# No-Regret Reinforcement Learning in Smooth MDPs

**Davide Maran**[1] **Alberto Maria Metelli**[1] **Matteo Papini**[1] **Marcello Restelli**[1]

## Abstract

Obtaining no-regret guarantees for reinforcement learning (RL) in the case of problems with continuous state and/or action spaces is still one of the major open challenges in the field. Recently, a variety of solutions have been proposed, but besides very specific settings, the general problem remains unsolved. In this paper, we introduce a novel structural assumption on the Markov decision processes (MDPs), namely $\nu-$smoothness, that generalizes most of the settings proposed so far (e.g., linear MDPs and Lipschitz MDPs). To face this challenging scenario, we propose two algorithms for regret minimization in $\nu-$smooth MDPs. Both algorithms build upon the idea of constructing an MDP representation through an orthogonal feature map based on Legendre polynomials. The first algorithm, LEGENDRE-ELEANOR, archives the no-regret property under weaker assumptions but is computationally inefficient, whereas the second one, LEGENDRE-LSVI, runs in polynomial time, although for a smaller class of problems. After analyzing their regret properties, we compare our results with state-of-the-art ones from RL theory, showing that our algorithms achieve the best guarantees.

## 1. Introduction

*Reinforcement learning* (RL) (Sutton & Barto, 2018) is a paradigm of artificial intelligence in which the agent interacts with an environment to maximize a reward signal in the long term. From the theoretical perspective, a lot of effort has been put into designing algorithms with small *(cumulative) regret*, which is an index of how much the policies (i.e., the behavior) played by the algorithm during the learning process are suboptimal. For the case of *tabular Markov decision processes* (MDPs), an optimal result was first proved by Azar et al. (2017), who showed a bound on the regret of order $\widetilde{\mathcal{O}}(H\sqrt{|\mathcal{S}||\mathcal{A}|K})$, where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space, and $K$ is the number of episodes, and $H$ the time horizon of every episode. This regret is minimax-optimal, in the sense that no algorithm can achieve smaller regret for every arbitrary tabular MDP. Unfortunately, assuming that the state-action space is finite is extremely restrictive, as the number of states and/or actions can be huge or even infinite in practice. This is especially critical for a large variety of real-world scenarios in which RL has achieved successful results, including robotics (Kober et al., 2013), autonomous driving (Kiran et al., 2021), and trading (Hambly et al., 2023). These scenarios are usually modeled as MDPs with continuous state and/or action spaces, as the underlying dynamics is too complex to be captured by a finite number of states and/or actions. It is not by chance that one of the most common benchmarks for RL algorithms, MUJOCO (Todorov et al., 2012; Brockman et al., 2016), is composed of environments characterized by continuous state and action spaces. This highlights the notable gap between the current maturity of theory and the pressing needs of practical application. For this reason, devising algorithms with regret bounds for RL in *continuous* spaces is currently one of the most important challenges of the whole field.

Since, without any further assumption, the RL problem in continuous spaces is *non-learnable*,[1] the modern literature revolves around searching for the weakest *structural assumptions* under which the problem can be solved efficiently. *Linear quadratic regulator* (LQR) (Bemporad et al., 2002) is a model for the environment that is widely used in control theory, where the state of the system evolves according to a linear dynamical system and the reward is quadratic. For the online control of this problem, when the system matrix is unknown, regret bound of order $\widetilde{\mathcal{O}}(\sqrt{K})$ were obtained by Abbasi-Yadkori & Szepesvári (2011) for a computationally inefficient algorithm. This limitation was then removed by Dean et al. (2018); Cohen et al. (2019). *Linear MDPs* (Yang & Wang, 2019; Jin et al., 2020) is a widespread setting in RL theory where another form of linearity is assumed. Different from LQRs, here the transition kernel of the MDP can be factorized as a scalar product

[*] Equal contribution [1] Politecnico di Milano, Milan, Italy. Correspondence to: Davide Maran <davide.maran@polimi.it>.

*Proceedings of the $41^{st}$ International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1] Think for example at searching for a maximum of a noisy reward function with infinitely many jumps.

between a feature map $\varphi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and an unknown vector of finite measures over $\mathcal{S}$. The reward function is typically assumed to be linear in the same features. When the feature map is known, regret bounds of order $\widetilde{\mathcal{O}}(\sqrt{d^3K})$ are possible (Jin et al., 2020). Still, these are two examples of *parametric* settings, which do not constitute a reasonable assumption for general continuous-space MDPs. A much wider family can be defined by just assuming that small variations in the state-action pair $(s, a)$ lead to $(i)$ small variations in the reward function $r(s, a)$ $(ii)$ small variations in the transition function $p(\cdot|s, a)$, as it is assumed in the setting of *Lipschitz MDPs* (Rachelson & Lagoudakis, 2010). Lipschitz MDPs have been applied to a number of different settings. Not only do they allow developing theoretically grounded algorithms (Pirotta et al., 2015; Asadi et al., 2018; Metelli et al., 2020), they also help to tackle generalizations of standard RL, such as RL with delayed feedback (Liotet et al., 2022) or configurable RL (Metelli, 2022), and auxiliary tasks for imitation learning (Damiani et al., 2022; Maran et al., 2023). The price of being very general is paid with a regret bound that is much worse than that of previous families. Indeed, no algorithm can achieve a better regret bound than $\Omega(K^{\frac{d+1}{d+2}})$, in terms of dependence on $K$, being $d$ the dimensionality of the state-action space. This entails a huge performance detriment compared with Linear MDPs and LQRs, where the order of the regret in $K$ is $\frac{1}{2}$, regardless of the dimension $d$. In fact, there is still a large gap in the theory between parametric families of MDPs and Lipschitz MDPs, and little is known about what lies in between. One last family of continuous-state MDPs for which regret bounds exist is that of *Kernelized MDPs* (Yang et al., 2020b), where both the reward function and the transition function belong to a *reproducing kernel Hilbert space* (RKHS) induced by a known kernel. In the typical application to continuous-state MDPs, the kernel is assumed to come from the Matérn covariance function with parameter $m > 0$. The higher the value of $m$, the more stringent the assumption, as the corresponding RHKS contains fewer functions. Coherently, regret bounds for this setting decrease with $m$. In particular, it was very recently proved (Vakili & Olkhovskaya, 2023) that an algorithm achieves regret $\widetilde{\mathcal{O}}(K^{\frac{m+d}{2m+d}})$ in this setting, approaching $\widetilde{\mathcal{O}}(\sqrt{K})$ as $m \to \infty$. In this paper, we aim to make one first step towards reaching an analogous result in the general case of any MDP endowed with some "smoothness" property.

**Why Smoothness?** The presence of mathematically elegant, smooth functions in real-world phenomena of the most diverse nature has always been a source of fascination and philosophical research (Wigner, 1990). Smooth functions, or even infinitely differentiable functions, play a crucial role in various scientific and engineering disciplines due to their versatility and analytical tractability. They are valuable tools for modeling complex phenomena and solving mathematical problems. In physics, for instance, smooth functions are widely employed to describe the behavior of physical systems, such as in the context of quantum mechanics and electromagnetic field theory (Shankar, 2012; Born & Wolf, 2013). In engineering, the utility of smooth functions is evident in control systems and signal processing, where they simplify the analysis and design of dynamic systems (Oppenheim et al., 1997; Ogata, 2010). Smooth functions are not just a formalism but a fundamental and practical mathematical framework for understanding and manipulating real-world phenomena. The reason why these functions are ubiquitous in the natural sciences can be attributed to their connection with partial differential equations. Many natural phenomena can be described by a limited number of partial differential equations, which have the characteristic of enforcing strong regularity conditions on solutions. In particular, thermal, electromagnetic, and wave phenomena are governed by three well-known different equations: the heat equation, the Laplace-Poisson equation, and the D'Alembert equation, respectively (Sobolev, 1964; Tikhonov & Samarskii, 2013; Salsa & Verzini, 2022). Each of these is characterized by inherent regularity properties so that solutions are infinitely differentiable under suitable boundary conditions.

**Our Contributions** In this paper, we introduce two very general classes of MDP based on the notion of $\nu$-smoothness either applied to the transition model and to the reward function (*Strongly Smooth MDP*) or to the Bellman operator (*Weakly Smooth MDP*) (Section 3). We develop a novel technique that builds upon results from the theory of *orthogonal functions* (specifically, Legendre polynomials) to design algorithms, LEGENDRE-LSVI and LEGENDRE-ELEANOR, characterized by different computational costs, for addressing regret minimization in smooth MDPs (Section 4). Then, we provide the theoretical analysis of the proposed algorithms showing that, under appropriate conditions on smoothness constant $\nu$, they fulfill the no-regret property with a regret rate depending on $\nu$ (Section 5). Finally, to compare our results with the state-of-the-art theoretical RL, we show that $(i)$ our setting includes the most common classes of problems for which no-regret guarantees have been shown $(ii)$ general-purpose RL algorithms that apply to our setting obtain worse regret guarantees than ours (Section 6). The proofs of all the results presented in the main paper are reported in Appendix B.

## 2. Preliminaries

**Markov decision processes and policies** We consider a finite-horizon Markov decision process (MDP) (Puterman, 2014) $M = (\mathcal{S}, \mathcal{A}, p, r, H)$, where $\mathcal{S} = [-1, 1]^{d_S}$

is the state space, $\mathcal{A} = [-1, 1]^{d_A}$ is the action space,[2] $p = \{p_h\}_{h=1}^{H-1}$ is the sequence of transition functions, each mapping a pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ to a probability distribution $p_h(\cdot|s, a)$ over $\mathcal{S}$, while the initial state $s_1$ may be chosen arbitrarily from the environment; $r = \{r_h\}_{h=1}^{H}$ is the sequence of reward functions, each mapping a pair $(s, a)$ to a real number $r_h(s, a)$, and $H$ is the time horizon. At each episode $k \in [K] := \{1, \dots, K\}$, the agent chooses a policy $\pi_k = \{\pi_{k,h}\}_{h=1}^{H}$, which is a sequence of mappings from $\mathcal{S}$ to the probability distributions over $\mathcal{A}$. For each stage $h \in [H]$, the action is chosen according to $a_h \sim \pi_{k,h}(\cdot|s_h)$, the agent gains reward $r_h(s_h, a_h) + \eta_h$, where $\eta_h$ is a $\sigma-$subgaussian noise independent of the past, and the environment transitions to the next state $s_{h+1} \sim p_h(\cdot|s_h, a_h)$. In this setting, it is useful to define the following quantities.

