reward, optimal cumulative expected-reward, or *value* $V_1^\star$ will also be quite small. In other words, the cost of making a "mistake" at any given episode results in a loss of at most $V_1^\star$ reward, and the cumulative loss associated with, say $\sqrt{K}$, mistakes, should also scale with this maximal penalty.

Motivated by this observation, there has been much recent interest in achieving so-called small-value, small-loss, or "first-order" regret bounds, which scale in proportion to $V_1^\star$: $\mathcal{R}_K \le$

*University of Washington, Seattle. Email: ajwagen@cs.washington.edu
†University of Washington, Seattle. Email: yifangc@cs.washington.edu
‡CSAIL, MIT. Email: msimchow@mit.edu
§University of Washington, Seattle. Email: ssdu@cs.washington.edu
¶University of Washington, Seattle. Email: jamieson@cs.washington.edu

$\mathcal{O}(\sqrt{V_1^\star \cdot \mathrm{poly}(d,H) \cdot K})$ (it is well know that the the scaling $\sqrt{V_1^\star K}$ is unimprovable in general, even in simple settings). Bounds of this form have received considerable attention in the online learning, bandits, and contextual bandits communities, and were responsible for initiating the study of a broad array of *instance-dependent* regret bounds in tabular (i.e. finite-state, finite-action) RL settings as well.

First-Order Regret Beyond Tabular RL. Though first-order regret has been achieved in both non-dynamic environments (e.g. contextual bandits) and in dynamic environments with finite state spaces (tabular RL) (Zanette & Brunskill, 2019; Foster & Krishnamurthy, 2021), extension to reinforcement learning in large state and action spaces has proven elusive. The main difficulty is that, even though the cumulative expected value of any policy is bounded as $V_1^\star$, the value-to-go associated with starting at some state s_h at step h, denoted $V_h^\star(s_h)$, may be considerably larger. Again, the paradigmatic example is when the reward is equal to 1 on a handful of very hard-to-reach states. This means that the variance of any learned predictor of the value function $V_h^\star(s_h)$ may also be highly nonuniform in the state s_h. In the RL setting, this becomes more challenging because the distribution across states evolves as the agent refines its policies. And while in tabular settings, one can address the non-uniformity by reasoning about each of the finitely-many states separately, there is no straightforward way to generalize the argument to larger state spaces.

Contributions and Techniques. In this paper, we provide first-order regret bounds for reinforcement learning in large state spaces, the first of their kind in this setting. Our results focus on the setting of MDPs with linear function approximation (Jin et al., 2020b), where the transition operators are described by linear functions in a known, d-dimensional featurization of a potentially infinite-cardinality state space. In this setting, we achieve the following regret bound.

Theorem 1 (Informal)**.** *Our proposed algorithm,* FORCE*, achieves the following first-order regret bound with high probability:* $\mathcal{R}_K \leq \widetilde{\mathcal{O}}(\sqrt{d^3H^3 \cdot V_1^\star \cdot K} + d^{3.5}H^3 \log K)$.

To our knowledge, FORCE is the first algorithm to achieve first-order regret for RL in large state spaces. Our algorithm builds on the LSVI-UCB algorithm of (Jin et al., 2020b) for worst-case (non-first-order) regret in linear MDPs. LSVI-UCB relies on solving successive linear regression problems to estimate the Bellman-backups of optimistic overestimates of the optimal value function. In that work, the analysis of the regression estimates relies on a so-called "self-normalized martingale" inequality for online least squares—a powerful tool which quantifies the refinement of a ridge-regularized least-squares estimator under an arbitrary sequence of regression covariates $\boldsymbol{\phi}_t$ to targets y_t satisfying $\mathbb{E}[y_t \mid \boldsymbol{\phi}_t] = \langle \boldsymbol{\phi}_t, \boldsymbol{\theta}_\star \rangle$, and under the assumption of sub-Gaussian noise. This tool has seen widespread application not only in linear RL, but in bandit and control domains as well (Abbasi-Yadkori et al., 2011; Sarkar & Rakhlin, 2019).

In the tabular RL setting, first-order regret bounds can be obtained by applying Bernstein-style concentration bounds, which allows the exploration level to adapt to the underlying problem difficulty. Towards achieving first-order regret in linear RL, we might hope that a similar approach could be used, and that developing variance-aware or Bernstein-style self-normalized bounds may provide the necessary refinements. A second challenge arises in the RL setting, however, since, as mentioned, the "noise" is inherently heteroscedastic (i.e., the noise variance changes with time)—the variance of y_t depends on $\boldsymbol{\phi}_t$. Thus, not only do we require a variance-aware self-normalized bound, but such a bound must be able to handle heteroscedastic noise as well.

The recent work of Zhou et al. (2020) addresses both of these issues—proposing a Bernstein-style self-normalized bound, and overcoming the heteroscedasticity by relying on a weighted least-squares

estimator which normalizes each sample by its variance. A naive application of these techniques, however, results in a scaling of $1/\sigma_{\min}$ in the regret bound, where $\sigma_{\min}$ is the *minimum noise variance across time*. While this dependence can be reduced somewhat, ultimately, it could be prohibitively large, and prevents us from achieving a first-order regret bound in the case when $V_1^\star$ is small.

The $1/\sigma_{\min}$ dependence arises because, if we normalize by the variance in our weighted least-squares estimate, the normalized "noise" has magnitude, in the worst case, of $\mathcal{O}(1/\sigma_{\min})$. In other words, we are paying for the "heavy tail" of the noise, rather than simply its variance. Obtaining concentration independent of such heavy tails is a problem well-studied in the robust statistics literature. Towards addressing this difficulty in the RL setting, we take inspiration from this literature, and propose applying the *robust Catoni estimator* (Catoni, 2012). In particular, we develop a novel self-normalized version of the Catoni estimator, as follows.

Proposition 2 (Self-Normalized Heteroscedastic Catoni Estimation, Informal)**.** *Given observations $y_t = \langle \boldsymbol{\theta}_\star, \boldsymbol{\phi}_t \rangle + \eta_t$ with $\mathbb{E}[y_t \mid \boldsymbol{\phi}_t] = \langle \boldsymbol{\phi}_t, \boldsymbol{\theta}_\star \rangle$, $\mathbb{E}[|\eta_t|^2 \mid \boldsymbol{\phi}_t] < \infty$, and $|\eta_t| < \infty$ with probability 1, let $\mathsf{cat}[\boldsymbol{v}]$ denote a Catoni estimate of $\boldsymbol{\theta}_\star$ in direction $\boldsymbol{v}$ from the observed data. Then, with high probability, for all $\boldsymbol{v}$ simultaneously:*

$$\left|\mathsf{cat}\left[\boldsymbol{v}\right] - \boldsymbol{v}^\top \boldsymbol{\theta}_\star\right| \lesssim \|\boldsymbol{v}\|_{\boldsymbol{\Lambda}_T^{-1}} \cdot \left(\sqrt{\log 1/\delta + d \cdot C_{\log}} + \sqrt{\lambda}\|\boldsymbol{\theta}_\star\|_2\right) + (\text{lower order term}).$$

where $\boldsymbol{\Lambda}_T = \lambda I + \sum_{t=1}^T \sigma_t^{-2} \boldsymbol{\phi}_t \boldsymbol{\phi}_t^\top$, σ_t^2 is an upper bound on $\mathbb{E}[y_t^2 \mid \boldsymbol{\phi}_t]$, $C_{\log}$ is logarithmic in problem parameters, and the 'lower order term' *can be made as small as T^{-q} for any constant $q > 0$.*

To apply Proposition 2, we take $\boldsymbol{\phi}_{h,k} = \boldsymbol{\phi}(s_{h,k}, a_{h,k})$ as the features, and $y_{h,k} = V_{h+1}^k(s_{h+1,k})$ as the targets, where $V_{h+1}^k(\cdot)$ is an optimistic overestimate of the value function. In particular, $V_{h+1}^k(\cdot)$ depends on, and may be correlated with, past data. Following Jin et al. (2020b), we address this issue by establishing an error bound which holds uniformly over possible value functions $V_{h+1}^k(\cdot)$. We call this guarantee the 'Heteroscedastic Self-Normalized Inequality with Function Approximation', and state it formally in Section 5. The proof combines Proposition 2 with a careful covering argument, which (unlike past approaches based on standard ridge-regularized least squares) requires a novel sensitivity analysis of the Catoni estimator.

2 Related Work

Worst-Case Regret Bounds in Tabular RL. A significant amount of work has been devoted to obtaining worst-case optimal bounds in the setting of tabular RL (Kearns & Singh, 2002; Kakade, 2003; Azar et al., 2017; Dann et al., 2017; Jin et al., 2018; Dann et al., 2019; Wang et al., 2020; Zhang et al., 2020b,a). These approaches fall into both the model-based (Azar et al., 2017; Dann et al., 2017) as well as the model-free category (Jin et al., 2018). While the exact bounds differ, they all take the form $\widetilde{\mathcal{O}}(\sqrt{\mathrm{poly}(H) \cdot SAK} + \mathrm{poly}(S, A, H))$. Recently, several works have focused on obtaining bounds that only scale logarithmically with the horizon, H, in the setting of time-invariant MDPs with rewards absolutely bounded by 1. Zhang et al. (2020a) answers the question of whether horizon-free learning is possible by proposing an algorithm with regret scaling as $\widetilde{\mathcal{O}}(\sqrt{SAK} + S^2A)$—independent of polynomial dependence on H. This is known to be worst-case minimax optimal.

RL with Function Approximation. In the last several years, there has been an explosion of interest in the RL community in obtaining provably efficient RL algorithms relying on function

approximation. An early work in this direction, Jiang et al. (2017), considers general function classes and shows that MDPs having small "Bellman rank" are efficiently learnable. Several recent works have extended their results significantly (Du et al., 2021; Jin et al., 2021). In the special case of linear function approximation, a vast body of recent work exists (Yang & Wang, 2019; Jin et al., 2020b; Wang et al., 2019; Du et al., 2019; Zanette et al., 2020a,b; Ayoub et al., 2020; Jia et al., 2020; Weisz et al., 2021; Zhou et al., 2020, 2021; Zhang et al., 2021; Wang et al., 2021). A variety of assumptions are made in these works, and we highlight two of them in particular. First, the linear MDP model of Jin et al. (2020b), which is the setting we consider in this work, assumes the transition probabilities and reward functions can both be parameterized as a linear function of a feature map. Second, the linear mixture MDP setting of (Jia et al., 2020; Ayoub et al., 2020; Zhou et al., 2020) makes no linearity assumption on the reward function, but assumes that the transition probabilities are the linear parameterization of d known transition kernels. Notably, the linear MDP assumption has infinite degrees of freedom, and as such model-free approaches are more appropriate, while the linear mixture MDP setting has only dH degrees of freedom, making model-based learning effective.

As mentioned above, of note in the linear function approximation literature is the work of Zhou et al. (2020), which proposes an algorithm with regret scaling as $\widetilde{\mathcal{O}}(\sqrt{(d^2H^3 + dH^3)\,K})$, which they show is minimax optimal when $d \geq H$. Their result relies on a Bernstein-style self-normalized confidence bound. While they show that the variance dependence of the Bernstein bound allows them to achieve minimax optimality, as noted, it is insufficient to achieve a first-order bound, motivating our use of the Catoni estimator.

First-Order and Problem-Dependent Regret Bounds in RL. The RL community has tended to pursue two primary directions towards obtaining problem-dependent regret bounds. The first is the aforementioned first-order bounds, the focus of this work. To our knowledge, the only work in the RL literature to obtain first-order regret is that of Zanette & Brunskill (2019), which only holds in the tabular setting. Zanette & Brunskill (2019) obtain several different forms of such a bound, showing that their algorithm, EULER, has regret which can be bounded as either

$$\mathcal{R}_K \leq \widetilde{\mathcal{O}}(\sqrt{\mathbb{Q}^* SAHK}) \quad \text{or} \quad \mathcal{R}_K \leq \widetilde{\mathcal{O}}(\sqrt{\mathcal{G}^2 SAK})$$

where $\mathbb{Q}^* = \max_{s,a,h} \left(\mathrm{Var}[R_h(s,a)] + \mathrm{Var}_{s' \sim P_h(\cdot|s,a)}[V^\star_{h+1}(s')]\right)$ and $\mathcal{G}$ is a deterministic upper bound on the maximum attainable reward on a single trajectory for any policy π: $\sum_{h=1}^H R(s_h, \pi_h(s_h)) \leq \mathcal{G}$. A subsequent work, Jin et al. (2020a), showed that a slight modification to the analysis of EULER allows one to obtain regret of[1]

$$\mathcal{R}_K \leq \widetilde{\mathcal{O}}(\sqrt{SAH \cdot V^\star_1 K}).$$

A second approach to instance-dependence, taken by (Simchowitz & Jamieson, 2019; Xu et al., 2021; Dann et al., 2021), seeks to obtain regret scaling with the suboptimality gaps. This yields regret bounds of the form $\mathcal{O}\left(\sum_{s,a,h:\Delta_h(s,a)>0} \frac{\mathrm{poly}(H)\cdot \log K}{\Delta_h(s,a)}\right)$ where $\Delta_h(s,a) := V^\star_h(s) - Q^\star_h(s,a)$ is the suboptimality of playing action a in state s at step h. While these works consider only the tabular setting, recently He et al. (2021) obtained regret in the linear MDP setting of $\mathcal{O}(\frac{d^3H^5\log(K)}{\Delta_{\min}})$ and in the linear mixture MDP setting of $\mathcal{O}(\frac{d^2H^5\log^3(K)}{\Delta_{\min}})$, where $\Delta_{\min}$ is the minimum non-zero gap in the MDP. Gap-dependent regret bounds allow for a characterization of the regret in terms of

[1] Note that this result was shown for an MDP where the reward function was non-zero only at a single (s,h). Their analysis can be extended to arbitrary reward functions, however, though extra H factors will be incurred.

fine-grained problem-dependent quantities. However, they typically capture the *total regret incurred to solve the problem*, and are therefore overly pessimistic over shorter time horizons.

First-Order Regret Beyond RL. A significant body of literature exists towards obtaining first-order regret bounds in settings other than RL. This work spans areas as diverse as statistical learning (Vapnik & Chervonenkis, 1971; Srebro et al., 2010), online learning (Freund & Schapire, 1997; Auer et al., 2002; Cesa-Bianchi et al., 2007; Luo & Schapire, 2015; Koolen & Van Erven, 2015; Foster et al., 2015), and multi-armed bandits, adversarial bandits, and semibandits (Allenberg et al., 2006; Hazan & Kale, 2011; Neu, 2015; Lykouris et al., 2018; Wei & Luo, 2018; Bubeck & Sellke, 2020; Ito et al., 2020).

We highlight in particular the work in the contextual bandit setting. A COLT 2017 open problem (Agarwal et al., 2017) posed the question of obtaining first-order bounds for contextual bandits to the community, which Allen-Zhu et al. (2018) subsequently addressed by obtaining a computationally inefficient algorithm achieving this. Foster & Krishnamurthy (2021) built on this, showing that it is possible to achieve such a bound with a computationally efficient algorithm. While Foster & Krishnamurthy (2021) considers function approximation, their regret bound scales with the number of actions, and is therefore not applicable to large action spaces.

Robust Mean Estimation. Our algorithm critically relies on robust mean estimation to obtain concentration bounds that avoid large lower-order terms. We rely in particular on the Catoni estimator, first proposed in Catoni (2012). While the original Catoni estimator assumes i.i.d. data, Wei et al. (2020) show that a martingale version of Catoni is possible, which is what we apply in this work. We remark that several applications of the Catoni estimator to linear bandits have been proposed recently (Camilleri et al., 2021; Lee et al., 2021). We refer the reader to the survey Lugosi & Mendelson (2019) for a discussion of other robust mean estimators.

3 Preliminaries

Notation. All logarithms are base-e unless otherwise noted. We let $\mathsf{logs}(x_1, x_2, \dots, x_n) := \sum_{i=1}^{n} \log(e + x_i)$ denote a term which is at most logarithmic in arguments $x_1, x_2, \dots, x_n \geq 0$. We let $\mathcal{B}^d(R) := \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| \leq R\}$ denote the ball of radius R in $\mathbb{R}^d$, and specialize $\mathcal{B}^d := \mathcal{B}^d(1)$ to denote the unit ball. $\mathcal{S}^{d-1}$ denotes the unit sphere in $\mathbb{R}^d$. We use $\lesssim$ to denote inequality up to absolute constants, $\mathcal{O}(\cdot)$ to hide absolute constants and lower-order terms, and $\widetilde{\mathcal{O}}(\cdot)$ to hide absolute constants, logarithmic terms, and lower-order terms. Throughout, we let bold characters refer to vectors and matrices and standard characters refer to scalars.

We also highlight MDP-specific notation; see below for further exposition. We let $s_{h,k}$ and $a_{h,k}$ denote the state and action at step h and episode k, and denote features and rewards $\boldsymbol{\phi}_{h,k} := \boldsymbol{\phi}(s_{h,k}, a_{h,k})$, $r_{h,k} := r_h(s_{h,k}, a_{h,k})$. π^k denotes the policy played at episode k. We use $\mathcal{F}_{h,k}$ to denote the σ-field $\sigma(\cup_{h'=1}^{H} \cup_{k'=1}^{k-1} \{(s_{h',k'}, a_{h',k'})\} \cup_{h'=1}^{h} \{(s_{h',k}, a_{h',k})\})$, so that $\boldsymbol{\phi}_{h,k}$ is $\mathcal{F}_{h,k}$-measurable. We will let $\mathbb{E}_h[V](s,a) = \mathbb{E}_{s' \sim P_h(\cdot|s,a)}[V(s')]$, so $\mathbb{E}_h[V](s,a)$ denotes the expected next-state value of V given that we are in state s and play action a at time h.

3.1 Markov Decision Processes

We consider finite-horizon, episodic Markov Decision Processes (MDPs) with time inhomogeneous transition kernel. An MDP is described by a tuple $(\mathcal{S}, \mathcal{A}, H, \{P_h\}_{h=1}^{H}, \{r_h\}_{h=1}^{H})$, with $\mathcal{S}$ the set of

states, $\mathcal{A}$ the set of actions, H the horizon, $P_h : \mathcal{S} \times \mathcal{A} \to \triangle(\mathcal{S})$ the probability transition kernel at time h, and $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ the reward function. We assume that $\{P_h\}_{h=1}^H$ is initially unknown to the learner, but that r_h is deterministic and known. Without loss of generality, we further assume the intial state s_1 is deterministic.

At each episode, the agents begins in state s_1; then for each time step $h \geq 1$, an agent in state s_h takes action a_h, receives reward $r_h(s_h, a_h)$ and transitions to state s' with probability $P_h(s'|s_h, a_h)$. This process continues for H steps, at which point the MDP resets and the process repeats.

A policy $\pi : \mathcal{S} \times [H] \to \triangle(\mathcal{A})$ is a mapping from states to distributions over actions. For deterministic policies ($\forall h, s,\ \pi_h(s)$ is supported on only 1 action) we let $\pi_h(s)$ denote the unique action in the support of the distribution $\pi_h(s)$. To an agent playing a policy π, at step h they choose an action $a_h \sim \pi_h(s_h)$. We let $\mathbb{E}_\pi[\cdot]$ denote the expectation over the joint distribution trajectories $(s_1, a_1, \ldots, s_H, a_H)$ induced by policy π.

Value Functions. Given a policy π, the *Q-value function* for policy π is defined as follows:

$$Q_h^\pi(s, a) := \mathbb{E}_\pi \left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) | s_h = s, a_h = a \right].$$

In words, $Q_h^\pi(s, a)$ denotes the expected reward we will acquire by taking action a in state s at time h and then playing π for all subsequent steps. We also denote the *value function* by $V_h^\pi(s) = \mathbb{E}_{a\sim\pi_h(s)}[Q_h^\pi(s, a)]$, which corresponds to the expected reward we will acquire by playing policy π from state s at time h. The Q-function satisfies the Bellman equation:

$$Q_h^\pi(s, a) = r_h(s, a) + \mathbb{E}_h[V_{h+1}^\pi](s, a).$$

We denote the optimal Q-value function by $Q_h^\star(s, a) = \sup_\pi Q_h^\pi(s, a)$, the optimal value function by $V_h^\star(s) = \sup_\pi V_h^\pi(s)$, and the optimal policy by $\pi^\star$. We define $V_{H+1}^\pi(s) = Q_{H+1}^\pi(s, a) = 0$ for all s and a. Finally, note that we always have that $Q_h^\pi(s, a) \leq H$, for all π, h, s, a, since we collect a reward of at most 1 at every step.

Episodic MDPs and Regret. In this paper, we study minimizing the *regret* over K episodes of interaction. At each episode k, the learning agent selects a policy π^k, and receives a trajectory $(s_{1,k}, a_{1,k}, \ldots, s_{H,k}, a_{H,k})$. Again, the transition kernels $(P_h)_{h=1}^H$ are *unknown* to the learner, whereas (as discussed above), the reward function is known. The regret is defined as the cumulative suboptimality of the learner's policies:

$$\mathcal{R}_K = \sum_{k=1}^{K} \left[V_1^\star(s_1) - V_1^{\pi_k}(s_1) \right].$$

As s_1 is fixed, we will denote the *value* of policy π as $V_1^\pi := V_1^\pi(s_1)$. Using this notation we can express the regret as $\mathcal{R}_K = \sum_{k=1}^K [V_1^\star - V_1^{\pi_k}]$.

3.2 Reinforcement Learning with Linear Function Approximation

In the tabular RL setting, it is assumed that $|\mathcal{S}|$ and $|\mathcal{A}|$ are both finite. This assumption is quite limited in practice, however, and is not able to model real-world settings where the state and action spaces may be infinite. Towards relaxing this assumption, we consider the linear MDP setting of Jin et al. (2020b), which allows for infinite state and action spaces. In particular, this setting is defined as follows.

Definition 3.1 (Linear MDPs)**.** We say that an MDP is a *d-dimensional linear MDP*, if there exists some (known) feature map $\boldsymbol{\phi}(s,a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and H (unknown) signed measures $\boldsymbol{\mu}_h \in \mathbb{R}^d$ over $\mathcal{S}$ such that:

$$P_h(\cdot|s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\mu}_h(\cdot) \rangle.$$

We will assume that $\|\boldsymbol{\phi}(s,a)\|_2 \leq 1$ for all s,a, and $\||\boldsymbol{\mu}_h|(\mathcal{S})\|_2 = \|\int_{s\in\mathcal{S}} |\mathrm{d}\boldsymbol{\mu}_h(s)|\|_2 \leq \sqrt{d}$.

Note that, unlike the standard definition of linear MDPs which assumes that the reward is also linear, $r_h(s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h \rangle$, we consider more general possibly non-linear (though bounded) reward functions. To accommodate this change we must assume that the reward is deterministic and known to the learner. We also consider time-varying reward in the appendix, and in the subsequent section remark on how unknown rewards can be accommodated, if we assume they are linear.

We further note that there cannot exist an (s,a) for which $\boldsymbol{\phi}(s,a) = \mathbf{0}$, for otherwise Definition 3.1 would imply that $P_h(\cdot|s,a)$ is not a valid distribution. As shown in Jin et al. (2020b), the linear MDP setting includes tabular MDPs, while also encompassing more general, non-tabular settings, for example where the feature space corresponds to the d-dimensional simplex. A key property of linear MDPs is the following.

Lemma 3.1 (Lemma 2.3 of Jin et al. (2020b))**.** *For a linear MDP and any policy π, there exists some set of weights $\{\boldsymbol{w}_h^\pi\}_{h=1}^H$ such that $Q_h^\pi(s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{w}_h^\pi \rangle$ for all s,a,h.*

Lemma 3.1 motivates us to consider linear policy classes in developing our algorithm. More generally, the linear structure of the MDP implies that $\mathbb{E}_h[V](s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{w}_V \rangle$ for some $\boldsymbol{w}_V$ and *any arbitrary* function $V : \mathcal{S} \to \mathbb{R}$.

3.3 Catoni Estimation

A key tool in our algorithm is the robust Catoni estimator (Catoni, 2012). The Catoni estimator is defined as follows.

Definition 3.2 (The Catoni Estimator)**.** Let $X_1, \ldots, X_T$ be a sequence of real-values. The *Catoni robust mean estimator* with parameter $\alpha > 0$, denoted $\mathsf{cat}_{T,\alpha}$, is the unique root z of the function

$$f_{\mathsf{cat}}(z; X_{1:T}, \alpha) := \sum_{t=1}^{T} \psi_{\mathsf{cat}}(\alpha(X_t - z)), \tag{3.1}$$

where $\psi_{\mathsf{cat}}(\cdot)$ is defined by

$$\psi_{\mathsf{cat}}(y) = \begin{cases} \log(1 + y + y^2) & y \geq 0 \\ -\log(1 - y + y^2) & y < 0 \end{cases}.$$

The following result illustrates the key property of the Catoni estimator.

Proposition 3 (Theorem 5 of Lugosi & Mendelson (2019))**.** *Let $X_1, \ldots, X_T$ be independent, identically distributed random variables with mean μ and finite variance $\sigma^2 < \infty$. Let $\delta \in (0,1)$ be such that $T \geq 2\log(1/\delta)$. Then the Catoni mean estimator $\mathsf{cat}_{T,\alpha}$ with parameter*

$$\alpha = \sqrt{\frac{2\log 1/\delta}{T\sigma^2(1 + \frac{2\log 1/\delta}{T - 2\log 1/\delta})}}$$

satisfies the following guarantee with probability $1 - 2\delta$,

$$|\mathsf{cat}_{T,\alpha} - \mu| < \sqrt{\frac{2\sigma^2 \log 1/\delta}{T - 2\log 1/\delta}}.$$

As Proposition 3 shows, the Catoni estimator requires only that the second moment of the distribution is bounded to obtain concentration, and has estimation error which scales only with the second moment and independent of other properties of the distribution. We make key use of this result in the following analysis, and state our novel extension of the Catoni estimator to general regression settings in Section 5.

4 First-Order Regret in Linear MDPs

We are now ready to present our algorithm, FORCE.

Summary of Key Parameters. Our algorithm applies the robust Catoni estimator to measure the next-state expectation of the value function. The Catoni estimator requires an estimated upper bound on the value function, which we denote as $\bar{\mathsf{v}}_{h,k}$ and describe in detail below. Throughout, we let $\mathsf{v}_{\min} = 1/K$ denote a lower floor on these estimates. Using these estimates, we introduce the value-normalized feature covariance, with its regularized analogue

$$\boldsymbol{\Sigma}_{h,k} = \sum_{\tau=1}^{k} \frac{1}{\bar{\mathsf{v}}_{h,\tau}^2} \boldsymbol{\phi}_{h,\tau} \boldsymbol{\phi}_{h,\tau}^\top, \quad \boldsymbol{\Lambda}_{h,k} = \lambda I + \boldsymbol{\Sigma}_{h,k} \tag{4.1}$$

For a given (k, h) and direction $\boldsymbol{v}$, we use the above covaraince to define a (directional) Catoni parameter

$$\alpha_{k,h}(\boldsymbol{v}) = \min\left\{\beta \cdot \|\boldsymbol{v}\|_{\boldsymbol{\Sigma}_{h,k-1}}^{-1}, \alpha_{\max}\right\},$$

where β is defined in the FORCE pseudocode, and we take $\alpha_{\max} = K/\mathsf{v}_{\min} = K^2$. For a given k, h and direction $\boldsymbol{v} \in \mathbb{R}^d$, we adopt the $\alpha_{k,h}(\boldsymbol{v})$ as the Catoni parameter, and set $\mathsf{cat}_{h,k}[\boldsymbol{v}]$ to refer to the associated Catoni estimate on the data

$$X_\tau = \boldsymbol{v}^\top \boldsymbol{\phi}_{h,\tau} V_{h+1}^k(s_{h+1,\tau}) / \bar{\mathsf{v}}_{h,\tau}^2, \quad \tau = 1, \dots, k-1.$$

Algorithm Description. FORCE proceeds similarly to the LSVI-UCB algorithm of Jin et al. (2020b) by approximating the classical value-iteration update:

$$Q_h^\star(s,a) \leftarrow r_h(s,a) + \mathbb{E}_h[\max_{a'} Q_{h+1}^\star(\cdot, a')](s,a), \quad \forall s, a. \tag{4.2}$$

It is known that this update converges to the optimal value function. While in practice we cannot evaluate the expectation directly, it stands to reason that an update approximating (4.2) may converge to an approximation of the optimal value function. As in Jin et al. (2020b), we therefore apply an *optimistic*, empirical variant of the value iteration update, which replaces $Q^\star$ with Q^k, the optimistic estimate of $Q^\star$ at round k, and the exact expectation with an empirical expectation. The key difference in our approach as compared to Jin et al. (2020b) is the setting of the optimistic

Algorithm 1 **F**irst-**O**rder **R**egret via **C**atoni **E**stimation (FORCE)

1: **input:** confidence δ, number of episodes K
2: $\lambda \leftarrow 1/H^2$, $\mathsf{v}_{\min} \leftarrow 1/K$, $c \leftarrow$ universal constant
3: $K_{\text{init}} \leftarrow c\left(d^2 \log(\max\{d, \mathsf{v}_{\min}^{-1}, K, H\}) + \log(2HK/\delta)\right)$
4: $\beta \leftarrow 6\sqrt{cd^2 \log\left(\max\{d, \mathsf{v}_{\min}^{-1}, H, K\}\right) + \log(2HK/\delta)}$
5: **for** $k = 1, 2, 3, \ldots, K$ **do**
6: **for** $h = H, H-1, \ldots, 1$ **do**
7: **if** $k \le K_{\text{init}}$ **then**
8: $\bar{\mathsf{v}}^2_{h,k-1} \leftarrow 2H^2$
9: **else**
10: $\bar{\mathsf{v}}^2_{h,k-1} \leftarrow \max\Big\{20H\mathsf{cat}_{h,k-1}[(k-2)\boldsymbol{\Lambda}^{-1}_{h,k-2}\boldsymbol{\phi}_{h,k-1}] + 20H\beta\|\boldsymbol{\phi}_{h,k-1}\|_{\boldsymbol{\Lambda}^{-1}_{h,k-2}}$
11: $+20H\mathsf{v}_{\min}\beta^2/(k-1)^2, \mathsf{v}^2_{\min}\Big\}$
12: Form $\boldsymbol{\Lambda}_{h,k-1} \leftarrow \lambda I + \sum_{\tau=1}^{k-1} \frac{1}{\bar{\mathsf{v}}^2_{h,\tau}} \boldsymbol{\phi}_{h,\tau}\boldsymbol{\phi}_{h,\tau}^\top$ as in Equation (4.1)
// $\widehat{\mathbb{E}}_h[V^k_{h+1}](\boldsymbol{v}) := \mathsf{cat}_{h,k}[(k-1)\boldsymbol{\Lambda}^{-1}_{h,k-1}\boldsymbol{v}]$
13: Compute linear summary $\widehat{\boldsymbol{w}}^k_h \leftarrow \arg\min_{\boldsymbol{w}} \sup_{\boldsymbol{v}\in\mathcal{B}^d\setminus\{\mathbf{0}\}} |\langle \boldsymbol{v}, \boldsymbol{w}\rangle - \widehat{\mathbb{E}}_h[V^k_{h+1}](\boldsymbol{v})|/\|\boldsymbol{v}\|_{\boldsymbol{\Lambda}^{-1}_{h,k-1}}$.
14: $Q^k_h(\cdot,\cdot) \leftarrow \min\{r_h(\cdot,\cdot) + \langle \boldsymbol{\phi}(\cdot,\cdot), \widehat{\boldsymbol{w}}^k_h\rangle + 6\beta\|\boldsymbol{\phi}(\cdot,\cdot)\|_{\boldsymbol{\Lambda}^{-1}_{h,k-1}} + 12\mathsf{v}_{\min}\beta^2/k^2, H\}$
15: $V^k_h(\cdot) \leftarrow \max_a Q^k_h(\cdot, a)$
16: **for** $h = 1, 2, \ldots, H$ **do**
17: Play $a_{h,k} = \arg\max_a Q^k_h(s_{h,k}, a)$, observe $r_{h,k}, s_{h+1,k}$

estimate. While Jin et al. (2020b) rely on a simple least-squares estimator to approximate the expectation, we rely on the Catoni estimator. We show that, with high probability:

$$|\mathsf{cat}_{h,k}[(k-1)\boldsymbol{\Lambda}^{-1}_{h,k-1}\boldsymbol{\phi}(s,a)] - \mathbb{E}_h[V^k_{h+1}](s,a)| \lesssim \beta\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}^{-1}_{h,k-1}}.$$

Thus, setting $\widehat{\mathbb{E}}_h[V^k_{h+1}](\boldsymbol{v})$, our estimate of the next-state expectation, to $\mathsf{cat}_{h,k}[(k-1)\boldsymbol{\Lambda}^{-1}_{h,k-1}\boldsymbol{v}]$ ensures that $\widehat{\mathbb{E}}_h[V^k_{h+1}](\boldsymbol{v})$ approximates the expectation in (4.2) for $\boldsymbol{v} = \boldsymbol{\phi}(s,a)$. As discussed in more detail in Section 5.4, instead of using $\widehat{\mathbb{E}}_h[V^k_{h+1}](\boldsymbol{v})$ directly, Line 13 summarizes $\widehat{\mathbb{E}}_h[V^k_{h+1}](\boldsymbol{v})$ with a linear approximation to it,

$$\widehat{\boldsymbol{w}}^k_h \leftarrow \arg\min_{\boldsymbol{w}} \sup_{\boldsymbol{v}\in\mathcal{B}^d\setminus\{\mathbf{0}\}} \frac{|\langle \boldsymbol{v}, \boldsymbol{w}\rangle - \widehat{\mathbb{E}}_h[V^k_{h+1}](\boldsymbol{v})|}{\|\boldsymbol{v}\|_{\boldsymbol{\Lambda}^{-1}_{h,k-1}}}. \tag{4.3}$$

This approximation, $\langle \boldsymbol{\phi}(s,a), \widehat{\boldsymbol{w}}^k_h\rangle$, is shown to be an accurate approximation of $\widehat{\mathbb{E}}_h[V^k_{h+1}](\boldsymbol{v})$ in Lemma 5.2, intuitively because the "ground truth" is itself linear. Solving the optimization on Line 13 may be computationally inefficient, so we provide a computationally-efficient modification in Section 4.2 which has only slightly larger regret.