**Value functions and Bellman operators.** The state-action value function (or *Q-function*) quantifies the expected sum of the rewards obtained under a policy $\pi$, starting from a state-stage pair $(s, h) \in \mathcal{S} \times [H]$ and fixing the first action to some $a \in \mathcal{A}$:

$$Q_h^\pi(s, a) := \mathbb{E}_\pi \left[ \sum_{\ell=h}^{H} r_\ell(s_\ell, a_\ell) \middle| s_0 = s, a_0 = a \right], \qquad (1)$$

where $\mathbb{E}_\pi$ denotes expectation w.r.t. to the stochastic process $a_h \sim \pi_h(\cdot|s_h)$ and $s_{h+1} \sim p_h(\cdot|s_h, a_h)$ for all $h \in [H]$. The state value function (or *V-function*) is defined as $V_h^\pi(s) := \mathbb{E}_{a\sim\pi_h(\cdot|s)}[Q_h^\pi(s, a)]$, for all $s \in \mathcal{S}$. The supremum of the value functions between all the policies take the name of optimal value functions, and are written as $Q_h^*(s, a) := \sup_\pi Q_h^\pi(s, a)$, $V_h^*(s) := \sup_\pi V_h^\pi(s)$.

In this work, as often done in the literature, we assume that the reward is normalized in a way that $|Q_h^\pi(s, a)| \leqslant 1$ for every $s \in \mathcal{S}, a \in A$ and $h \in [H]$.[3] The evaluation of the expected return is linked to the notion of Bellman operators. For a policy $\pi$, the corresponding *Bellman operator* $\mathcal{T}^\pi$ is defined as follows, for every $h \in [H]$ and every function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$:

$$\mathcal{T}_h^\pi f(s, a) := r_h(s, a) + \mathop{\mathbb{E}}_{\substack{s'\sim p_h(\cdot|s,a) \\ a'\sim\pi_h(\cdot|s)}} [f(s', a')].$$

Even more crucial for control is the *Bellman optimality operator*, which, instead of fixing the policy, chooses the maximum of $f$ for the next state:

$$\mathcal{T}_h^* f(s, a) := r_h(s, a) + \mathop{\mathbb{E}}_{s'\sim p_h(\cdot|s,a)} \left[ \sup_{a'\in\mathcal{A}} f(s', a') \right].$$

**Agent's goal.** The agent aims at choosing a sequence of policies $\pi_k$ in order to minimize the cumulative difference between the expected return of their policies $J^{\pi_k}$ and the optimal one, given the initial state chosen by the environment. This quantity takes the name of *(cumulative) regret*:

$$R_K := \sum_{k=1}^{K} \left( V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \right).$$

Note that if $R_K = o(K)$ for any $K$ with some probability, then, with the same probability, $J^{\pi_k} \to \sup_\pi J^\pi$ as $K \to \infty$. An algorithm choosing a sequence of policies with this property is called *no-regret*.

**Smoothness of real functions.** Let $\Omega \subset [-1, 1]^d$ and $f : \Omega \to \mathbb{R}$. We say that $f \in \mathcal{C}^{\nu,1}(\Omega)$ if there exists a constant $L < +\infty$ such that $f$ is $\nu-$differentiable (i.e., differentiable $\nu$ times), and for every multi-index $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_d)$ with $|\boldsymbol{\alpha}| := \sum_{i=1}^{d} \alpha_i \leqslant \nu$ we have:

$$\forall x, y \in \Omega : \quad |D^{\boldsymbol{\alpha}} f(x) - D^{\boldsymbol{\alpha}} f(y)| \leqslant L\|x - y\|_2, \qquad (2)$$

where the multi-index derivative is defined as follows $D^{\boldsymbol{\alpha}} f := \frac{\partial^{\alpha_1+\dots+\alpha_d}}{\partial x_1^{\alpha_1}\dots\partial x_d^{\alpha_d}} f$. The set $\mathcal{C}^{\nu,1}(\Omega)$ forms, for every value of $\nu$, a normed vector space, and more precisely, a Banach space (Kolmogorov & Fomin, 1975). A norm for which this holds is given by $\|f\|_{\mathcal{C}^{\nu,1}} := \max_{|\boldsymbol{\alpha}|\leqslant\nu+1} \|D^{\boldsymbol{\alpha}} f\|_{L_\infty}$. This definition may seem counterintuitive since derivatives up to order $\nu + 1$ appear, but in fact, this is because the derivative of a Lipschitz function is defined almost everywhere, and its $L_\infty-$norm equals the Lipschitz constant itself (Rudin, 1974). The most straightforward case of this definition is given by $\mathcal{C}^{0,1}(\Omega)$, corresponding to the space of Lipschitz continuous functions, where the semi-norm $\|\cdot\|_{\mathcal{C}^{0,1}}$ corresponds exactly to the *Lipschitz constant* of a function. The concept of Wasserstein metric $\mathcal{W}(\cdot, \cdot)$, a notion of distance for probability measures on metric spaces, is strictly related to Lipschitz functions. For two measures $\mu, \zeta$ on a metric space $\Omega$, this distance is defined as $\mathcal{W}(\mu, \zeta) := \sup_{f\in\mathcal{C}^{0,1}:\|f\|_{\mathcal{C}^{0,1}}=1} \int_\Omega f(\omega) d(\mu - \zeta)(\omega)$. We will also make use of the space $\mathcal{C}^\infty(\Omega) := \cap_{\nu=1}^{\infty} \mathcal{C}^{\nu,1}(\Omega)$ of indefinitely differentiable functions. Despite assuming a function is $\mathcal{C}^\infty(\Omega)$ seems restrictive, this class includes polynomial, trigonometric, and exponential functions.

## 3. Smoothness in MDPs

In this section, we introduce two sets of assumptions concerning the MDP. We start with the simpler one, which we call *Strongly Smooth* MDP. This assumption is similar to the kernelized MDP setting (Yang et al., 2020b), as it bounds the norm of the transition function and the reward function in a given space, but without specifying an explicit structure. Instead of assuming that they belong to a given RKHS, we limit our assumption to their smoothness.

**Assumption 1.** *(Strongly Smooth MDP). An MDP is a*

[2] Choosing these compacts set is without loss of generality as, provided a suitable rescaling, any compact set could be used.

[3] Sometimes it is instead assumed that the reward lies in $[-1, 1]$ so that the total return is bounded in $[-H, H]$.

Strongly Smooth *of order $\nu$ if:*

$$\forall h \in [H]\ \forall s' \in \mathcal{S}, \quad r_h(\cdot,\cdot), p_h(s'|\cdot,\cdot) \in \mathcal{C}^{\nu,1}(\mathcal{S} \times \mathcal{A}),$$

*with* $\sup_{h,s'} \|p_h(s'|\cdot,\cdot)\|_{\mathcal{C}^{\nu,1}}, \sup_{h,s'} \|r_h(\cdot,\cdot)\|_{\mathcal{C}^{\nu,1}} < +\infty.$

Note that assuming the finiteness of the norms but not the knowledge of its upper bound, is different from what is asked in the analogous assumption for kernelized MDPs.

Assuming this form of regularity of the reward function seems fair. Being very often a human-designed function, we can expect it to be indefinitely differentiable most of the time. For what concerns the transition function, this requirement is more tricky. Indeed, nontrivial transition functions for deterministic MDPs often take the form $p(s'|s,a) = \delta(s' = f(s,a))$, for some function $f : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$. This function does not satisfy Strong Smoothness, even when $f$ is itself very smooth, as the Dirac delta $\delta(\cdot)$ is not a continuous function. For this reason, we introduce a more general assumption, which directly concerns the Bellman optimality operator.

**Assumption 2.** *(Weakly Smooth MDP). An MDPs is* Weakly Smooth *of order $\nu$ if, for every $h \in [H]$ the Bellman optimality operator $\mathcal{T}_h^*$ is bounded on $\mathcal{C}^{\nu,1}(\mathcal{S}\times\mathcal{A}) \to \mathcal{C}^{\nu,1}(\mathcal{S}\times\mathcal{A})$.*

Boundedness over $\mathcal{C}^{\nu,1}(\mathcal{S} \times \mathcal{A}) \to \mathcal{C}^{\nu,1}(\mathcal{S} \times \mathcal{A})$ means that the operator cannot output a function that is not $\mathcal{C}^{\nu,1}$ when receiving a function from the same set. Moreover, there exists a constant $C_{\mathcal{T}^*} < +\infty$ such that $\|\mathcal{T}_h^* f\|_{\mathcal{C}^{\nu,1}} \leqslant C_{\mathcal{T}^*}(\|f\|_{\mathcal{C}^{\nu,1}} + 1)$ for every $h \in [H]$ and every function $f \in \mathcal{C}^{\nu,1}(\mathcal{S} \times \mathcal{A})$. In Appendix B.1, we show that Weak Smoothness is (much) weaker than Strong Smoothness.

## 4. Orthogonal Function Representations

Our approach to the solution of Strongly and Weakly Smooth MDPs is based on the idea of finding a representation of the state-action space $\mathcal{S} \times \mathcal{A}$ such that the problem is reduced to a Linear MDP in a feature space. To achieve this result, we will use a particular class of feature maps based on *Legendre polynomials* (Quarteroni et al., 2010).