With the aforementioned linear approximation, we define our optimistic overestimate of the Q-function on Line 14 as

$$Q^k_h(\cdot,\cdot) \leftarrow \min\{r_h(\cdot,\cdot) + \langle \boldsymbol{\phi}(\cdot,\cdot), \widehat{\boldsymbol{w}}^k_h\rangle + 6\beta\|\boldsymbol{\phi}(\cdot,\cdot)\|_{\boldsymbol{\Lambda}^{-1}_{h,k-1}} + 12\mathsf{v}_{\min}\beta^2/k^2, H\},$$

approximating the value-iteration update of (4.2) with additional bonuses to account for the approximation error.

A Note On Scaling. To achieve first-order regret, we need both the *errors in our estimates* and the *magnitude of the bonuses* to scale with the magnitude of the value function. To accomplish this, we ensure the bonuses scale with $\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}}$, where $\boldsymbol{\Lambda}_{h,k-1}$ is the regularized variance-normalized covariance in (4.1), and $\bar{\mathsf{v}}_{h,k-1}^2$ is defined as

$$\bar{\mathsf{v}}_{h,k-1}^2 := \max\left\{20H\mathsf{cat}_{h,k-1}[(k-2)\boldsymbol{\Lambda}_{h,k-2}^{-1}\boldsymbol{\phi}_{h,k-1}] + 20H\beta(\|\boldsymbol{\phi}_{h,k-1}\|_{\boldsymbol{\Lambda}_{h,k-2}^{-1}} + \tfrac{\mathsf{v}_{\min}\beta}{(k-1)^2}),\, \mathsf{v}_{\min}^2\right\} \quad (4.4)$$

so that, up to effectively lower-order terms accounting for the estimation error,

$$\mathbb{E}_h[(V_{h+1}^\tau)^2](s_{h,\tau}, a_{h,\tau}) \approx \bar{\mathsf{v}}_{h,\tau}^2.$$

As we will show, this choice of $\bar{\mathsf{v}}_{h,\tau}^2$ is sufficiently large to ensure our Catoni estimate, $\widehat{\mathbb{E}}_h[V_{h+1}^k](\boldsymbol{v})$, concentrates. At the same time, when $\mathbb{E}_h[(V_{h+1}^\tau)^2](s_{h,\tau}, a_{h,\tau})$ is small for some τ, the variance-normalized regularization $\boldsymbol{\Lambda}_{h,k-1}$ ensures the bonus $\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}}$ is small as well.

4.1 Formal Regret Guarantee

We analyze two regret bounds for FORCE. In the first bound, we analyze the description given in Algorithm 1, which gives a sharper regret guarantee at the expense of computational inefficiency:

Theorem 4 (Main Regret Bound)**.** *Fix a failure probability $\delta \in (0,1)$ and $K \in \mathbb{N}$. Then, the regret of* FORCE *as specified in Algorithm 1 satisfies the following bound with probability at least $1-3\delta$:*

$$\mathcal{R}_K \le c_1\sqrt{d^3H^3V_1^\star K \cdot \log^3(HK/\delta)} + c_2 d^{7/2}H^3\log^{7/2}(HK/\delta)$$

for universal constants c_1, c_2.

As Theorem 4 shows, up to lower order terms scaling only polynomially in $d, H, \log K$, and $\log 1/\delta$, FORCE achieves a first-order scaling in its leading order term of $\mathcal{O}(\sqrt{V_1^\star K})$. We sketch the proof of Theorem 4 in Section 6 and defer the full proof to Appendix B.

Comparison to Jin et al. (2020b). Note that $V_1^\star \le H$, since we assume that the reward at each step is bounded by 1. Thus, Theorem 4 shows that in the worst case FORCE has regret scaling as $\widetilde{\mathcal{O}}(\sqrt{d^3H^4K})$. This exactly matches the regret of LSVI-UCB given in Jin et al. (2020b). However, we could have that $V_1^\star \ll H$, in which case FORCE significantly improves on LSVI-UCB. Note also that the minimax lower bound scales at least as $\Omega(\sqrt{d^2K})$ (Zanette et al., 2020b)—while we do not match this in general, our d dependence does match the d-dependence of the best-known computationally efficient algorithm (Jin et al., 2020b).

Extension to Linear Mixture MDPs. While we have focused on the linear MDP setting in this work, we believe our techniques and use of the Catoni estimator could be easily extended to obtain first-order regret bounds in the linear mixture MDP setting. As noted, while Zhou et al. (2020) achieves nearly minimax optimal regret, their techniques do not easily generalize to obtain a first-order regret bound. We leave extending our method to linear mixture MDPs to future work.

Handling Unknown and Linear Rewards. We have assumed that the reward function is known, but that it may be nonlinear. If we are willing to make the additional assumption that the reward is linear, $r_h(s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h \rangle$ for some $\boldsymbol{\theta}_h$, we can handle unknown reward by modifying the Catoni estimator on Line 12 to use the data

$$X_\tau = \boldsymbol{v}^\top \boldsymbol{\phi}_{h,t}(r_{h,\tau} + V^k_{h+1}(s_{h+1,\tau}))/\bar{\mathsf{v}}^2_{h,\tau}, \quad \tau = 1, \dots, k-1,$$

and adding $4r^2_{h,k-1}$ to $\bar{\mathsf{v}}^2_{h,k-1}$. With this small modification, FORCE is able to handle unknown rewards and achieves the same regret as given in Theorem 4.

4.2 Computationally Efficient Implementation

As noted, FORCE is not computationally efficient because it is not clear how to efficiently solve the optimization on Line 13[2]. In this section, we provide a computationally efficient alternative, which only suffers slightly worse regret.

To obtain a computationally efficient variant of FORCE, we propose replacing Lines 12 and 13 with the following update:

$$\begin{aligned} &\widehat{\mathbb{E}}_h[V^k_{h+1}](\boldsymbol{u}_i) \leftarrow \mathsf{cat}_{h,k}[(k-1)\boldsymbol{\Lambda}^{-1}_{h,k-1}\boldsymbol{u}_i] \text{ for } i = 1, \dots, d \\ &\widehat{\boldsymbol{w}}^k_h \leftarrow [\boldsymbol{u}_1, \dots, \boldsymbol{u}_d] \cdot [\widehat{\mathbb{E}}_h[V^k_{h+1}](\boldsymbol{u}_1), \dots \widehat{\mathbb{E}}_h[V^k_{h+1}](\boldsymbol{u}_d)]^\top \end{aligned} \tag{4.5}$$

where $\boldsymbol{U} = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_d]$ denotes the eigenvectors of $\boldsymbol{\Lambda}_{h,k-1}$. This update *is* computationally efficient, as it involves only an eigendecomposition, the computation of d Catoni estimates (which can be computed efficiently), and a matrix-vector multiplication. The above approach satisfies the following guarantee:

Theorem 5 (Computationally Efficient Regret Bound)**.** *Consider the variant of* FORCE *with Lines 12 and 13 replaced by (4.5), and the update to $Q^k_h(s,a)$ on Line 14 replaced with*

$$Q^k_h(\cdot,\cdot) = \min\{r_h(\cdot,\cdot) + \langle \boldsymbol{\phi}(\cdot,\cdot), \widehat{\boldsymbol{w}}^k_h \rangle + 3(\sqrt{d}+2)\beta \|\boldsymbol{\phi}(\cdot,\cdot)\|_{\boldsymbol{\Lambda}^{-1}_{h,k-1}} + 3(\sqrt{d}+2)^2 \mathsf{v}_{\min}\beta^2/k^2, H\}.$$

Then with probability at least $1-3\delta$, the regret is at most

$$\mathcal{R}_K \le c_1 \sqrt{d^4H^3V^\star_1 K \cdot \log^3(HK/\delta)} + c_2 d^4 H^3 \log^{7/2}(HK/\delta)$$

and computation scales polynomially in d, H, K, and $\min\{|\mathcal{A}|, \mathcal{O}(2^d)\}$.

If we are willing to pay an additional factor of $\sqrt{d}$, it follows that we can run FORCE in a computationally efficient manner, assuming $|\mathcal{A}|$ is small. The dependence on $|\mathcal{A}|$ seems unavoidable and will be suffered by Jin et al. (2020b) as well, since computing the best action to play, $a_{h,k}$, on Line 17 will require enumerating all possible choices of a. If $|\mathcal{A}|$ is infinite, we can reduce this to only $2^{\mathcal{O}(d)}$ by covering all possible directions of $\boldsymbol{\phi}(s_{h,k}, a)$, but it is not clear if this can be reduced further in general.

[2]Note that one could also solve Line 13 by approximating the sup over $\mathcal{B}^d\backslash\{\mathbf{0}\}$ to a max over a sufficiently-fine ϵ-net of $\mathcal{B}^d\backslash\{\mathbf{0}\}$. Using standard covering estimates, this would require an exponentially-large-in-d cover of the ball, and thus require computing $2^{\Omega(d)}$ Catoni estimates.

5 Catoni Estimation in General Regression Settings

In this section we develop a set of results that extend the standard Catoni estimator to general martingale and heteroscedastic regression settings. The results presented here are critical to obtaining the first-order regret scaling of FORCE. We remark that the results in this section are based on a martingale version of the Catoni estimator first proposed in Wei et al. (2020).

5.1 Martingale Catoni Estimation

We begin by formalizing a martingale-linear regression setting in which our bounds (without function approximation) apply. The setting is reminiscent of that considered in Abbasi-Yadkori et al. (2011), but with two key generalizations: (a) the targets y_t can be *heavy-tailed*, we only require they have finite-variance, and (b) for each target y_t we have an associated upper bound σ_t^2 on its *conditional* square expectation. This latter point is crucial for modeling heteroscedastic noise.

Definition 5.1 (Heteroscedastic Heavy-Tailed Martingale Linear Regression)**.** Let $(\mathcal{F}_t)_{t\geq 0}$ denote a filtration, let $\boldsymbol{\phi}_t \in \mathbb{R}^d$ be a sequence of random $\mathcal{F}_{t-1}$-measurable vectors, and $y_t \in \mathbb{R}$ be $\mathcal{F}_t$-measurable random scalars satisfying

$$y_t = \langle \boldsymbol{\phi}_t, \boldsymbol{\theta}_\star \rangle + \eta_t, \quad \mathbb{E}[y_t|\mathcal{F}_{t-1}] = \langle \boldsymbol{\phi}_t, \boldsymbol{\theta}_\star \rangle$$

for some $\boldsymbol{\theta}_\star \in \mathbb{R}^d$ and η_t satisfying $\mathbb{E}[\eta_t|\mathcal{F}_{t-1}] = 0$, and $\mathbb{E}[\eta_t^2|\mathcal{F}_{t-1}] < \infty$, but otherwise arbitrary (as such, the distribution of η_t may depend on $\boldsymbol{\phi}_t$). Furthermore, let σ_t^2 be a $\mathcal{F}_{t-1}$-measurable sequence of scalars satisfying $\sigma_t^2 \geq \mathbb{E}[y_t^2|\mathcal{F}_{t-1}]$, and let

$$\boldsymbol{\Sigma}_T := \sum_{t=1}^{T} \sigma_t^{-2} \boldsymbol{\phi}_t \boldsymbol{\phi}_t^\top.$$

In the regression setting of Definition 5.1, our goal is to estimate $\boldsymbol{\theta}_\star$ in a particular direction, $\boldsymbol{v} \in \mathbb{R}^d$, given observations $\{(\boldsymbol{\phi}_t, y_t, \sigma_t)\}_{t=1}^T$. As a warmup, the following lemma bounds certain *directional* Catoni estimates.

Lemma 5.1 (Heteroscedastic Catoni Estimator)**.** *Assume we are in the regression setting of Definition 5.1. For a fixed vector $\boldsymbol{v} \in \mathbb{R}^d$ let $\mathsf{cat}[\boldsymbol{v}]$ denote the Catoni estimate applied to $(X_t)_{t=1}^T$ where $X_t := \boldsymbol{v}^\top \boldsymbol{\phi}_t y_t / \sigma_t^2$, with a fixed (deterministic) parameter $\alpha > 0$. Then, for any failure probability $\delta > 0$ and fixed $\alpha_{\max} > 0$, if our deterministic α can be written as*

$$\alpha = \min\left\{ \gamma \cdot \frac{\sqrt{\log(2/\delta)}}{\|\boldsymbol{v}\|_{\boldsymbol{\Sigma}_T}}, \alpha_{\max} \right\} \tag{5.1}$$

for some (possibly random) $\gamma \geq 1$, then

$$\left| \mathsf{cat}[\boldsymbol{v}] - \frac{1}{T} \boldsymbol{v}^\top \boldsymbol{\Sigma}_T \cdot \boldsymbol{\theta}_\star \right| \leq (2 + 2\gamma) \|\boldsymbol{v}\|_{\boldsymbol{\Sigma}_T} \sqrt{\frac{\log \frac{2}{\delta}}{T^2}} + \frac{2 \log \frac{2}{\delta}}{\alpha_{\max} T}$$

provided that $T \geq (2 + 2\gamma^2) \log \frac{2}{\delta}$.

Note that the introduction of the slack parameter $\gamma \geq 1$ accounts for the fact that α is assumed to be chosen deterministically, while $\boldsymbol{\Sigma}_T$ is random. Lemma 5.1 shows that we can apply the Catoni

estimator to estimate $\boldsymbol{\theta}_\star$ in a particular direction, with estimation error scaling only with an upper bound on $\mathbb{E}[\eta_t^2|\mathcal{F}_{t-1}]$ and independent of other properties of η_t, such as its magnitude. This is in contrast to Bernstein-style bounds which exhibit lower-order terms scaling with the absolute magnitude of η_t. Lemma 5.1 serves as a building block for our subsequent estimation bounds, where the choice of finite $\alpha_{\max}$ serves a useful technical purpose.

5.2 Self-Normalized Catoni Inequality

Next, we bootstrap Lemma 5.1 into a full-fledged self-normalized inequality for heteroscedastic noise. To do so, we need to address two technical points:

- The ideal choice of α (for which γ is close to 1) is not deterministic, but data-dependent.
- To estimate $\boldsymbol{\theta}_\star$ in direction $\boldsymbol{v}$, we would like to consider $\mathsf{cat}[\widetilde{\boldsymbol{v}}]$, where $\widetilde{\boldsymbol{v}} = T\boldsymbol{\Sigma}_T^{-1}\boldsymbol{v}$, since then $\frac{1}{T}\widetilde{\boldsymbol{v}}^\top \boldsymbol{\Sigma}_T \cdot \boldsymbol{\theta}_\star = \boldsymbol{v}^\top \boldsymbol{\theta}_\star$. However, this choice of $\boldsymbol{v}$ introduces correlations between $\boldsymbol{v}$ and our observations $\{(\boldsymbol{\phi}_t, y_t, \sigma_t)\}_{t=1}^T$, which prevents us from applying Lemma 5.1 directly.

We adress both via a uniform-convergence-style argument and argue that a bound of the form given in Lemma 5.1 holds for *all* $\boldsymbol{v}$ simultaneously. This requires a subtle argument to bound the sensitivity of the Catoni estimator, given in Appendix A.5. With this bound in hand, we establish the following *truly heteroscedastic self-normalized concentration inequality*, the formal statement of Proposition 2 in the introduction.

Corollary 1 (Self-Normalized Heteroscedastic Catoni Estimation)**.** *Consider the setting of Definition 5.1, and suppose that with probability* 1, $|\eta_t| \le \beta_\eta < \infty$ *and* $\sigma_t^2 \ge \sigma_{\min}^2 > 0$ *for all* t*. For a fixed regularization parameter* $\lambda > 0$*, define the effective dimension*

$$d_T := c \cdot d \cdot \mathsf{logs}\left(T, \alpha_{\max}^2, \lambda^{-1}, \sigma_{\min}^{-2}, \beta_\eta\right).$$

Let $\mathsf{cat}\left[T\boldsymbol{\Lambda}_T^{-1}\boldsymbol{v}\right]$ *denote the Catoni estimate applied to* $(X_t)_{t=1}^T$ *and parameter* α *given by*

$$X_t = T\boldsymbol{v}^\top \boldsymbol{\Lambda}_T^{-1}\boldsymbol{\phi}_t y_t/\sigma_t^2, \quad \alpha = \min\left\{\sqrt{\|T\boldsymbol{\Lambda}_T^{-1}\boldsymbol{v}\|_{\boldsymbol{\Sigma}_T}^{-2} \cdot (d_T + \log 1/\delta)},\, \alpha_{\max}\right\}$$

and for $\boldsymbol{\Lambda}_T = \lambda I + \sum_{t=1}^T \sigma_t^{-2}\boldsymbol{\phi}_t\boldsymbol{\phi}_t^\top$*. Then, as long as* $T \ge 5(\log 1/\delta + d_T)$*, with probability at least* $1-\delta$*, for all* $\boldsymbol{v} \in \mathcal{B}^d$ *simultaneously,*

$$\left|\mathsf{cat}\left[T\boldsymbol{\Lambda}_T^{-1}\boldsymbol{v}\right] - \boldsymbol{v}^\top\boldsymbol{\theta}_\star\right| \le 5\|\boldsymbol{v}\|_{\boldsymbol{\Lambda}_T^{-1}} \cdot \left(\sqrt{\log 1/\delta + d_T} + \sqrt{\lambda}\|\boldsymbol{\theta}_\star\|_2\right) + \frac{3(\log\frac{1}{\delta} + d_T)}{\alpha_{\max} T}. \tag{5.2}$$

In contrast to Lemma A.5, Corollary 1 only adds the requirement that η_t and $\sigma_{\min}^2$ satisfy probability-one upper and lower bounds, respectively, which enter only logarithmically into our final bound[3]. Similarly, the parameter $\alpha_{\max}$ also enters at most logarithmically into the final bound, and hence can also be chosen suitably large to make the second term in Equation (5.2) suitably small. Intuitively, $\alpha_{\max}$ ensures that the Catoni estimator is sufficiently robust to perturbation, which is necessary for our uniform convergence arguments.

[3]In the case when the noise is unbounded, note that, by Chebyshev's inequality, one can just take $\beta_u \le \sqrt{\max_t \sigma_t^2/\delta}$, at the expense of at most $\delta > 0$ failure probability, whilst maintaining a logarithmic dependence on $1/\delta$ in the final bound.

Corollary 1 is a special case of a more general result, Theorem 6, whose statement and proof we detail in the following subsection. Up to logarithmic factors our guarantee matches that of Abbasi-Yadkori et al. (2011). The key difference is that, whereas Abbasi-Yadkori et al. (2011) considers the a norm in a covariance *not weighted by the variance* $\|\boldsymbol{v}\|_{\widetilde{\boldsymbol{\Lambda}}_T^{-1}}$ with $\widetilde{\boldsymbol{\Lambda}}_T := \lambda I + \sum_{t=1}^T \boldsymbol{\phi}_t \boldsymbol{\phi}_t^\top$, our guarantee uses the weighted-covariance norm $\boldsymbol{\Lambda}_T := \lambda I + \sum_{t=1}^T \sigma_t^{-2} \boldsymbol{\phi}_t \boldsymbol{\phi}_t^\top$. It is clear that the latter is much *larger* when σ_t^2 are small, leading to a smaller error bound.

Our bound is similar in spirit to another self-normalized heteroscedastic inequality recently provided by Zhou et al. (2020). The key distinction is that the Catoni estimator lets us obtain estimates that scale with the standard deviation of the noise, σ, and only logarithmically with the absolute magnitude, β_η. This is in contrast to the bound obtained in Zhou et al. (2020), which scale only with σ in the leading order term, but scales with β_η in the lower order term. In situations where β_η is large, which will be the case when deriving first-order bounds for linear RL, this scaling could be significantly worse. To make this concrete, the following example illustrates Corollary 1 on a simple problem.

Example 5.1 (Regression with Bounded Noise)**.** Consider the linear regression setting where we receive observations

$$y_t = \langle \boldsymbol{\phi}_t, \boldsymbol{\theta}_\star \rangle + \eta_t$$

for some $\mathcal{F}_{t-1}$-measurable $\boldsymbol{\phi}_t$ and noise η_t satisfying $\mathbb{E}[\eta_t | \mathcal{F}_{t-1}] = 0$, $\mathrm{Var}[\eta_t \mid \mathcal{F}_{t-1}] = \sigma^2$, and $|\eta_t| \le \beta_\eta$ almost surely for some β_η. Assume σ^2 is known and that $|\langle \boldsymbol{\theta}_\star, \boldsymbol{\phi}_t \rangle| \le \epsilon$ for all t. Define $\sigma_t^2 = 2(\epsilon^2 + \sigma^2)$ for all t and note that

$$\sigma_t^2 = 2(\epsilon^2 + \sigma^2) \ge 2(\langle \boldsymbol{\theta}_\star, \boldsymbol{\phi}_t \rangle^2 + \mathbb{E}[\eta_t^2 \mid \mathcal{F}_{t-1}]) = 2\mathbb{E}[y_t^2 \mid \mathcal{F}_{t-1}].$$

Now take some $\boldsymbol{v} \in \mathbb{R}^d$ and consider applying the Catoni estimator to the data

$$X_t = T \boldsymbol{v}^\top \boldsymbol{\Lambda}_T^{-1} \boldsymbol{\phi}_t y_t / (2\epsilon^2 + 2\sigma^2), \quad \boldsymbol{\Lambda}_T = \frac{1}{2\epsilon^2 + 2\sigma^2} \left(I + \sum_{t=1}^T \boldsymbol{\phi}_t \boldsymbol{\phi}_t^\top \right)$$

and with α set as in Corollary 1. We can then apply Corollary 1 to get that, with probability $1 - \delta$,

$$\left| \mathsf{cat}\left[T \boldsymbol{\Lambda}_T^{-1} \boldsymbol{v} \right] - \boldsymbol{v}^\top \boldsymbol{\theta}_\star \right| \lesssim \|\boldsymbol{v}\|_{\boldsymbol{\Lambda}_T^{-1}} \cdot \sqrt{\log 1/\delta + d_T} + \frac{\log 1/\delta + d_T}{\alpha_{\max} T}.$$

Note that, given our setting of $\boldsymbol{\Lambda}_T$, we have

$$\|\boldsymbol{v}\|_{\boldsymbol{\Lambda}_T^{-1}} = \sqrt{2\epsilon^2 + 2\sigma^2} \|\boldsymbol{v}\|_{(I + \boldsymbol{\Sigma}_T)^{-1}}$$

and we can set $\alpha_{\max} = T$, $\sigma_{\min} = 1/T$, so $d_T = \mathcal{O}(d \cdot \mathsf{logs}(T, \sigma^2, \beta_\eta))$. We conclude that

$$\begin{aligned} \left| \mathsf{cat}\left[T \boldsymbol{\Lambda}_T^{-1} \boldsymbol{v} \right] - \boldsymbol{v}^\top \boldsymbol{\theta}_\star \right| &\lesssim \|\boldsymbol{v}\|_{(I + \boldsymbol{\Sigma}_T)^{-1}} \cdot (\epsilon + \sigma) \sqrt{\log 1/\delta + d \cdot \mathsf{logs}(T, \sigma^2, \beta_\eta)} + \frac{\log 1/\delta + d \cdot \mathsf{logs}(T, \sigma^2, \beta_\eta)}{T^2} \\ &= \widetilde{\mathcal{O}}\left(\|\boldsymbol{v}\|_{(I + \boldsymbol{\Sigma}_T)^{-1}} \cdot (\epsilon + \sigma) \sqrt{d} \right). \end{aligned}$$

By Corollary 1 this holds for all $\boldsymbol{v}$ simultaneously.

In contrast to this, using the same regularization as above, the Bernstein self-normalized bound of Zhou et al. (2020) will scale as (hiding logarithmic terms),

$$\left| \boldsymbol{v}^\top (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_\star) \right| \le \widetilde{\mathcal{O}}\left(\|\boldsymbol{v}\|_{(I + \boldsymbol{\Sigma}_T)^{-1}} \cdot \left(\sigma \sqrt{d} + \beta_\eta \right) \right)$$

where $\widehat{\boldsymbol{\theta}}$ denotes the least-squares estimate.

Example 5.1 could model, for example, a linear bandit problem where the value of the optimal arm is 0 (which is always achievable by shifting the problem), and we are in the regime where we are playing near-optimally, so that $\langle \boldsymbol{\theta}_\star, \boldsymbol{\phi}_t \rangle \approx 0$. In this regime, $\epsilon \approx 0$, so the dominant scaling will simply be $\widetilde{\mathcal{O}}(\|\boldsymbol{v}\|_{(I+\boldsymbol{\Sigma}_T)^{-1}} \cdot \sigma\sqrt{d})$.

5.3 Self-Normalized Catoni Estimation with Function Approximation

To apply our self-normalized bound in the linear RL setting, we need to allow for regression targets which are potentially *correlated* with the features $\boldsymbol{\phi}_t$ in a verify specific way. More precisely, the targets y_t take the form $y_t = \langle \boldsymbol{u}_\star, \boldsymbol{\phi}_t \rangle + f_\star(\boldsymbol{\phi}'_t)$ where $\boldsymbol{\phi}'_t$ is a $\mathcal{F}_t$-measurable feature vector, and $f_\star$ is a function which may *depend on all the data* $\{(\boldsymbol{\phi}_t, y_t, \sigma_t)\}_{t=1}^T$. The function $f_\star$ is therefore *not* $\mathcal{F}_t$-measurable, and so y_t does not satisfy the condition of Definition 5.1. To handle these challenges, we introduce the following regression setting, which specifies the precise conditions needed for our most general result.

Definition 5.2 (Heteroscedastic Regression with Function Approximation)**.** Given dimension parameters $d, d', p \in \mathbb{N}$, scaling parameters $H, \beta_u, \beta_\mu > 0$, and minimal varaince $\sigma^2_{\min}$, the heteroscedastic regression with function approximation setting is defined as follows. Let $(\mathcal{F}_t)_{t\geq 0}$ be a filtration, and consider a sequence of random vectors $(\boldsymbol{\phi}_t, \boldsymbol{\phi}'_t)_{t=1}^T$ and random scalar outuits $(y_t)_{t=1}^T$ and noises $(\eta_t)_{t=1}^T$ and variance bounds $(\sigma_t^2)_{t=1}^T$ such that

- $\boldsymbol{\phi}_t \in \mathbb{R}^d$ is $\mathcal{F}_{t-1}$-measurable, $\boldsymbol{\phi}'_t \in \mathbb{R}^{d'}$ is $\mathcal{F}_t$ measurable, and $\|\boldsymbol{\phi}_t\|_2, \|\boldsymbol{\phi}'_t\|_2 \leq 1$.
- There exists a signed measure $\boldsymbol{\mu}$ over $\mathcal{B}^{d'}$ with total mass $\||\boldsymbol{\mu}|(\mathcal{B}^{d'})\|_2 \leq \beta_\mu$ such that, for all t, the conditional distribution of $\boldsymbol{\phi}'_t$ given $\mathcal{F}_{t-1}$ ensures that, for all bounded functions f,

$$\mathbb{E}\left[f(\boldsymbol{\phi}'_t) \mid \mathcal{F}_{t-1}\right] = \langle \boldsymbol{\phi}_t, \int f(\boldsymbol{\phi}') \mathrm{d}\boldsymbol{\mu}(\boldsymbol{\phi}') \rangle. \tag{5.3}$$

- $|\eta_t| \leq \beta_\eta$ with probability 1, and $\mathbb{E}[\eta_t \mid \mathcal{F}_{t-1}] = 0$.
- There exist a parameter $\boldsymbol{u}_\star \in \mathbb{R}^d$ with $\|\boldsymbol{u}_\star\|_2 \leq \beta_u$, a function class $\mathscr{F}$ of functions $f : \mathbb{R}^{d'} \to [-H, H]$, and a function $f_\star \in \mathscr{F}$ which may be random and dependent on $(\boldsymbol{\phi}_t, \boldsymbol{\phi}'_t)_{t=1}^T$ such that, for all t, $y_t = \langle \boldsymbol{u}_\star, \boldsymbol{\phi}_t \rangle + f_\star(\boldsymbol{\phi}'_t) + \eta_t$. Thus,

$$\mathbb{E}[y_t | \mathcal{F}_{t-1}] = \langle \boldsymbol{\phi}_t, \boldsymbol{\theta}_\star \rangle \quad \text{for} \quad \boldsymbol{\theta}_\star = \boldsymbol{u}_\star + \int f_\star(\boldsymbol{\phi}') \mathrm{d}\boldsymbol{\mu}(\boldsymbol{\phi}').$$

- σ_t are uniformly lower bounded by $\sigma_{\min}$, finite, $\mathcal{F}_{t-1}$ measurable, and satisfy

$$\mathbb{E}\left[\left(\langle \boldsymbol{\phi}_t, \boldsymbol{u}_\star \rangle + f_\star(\boldsymbol{\phi}'_t) + \eta_t\right)^2 \mid \mathcal{F}_{t-1}\right] \leq \frac{1}{2}\sigma_t^2. \tag{5.4}$$

- The covering numbers of $\mathscr{F}$ are *parameteric*, in the sense that there exists a $p \in \mathbb{N}$ and $R > 0$ such that, for $\epsilon > 0$, the ϵ-covering number of $\mathscr{F}$ in the metric $\mathrm{dist}_\infty(f, f') := \sup_{\boldsymbol{\phi}' \in \mathcal{B}^{d'}} |f(\boldsymbol{\phi}') - f'(\boldsymbol{\phi}')|$ is bounded as $\mathsf{N}(\mathscr{F}, \mathrm{dist}_\infty, \epsilon) \leq p \log(1 + \frac{2R}{\epsilon})$, where $\mathsf{N}(\mathscr{F}, \mathrm{dist}_\infty, \epsilon)$ is the ϵ-covering number of $\mathscr{F}$ in the norm dist_∞.

Note that Definition 5.2 strictly generalizes Definition 5.1 since we can always choose $f_\star = 0$ to be a fixed function, and are left only with the noise η_t. For this most general setting, we attain the following result:

Theorem 6 (Heteroscedastic Catoni Estimation with Function Approximation)**.** *Assume that we are in the setting of Definition 5.2. Define*

$$d_T := c \cdot (p+d) \cdot \mathsf{logs}\left(T, \alpha_{\max}^2, \lambda^{-1}, \sigma_{\min}^{-2}, \beta_\mu, \beta_u, \beta_\eta, R, H\right).$$

Let $\mathsf{cat}\left[T\boldsymbol{\Lambda}_T^{-1}\boldsymbol{v}\right]$ *denote the Catoni estimate applied to* $(X_t)_{t=1}^T$ *and parameter* α *given by*

$$X_t = T\boldsymbol{v}^\top \boldsymbol{\Lambda}_T^{-1}\boldsymbol{\phi}_t y_t/\sigma_t^2, \quad \alpha = \min\left\{\sqrt{\|T\boldsymbol{\Lambda}_T^{-1}\boldsymbol{v}\|_{\boldsymbol{\Sigma}_T}^{-2} \cdot (d_T + \log 1/\delta)},\, \alpha_{\max}\right\}$$

and for $\boldsymbol{\Lambda}_T = \lambda I + \sum_{t=1}^T \sigma_t^{-2}\boldsymbol{\phi}_t\boldsymbol{\phi}_t^\top$*. Then, as long as* $T \geq 6(\log 1/\delta + d_T)$*, with probability at least* $1-\delta$*, for all* $\boldsymbol{v} \in \mathcal{B}^d$ *simultaneously,*

$$\left|\mathsf{cat}\left[\boldsymbol{\Lambda}_T^{-1}\boldsymbol{v}\right] - \boldsymbol{v}^\top\boldsymbol{\theta}_\star\right| \leq 5\|\boldsymbol{v}\|_{\boldsymbol{\Lambda}_T^{-1}} \cdot \left(\sqrt{\log 1/\delta + d_T} + \sqrt{\lambda}\|\boldsymbol{\theta}_\star\|_2\right) + \frac{3(\log\frac{1}{\delta} + d_T)}{\alpha_{\max} T}. \tag{5.5}$$

A couple remarks are in order. First, Corollary 1 is just the special case obtained by setting $f_\star(\cdot) = 0$ to be the zero function, and sole element of $\mathscr{F} = \{f_\star\}$. Second, as will be observed, the assumptions in Definition 5.2 precisely line up with those required for linear RL. The proof of Theorem 6, detailed in Appendix A.3, follows by applying Lemma 5.1 and carefully union bounding over the parameter space. It invokes a novel perturbation analysis of the Catoni estimator, given in Appendix A.5, which may be of independent interest. Again, we remark $\alpha_{\max}$ can be chosen suitably large that estimation error of the Catoni estimator scales primarily as $\|\boldsymbol{v}\|_{\boldsymbol{\Lambda}_T^{-1}} \cdot \sqrt{\log(1/\delta) + d_T}$.

5.4 Linear Approximation to the Catoni Estimator

In the linear RL setting, we will rely on the Catoni estimator to form an optimistic estimate, $Q_h^k(s,a)$, of $Q_h^\star(s,a)$. To construct this estimator, we will set $y_t = V_{h+1}^k(s_{h+1,t})$—thus, $f_\star$ will itself be an optimistic Q-value estimate. In order to apply Theorem 6 directly to the linear RL setting, we therefore need to cover the space of *all* Catoni estimates. It is not clear how to do this in general without covering all $\mathcal{O}(dT)$ parameters the Catoni estimator takes as input, which will result in suboptimal K dependence in the final regret bound.

To overcome this challenge, we make the critical observation that (5.5) implies that, up to some tolerance, *there exists a linear function which approximates* $\mathsf{cat}[T\boldsymbol{\Lambda}_T^{-1}\boldsymbol{v}]$ *for all* $\boldsymbol{v}$, namely $\langle \boldsymbol{v}, \boldsymbol{\theta}_\star\rangle$. As we do not know $\boldsymbol{\theta}_\star$, we cannot compute this function directly. However, the following result shows that we can exploit the fact that there *exists* such a linear approximation in order to come up with our own linear approximation:

Lemma 5.2. *Let* $\mathsf{cat}[\boldsymbol{\Lambda}^{-1}\boldsymbol{v}]$ *denote a Catoni estimate, as defined in Lemma 5.1. Assume that, for all* $\boldsymbol{v} \in \mathcal{V}$ *for some* $\mathcal{V} \subseteq \mathbb{R}^d$*,* $\mathbf{0} \notin \mathcal{V}$*, we have*

$$|\mathsf{cat}[\boldsymbol{\Lambda}^{-1}\boldsymbol{v}] - \langle \boldsymbol{v}, \boldsymbol{\theta}_\star\rangle| \leq C_1\|\boldsymbol{v}\|_{\boldsymbol{\Lambda}^{-1}} + C_2/T \tag{5.6}$$

for some C_1, C_2*. Set*

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sup_{\boldsymbol{v}\in\mathcal{V}} \frac{|\langle \boldsymbol{\theta}, \boldsymbol{v}\rangle - \mathsf{cat}[\boldsymbol{\Lambda}^{-1}\boldsymbol{v}]|}{\|\boldsymbol{v}\|_{\boldsymbol{\Lambda}^{-1}}}. \tag{5.7}$$

Then, for all $\boldsymbol{v} \in \mathcal{V}$*, we have*

$$|\langle \widehat{\boldsymbol{\theta}}, \boldsymbol{v}\rangle - \mathsf{cat}[\boldsymbol{\Lambda}^{-1}\boldsymbol{v}]| \leq C_1\|\boldsymbol{v}\|_{\boldsymbol{\Lambda}^{-1}} + C_2/T, \qquad |\langle \widehat{\boldsymbol{\theta}}, \boldsymbol{v}\rangle - \langle \boldsymbol{v}, \boldsymbol{\theta}_\star\rangle| \leq 2C_1\|\boldsymbol{v}\|_{\boldsymbol{\Lambda}^{-1}} + 2C_2/T.$$

Given this result, if we approximate our Q-functions by Catoni estimates, $\widehat{\mathbb{E}}_h[V_{h+1}^k](s,a)$, instead of directly using $\widehat{\mathbb{E}}_h[V_{h+1}^k](s,a)$ we can rely on a linear approximation to it, $\langle \widehat{\boldsymbol{w}}_h^k, \boldsymbol{\phi}(s,a)\rangle$. By Lemma 5.2, this will be an accurate approximation for *all* s,a. As we can easily cover the space of d-dimensional vectors, this allows us to cover the space of all of our Q-function estimates. As we will see, in practice we rely on optimistic Q functions which also depend on some $\boldsymbol{\Lambda} \succeq 0$, so we will ultimately choose $\mathscr{F}$ in Definition 5.2 so that $p = \mathcal{O}(d^2)$.

Note that solving (5.7) is not computationally efficient in general, yet as we described in Section 4.2 and show in more detail in Appendix A.4, a linear approximation to a Catoni estimator can be found in a computationally efficient manner if we are willing to pay an extra factor of $\sqrt{d}$ in the approximation error.

6 Regret Bound Proof Sketch

We turn now to applying the Catoni estimation results of Section 5 in the setting of linear RL. We defer the full proofs to Appendix B.

6.1 Failure of Least Squares Estimation

We first describe in more detail why least squares estimation is insufficient to obtain first-order regret. Building on Jin et al. (2020b), our goal in the RL setting will be to construct optimistic estimators, $Q_h^k(s,a)$, to the optimal value function, $Q_h^\star(s,a)$, satisfying $Q_h^k(s,a) \geq Q_h^\star(s,a)$. Jin et al. (2020b) construct such estimators recursively by applying a least-squares value iteration update and solving

$$\widetilde{\boldsymbol{w}}_h^k = \operatorname*{arg\,min}_{\boldsymbol{w}\in\mathbb{R}^d} \sum_{\tau=1}^{k-1} \left(V_{h+1}^k(s_{h+1,\tau}) - \boldsymbol{w}^\top \boldsymbol{\phi}_{h,\tau}\right)^2 + \lambda \|\boldsymbol{w}\|_2^2.$$

Intuitively, if enough data has been collected, this update will produce a $\widetilde{\boldsymbol{w}}_h^k$ which accurately approximates the expectation over the next state. Indeed, Jin et al. (2020b) show that, for any π,[4]

$$\langle \widetilde{\boldsymbol{w}}_h^k, \boldsymbol{\phi}(s,a)\rangle + r_h(s,a) - Q_h^\pi(s,a) = \mathbb{E}_h[V_{h+1}^k - V_{h+1}^\pi](s,a) + \xi_h(s,a)$$

for some $\xi_h(s,a) \lesssim dH\|\boldsymbol{\phi}(s,a)\|_{\widetilde{\boldsymbol{\Lambda}}_{h,k-1}^{-1}}$, where $\widetilde{\boldsymbol{\Lambda}}_{h,k-1} = \lambda I + \sum_{\tau=1}^{k-1} \boldsymbol{\phi}_{h,\tau}\boldsymbol{\phi}_{h,\tau}^\top$. Applying this estimator, Jin et al. (2020b) are able to construct a value function guaranteed to be optimistic, and ultimately obtains regret of $\widetilde{\mathcal{O}}(\sqrt{d^3H^4K})$. This is fundamentally a Hoeffding-style estimator, however, and does not scale with the variance of the next-state value function. As such, it does not appear that tighter regret bounds can be obtained using this approach.

A natural modification of this estimator would be the *weighted least squares estimate*:

$$\widetilde{\boldsymbol{w}}_h^k = \operatorname*{arg\,min}_{\boldsymbol{w}\in\mathbb{R}^d} \sum_{\tau=1}^{k-1} \left(V_{h+1}^k(s_{h+1,\tau}) - \boldsymbol{w}^\top \boldsymbol{\phi}_{h,\tau}\right)^2 / \widehat{\sigma}_{h,\tau}^2 + \lambda \|\boldsymbol{w}\|_2^2 \tag{6.1}$$

[4] In fact, Jin et al. (2020b) uses a slightly different update, including $r_{h,\tau}$ in the regression problem so that $\langle \widetilde{\boldsymbol{w}}_h^k, \phi(s,a)\rangle$ estimates the reward and next-state expectation. In contrast, the setting of $\widetilde{\boldsymbol{w}}_h^k$ given here estimates only the next-state expectation. Jin et al. (2020b) assume that the reward is unknown and is linear, motivating their inclusion of it in the regression problem. The direct extension of their approach to known but nonlinear reward is the update stated above.

for $\widehat{\sigma}^2_{h,\tau}$ an upper bound on $\mathrm{Var}_{s'\sim P_h(\cdot|s_{h,\tau},a_{h,\tau})}[V^k_{h+1}(s')]$. An approach similar to this is taken in the linear mixture MDP setting of Zhou et al. (2020), where it is shown that this approach does indeed yield variance-dependent bounds when a Bernstein-style self-normalized bound is applied. However, as noted, this Bernstein-style bound still scales with the *magnitude* of the "noise" in its lower-order term, which here will be of order $H/\sigma_{\min}$. Carrying their analysis through, we see that the leading order term of the regret is at least on order $(\sqrt{d} + \frac{H}{\sigma_{\min}})\sqrt{\sigma^2_{\min} HK} \geq H\sqrt{HK}$. Thus, while this approach may yield an improved d and H dependence, it is unable to obtain a first-order scaling of $\mathcal{O}(\sqrt{V_1^\star K})$ when $V_1^\star$ is small.

6.2 From Catoni Estimation to Optimism

Note that $\widetilde{\boldsymbol{w}}^k_h$ in (6.1) can be written as

$$\widetilde{\boldsymbol{w}}^k_h = \sum_{\tau=1}^{k-1} \boldsymbol{\Lambda}^{-1}_{h,k-1} \boldsymbol{\phi}_{h,\tau} V^k_{h+1}(s_{h+1,\tau})/\bar{\mathsf{v}}^2_{h,\tau}.$$

In other words, $\widetilde{\boldsymbol{w}}^k_h$ is simply the sample mean. This motivates applying the Catoni estimator to the problem. Indeed, consider setting $\widehat{\mathbb{E}}_h[V^k_{h+1}](s,a) = \mathsf{cat}_{h,k-1}[(k-1)\boldsymbol{\Lambda}^{-1}_{h,k-1}\boldsymbol{\phi}(s,a)]$. By Definition 3.1, we can set $\boldsymbol{\theta}_\star$ in Definition 5.2 as

$$\boldsymbol{\theta}_\star \leftarrow \int V^k_{h+1}(s')\mathrm{d}\boldsymbol{\mu}_h(s')$$

and will have that $\langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_\star \rangle = \mathbb{E}_h[V^k_{h+1}](s,a)$. Theorem 6 then immediately gives that, for all s, a, h, k,

$$|\widehat{\mathbb{E}}_h[V^k_{h+1}](s,a) - \mathbb{E}_h[V^k_{h+1}](s,a)| \lesssim (1 + H\sqrt{\lambda})\beta \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}^{-1}_{h,k-1}} + \mathsf{v}_{\min}\beta^2/k^2 \tag{6.2}$$

where here $\beta = 6\sqrt{\log 1/\delta + d_T}$ for $d_T = \widetilde{\mathcal{O}}(d + p_{\mathrm{mdp}})$ and p_{mdp} the covering number of the set of functions $V^k_{h+1}(\cdot)$. Recall that we chose $\alpha_{\max} = K/\mathsf{v}_{\min}$ in the linear RL setting. Thus, the lower-order term of $3\beta^2/(\alpha_{\max}k)$ of Theorem 6 can be upper bounded as $3\mathsf{v}_{\min}\beta^2/k^2$, as in (6.2).

Given this $\widehat{\mathbb{E}}_h[V^k_{h+1}](s,a)$, let $\widehat{\boldsymbol{w}}^k_h$ denote the linear approximation to $\widehat{\mathbb{E}}_h[V^k_{h+1}](s,a)$, as described in Lemma 5.2. By Lemma 5.2, it follows that for all s, a,

$$|\langle \widehat{\boldsymbol{w}}^k_h, \boldsymbol{\phi}(s,a)\rangle - \mathbb{E}_h[V^k_{h+1}](s,a)| \lesssim (1 + H\sqrt{\lambda})\beta \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}^{-1}_{h,k-1}} + \mathsf{v}_{\min}\beta^2/k^2. \tag{6.3}$$

Constructing Optimistic Estimators. Fix some h and k and assume that (6.3) holds for all s, a. Let

$$Q^k_h(s,a) = \min\left\{ r_h(s,a) + \langle \boldsymbol{\phi}(s,a), \widehat{\boldsymbol{w}}^k_h\rangle + 3(1 + H\sqrt{\lambda})\beta\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}^{-1}_{h,k-1}} + 3\mathsf{v}_{\min}\beta^2/k^2, H \right\}.$$

Assume that $V^k_{h+1}(s)$ is optimistic, that is, $V^k_{h+1}(s) \geq V^\star_{h+1}(s)$ for all s. Then (6.3) and this assumption imply that

$$\begin{aligned} Q^k_h(s,a) &= \min\left\{ r_h(s,a) + \langle \boldsymbol{\phi}(s,a), \widehat{\boldsymbol{w}}^k_h\rangle + 3(1 + H\sqrt{\lambda})\beta\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}^{-1}_{h,k-1}} + 3\mathsf{v}_{\min}\beta^2/k^2, H \right\} \\ &\geq \min\left\{ r_h(s,a) + \mathbb{E}_h[V^k_{h+1}](s,a) + 2(1 + H\sqrt{\lambda})\beta\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}^{-1}_{h,k-1}} + 2\mathsf{v}_{\min}\beta^2/k^2, H \right\} \end{aligned}$$

$$
\begin{aligned}
&\geq \min\left\{r_h(s,a) + \mathbb{E}_h[V^\star_{h+1}](s,a) + 2(1+H\sqrt{\lambda})\beta\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}^{-1}_{h,k-1}} + 2\mathsf{v}_{\min}\beta^2/k^2, H\right\} \\
&\geq \min\left\{r_h(s,a) + \mathbb{E}_h[V^\star_{h+1}](s,a), H\right\} \\
&= Q^\star_h(s,a).
\end{aligned}
$$

In other words, given that $\widehat{\boldsymbol{w}}^k_h$ accurately approximates the next state expectation, (6.3), and that $V^k_{h+1}(s)$ is optimistic, it immediately follows that $Q^k_h(s,a)$ is also optimistic.

Defining the Function Class. It remains to determine the value of p_{mdp}. Applying the above argument inductively, we see that to form an optimistic estimate, it suffices to consider functions in the set

$$
\mathscr{F}_{\text{mdp}} = \left\{f(\cdot) = \min\{\langle\cdot, \boldsymbol{w}\rangle + \bar{\beta}\|\cdot\|_{\boldsymbol{\Lambda}^{-1}} + \bar{c}, H\} \;:\; \|\boldsymbol{w}\|_2 \leq \beta_{\boldsymbol{w}},\ \boldsymbol{\Lambda} \succeq \lambda I\right\}.
$$

for some $\bar{\beta}, \bar{c}$, and $\beta_{\boldsymbol{w}}$. $\mathscr{F}_{\text{mdp}}$ depends on two parameters—the d-dimensional $\boldsymbol{w}$ and $d\times d$ dimensional $\boldsymbol{\Lambda}$. Thus, using standard covering arguments, it's easy to see that $\mathsf{N}(\mathscr{F}_{\text{mdp}}, \text{dist}_\infty, \epsilon) = \mathcal{O}(d^2\log(1+1/\epsilon))$, so it suffices to take $p_{\text{mdp}} = \mathcal{O}(d^2)$. Given this and the definition of d_T, we see that in our setting we will have that $d_T = \widetilde{\mathcal{O}}(d^2)$, so $\beta = \widetilde{\mathcal{O}}(\sqrt{\log 1/\delta + d^2})$.

6.3 Proving the Regret Bound

Henceforth, we will assume that (6.2) holds for all s, a, h, and k. We turn now to showing how the above results can be used to prove a regret bound. The following lemma, which is a simple consequence of (6.2), will be useful in decomposing the regret.

Lemma 6.1 (Informal)**.** *Let* $\delta^k_h = V^k_h(s^k_h) - V^{\pi_k}_h(s^k_h)$ *and* $\zeta^k_{h+1} = \mathbb{E}_h[\delta^k_{h+1}](s_{h,k}, a_{h,k}) - \delta^k_{h+1}$*. Then, with high probability,*

$$
\delta^k_h \leq \delta^k_{h+1} + \zeta^k_{h+1} + \min\{5(1+H\sqrt{\lambda})\beta\|\boldsymbol{\phi}_{h,k}\|_{\boldsymbol{\Lambda}^{-1}_{h,k-1}} + 5\mathsf{v}_{\min}\beta^2/k^2, H\}.
$$

By definition of $\mathcal{R}_K$, the optimism of $V^k_h(s)$, and Lemma 6.1, we can bound

$$
\begin{aligned}
\mathcal{R}_K &\leq \sum_{k=1}^K (V^\star_1(s_1) - V^{\pi_k}_1(s_1)) \\
&\leq \sum_{k=1}^K (V^k_1(s_1) - V^{\pi_k}_1(s_1)) \\
&\lesssim \sum_{k=1}^K\sum_{h=1}^H \zeta^k_h + \sum_{k=1}^K\sum_{h=1}^H \min\{(1+H\sqrt{\lambda})\beta\|\boldsymbol{\phi}_{h,k}\|_{\boldsymbol{\Lambda}^{-1}_{h,k-1}} + \mathsf{v}_{\min}\beta^2/k^2, H\}.
\end{aligned}
$$

$\sum_{k=1}^K\sum_{h=1}^H \zeta^k_h$ is a martingale-difference sequence and can be bounded using Freedman's Inequality to obtain the desired $V^\star_0$ dependence. In particular, we have, with high probability

$$
\sum_{k=1}^K\sum_{h=1}^H \zeta^k_h \lesssim \sqrt{H^2 V^\star_1 K \cdot \log 1/\delta} + (\text{lower order terms}).
$$

In addition, $\mathsf{v}_{\min}\beta^2/k^2$ sums to a term that is $\text{poly}(d, H)$, so we ignore it for future calculations. We focus our attention on the term:

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\min\{(1+H\sqrt{\lambda})\beta\|\boldsymbol{\phi}_{h,k}\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}}, H\}$$

which can be expressed as:

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\mathsf{v}}_{h,k}\min\{(1+H\sqrt{\lambda})\beta\|\boldsymbol{\phi}_{h,k}/\bar{\mathsf{v}}_{h,k}\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}}, H/\bar{\mathsf{v}}_{h,k}\}. \tag{6.4}$$

Typically, terms such as this are handled via the Elliptic Potential Lemma. However, to apply the Elliptic Potential Lemma (Abbasi-Yadkori et al., 2011) here, we need to choose $\lambda = 1/\mathsf{v}_{\min}^2$ to guarantee $\lambda \geq \max_{h,k}\|\boldsymbol{\phi}_{h,k}/\bar{\mathsf{v}}_{h,k}\|_2^2$. Due to the $\sqrt{\lambda}$ dependence, this will result in a $1/\mathsf{v}_{\min}$ scaling in the final regret bound, which is prohibitively large. To overcome this, we instead apply the following result, to control the number of times $\|\boldsymbol{\phi}_{h,k}/\bar{\mathsf{v}}_{h,k}\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}}$ can be large:

Lemma 6.2. *Consider a sequence of vectors $(\boldsymbol{x}_t)_{t=1}^T, \boldsymbol{x}_t \in \mathbb{R}^d$, and assume that $\|\boldsymbol{x}_t\|_2 \leq a$ for all t. Let $\mathbf{V}_t = \lambda I + \sum_{s=1}^t \boldsymbol{x}_s\boldsymbol{x}_s^\top$ for some $\lambda > 0$. Then, we will have that $\|\boldsymbol{x}_t\|_{\mathbf{V}_{t-1}^{-1}} > b$ at most*

$$d\log(1+a^2T/\lambda)/\log(1+b)$$

times.

Let $\mathcal{K}_h = \{k \ : \ \|\boldsymbol{\phi}_{h,k}/\bar{\mathsf{v}}_{h,k}\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}} \leq 1\}$. Then we can bound (6.4) as

$$\begin{aligned}(6.4) &\lesssim \sum_{h=1}^{H}\sum_{k\in\mathcal{K}_h}(1+H\sqrt{\lambda})\beta\bar{\mathsf{v}}_{h,k}\min\{\|\boldsymbol{\phi}_{h,k}/\bar{\mathsf{v}}_{h,k}\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}}, 1\} + \sum_{h=1}^{H}H|\mathcal{K}_h^c| \\ &\lesssim \sum_{h=1}^{H}\sum_{k\in\mathcal{K}_h}(1+H\sqrt{\lambda})\beta\bar{\mathsf{v}}_{h,k}\min\{\|\boldsymbol{\phi}_{h,k}/\bar{\mathsf{v}}_{h,k}\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}}, 1\} + dH^2\log(1+K/(\lambda\mathsf{v}_{\min}^2))\end{aligned}$$

where the first inequality holds by definition of $\mathcal{K}_h$, and the second holds by Lemma 6.2. By Cauchy-Schwarz, the first term can be bounded as

$$\lesssim (1+H\sqrt{\lambda})\beta\sqrt{\sum_{h=1}^{H}\sum_{k=1}^{K}\bar{\mathsf{v}}_{h,k}^2}\sqrt{\sum_{h=1}^{H}\sum_{k=1}^{K}\min\{\|\boldsymbol{\phi}_{h,k}/\bar{\mathsf{v}}_{h,k}\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}}^2, 1\}}.$$

As we take the min over 1 and $\|\boldsymbol{\phi}_{h,k}/\bar{\mathsf{v}}_{h,k}\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}}^2$, regardless of the choice of λ we can now apply the Elliptic Potential Lemma to get

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\min\{\|\boldsymbol{\phi}_{h,k}/\bar{\mathsf{v}}_{h,k}\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}}^2, 1\} \lesssim dH\log(1+K/(d\lambda\mathsf{v}_{\min}^2)).$$

Choosing $\lambda = 1/H^2$, we then have that the regret is bounded as

$$\lesssim \beta\sqrt{dH\log(1+HK/(d\mathsf{v}_{\min}^2))}\sqrt{\sum_{h=1}^{H}\sum_{k=1}^{K}\bar{\mathsf{v}}_{h,k}^2} + \text{poly}(d, H, \log K).$$

It remains to bound $\bar{\mathrm{v}}_{h,k}^2$. After some manipulation, and using the definition of $\bar{\mathrm{v}}_{h,k}$ given in FORCE, we can bound

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\bar{\mathrm{v}}_{h,k}^2 \lesssim H^2 V_1^\star K + (\text{lower order terms}).$$

Putting this together yields a final regret bound of

$$\beta\sqrt{dH\log(1+HK/(d\mathrm{v}_{\min}^2))}\sqrt{H^2V_1^\star K} + (\text{lower order terms}).$$

In the proof, slightly more care must be taken with handling $\bar{\mathrm{v}}_{h,k}^2$ to avoid a lower order $K^{1/4}$ term, but we defer the details of this to the appendix.

7 Conclusion

In this work we have shown that it is possible to obtain first-order regret in reinforcement learning with large state spaces. Our algorithm, FORCE, critically relies on the robust Catoni estimator, and our analysis establishes novel results on uniform Catoni estimation in general martingale regression settings, which may be of independent interest.

Several questions remain open for future work. First, while we show that it is possible to obtain a computationally efficient version of FORCE, doing so incurs an additional $\sqrt{d}$ factor. Removing this factor while maintaining computational efficiency would be an interesting direction and may require new techniques. More broadly, obtaining a computationally efficient algorithm with regret scaling as $\sqrt{d^2}$ would be an interesting future direction. Zanette et al. (2020b) show that it is possible to obtain a $\sqrt{d^2}$ scaling, but their algorithm is computationally inefficient. In addition, obtaining optimal H dependence is of much interest. While FORCE will achieve this for $V_1^\star \leq 1$, technical challenges remain to showing this holds in general. We believe our use of the Catoni estimator could be a key step towards achieving this, but leave this for future work. Finally, developing first-order regret bounds for more general function approximation settings Jiang et al. (2017); Du et al. (2021) is an exciting direction. The results in this work rely strongly on the linearity of the MDP, yet, as a first step, it may be possible to extend our techniques to bilinear classes Du et al. (2021), which also exhibit a certain linear structure.

Acknowledgements

The work of AW is supported by an NSF GFRP Fellowship DGE-1762114. The work of SSD is in part supported by grants NSF IIS-2110170. The work of KJ was funded in part by the AFRL and NSF TRIPODS 2023166.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011.

Agarwal, A., Krishnamurthy, A., Langford, J., Luo, H., et al. Open problem: First-order regret bounds for contextual bandits. In *Conference on Learning Theory*, pp. 4–7. PMLR, 2017.

Allen-Zhu, Z., Bubeck, S., and Li, Y. Make the minority great again: First-order regret bound for contextual bandits. In *International Conference on Machine Learning*, pp. 186–194. PMLR, 2018.

Allenberg, C., Auer, P., Györfi, L., and Ottucsák, G. Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In *International Conference on Algorithmic Learning Theory*, pp. 229–243. Springer, 2006.

Auer, P., Cesa-Bianchi, N., and Gentile, C. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64(1):48–75, 2002.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pp. 463–474. PMLR, 2020.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pp. 263–272. PMLR, 2017.

Bubeck, S. and Sellke, M. First-order bayesian regret analysis of thompson sampling. In *Algorithmic Learning Theory*, pp. 196–233. PMLR, 2020.

Camilleri, R., Jamieson, K., and Katz-Samuels, J. High-dimensional experimental design and kernel bandits. In *International Conference on Machine Learning*, pp. 1227–1237. PMLR, 2021.

Catoni, O. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pp. 1148–1185, 2012.

Cesa-Bianchi, N., Mansour, Y., and Stoltz, G. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2):321–352, 2007.

Dann, C., Lattimore, T., and Brunskill, E. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *arXiv preprint arXiv:1703.07710*, 2017.

Dann, C., Li, L., Wei, W., and Brunskill, E. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pp. 1507–1516. PMLR, 2019.

Dann, C., Marinov, T. V., Mohri, M., and Zimmert, J. Beyond value-function gaps: Improved instance-dependent regret bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Du, S. S., Kakade, S. M., Wang, R., and Yang, L. F. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019.

Du, S. S., Kakade, S. M., Lee, J. D., Lovett, S., Mahajan, G., Sun, W., and Wang, R. Bilinear classes: A structural framework for provable generalization in rl. *arXiv preprint arXiv:2103.10897*, 2021.

Foster, D. J. and Krishnamurthy, A. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

Foster, D. J., Rakhlin, A., and Sridharan, K. Adaptive online learning. *arXiv preprint arXiv:1508.05170*, 2015.

Freedman, D. A. On tail probabilities for martingales. *the Annals of Probability*, pp. 100–118, 1975.

Freund, Y. and Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Hazan, E. and Kale, S. Better algorithms for benign bandits. *Journal of Machine Learning Research*, 12(4), 2011.

He, J., Zhou, D., and Gu, Q. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pp. 4171–4180. PMLR, 2021.

Ito, S., Hirahara, S., Soma, T., and Yoshida, Y. Tight first-and second-order regret bounds for adversarial linear bandits. *Advances in Neural Information Processing Systems*, 33:2028–2038, 2020.

Jia, Z., Yang, L., Szepesvari, C., and Wang, M. Model-based reinforcement learning with value-targeted regression. In *Learning for Dynamics and Control*, pp. 666–686. PMLR, 2020.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2017.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 4868–4878, 2018.

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pp. 4870–4879. PMLR, 2020a.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020b.

Jin, C., Liu, Q., and Miryoosefi, S. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *arXiv preprint arXiv:2102.00815*, 2021.

Kakade, S. M. *On the sample complexity of reinforcement learning*. PhD thesis, UCL (University College London), 2003.

Kearns, M. and Singh, S. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2):209–232, 2002.

Kim, Y., Yang, I., and Jun, K.-S. Improved regret analysis for variance-adaptive linear bandits and horizon-free linear mixture mdps. *arXiv preprint arXiv:2111.03289*, 2021.

Koolen, W. M. and Van Erven, T. Second-order quantile methods for experts and combinatorial games. In *Conference on Learning Theory*, pp. 1155–1175. PMLR, 2015.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Lee, C.-W., Luo, H., Wei, C.-Y., Zhang, M., and Zhang, X. Achieving near instance-optimality and minimax-optimality in stochastic and adversarial linear bandits simultaneously. *arXiv preprint arXiv:2102.05858*, 2021.

Lugosi, G. and Mendelson, S. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

Luo, H. and Schapire, R. E. Achieving all with no parameters: Adanormalhedge. In *Conference on Learning Theory*, pp. 1286–1304. PMLR, 2015.

Lykouris, T., Sridharan, K., and Tardos, É. Small-loss bounds for online learning with partial information. In *Conference on Learning Theory*, pp. 979–986. PMLR, 2018.

Neu, G. First-order regret bounds for combinatorial semi-bandits. In *Conference on Learning Theory*, pp. 1360–1375. PMLR, 2015.

Sarkar, T. and Rakhlin, A. Near optimal finite time identification of arbitrary linear dynamical systems. In *International Conference on Machine Learning*, pp. 5610–5618. PMLR, 2019.

Simchowitz, M. and Jamieson, K. Non-asymptotic gap-dependent regret bounds for tabular mdps. *arXiv preprint arXiv:1905.03814*, 2019.

Srebro, N., Sridharan, K., and Tewari, A. Smoothness, low noise and fast rates. *Advances in neural information processing systems*, 23, 2010.

Vapnik, V. N. and Chervonenkis, A. Y. On the uniform convergence of relative frequencies of events to their probabilities. 1971.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Wang, R., Du, S. S., Yang, L. F., and Kakade, S. M. Is long horizon reinforcement learning more difficult than short horizon reinforcement learning? *arXiv preprint arXiv:2005.00527*, 2020.

Wang, Y., Wang, R., Du, S. S., and Krishnamurthy, A. Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*, 2019.

Wang, Y., Wang, R., and Kakade, S. M. An exponential lower bound for linearly-realizable mdps with constant suboptimality gap. *arXiv preprint arXiv:2103.12690*, 2021.

Wei, C.-Y. and Luo, H. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, pp. 1263–1291. PMLR, 2018.

Wei, C.-Y., Luo, H., and Agarwal, A. Taking a hint: How to leverage loss predictors in contextual bandits? In *Conference on Learning Theory*, pp. 3583–3634. PMLR, 2020.

Weisz, G., Amortila, P., and Szepesvári, C. Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory*, pp. 1237–1264. PMLR, 2021.

Xu, H., Ma, T., and Du, S. S. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. *arXiv preprint arXiv:2102.04692*, 2021.

Yang, L. and Wang, M. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pp. 6995–7004. PMLR, 2019.

Zanette, A. and Brunskill, E. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pp. 7304–7312. PMLR, 2019.

Zanette, A., Brandfonbrener, D., Brunskill, E., Pirotta, M., and Lazaric, A. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pp. 1954–1964. PMLR, 2020a.

Zanette, A., Lazaric, A., Kochenderfer, M., and Brunskill, E. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pp. 10978–10989. PMLR, 2020b.

Zhang, Z., Ji, X., and Du, S. S. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. *arXiv preprint arXiv:2009.13503*, 2020a.

Zhang, Z., Zhou, Y., and Ji, X. Almost optimal model-free reinforcement learningvia reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33, 2020b.

Zhang, Z., Yang, J., Ji, X., and Du, S. S. Variance-aware confidence set: Variance-dependent bound for linear bandits and horizon-free bound for linear mixture mdp. *arXiv preprint arXiv:2101.12745*, 2021.

Zhou, D., Gu, Q., and Szepesvari, C. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. *arXiv preprint arXiv:2012.08507*, 2020.

Zhou, D., He, J., and Gu, Q. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, pp. 12793–12802. PMLR, 2021.

A Technical Results

A.1 Covering and Elliptical Potential Lemmas

Definition A.1 (Covering Number)**.** Let $\mathcal{X}$ be a set with metric $\text{dist}(\cdot,\cdot)$. Given $\epsilon > 0$, the ϵ-covering number of $\mathcal{X}$ in dist, $\mathsf{N}(\mathcal{X}, \text{dist}, \epsilon)$, is defined as the minimal cardinality of a set $\mathcal{N} \subset \mathcal{X}$ such that, for all $x \in \mathcal{X}$, there exists an $x' \in \mathcal{N}$ with $\text{dist}(x, x') \le \epsilon$.

Lemma A.1 (Vershynin (2010))**.** *For any $\epsilon > 0$, the ϵ-covering number of the Euclidean ball $\mathcal{B}^d(R) := \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 = 1\}$ with radius $R > 0$ in the Euclidean metric is upper bounded by $(1 + 2R/\epsilon)^d$.*

Lemma A.2 (Lemma D.6 of Jin et al. (2020b))**.** *Consider the class of functions from $\mathbb{R}^d$ to $\mathbb{R}$ of the form*

$$f(\boldsymbol{\phi}) = \min\{\langle \boldsymbol{w}, \boldsymbol{\phi}\rangle + \beta\|\boldsymbol{\phi}\|_{\boldsymbol{\Lambda}^{-1}}, H\}$$

where the parameters $\boldsymbol{w}, \beta, \boldsymbol{\Lambda}$ satisfy $\|\boldsymbol{w}\|_2 \le \beta_{\boldsymbol{w}}$, $\beta \in [0, B]$, and $\boldsymbol{\Lambda} \succeq \lambda I$. Let $\mathcal{N}_\epsilon$ be an ϵ-covering of this set with respect to the norm $\text{dist}_\infty(f, f') := \sup_{\boldsymbol{\phi}\in\mathcal{B}^d} |f(\boldsymbol{\phi}) - f'(\boldsymbol{\phi})|$. Then,

$$\log|\mathcal{N}_\epsilon| \le d\log(1 + 4\beta_{\boldsymbol{w}}/\epsilon) + d^2\log(1 + 8\sqrt{d}B^2/(\lambda\epsilon^2)).$$

Lemma A.3 (Elliptic Potential Lemma, Lemma 11 of Abbasi-Yadkori et al. (2011))**.** *Under the same assumptions as Lemma 6.2, for any choice of $\lambda > 0$, we will have that*

$$\sum_{t=1}^{T} \min\{1, \|\boldsymbol{x}_t\|^2_{\mathbf{V}_{t-1}^{-1}}\} \le 2d\log(1 + a^2T/(d\lambda)).$$

Furthermore, if $\lambda \ge \max\{1, a^2\}$,

$$\sum_{t=1}^{T} \|\boldsymbol{x}_t\|^2_{\mathbf{V}_{t-1}^{-1}} \le 2d\log(1 + a^2T/(d\lambda)).$$

Lemma A.4 (Freedman's Inequality (Freedman, 1975))**.** *$\mathcal{F}_0 \subset \mathcal{F}_1 \subset \ldots \subset \mathcal{F}_T$ be a filtration and let $X_1, X_2, \ldots, X_T$ be real random variables such that X_t is $\mathcal{F}_t$-measurable, $\mathbb{E}[X_t|\mathcal{F}_{t-1}] = 0$, $|X_t| \le b$ almost surely, and $\sum_{t=1}^{T}\mathbb{E}[X_t^2|\mathcal{F}_{t-1}] \le V$ for some fixed $V > 0$ and $b > 0$. Then for any $\delta \in (0, 1)$, we have with probability at least $1 - \delta$,*

$$\sum_{t=1}^{T} X_t \le 2\sqrt{V\log(1/\delta)} + b\log(1/\delta).$$

Proof of Lemma 6.2. Our goal is to bound the number of times that $\|\boldsymbol{x}_t\|_{\mathbf{V}_{t-1}^{-1}} > b$. A now-standard determinant computation (see, e.g. Abbasi-Yadkori et al. (2011)) based on the Sherman-Morrison identity yields

$$\det(\mathbf{V}_t) = \det(\mathbf{V}_{t-1})(1 + \|\boldsymbol{x}_t\|^2_{\mathbf{V}_{t-1}^{-1}}) \implies \|\boldsymbol{x}_t\|^2_{\mathbf{V}_{t-1}^{-1}} = \frac{\det(\mathbf{V}_t)}{\det(\mathbf{V}_{t-1})} - 1.$$

It follows that, whenever $\|\boldsymbol{x}_t\|_{V_{t-1}^{-1}} > b$, it must also be the case that

$$\frac{\det(\mathbf{V}_t)}{\det(\mathbf{V}_{t-1})} - 1 > b \iff \det(\mathbf{V}_t) > (1 + b)\det(\mathbf{V}_{t-1}).$$

In particular, if N denotes the number of times that $\|\boldsymbol{x}_t\|_{\mathbf{V}_{t-1}^{-1}} > b$ for $t \in \{1, \ldots, T\}$, then it follows that $\det(\mathbf{V}_T) > (1+b)^N \det(\mathbf{V}_0) = (1+b)^N \lambda^d$. At the same time,

$$\begin{aligned}\det(\mathbf{V}_T) &= \det\left(\lambda I + \sum_{s=1}^{T} \boldsymbol{x}_s \boldsymbol{x}_s^\top\right) \\ &\leq \left(\|\lambda I + \sum_{s=1}^{T} \boldsymbol{x}_s \boldsymbol{x}_s^\top\|_{\mathrm{op}}\right)^d \\ &\leq (\lambda + a^2 T)^d.\end{aligned}$$

Combining these inequalities gives:

$$(1+b)^N \lambda^d < (\lambda + a^2 T)^d \iff N < \frac{d\log(\lambda + a^2 T) - d\log(\lambda)}{\log(1+b)}.$$

□

We remark that a variant of Lemma 6.2 appeared in concurrent work (Kim et al., 2021), and originally as an exercise in Lattimore & Szepesvári (2020).