**Definition 4.1** (Legendre feature map)**.** Let $\varphi_{L,n}(x)$ be the $n$-th order Legendre polynomial, we define, for every $N \in \mathbb{N}$, the feature map $\boldsymbol{\varphi}_{L,N} : [-1,1] \to \mathbb{R}^N$ as follows:

$$\boldsymbol{\varphi}_{L,N}(x) := N^{-1/2}(\varphi_{L,0}(x), \ldots, \varphi_{L,N}(x)).$$

The importance of these feature maps, not apparent from their definition, lies in their *orthogonality*. In fact, Legendre polynomials are such that $\int_{-1}^1 \varphi_{L,i}(x)\varphi_{L,j}(x)\ \mathrm{d}x = \delta_{ij}$, which is 1 if $i = j$, 0 otherwise.[4] The multidimensional generalization of this map to $[-1,1]^d$, which preserves the orthogonality property, is obtained by a Cartesian product operation. Precisely, we call the generalization of the Legendre map to $[-1,1]^d$ as $\widetilde{N}^{-1/2}\boldsymbol{\varphi}_{L,N}^d(x_1,\ldots,x_d)$, where $\boldsymbol{\varphi}_{L,N}^d(x_1,\ldots,x_d)$ stacks, in its $\widetilde{N}$ components, all possible products of Legendre polynomials in the variables $x_1,\ldots,x_d$ such that the total degree (sum of the degrees of the single polynomials) does not exceed $N$ (and $\widetilde{N}^{-1/2}$ is just a normalization term). A formal definition of the feature map is given in the following:

$$\mathcal{L}_N = \{(g_1,\ldots,g_d) \in \{0,\ldots,N\}^d : \sum_{i=1}^d g_i \leqslant N\} \tag{3}$$

$$\boldsymbol{\varphi}_{L,N}^d(x_1,\ldots,x_d) = \left(\prod_{i=1}^d \varphi_{L,g_i}(x_i)\right)_{(g_1,\ldots,g_d)\in\mathcal{L}_N} \tag{4}$$

This definition draws an analogy with Fourier series (Katznelson, 2004), which are built on another family of orthogonal functions. As for the Fourier series, we can use a linear combination of Legendre polynomials to approximate any smooth function. However, while the convergence of the Fourier series is only guaranteed for periodic functions, Legendre polynomials are not affected by this limitation.

### 4.1. Weakly Smooth MDPs: Legendre-Eleanor

We start from the most general case, i.e., that of Weakly Smooth MDPs. We can show that the pair given by an MDP of this class and our Legendre feature map forms a process with *low inherent Bellman error* (Zanette et al., 2020). Given any feature map $\boldsymbol{\varphi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and a sequence of compact sets $\mathcal{B}_h \subset \mathbb{R}^d$ for $h \in [H]$, calling $Q_\theta(s,a)$ the function $\boldsymbol{\varphi}(s,a)^\top\theta$, the inherent Bellman error w.r.t. $\{\mathcal{B}_h\}_h$ is defined as:

$$\mathcal{I} := \max_{h\in[H]} \sup_{\theta\in\mathcal{B}_{h+1}} \inf_{\theta'\in\mathcal{B}_h} \|\boldsymbol{\varphi}(s,a)^\top\theta' - \mathcal{T}^*Q_\theta(s,a)\|_\infty. \tag{5}$$

This definition illustrates, intuitively, that starting from a $Q$-function in the span of $\boldsymbol{\varphi}$, the Bellman optimality operator produces another one which is $\mathcal{I}$-close to the span of $\boldsymbol{\varphi}$. We can prove that a Weakly Smooth MDP equipped with a Legendre feature map $\boldsymbol{\varphi}_{L,N}^d$ has bounded inherent Bellman error $\mathcal{I}$, where the bound depends on the order of smoothness $\nu$ (Theorem 9 in the appendix). Using the Legendre representation along with Eleanor (Zanette et al., 2020), an algorithm designed for MDPs with low inherent Bellman error, we can achieve the no-regret property under mild assumptions. We call the resulting algorithm Legendre-Eleanor, and we will theoretically analyze it in Section 5.

### 4.2. Strongly Smooth MDPs: Legendre-LSVI

In the previous section, we have presented an approach to solving Weakly Smooth MDPs. Still, the algorithm that

[4] We use Legendre polynomials which are normalized in the space $L^2$, while some authors normalize them differently.

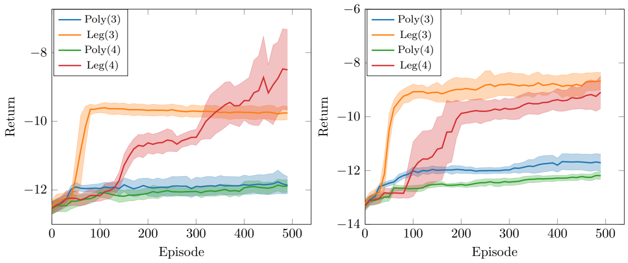

*Figure 1.* Curve of the episodic return for the simulation in Section 4.3 with 95% confidence intervals over five random seeds.

we introduced is based on Eleanor, which is known to be computationally inefficient (Zanette et al., 2020). A major challenge in RL is to devise algorithms that are both no-regret and have a running time polynomial in the task horizon, problem dimension, and number of episodes. This motivates the search for a *polynomial-time* algorithm that can achieve no regret under the Strongly Smooth assumption. This is possible thanks to the fact that, when we apply the Legendre representation of a Strongly Smooth MDP, we get not only an MDP with low inherent Bellmann error, but a Linear MDP, for which computationally feasible algorithms, such as LSVI-UCB (Jin et al., 2020), are known.

We call the resulting algorithm Legendre-LSVI. To analyze its computational complexity, remembering that we have called $\widetilde{N} = \binom{N+d}{N} \leqslant N^d$ the dimension of the feature map used, we can just replace this value in the computational complexity of LSVI-UCB. As it is well-known that the time complexity of LSVI-UCB is $\mathcal{O}(K^2H + \widetilde{N}^3KH)$, Legendre-LSVI has polynomial time complexity, provided that we choose $N$ so that term $\widetilde{N}$ is not exponential in the relevant quantities.

### 4.3. Why Orthogonal Features?

The reader may wonder how crucial is the choice of an *orthogonal* feature representation. Before moving to the theoretical analysis that will analytically justify this choice, in this subsection, we empirically show, on an illustrative problem, that the use of orthogonal features has beneficial effects on learning performance. We employ two modified versions of the LQR, in which the state, after the linear dynamic transition, is pushed towards the origin in a way that prevents it from escaping from a given compact set. Precisely, using the same formalism of the LQR, we have: $s_{h+1} = g(As_h + Ba_h + \xi_h)$, $r_h = -s_h^\top Q s_h - a_h^\top R a_h$, where $g(x) := \frac{x}{1+\|x\|_2}$ and $\xi_h$ is a Gaussian noise. As the support of the Gaussian distribution is the full $\mathbb{R}^d$, after applying $g(\cdot)$, the possible set of new states is the ball of radius one. We performed two experiments with different parameter values and with horizon $H = 20$, whose details can be found in the appendix C. In Figure 1, we can see plots showing the episodic return of the algorithms as a function of the number of learning episodes. As a learning algorithm, we can see two "correct" versions of Legendre-LSVI, which are called Leg(3) and Leg(4), against two "naïve" versions of the same algorithm, Poly(3) and Poly(4). For all the algorithms, the number between the brackets corresponds to the degree of the polynomials used so that the approximation order and the length of the feature vector are equal in the two cases. The difference between the Poly() and the corresponding Leg() algorithm lies in the fact that the former is the standard basis of multivariate polynomials (e.g., $\{1, x, y, x^2, y^2, xy, \dots\}$), while the latter corresponds to the (orthogonal) Legendre basis, for which theoretical guarantees hold. Using standard polynomials as feature maps is common in practice. However, the results show that baselines using Legendre polynomials achieve much superior episodic return compared with the analog with standard polynomials, as it is predicted from the theory behind our results. The latter, in green and blue, failed to achieve significant learning throughout 500 episodes in either environment. On the contrary, Leg(3), in orange, is able to learn a good policy suddenly and subsequently settles down, obtaining an almost constant return in all the following episodes. Leg(4) proves to learn more slowly than Leg(3), but in the first environment, it obtains a higher return value, while in the second environment, it obtains a comparable one. These results are consistent with the theory, as increasing the dimensionality of the feature map considered and increasing the degree of the polynomial from 3 to 4 has the effect of slowing down learning but improving the order of approximation to converge to a higher return.

## 5. Theoretical Guarantees

In this section, we derive the regret bounds for our two algorithms Legendre-Eleanor and Legendre-LSVI. The former is able to achieve the no-regret property for Weakly Smooth MDPs under the assumption that $2\nu \geqslant d-2$, as shown in the following result.