A.2 Martingale Catoni Estimation

Lemma A.5 (Martingale Catoni Estimator, Lemma 13 of Wei et al. (2020))**.** *Let $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \ldots \subset \mathcal{F}_T$ be a filtration and let $X_1, X_2, \ldots, X_T$ be square-integrable real random variables such that X_t is $\mathcal{F}_t$-measurable, and*

- *Conditional means $\mathbb{E}[X_t|\mathcal{F}_{t-1}] = \zeta_t$ for some fixed (non-random) ζ_t.*
- *Average conditional mean $\zeta := \frac{1}{T}\sum_{t=1}^{T} \zeta_t$.*
- *Conditional variances $\sum_{t=1}^{T} \mathbb{E}[(X_t - \zeta_t)^2|\mathcal{F}_{t-1}] \leq V$ for some fixed (non-random) $V > 0$.*

Then for any confidence $\delta \in (0,1)$ and sample size $T \geq \alpha^2(V + \sum_{t=1}^{T}(\zeta_t - \zeta)^2) + 2\log\frac{1}{\delta}$, we have with probability at least $1 - 2\delta$, the Catoni estimator $\mathsf{cat}_{T,\alpha}$ satisfies

$$|\mathsf{cat}_{T,\alpha} - \zeta| \leq \frac{\alpha\left(V + \sum_{t=1}^{T}(\zeta_t - \zeta)^2\right)}{T} + \frac{2\log\frac{1}{\delta}}{\alpha T}.$$

We recall the heteroscedastic heavy-tailed martingale linear regression setting as defined in Definition 5.1.

Definition 5.1 (Heteroscedastic Heavy-Tailed Martingale Linear Regression)**.** Let $(\mathcal{F}_t)_{t\geq 0}$ denote a filtration, let $\boldsymbol{\phi}_t \in \mathbb{R}^d$ be a sequence of random $\mathcal{F}_{t-1}$-measurable vectors, and $y_t \in \mathbb{R}$ be $\mathcal{F}_t$-measurable random scalars satisfying

$$y_t = \langle \boldsymbol{\phi}_t, \boldsymbol{\theta}_\star \rangle + \eta_t, \quad \mathbb{E}[y_t|\mathcal{F}_{t-1}] = \langle \boldsymbol{\phi}_t, \boldsymbol{\theta}_\star \rangle$$

for some $\boldsymbol{\theta}_\star \in \mathbb{R}^d$ and η_t satisfying $\mathbb{E}[\eta_t|\mathcal{F}_{t-1}] = 0$, and $\mathbb{E}[\eta_t^2|\mathcal{F}_{t-1}] < \infty$, but otherwise arbitrary (as such, the distribution of η_t may depend on $\boldsymbol{\phi}_t$). Furthermore, let σ_t^2 be a $\mathcal{F}_{t-1}$-measurable sequence of scalars satisfying $\sigma_t^2 \geq \mathbb{E}[y_t^2|\mathcal{F}_{t-1}]$, and let

$$\boldsymbol{\Sigma}_T := \sum_{t=1}^{T} \sigma_t^{-2} \boldsymbol{\phi}_t \boldsymbol{\phi}_t^\top.$$

Lemma 5.1 (Heteroscedastic Catoni Estimator)**.** *Assume we are in the regression setting of Definition 5.1. For a fixed vector $\boldsymbol{v} \in \mathbb{R}^d$ let $\mathsf{cat}[\boldsymbol{v}]$ denote the Catoni estimate applied to $(X_t)_{t=1}^T$ where $X_t := \boldsymbol{v}^\top \boldsymbol{\phi}_t y_t / \sigma_t^2$, with a fixed (deterministic) parameter $\alpha > 0$. Then, for any failure probability $\delta > 0$ and fixed $\alpha_{\max} > 0$, if our deterministic α can be written as*

$$\alpha = \min\left\{ \gamma \cdot \frac{\sqrt{\log(2/\delta)}}{\|\boldsymbol{v}\|_{\boldsymbol{\Sigma}_T}}, \alpha_{\max} \right\} \tag{5.1}$$

for some (possibly random) $\gamma \geq 1$, then

$$\left| \mathsf{cat}[\boldsymbol{v}] - \frac{1}{T} \boldsymbol{v}^\top \boldsymbol{\Sigma}_T \cdot \boldsymbol{\theta}_\star \right| \leq (2 + 2\gamma) \|\boldsymbol{v}\|_{\boldsymbol{\Sigma}_T} \sqrt{\frac{\log \frac{2}{\delta}}{T^2}} + \frac{2 \log \frac{2}{\delta}}{\alpha_{\max} T}$$

provided that $T \geq (2 + 2\gamma^2) \log \frac{2}{\delta}$.

Proof. We apply Lemma A.5 to the scalar data $X_t := \boldsymbol{v}^\top \boldsymbol{\phi}_t y_t / \sigma_t^2$. Note that with this choice of X_t,

$$\mathbb{E}[X_t | \mathcal{F}_{t-1}] = \frac{1}{\sigma_t^2} \boldsymbol{v}^\top \boldsymbol{\phi}_t \boldsymbol{\phi}_t^\top \boldsymbol{\theta}_\star =: \zeta_t[\boldsymbol{v}]$$

so we will have that

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E}[X_t | \mathcal{F}_{t-1}] = \frac{1}{T} \sum_{t=1}^T \frac{1}{\sigma_t^2} v^\top \boldsymbol{\phi}_t \boldsymbol{\phi}_t^\top \boldsymbol{\theta} = \zeta[\boldsymbol{v}]$$

Applying Lemma A.5 gives that, with probability at least $1 - \delta$,

$$|\mathsf{cat}[\boldsymbol{v}] - \zeta[\boldsymbol{v}]| \leq \frac{\alpha \left(V + \sum_{t=1}^T (\zeta_t[\boldsymbol{v}] - \zeta[\boldsymbol{v}])^2 \right)}{T} + \frac{2 \log 2/\delta}{\alpha T},$$

where $V > 0$ is any fixed upper bound on the quantity

$$\begin{aligned} V \geq \sum_{t=1}^{\mathbb{E}} [(X_t - \zeta_t[\boldsymbol{v}])^2 \mid \mathcal{F}_{t-1}] &= \sum_{t=1}^T \mathbb{E}\left[\sigma_t^{-4} \left(\boldsymbol{v}^\top \boldsymbol{\phi}_t y_t - \boldsymbol{v}^\top \boldsymbol{\phi}_t \boldsymbol{\phi}_t^\top \boldsymbol{\theta}_\star \right)^2 | \mathcal{F}_{t-1} \right] \\ &= \sum_{t=1}^T \left(\boldsymbol{v}^\top \boldsymbol{\phi}_t / \sigma_t \right)^2 \mathbb{E}\left[(y_t - \boldsymbol{\phi}_t^\top \boldsymbol{\theta}_\star)^2 / \sigma_t^2 | \mathcal{F}_{t-1} \right] \end{aligned}$$

Our assumption on σ_t ensures that

$$\begin{aligned} 1 \geq \sigma_t^{-2} \mathbb{E}[y_t^2 \mid \mathcal{F}_{t-1}] &= \sigma_t^{-2} \mathrm{Var}[y_t] + \sigma_t^{-2} \mathbb{E}[y_t \mid \mathcal{F}_{t-1}]^2 \\ &= \mathbb{E}\left[(y_t - \boldsymbol{\phi}_t^\top \boldsymbol{\theta}_\star)^2 / \sigma_t^2 \mid \mathcal{F}_{t-1} \right] + \sigma_t^{-2} (\boldsymbol{\phi}_t^\top \boldsymbol{\theta}_\star)^2, \end{aligned} \tag{A.1}$$

so it suffices that we select V to be

$$V = \sum_{t=1}^T \left(\boldsymbol{v}^\top \boldsymbol{\phi}_t / \sigma_t \right)^2 = \|\boldsymbol{v}\|_{\boldsymbol{\Sigma}_T}^2.$$

Furthermore, since $\zeta[\boldsymbol{v}]$ is the average of the terms $\zeta_t[\boldsymbol{v}]$, we can upper bound

$$\sum_{t=1}^T (\zeta_t[\boldsymbol{v}] - \zeta[\boldsymbol{v}])^2 \leq \sum_{t=1}^T \zeta_t[\boldsymbol{v}]^2 = \sum_{t=1}^T (\boldsymbol{v}^\top \boldsymbol{\phi}_t \boldsymbol{\phi}_t^\top \boldsymbol{\theta}_\star / \sigma_t^2)^2$$

$$= \sum_{t=1}^{T} (\boldsymbol{v}^\top \boldsymbol{\phi}_t / \sigma_t)^2 (\boldsymbol{\phi}_t^\top \boldsymbol{\theta}_\star / \sigma_t)^2$$

Again, our assumption on σ_t ensures $(\boldsymbol{\phi}_t^\top \boldsymbol{\theta}_\star / \sigma_t)^2 \le 1$ via Equation (A.1), so we can bound

$$\sum_{t=1}^{T} (\zeta_t[\boldsymbol{v}] - \zeta[\boldsymbol{v}])^2 \le \sum_{t=1}^{T} (\boldsymbol{v}^\top \boldsymbol{\phi}_t / \sigma_t)^2 = \|\boldsymbol{v}\|_{\boldsymbol{\Sigma}_T}^2.$$

Putting these two bounds together, we have that

$$\begin{aligned} |\mathsf{cat}[\boldsymbol{v}] - \zeta[\boldsymbol{v}]| &\le \frac{\alpha \left(V + \sum_{t=1}^{T} (\zeta_t[\boldsymbol{v}] - \zeta[\boldsymbol{v}])^2 \right)}{T} + \frac{2 \log \frac{1}{\delta}}{\alpha T}. \\ &\le \frac{2\alpha \|\boldsymbol{v}\|_{\boldsymbol{\Sigma}_T}^2}{T} + \frac{2 \log \frac{2}{\delta}}{\alpha T}. \end{aligned} \tag{A.2}$$

provided that

$$T \ge 2\alpha^2 \|\boldsymbol{v}\|_{\boldsymbol{\Sigma}_T}^2 + 2 \log \frac{2}{\delta} \ge \alpha^2 (V + \sum_{t=1}^{T} (\zeta_t[\boldsymbol{v}] - \zeta[\boldsymbol{v}])^2) + 2 \log \frac{2}{\delta}.$$

Introduce $\alpha_0 := \sqrt{\frac{\log 2/\delta}{\|\boldsymbol{v}\|_{\boldsymbol{\Sigma}_T}^2}}$, so that $\alpha = \min\{\gamma \alpha_0, \alpha_{\max}\}$ (recall $\gamma \ge 1$ is a possibly random scalar but that α is deterministic). Then, $\gamma \alpha_0 \ge \alpha$. Hence, it is enough that

$$T \ge 2\gamma^2 \alpha_0^2 \|\boldsymbol{v}\|_{\boldsymbol{\Sigma}_T}^2 + 2 \log \frac{2}{\delta} = (2 + 2\gamma^2) \log \frac{2}{\delta}.$$

Moreover, using $\alpha = \min\{\gamma \alpha_0, \alpha_{\max}\}$ and $\gamma \ge 1$, we can continue the bound in Equation (A.2) via

$$\begin{aligned} |\mathsf{cat}[\boldsymbol{v}] - \zeta[\boldsymbol{v}]| &\le \frac{2 \min\{\alpha_{\max}, \gamma \alpha_0\} \|\boldsymbol{v}\|_{\boldsymbol{\Sigma}_T}^2}{T} + \frac{2 \log \frac{2}{\delta}}{\min\{\alpha_{\max}, \gamma \alpha_0\} T} \\ &\le \frac{2\gamma \alpha_0 \|\boldsymbol{v}\|_{\boldsymbol{\Sigma}_T}^2}{T} + \frac{2 \log \frac{2}{\delta}}{\gamma \alpha_0 T} + \frac{2 \log \frac{2}{\delta}}{\alpha_{\max} T} \\ &\le (2 + 2\gamma) \|\boldsymbol{v}\|_{\boldsymbol{\Sigma}_T} \sqrt{\frac{\log 2/\delta}{T^2}} + \frac{2 \log \frac{2}{\delta}}{\alpha_{\max} T}. \end{aligned}$$

□

A.3 Self-Normalized Catoni Estimation

Definition 5.2 (Heteroscedastic Regression with Function Approximation)**.** Given dimension parameters $d, d', p \in \mathbb{N}$, scaling parameters $H, \beta_u, \beta_\mu > 0$, and minimal varaince $\sigma_{\min}^2$, the heteroscedastic regression with function approximation setting is defined as follows. Let $(\mathcal{F}_t)_{t \ge 0}$ be a filtration, and consider a sequence of random vectors $(\boldsymbol{\phi}_t, \boldsymbol{\phi}'_t)_{t=1}^T$ and random scalar outuds $(y_t)_{t=1}^T$ and noises $(\eta_t)_{t=1}^T$ and variance bounds $(\sigma_t^2)_{t=1}^T$ such that

- $\boldsymbol{\phi}_t \in \mathbb{R}^d$ is $\mathcal{F}_{t-1}$-measurable, $\boldsymbol{\phi}'_t \in \mathbb{R}^{d'}$ is $\mathcal{F}_t$ measurable, and $\|\boldsymbol{\phi}_t\|_2, \|\boldsymbol{\phi}'_t\|_2 \le 1$.

- There exists a signed measure $\boldsymbol{\mu}$ over $\mathcal{B}^{d'}$ with total mass $\||\boldsymbol{\mu}|(\mathcal{B}^{d'})\|_2 \le \beta_\mu$ such that, for all t, the conditional distribution of $\boldsymbol{\phi}'_t$ given $\mathcal{F}_{t-1}$ ensures that, for all bounded functions f,

$$\mathbb{E}\left[f(\boldsymbol{\phi}'_t) \mid \mathcal{F}_{t-1}\right] = \langle \boldsymbol{\phi}_t, \int f(\boldsymbol{\phi}')\mathrm{d}\boldsymbol{\mu}(\boldsymbol{\phi}')\rangle. \tag{5.3}$$

- $|\eta_t| \le \beta_\eta$ with probability 1, and $\mathbb{E}[\eta_t \mid \mathcal{F}_{t-1}] = 0$.

- There exist a parameter $\boldsymbol{u}_\star \in \mathbb{R}^d$ with $\|\boldsymbol{u}_\star\|_2 \le \beta_u$, a function class $\mathscr{F}$ of functions $f : \mathbb{R}^{d'} \to [-H, H]$, and a function $f_\star \in \mathscr{F}$ which may be random and dependent on $(\boldsymbol{\phi}_t, \boldsymbol{\phi}'_t)_{t=1}^T$ such that, for all t, $y_t = \langle \boldsymbol{u}_\star, \boldsymbol{\phi}_t\rangle + f_\star(\boldsymbol{\phi}'_t) + \eta_t$. Thus,

$$\mathbb{E}[y_t|\mathcal{F}_{t-1}] = \langle \boldsymbol{\phi}_t, \boldsymbol{\theta}_\star\rangle \quad \text{for} \quad \boldsymbol{\theta}_\star = \boldsymbol{u}_\star + \int f_\star(\boldsymbol{\phi}')\mathrm{d}\boldsymbol{\mu}(\boldsymbol{\phi}').$$

- σ_t are uniformly lower bounded by $\sigma_{\min}$, finite, $\mathcal{F}_{t-1}$ measurable, and satisfy

$$\mathbb{E}\left[\left(\langle \boldsymbol{\phi}_t, \boldsymbol{u}_\star\rangle + f_\star(\boldsymbol{\phi}'_t) + \eta_t\right)^2 \mid \mathcal{F}_{t-1}\right] \le \frac{1}{2}\sigma_t^2. \tag{5.4}$$

- The covering numbers of $\mathscr{F}$ are *parameteric*, in the sense that there exists a $p \in \mathbb{N}$ and $R > 0$ such that, for $\epsilon > 0$, the ϵ-covering number of $\mathscr{F}$ in the metric $\mathrm{dist}_\infty(f, f') := \sup_{\boldsymbol{\phi}' \in \mathcal{B}^{d'}} |f(\boldsymbol{\phi}') - f'(\boldsymbol{\phi}')|$ is bounded as $\mathsf{N}(\mathscr{F}, \mathrm{dist}_\infty, \epsilon) \le p \log(1 + \frac{2R}{\epsilon})$, where $\mathsf{N}(\mathscr{F}, \mathrm{dist}_\infty, \epsilon)$ is the ϵ-covering number of $\mathscr{F}$ in the norm dist_∞.

We now state an intermediate technical proposition, from which derive our main self-normalized guarantee as a special case:

Proposition 7. *Let $c > 0$ denote a universal constant, take parameters $\lambda > 0$ and $\alpha_{\max} \ge 1$, and consider the regression with function approximation of Definition 5.2 with parameters $d, p, \sigma^2_{\min}, \beta_u, \beta_\mu, R, H$. For a sample size $T \in \mathbb{N}$ introduce the effective dimension*

$$d_T := c \cdot (p + d) \cdot \mathsf{logs}\left(T, \alpha^2_{\max}, \lambda^{-1}, \sigma^{-2}_{\min}, \beta_\mu, \beta_u, \beta_\eta, R, H\right).$$

For vectors $\widetilde{\boldsymbol{v}} \in \mathbb{R}^d$, define the mean parameter

$$\zeta[\widetilde{\boldsymbol{v}}] := \frac{1}{T}\widetilde{\boldsymbol{v}}^\top \boldsymbol{\Sigma}_T \cdot \boldsymbol{\theta}_\star$$

and let $\mathsf{cat}[\widetilde{\boldsymbol{v}}]$ denote the Catoni estimator using features and $\alpha[\widetilde{\boldsymbol{v}}]$ parameter

$$X_t = \widetilde{\boldsymbol{v}}^\top \boldsymbol{\phi}_t y_t / \sigma_t^2, \quad \alpha[\widetilde{\boldsymbol{v}}] = \min\left\{\sqrt{\tfrac{d_T + \log 1/\delta}{\|\widetilde{\boldsymbol{v}}\|^2_{\boldsymbol{\Sigma}_T}}}, \alpha_{\max}\right\}.$$

Then, if $T \ge 6(\log \frac{1}{\delta} + d_T)$, with probability $1 - \delta$, it holds that $\forall \boldsymbol{v} \in \mathcal{B}^d$ and for all $\boldsymbol{\Lambda} \succeq \lambda I$,

$$\left|\mathsf{cat}\left[\boldsymbol{\Lambda}^{-1}\boldsymbol{v}\right] - \zeta\left[\boldsymbol{\Lambda}^{-1}\boldsymbol{v}\right]\right| \le 4\|\boldsymbol{\Lambda}^{-1}\boldsymbol{v}\|_{\boldsymbol{\Sigma}_T}\sqrt{\frac{\log \frac{1}{\delta} + d_T}{T^2}} + \frac{3(\log \frac{1}{\delta} + d_T)}{\alpha_{\max} T}.$$

Proof of Theorem 6. We instantiate Proposition 7 with $\boldsymbol{\Lambda} = \frac{1}{T}(\lambda I + \boldsymbol{\Sigma}_T) \succeq \frac{1}{T}\boldsymbol{\Sigma}_T$. For this choice of $\boldsymbol{\Lambda}$, it holds that

$$\|\boldsymbol{\Lambda}^{-1}\boldsymbol{v}\|^2_{\boldsymbol{\Sigma}_T} = \boldsymbol{v}^\top \boldsymbol{\Lambda}^{-1}\boldsymbol{\Sigma}_T\boldsymbol{\Lambda}^{-1}\boldsymbol{v} \le T\|\boldsymbol{v}\|^2_{\boldsymbol{\Lambda}^{-1}}.$$

Moreover, $\mathbf{\Lambda} \succeq \frac{1}{T} \cdot \lambda I$, so taking $\lambda \leftarrow \lambda/T$, d_T still has the same form for a possibly larger constant $c > 0$. Next,

$$\begin{aligned}\zeta[\mathbf{\Lambda}^{-1}\boldsymbol{v}] &= \frac{1}{T}\boldsymbol{v}^\top \mathbf{\Lambda}^{-1}\mathbf{\Sigma}_T \cdot \boldsymbol{\theta}_\star \\ &= \boldsymbol{v}^\top(\mathbf{\Sigma}_T + \lambda I)^{-1}\mathbf{\Sigma}_T \cdot \boldsymbol{\theta}_\star \\ &= \boldsymbol{v}^\top \boldsymbol{\theta}_\star - \lambda \boldsymbol{v}^\top(\mathbf{\Sigma}_T + \lambda I)^{-1} \cdot \boldsymbol{\theta}_\star.\end{aligned}$$

It follows that

$$\left|\mathsf{cat}\left[\mathbf{\Lambda}_T^{-1}\boldsymbol{v}\right] - \boldsymbol{v}^\top\boldsymbol{\theta}_\star\right| \le \left|\mathsf{cat}\left[\mathbf{\Lambda}_T^{-1}\boldsymbol{v}\right] - \zeta[\mathbf{\Lambda}^{-1}\boldsymbol{v}]\right| + |\lambda \boldsymbol{v}^\top(\mathbf{\Sigma}_T + \lambda I)^{-1} \cdot \boldsymbol{\theta}_\star|.$$

We bound $\left|\mathsf{cat}\left[\mathbf{\Lambda}_T^{-1}\boldsymbol{v}\right] - \zeta[\mathbf{\Lambda}^{-1}\boldsymbol{v}]\right|$ by Proposition 7 and bound

$$\begin{aligned}|\lambda \boldsymbol{v}^\top(\mathbf{\Sigma}_T + \lambda I)^{-1} \cdot \boldsymbol{\theta}_\star| &\le \lambda \|\boldsymbol{v}^\top(\mathbf{\Sigma}_T + \lambda I)^{-1/2}\|_2 \|(\mathbf{\Sigma}_T + \lambda I)^{-1/2}\|_{\mathrm{op}} \|\boldsymbol{\theta}_\star\|_2 \\ &\le \sqrt{\lambda}\|\boldsymbol{v}\|_{\mathbf{\Lambda}^{-1}}\|\boldsymbol{\theta}_\star\|_2.\end{aligned}$$

□

Proof of Proposition 7. The proof requires a careful covering of directions $\boldsymbol{v}^\top\mathbf{\Lambda}^{-1}$, and regression functions $f \in \mathscr{F}$.

Notation. Let us establish some notation to facilliate the covering. Given $f \in \mathscr{F}$, we define the associated targets

$$\widetilde{y}_t(f) := \langle \boldsymbol{\phi}_t, \boldsymbol{u}_\star\rangle + f(\boldsymbol{\phi}_t') + \eta_t, \quad \tilde{\boldsymbol{\theta}}(f) := \boldsymbol{u}_\star + \int f(\boldsymbol{\phi}')\mathrm{d}\boldsymbol{\mu}(\boldsymbol{\phi}').$$

Given $\widetilde{\boldsymbol{v}} \in \mathbb{R}^d$ and $f \in \mathscr{F}$, define

$$\zeta[\widetilde{\boldsymbol{v}}, f] := \widetilde{\boldsymbol{v}}^\top \mathbf{\Sigma}_T \boldsymbol{\theta}_\star,$$

and let $\mathsf{cat}[\widetilde{\boldsymbol{v}}, f, \tilde{\alpha}]$ to denote the Catoni estimator using parameter $\tilde{\alpha}$ and features

$$X_t[\widetilde{\boldsymbol{v}}, f] := \frac{1}{\sigma_t^2}\widetilde{\boldsymbol{v}}^\top \boldsymbol{\phi}_t \widetilde{y}_t(f) \tag{A.3}$$

Over loading notation, define $\mathsf{cat}[\widetilde{\boldsymbol{v}}, f]$ to denote the following estimate using the correct, data-dependent :

$$\mathsf{cat}[\widetilde{\boldsymbol{v}}, f] = \mathsf{cat}[\widetilde{\boldsymbol{v}}, f, \alpha[\widetilde{\boldsymbol{v}}]], \quad \alpha[\widetilde{\boldsymbol{v}}] = \min\left\{\frac{\sqrt{\log\frac{2M}{\delta}}}{\|\widetilde{\boldsymbol{v}}\|_{\mathbf{\Sigma}_T}}, \alpha_{\max}\right\}, \tag{A.4}$$

Note that the correspondence between the original notation parameterized by direction $\mathbf{\Lambda}^{-1}\boldsymbol{v}$ and the new notation is given by

$$\mathsf{cat}[\widetilde{\boldsymbol{v}}, f_\star] = \mathsf{cat}[\mathbf{\Lambda}^{-1}\boldsymbol{v}], \quad \zeta[\widetilde{\boldsymbol{v}}, f_\star] = \zeta^\star[\mathbf{\Lambda}^{-1}\boldsymbol{v}], \quad \widetilde{\boldsymbol{v}} = \mathbf{\Lambda}^{-1}\boldsymbol{v}.. \tag{A.5}$$

We note that by the assumption that $\|\boldsymbol{v}\|_2 \le 1$ and $\mathbf{\Lambda} \succeq \lambda I$, it suffices to consider $\widetilde{\boldsymbol{v}}$ in the set

$$\mathcal{V} := \{\widetilde{\boldsymbol{v}} : \|\widetilde{\boldsymbol{v}}\|_2 \le \beta_{\tilde{v}}, \quad \beta_{\tilde{v}} := 1/\lambda\}.$$

Lastly, we define the interval

$$\mathcal{A} := \{\alpha : \frac{\lambda}{T} \le \alpha \le \alpha_{\max}\}$$

Rounding α. To handle that $\alpha[\widetilde{\boldsymbol{v}}]$ is data-dependent, we will build a cover using the Catoni estimator with rounded values of $\alpha[\widetilde{\boldsymbol{v}}]$. Note that this step is purely for the analysis, and does not need to be incorporated into the algorithm. For $\epsilon > 0$ and scalar k, set

$$\mathsf{round}(x, \epsilon) := \inf\{(1+\epsilon)^k : (1+\epsilon)^k \geq x, \quad k \in \mathbb{N}\}.$$

Fixing an $\epsilon_{\mathsf{rnd}} \in (0, 1/4)$ to be chosen, set

$$\begin{aligned} \mathsf{cat}_{\mathsf{rnd}}[\widetilde{\boldsymbol{v}}, f] &= \mathsf{cat}[\widetilde{\boldsymbol{v}}, f, \alpha_{\mathsf{rnd}}[\widetilde{\boldsymbol{v}}]], \\ \text{where } \alpha_{\mathsf{rnd}}[\widetilde{\boldsymbol{v}}] &= \min\left\{ \mathsf{round}\left(\frac{\sqrt{\log \frac{2M}{\delta}}}{\|\widetilde{\boldsymbol{v}}\|_{\boldsymbol{\Sigma}_T}}, \epsilon_{\mathsf{rnd}} \right), \alpha_{\max} \right\} \end{aligned} \tag{A.6}$$

Note that since $\|\boldsymbol{\phi}_t\|_2 \leq 1$ and $\|\widetilde{\boldsymbol{v}}\|_2 \leq 1/\lambda$ for $\widetilde{\boldsymbol{v}} \in \mathcal{V}$, we have that $\alpha[\widetilde{\boldsymbol{v}}] \in \mathcal{A}$ for $\widetilde{\boldsymbol{v}} \in \mathcal{V}$. Note then that the rounded Catoni parameters lie in the finite set

$$\alpha_{\mathsf{rnd}}[\widetilde{\boldsymbol{v}}] \in \mathcal{A}_{\mathsf{rnd}}, \quad \text{where } \mathcal{A}_{\mathsf{rnd}} := \left\{ (1+\epsilon_{\mathsf{rnd}})^k : \frac{\lambda}{T} \leq (1+\epsilon_{\mathsf{rnd}})^k \leq \alpha_{\max} \right\} \cup \{\alpha_{\max}\}. \tag{A.7}$$

Furthermore, the cardinality of $\mathcal{A}_{\mathsf{rnd}}$ can be crudely bounded by

$$\begin{aligned} |\mathcal{A}_{\mathsf{rnd}}| \leq \log_{1+\epsilon_{\mathsf{rnd}}}(T\alpha_{\max}/\lambda) &= 1 + \frac{\log(T\alpha_{\max}/\lambda)}{\log(1+\epsilon_{\mathsf{rnd}})} \\ &\leq 1 + (T\alpha_{\max}/\lambda) \cdot 2/\epsilon_{\mathsf{rnd}} \\ &\leq 1 + \frac{2T\alpha_{\max}}{\lambda\epsilon_{\mathsf{rnd}}}. \end{aligned} \tag{A.8}$$

where we used the crude bound $\log x \leq 1 + x$ for $x \geq 1$, and $\log(1+\epsilon) \geq \epsilon/2$ for $\epsilon \in (0, 1/4)$.

Uniform bound on a cover. Let $\mathcal{N}_1 \subset \mathcal{V} \subset \mathbb{R}^d$ and $\mathcal{N}_2 \subset \mathscr{F}$ denote fixed (deterministic), finite sets whose product $\mathcal{N} = \mathcal{N}_1 \times \mathcal{N}_2$ has cardinality at most $|\mathcal{A}_{\mathsf{rnd}}| \cdot |\mathcal{N}| \leq M$. We use Lemma 5.1 to establish a uniform bound on the errors $|\mathsf{cat}[\widetilde{\boldsymbol{v}}, f, \tilde{\alpha}] - \zeta[\widetilde{\boldsymbol{v}}, f, \tilde{\alpha}]|$ of the Catoni esimator corresponding to pairs $(\widetilde{\boldsymbol{v}}, f) \in \mathcal{N}$ and $\tilde{\alpha} \in \mathcal{A}_{\mathsf{rnd}}$. To do this, we have to be somewhat careful, because we require that conditional variances are upper bounded by σ_t^2. To this end, we argue a bound on the Catoni error when the following *random event* holds:

$$\mathcal{E}_f := \{\mathbb{E}\left[\widetilde{y}_t(f)^2 \mid \mathcal{F}_{t-1}\right] \leq \sigma_t^2, \ \forall t\},$$

Note that $\mathcal{E}_f$ is indeed random because σ_t^2 are random. Using linearity of expectation, we have

$$\begin{aligned} \mathbb{E}[\widetilde{y}_t(f)|\mathcal{F}_{t-1}] &= \langle \boldsymbol{\phi}_t, \boldsymbol{u}_\star \rangle + \left\langle \boldsymbol{\phi}_t, \int f(\boldsymbol{\phi}') \mathrm{d}\boldsymbol{\mu}(\boldsymbol{\phi}') \right\rangle \\ &= \left\langle \boldsymbol{\phi}_t, \boldsymbol{u}_\star + \int f(\boldsymbol{\phi}') \mathrm{d}\boldsymbol{\mu}(\boldsymbol{\phi}') \right\rangle \\ &= \left\langle \boldsymbol{\phi}_t, \tilde{\boldsymbol{\theta}}(f) \right\rangle. \end{aligned}$$

so the linearity of expectation assumption required by Lemma 5.1 will be met. Hence, for all pairs $(\widetilde{\boldsymbol{v}}, f) \in \mathcal{N}$ such that $\mathcal{E}_f$ holds, and all $\tilde{\alpha} \in \mathcal{A}_{\mathsf{rnd}}$ such that

- $\tilde{\alpha} \geq \alpha[\widetilde{\boldsymbol{v}}]$

- $\tilde{\alpha}$ can be expressed as $\min\left\{\tilde{\gamma}\cdot\frac{\sqrt{\log\frac{2M}{\delta}}}{\|\widetilde{\boldsymbol{v}}\|_{\boldsymbol{\Sigma}_T}}, \alpha_{\max}\right\}$
- $T \geq (2+2\tilde{\gamma})\log\frac{2M}{\delta}$.

then it holds that with probability $1-\frac{\delta}{M}$,

$$|\mathsf{cat}[\widetilde{\boldsymbol{v}}, f, \tilde{\alpha}] - \zeta[\widetilde{\boldsymbol{v}}, f]| \leq (2+2\tilde{\gamma})\|\widetilde{\boldsymbol{v}}\|_{\boldsymbol{\Sigma}_T}\sqrt{\frac{\log\frac{2M}{\delta}}{T^2}} + \frac{2\log\frac{2}{M\delta}}{\alpha_{\max}T}, \quad \text{whenever } \mathcal{E}_f \text{ holds.}$$

In particular, selecting $\tilde{\alpha} = \alpha_{\mathsf{rnd}}[\widetilde{\boldsymbol{v}}]$, we can choose $\tilde{\gamma} = 1+\epsilon_{\mathsf{rnd}}$. As ϵ_{rnd} was chosen such that $\epsilon_{\mathsf{rnd}} \leq 1/4$, $2+2\tilde{\gamma} \leq 5$. Hence, we find that with probability at least $1-\delta$,

$$|\mathsf{cat}_{\mathsf{rnd}}[\widetilde{\boldsymbol{v}}, f] - \zeta[\widetilde{\boldsymbol{v}}, f]| \leq 5\|\widetilde{\boldsymbol{v}}\|_{\boldsymbol{\Sigma}_T}\sqrt{\frac{\log\frac{2M}{\delta}}{T^2}} + \frac{2\log\frac{2M}{\delta}}{\alpha_{\max}T}, \quad \forall (f,\widetilde{\boldsymbol{v}}) \in \mathcal{N} \text{ such that } \mathcal{E}_f \text{ holds,}$$

provided that $T \geq (2+2\tilde{\gamma}^2)\log\frac{2M}{\delta}$. Given our setting of $\tilde{\gamma}$, it suffices to take $T \geq 6\log\frac{2M}{\delta}$.

Approximation by covering. Having achieved a pointwise bound, we observe that, on the $1-\delta$ event above, for any $(\widetilde{\boldsymbol{v}}, f) \in \mathcal{V}\times\mathscr{F}$, and any $(\widetilde{\boldsymbol{v}}_0, f_0) \in \mathcal{N}$ for which $\mathcal{E}_{f_0}$ holds,

$$\begin{aligned}
&|\mathsf{cat}[\widetilde{\boldsymbol{v}}, f] - \zeta[\widetilde{\boldsymbol{v}}, f]| \\
&\quad\leq |\mathsf{cat}[\widetilde{\boldsymbol{v}}, f] - \mathsf{cat}_{\mathsf{rnd}}[\widetilde{\boldsymbol{v}}_0, f_0]| + |\zeta[\widetilde{\boldsymbol{v}}, f] - \zeta[\widetilde{\boldsymbol{v}}_0, f_0]| + |\mathsf{cat}_{\mathsf{rnd}}[\widetilde{\boldsymbol{v}}_0, f_0] - \zeta[\widetilde{\boldsymbol{v}}_0, f_0]| \\
&\quad\leq |\mathsf{cat}[\widetilde{\boldsymbol{v}}, f] - \mathsf{cat}_{\mathsf{rnd}}[\widetilde{\boldsymbol{v}}_0, f_0]| + |\zeta[\widetilde{\boldsymbol{v}}, f] - \zeta[\widetilde{\boldsymbol{v}}_0, f_0]| + 5\|\widetilde{\boldsymbol{v}}_0\|_{\boldsymbol{\Sigma}_T}\sqrt{\frac{\log\frac{2M}{\delta}}{T^2}} + \frac{2\log\frac{2M}{\delta}}{\alpha_{\max}T} \\
&\quad\leq 5\|\widetilde{\boldsymbol{v}}\|_{\boldsymbol{\Sigma}_T}\sqrt{\frac{\log\frac{2M}{\delta}}{T^2}} + \frac{2\log\frac{2M}{\delta}}{\alpha_{\max}T} + \underbrace{5\sqrt{\frac{\log\frac{2M}{\delta}}{T^2}}\cdot\left|\|\widetilde{\boldsymbol{v}}\|_{\boldsymbol{\Sigma}_T} - \|\widetilde{\boldsymbol{v}}_0\|_{\boldsymbol{\Sigma}_T}\right|}_{(i)} \\
&\qquad + \underbrace{|\zeta[\widetilde{\boldsymbol{v}}, f] - \zeta[\widetilde{\boldsymbol{v}}_0, f_0]|}_{(ii)} + \underbrace{|\mathsf{cat}[\widetilde{\boldsymbol{v}}, f] - \mathsf{cat}_{\mathsf{rnd}}[\widetilde{\boldsymbol{v}}_0, f_0]|}_{(iii)}
\end{aligned} \tag{A.9}$$

Recall $\beta_u = \|\boldsymbol{u}_\star\|_2$, $\beta_\mu = \||\boldsymbol{\mu}|(\mathcal{B}^{d'})\|_2$, $\sigma_{\min}^2 \leq \sigma_t^2$, and $\mathrm{dist}_\infty(f, f_0) := \sup_{\boldsymbol{\phi}'\in\mathcal{B}_{d'}}|f(\boldsymbol{\phi}') - f_0(\boldsymbol{\phi}')|$. We further assume that $\|\widetilde{\boldsymbol{v}}_0\|_2, \|\widetilde{\boldsymbol{v}}\|_2 \leq \beta_{\tilde{v}}$. We show that, for these scalings, it suffices to ensure that $\|\widetilde{\boldsymbol{v}} - \widetilde{\boldsymbol{v}}_0\|_2$ and $\mathrm{dist}_\infty(f, f_0)$ are at most polynomial in relevant problem parameters:

Lemma A.6. *There exists a constant* $\mathcal{C}_{\mathrm{poly}} = \mathrm{poly}(\sigma_{\min}^{-2}, T, \beta_{\tilde{v}}, \alpha_{\max}, H, \beta_u, \beta_\mu, \beta_\eta)$ *such that, if*

$$\max\{\|\widetilde{\boldsymbol{v}} - \widetilde{\boldsymbol{v}}_0\|_2, \mathrm{dist}_\infty(f, f_0), \epsilon_{\mathsf{rnd}}\} \leq 1/\mathcal{C}_{\mathrm{poly}},$$

then

$$\text{Term } (i) + \text{Term } (ii) + \text{Term } (iii) \leq \frac{\log\frac{2M}{\delta}}{\alpha_{\max}T}.$$

We defer the proof of Lemma A.6 to the end of the section. In addition, we show that if $\mathrm{dist}_\infty(f, f_0)$ is sufficiently small, then the element f_0 from the covering satisfies the desired variance upper bound:

Lemma A.7. *Suppose that f satisfies the variance bound in Equation (5.4), that is,*

$$\mathbb{E}[\big(\langle \boldsymbol{\phi}_t, \boldsymbol{u}_\star\rangle + f(\boldsymbol{\phi}'_t) + \eta_t\big)^2 \mid \mathcal{F}_{t-1}] \le \frac{1}{2}\sigma_t^2. \tag{A.10}$$

Then, if $\mathrm{dist}_\infty(f, f_0) \le 1/\mathcal{C}_{\mathrm{poly}}$ for an appropriate choice of $\mathcal{C}_{\mathrm{poly}}$ as in Lemma A.6, $\mathcal{E}_{f_0}$ holds, i.e.,

$$\mathbb{E}[\big(\langle \boldsymbol{\phi}_t, \boldsymbol{u}_\star\rangle + f_0(\boldsymbol{\phi}'_t) + \eta_t\big)^2 \mid \mathcal{F}_{t-1}] \le \sigma_t^2.$$

Proof. For any f satisfying Equation (A.10),

$$\begin{aligned}\mathbb{E}[\big(\langle \boldsymbol{\phi}_t, \boldsymbol{u}_\star\rangle + f_0(\boldsymbol{\phi}'_t) + \eta_t\big)^2 \mid \mathcal{F}_{t-1}] &\le 2\mathbb{E}[\big(\langle \boldsymbol{\phi}_t, \boldsymbol{u}_\star\rangle + f(\boldsymbol{\phi}'_t) + \eta_t\big)^2 \mid \mathcal{F}_{t-1}] + 2\mathbb{E}[\big(f(\boldsymbol{\phi}'_t) - f_0(\boldsymbol{\phi}'_t)\big)^2 \mid \mathcal{F}_{t-1}] \\ &\le \frac{1}{2}\sigma_t^2 + 2\mathrm{dist}_\infty(f, f_0)^2.\end{aligned}$$

Since $\sigma_t^2 \ge \sigma_{\min}^2$, it is enough that $\mathrm{dist}_\infty(f, f_0)^2 \le \frac{\sigma_{\min}^2}{2}$, which is ensured by an appropriate choice of $\mathcal{C}_{\mathrm{poly}}$. □

Concluding the proof Let us summarize our current findings. We see that if $\mathcal{N} \subset \mathcal{V} \times \mathscr{F}$ is a collection of pairs $(\widetilde{\boldsymbol{v}}_0, f_0)$ satisfying

- The cardinality bound $|\mathcal{A}_{\mathsf{rnd}}| \cdot |\mathcal{N}| \le M$
- The approximation bound that,

  $$\forall \widetilde{\boldsymbol{v}} \in \mathcal{V}, f \in \mathscr{F}, \quad \exists (\widetilde{\boldsymbol{v}}_0, f_0) \in \mathcal{N} \text{ such that } \max\{\|\widetilde{\boldsymbol{v}} - \widetilde{\boldsymbol{v}}_0\|_2, \mathrm{dist}_\infty(f, f_0)\} \le 1/\mathcal{C}_{\mathrm{poly}},$$

  and that $\epsilon_{\mathsf{rnd}} \le 1/\mathcal{C}_{\mathrm{poly}}$, where again

  $$\mathcal{C}_{\mathrm{poly}} = \mathrm{poly}(\sigma_{\min}^{-2}, T, \beta_{\tilde{v}}, \alpha_{\max}, H, \beta_\mu, \beta_u, \beta_\eta) = \mathrm{poly}(\sigma_{\min}^{-2}, T, 1/\lambda, \alpha_{\max}, H, \beta_u, \beta_\mu, \beta_\eta).$$

Then, Equation (A.9), the fact that $f_\star$ satisfies Equation (A.10), and Lemmas A.6 and A.7 imply that with probability $1-\delta$,

$$\forall \widetilde{\boldsymbol{v}} \in \mathcal{V}, \quad |\mathsf{cat}(\widetilde{\boldsymbol{v}}, f_\star) - \zeta(\widetilde{\boldsymbol{v}}, f_\star)| \le 5\|\widetilde{\boldsymbol{v}}\|_{\boldsymbol{\Sigma}_T}\sqrt{\frac{\log\frac{2M}{\delta}}{T^2}} + \frac{3\log\frac{2M}{\delta}}{\alpha_{\max}T},$$

provided that $T \ge 6\log\frac{2M}{\delta}$. We now find an M sufficiently large to ensure the covering conditions hold. To this end, it suffices to ensure that $\mathcal{N} = \mathcal{N}_1 \times \mathcal{N}_2$, where $\mathcal{N}_1$ is an $\epsilon = 1/\mathcal{C}_{\mathrm{poly}}$ net of $\mathcal{V}$, and $\mathcal{N}_2$ is an $\epsilon = 1/\mathcal{C}_{\mathrm{poly}}$ net of $\mathscr{F}$ in the norm $\mathrm{dist}_\infty(\cdot,\cdot)$. By Lemma A.1 and the fact that $\mathcal{V}$ is a Euclidean ball of radius $\beta_{\tilde{v}} = 1/\lambda$, it suffices to take

$$\log|\mathcal{N}_1| \le d\log(1 + \frac{2\mathcal{C}_{\mathrm{poly}}}{\lambda}).$$

Similarly, by assumption that the covering numbers of $\mathscr{F}$ are $\mathsf{N}(\mathscr{F}, \mathrm{dist}_\infty(\cdot,\cdot), \epsilon) \le p\log(1 + \frac{R}{\epsilon})$,

$$\log|\mathcal{N}_2| \le p\log(1 + 2R\mathcal{C}_{\mathrm{poly}}).$$

Finally, using the bound on $|\mathcal{A}_{\mathsf{rnd}}|$ from Equation (A.8),

$$\log|\mathcal{A}_{\mathsf{rnd}}| \le \log(1 + \frac{2T\alpha_{\max}}{\lambda\epsilon_{\mathsf{rnd}}}) \le \log(1 + \frac{2\mathcal{C}_{\mathrm{poly}}T\alpha_{\max}}{\lambda}).$$

Hence, we can bound, for universal constants $c', c > 0$,

$$\begin{aligned}\log\frac{2M}{\delta} &= \log\frac{2|\mathcal{A}_{\mathsf{rnd}}||\mathcal{N}_1||\mathcal{N}_2|}{\delta} \\ &\leq \log\frac{1}{\delta} + \log 2 + (p+d)\cdot c' \,\mathsf{logs}\left(R, T, \alpha_{\max}, \lambda^{-1}, \mathcal{C}_{\mathrm{poly}}\right) \\ &= \log\frac{1}{\delta} + \underbrace{(p+d)\cdot c\cdot \mathsf{logs}\left(\frac{1}{\delta}, T, R, \lambda^{-1}, \sigma_{\min}^{-2}, \alpha_{\max}^2, H, \beta_\mu, \beta_u, \beta_\eta\right)}_{:=d_T}.\end{aligned}$$

For this choice of M,

$$\forall \widetilde{\boldsymbol{v}} \in \mathcal{V}, \quad |\mathsf{cat}(\widetilde{\boldsymbol{v}}, f_\star) - \zeta(\widetilde{\boldsymbol{v}}, f_\star)| \leq 5\|\widetilde{\boldsymbol{v}}\|_{\boldsymbol{\Sigma}_T}\sqrt{\frac{\log\frac{2}{\delta} + d_T}{T^2}} + \frac{3\log\frac{2M}{\delta}}{\alpha_{\max}T},$$

which, returning to the orginal notation parameterized by $(\boldsymbol{v}, \boldsymbol{\Lambda})$ and noting the equivalence of notation in Equation (A.5), we see that with probability $1-\delta$, it holds that $\forall \boldsymbol{v} \in \mathcal{B}^d$ and $\boldsymbol{\Lambda} \succeq \lambda I$ (ensuring $\boldsymbol{\Lambda}^{-1}\boldsymbol{v} \in \mathcal{V}$)

$$|\mathsf{cat}[\boldsymbol{\Lambda}^{-1}\boldsymbol{v}] - \zeta[\boldsymbol{\Lambda}^{-1}\boldsymbol{v}]| \lesssim \|\boldsymbol{\Lambda}^{-1}\boldsymbol{v}\|_{\boldsymbol{\Sigma}_T}\sqrt{\frac{\log\frac{1}{\delta} + d_T}{T^2}} + \frac{\log\frac{1}{\delta} + d_T}{\alpha_{\max}T}.$$

□

A.3.1 Proofs supporting Proposition 7

Proof of Lemma A.6. Recall $\beta_u = \|\boldsymbol{u}_\star\|_2$, $\beta_\mu = \||\boldsymbol{\mu}|(\mathcal{B}^{d'})\|_2$, $\sigma_{\min}^2 \leq \sigma_t^2$, $\beta_\eta \geq |\eta_t|$ with probability 1, and $\mathrm{dist}_\infty(f, f_0) := \sup_{\boldsymbol{\phi}' \in \mathcal{B}_{d'}} |f(\boldsymbol{\phi}') - f_0(\boldsymbol{\phi}')|$. We further assume that $\widetilde{\boldsymbol{v}}_0, \widetilde{\boldsymbol{v}} \in \mathcal{V}$, i.e.

$$\|\widetilde{\boldsymbol{v}}_0\|_2, \|\widetilde{\boldsymbol{v}}\|_2 \leq \beta_{\tilde{v}},$$

and that we may choose $\mathcal{C}_{\mathrm{poly}} = \mathrm{poly}(\sigma_{\min}^{-2}, T, \beta_{\tilde{v}}, \alpha_{\max}, H, \beta_u, \beta_\mu, \beta_\eta)$ to be an aritrary polynomial in these quantities. We move term by term, showing we can make each at most $\frac{\log(2M/\delta)}{3\alpha_{\max}T}$ by selecting $\mathcal{C}_{\mathrm{poly}}$ appropriately. Throughout, we use the fact that, for $\delta \in (0, 1/2)$ and $M \geq 1$, $\log(2M/\delta) \geq 1$.

Claim A.8 (Bounding Term (i))**.** Term (i) *is at most* $\sqrt{\frac{\log(2M/\delta)}{T\sigma_{\min}^2}} \cdot \|\widetilde{\boldsymbol{v}} - \widetilde{\boldsymbol{v}}_0\|_2$. *Hence, for an appropriate choice of* $\mathcal{C}_{\mathrm{poly}}$*, the above is at most* Term (i) *is at most* $\frac{\sqrt{\log(2M/\delta)}}{3\alpha_{\max}T} \leq \frac{\log(2M/\delta)}{3\alpha_{\max}T}$.

Proof. We have

$$\begin{aligned}\text{Term } (i) &:= 5\sqrt{\frac{\log\frac{2M}{\delta}\cdot}{T^2}} \cdot |\|\widetilde{\boldsymbol{v}}\|_{\boldsymbol{\Sigma}_T} - \|\widetilde{\boldsymbol{v}}_0\|_{\boldsymbol{\Sigma}_T}| \\ &\leq 5\sqrt{\frac{\log\frac{2M}{\delta}\cdot}{T^2}} \cdot \sqrt{\|\boldsymbol{\Sigma}_T\|_{\mathrm{op}}} \cdot \|\widetilde{\boldsymbol{v}} - \widetilde{\boldsymbol{v}}_0\|_2.\end{aligned}$$

Since $\|\boldsymbol{\phi}_t\| \leq 1$ and $\sigma_t \geq \sigma_{\min}$ by assumption, $\|\boldsymbol{\Sigma}_T\|_{\mathrm{op}} = \|\sum_{t=1}^T \sigma_t^{-2}\boldsymbol{\phi}_t\boldsymbol{\phi}_t^\top\|_{\mathrm{op}} \leq T\sigma_{\min}^{-2}$. The bound follows. □

Claim A.9 (Bounding Term (*ii*))**.** *We can bound*

$$\text{Term (ii)} := |\zeta[\widetilde{\boldsymbol{v}}, f] - \zeta[\widetilde{\boldsymbol{v}}_0, f_0]| \leq \frac{1}{\sigma_{\min}^2}\left(\beta_\mu \beta_{\tilde{v}} \cdot \mathrm{dist}_\infty(f, f_0) + (\beta_u + H\beta_\mu)\|\widetilde{\boldsymbol{v}} - \widetilde{\boldsymbol{v}}_0\|_2\right). \tag{A.11}$$

Hence, the appropriate choice of $\mathcal{C}_{\mathrm{poly}}$ *ensures* Term (*ii*) *is at most* $\frac{\log(2M/\delta)}{3\alpha_{\max}T}$.

Proof. We expand

$$\begin{aligned}\zeta[\widetilde{\boldsymbol{v}}, f] - \zeta[\widetilde{\boldsymbol{v}}_0, f_0] &= \frac{1}{T}\left(\widetilde{\boldsymbol{v}}^\top \boldsymbol{\Sigma}_T \tilde{\boldsymbol{\theta}}(f) - \widetilde{\boldsymbol{v}}_0^\top \boldsymbol{\Sigma}_T \tilde{\boldsymbol{\theta}}(f_0)\right) \\ &= \frac{1}{T}\widetilde{\boldsymbol{v}}^\top \boldsymbol{\Sigma}_T\left(\tilde{\boldsymbol{\theta}}(f) - \tilde{\boldsymbol{\theta}}(f_0)\right) + \frac{1}{T}\left(\widetilde{\boldsymbol{v}} - \widetilde{\boldsymbol{v}}_0\right)^\top \boldsymbol{\Sigma}_T \tilde{\boldsymbol{\theta}}(f_0).\end{aligned}$$

Hence, using the bound $\|\boldsymbol{\Sigma}_T\|_{\mathrm{op}} \leq T/\sigma_{\min}^2$ developed above,

$$|\zeta[\widetilde{\boldsymbol{v}}, f] - \zeta[\widetilde{\boldsymbol{v}}_0, f_0]| \leq \frac{1}{\sigma_{\min}^2}(\|\widetilde{\boldsymbol{v}}\||\tilde{\boldsymbol{\theta}}(f) - \tilde{\boldsymbol{\theta}}(f_0)| + |\tilde{\boldsymbol{\theta}}(f_0)|\|\widetilde{\boldsymbol{v}} - \widetilde{\boldsymbol{v}}_0\|).$$

Note that $\|\widetilde{\boldsymbol{v}}\| \leq \beta_{\tilde{v}}$ by assumption. Further, we bound

$$\|\tilde{\boldsymbol{\theta}}(f) - \tilde{\boldsymbol{\theta}}(f_0)\|_2 = \left\|\int (f - f_0)(\boldsymbol{\phi}')\mathrm{d}\boldsymbol{\mu}(\boldsymbol{\phi}')\right\| \leq \||\boldsymbol{\mu}|(\mathcal{B}_{d'})\|_2 \cdot \|f - f_0\|_{\mathcal{L}_\infty(\mathcal{B}_{d'})} \leq \beta_\mu \mathrm{dist}_\infty(f, f_0). \tag{A.12}$$

and moreover,

$$\|\tilde{\boldsymbol{\theta}}(f_0)\|_2 \leq \|\boldsymbol{u}_\star\|_2 + \left\|\int f_0(\boldsymbol{\phi}')\mathrm{d}\boldsymbol{\mu}(\boldsymbol{\phi}')\right\| \leq \beta_u + \||\boldsymbol{\mu}|(\mathcal{B}_{d'})\|_2 \max_{\boldsymbol{\phi}' \in \mathcal{B}_{d'}} |f_0(\boldsymbol{\phi}')| \leq \beta_u + H\beta_\mu. \tag{A.13}$$

Combining the bounds concludes the proof. □

Claim A.10 (Bounding Term (*iii*))**.** *An appropriate choice of* $\mathcal{C}_{\mathrm{poly}}$ *ensures* Term (*iii*) *is at most* $\frac{\log(2M/\delta)}{3\alpha_{\max}T}$.

Proof. Recall that Term (*iii*) $= |\mathsf{cat}[\widetilde{\boldsymbol{v}}, f] - \mathsf{cat}_{\mathsf{rnd}}[\widetilde{\boldsymbol{v}}_0, f_0]|$. Recall that $\mathsf{cat}[\widetilde{\boldsymbol{v}}, f]$ uses the data and parameter

$$X_t[\widetilde{\boldsymbol{v}}, f] := \frac{1}{\sigma_t^2}\widetilde{\boldsymbol{v}}^\top \boldsymbol{\phi}_t \widetilde{y}_t(f) \quad \alpha[\widetilde{\boldsymbol{v}}] = \min\left\{\frac{\sqrt{\log 2M/\delta}}{\|\widetilde{\boldsymbol{v}}\|_{\boldsymbol{\Sigma}_T}}, \alpha_{\max}\right\},$$

whereas $\mathsf{cat}_{\mathsf{rnd}}[\widetilde{\boldsymbol{v}}_0, f_0]$ uses the same X_t, replaced with $\widetilde{\boldsymbol{v}}_0$ and f_0, and uses the rounded version $\alpha_{\mathsf{rnd}}[\widetilde{\boldsymbol{v}}]$. To compute the sensitivity bound, we consider differences between various quantities of interest. Throughout, we use

$$|\widetilde{y}_t(f)| := |\langle \boldsymbol{\phi}_t, \boldsymbol{u}_\star\rangle + f(\boldsymbol{\phi}_t') + \eta_t| \leq \beta_u + H + \beta_\eta$$

Difference in scalar data. We have

$$\begin{aligned}|X_t[\widetilde{\boldsymbol{v}}, f] - X_t[\widetilde{\boldsymbol{v}}_0, f_0]| &= \left|\frac{1}{\sigma_t^2}\widetilde{\boldsymbol{v}}^\top \boldsymbol{\phi}_t \widetilde{y}_t(f) - \frac{1}{\sigma_t^2}\widetilde{\boldsymbol{v}}_0^\top \boldsymbol{\phi}_t \widetilde{y}_t(f_0)\right| \\ &\leq \frac{\|\boldsymbol{\phi}_t\|_2}{\sigma_t^2} \cdot (|\widetilde{y}_t(f_0)|\|\widetilde{\boldsymbol{v}} - \widetilde{\boldsymbol{v}}_0\|_2 + \|\widetilde{\boldsymbol{v}}\|_2|\widetilde{y}_t(f) - \widetilde{y}_t(f_0)|) \\ &\leq \frac{1}{\sigma_{\min}^2} \cdot \left(H\|\widetilde{\boldsymbol{v}} - \widetilde{\boldsymbol{v}}_0\|_2 + \|\widetilde{\boldsymbol{v}}\|_2|f(\boldsymbol{\phi}_t') - f_0(\boldsymbol{\phi}_t')|\right) \\ &\leq \epsilon_X := \frac{1}{\sigma_{\min}^2} \cdot ((\beta_u + H + \beta_\eta)\|\widetilde{\boldsymbol{v}} - \widetilde{\boldsymbol{v}}_0\|_2 + \beta_{\tilde{v}}\mathrm{dist}_\infty(f, f_0)).\end{aligned}$$

Difference in Catoni parameters. Setting $c = \sqrt{\log 2M/\delta}$ and $a(\widetilde{\boldsymbol{v}}) := \|\widetilde{\boldsymbol{v}}\|_{\boldsymbol{\Sigma}_T}$, we can express

$$\alpha[\widetilde{\boldsymbol{v}}] = \min\left\{\frac{c}{a(\widetilde{\boldsymbol{v}})}, \alpha_{\max}\right\} = \frac{c}{\max\{a(\widetilde{\boldsymbol{v}}), c/\alpha_{\max}\}}.$$

This gives that the difference between the unrounded parameter with $\widetilde{\boldsymbol{v}}$, $\alpha[\widetilde{\boldsymbol{v}}]$, and the *also unrounded* parameter $\alpha[\widetilde{\boldsymbol{v}}_0]$ with $\widetilde{\boldsymbol{v}}_0$ are bounded as

$$\begin{aligned}
|\alpha[\widetilde{\boldsymbol{v}}] - \alpha[\widetilde{\boldsymbol{v}}_0]| &= c\left|\frac{1}{\max\{a(\widetilde{\boldsymbol{v}}), c/\alpha_{\max}\}} - \frac{1}{\max\{a(\widetilde{\boldsymbol{v}}), c/\alpha_{\max}\}}\right| \\
&= c \cdot \left|\frac{\max\{a(\widetilde{\boldsymbol{v}}), \frac{c}{\alpha_{\max}}\} - \max\{a(\widetilde{\boldsymbol{v}}_0), \frac{c}{\alpha_{\max}}\}}{\max\{a(\widetilde{\boldsymbol{v}}), \frac{c}{\alpha_{\max}}\} \cdot \max\{a(\widetilde{\boldsymbol{v}}_0), \frac{c}{\alpha_{\max}}\}}\right| \\
&\le c\frac{|a(\widetilde{\boldsymbol{v}}) - a(\widetilde{\boldsymbol{v}}_0)|}{\frac{c^2}{\alpha_{\max}^2}} = \frac{\alpha_{\max}^2|a(\widetilde{\boldsymbol{v}}) - a(\widetilde{\boldsymbol{v}}_0)|}{c} \\
&\le \alpha_{\max}^2\sqrt{\sigma_{\min}^{-2}T} \cdot \|\widetilde{\boldsymbol{v}} - \widetilde{\boldsymbol{v}}_0\|_2
\end{aligned}$$

where the last line uses the definition of c, and the argument of Claim A.8 to bound $|\alpha[\widetilde{\boldsymbol{v}}] - \alpha[\widetilde{\boldsymbol{v}}_0]|$, as well as $\log(2M/\delta) \le 1$.

Note however the $\mathsf{cat}_{\mathsf{rnd}}[\widetilde{\boldsymbol{v}}_0, f_0]$ uses the *rounded* parameter $\alpha_{\mathsf{rnd}}[\widetilde{\boldsymbol{v}}_0]$. Directly from its definition, we can see that

$$\alpha[\widetilde{\boldsymbol{v}}_0] \le \alpha_{\mathsf{rnd}}[\widetilde{\boldsymbol{v}}_0] \le (1 + \epsilon_{\mathsf{rnd}})\alpha[\widetilde{\boldsymbol{v}}_0],$$

so that

$$|\alpha[\widetilde{\boldsymbol{v}}_0] - \alpha_{\mathsf{rnd}}[\widetilde{\boldsymbol{v}}_0]| \le \epsilon_{\mathsf{rnd}}\alpha[\widetilde{\boldsymbol{v}}_0] \le \epsilon_{\mathsf{rnd}}\alpha_{\max}.$$

By the triangle inequality, we therefore conclude

$$\begin{aligned}
|\alpha[\widetilde{\boldsymbol{v}}] - \alpha_{\mathsf{rnd}}[\widetilde{\boldsymbol{v}}_0]| &\le |\alpha[\widetilde{\boldsymbol{v}}] - \alpha[\widetilde{\boldsymbol{v}}_0]| + |\alpha[\widetilde{\boldsymbol{v}}_0] - \alpha_{\mathsf{rnd}}[\widetilde{\boldsymbol{v}}_0]| \\
&\le \underbrace{\alpha_{\max}^2\sqrt{\sigma_{\min}^{-2}T} \cdot \|\widetilde{\boldsymbol{v}} - \widetilde{\boldsymbol{v}}_0\|_2 + \epsilon_{\mathsf{rnd}}\alpha_{\max}}_{:=\epsilon_\alpha}.
\end{aligned}$$

Upper bounding data norms and lower bound $\alpha[\widetilde{\boldsymbol{v}}]$. We have

$$\begin{aligned}
\max\{|X_t[\widetilde{\boldsymbol{v}}, f]|, |X_t[\widetilde{\boldsymbol{v}}_0, f_0]|\} &\le \sigma_t^{-2}\|\boldsymbol{\phi}\|_2 \max\{\|\widetilde{\boldsymbol{v}}\|_2, \|\widetilde{\boldsymbol{v}}_0\|\} \cdot \max\{|\widetilde{y}_t(f), \widetilde{y}_f(f_0)|\} \\
&\le \sigma_{\min}^{-2}\beta_{\tilde{v}}(H + \beta_u + \beta_\eta) := \gamma_X.
\end{aligned}$$

and, upper bounding $\|\widetilde{\boldsymbol{v}}\|_{\boldsymbol{\Sigma}_T}^2 \le \beta_{\tilde{v}}^2 T/\sigma_{\min}^2$,

$$\begin{aligned}
\alpha[\widetilde{\boldsymbol{v}}] = \min\left\{\frac{\sqrt{\log 2M/\delta}}{\|\widetilde{\boldsymbol{v}}\|_{\boldsymbol{\Sigma}_T}}, \alpha_{\max}\right\} &\ge \min\left\{\sqrt{\frac{\log 2M/\delta}{\beta_{\tilde{v}}^2 T/\sigma_{\min}^2}}, \alpha_{\max}\right\} \\
&\ge \alpha_- := \min\left\{\frac{1}{\beta_{\tilde{v}}\sqrt{T\sigma_{\min}^2}}, \alpha_{\max}\right\},
\end{aligned}$$

We now invoke a perturbation bound for the Catoni estimator (Lemma A.13), which ensures that, as long as

$$\epsilon := \alpha(\widetilde{\boldsymbol{v}})\epsilon_X + 3\gamma_X\epsilon_\alpha \le \frac{1}{18}\min\left\{1, \alpha(\widetilde{\boldsymbol{v}})^2\gamma^2\right\},$$

for which it suffices that

$$\alpha_{\max}\epsilon_X + 3\gamma_X\epsilon_\alpha \le \frac{1}{18}\min\left\{1, \frac{\gamma_X^2}{\alpha_{\max}^2}, \sigma_{\min}^{-2}(H+\beta_u)^2/T\right\},$$

we will have

$$|z^* - \widetilde{z}^*| \le \frac{1+2\alpha(\widetilde{\boldsymbol{v}})\gamma}{\alpha(\widetilde{\boldsymbol{v}})}\epsilon + \sqrt{\frac{2\epsilon}{\alpha(\widetilde{\boldsymbol{v}})^2}} \le \frac{1+2\alpha_{\max}\gamma}{\alpha_-}\epsilon + \sqrt{\frac{2\epsilon}{\alpha_-}}$$

Examining the above bounds, we have that $|\mathsf{cat}(\widetilde{\boldsymbol{v}}, f) - \mathsf{cat}(\widetilde{\boldsymbol{v}}_0, f_0)| \le \epsilon_0$ provided

$$\max\{\|\widetilde{\boldsymbol{v}} - \widetilde{\boldsymbol{v}}_0\|_2, \mathrm{dist}_\infty(f, f_0), \epsilon_{\mathsf{rnd}}\} \le \epsilon_0^2 \cdot 1/\mathrm{poly}(\sigma_{\min}^{-2}, T, \beta_{\tilde{v}}, \alpha_{\max}, H, \beta_u, \beta_\eta), \quad \epsilon_0 \le 1. \tag{A.14}$$

The bound follows by taking ϵ_0 to be $\frac{1}{3\alpha_{\max}T} \le \frac{\log(2M/\delta)}{3\alpha_{\max}T}$. □

□

A.4 Linear Approximation to Catoni

Proof of Lemma 5.2. By definition of $\widehat{\boldsymbol{\theta}}$ and (5.6), we will have that

$$\begin{aligned}
\sup_{\boldsymbol{v}\in\mathcal{V}} \frac{|\langle\widehat{\boldsymbol{\theta}}, \boldsymbol{v}\rangle - \mathsf{cat}[\boldsymbol{\Lambda}^{-1}\boldsymbol{v}]|}{\|\boldsymbol{v}\|_{\boldsymbol{\Lambda}^{-1}}} &= \min_{\boldsymbol{\theta}} \sup_{\boldsymbol{v}\in\mathcal{V}} \frac{|\langle\boldsymbol{\theta}, \boldsymbol{v}\rangle - \mathsf{cat}[\boldsymbol{\Lambda}^{-1}\boldsymbol{v}]|}{\|\boldsymbol{v}\|_{\boldsymbol{\Lambda}^{-1}}} \\
&\le \sup_{\boldsymbol{v}\in\mathcal{V}} \frac{|\langle\boldsymbol{\theta}_\star, \boldsymbol{v}\rangle - \mathsf{cat}[\boldsymbol{\Lambda}^{-1}\boldsymbol{v}]|}{\|\boldsymbol{v}\|_{\boldsymbol{\Lambda}^{-1}}} \\
&\le C_1 + \frac{C_2}{T\|\boldsymbol{v}\|_{\boldsymbol{\Lambda}^{-1}}}.
\end{aligned}$$

Rearranging this implies that for all $\boldsymbol{v} \in \mathcal{V}$,

$$\frac{|\langle\widehat{\boldsymbol{\theta}}, \boldsymbol{v}\rangle - \mathsf{cat}[\boldsymbol{\Lambda}^{-1}\boldsymbol{v}]|}{\|\boldsymbol{v}\|_{\boldsymbol{\Lambda}^{-1}}} \le C_1 + \frac{C_2}{T\|\boldsymbol{v}\|_{\boldsymbol{\Lambda}^{-1}}} \iff |\langle\widehat{\boldsymbol{\theta}}, \boldsymbol{v}\rangle - \mathsf{cat}[\boldsymbol{\Lambda}^{-1}\boldsymbol{v}]| \le C_1\|\boldsymbol{v}\|_{\boldsymbol{\Lambda}^{-1}} + \frac{C_2}{T}.$$