**Theorem 1.** *Let us consider a Weakly Smooth MDP $M$ with state action space $[-1,1]^d$. Under the condition that $\nu > d/2 - 1$, Legendre-Eleanor initialized with $N = \lceil K^{\frac{1}{d+2(\nu+1)}} \rceil$, with probability at least $1-\delta$, suffers a regret of order at most:*

$$R_K \leqslant \widetilde{\mathcal{O}}\left(C_{\mathrm{ELE}}^{H} K^{\frac{3d/2+\nu+1}{d+2(\nu+1)}}\right),$$

*where the constant depends only on d and $\nu$ and the $\widetilde{\mathcal{O}}$ hides logarithmic functions of $K$, $\delta$.*

The proof is provided in Appendix B.5. The fact that the regret grows exponentially in $H$ is annoying but unavoid-

| Algorithm | **W.ly Smooth** | Lipschitz | **S.ly Smooth** | Kernelized | LQR | LinearMDP |
|---|---|---|---|---|---|---|
| LEGENDRE-ELEANOR (4.1) | $K^{\frac{3d/2+\nu+1}{d+2(\nu+1)}}$ | $K^{\frac{3d/2+1}{d+2}}$ | $K^{\frac{3d/2+\nu+1}{d+2(\nu+1)}}$ | $K^{\frac{3d/2+\lceil m\rceil}{d+2\lceil m\rceil}}$ | $K^{\frac{1}{2}}$ | $K^{\frac{1}{2}}$ |
| (Jin et al., 2021) | $K^{\frac{2\nu+3d+2}{4\nu+4}}$ | $K^{\frac{3d+2}{4}}$ | $K^{\frac{2\nu+3d+2}{4\nu+4}}$ | $K^{\frac{2\lceil m\rceil+3d}{\lceil m\rceil}}$ | $K^{\frac{1}{2}}$ | $K^{\frac{1}{2}}$ |
| (Song & Sun, 2019) | ✗ | $K^{\frac{d+1}{d+2}}$ | $K^{\frac{d+1}{d+2}}$ | $K^{\frac{d+1}{d+2}}$ | $K^{\frac{d+1}{d+2}}$ | $K^{\frac{d+1}{d+2}}$ |
| LEGENDRE-LSVI (4.2) | ✗ | ✗ | $K^{\frac{2d+\nu+1}{d+2(\nu+1)}}$ | $K^{\frac{2d+\lceil m\rceil}{d+2\lceil m\rceil}}$ | $K^{\frac{1}{2}}$ | $K^{\frac{1}{2}}$ |
| (Vakili & Olkhovskaya, 2023) | ✗ | ✗ | ✗ | $K^{\frac{d+m+1}{d+2(m+1)}}$ | ✗ | $K^{\frac{1}{2}}$ |
| (Dean et al., 2018) | ✗ | ✗ | ✗ | ✗ | $K^{\frac{1}{2}}$ | ✗ |
| (Jin et al., 2020) | ✗ | ✗ | ✗ | ✗ | ✗ | $K^{\frac{1}{2}}$ |

*Table 1.* Table containing the regret guarantee of each algorithm presented in the main paper for each setting. For convenience, we recall that Song & Sun (2019); Vakili & Olkhovskaya (2023); Dean et al. (2018); Jin et al. (2020) represent the state of the art for Lipschitz MDPs, Kernelized MDPs, LQRs and LinearMDPs respectively. On the columns we have the algorithm, and on the rows the setting in which it is tested. Some specifications are needed: 1) we have only reported the order of the regret in $K$, ignoring the other parameters of the problems and the logarithmic terms 2) for the linear MDP setting we have assumed that the feature map is indefinitely differentiable, an assumption explained in detail in the corresponding section of the main paper 3) for the linear MDP, the algorithms SOTA-Linear and SOTA-Kern assume to know the feature map, while our algorithm do not have this requirement 4) Kernelized MDP assume Matérn kernel of order $m$. As a result of the table, we can see that our algorithm LEGENDRE-ELEANOR is the best performing between the ones having no-regret guarantees for all the settings: the only algorithms that are able to surpass its performance are designed for settings that are much more specific.

able. Indeed, we derive a lower bound that shows that any MDP class that is rich enough to capture Lipschitz MDPs must have a regret bound which is exponential in $H$ (see Appendix B.6). No surprise, all related works on Lipschtz MDPs are affected by the same problem. Apart from that, Theorem 1 shows the bound we aimed for. In the "good" regime, where $\nu > d/2 - 1$, we are able to prove a regret bound that is monotonically decreasing in $\nu$, and approaches $\sqrt{K}$ for $\nu \to \infty$. Therefore, our model can cover both general (Lipschitz MDPs and Kernelized MDPs) and specific (LQRs, Linear MDPs) models, with a regret bound that is adaptive to the higher smoothness.

We now turn to LEGENDRE-LSVI. Its guarantees are restricted to Strongly Smooth MDPs, and it only achieves no regret under the more demanding requirement $\nu \geqslant d - 1$. Still, its value lies both in its polynomial computational complexity and in its polynomial dependence on the horizon $H$.

**Theorem 2.** *Let us consider a Strongly Smooth MDP $M$ with state action space $[-1, 1]^d$. Under the condition that $d \leqslant \nu + 1$, LEGENDRE-LSVI initialized with $N = \lceil K^{\frac{1}{d+2(\nu+1)}} \rceil$, with probability at least $1 - \delta$, suffers a regret of order at most:*

$$R_K \leqslant \widetilde{\mathcal{O}}\left(H^{3/2} K^{\frac{2d+\nu+1}{d+2(\nu+1)}}\right),$$

*where $\widetilde{\mathcal{O}}$ hides logarithmic functions of $K$, $\delta$, and $H$.*

The order of the regret in $K$ is worse than the one of LEGENDRE-ELEANOR but we still have $\sqrt{K}$ in the limit $\nu \to +\infty$ and, as anticipated, the exponential growth in $H$ is avoided. The proof is provided in Appendix B.7.

Note that the choice $N = \lceil K^{\frac{1}{d+2(\nu+1)}} \rceil$ makes the running time polynomial. Indeed, we have seen in Section 4.2 that the latter scales as $\mathcal{O}(K^2 H + \widetilde{N}^3 K H)$, so that this choice of $N$ allows bound the time complexity as $\mathcal{O}(K^2 H + K^{1+\frac{3d}{d+2(\nu+1)}} H) = \mathcal{O}(K^2 H)$.

## 6. Comparison with Related Literature

Achieving regret or sample complexity guarantees for MDPs with continuous state and action spaces has been one of the main challenges of theoretical RL in recent years. Among the numerous papers dealing with this problem, we provide a brief overview of the most significant results achieved under the most common assumptions proposed in the literature. This way, we show how very different problems studied so far are included in our setting. The overall inclusion relationships between the settings are summarized in Figure 2, while the relations between the regret bounds in are shown in table 1.

### 6.1. Lipschitz MDPs

Lipschitz MDPs assume that the transition function of the model, as well as the reward function, are Lipschitz continuous with constants $L_p$ and $L_r$, respectively. For the reward function, this is just expressed by enforcing for all $h \in [H]$ $s, s' \in \mathcal{S}$, $a, a' \in \mathcal{A}$:

$$|r(s, a) - r(s', a')| \leqslant L_r(\|s - s'\|_2 + \|a - a'\|_2).$$

While, as the transition function maps a state-action pair into a probability distribution, we need a metric for probability distribution to define Lipschitzness. This is commonly done by means of the Wasserstein metric $\mathcal{W}(\cdot,\cdot)$ (Rachelson & Lagoudakis, 2010):

$$\mathcal{W}(p_h(\cdot|s,a), p_h(\cdot|s',a')) \leqslant L_p(\|s-s'\|_2 + \|a-a'\|_2).$$

Learning in Lipschitz MDPs is a very hot topic (Ortner & Ryabko, 2012; Sinclair et al., 2019; Song & Sun, 2019; Sinclair et al., 2020; Domingues et al., 2020; Le Lan et al., 2021) and many regret bounds of order $K^{\frac{d+1}{d+2}}$ have been proved.

**Lipschitz MDPs are Weakly Smooth but not Strongly Smooth.** Lipschitz MDPs represent the most general class of continuous space MDPs studied in the literature. These processes are not necessarily Strongly Smooth, as they include deterministic processes that cannot be Strongly Smooth due to the discrete nature of the transition function $p$. Still, it can be proved that Lipschitz MDPs are Weakly Smooth for $\nu = 0$, as shown in the Appendix B.2.

### 6.2. Linear Quadratic Regulator

Linear Quadratic Regulator (LQR) is a model of the environment s where the state and action space are $\mathcal{S} = \mathbb{R}^{d_S}, \mathcal{A} = \mathbb{R}^{d_A}$, and there exist two matrices $A \in \mathbb{R}^{d_S \times d_S}$ and $B \in \mathbb{R}^{d_S \times d_A}$ defining the transition model $s_{h+1} = As_h + Ba_h + \xi_h$, where $\xi_h \sim \mathcal{N}(0, \Sigma)$ is a Gaussian noise. Also, there are other two positive semi-definite matrices $Q \in \mathbb{R}^{d_S \times d_S}$ and $R \in \mathbb{R}^{d_S \times d_S}$ such that the reward function is given by $r_h = -s_h^\top Q s_h - a_h^\top R a_h$. Regret bounds of order $\sqrt{K}$ were obtained for this kind of MDPs Abbasi-Yadkori & Szepesvári (2011); Dean et al. (2018); Cohen et al. (2019). This result has been generalized by Kakade et al. (2020), preserving the optimal $\sqrt{K}$ regret order when the linear dynamics are composed with a known feature map. Note that, differently from the previous case, the dimension $d$ of the state-action space does not affect the regret order, as its dependence in the regret takes the form $\text{poly}(d_{\mathcal{S}}, d_{\mathcal{A}})\sqrt{K}$.

**LQRs are Strongly Smooth for $\nu = +\infty$.** The class of LQRs contains only processes which are indefinitely differentiable. Indeed, the reward function is quadratic, while the transition one can be written as $p_h(s'|s,a) = \mathcal{N}(s'; As_h + Ba_h, \Sigma)$. The last function is $\propto \exp\left(-(s' - As_h - Ba_h)^\top \Sigma^{-1}(s' - As_h - Ba_h)\right)$, which is indefinitely differentiable is all its variables. For $\nu = +\infty$ our Theorem 6.3 ensures that regret of order $\sqrt{K}$ can be achieved, which is coherent with the results from the literature.

### 6.3. Linear MDPs

Linear MDPs are processes satisfying, for every $h \in [H]$ $s, s' \in \mathcal{S}$, $a \in \mathcal{A}$: $p_h(s'|s,a) = \langle \boldsymbol{\mu}_h(s'), \boldsymbol{\varphi}(s,a)\rangle$, where $\boldsymbol{\varphi}(s,a)$ (resp. $\boldsymbol{\mu}(s')$) are fixed functions from $\mathcal{S} \times \mathcal{A}$ (resp. $\mathcal{S}$) to $\mathbb{R}^{d_{\boldsymbol{\varphi}}}$. Similarly, the reward function factorizes as $r_h(s,a) = \langle \boldsymbol{\theta}_h, \boldsymbol{\varphi}(s,a)\rangle$, for some vector $\boldsymbol{\theta}_h \in \mathbb{R}^{d_{\boldsymbol{\varphi}}}$. For linear MDPs, if the feature map is given, the optimal regret order of $\sqrt{K}$ can be achieved. For example, in (Jin et al., 2020) the regret takes the form of $d_{\boldsymbol{\varphi}}^{3/2} K^{1/2}$, so that the order of the regret is not affected by the magnitude of $d$ or $d_{\boldsymbol{\varphi}}$. When the feature map is not known in advance, the problem becomes significantly harder (Agarwal et al., 2020; Uehara et al., 2021). In fact, there is currently no work able to prove regret guarantees for this setting, although sample complexity guarantees are available.

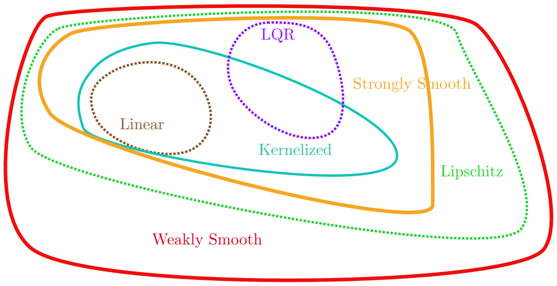

*Figure 2.* A schematic summarizing relations among families of continuous space RL problems. Our assumptions correspond to the red and orange sets.

**Linear MDPs are Strongly Smooth.** Linear MDPs are Strongly Smooth with an order $\nu$ depending on the smoothness of the feature map $\boldsymbol{\varphi}$ (see Appendix B.3). If this function is handcrafted, such as a polynomial, exponential, or trigonometric function, we have $\nu = +\infty$. The most favorable case also happens when $\boldsymbol{\varphi}$ is a fully connected neural network with either sigmoid (Narayan, 1997), tanh (Abdelouahab et al., 2017), or softplus (Zheng et al., 2015) activation function, as these activations are all infinitely differentiable. Instead, using ReLU (Schmidt-Hieber, 2020) activation, which is only Lipschitz continuous, leads to $\nu = 0$. Applying our theorem ensures that when using a feature map that is indefinitely smooth, LEGENDRE-LSVI has a regret order of $\sqrt{K}$, ***even if the feature map is not explicitly known***. This is a very strong result that has independent interest for the literature of linear MDPs.

### 6.4. Kernelized MDPs

Kernelized MDPs are built on a completely different assumption with respect to the previous methods. In fact, they start from a given kernel $k(\cdot,\cdot)$ and assume that both the transition function and the reward function belong to the RKHS $\mathcal{H}_k$ corresponding to the kernel $k$:

$$\forall h \in [H],\ \forall s' \in \mathcal{S}: \qquad r_h(\cdot,\cdot), p_h(s'|\cdot,\cdot) \in \mathcal{H}_k.$$

Moreover, as $\mathcal{H}_k$ is endowed with a norm $\|\cdot\|_{\mathcal{H}_k}$, an upper bound on the norm of the reward and of the transition function is assumed to be known. The most common family of kernels is given by the Matérn kernels, which depend on a parameter $m > 0$. Interestingly, the corresponding $\mathcal{H}_k$ gets smaller the higher value of $m$, and the function there contained becomes progressively smoother. For this family of MDPs, a regret bound of order $\widetilde{\mathcal{O}}(K^{\frac{m+d}{2m+d}})$ is known, which has been shown to be optimal for both the case $m \to 0$ and for $m \to +\infty$ (Vakili & Olkhovskaya, 2023). Other papers that dealt with this setting are (Chowdhury & Gopalan, 2019; Yang et al., 2020a; Domingues et al., 2021).

**Kernelized MDPs are Strongly Smooth.** In fact, Kernelized MDPs are just a particular case of Strongly Smooth MDPs, as directly follows from the fact that functions in the most studied RKHS are smooth. A general result (Theorem 10.45 from Wendland (2004)) shows that whenever we have a $\nu-$times differentiable kernel $k$, the corresponding RKHS contains only functions that are $\nu/2-$times differentiable, and thus contained in $\mathcal{C}^{\nu/2-1,1}(\Omega)$. The result can be specialized to the Matérn family of Kernels. We show in Appendix B.4 that the Matérn kernel of parameter $m$ generates an RKHS which is contained in $\mathcal{C}^{\nu-1,1}(\Omega)$ for every $\nu < m$, provided that the domain $\Omega$ satisfies some reasonable assumption there specified.

### 6.5. Other Structural Assumptions: Comparison with General-Purpose Algorithms

In recent years, RL theory has seen a rush in searching for the weakest assumptions under which sample-efficient RL is possible. These assumptions typically generalize linear MDPs. In particular, we cite MDPs with low inherent Bellman error (Zanette et al., 2020), MDPs with Gaussian noise (Ren et al., 2022), and MDPs with low Bellman-Eluder dimension (Jin et al., 2021). Another strong structural assumption is introduced in (Du et al., 2021), but no regret bounds are known for problems of that family.

Among the most general algorithms with regret guarantees for general RL, there is Algorithm 1 from (Ren et al., 2022). The latter is built on the idea that if an MDP has a transition function given by deterministic dynamics plus a Gaussian noise (like in LQRs, but without the linear structure), we can exploit the properties of the Gaussian function to our advantage. Applying this algorithm in the setting of Strongly Smooth MDPs, with the additional assumption that the noise is Gaussian, we can prove what follows.

**Theorem 3.** *The regret of Algorithm 1 from (Ren et al., 2022) in a Strongly Smooth MDP, supposing that the transition function is given by a deterministic function plus a Gaussian noise, is bounded, with probability at least $1-\delta$, by:*

$$R_K \leqslant \widetilde{\mathcal{O}}\left(H^{\frac{3\nu+d+3}{2\nu+2}} K^{\frac{\nu+d+1}{2\nu+2}}\right),$$

*assuming that $d < \nu + 1$.*

For the proof, see the Appendix B.8. The regret bound is similar to the one of our LEGENDRE-LSVI from Theorem 6.3. Both ensure no regret if $d < \nu + 1$ and are polynomial in $H$ with very similar exponents, both in $K$ and in $H$. Still, the latter result has two major drawbacks: the noise must be Gaussian, and the algorithm is *not* computationally efficient, as opposed to our LEGENDRE-LSVI, which runs in $\mathcal{O}(K^2 H)$ time.

Another comparison for our algorithms is GOLF from (Jin et al., 2021). The latter is guaranteed to work in a setting that is extremely general, as it only assumes the MDP to have a low Bellman-Eluder dimension. Therefore, the following result holds.

**Theorem 4.** *The regret of* GOLF *on a Weakly Smooth MDP, provided that $d < \frac{2}{3}\nu + \frac{2}{3}$, is bounded, with probability at least $1-\delta$, by:*

$$R_K \leqslant \widetilde{\mathcal{O}}\left(C_{\text{GOLF}}^H K^{\frac{2\nu+3d+2}{4\nu+4}}\right).$$

The proof is reported in Appendix B.8. This time, no assumption is made more than the fact we are in a Weakly Smooth MDP. Therefore, the regret bound can be compared to the one of our LEGENDRE-ELEANOR from Theorem 1. In fact, this result only guarantees no regret under the strong assumption that $d < \frac{2}{3}\nu + \frac{2}{3}$, as opposed to Theorem 1, which only requires $d < 2\nu + 2$. The order in $K$ is also much better for our algorithm for every possible smoothness parameter $\nu$, while both suffer exponential dependence on the time horizon, as it was already clear from the lower bound (see Appendix B.6). Lastly, both GOLF and LEGENDRE-LSVI are computationally inefficient.

## 7. Conclusions

In this study, we defined two broad classes of MDPs distinguished by varying degrees of smoothness, which generalize most of the settings for which no-regret RL algorithms exist in the literature. Furthermore, we introduced two novel algorithms based on the theory of orthogonal functions, which are able to deal with our general setting, achieving a better regret guarantee than any previous algorithm having the same level of generality.

**Future works.** Despite the generality of our results, it is not clear if the proposed algorithms are optimal for our general setting, or if a better regret bound is possible. Therefore, the main objective of future work is to close this gap, by either finding a lower bound for the regret of any algorithm or by proving an improved upper bound.

## Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

## Acknowledgements

Funded by the European Union – Next Generation EU within the project NRPP M4C2, Investment 1.,3 DD. 341 - 15 march 2022 – FAIR – Future Artificial Intelligence Research – Spoke 4 - PE00000013 - D53C22002380006.

## References

Abbasi-Yadkori, Y. and Szepesvári, C. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 1–26. JMLR Workshop and Conference Proceedings, 2011.

Abdelouahab, K., Pelcat, M., and Berry, F. Why tanh is a hardware friendly activation function for cnns. In *Proceedings of the 11th international conference on distributed smart cameras*, pp. 199–201, 2017.

Agarwal, A., Kakade, S., Krishnamurthy, A., and Sun, W. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 33:20095–20107, 2020.

Asadi, K., Misra, D., and Littman, M. Lipschitz continuity in model-based reinforcement learning. In *International Conference on Machine Learning*, pp. 264–273. PMLR, 2018.

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.

Bagby, T., Bos, L., and Levenberg, N. Multivariate simultaneous approximation. *Constructive approximation*, 18 (4):569–577, 2002.

Bemporad, A., Morari, M., Dua, V., and Pistikopoulos, E. N. The explicit linear quadratic regulator for constrained systems. *Automatica*, 38(1):3–20, 2002.

Born, M. and Wolf, E. *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light*. Elsevier, 2013.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Chowdhury, S. R. and Gopalan, A. Online learning in kernelized markov decision processes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3197–3205. PMLR, 2019.