Similarly,

$$\sup_{\boldsymbol{v}\in\mathcal{V}} \frac{|\langle\widehat{\boldsymbol{\theta}}, \boldsymbol{v}\rangle - \zeta[\boldsymbol{\Lambda}^{-1}\boldsymbol{v}]|}{\|\boldsymbol{v}\|_{\Lambda^{-1}}} \le \sup_{\boldsymbol{v}\in\mathcal{V}} \frac{|\langle\widehat{\boldsymbol{\theta}}, \boldsymbol{v}\rangle - \mathsf{cat}[\boldsymbol{\Lambda}^{-1}\boldsymbol{v}]| + |\mathsf{cat}[\boldsymbol{\Lambda}^{-1}\boldsymbol{v}] - \zeta[\boldsymbol{\Lambda}^{-1}\boldsymbol{v}]|}{\|\boldsymbol{v}\|_{\boldsymbol{\Lambda}^{-1}}} \le 2C_1 + \frac{2C_2}{T\|\boldsymbol{v}\|_{\boldsymbol{\Lambda}^{-1}}}.$$

□

Lemma A.11. *Assume that, for all $\boldsymbol{v} \in \mathcal{V}$ we have*

$$|\mathsf{cat}[\boldsymbol{\Lambda}^{-1}\boldsymbol{v}] - \zeta[\boldsymbol{\Lambda}^{-1}\boldsymbol{v}]| \le C_1\|\boldsymbol{v}\|_{\boldsymbol{\Lambda}^{-1}} + C_2/T$$

for some $\mathcal{V} \subseteq \mathbb{R}^d$, $\mathbf{0} \notin \mathcal{V}$, *and* $\zeta[\mathbf{\Lambda}^{-1}\boldsymbol{v}] = \langle \boldsymbol{v}, \boldsymbol{\theta}_\star \rangle$. *Let* $\boldsymbol{u}_1, \dots, \boldsymbol{u}_d$ *denote the eigenvectors of* $\mathbf{\Lambda}$, *and set*

$$\widehat{\boldsymbol{\theta}} = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_d] \cdot [\mathsf{cat}[\mathbf{\Lambda}^{-1}\boldsymbol{u}_1], \dots, \mathsf{cat}[\mathbf{\Lambda}^{-1}\boldsymbol{u}_d]]^\top.$$

Then, for all $\boldsymbol{v} \in \mathcal{V}$,

$$|\langle \boldsymbol{v}, \widehat{\boldsymbol{\theta}} \rangle - \mathsf{cat}[\mathbf{\Lambda}^{-1}\boldsymbol{v}]| \le (\sqrt{d}+1)C_1 \|\boldsymbol{v}\|_{\mathbf{\Lambda}^{-1}} + (\sqrt{d} \cdot \sup_{\boldsymbol{v}' \in \mathcal{V}} \|\boldsymbol{v}'\|_2 + 1)C_2/T,$$
$$|\langle \boldsymbol{v}, \widehat{\boldsymbol{\theta}} \rangle - \zeta[\mathbf{\Lambda}^{-1}\boldsymbol{v}]| \le (\sqrt{d}+2)C_1 \|\boldsymbol{v}\|_{\mathbf{\Lambda}^{-1}} + (\sqrt{d} \cdot \sup_{\boldsymbol{v}' \in \mathcal{V}} \|\boldsymbol{v}'\|_2 + 2)C_2/T.$$

Proof. Let

$$\widetilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \max_{\boldsymbol{v} \in \mathcal{V}} \frac{|\langle \boldsymbol{\theta}, \boldsymbol{v} \rangle - \mathsf{cat}[\mathbf{\Lambda}^{-1}\boldsymbol{v}]|}{\|\boldsymbol{v}\|_{\mathbf{\Lambda}^{-1}}}.$$

Fix some $\boldsymbol{v} \in \mathcal{V}$, and express $\boldsymbol{v}$ as $\boldsymbol{v} = \sum_{i=1}^d a_i \boldsymbol{u}_i$. Then,

$$|\langle \boldsymbol{v}, \widehat{\boldsymbol{\theta}} \rangle - \langle \boldsymbol{v}, \widetilde{\boldsymbol{\theta}} \rangle| = |\sum_{i=1}^d a_i \langle \boldsymbol{u}_i, \widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}} \rangle| \le \sum_{i=1}^d |a_i| |\langle \boldsymbol{u}_i, \widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}} \rangle|.$$

By construction, we will have that $\langle \boldsymbol{u}_i, \widehat{\boldsymbol{\theta}} \rangle = \mathsf{cat}[\mathbf{\Lambda}^{-1}\boldsymbol{u}_i]$. Furthermore, by Lemma 5.2, $\widetilde{\boldsymbol{\theta}}$ will satisfy $|\langle \boldsymbol{u}_i, \widetilde{\boldsymbol{\theta}} \rangle - \mathsf{cat}[\mathbf{\Lambda}^{-1}\boldsymbol{u}_i]| \le C_1 \|\boldsymbol{u}_i\|_{\mathbf{\Lambda}^{-1}} + C_2/T$. Thus,

$$\begin{aligned}
\sum_{i=1}^d |a_i| |\langle \boldsymbol{u}_i, \widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}} \rangle| &= \sum_{i=1}^d |a_i| |\mathsf{cat}[\mathbf{\Lambda}^{-1}\boldsymbol{u}_i] - \langle \boldsymbol{u}_i, \widetilde{\boldsymbol{\theta}} \rangle| \\
&\le \sum_{i=1}^d |a_i| (C_1 \|\boldsymbol{u}_i\|_{\mathbf{\Lambda}^{-1}} + C_2/T) \\
&\le C_1 \sqrt{d} \sqrt{\sum_{i=1}^d a_i^2 \|\boldsymbol{u}_i\|_{\mathbf{\Lambda}^{-1}}^2} + \frac{C_2}{T} \sum_{i=1}^d |a_i|
\end{aligned}$$

where the last inequality follows by Cauchy-Schwarz. Since $\boldsymbol{u}_i$ are the eigenvectors of $\mathbf{\Lambda}_T$ and are therefore orthogonal, we will have that

$$\sqrt{\sum_{i=1}^d a_i^2 \|\boldsymbol{u}_i\|_{\mathbf{\Lambda}_T^{-1}}^2} = \sqrt{(\sum_{i=1}^d a_i \boldsymbol{u}_i)^\top \mathbf{\Lambda}_T^{-1} (\sum_{i=1}^d a_i \boldsymbol{u}_i)} = \|\boldsymbol{v}\|_{\mathbf{\Lambda}_T^{-1}}.$$

Finally, we can bound

$$\sum_{i=1}^d |a_i| \le \sqrt{d} \sqrt{\sum_{i=1}^d a_i^2} = \sqrt{d} \sqrt{\|\boldsymbol{v}\|_2^2} \le \sqrt{d} \cdot \sup_{\boldsymbol{v}' \in \mathcal{V}} \|\boldsymbol{v}'\|_2.$$

The first result follows by upper bounding

$$|\langle \boldsymbol{v}, \widehat{\boldsymbol{\theta}} \rangle - \mathsf{cat}[\mathbf{\Lambda}^{-1}\boldsymbol{v}]| \le |\langle \boldsymbol{v}, \widehat{\boldsymbol{\theta}} \rangle - \langle \boldsymbol{v}, \widetilde{\boldsymbol{\theta}} \rangle| + |\langle \boldsymbol{v}, \widetilde{\boldsymbol{\theta}} \rangle - \mathsf{cat}[\mathbf{\Lambda}^{-1}\boldsymbol{v}]|$$

and again applying Lemma 5.2. The second result follows since

$$|\langle \boldsymbol{v}, \widehat{\boldsymbol{\theta}} \rangle - \zeta[\mathbf{\Lambda}^{-1}\boldsymbol{v}]| \le |\langle \boldsymbol{v}, \widehat{\boldsymbol{\theta}} \rangle - \mathsf{cat}[\mathbf{\Lambda}^{-1}\boldsymbol{v}]| + |\mathsf{cat}[\mathbf{\Lambda}^{-1}\boldsymbol{v}] - \zeta[\mathbf{\Lambda}^{-1}\boldsymbol{v}]|$$

and using the first result and the assumption on $|\mathsf{cat}[\mathbf{\Lambda}^{-1}\boldsymbol{v}] - \zeta[\mathbf{\Lambda}^{-1}\boldsymbol{v}]|$. □

Lemma A.12. *Let* $\mathsf{cat}[\mathbf{\Lambda}^{-1}\boldsymbol{v}]$ *denote a Catoni estimate as defined in Lemma 5.1, and assume that the data used to form* $\mathsf{cat}[\mathbf{\Lambda}^{-1}\boldsymbol{v}]$, $\{X_t(\boldsymbol{v},\mathbf{\Lambda})\}_{t=1}^T$, *satisfies* $|X_t(\boldsymbol{v},\mathbf{\Lambda})| \le \gamma\|\boldsymbol{v}\|_{\mathbf{\Lambda}^{-1}}$ *for all* t *and* $\boldsymbol{v}$, *and that* $\|\boldsymbol{v}\|_{\mathbf{\Lambda}^{-1}} \le \|\boldsymbol{v}\|_2/\sqrt{\lambda}$. *Then for* $\widehat{\boldsymbol{\theta}}$ *as defined in Lemma 5.2, if we have* $\mathcal{S}^{d-1} \subseteq \mathcal{V}$, *for* $\mathcal{S}^{d-1}$ *the unit sphere, we have:*

$$\|\widehat{\boldsymbol{\theta}}\|_2 \le 2\gamma/\sqrt{\lambda}$$

and for $\widehat{\boldsymbol{\theta}}$ *as defined in Lemma A.11 and any* $\mathcal{V}$*:*

$$\|\widehat{\boldsymbol{\theta}}\|_2 \le \sqrt{d}\gamma/\sqrt{\lambda}.$$

Proof. By Claim A.14 and our assumption that $|X_t(\boldsymbol{v},\mathbf{\Lambda})| \le \gamma\|\boldsymbol{v}\|_2$, we can bound $|\mathsf{cat}[\mathbf{\Lambda}^{-1}\boldsymbol{v}]| \le \gamma\|\boldsymbol{v}\|_2$. First consider setting $\widehat{\boldsymbol{\theta}}$ as in Lemma 5.2. Fix $\boldsymbol{v} \in \mathcal{S}^{d-1}$. By assumption $\boldsymbol{v} \in \mathcal{V}$, so we have

$$\begin{aligned}
|\langle\widehat{\boldsymbol{\theta}},\boldsymbol{v}\rangle| &\le \|\boldsymbol{v}\|_{\mathbf{\Lambda}^{-1}} \frac{|\langle\widehat{\boldsymbol{\theta}},\boldsymbol{v}\rangle - \mathsf{cat}[\mathbf{\Lambda}^{-1}\boldsymbol{v}]|}{\|\boldsymbol{v}\|_{\mathbf{\Lambda}^{-1}}} + |\mathsf{cat}[\mathbf{\Lambda}^{-1}\boldsymbol{v}]| \\
&\le \|\boldsymbol{v}\|_{\mathbf{\Lambda}^{-1}} \left[\sup_{\boldsymbol{v}'\in\mathcal{V}} \frac{|\langle\widehat{\boldsymbol{\theta}},\boldsymbol{v}'\rangle - \mathsf{cat}[\mathbf{\Lambda}^{-1}\boldsymbol{v}']|}{\|\boldsymbol{v}'\|_{\mathbf{\Lambda}^{-1}}}\right] + |\mathsf{cat}[\mathbf{\Lambda}^{-1}\boldsymbol{v}]| \\
&= \|\boldsymbol{v}\|_{\mathbf{\Lambda}^{-1}} \left[\min_{\boldsymbol{\theta}} \sup_{\boldsymbol{v}'\in\mathcal{V}} \frac{|\langle\boldsymbol{\theta},\boldsymbol{v}'\rangle - \mathsf{cat}[\mathbf{\Lambda}^{-1}\boldsymbol{v}']|}{\|\boldsymbol{v}'\|_{\mathbf{\Lambda}^{-1}}}\right] + |\mathsf{cat}[\mathbf{\Lambda}^{-1}\boldsymbol{v}]| \\
&\le \|\boldsymbol{v}\|_{\mathbf{\Lambda}^{-1}} \left[\sup_{\boldsymbol{v}'\in\mathcal{V}} \frac{|\mathsf{cat}[\mathbf{\Lambda}^{-1}\boldsymbol{v}']|}{\|\boldsymbol{v}'\|_{\mathbf{\Lambda}^{-1}}}\right] + |\mathsf{cat}[\mathbf{\Lambda}^{-1}\boldsymbol{v}]| \\
&\le \|\boldsymbol{v}\|_{\mathbf{\Lambda}^{-1}} \left[\sup_{\boldsymbol{v}'\in\mathcal{V}} \frac{\gamma\|\boldsymbol{v}'\|_{\mathbf{\Lambda}^{-1}}}{\|\boldsymbol{v}'\|_{\mathbf{\Lambda}^{-1}}}\right] + \gamma\|\boldsymbol{v}\|_{\mathbf{\Lambda}^{-1}} \\
&\le \gamma\|\boldsymbol{v}\|_{\mathbf{\Lambda}^{-1}} + \gamma\|\boldsymbol{v}\|_{\mathbf{\Lambda}^{-1}} \\
&\le 2\gamma/\sqrt{\lambda}.
\end{aligned}$$

As this holds for all $\boldsymbol{v} \in \mathcal{S}^{d-1}$, it follows that $\|\widehat{\boldsymbol{\theta}}\|_2 \le 2\gamma/\sqrt{\lambda}$.

If $\widehat{\boldsymbol{\theta}}$ is set as in Lemma A.11, we have that

$$\|\widehat{\boldsymbol{\theta}}\|_2 \le \sqrt{\sum_{i=1}^d \mathsf{cat}[\mathbf{\Lambda}^{-1}\boldsymbol{u}_i]^2} \le \sqrt{\sum_{i=1}^d \gamma^2\|\boldsymbol{u}_i\|^2_{\mathbf{\Lambda}^{-1}}} = \sqrt{d/\lambda}\gamma.$$

□

A.5 Catoni Perturbation Analysis

Lemma A.13. *Consider some fixed* $\mathcal{X} := \{X_t\}_{t=1}^T, \widetilde{\mathcal{X}} := \{\widetilde{X}_t\}_{t=1}^T$ *satisfying* $|X_t| \le \gamma, |\widetilde{X}_t| \le \gamma$ *for all* t, *and some fixed* $\alpha > 0, \widetilde{\alpha} > 0$. *Let* z^* *denote the root of the function* $f_{\mathsf{cat}}(z;\mathcal{X},\alpha)$ *and* $\widetilde{z}^*$ *the root of* $f_{\mathsf{cat}}(z;\widetilde{\mathcal{X}},\widetilde{\alpha})$. *Then, assuming that*

$$\epsilon := \frac{1}{T}\sum_{t=1}^T \alpha|X_t - \widetilde{X}_t| + 3\gamma|\alpha - \widetilde{\alpha}| \le \frac{1}{18}\min\left\{1, \alpha^2\gamma^2\right\}$$

we will have

$$|z^* - \widetilde{z}^*| \le \frac{1+2\alpha\gamma}{\alpha}\epsilon + \sqrt{\frac{2\epsilon}{\alpha^2}}.$$

Proof. For simplicity, we will denote $f(z) := f_{\mathsf{cat}}(z; \mathcal{X}, \alpha)$ and $\widetilde{f}(z) := f_{\mathsf{cat}}(z; \widetilde{\mathcal{X}}, \widetilde{\alpha})$. Fix some $\Delta \geq 0$ with $\Delta \leq \gamma$. Note that $f(z)$ is differentiable, even at $z = 0$, and

$$\frac{d}{dz} f(z) = \sum_{t=1}^{T} -\alpha \psi'_{\mathsf{cat}}(\alpha(X_t - z)), \qquad \psi'_{\mathsf{cat}}(y) = \begin{cases} \frac{1+y}{1+y+y^2/2} & y \geq 0 \\ \frac{1-y}{1-y+y^2/2} & y < 0 \end{cases}$$

By the Mean Value Theorem,

$$f(z^* + \Delta) = f(z^* + \Delta) - f(z^*) = f'(y)\Delta$$

for some $y \in [z^*, z^* + \Delta]$ which implies that

$$0 \geq f(z^* + \Delta) - \Delta \sup_{z \in [z^*, z^* + \Delta]} f'(z).$$

Note that $\psi'_{\mathsf{cat}}(y) \geq 0$ for all y, that $\psi'_{\mathsf{cat}}(y)$ decreases as $|y|$ increases, and that $\psi'_{\mathsf{cat}}(y) = \psi'_{\mathsf{cat}}(-y)$. It follows that

$$\sup_{z \in [z^*, z^* + \Delta]} -\alpha \psi'_{\mathsf{cat}}(\alpha(X_t - z)) \leq -\alpha \psi'_{\mathsf{cat}}(\alpha|X_t| + \alpha|z^*| + \alpha\Delta).$$

Claim A.14. $z^* \in [-\gamma, \gamma]$.

Proof of Claim A.14. Recall that, by assumption, $|X_t| \leq \gamma$. Furthermore, note that if $X_t - z^* < 0$ for all t, then $f(z^*) < 0$, and similarly, if $X_t - z^* > 0$ for all t, then $f(z^*) > 0$. Since $f(z^*) = 0$, this implies that $\max_t X_t \geq z^*$ and $\min_t X_t \leq z^*$, which implies that $z^* \in [-\gamma, \gamma]$, and so $|z^*| \leq \gamma$. □

By Claim A.14, we can upper bound

$$\begin{aligned} -\alpha \psi'_{\mathsf{cat}}(\alpha|X_t| + \alpha|z^*| + \alpha\Delta) &\leq -\alpha \psi'_{\mathsf{cat}}(2\alpha\gamma + \alpha\Delta) \\ &= -\alpha \cdot \frac{1 + \alpha(2\gamma + \Delta)}{1 + \alpha(2\gamma + \Delta) + \alpha^2(2\gamma + \Delta)^2/2} \end{aligned}$$

which implies that

$$\begin{aligned} \sup_{z \in [z^*, z^* + \Delta]} f'(z) &\leq \sum_{t=1}^{T} \sup_{z \in [z^*, z^* + \Delta]} -\alpha \psi'_{\mathsf{cat}}(\alpha(X_t - z)) \\ &\leq -\frac{T\alpha + T\alpha^2(2\gamma + \Delta)}{1 + \alpha(2\gamma + \Delta) + \alpha^2(2\gamma + \Delta)^2/2} \end{aligned}$$

so

$$0 = f(z^*) \geq f(z^* + \Delta) + \frac{T\alpha\Delta(1 + \alpha(2\gamma + \Delta))}{1 + \alpha(2\gamma + \Delta) + \alpha^2(2\gamma + \Delta)^2/2}. \tag{A.15}$$

Note that $|\psi'_{\mathsf{cat}}(y)| \leq 1$ for all y, which implies that $|\psi_{\mathsf{cat}}(y) - \psi_{\mathsf{cat}}(y')| \leq |y - y'|$. It follows that

$$\begin{aligned} |f(z^* + \Delta) - \widetilde{f}(z^* + \Delta)| &\leq \sum_{t=1}^{T} |\psi_{\mathsf{cat}}(\alpha(X_t - z^* - \Delta)) - \psi_{\mathsf{cat}}(\widetilde{\alpha}(\widetilde{X}_t - z^* - \Delta))| \\ &\leq \sum_{t=1}^{T} |\alpha(X_t - z^* - \Delta) - \widetilde{\alpha}(\widetilde{X}_t - z^* - \Delta)| \end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{t=1}^{T} \alpha |X_t - \widetilde{X}_t| + \sum_{t=1}^{T} |\alpha - \widetilde{\alpha}||\widetilde{X}_t| + T|\alpha - \widetilde{\alpha}||z^* + \Delta| \\
&\leq \sum_{t=1}^{T} \alpha |X_t - \widetilde{X}_t| + 3T\gamma|\alpha - \widetilde{\alpha}| \\
&=: \epsilon \cdot T
\end{aligned}
$$

Thus,

$$
\begin{aligned}
f(z^* + \Delta) + \frac{T\alpha\Delta(1 + \alpha(2\gamma + \Delta))}{1 + \alpha(2\gamma + \Delta) + \alpha^2(2\gamma + \Delta)^2/2} \\
\geq \widetilde{f}(z^* + \Delta) + \frac{T\alpha\Delta(1 + \alpha(2\gamma + \Delta))}{1 + \alpha(2\gamma + \Delta) + \alpha^2(2\gamma + \Delta)^2/2} - \epsilon T.
\end{aligned}
$$

If

$$
\frac{\alpha\Delta(1 + \alpha(2\gamma + \Delta))}{1 + \alpha(2\gamma + \Delta) + \alpha^2(2\gamma + \Delta)^2/2} - \epsilon \geq 0, \tag{A.16}
$$

then by (A.15) it follows that $0 \geq \widetilde{f}(z^* + \Delta)$. Since $\widetilde{f}$ is monotonically decreasing in z and $\widetilde{f}(\widetilde{z}^*) = 0$, $\widetilde{z}^* \leq z^* + \Delta$.

It remains to determine what choice of Δ is sufficient. Solving (A.16) for Δ, we will have that (A.16) is met as long as

$$
\Delta \geq \frac{(1 + 2\alpha\gamma)\epsilon - (1 + 2\alpha\gamma) + \sqrt{-\epsilon^2 + 2\epsilon + (1 + 4\alpha\gamma + 4\alpha^2\gamma^2)}}{2\alpha - \alpha\epsilon}.
$$

By assumption we have that $1 \geq \epsilon$, so $-\epsilon^2 + 2\epsilon + (1 + 4\alpha\gamma + 4\alpha^2\gamma^2)$ is non-negative. We can then bound

$$
\begin{aligned}
&\frac{(1 + 2\alpha\gamma)\epsilon - (1 + 2\alpha\gamma) + \sqrt{-\epsilon^2 + 2\epsilon + (1 + 4\alpha\gamma + 4\alpha^2\gamma^2)}}{2\alpha - \alpha\epsilon} \\
&\leq \frac{(1 + 2\alpha\gamma)\epsilon - (1 + 2\alpha\gamma) + \sqrt{-\epsilon^2 + 2\epsilon} + \sqrt{(1 + 4\alpha\gamma + 4\alpha^2\gamma^2)}}{2\alpha - \alpha\epsilon} \\
&\leq \frac{(1 + 2\alpha\gamma)\epsilon + \sqrt{-\epsilon^2 + 2\epsilon}}{\alpha} \\
&\leq \frac{1 + 2\alpha\gamma}{\alpha}\epsilon + \sqrt{\frac{2\epsilon}{\alpha^2}}.
\end{aligned}
$$

A sufficient condition to meet (A.16) is then

$$
\Delta = \frac{1 + 2\alpha\gamma}{\alpha}\epsilon + \sqrt{\frac{2\epsilon}{\alpha^2}}.
$$

Thus,

$$
\widetilde{z}^* \geq z^* + \frac{1 + 2\alpha\gamma}{\alpha}\epsilon + \sqrt{\frac{2\epsilon}{\alpha^2}}.
$$

We have required that $\Delta \leq \gamma$, but note that this is met for this choice of Δ since we have assumed that

$$
\epsilon \leq \min\left\{\frac{1}{6}, \frac{\alpha\gamma}{3}, \frac{\alpha^2\gamma^2}{18}\right\}
$$

and ϵ satisfying this will ensure that $\Delta \leq \gamma$. Notice that the above condition is satisfied when

$$\epsilon \leq \frac{1}{18} \min\left\{1, \alpha^2\gamma^2\right\}.$$

The result follows by repeating this argument in the opposite direction. □

B Regret Analysis

We will consider a slightly more general setup here than that considered in the main text. In particular, we will allow for the reward function to be time-varying: at episodes k, the reward is specified by $r_h^k(s,a)$. We will make several assumptions on this reward.

Assumption 1 (Time-Varying Reward)**.** *The reward function $r_h^k(s,a) \in [0,1]$ is $\mathcal{F}_{k-1}$-measurable, and non-increasing in k: $r_h^k(s,a) \leq r_h^{k-1}(s,a)$ for all s,a,h,k. Furthermore, for each h,k, $r_h^k \in \mathscr{R}$ for some function class $\mathscr{R}$, and $\mathscr{R}$ has covering number bounded as $\mathsf{N}(\mathscr{R}, \mathrm{dist}_\infty, \epsilon) \leq d_{\mathscr{R}} \log(1+\frac{2R_{\mathscr{R}}}{\epsilon})$.*

As the reward function changes at each step, we will denote the value function for policy π at episode k by $Q_h^{k,\pi}(s,a)$ (and similarly $V_h^{k,\pi}(s)$). We will also redefine regret as

$$\mathcal{R}_K := \sum_{k=1}^{K} (V_1^{k,\star} - V_1^{k,\pi_k})$$

for $V_1^{k,\star}$ the optimal value function for reward r^k. To accommodate time-varying reward in FORCE, the update of the optimistic Q-estimate on Line 14 must be changed to:

$$Q_h^k(\cdot,\cdot) \leftarrow \min\{r_h^k(\cdot,\cdot) + \langle \boldsymbol{\phi}(\cdot,\cdot), \widehat{\boldsymbol{w}}_h^k \rangle + 6\beta \|\boldsymbol{\phi}(\cdot,\cdot)\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}} + 12\mathsf{v}_{\min}\beta^2/k^2, H\}$$

and the following settings of K_{init} and β must be used:

$$K_{\mathrm{init}} \leftarrow c\left((d^2 + d_{\mathscr{R}}) \log(\max\{d, \mathsf{v}_{\min}^{-1}, K, H, R_{\mathscr{R}}\}) + \log(2HK/\delta)\right)$$
$$\beta \leftarrow 6\sqrt{c(d^2 + d_{\mathscr{R}}) \log\left(\max\{d, \mathsf{v}_{\min}^{-1}, H, K, R_{\mathscr{R}}\}\right) + \log(2HK/\delta)}.$$

We then have the following result.

Theorem 8 (Regret Bound for Time-Varying Reward)**.** *Fix a failure probability $\delta \in (0,1)$ and $K \in \mathbb{N}$, and assume that the reward satisfies Assumption 1. Then, the regret of* FORCE*, modified to handle time-varying rewards as outlined above, satisfies the following bound with probability at least $1-3\delta$:*

$$\mathcal{R}_K \leq c_1 \sqrt{d(d^2 + d_{\mathscr{R}})H^3 \cdot \log(R_{\mathscr{R}}HK/\delta) \log^2(HK/\delta) \cdot \sum_{k=1}^{K} V_1^{k,\star}} + c_2 \sqrt{d}(d^2 + d_{\mathscr{R}})^{3/2} H^3 \log^{3/2}(R_{\mathscr{R}}HK/\delta) \log^2(HK/\delta)$$

for universal constants c_1, c_2. Furthermore, if we use the computationally efficient update as outlined in Theorem 5, with probability at least $1-3\delta$, the regret is bounded by

$$\mathcal{R}_K \leq c_1 \sqrt{d^2(d^2 + d_{\mathscr{R}})H^3 \cdot \log(R_{\mathscr{R}}HK/\delta) \log^2(HK/\delta) \cdot \sum_{k=1}^{K} V_1^{k,\star}} + c_2 d(d^2 + d_{\mathscr{R}})^{3/2} H^3 \log^{3/2}(R_{\mathscr{R}}HK/\delta) \log^2(HK/\delta)$$

and computation will scale polynomially in d, H, K, and $\min\{|\mathcal{A}|, \mathcal{O}(2^d)\}$.

Theorem 4 and Theorem 5 are direct corollaries of Theorem 8, where we simply set $r_h^k = r_h$ for all k, replace $\sum_{k=1}^K V_1^{k,\star}$ with $KV_1^\star$, and note that since the reward is deterministic in this case, no cover over reward functions is necessary, so the regret scales independently of $d_{\mathscr{R}}$ and $R_{\mathscr{R}}$. Throughout the remainder of this section, we will consider this more general time-varying reward setting.

B.1 Preliminaries and Notation

Define the following events:

$$A_{k,h} := \left\{ \left| \mathsf{cat}_{h,k}\left[(k-1)\mathbf{\Lambda}_{h,k-1}^{-1}\boldsymbol{\phi}_{h,k}\right] - \mathbb{E}_h[V_{h+1}^k](s_{h,k},a_{h,k}) \right| \le \beta\|\boldsymbol{\phi}_{h,k}\|_{\mathbf{\Lambda}_{h,k-1}^{-1}} + \mathsf{v}_{\min}\beta^2/k^2 \right\}$$

$$B_{k,h} := \left\{ \forall \boldsymbol{v} \in \mathcal{B}^d \ : \ \left| \widehat{\mathbb{E}}_h[V_{h+1}^k](\boldsymbol{v}) - \mathbb{E}_h[V_{h+1}^k](\boldsymbol{v}) \right| \le \beta\|\boldsymbol{v}\|_{\mathbf{\Lambda}_{h,k-1}^{-1}} + \mathsf{v}_{\min}\beta^2/k^2 \right\}$$

$$\mathcal{E} := \bigcap_{k=K_{\text{init}}}^{K} \bigcap_{h=1}^{H} (B_{k,h} \cap A_{k,h})$$

where we denote $\widehat{\mathbb{E}}_h[V_{h+1}^k](\boldsymbol{v}) = \mathsf{cat}_{h,k}[(k-1)\mathbf{\Lambda}_{h,k-1}^{-1}\boldsymbol{v}]$, $\beta = 6\sqrt{C_{\text{mdp}} + \log(2HK/\delta)}$, and

$$C_{\text{mdp}} := c(d^2 + d_{\mathscr{R}}) \cdot \mathsf{logs}\left(d, \mathsf{v}_{\min}^{-1}, H, 1/\lambda, K, R_{\mathscr{R}}\right)$$

for a universal constant c. Here we overload notation slightly and define:

$$\mathbb{E}_h[V_{h+1}^k](\boldsymbol{v}) = \langle \boldsymbol{v}, \int V_{h+1}^k(s')\mathrm{d}\boldsymbol{\mu}_h(s') \rangle.$$

We will also define $r_h(\boldsymbol{v}) = r_h(s,a)$ if $\boldsymbol{v} = \boldsymbol{\phi}(s,a)$, and 0 otherwise. Throughout this section, we will also denote $\widehat{\mathbb{E}}_h[V_{h+1}^k](s,a) := \widehat{\mathbb{E}}_h[V_{h+1}^k](\boldsymbol{\phi}(s,a))$.

The analysis of the computationally inefficient and computationally efficient versions of FORCE are nearly identical, and we therefore prove them in tandem. To facilitate this, we will define the parameter

$$\widetilde{\beta} := \begin{cases} 2\beta & \texttt{efficient} = \texttt{false} \\ (\sqrt{d}+2)\beta & \texttt{efficient} = \texttt{true} \end{cases}$$

where the `efficient` flag corresponds to which version of the algorithm we are running: `efficient` = `false` corresponds to running the version of FORCE as stated in Algorithm 1, and `efficient` = `true` corresponds to running the computationally efficient version as described in Section 4.2. Given the definition of $\widetilde{\beta}$, we can then write the update to Q_h^k as

$$Q_h^k(\cdot,\cdot) \leftarrow \min\{r_h^k(\cdot,\cdot) + \langle \boldsymbol{\phi}(\cdot,\cdot), \widehat{\boldsymbol{w}}_h^k \rangle + 3\widetilde{\beta}\|\boldsymbol{\phi}(\cdot,\cdot)\|_{\mathbf{\Lambda}_{h,k-1}^{-1}} + 3\mathsf{v}_{\min}\widetilde{\beta}^2/k^2, H\},$$

and this update holds in either the efficient or inefficient case. We will use $\widetilde{\beta}$ throughout the analysis, and set $\mathsf{v}_{\min} = 1/K$, $\alpha_{\max} = K/\mathsf{v}_{\min}$ as in FORCE.

B.2 Catoni Estimation is Correct for Linear MDPs

Lemma B.1. *Consider the function class*

$$\mathscr{F}_{\text{mdp}} = \left\{ f(\cdot) = \min\{r(\cdot) + \langle \cdot, \boldsymbol{w} \rangle + 3\widetilde{\beta}\|\cdot\|_{\mathbf{\Lambda}^{-1}} + \bar{c}, H\} \ : \ \|\boldsymbol{w}\|_2 \le 4H\sqrt{d}/(\lambda \mathsf{v}_{\min}^2), \ \mathbf{\Lambda} \succeq \lambda I, r \in \mathscr{R} \right\}$$

and assume $\bar{c} \geq 0$, $\mathsf{N}(\mathscr{R}, \mathrm{dist}_\infty, \epsilon) \leq d_{\mathscr{R}} \log(1 + \frac{2R_{\mathscr{R}}}{\epsilon})$. *Then,*

$$\mathsf{N}(\mathscr{F}_{\mathrm{mdp}}, \mathrm{dist}_\infty, \epsilon) \leq (4d^2 + d_{\mathscr{R}}) \log\left(1 + \frac{\sqrt{d}(288\widetilde{\beta}^2 + 8H/\mathsf{v}_{\min}^2)/\lambda + 4R_{\mathscr{R}}}{\epsilon}\right).$$

Furthermore, conditioned on the event $\cap_{\tau=K_{\mathrm{init}}}^{k-2} \cap_{h'=h+1}^{H} (B_{\tau,h'} \cap B_{k-1,h'}) \cap A_{\tau,h}$, *the Catoni estimation problems on Line 10 at episode* k *of* FORCE *are instances of the regression with function approximation setting of Definition 5.2 for*

$$\beta_\mu = \sqrt{d}, \quad \beta_u = 0, \quad \mathscr{F} = \mathscr{F}_{\mathrm{mdp}}.$$

Similarly, conditioned on the event $\cap_{\tau=K_{\mathrm{init}}}^{k-1} \cap_{h'=h+1}^{H} (B_{\tau,h'} \cap B_{k,h'}) \cap A_{\tau,h}$, *the Catoni estimation problems on Line 13 and in the computationally efficient update of Equation (4.5), are instances of the regression with function approximation setting of Definition 5.2 for*

$$\beta_\mu = \sqrt{d}, \quad \beta_u = 0, \quad \mathscr{F} = \mathscr{F}_{\mathrm{mdp}}.$$

Proof. We will instantiate Definition 5.2 with $\boldsymbol{\phi}_\tau = \boldsymbol{\phi}_{h,\tau}$, $\boldsymbol{\phi}'_\tau = \boldsymbol{\phi}_{h+1,\tau}$, $\boldsymbol{\mu} = \boldsymbol{\mu}_h$, $\sigma_\tau = \bar{\mathsf{v}}_{h,\tau}$, and the function class $\mathscr{F} = \mathscr{F}_{\mathrm{mdp}}$. FORCE solves two different forms of regression problems. In the first setting, when solving for $\bar{\mathsf{v}}^2_{h,k-1}$ on Line 10, we consider $y_\tau = V^{k-1}_{h+1}(s_{h+1,\tau})$, and $\boldsymbol{u}_\star = 0$. In the second, when solving either $\mathsf{cat}_{h,k}[(k-1)\boldsymbol{\Lambda}^{-1}_{h,k-1}\boldsymbol{v}]$ or $\mathsf{cat}_{h,k}[(k-1)\boldsymbol{\Lambda}^{-1}_{h,k-1}\boldsymbol{u}_i]$, we set $y_\tau = V^k_{h+1}(s_{h+1,\tau})$ and set $\boldsymbol{u}_\star = 0$.