Cohen, A., Koren, T., and Mansour, Y. Learning linear-quadratic regulators efficiently with only $\sqrt{t}$ regret. In *International Conference on Machine Learning*, pp. 1300–1309. PMLR, 2019.

Damiani, A., Manganini, G., Metelli, A. M., and Restelli, M. Balancing sample efficiency and suboptimality in inverse reinforcement learning. In *International Conference on Machine Learning*, pp. 4618–4629. PMLR, 2022.

Dean, S., Mania, H., Matni, N., Recht, B., and Tu, S. Regret bounds for robust adaptive control of the linear quadratic regulator. *Advances in Neural Information Processing Systems*, 31, 2018.

Dlotko, T. Sobolev spaces and embedding theorems. *Silesian University, Poland*, 2014.

Domingues, O. D., Ménard, P., Pirotta, M., Kaufmann, E., and Valko, M. Regret bounds for kernel-based reinforcement learning. In *International Conference on Machine Learning*, 2020.

Domingues, O. D., Ménard, P., Pirotta, M., Kaufmann, E., and Valko, M. Kernel-based reinforcement learning: A finite-time analysis. In *International Conference on Machine Learning*, pp. 2783–2792. PMLR, 2021.

Du, S., Kakade, S., Lee, J., Lovett, S., Mahajan, G., Sun, W., and Wang, R. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pp. 2826–2836. PMLR, 2021.

Folland, G. B. *Real analysis: modern techniques and their applications*, volume 40. John Wiley & Sons, 1999.

Grant, J. and Leslie, D. On thompson sampling for smoother-than-lipschitz bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 2612–2622. PMLR, 2020.

Hambly, B., Xu, R., and Yang, H. Recent advances in reinforcement learning in finance. *Mathematical Finance*, 33(3):437–503, 2023.

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, 2020.

Jin, C., Liu, Q., and Miryoosefi, S. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021.

Kakade, S., Krishnamurthy, A., Lowrey, K., Ohnishi, M., and Sun, W. Information theoretic regret bounds for online nonlinear control. *Advances in Neural Information Processing Systems*, 33:15312–15325, 2020.

Katznelson, Y. *An introduction to harmonic analysis*. Cambridge University Press, 2004.

Kiran, B. R., Sobh, I., Talpaert, V., Mannion, P., Al Sallab, A. A., Yogamani, S., and Pérez, P. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23 (6):4909–4926, 2021.

Kober, J., Bagnell, J. A., and Peters, J. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

Kolmogorov, A. N. and Fomin, S. V. *Introductory real analysis*. Courier Corporation, 1975.

Le Lan, C., Bellemare, M. G., and Castro, P. S. Metrics and continuity in reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8261–8269, 2021.

Liotet, P., Maran, D., Bisi, L., and Restelli, M. Delayed reinforcement learning by imitation. In *International Conference on Machine Learning*, pp. 13528–13556. PMLR, 2022.

Liu, Y., Wang, Y., and Singh, A. Smooth bandit optimization: generalization to holder space. In *International Conference on Artificial Intelligence and Statistics*, pp. 2206–2214. PMLR, 2021.

Maran, D., Metelli, A. M., and Restelli, M. Tight performance guarantees of imitator policies with continuous actions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9073–9080, 2023.

Metelli, A. M. *Exploiting environment configurability in reinforcement learning*, volume 361. IOS Press, 2022.

Metelli, A. M., Mazzolini, F., Bisi, L., Sabbioni, L., and Restelli, M. Control frequency adaptation via action persistence in batch reinforcement learning. In *International Conference on Machine Learning*, pp. 6862–6873. PMLR, 2020.

Narayan, S. The generalized sigmoid activation function: Competitive supervised learning. *Information sciences*, 99(1-2):69–82, 1997.

Ogata, K. *Modern control engineering fifth edition*. 2010.

Oppenheim, A. V., Willsky, A. S., Nawab, S. H., and Ding, J.-J. *Signals and systems*, volume 2. Prentice hall Upper Saddle River, NJ, 1997.

Ortner, R. and Ryabko, D. Online regret bounds for undiscounted continuous reinforcement learning. *Advances in Neural Information Processing Systems*, 25, 2012.

Pirotta, M., Restelli, M., and Bascetta, L. Policy gradient in lipschitz markov decision processes. *Machine Learning*, 100:255–283, 2015.

Pleśniak, W. Multivariate jackson inequality. *Journal of computational and applied mathematics*, 233(3):815–820, 2009.

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

Quarteroni, A., Sacco, R., and Saleri, F. *Numerical mathematics*, volume 37. Springer Science & Business Media, 2010.

Rachelson, E. and Lagoudakis, M. G. On the locality of action domination in sequential decision making. 2010.

Ren, T., Zhang, T., Szepesvári, C., and Dai, B. A free lunch from the noise: Provable and practical exploration for representation learning. In *Uncertainty in Artificial Intelligence*, pp. 1686–1696. PMLR, 2022.

Rudin, W. Real and complex analysis, mcgraw-hill. *Inc.*,, 1974.

Russo, D. and Van Roy, B. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

Salsa, S. and Verzini, G. *Partial differential equations in action: from modelling to theory*, volume 147. Springer Nature, 2022.

Schmidt-Hieber, J. Nonparametric regression using deep neural networks with relu activation function. 2020.

Schultz, M. H. $l^\infty$-multivariate approximation theory. *SIAM Journal on Numerical Analysis*, 6(2):161–183, 1969.

Shankar, R. *Principles of quantum mechanics*. Springer Science & Business Media, 2012.

Sinclair, S., Wang, T., Jain, G., Banerjee, S., and Yu, C. Adaptive discretization for model-based reinforcement learning. *Advances in Neural Information Processing Systems*, 33:3858–3871, 2020.

Sinclair, S. R., Banerjee, S., and Yu, C. L. Adaptive discretization for episodic reinforcement learning in metric spaces. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(3):1–44, 2019.

Sobolev, S. *Partial differential equations of mathematical physics*, volume 56. Courier Corporation, 1964.

Song, Z. and Sun, W. Efficient model-free reinforcement learning in metric spaces. *arXiv preprint arXiv:1905.00475*, 2019.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Tikhonov, A. N. and Samarskii, A. A. *Equations of mathematical physics*. Courier Corporation, 2013.

Todorov, E., Erez, T., and Tassa, Y. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 5026–5033. IEEE, 2012.

Tsybakov, A. B. *An introduction to harmonic analysis*. Springer Science and Business Media, 2008.

Tuo, R. and Jeff Wu, C. A theoretical framework for calibration in computer models: Parametrization, estimation and convergence properties. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):767–795, 2016.

Uehara, M., Zhang, X., and Sun, W. Representation learning for online and offline rl in low-rank mdps. *arXiv preprint arXiv:2110.04652*, 2021.

Vakili, S. and Olkhovskaya, J. Kernelized reinforcement learning with order optimal regret bounds. *arXiv preprint arXiv:2306.07745*, 2023.

Wendland, H. *Scattered data approximation*, volume 17. Cambridge university press, 2004.

Wigner, E. P. The unreasonable effectiveness of mathematics in the natural sciences. In *Mathematics and science*, pp. 291–306. World Scientific, 1990.

Yang, L. and Wang, M. Sample-optimal parametric q-learning using linearly additive features. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6995–7004. PMLR, 2019.

Yang, Z., Jin, C., Wang, Z., Wang, M., and Jordan, M. I. On function approximation in reinforcement learning: Optimism in the face of large state spaces. *Advances in Neural Information Processing Systems*, 2020, 2020a.

Yang, Z., Jin, C., Wang, Z., Wang, M., and Jordan, M. I. Provably efficient reinforcement learning with kernel and neural function approximations. In *NeurIPS*, 2020b.

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, 2020.

Zheng, H., Yang, Z., Liu, W., Liang, J., and Li, Y. Improving deep neural networks using softplus units. In *2015 International joint conference on neural networks (IJCNN)*, pp. 1–4. IEEE, 2015.

## A. Notation

In this section, we leave, for the reader's convenience, two tables of the notations introduced in this paper. We start from one with standard RL notation:

| | |
|---|---|
| $\mathcal{S}$ | State space of an MDP |
| $\mathcal{A}$ | Action space of an MDP |
| $H$ | Time horizon of an MDP |
| $p_h$ | Transition function of an MDP at step $h$ |
| $r_h$ | Transition function of an MDP at step $h$ |
| $H$ | Time horizon of an MDP |
| $K$ | Number of interaction episodes between an MDP and a learning algorithm |
| $R_K$ | Cumulative regret after $K$ episodes |
| $\pi_h^k$ | Policy selected by the algorithm after $k$ episodes for step $h$ |
| $Q_h^\pi(s,a)$ | State-action value function for policy $\pi$ at step $h$ |
| $Q_h^*(s,a)$ | Optimal state-action value function at step $h$ |
| $V_h^\pi(s,a)$ | State value function for policy $\pi$ at step $h$ |
| $\mathcal{T}_h^\pi$ | Bellman operator for policy $\pi$ at step $h$ |
| $\mathcal{T}_h^*$ | Bellman optimality operator at step $h$ |

Then, we have one related to the notation coming from mathematical analysis.

| | |
|---|---|
| $d_{\mathcal{S}}$ | Vector space dimension of $\mathcal{S}$ |
| $d_{\mathcal{A}}$ | Vector space dimension of $\mathcal{A}$ |
| $d$ | Vector space dimension of $\mathcal{S} \times \mathcal{A}$ |
| $D^{\boldsymbol{\alpha}} f$ | Derivative of function $f$ with respect to the multi index $\boldsymbol{\alpha}$ |
| $\nu$ | order to smoothness of a function |
| $\mathcal{C}^\nu(\Omega)$ | Space of $\nu-$times differentiable functions |
| $\mathcal{C}^{\nu,1}(\Omega)$ | Banach space of $\nu-$times differentiable functions with last derivative which is Lipschitz continuous |
| $\lVert \cdot \rVert_{\mathcal{C}^{\nu,1}}$ | Norm in the Banach space $\mathcal{C}^{\nu,1}(\Omega)$, so that $\lVert f \rVert_{\mathcal{C}^{\nu,1}} := \max_{\lvert\boldsymbol{\alpha}\rvert \leqslant \nu+1} \lVert D^{\boldsymbol{\alpha}} f \rVert_\infty$ |
| $\mathcal{W}(\cdot,\cdot)$ | Wasserstein distance between measures |
| $\mathcal{N}(\cdot; x, \Sigma)$ | Multivariate normal distribution of mean $x$ and covariance matrix $\Sigma$ |
| $I$ | Identity matrix |

Lastly, we have the notation which is specific for our paper and the related works.

| | |
|---|---|
| $L_p$ | Lipschtz constant of the transition function in a Lipschitz MDP |
| $L_r$ | Lipschtz constant of the reward function in a Lipschitz MDP |
| $\boldsymbol{\varphi}_N$ | generic feature map of degree $N$ |
| $\varphi_n$ | $n-$th element of a generic feature map $\boldsymbol{\varphi}_N$ |
| $\boldsymbol{\varphi}_{L,N}$ | Legendre feature map of degree $N$ (1 dimension) |
| $\boldsymbol{\varphi}_{L,N}^d$ | Legendre feature map of degree $N$ ($d$ dimensions) |
| $N$ | Degree of a feature map |
| $\widetilde{N}$ | Length of a feature map ($= N$ if $d = 1$) |
| $\mathcal{I}$ | inherent Bellman error |
| $\theta$ | Linear parameter in a Linear MDP/ Bellman complete MDP |
| $\dim_E(\mathcal{F}, \varepsilon)$ | Eluder dimension of the function class $\mathcal{F}$ with respect to the threshold $\varepsilon > 0$ |
| $\mathcal{N}_\infty(\mathcal{F}, \varepsilon)$ | Covering number of the function class $\mathcal{F}$ with respect to the threshold $\varepsilon > 0$ in $L^\infty$ norm |

## B. Omitted Proofs

### B.1. Strongly Smooth $\implies$ Weakly Smooth

Let us assume an MDP is Strongly Smooth. Indeed, for every function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ we have:

$$\begin{aligned}\mathcal{T}_h^* f(s,a) &= r_h(s,a) + \mathop{\mathbb{E}}_{s' \sim p_h(\cdot|s,a)}[\max_{a' \in \mathcal{A}} f(s',a')] \\ &= r_h(s,a) + \int_{\mathcal{S}} \max_{a' \in \mathcal{A}} f(s',a') p_h(s'|s,a)\, ds'.\end{aligned}$$

By triangular inequality, this entails:

$$\|\mathcal{T}_h^* f\|_{\mathcal{C}^{\nu,1}} \leqslant \|r\|_{\mathcal{C}^{\nu,1}} + \left\| \int_{\mathcal{S}} \max_{a' \in \mathcal{A}} f(s',a') p_h(s'|s,a)\, ds' \right\|_{\mathcal{C}^{\nu,1}},$$

where the first term is bounded by assumption, so that we can focus on the second one. If we can apply the theorem of exchange between integral and derivative, we have, for every multi-index with $|\boldsymbol{\alpha}| \leqslant \nu$:

$$D^{\boldsymbol{\alpha}} \int_{\mathcal{S}} \max_{a' \in \mathcal{A}} f(s',a') p_h(s'|s,a)\, ds' = \int_{\mathcal{S}} \max_{a' \in \mathcal{A}} f(s',a') D^{\boldsymbol{\alpha}} p_h(s'|s,a)\, ds'; \tag{6}$$

In such case, using the abbreviation $z = (s,a)$ and $\tilde{f}(s') = \max_{a' \in \mathcal{A}} f(s',a')$ we get:

$$\begin{aligned}D^{\boldsymbol{\alpha}} \int_{\mathcal{S}} \tilde{f}(s') p_h(s'|z_1)\, ds' - D^{\boldsymbol{\alpha}} \int_{\mathcal{S}} \tilde{f}(s') p_h(s'|z_2)\, ds' &= \int_{\mathcal{S}} \tilde{f}(s')(D^{\boldsymbol{\alpha}} p_h(s'|z_1) - D^{\boldsymbol{\alpha}} p_h(s'|z_2))\, ds' \\ &\leqslant \int_{\mathcal{S}} \tilde{f}(s') C_p \|z_1 - z_2\|_2\, ds' \\ &\leqslant \text{Vol}(\mathcal{S}) C_p \|z_1 - z_2\|_2 \|\tilde{f}\|_\infty \\ &\leqslant \text{Vol}(\mathcal{S}) C_p \|z_1 - z_2\|_2 \|f\|_\infty \\ &\leqslant \text{Vol}(\mathcal{S}) C_p \|z_1 - z_2\|_2 \|f\|_{\mathcal{C}^{\nu,1}} \\ &= 2^{d_S} C_p \|z_1 - z_2\|_2 \|f\|_{\mathcal{C}^{\nu,1}}\end{aligned}$$

where the second step comes from the Strongly Smoothness assumption on $p_h$ and the third one from the fact that $\max_{s \in \mathcal{S}} |\max_{a \in \mathcal{A}} f(s,a)| \leqslant \max_{s \in \mathcal{S}} \max_{a \in \mathcal{A}} |f(s,a)|$. This proves that the norm of the operator $\mathcal{T}_h^* f$ is bounded by $\text{Vol}(\mathcal{S}) C_p \|f\|_{\mathcal{C}^{\nu,1}} + \|r\|_{\mathcal{C}^{\nu,1}}$, so that, tanking $C_{\mathcal{T}*} = \max\{1, \text{Vol}(\mathcal{S}) C_p\}$ we have the boundedness of operator $\mathcal{T}^*$. It remains to justify Equation 6, by ensuring that we can differentiate under the integral. This holds (Folland, 1999) under the condition that:

$$\exists g \geqslant 0, \int_{\mathcal{S}} g(s') ds' \leqslant +\infty, \qquad \forall z \quad \left| D^{\boldsymbol{\alpha}} \tilde{f}(s') p_h(s'|z) \right| \leqslant g(s'),$$

where the derivative is intended w.r.t. $z$, as before. Taking $g(s') = C_p \|f\|_{\mathcal{C}^{\nu,1}}$ is already sufficient.

### B.2. Lipschitz MDPs are Weakly Smooth but not Strongly Smooth

We only prove that all Lipschitz MDPs are Weakly Smooth, as the fact that they are not always Strongly Smooth can be proved by simply taking a Lipschitz MDPs that is also deterministic; indeed, no smoothness condition can be imposed if $p_h(\cdot|s,a)$ is a Dirac delta.

**Theorem 5.** *Let $M = (\mathcal{S}, \mathcal{A}, p, r, H)$ be a Lipschitz MDP with constants $L_r, L_p$. Then, the same MDP is Weakly Smooth for $\nu = 0$.*

*Proof.* Let $f \in \mathcal{C}^{0,1}(\mathcal{S} \times \mathcal{A})$ (which corresponds to the space of Lipschitz functions). We start from:

$$\begin{aligned}\mathcal{T}_h^* f(s,a) &= r_h(s,a) + \mathop{\mathbb{E}}_{s' \sim p_h(\cdot|s,a)}[\max_{a' \in \mathcal{A}} f(s',a')]\\ &= r_h(s,a) + \int_{\mathcal{S}} \max_{a' \in \mathcal{A}} f(s',a') p_h(s'|s,a)\, ds'.\end{aligned}$$

By triangular inequality we have:

$$\begin{aligned}\|\mathcal{T}_h^* f\|_{\mathcal{C}^{0,1}} &\leqslant \|r\|_{\mathcal{C}^{0,1}} + \left\|\int_{\mathcal{S}} \max_{a' \in \mathcal{A}} f(s',a') p_h(s'|s,a)\, ds'\right\|_{\mathcal{C}^{0,1}}\\ &= \max\{1, L_r\} + \left\|\int_{\mathcal{S}} \max_{a' \in \mathcal{A}} f(s',a') p_h(s'|s,a)\, ds'\right\|_{\mathcal{C}^{0,1}}.\end{aligned}$$

We have now to evaluate the second part. Indeed, we can bound the infinity norm in this way:

$$\left\|\int_{\mathcal{S}} \max_{a' \in \mathcal{A}} f(s',a') p_h(s'|s,a)\, ds'\right\|_{\infty} \leqslant \|f\|_\infty,$$

so that we have only to bound its Lipschitz constant to bound its norm in $\mathcal{C}^{0,1}$. We have:

$$\begin{aligned}\left|\int_{\mathcal{S}} \max_{a' \in \mathcal{A}} f(s',a') p_h(s'|s_1,a_1)\, ds' - \int_{\mathcal{S}} \max_{a' \in \mathcal{A}} f(s',a') p_h(s'|s_2,a_2) ds'\right| &= \left|\int_{\mathcal{S}} \max_{a' \in \mathcal{A}} f(s',a') \big(p_h(s'|s_1,a_1) - p_h(s'|s_2,a_2)\big)\right|\\ &\leqslant \mathcal{W}(p_h(\cdot|s_1,a_1), p_h(\cdot|s_2,a_2)) \text{Lip}(\max_{a' \in \mathcal{A}} f)\\ &\leqslant \mathcal{W}(p_h(\cdot|s_1,a_1), p_h(\cdot|s_2,a_2)) \text{Lip}(f)\\ &\leqslant \mathcal{W}(p_h(\cdot|s_1,a_1), p_h(\cdot|s_2,a_2)) \|f\|_{\mathcal{C}^{0,1}}\\ &\leqslant L_p \|f\|_{\mathcal{C}^{0,1}}(\|s_1 - s_2\|_2 + \|a_1 - a_2\|_2),\end{aligned}$$

the second passage being valid by definition of Wasserstein distance, the third since the Lipschitz constant of $\max_{a' \in \mathcal{A}} f$ is at most equal to the one of $f$, the fourth by definition of $\|\cdot\|_{\mathcal{C}^{0,1}}$ norm and the last one by definition of Lipschitz MDP. This proves that:

$$\left\|\int_{\mathcal{S}} \max_{a' \in \mathcal{A}} f(s',a') p_h(s'|s,a)\, ds'\right\|_{\mathcal{C}^{0,1}} \leqslant \max\{1, L_p\} \|f\|_{\mathcal{C}^{0,1}}.$$