We verify that this meets the criteria of Definition 5.2. First, note that by definition of $\mathcal{F}_{h,\tau}$, we will have that $\boldsymbol{\phi}_{h,\tau}$ is $\mathcal{F}_{h,\tau}$-measurable and that $\boldsymbol{\phi}_{h+1,\tau}$ is $\mathcal{F}_{h+1,\tau}$-measurable. In addition, $\|\boldsymbol{\phi}_{h,\tau}\|_2 \leq 1$ and $\|\boldsymbol{\phi}_{h+1,\tau}\|_2 \leq 1$ by assumption. Given the linear MDP structure of Definition 3.1, for any bounded function f,

$$\mathbb{E}[f(\boldsymbol{\phi}_{h+1,\tau}) \mid \mathcal{F}_{h,\tau}] = \langle \boldsymbol{\phi}_{h,\tau}, \int f(\boldsymbol{\phi}(s', \pi^\tau_{h+1}(s')))\mathrm{d}\boldsymbol{\mu}_h(s')\rangle.$$

Note that we can think of $\boldsymbol{\mu}_h(\cdot)$ as a measure over $\mathbb{R}^d$, as required by Definition 5.2, by associating s' with $\boldsymbol{\phi}(s', \pi^\tau_{h+1}(s'))$, and putting a measure of 0 on all vectors $\boldsymbol{v}$ such that there does not exist s, a with $\boldsymbol{v} = \boldsymbol{\phi}(s,a)$. In addition, by assumption $\||\boldsymbol{\mu}_h|(\mathcal{S})\|_2 \leq \sqrt{d}$, so we can take $\beta_\mu = \sqrt{d}$.

In both settings, since $\boldsymbol{u}_\star = 0$, it suffices to take $\beta_u = 0$. Note that for any s, a, h, k, τ, we can bound

$$\begin{aligned}
&|\boldsymbol{\phi}(s,a)^\top \boldsymbol{\Lambda}^{-1}_{h,k-1}\boldsymbol{\phi}_{h,\tau} V^k_{h+1}(s_{h+1,\tau})/\bar{\mathsf{v}}^2_{h,\tau}| \\
&\qquad \leq \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}^{-1}_{h,k-1}} \cdot \|\boldsymbol{\Lambda}^{-1/2}_{h,k-1}\|_{\mathrm{op}} \|\boldsymbol{\phi}_{h,\tau}\|_2 |V^k_{h+1}(s_{h+1,k})|/\bar{\mathsf{v}}^2_{h,\tau} \\
&\qquad \leq \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}^{-1}_{h,k-1}} \cdot \frac{H}{\sqrt{\lambda}\mathsf{v}^2_{\min}}
\end{aligned}$$

so by Lemma A.12, we will have that $\|\widehat{\boldsymbol{w}}^k_{h+1}\|_2 \leq \frac{4H\sqrt{d}}{\lambda \mathsf{v}^2_{\min}}$. It follows that, by construction of $V^k_{h+1}(\cdot)$, we will have $V^k_{h+1}(\cdot) \in \mathscr{F}_{\mathrm{mdp}}$.

It remains to show that the condition on $\bar{\mathsf{v}}_{h,\tau}$, (5.4), is met at round k. In our setting, for the Catoni estimation on Line 10 at episode k, (5.4) is equivalent to

$$\mathbb{E}_h[(V^{k-1}_{h+1})^2](s_{h,\tau}, a_{h,\tau}) \leq \frac{1}{2}\bar{\mathsf{v}}^2_{h,\tau}.$$

However, by Lemma B.5, this holds for all $\tau \geq K_{\text{init}}$ on the $\cap_{\tau=K_{\text{init}}}^{k-2} \cap_{h'=h+1}^{H} (B_{\tau,h'} \cap B_{k-1,h'}) \cap A_{\tau,h}$. For $\tau \leq K_{\text{init}}$ it trivially as we set $\bar{\mathsf{v}}_{h,\tau}^2 = 2H^2$ and since $V_{h+1}^k(s') \in [0, H]$. For the Catoni estimation on Line 13 or in Equation (4.5), (5.4) is equivalent to

$$\mathbb{E}_h[(V_{h+1}^k)^2](s_{h,\tau}, a_{h,\tau}) \leq \frac{1}{2}\bar{\mathsf{v}}_{h,\tau}^2.$$

Again by Lemma B.5, this holds on the event $\cap_{\tau=K_{\text{init}}}^{k-1} \cap_{h'=h+1}^{H} (B_{\tau,h'} \cap B_{k,h'}) \cap A_{\tau,h}$ for $\tau \geq K_{\text{init}}$. For $\tau \leq K_{\text{init}}$ this trivially holds since $\bar{\mathsf{v}}_{h,\tau}^2 = 2H^2$.

Finally, we bound the covering number of $\mathscr{F}_{\text{mdp}}$. Consider $f_1, f_2 \in \mathscr{F}_{\text{mdp}}$, then

$$\begin{aligned}
\text{dist}_\infty(f_1, f_2) &= \sup_{\boldsymbol{\phi} \in \mathcal{B}^d} |f_1(\boldsymbol{\phi}) - f_2(\boldsymbol{\phi})| \\
&= \sup_{\boldsymbol{\phi} \in \mathcal{B}^d} \Big| \min\{r_1(\boldsymbol{\phi}) + \langle \boldsymbol{\phi}, \boldsymbol{w}_1 \rangle + 3\widetilde{\beta}\|\boldsymbol{\phi}\|_{\boldsymbol{\Lambda}_1^{-1}} + \bar{c}, H\} \\
&\qquad\qquad - \min\{r_2(\boldsymbol{\phi}) + \langle \boldsymbol{\phi}, \boldsymbol{w}_2 \rangle + 3\widetilde{\beta}\|\boldsymbol{\phi}\|_{\boldsymbol{\Lambda}_2^{-1}} + \bar{c}, H\}\Big|.
\end{aligned}$$

Assume that $f_1(\boldsymbol{\phi}) = f_2(\boldsymbol{\phi}) = H$, then we can clearly bound

$$|f_1(\boldsymbol{\phi}) - f_2(\boldsymbol{\phi})| \leq |r_1(\boldsymbol{\phi}) - r_2(\boldsymbol{\phi})| + |\min\{\langle \boldsymbol{\phi}, \boldsymbol{w}_1 \rangle + 3\widetilde{\beta}\|\boldsymbol{\phi}\|_{\boldsymbol{\Lambda}_1^{-1}}, H\} - \min\{\langle \boldsymbol{\phi}, \boldsymbol{w}_2 \rangle + 3\widetilde{\beta}\|\boldsymbol{\phi}\|_{\boldsymbol{\Lambda}_2^{-1}}, H\}|.$$

If $f_1(\boldsymbol{\phi}) < H, f_2(\boldsymbol{\phi}) < H$, using that $\bar{c}, r_1(\boldsymbol{\phi}), r_2(\boldsymbol{\phi}) \geq 0$, we can bound

$$\begin{aligned}
|f_1(\boldsymbol{\phi}) - f_2(\boldsymbol{\phi})| &= |r_1(\boldsymbol{\phi}) + \langle \boldsymbol{\phi}, \boldsymbol{w}_1 \rangle + 3\widetilde{\beta}\|\boldsymbol{\phi}\|_{\boldsymbol{\Lambda}_1^{-1}} - (r_2(\boldsymbol{\phi}) + \langle \boldsymbol{\phi}, \boldsymbol{w}_2 \rangle + 3\widetilde{\beta}\|\boldsymbol{\phi}\|_{\boldsymbol{\Lambda}_2^{-1}})| \\
&= |r_1(\boldsymbol{\phi}) + \min\{\langle \boldsymbol{\phi}, \boldsymbol{w}_1 \rangle + 3\widetilde{\beta}\|\boldsymbol{\phi}\|_{\boldsymbol{\Lambda}_1^{-1}}, H\} - (r_2(\boldsymbol{\phi}) + \min\{\langle \boldsymbol{\phi}, \boldsymbol{w}_2 \rangle + 3\widetilde{\beta}\|\boldsymbol{\phi}\|_{\boldsymbol{\Lambda}_2^{-1}}, H\})| \\
&\leq |r_1(\boldsymbol{\phi}) - r_2(\boldsymbol{\phi})| + |\min\{\langle \boldsymbol{\phi}, \boldsymbol{w}_1 \rangle + 3\widetilde{\beta}\|\boldsymbol{\phi}\|_{\boldsymbol{\Lambda}_1^{-1}}, H\} - \min\{\langle \boldsymbol{\phi}, \boldsymbol{w}_2 \rangle + 3\widetilde{\beta}\|\boldsymbol{\phi}\|_{\boldsymbol{\Lambda}_2^{-1}}, H\}|.
\end{aligned}$$

If $f_1(\boldsymbol{\phi}) = H, f_2(\boldsymbol{\phi}) < H$,

$$\begin{aligned}
|f_1(\boldsymbol{\phi}) - f_2(\boldsymbol{\phi})| &\leq |r_1(\boldsymbol{\phi}) + \min\{\langle \boldsymbol{\phi}, \boldsymbol{w}_1 \rangle + 3\widetilde{\beta}\|\boldsymbol{\phi}\|_{\boldsymbol{\Lambda}_1^{-1}}, H\} - (r_2(\boldsymbol{\phi}) + \langle \boldsymbol{\phi}, \boldsymbol{w}_2 \rangle + 3\widetilde{\beta}\|\boldsymbol{\phi}\|_{\boldsymbol{\Lambda}_2^{-1}})| \\
&= |r_1(\boldsymbol{\phi}) + \min\{\langle \boldsymbol{\phi}, \boldsymbol{w}_1 \rangle + 3\widetilde{\beta}\|\boldsymbol{\phi}\|_{\boldsymbol{\Lambda}_1^{-1}}, H\} - (r_2(\boldsymbol{\phi}) + \min\{\langle \boldsymbol{\phi}, \boldsymbol{w}_2 \rangle + 3\widetilde{\beta}\|\boldsymbol{\phi}\|_{\boldsymbol{\Lambda}_2^{-1}}, H\})| \\
&\leq |r_1(\boldsymbol{\phi}) - r_2(\boldsymbol{\phi})| + |\min\{\langle \boldsymbol{\phi}, \boldsymbol{w}_1 \rangle + 3\widetilde{\beta}\|\boldsymbol{\phi}\|_{\boldsymbol{\Lambda}_1^{-1}}, H\} - \min\{\langle \boldsymbol{\phi}, \boldsymbol{w}_2 \rangle + 3\widetilde{\beta}\|\boldsymbol{\phi}\|_{\boldsymbol{\Lambda}_2^{-1}}, H\}|.
\end{aligned}$$

The same argument holds of $f_1(\boldsymbol{\phi}) < H, f_2(\boldsymbol{\phi}) = H$. Altogether then,

$$\begin{aligned}
\text{dist}_\infty(f_1, f_2) \leq &\sup_{\boldsymbol{\phi} \in \mathcal{B}^d} |r_1(\boldsymbol{\phi}) - r_2(\boldsymbol{\phi})| \\
&+ \sup_{\boldsymbol{\phi} \in \mathcal{B}^d} |\min\{\langle \boldsymbol{\phi}, \boldsymbol{w}_1 \rangle + 3\widetilde{\beta}\|\boldsymbol{\phi}\|_{\boldsymbol{\Lambda}_1^{-1}}, H\} - \min\{\langle \boldsymbol{\phi}, \boldsymbol{w}_2 \rangle + 3\widetilde{\beta}\|\boldsymbol{\phi}\|_{\boldsymbol{\Lambda}_2^{-1}}, H\}|.
\end{aligned}$$

It follows that we can construct $\epsilon/2$-nets of $\mathscr{R}$ and the class

$$\widetilde{\mathscr{F}} := \left\{ f(\cdot) = \min\{\langle \cdot, \boldsymbol{w} \rangle + 3\widetilde{\beta}\|\cdot\|_{\boldsymbol{\Lambda}^{-1}}, H\} \;:\; \|\boldsymbol{w}\|_2 \leq 4H\sqrt{d}/(\lambda \mathsf{v}_{\min}^2),\ \boldsymbol{\Lambda} \succeq \lambda I \right\}$$

separately, and the union of these nets will serve as an ϵ-net of $\mathscr{F}_{\text{mdp}}$. By assumption, we have $\mathsf{N}(\mathscr{R}, \text{dist}_\infty, \epsilon/2) \leq d_{\mathscr{R}} \log(1 + \frac{4R_{\mathscr{R}}}{\epsilon})$. Furthermore, $\widetilde{\mathscr{F}}$ is identical to the function class considered in Lemma A.2, so

$$\mathsf{N}(\widetilde{\mathscr{F}}, \text{dist}_\infty, \epsilon/2) \leq d \log\left(1 + \frac{8H\sqrt{d}}{\lambda \mathsf{v}_{\min}^2 \epsilon}\right) + d^2 \log\left(1 + \frac{288\sqrt{d}\widetilde{\beta}^2}{\lambda \epsilon^2}\right)$$

$$\leq 4d^2 \log\left(1 + \frac{\sqrt{d}(288\widetilde{\beta}^2 + 8H/\mathsf{v}_{\min}^2)}{\lambda\epsilon}\right).$$

This implies that (since log-covering numbers are additive)

$$\begin{aligned}\mathsf{N}(\mathscr{F}_{\mathrm{mdp}}, \mathrm{dist}_\infty, \epsilon) &\leq 4d^2 \log\left(1 + \frac{\sqrt{d}(288\widetilde{\beta}^2 + 8H/\mathsf{v}_{\min}^2)}{\lambda\epsilon}\right) + d_{\mathscr{R}} \log\left(1 + \frac{4R_{\mathscr{R}}}{\epsilon}\right) \\ &\leq (4d^2 + d_{\mathscr{R}}) \log\left(1 + \frac{\sqrt{d}(288\widetilde{\beta}^2 + 8H/\mathsf{v}_{\min}^2)/\lambda + 4R_{\mathscr{R}}}{\epsilon}\right).\end{aligned}$$

□

Lemma B.2. *Assume we are in the linear MDP setting and are running Algorithm 1 with $\lambda \leq 1/H^2$. Then as long as $K \geq K_{\mathrm{init}}$, we will have that $\mathbb{P}[\mathcal{E}] \geq 1 - \delta$.*

Proof. First, note that

$$\mathcal{E}^c = \bigcup_{k=K_{\mathrm{init}}}^{K} \bigcup_{h=1}^{H} (B_{k,h}^c \cup A_{k,h}^c).$$

Then,

Claim B.3.

$$\begin{aligned}\bigcup_{k=K_{\mathrm{init}}}^{K} \bigcup_{h=1}^{H} (B_{k,h}^c \cup A_{k,h}^c) = &\bigcup_{k=K_{\mathrm{init}}}^{K} \bigcup_{h=1}^{H} \left[B_{k,h}^c \cap \left(\cap_{k'=K_{\mathrm{init}}}^{k-1} \cap_{h'=1}^{H} (B_{k',h'} \cap A_{k',h'})\right) \cap \left(\cap_{h'=h+1}^{H} B_{k,h'}\right)\right] \\ &\cup \bigcup_{k=K_{\mathrm{init}}}^{K} \bigcup_{h=1}^{H} \left[A_{k,h}^c \cap \left(\cap_{k'=K_{\mathrm{init}}}^{k-1} \cap_{h'=1}^{H} (B_{k',h'} \cap A_{k',h'})\right) \cap \left(\cap_{h'=h+1}^{H} B_{k,h'}\right)\right].\end{aligned}$$

Claim B.3 and a union bound imply that

$$\begin{aligned}\mathbb{P}[\mathcal{E}^c] &= \mathbb{P}\left[\bigcup_{k=K_{\mathrm{init}}}^{K} \bigcup_{h=1}^{H} (B_{k,h}^c \cup A_{k,h}^c)\right] \\ &\leq \sum_{k=K_{\mathrm{init}}}^{K} \sum_{h=1}^{H} \mathbb{P}\left[B_{k,h}^c \cap \left(\cap_{k'=K_{\mathrm{init}}}^{k-1} \cap_{h'=1}^{H} (B_{k',h'} \cap A_{k',h'})\right) \cap \left(\cap_{h'=h+1}^{H} B_{k,h'}\right)\right] \\ &\quad + \sum_{k=K_{\mathrm{init}}}^{K} \sum_{h=1}^{H} \mathbb{P}\left[A_{k,h}^c \cap \left(\cap_{k'=K_{\mathrm{init}}}^{k-1} \cap_{h'=1}^{H} (B_{k',h'} \cap A_{k',h'})\right) \cap \left(\cap_{h'=h+1}^{H} B_{k,h'}\right)\right] \\ &\leq \sum_{k=K_{\mathrm{init}}}^{K} \sum_{h=1}^{H} \mathbb{P}\left[B_{k,h}^c | \left(\cap_{k'=K_{\mathrm{init}}}^{k-1} \cap_{h'=1}^{H} (B_{k',h'} \cap A_{k',h'})\right) \cap \left(\cap_{h'=h+1}^{H} B_{k,h'}\right)\right] \\ &\quad + \sum_{k=K_{\mathrm{init}}}^{K} \sum_{h=1}^{H} \mathbb{P}\left[A_{k,h}^c | \cap_{k'=K_{\mathrm{init}}}^{k-1} \cap_{h'=1}^{H} (B_{k',h'} \cap A_{k',h'}) \cap \left(\cap_{h'=h+1}^{H} B_{k,h'}\right)\right].\end{aligned}$$

We first bound

$$\mathbb{P}\left[A_{k,h}^c | \cap_{k'=K_{\text{init}}}^{k-1} \cap_{h'=1}^{H} (B_{k',h'} \cap A_{k',h'}) \cap \left(\cap_{h'=h+1}^{H} B_{k,h'}\right)\right].$$

By Lemma B.1, we have the regression estimate $\mathsf{cat}_{h,k}\left[(k-1)\mathbf{\Lambda}_{h,k-1}^{-1}\boldsymbol{\phi}_{h,k}\right]$ satisfies Definition 5.2 conditioned on the event $\cap_{k'=K_{\text{init}}}^{k-1} \cap_{h'=1}^{H} (B_{k',h'} \cap A_{k',h'}) \cap \left(\cap_{h'=h+1}^{H} B_{k,h'}\right)$. We now apply Theorem 6. First note that since $\alpha_{\max} = K/\mathsf{v}_{\min}$, $\|\boldsymbol{\theta}_h\|_2 \leq \sqrt{d}$, and using the values for β_u and β_μ from Lemma B.1, as well as the covering number bound of $\mathscr{F}_{\text{mdp}}$, it follows that C_{mdp} upper bounds d_T[5]. Since we have assumed $K \geq K_{\text{init}}$, by our choice of $K_{\text{init}} = c(\log(2HK/\delta) + C_{\text{mdp}})$, it follows that the minimum sample condition of Theorem 6 is met for $k \geq K_{\text{init}}$. Finally, note that in this setting, using the definition of linear MDPs, Definition 3.1, we will have that

$$\boldsymbol{\theta}_\star = \int V_{h+1}^k(s')\mathrm{d}\boldsymbol{\mu}_h(s').$$

Thus, by Theorem 6, with probability at least $1 - \delta/(2HK)$,

$$\begin{aligned}
&\left|\mathsf{cat}_{h,k}\left[(k-1)\mathbf{\Lambda}_{h,k-1}^{-1}\boldsymbol{\phi}_{h,k}\right] - \langle \boldsymbol{\phi}_{h,k}, \int V_{h+1}^k(s')d\boldsymbol{\mu}_h(s')\rangle\right| \\
&\qquad \leq 5\|\boldsymbol{\phi}_{h,k}\|_{\mathbf{\Lambda}_{h,k-1}^{-1}}\left(\sqrt{C_{\text{mdp}} + \log(2HK/\delta)} + \sqrt{\lambda}\|\boldsymbol{\theta}_\star\|_2\right) + \frac{3(C_{\text{mdp}} + \log(2HK/\delta))}{\alpha_{\max}(k-1)}
\end{aligned}$$

Note that,

$$\|\boldsymbol{\theta}_\star\|_2 = \|\int V_{h+1}^k(s')\mathrm{d}\boldsymbol{\mu}_h(s')\|_2 \leq H\|\int |\mathrm{d}\boldsymbol{\mu}_h(s')|\|_2 \leq H\sqrt{d}$$

where the last inequality follows by Definition 3.1. It follows that for $\lambda \leq 1/H^2$ and proper choice of the universal constant in C_{mdp}, we can bound

$$\sqrt{\lambda}\|\boldsymbol{\theta}_\star\|_2 \leq \frac{1}{5}\sqrt{C_{\text{mdp}} + \log(2HK/\delta)}.$$

As $\langle \boldsymbol{\phi}_{h,k}, \int \boldsymbol{\mu}_h(s')V_{h+1}^k(s')ds'\rangle = \mathbb{E}_h[V_{h+1}^k](s_{h,k}, a_{h,k})$ by Definition 3.1, by our choice of β we conclude that with probability at least $1 - \delta/(2HK)$[6],

$$\left|\mathsf{cat}_{h,k}\left[(k-1)\mathbf{\Lambda}_{h,k-1}^{-1}\boldsymbol{\phi}_{h,k}\right] - \mathbb{E}_h[V_{h+1}^k](s_{h,k}, a_{h,k})\right| \leq \beta\|\boldsymbol{\phi}_{h,k}\|_{\mathbf{\Lambda}_{h,k-1}^{-1}} + \mathsf{v}_{\min}\beta^2/k^2.$$

This is precisely the definition of $A_{k,h}$, however, so it follows that

$$\mathbb{P}\left[A_{k,h}^c | \cap_{k'=K_{\text{init}}}^{k-1} \cap_{h'=1}^{H} (B_{k',h'} \cap A_{k',h'}) \cap \left(\cap_{h'=h+1}^{H} B_{k,h'}\right)\right] \leq \frac{\delta}{2HK}.$$

The bound on

$$\mathbb{P}\left[B_{k,h}^c | \left(\cap_{k'=K_{\text{init}}}^{k-1} \cap_{h'=1}^{H} (B_{k',h'} \cap A_{k',h'})\right) \cap \left(\cap_{h'=h+1}^{H} B_{k,h'}\right)\right]$$

can be shown almost identically. As such, we omit the calculation and conclude that

$$\mathbb{P}\left[B_{k,h}^c | \left(\cap_{k'=K_{\text{init}}}^{k-1} \cap_{h'=1}^{H} (B_{k',h'} \cap A_{k',h'})\right) \cap \left(\cap_{h'=h+1}^{H} B_{k,h'}\right)\right] \leq \frac{\delta}{2HK}.$$

Combining these bounds gives that $\mathbb{P}[\mathcal{E}^c] \leq \delta$. □

[5]Note that if $k \geq K_{\text{init}} = 4(C_{\text{mdp}} + \log(2HK/\delta))$, then $K \geq \widetilde{\beta}$, so we can remove $\widetilde{\beta}$ from the definition of d_T as it will be dominated by K.

[6]We have replaced $1/(k-1)$ in the lower order term with $1/k$ for future notational convenience. Note that this is valid since $k \geq K_{\text{init}} > 1$ so we can accommodate this change by slightly increasing the constant in β.

Lemma B.4. *Fix h, $k \geq K_{\text{init}}$, and $k' > k$. Then if $B_{k,h'}$ and $B_{k',h'}$ hold for all $h' \in [h, H]$, we will have*

$$5Q^k_{h'}(s,a) \geq Q^{k'}_{h'}(s,a)$$

for all $h' \in [h, H]$. In particular, $5V^k_{h'}(s) \geq V^{k'}_{h'}(s)$.

Proof. We will prove this by induction. In the base case, take $h' = H$. On $B_{k,H} \cap B_{k',H}$, we have

$$\begin{aligned}
5Q^k_H(s,a) &= \min\left\{5r^k_H(s,a) + 5\langle \boldsymbol{\phi}(s,a), \widehat{\boldsymbol{w}}^k_H\rangle + 15\widetilde{\beta}\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}^{-1}_{H,k-1}} + 15\mathsf{v}_{\min}\widetilde{\beta}^2/k^2, 5H\right\} \\
&\overset{(a)}{\geq} \min\left\{5r^k_h(s,a) + \mathbb{E}_H[5V^k_{H+1}](s,a) + 5\widetilde{\beta}\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}^{-1}_{H,k-1}} + 5\mathsf{v}_{\min}\widetilde{\beta}^2/k^2, 5H\right\} \\
&\overset{(b)}{=} \min\left\{5r^k_h(s,a) + 5\widetilde{\beta}\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}^{-1}_{H,k-1}} + 5\mathsf{v}_{\min}\widetilde{\beta}^2/k^2, 5H\right\} \\
&\overset{(c)}{\geq} \min\left\{r^{k'}_h(s,a) + 5\widetilde{\beta}\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}^{-1}_{H,k'-1}} + 5\mathsf{v}_{\min}\widetilde{\beta}^2/(k')^2, H\right\} \\
&\overset{(d)}{\geq} \min\left\{r^{k'}_h(s,a) + \langle \boldsymbol{\phi}(s,a), \widehat{\boldsymbol{w}}^{k'}_H\rangle + 3\widetilde{\beta}\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}^{-1}_{H,k'-1}} + 3\mathsf{v}_{\min}\widetilde{\beta}^2/(k')^2, H\right\} \\
&= Q^{k'}_H(s,a)
\end{aligned}$$

where (a) follows since we are on $B_{k,H}$ and by Lemma B.6, (b) follows since $V^k_{H+1}(s') = 0$ by definition, (c) follows since reward is non-increasing in k so $r^{k'}_h(s,a) \leq r^k_h(s,a)$, and since $\boldsymbol{\Lambda}_{H,k'-1} \succeq \boldsymbol{\Lambda}_{H,k}$, and (d) follows since we are on $B_{k',H}$, and by Lemma B.6. This implies that, for all s,

$$5V^k_H(s) = 5Q^k_H(s, \pi^k_H(s)) \geq 5Q^k_H(s, \pi^{k'}_H(s)) \geq Q^{k'}_H(s, \pi^{k'}_H(s)) = V^{k'}_H(s). \tag{B.1}$$

For the inductive step, assume that $5V^k_{h'+1}(s) \geq V^{k'}_{h'+1}(s)$ for all s and that $B_{k,h'} \cap B_{k',h'}$ holds. Then we can repeat the above calculation, but now lower bounding

$$\mathbb{E}_{h'}[5V^k_{h'+1}](s,a) \geq \mathbb{E}_{h'}[V^{k'}_{h'+1}](s,a).$$

In full detail,

$$\begin{aligned}
5Q^k_{h'}(s,a) &= \min\left\{5r^k_{h'}(s,a) + 5\langle \boldsymbol{\phi}(s,a), \widehat{\boldsymbol{w}}^k_{h'}\rangle + 15\widetilde{\beta}\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}^{-1}_{h',k-1}} + 15\mathsf{v}_{\min}\widetilde{\beta}^2/k^2, c_{\bar{\mathsf{v}}}H\right\} \\
&\geq \min\left\{5r^k_{h'}(s,a) + \mathbb{E}_{h'}[5V^k_{h'+1}](s,a) + 5\widetilde{\beta}\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}^{-1}_{h',k-1}} + 5\mathsf{v}_{\min}\widetilde{\beta}^2/k^2, 5H\right\} \\
&\geq \min\left\{5r^k_{h'}(s,a) + \mathbb{E}_{h'}[V^{k'}_{h'+1}](s,a) + 5\widetilde{\beta}\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}^{-1}_{h',k-1}} + 5\mathsf{v}_{\min}\widetilde{\beta}^2/k^2, 5H\right\} \\
&\geq \min\left\{r^{k'}_{h'}(s,a) + \langle \boldsymbol{\phi}(s,a), \widehat{\boldsymbol{w}}^{k'}_{h'}\rangle + 3\widetilde{\beta}\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}^{-1}_{h',k'-1}} + 3\mathsf{v}_{\min}\widetilde{\beta}^2/(k')^2, H\right\} \\
&= Q^{k'}_{h'}(s,a).
\end{aligned}$$

It follows that $5V^k_{h'}(s) \geq V^{k'}_{h'}(s)$ by the same argument as in (B.1). This proves the inductive step, so the result follows. □

Lemma B.5. *Set*

$$\bar{\mathsf{v}}^2_{h,k} = \max\left\{20H\mathsf{cat}_{h,k}\left[(k-1)\boldsymbol{\Lambda}^{-1}_{h,k-1}\boldsymbol{\phi}_{h,k}\right] + 20H\beta\|\boldsymbol{\phi}_{h,k}\|_{\boldsymbol{\Lambda}^{-1}_{h,k-1}} + 20H\mathsf{v}_{\min}\beta^2/k^2, \mathsf{v}^2_{\min}\right\}.$$

Then $\bar{\mathsf{v}}_{h,k}^2$ *is* $\mathcal{F}_{h,k}$*-measurable, and, for any* $k' \geq k$*, on the event* $\cap_{h'=h+1}^{H}(B_{k,h'} \cap B_{k',h'}) \cap A_{k,h}$*, we have*

$$\begin{aligned} \mathbb{E}_h[(V_{h+1}^{k'})^2](s_{h,k}, a_{h,k}) &\leq \frac{1}{2}\bar{\mathsf{v}}_{h,k}^2, \\ 4H\mathbb{E}_h[V_{h+1}^{k'}](s_{h,k}, a_{h,k}) &\leq \bar{\mathsf{v}}_{h,k}^2 \end{aligned} \tag{B.2}$$

and

$$\bar{\mathsf{v}}_{h,k}^2 \leq \max\left\{20H\mathbb{E}_h[V_{h+1}^{k}](s_{h,k}, a_{h,k}) + 40H\beta\|\boldsymbol{\phi}_{h,k}\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}} + 40H\mathsf{v}_{\min}\beta^2/k^2, \mathsf{v}_{\min}^2\right\}. \tag{B.3}$$

Proof. By definition $\boldsymbol{\phi}_{h,k}, \boldsymbol{\Lambda}_{h,k-1}$, and $r_{h,k}$ are $\mathcal{F}_{h,k}$-measurable. As we only rely on data up to episode $k-1$, it follows that $\boldsymbol{\phi}_{h,\tau}$ and $s_{h+1,\tau}$ are also $\mathcal{F}_{h,k}$-measurable. Finally, we see from the definition of Algorithm 1 that V_{h+1}^k is formed using only data up to and including episode $k-1$. It follows that $\bar{\mathsf{v}}_{h,k}$ is $\mathcal{F}_{h,k}$-measurable.

Note that we can trivially bound

$$\mathbb{E}_h[(V_{h+1}^{k'})^2](s_{h,k}, a_{h,k}) \leq H\mathbb{E}_h[V_{h+1}^{k'}](s_{h,k}, a_{h,k})$$

where the last inequality follows since $V_{h+1}^{k'}(s') \in [0, H]$. By Lemma B.4, on the event $\cap_{h'=h+1}^{H}(B_{k,h'} \cap B_{k',h'})$, we will have that

$$H\mathbb{E}_h[V_{h+1}^{k'}](s_{h,k}, a_{h,k}) \leq 5H\mathbb{E}_h[V_{h+1}^{k}](s_{h,k}, a_{h,k}).$$

On the event $A_{k,h}$, we can bound

$$\mathbb{E}_h[V_{h+1}^{k}](s_{h,k}, a_{h,k}) \leq \mathsf{cat}_{h,k}\left[(k-1)\boldsymbol{\Lambda}_{h,k-1}^{-1}\boldsymbol{\phi}_{h,k}\right] + \beta\|\boldsymbol{\phi}_{h,k}\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}} + \mathsf{v}_{\min}\beta^2/k^2.$$

The lower bound (B.2) follows by our choice of $\bar{\mathsf{v}}_{h,k}^2$. The upper bound (B.3) follows since, on $A_{k,h}$, we have

$$\mathsf{cat}_{h,k}\left[(k-1)\boldsymbol{\Lambda}_{h,k-1}^{-1}\boldsymbol{\phi}_{h,k}\right] \leq \mathbb{E}_h[V_{h+1}^{k}](s_{h,k}, a_{h,k}) + \beta\|\boldsymbol{\phi}_{h,k}\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}} + \mathsf{v}_{\min}\beta^2/k^2.$$

□

Lemma B.6. *On the event* $B_{k,h}$*, if we are running Algorithm 1, we will have that*

$$\begin{aligned} |\langle\boldsymbol{\phi}(s,a), \widehat{\boldsymbol{w}}_h^k\rangle - \widehat{\mathbb{E}}_h[V_{h+1}^k](s,a)| &\leq \widetilde{\beta}\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}} + \mathsf{v}_{\min}\widetilde{\beta}^2/k^2, \\ |\langle\boldsymbol{\phi}(s,a), \widehat{\boldsymbol{w}}_h^k\rangle - \mathbb{E}_h[V_{h+1}^k](s,a)| &\leq \widetilde{\beta}\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}} + \mathsf{v}_{\min}\widetilde{\beta}^2/k^2 \end{aligned}$$

for all s *and* a*.*

Proof. This follows directly from Lemma 5.2 and Lemma A.11, the definition of $\widetilde{\beta}$ and $B_{k,h}$, and since $\mathbb{E}_h[V_{h+1}^k](s,a)$ is linear in $\boldsymbol{\phi}(s,a)$ and we assume that $\|\boldsymbol{\phi}(s,a)\|_2 \leq 1$ for all s, a and that there does not exist s, a such that $\boldsymbol{\phi}(s,a) = \mathbf{0}$. □

Proof of Claim B.3. Clearly,

$$\bigcup_{k=K_{\text{init}}}^{K}\bigcup_{h=1}^{H}(B_{k,h}^c \cup A_{k,h}^c)$$

$$= \bigcup_{k=K_{\text{init}}}^{K} \bigcup_{h=1}^{H} \left[(B_{k,h}^c \cup A_{k,h}^c) \setminus \left((\cup_{k'=K_{\text{init}}}^{k-1} \cup_{h'=1}^{H} (B_{k',h'}^c \cup A_{k',h'}^c)) \cup (\cup_{h'=h+1}^{H} (B_{k,h'}^c \cup A_{k,h'}^c)) \right) \right]$$

$$= \bigcup_{k=K_{\text{init}}}^{K} \bigcup_{h=1}^{H} \left[B_{k,h}^c \setminus \left((\cup_{k'=K_{\text{init}}}^{k-1} \cup_{h'=1}^{H} (B_{k',h'}^c \cup A_{k',h'}^c)) \cup (\cup_{h'=h+1}^{H} (B_{k,h'}^c \cup A_{k,h'}^c)) \right) \right]$$

$$\cup \bigcup_{k=K_{\text{init}}}^{K} \bigcup_{h=1}^{H} \left[A_{k,h}^c \setminus \left((\cup_{k'=K_{\text{init}}}^{k-1} \cup_{h'=1}^{H} (B_{k',h'}^c \cup A_{k',h'}^c)) \cup (\cup_{h'=h+1}^{H} (B_{k,h'}^c \cup A_{k,h'}^c)) \right) \right].$$

Noting that

$$X \setminus \left((\cup_{k'=K_{\text{init}}}^{k-1} \cup_{h'=1}^{H} (B_{k',h'}^c \cup A_{k',h'}^c)) \cup (\cup_{h'=h+1}^{H} (B_{k,h'}^c \cup A_{k,h'}^c)) \right)$$
$$= X \cap \left((\cap_{k'=K_{\text{init}}}^{k-1} \cap_{h'=1}^{H} (B_{k',h'} \cap A_{k',h'})) \cap (\cap_{h'=h+1}^{H} (B_{k,h'} \cap A_{k,h'})) \right)$$

for any X completes the proof. □

B.3 Optimism

Lemma B.7. *On the event $\mathcal{E}$, for all s, a, h, and $k \geq K_{\text{init}}$ and any π, we have*

$$\widehat{\mathbb{E}}_h[V_{h+1}^k](s,a) + r_h^k(s,a) - Q_h^{k,\pi}(s,a) = \mathbb{E}_h[V_{h+1}^k - V_{h+1}^{k,\pi}](s,a) + \xi_h^k(s,a)$$

where $\xi_h^k(s,a)$ satisfies $|\xi_h^k(s,a)| \leq \beta \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}} + \mathsf{v}_{\min}\beta^2/k^2$.

Proof. By definition, we have that

$$Q_h^{k,\pi}(s,a) = r_h^k(s,a) + \mathbb{E}_h[V_{h+1}^{k,\pi}](s,a).$$

On $\mathcal{E}$, we have that

$$\left| \widehat{\mathbb{E}}_h[V_{h+1}^k](s,a) - \mathbb{E}_h[V_{h+1}^k](s,a) \right| \leq \beta \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}} + \beta^2/k$$

so we can therefore write

$$\widehat{\mathbb{E}}_h[V_{h+1}^k](s,a) = \mathbb{E}_h[V_{h+1}^k](s,a) + \xi_h^k(s,a)$$

for a term $\xi_h^k(s,a)$ satisfying

$$|\xi_h^k(s,a)| \leq \beta \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}} + \mathsf{v}_{\min}\beta^2/k^2.$$

It follows that

$$\widehat{\mathbb{E}}_h[V_{h+1}^k](s,a) + r_h^k(s,a) - Q_h^{\pi}(s,a) = \mathbb{E}_h[V_{h+1}^k - V_{h+1}^{k,\pi}](s,a) + \xi_h^k(s,a).$$

□

Lemma B.8. *On the event $\mathcal{E}$, for all s, a, h, and $k \geq K_{\text{init}}$, we have that $Q_h^k(s,a) \geq Q_h^{k,\star}(s,a)$.*

Proof. We will prove this by induction for a fixed k. First, take $h = H$. Since $V_{H+1}(s) = V^\star_{H+1}(s) = 0$ by definition, by Lemma B.7 we have

$$|\widehat{\mathbb{E}}_H[V^k_{H+1}](s,a) + r^k_H(s,a) - Q^{k,\star}_H(s,a)| \le \beta\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}^{-1}_{H,k-1}} + \mathsf{v}_{\min}\beta^2/k^2$$

which implies

$$\begin{aligned}
Q^{k,\star}_H(s,a) &\le \min\{r^k_h(s,a) + \widehat{\mathbb{E}}_H[V^k_{H+1}](s,a) + \beta\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}^{-1}_{H,k-1}} + \mathsf{v}_{\min}\beta^2/k^2, H\} \\
&\le \min\{r^k_h(s,a) + \langle\boldsymbol{\phi}(s,a), \widehat{\boldsymbol{w}}^k_H\rangle + (\beta+\widetilde{\beta})\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}^{-1}_{H,k-1}} + \mathsf{v}_{\min}(\beta^2+\widetilde{\beta}^2)/k^2, H\} \\
&\le \min\{r^k_h(s,a) + \langle\boldsymbol{\phi}(s,a), \widehat{\boldsymbol{w}}^k_H\rangle + 3\widetilde{\beta}\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}^{-1}_{H,k-1}} + 3\mathsf{v}_{\min}\widetilde{\beta}^2/k^2, H\} \\
&= Q^k_H(s,a)
\end{aligned}$$

where we have used Lemma B.6 and that $\beta \le \widetilde{\beta}$. Now assume that $Q^k_{h+1}(s,a) \ge Q^{k,\star}_{h+1}(s,a)$ for all (s,a) and some h. Again by Lemma B.7, we have that

$$|\widehat{\mathbb{E}}_h[V^k_{h+1}](s,a) - Q^{k,\star}_h(s,a) - \mathbb{E}_h[V^k_{h+1} - V^{k,\star}_{h+1}](s,a)| \le \beta\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}^{-1}_{h,k-1}} + \mathsf{v}_{\min}\beta^2/k^2.$$

By the inductive hypothesis $\mathbb{E}_h[V^k_{h+1} - V^{k,\star}_{h+1}](s,a) \ge 0$, so

$$\begin{aligned}
Q^{k,\star}_h(s,a) &\le \min\{r^k_h(s,a) + \widehat{\mathbb{E}}_h[V^k_{h+1}](s,a) + \beta\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}^{-1}_{h,k-1}} + \mathsf{v}_{\min}\beta^2/k^2, H\} \\
&\le \min\{r^k_h(s,a) + \langle\boldsymbol{\phi}(s,a), \widehat{\boldsymbol{w}}^k_h\rangle + (\beta+\widetilde{\beta})\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}^{-1}_{h,k-1}} + \mathsf{v}_{\min}(\beta^2+\widetilde{\beta}^2)/k^2, H\} \\
&= \min\{r^k_h(s,a) + \langle\boldsymbol{\phi}(s,a), \widehat{\boldsymbol{w}}^k_h\rangle + 3\widetilde{\beta}\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}^{-1}_{h,k-1}} + 3\mathsf{v}_{\min}\widetilde{\beta}^2/k^2, H\} \\
&= Q^k_h(s,a).
\end{aligned}$$

This proves the inductive hypothesis so the result follows. □

Lemma B.9 (Formal version of Lemma 6.1)**.** *Let $\delta^k_h = V^k_h(s^k_h) - V^{k,\pi_k}_h(s^k_h)$ and $\zeta^k_{h+1} = \mathbb{E}_h[\delta^k_{h+1}](s_{h,k}, a_{h,k}) - \delta^k_{h+1}$. Then, on the event $\mathcal{E}$, for any $k \ge K_{\text{init}}$,*

$$\delta^k_h \le \delta^k_{h+1} + \zeta^k_{h+1} + \min\{5\widetilde{\beta}\|\boldsymbol{\phi}_{h,k}\|_{\boldsymbol{\Lambda}^{-1}_{h,k-1}} + 5\mathsf{v}_{\min}\widetilde{\beta}^2/k^2, H\}.$$

Proof. We have

$$\begin{aligned}
Q^k_h(s,a) - Q^{k,\pi_k}_h(s,a) &\overset{(a)}{=} \min\{r^k_h(s,a) + \langle\boldsymbol{\phi}(s,a), \widehat{\boldsymbol{w}}^k_h\rangle + 3\beta\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}^{-1}_{h,k-1}} + 3\mathsf{v}_{\min}\beta^2/k^2, H\} - Q^{k,\pi_k}_h(s,a) \\
&\overset{(b)}{\le} \min\{r^k_h(s,a) + \langle\boldsymbol{\phi}(s,a), \widehat{\boldsymbol{w}}^k_h\rangle - Q^{k,\pi_k}_h(s,a) + 3\widetilde{\beta}\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}^{-1}_{h,k-1}} + 3\mathsf{v}_{\min}\widetilde{\beta}^2/k^2, H\} \\
&\overset{(c)}{\le} \min\{r^k_h(s,a) + \widehat{\mathbb{E}}_h[V^k_{h+1}](s,a) - Q^{k,\pi_k}_h(s,a) + 4\widetilde{\beta}\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}^{-1}_{h,k-1}} + 4\mathsf{v}_{\min}\widetilde{\beta}^2/k^2, H\} \\
&\overset{(d)}{\le} \min\{\mathbb{E}_h[V^k_{h+1} - V^{k,\pi_k}_{h+1}](s,a) + 5\widetilde{\beta}\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}^{-1}_{h,k-1}} + 5\mathsf{v}_{\min}\widetilde{\beta}^2/k^2, H\} \\
&\overset{(e)}{\le} \mathbb{E}_h[V^k_{h+1} - V^{k,\pi_k}_{h+1}](s,a) + \min\{5\widetilde{\beta}\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}^{-1}_{h,k-1}} + 5\mathsf{v}_{\min}\widetilde{\beta}^2/k^2, H\}
\end{aligned}$$

where (a) is by definition of $Q^k_h(s,a)$, (b) holds since $Q^{k,\pi_k}_h(s,a) \ge 0$, (c) holds by Lemma B.6, (d) follows by Lemma B.7, and (e) follows since $\mathbb{E}_h[V^k_{h+1} - V^{k,\pi_k}_{h+1}](s,a) \ge 0$ by Lemma B.8.

Now note that since at episode k we play action $a_h^k = \arg\max_a Q_h^k(s_h^k, a)$, we will have that

$$\delta_h^k = Q_h^k(s_h^k, a_h^k) - Q_h^{k,\pi_k}(s_h^k, a_h^k).$$

The result follows by the definition of $V_h^k(s)$ and $V_h^{k,\pi_k}(s)$. □

B.4 Regret Bound

Lemma B.10. *With probability at least $1-\delta$, we can bound*

$$\sum_{k=K_{\text{init}}}^{K}\sum_{h=1}^{H}\zeta_h^k \le 2\sqrt{8H\sum_{k=K_{\text{init}}}^{K}\sum_{h=1}^{H}\mathbb{E}_{h-1}[V_h^k](s_{h-1,k}, a_{h-1,k})\cdot\log 1/\delta} + 2H\log 1/\delta.$$

Proof. This is a direct application of Lemma A.4, Freedman's inequality. Recall that

$$\zeta_h^k = \mathbb{E}_{h-1}[\delta_h^k](s_{h-1,k}, a_{h-1,k}) - \delta_h^k = \mathbb{E}_{h-1}[V_h^k - V_h^{k,\pi_k}](s_{h-1,k}, a_{h-1,k}) - (V_h^k(s_h^k) - V_h^{k,\pi_k}(s_h^k)).$$

Thus, we can bound

$$|\zeta_h^k| \le 2H$$

since the value function will always be bounded in $[0, H]$. Next, note that

$$\begin{aligned}
\mathbb{E}_{h-1}[(\zeta_h^k)^2](s_{h-1,k}, a_{h-1,k}) &\le 2\mathbb{E}_{h-1}[(V_h^k - V_h^{k,\pi_k})^2](s_{h-1,k}, a_{h-1,k}) \\
&\le 4\mathbb{E}_{h-1}[(V_h^k)^2 + (V_h^{k,\pi_k})^2](s_{h-1,k}, a_{h-1,k}) \\
&\le 8\mathbb{E}_{h-1}[(V_h^k)^2](s_{h-1,k}, a_{h-1,k}) \\
&\le 8H\mathbb{E}_{h-1}[V_h^k](s_{h-1,k}, a_{h-1,k})
\end{aligned}$$

where the second to last inequality uses Lemma B.8. Using these bounds the result then follows directly from Lemma A.4. □

Lemma B.11. *With probability at least $1-\delta$, we have*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{h-1}[V_h^k](s_{h-1,k}, a_{h-1,k}) \le H\cdot\left(\sum_{k=K_{\text{init}}}^{K} V_1^{k,\star} + \widetilde{\mathcal{R}}_K + 2\sqrt{\left(\sum_{k=K_{\text{init}}}^{K} V_1^{k,\star} + \widetilde{\mathcal{R}}_K\right)\cdot\log 1/\delta} + \log 1/\delta\right)$$

where $\widetilde{\mathcal{R}}_K = \sum_{k=1}^{K}(V_1^k(s_1) - V_1^{k,\pi_k}(s_1))$.

Proof. By Lemma B.6, on $\mathcal{E}$,

$$|\langle\boldsymbol{\phi}(s,a), \widehat{\boldsymbol{w}}_h^k\rangle - \mathbb{E}_h[V_{h+1}^k](s,a)| \le \widetilde{\beta}\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}} + \mathsf{v}_{\min}\widetilde{\beta}^2/k^2$$

which implies that

$$\mathbb{E}_h[V_{h+1}^k](s,a) \le \langle\boldsymbol{\phi}(s,a), \widehat{\boldsymbol{w}}_h^k\rangle + \widetilde{\beta}\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}} + \mathsf{v}_{\min}\widetilde{\beta}^2/k^2.$$

Thus,

$$Q_h^k(s,a) = \min\{r_h^k(s,a) + \langle\boldsymbol{\phi}(s,a), \widehat{\boldsymbol{w}}_h^k\rangle + 3\widetilde{\beta}\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}} + 3\mathsf{v}_{\min}\widetilde{\beta}^2/k^2, H\}$$

$$\geq \min\{r_h^k(s,a) + \mathbb{E}_h[V_{h+1}^k](s,a), H\}.$$

Since $\pi_h^k(s) = \arg\max_a Q_h^k(s,a)$, we have that $V_{h+1}^k(s') = Q_{h+1}^k(s', \pi_{h+1}^k(s'))$. Using that reward is always nonnegative, we can therefore unroll $V_1^k(s_1)$ backwards as:

$$\begin{aligned}
V_1^k(s_1) &= Q_1^k(s_1, \pi_1^k(s_1)) \\
&\geq \min\{r_1^k(s_1, \pi_1^k(s_1)) + \mathbb{E}_{s_2}[Q_2^k(s_2, \pi_2^k(s_2)) \mid s_1, \pi_1^k(s_1)], H\} \\
&\geq \min\{\mathbb{E}_{s_2}[Q_2^k(s_2, \pi_2^k(s_2)) \mid s_1, \pi_1^k(s_1)], H\} \\
&= \mathbb{E}_{s_2}[Q_2^k(s_2, \pi_2^k(s_2)) \mid s_1, \pi_1^k(s_1)] \\
&\geq \mathbb{E}_{s_2}[\min\{r_2^k(s_2, \pi_2^k(s_2)) + \mathbb{E}_{s_3}[Q_3^k(s_3, \pi_3^k(s_3)) \mid s_2, \pi_2^k(s_2)], H\} \mid s_1, \pi_1^k(s_1)] \\
&\geq \mathbb{E}_{s_2}[\mathbb{E}_{s_3}[Q_3^k(s_3, \pi_3^k(s_3)) \mid s_2, \pi_2^k(s_2)] \mid s_1, \pi_1^k(s_1)] \\
&= \mathbb{E}_{\pi^k}[Q_3^k(s_3, \pi_3^k(s_3))] \\
&\vdots \\
&\geq \mathbb{E}_{\pi^k}[Q_h^k(s_h, \pi_h^k(s_h))] \\
&= \mathbb{E}_{\pi^k}[V_h^k(s_h)]
\end{aligned}$$

where here $\mathbb{E}_{s'}[\cdot \mid s, a]$ denotes taking the expectation over the next state s' given that we are in (s,a), and $\mathbb{E}_{\pi^k}[\cdot]$ denotes the expectation over trajectories generated by π_k. We conclude

$$V_1^k(s_1) \geq \mathbb{E}_{\pi^k}[V_h^k(s_h)]$$

for any h. Given this, since we play policy π_k at episode k, we will have that

$$\mathbb{E}_{\pi_k}\mathbb{E}_{h-1}[V_h^k](s_{h-1,k}, a_{h-1,k}) = \mathbb{E}_{\pi_k}[V_h^k(s_h^k)] \leq \mathbb{E}_{\pi_k}[V_1^k(s_1^k)] = V_1^k(s_1)$$

which allows us to bound

$$\begin{aligned}
&\sum_{k=K_{\text{init}}}^{K} \sum_{h=1}^{H} \mathbb{E}_{h-1}[V_h^k](s_{h-1,k}, a_{h-1,k}) \\
&= \sum_{k=K_{\text{init}}}^{K} \sum_{h=1}^{H} \mathbb{E}_{\pi_k}\mathbb{E}_{h-1}[V_h^k](s_{h-1,k}, a_{h-1,k}) + \sum_{k=K_{\text{init}}}^{K} \sum_{h=1}^{H} (\mathbb{E}_{h-1}[V_h^k](s_{h-1,k}, a_{h-1,k}) - \mathbb{E}_{\pi_k}\mathbb{E}_{h-1}[V_h^k](s_{h-1,k}, a_{h-1,k})) \\
&\leq H \sum_{k=K_{\text{init}}}^{K} V_1^k(s_1) + \sum_{k=K_{\text{init}}}^{K} \sum_{h=1}^{H} (\mathbb{E}_{h-1}[V_h^k](s_{h-1,k}, a_{h-1,k}) - \mathbb{E}_{\pi_k}\mathbb{E}_{h-1}[V_h^k](s_{h-1,k}, a_{h-1,k})).
\end{aligned}$$

By definition of $\widetilde{\mathcal{R}}_K$,

$$H \sum_{k=K_{\text{init}}}^{K} V_1^k(s_1) = H \sum_{k=K_{\text{init}}}^{K} V_1^{k,\pi_k}(s_1) + H\widetilde{\mathcal{R}}_K \leq H \sum_{k=K_{\text{init}}}^{K} V_1^{k,\star}(s_1) + H\widetilde{\mathcal{R}}_K.$$

It remains to bound

$$\sum_{k=K_{\text{init}}}^{K} \sum_{h=1}^{H} (\mathbb{E}_{h-1}[V_h^k](s_{h-1,k}, a_{h-1,k}) - \mathbb{E}_{\pi_k}\mathbb{E}_{h-1}[V_h^k](s_{h-1,k}, a_{h-1,k})).$$

Note first that $|\mathbb{E}_{h-1}[V_h^k](s_{h-1,k}, a_{h-1,k}) - \mathbb{E}_{\pi_k}\mathbb{E}_{h-1}[V_h^k](s_{h-1,k}, a_{h-1,k})| \le H$ almost surely,

$$\mathbb{E}_{\pi_k}[\mathbb{E}_{h-1}[V_h^k](s_{h-1,k}, a_{h-1,k}) - \mathbb{E}_{\pi_k}\mathbb{E}_{h-1}[V_h^k](s_{h-1,k}, a_{h-1,k})] = 0,$$

and

$$\begin{aligned}\mathbb{E}_{\pi_k}[(\mathbb{E}_{h-1}[V_h^k](s_{h-1,k}, a_{h-1,k}) - \mathbb{E}_{\pi_k}\mathbb{E}_{h-1}[V_h^k](s_{h-1,k}, a_{h-1,k}))^2] &\le \mathbb{E}_{\pi_k}[\mathbb{E}_{h-1}[V_h^k](s_{h-1,k}, a_{h-1,k})^2] \\ &\le H\mathbb{E}_{\pi_k}[\mathbb{E}[\mathbb{E}_{h-1}[V_h^k](s_{h-1,k}, a_{h-1,k})] \\ &\le HV_1^k(s_1)\end{aligned}$$

where the last inequality follows by what we have shown above. Applying Freedman's inequality (Lemma A.4), we can then bound, with probability at least $1-\delta$,

$$\begin{aligned}&\sum_{k=K_{\text{init}}}^{K}\sum_{h=1}^{H}(\mathbb{E}_{h-1}[V_h^k](s_{h-1,k}, a_{h-1,k}) - \mathbb{E}_{\pi_k}\mathbb{E}_{h-1}[V_h^k](s_{h-1,k}, a_{h-1,k})) \\ &\qquad\le 2\sqrt{H^2\sum_{k=K_{\text{init}}}^{K} V_1^k(s_1)\cdot\log 1/\delta + H\log 1/\delta} \\ &\qquad\le 2\sqrt{(H^2\sum_{k=K_{\text{init}}}^{K} V_1^{k,\star}(s_1) + H^2\widetilde{\mathcal{R}}_K)\cdot\log 1/\delta + H\log 1/\delta}\end{aligned}$$

where the last inequality follows by what we have shown above. □

Proof of Theorem 8. By definition of $\mathcal{R}_K$ and Lemma B.8,

$$\mathcal{R}_K = \sum_{k=1}^{K}(V_1^{k,\star}(s_1) - V_1^{k,\pi_k}(s_1)) \le HK_{\text{init}} + \sum_{k=K_{\text{init}}}^{K}(V_1^k(s_1) - V_1^{k,\pi_k}(s_1)) =: HK_{\text{init}} + \widetilde{\mathcal{R}}_K.$$

Decomposing the regret. By Lemma B.9,

$$\begin{aligned}\widetilde{\mathcal{R}}_K &\le \sum_{k=K_{\text{init}}}^{K}\sum_{h=1}^{H}\zeta_h^k + \sum_{k=K_{\text{init}}}^{K}\sum_{h=1}^{H}\min\{5\widetilde{\beta}\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}} + 5\mathsf{v}_{\min}\widetilde{\beta}^2/k^2, H\} \\ &\le \sum_{k=K_{\text{init}}}^{K}\sum_{h=1}^{H}\zeta_h^k + \sum_{k=K_{\text{init}}}^{K}\sum_{h=1}^{H}\min\{5\widetilde{\beta}\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}}, H\} + \sum_{k=K_{\text{init}}}^{K}\sum_{h=1}^{H}5\mathsf{v}_{\min}\widetilde{\beta}^2/k^2.\end{aligned}$$

By Lemma B.10, with probability $1-\delta$, $\sum_{k=K_{\text{init}}}^{K}\sum_{h=1}^{H}\zeta_h^k$ can be bounded as

$$\sum_{k=K_{\text{init}}}^{K}\sum_{h=1}^{H}\zeta_h^k \le \sqrt{32H\sum_{k=K_{\text{init}}}^{K}\sum_{h=1}^{H}\mathbb{E}_{h-1}[V_h^k](s_{h-1,k}, a_{h-1,k})\cdot\log 1/\delta + 2H\log 1/\delta}.$$

Furthermore,

$$\sum_{k=K_{\text{init}}}^{K}\sum_{h=1}^{H}5\mathsf{v}_{\min}\widetilde{\beta}^2/k^2 \le 10\widetilde{\beta}^2 H\mathsf{v}_{\min}$$

Controlling the optimistic bonuses. Let $\mathcal{K}_h = \{k \geq K_{\text{init}} : \|\boldsymbol{\phi}_{h,k}/\bar{\mathsf{v}}_{h,k}\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}} \leq 1\}$ and $\mathcal{K}_h^c = \{K_{\text{init}}, \dots, K\} \backslash \mathcal{K}_h$. Then,

$$\begin{aligned}\sum_{k=K_{\text{init}}}^{K}\sum_{h=1}^{H} \min\{5\widetilde{\beta}\|\boldsymbol{\phi}_{h,k}\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}}, H\} &= \sum_{k=K_{\text{init}}}^{K}\sum_{h=1}^{H} \bar{\mathsf{v}}_{h,k} \min\{5\widetilde{\beta}\|\boldsymbol{\phi}_{h,k}/\bar{\mathsf{v}}_{h,k}\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}}, H/\bar{\mathsf{v}}_{h,k}\} \\ &\leq 5\widetilde{\beta}\sum_{h=1}^{H}\sum_{k\in\mathcal{K}_h} \bar{\mathsf{v}}_{h,k}\|\boldsymbol{\phi}_h^k/\bar{\mathsf{v}}_{h,k}\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}} + \sum_{h=1}^{H} H|\mathcal{K}_h^c|.\end{aligned}$$

By Lemma 6.2, and since $\|\boldsymbol{\phi}_h^k/\bar{\mathsf{v}}_{h,k}\|_2 \leq 1/\mathsf{v}_{\min}$ almost surely, we can bound $|\mathcal{K}_h^c| \leq 2d\log(1 + K/(\lambda\mathsf{v}_{\min}^2))$, which implies that

$$\sum_{h=1}^{H} H|\mathcal{K}_h^c| \leq 2dH^2\log(1 + K/(\lambda\mathsf{v}_{\min}^2)).$$

Denote

$$\eta_{h,\tau} := 20H\mathbb{E}_h[V_{h+1}^{\tau}](s_{h,\tau}, a_{h,\tau}).$$

By Lemma B.5 and the definition of $\bar{\mathsf{v}}_{h,\tau}^2$, we can bound

$$\begin{aligned}&5\widetilde{\beta}\sum_{h=1}^{H}\sum_{k\in\mathcal{K}_h} \bar{\mathsf{v}}_{h,k}\|\boldsymbol{\phi}_{h,k}/\bar{\mathsf{v}}_{h,k}\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}} \\ &\quad= 5\widetilde{\beta}\sum_{h=1}^{H}\sum_{k\in\mathcal{K}_h} \frac{\bar{\mathsf{v}}_{h,k}^2}{\bar{\mathsf{v}}_{h,k}}\|\boldsymbol{\phi}_{h,k}/\bar{\mathsf{v}}_{h,k}\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}} \\ &\quad\leq 5\widetilde{\beta}\sum_{h=1}^{H}\sum_{k\in\mathcal{K}_h} \frac{\eta_{h,k} + 20H\beta\|\boldsymbol{\phi}_{h,k}\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}} + 20H\mathsf{v}_{\min}\beta^2/k^2 + \mathsf{v}_{\min}^2}{\bar{\mathsf{v}}_{h,k}}\|\boldsymbol{\phi}_{h,k}/\bar{\mathsf{v}}_{h,k}\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}} \\ &\quad\leq 5\widetilde{\beta}\sum_{h=1}^{H}\sum_{k\in\mathcal{K}_h} \left((\sqrt{5\eta_{h,k}} + 20H\beta^2/k^2 + \mathsf{v}_{\min})\|\boldsymbol{\phi}_{h,k}/\bar{\mathsf{v}}_{h,k}\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}} + 20H\beta\|\boldsymbol{\phi}_{h,k}/\bar{\mathsf{v}}_{h,k}\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}}^2\right)\end{aligned}$$

where the final inequality follows since, by Lemma B.5, we can lower bound $\bar{\mathsf{v}}_{h,k}^2 \geq \eta_{h,k}/5$, and since we can always lower bound $\bar{\mathsf{v}}_{h,k} \geq \mathsf{v}_{\min}$.

Recalling the definition of $\mathcal{K}_h$, we can bound

$$\begin{aligned}5\widetilde{\beta}\sum_{h=1}^{H}\sum_{k\in\mathcal{K}_h} 20H\beta\|\boldsymbol{\phi}_{h,k}/\bar{\mathsf{v}}_{h,k}\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}}^2 &\leq 100H\widetilde{\beta}\beta\sum_{h=1}^{H}\sum_{k=1}^{K}\min\{\|\boldsymbol{\phi}_h^k/\bar{\mathsf{v}}_{h,k}\|_{\boldsymbol{\Lambda}_{h,k}^{-1}}^2, 1\} \\ &\leq 200H^2\widetilde{\beta}\beta d\log(1 + K/(d\lambda\mathsf{v}_{\min}^2))\end{aligned}$$

where the last inequality follows by Lemma A.3. By Cauchy-Schwarz and again using the definition of $\mathcal{K}_h$, we can bound

$$5\widetilde{\beta}\sum_{h=1}^{H}\sum_{k\in\mathcal{K}_h} (\sqrt{5\eta_{h,k}} + 20H\beta^2/k^2 + \mathsf{v}_{\min})\|\boldsymbol{\phi}_h^k/\bar{\mathsf{v}}_{h,k}\|_{\boldsymbol{\Lambda}_{h,k}^{-1}}$$

$$\leq 5\widetilde{\beta}\sqrt{4\sum_{h=1}^{H}\sum_{k\in\mathcal{K}_h}(5\eta_{h,k}+400H^2\beta^4/k^4+\mathsf{v}_{\min}^2)}\sqrt{\sum_{h=1}^{H}\sum_{k\in\mathcal{K}_h}\|\boldsymbol{\phi}_h^k/\bar{\mathsf{v}}_{h,k}\|^2_{\boldsymbol{\Lambda}_{h,k}^{-1}}}$$

$$\leq 5\widetilde{\beta}\sqrt{4\sum_{h=1}^{H}\sum_{k\in\mathcal{K}_h}(5\eta_{h,k}+400H^2\beta^4/k^4+\mathsf{v}_{\min}^2)}\sqrt{\sum_{h=1}^{H}\sum_{k=1}^{K}\min\{\|\boldsymbol{\phi}_h^k/\bar{\mathsf{v}}_{h,k}\|^2_{\boldsymbol{\Lambda}_{h,k}^{-1}},1\}}$$

$$\leq 5\widetilde{\beta}\sqrt{2dH\log(1+K/(d\lambda\mathsf{v}_{\min}^2))}\sqrt{40\sum_{h=1}^{H}\sum_{k=K_{\text{init}}}^{K}\eta_{h,k}+3200H^3\beta^4+4HK\mathsf{v}_{\min}^2}$$

$$\leq 5\widetilde{\beta}\sqrt{2dH\log(1+K/(d\lambda\mathsf{v}_{\min}^2))}\left(\sqrt{40\sum_{h=1}^{H}\sum_{k=K_{\text{init}}}^{K}\eta_{h,k}}+60H^{3/2}\beta^2+2\sqrt{HK\mathsf{v}_{\min}^2}\right)$$

where we again apply Lemma A.3 and use that $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$ for $a,b\geq 0$.

Finishing the Proof. By definition,

$$\sum_{h=1}^{H}\sum_{k=K_{\text{init}}}^{K}\eta_{h,k}=\sum_{h=1}^{H}\sum_{k=K_{\text{init}}}^{K}20H\mathbb{E}_h[V_{h+1}^k](s_{h,k},a_{h,k})\leq\sum_{h=1}^{H}\sum_{k=K_{\text{init}}}^{K}20H\mathbb{E}_{h-1}[V_h^k](s_{h-1,k},a_{h-1,k}).$$

Collecting terms, we have then shown that,

$$\begin{aligned}\widetilde{\mathcal{R}}_K &\leq c_1\widetilde{\beta}\sqrt{dH\log(1+K/(d\lambda\mathsf{v}_{\min}^2))}\sqrt{H\sum_{h=1}^{H}\sum_{k=K_{\text{init}}}^{K}\mathbb{E}_{h-1}[V_h^k](s_{h-1,k},a_{h-1,k})}\\&\quad+c_2\widetilde{\beta}\sqrt{dH\log(1+K/(d\lambda\mathsf{v}_{\min}^2))}\sqrt{H\mathsf{v}_{\min}^2K}\\&\quad+c_3\widetilde{\beta}\beta^2H^2\sqrt{d}\log(1+K/(d\lambda\mathsf{v}_{\min}^2))\end{aligned}$$

for universal constants c_1,c_2,c_3. By Lemma B.11 we can bound, with probability at least $1-\delta$,

$$\begin{aligned}\sum_{k=K_{\text{init}}}^{K}\sum_{h=1}^{H}\mathbb{E}[V_{h-1}^k](s_{h-1,k},a_{h-1,k})&\leq H\cdot\left(\sum_{k=1}^{K}V_1^{k,\star}+\widetilde{\mathcal{R}}_K+2\sqrt{\left(\sum_{k=1}^{K}V_1^{k,\star}+\widetilde{\mathcal{R}}_K\right)\cdot\log 1/\delta}+\log 1/\delta\right)\\&\leq 4H\log 1/\delta\cdot\left(\sum_{k=1}^{K}V_1^{k,\star}+\widetilde{\mathcal{R}}_K\right)\end{aligned}$$

so

$$\begin{aligned}\widetilde{\mathcal{R}}_K &\leq c_1\widetilde{\beta}\sqrt{dH\log(1+K/(d\lambda\mathsf{v}_{\min}^2))}\left(\sqrt{H^2\log 1/\delta\cdot\left(\sum_{k=1}^{K}V_1^{k,\star}+\widetilde{\mathcal{R}}_K\right)}+\sqrt{H\mathsf{v}_{\min}^2K}\right)\\&\quad+c_3\widetilde{\beta}\beta^2H^2\sqrt{d}\log(1+K/(d\lambda\mathsf{v}_{\min}^2)).\end{aligned}$$

Finally, choosing $\mathsf{v}_{\min}^2=1/K$ and solving the above for $\widetilde{\mathcal{R}}_K$ gives

$$\widetilde{\mathcal{R}}_K\leq c_1\widetilde{\beta}\sqrt{dH\log(1+K/(d\lambda\mathsf{v}_{\min}^2))}\sqrt{H^2\log 1/\delta\cdot\sum_{k=1}^{K}V_1^{k,\star}}+c_2\widetilde{\beta}\beta^2H^3\sqrt{d}\log(1+K/(d\lambda\mathsf{v}_{\min}^2))\cdot\log 1/\delta.$$

Since $\mathcal{R}_K \le HK_{\text{init}} + \widetilde{\mathcal{R}}_K$, union bounding over $\mathcal{E}$, which holds with probability at least $1-\delta$ by Lemma B.2, and the two additional events stated above, and using that $\beta = 6\sqrt{C_{\text{mdp}} + \log(2HK/\delta)}$ and

$$C_{\text{mdp}} := c(d^2 + d_{\mathscr{R}}) \cdot \mathsf{logs}\left(d, \mathsf{v}_{\min}^{-1}, H, 1/\lambda, K, R_{\mathscr{R}}\right),$$

and the definition of $\widetilde{\beta}$, and setting $\lambda = 1/H^2$, gives the final result. □