Overall, this entails:

$$\|\mathcal{T}_h^* f\|_{\mathcal{C}^{0,1}} \leqslant \max\{1, L_r\} + \max\{1, L_p\} \|f\|_{\mathcal{C}^{0,1}},$$

which proves the boundedness of the operator $\mathcal{T}_h^*$. □

### B.3. LinearMDPs are Strongly Smooth

**Theorem 6.** *Let $M = (\mathcal{S}, \mathcal{A}, p, r, H)$ be a Linear MDP with feature map $\boldsymbol{\varphi}$. Then, the same MDP is Strongly Smooth for an order $\nu$ corresponding to the smoothness of $\boldsymbol{\varphi}$.*

*Proof.* By definition, the Linear MDP satisfies, $\forall h \in [H]\ s, s' \in \mathcal{S},\ a \in \mathcal{A}$,

$$r_h(s,a) = \langle \boldsymbol{\theta}_h, \boldsymbol{\varphi}(s,a) \rangle \qquad p_h(s'|s,a) = \langle \boldsymbol{\mu}_h(s'), \boldsymbol{\varphi}(s,a) \rangle.$$

Assuming that the feature map $\boldsymbol{\varphi} \in \mathcal{C}^{\nu,1}(\mathcal{S} \times \mathcal{A}, \mathbb{R}^{d_{\boldsymbol{\varphi}}})$ (this is the space of functions $\mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_{\boldsymbol{\varphi}}}$ such that each of the $d_{\boldsymbol{\varphi}}$ components is in $\mathcal{C}^{\nu,1}(\mathcal{S} \times \mathcal{A})$), we have:

- $r_h \in \mathcal{C}^{\nu,1}(\mathcal{S}\times\mathcal{A})$, being a linear combination of $d_{\boldsymbol{\varphi}}$ functions in $\mathcal{C}^{\nu,1}(\mathcal{S}\times\mathcal{A})$. Moreover, $\|r_h\|_{\mathcal{C}^{\nu,1}} \leqslant \|\boldsymbol{\theta}_h\|_1 \max_{i=1,\dots,d_{\boldsymbol{\varphi}}} \|\boldsymbol{\varphi}_i\|_{\mathcal{C}^{\nu,1}}$.
- For every $s'$, $p_h(s'|\cdot,\cdot) \in \mathcal{C}^{\nu,1}(\mathcal{S}\times\mathcal{A})$ for the same reason, and $\sup_{s'\in\mathcal{S}} \|p_h(s'|\cdot,\cdot)\|_{\mathcal{C}^{\nu,1}} \leqslant \sup_{s'\in\mathcal{S}} \|\boldsymbol{\mu}(s')\|_1 \max_{i=1,\dots,d_{\boldsymbol{\varphi}}} \|\boldsymbol{\varphi}_i\|_{\mathcal{C}^{\nu,1}}$.

As, for linear MDPs, it is assumed that $\max\{\|\boldsymbol{\theta}\|_2, \|\boldsymbol{\mu}(s')\|_2\} \leqslant \sqrt{d_{\boldsymbol{\varphi}}}$ (Jin et al., 2020; Uehara et al., 2021), this ends the proof. □

### B.4. Kernelized MDPs are Strongly Smooth

In this section, we are going to prove that under the cone property, a standard assumption in mathematical analysis, any RKHS with Matérn kernel contains function that are smooth up to a certain degree.

**Proposition 7.** *Let $k_m$ be the Matérn kernel of order $m > 1$. Then, if the domain $\Omega$ satisfies the cone property (see Definition 1 in (Dlotko, 2014)), the corresponding RKHS $\mathcal{H}_{k_m} \subset \mathcal{C}^{\nu-1,1}(\Omega)$ for every $\nu < m$.*

*Proof.* We will actually prove a stronger statement, that is $\mathcal{H}_{k_m} \subset \mathcal{C}^{\nu}(\Omega)$. First, we apply Corollary A.6 from (Tuo & Jeff Wu, 2016), which, under the condition $\lfloor m + d/2 \rfloor > d/2$, which is automatically verified being $m > 1$, ensures that

$$\mathcal{H}_{k_n} \subset W^{m+d/2}(\Omega),$$

where $W^{m+d/2}(\Omega)$ denotes the Sobolev space of order $m + d/2$, containing functions that have $m + d/2$ derivatives in $L^2(\Omega)$. Therefore, we can apply theorems that embed Sobolev spaces into spaces of continuous functions to get the result. Precisely, from Proposition 2 in (Dlotko, 2014), we have, for each $j$ such that $2(m + d/2 - j) > d$,

$$W^{m+d/2}(\Omega) \subset \mathcal{C}^{j}(\Omega).$$

By taking $j = \nu$, we have $2(m + d/2 - \nu) = d + 2(m - \nu) > d$, so that

$$W^{m+d/2}(\Omega) \subset \mathcal{C}^{\nu}(\Omega),$$

provided that $\Omega$ satisfies the cone property. This ends the proof. □

### B.5. Proof of the regret bound for LEGENDRE-ELEANOR (Theorem 1)

Before coming to the actual proof, it is necessary to introduce a result from approximation theory (Schultz, 1969; Bagby et al., 2002; Pleśniak, 2009). We start with a simple lemma of functional analysis, which allows us to draw a relation between the space $\mathcal{C}^{\nu,1}$ we have defined in this article and the more common $\mathcal{C}^{\nu+1}$.

**Lemma 1.** *Let $f \in \mathcal{C}^{\nu,1}(\mathbb{R}^d)$. Then, for every $\varepsilon > 0$, there is $f_\varepsilon \in \mathcal{C}^{\nu+1}(\mathbb{R}^d)$ such that*

$$\|f - f_\varepsilon\|_{L^\infty} \leqslant \varepsilon$$

*and $\|f_\varepsilon\|_{\mathcal{C}^{\nu,1}} \leqslant \|f\|_{\mathcal{C}^{\nu,1}}$.*

*Proof.* Fix $\varepsilon > 0$,and let $\varepsilon' = \|f\|_{\mathcal{C}^{\nu,1}}^{-1}\varepsilon$. Let $\chi(x)$ be the standard mollifier in $\mathbb{R}^d$:

$$\chi(x) = \begin{cases} C\exp\left(\frac{1}{\|x\|^2-1}\right) & \|x\| < 1 \\ 0 & \|x\| \geqslant 1 \end{cases},$$

For a constant $C$ such that the function integrates to one. If we define $\chi_{\varepsilon'}(x) := \frac{1}{\varepsilon'^d}\chi(x/\varepsilon')$, we can take

$$f_{\varepsilon'}(x) := f * \chi_{\varepsilon'}(x) = \int_{\mathbb{R}^d} f(y)\chi_{\varepsilon'}(x-y)\,dy.$$

By the properties of convolution, as $\chi(\cdot)$ is $\mathcal{C}^\infty$, the function $f_{\varepsilon'}(x) \in \mathcal{C}^\infty(\mathbb{R}^d)$. Then, we have

- The bound on the norm difference:

$$\begin{aligned}\|f - f_{\varepsilon'}\|_{L^\infty} &= \|f * \delta_0(x) - f * \chi_{\varepsilon'}(x)\|_{L^\infty}\\ &\leqslant \|f\|_{\mathcal{C}^{\nu,1}} \mathcal{W}(\delta_0(\cdot), \chi_{\varepsilon'}(\cdot))\\ &= \|f\|_{\mathcal{C}^{\nu,1}} \int_{\mathbb{R}^d} \|x\| \chi_{\varepsilon'}(x) dx\\ &\leqslant \varepsilon' \|f\|_{\mathcal{C}^{\nu,1}} = \varepsilon.\end{aligned}$$

  Where $\delta_0(\cdot)$ is the Dirac's delta, the second passage is valid by definition of Wassertein distance thanks to the fact that the Lipschitz constant of $f$ is bounded by $\|f\|_{\mathcal{C}^{\nu,1}}$, and the last is due to the fact that $\chi_{\varepsilon'}(\cdot)$ has integral $1$, support in a ball of radius $\varepsilon'$, and center in the origin.

- The bound on the derivatives: for every $|\boldsymbol{\alpha}| \leqslant \nu + 1$,

$$\|D^{\boldsymbol{\alpha}} f_{\varepsilon'}\|_{L^\infty} = \|\chi_{\varepsilon'} * D^{\boldsymbol{\alpha}} f\|_{L^\infty} \leqslant \|D^{\boldsymbol{\alpha}} f\|_{L^\infty},$$

  the last being valid since $\chi_{\varepsilon'}$ has integral one. Therefore,

$$\|f_{\varepsilon'}\|_{\mathcal{C}^{\nu,1}} = \max_{|\boldsymbol{\alpha}| \leqslant \nu+1} \|D^{\boldsymbol{\alpha}} f_{\varepsilon'}\|_{L^\infty} \leqslant \max_{|\boldsymbol{\alpha}| \leqslant \nu+1} \|D^{\boldsymbol{\alpha}} f\|_{L^\infty} = \|f\|_{\mathcal{C}^{\nu,1}}.$$

This ends the proof. □

One key ingredient of our next results will be a theorem from approximation theory. This theorem, which is built on a family of result known as Jackson's theorems, ensures that we are able to approximate smooth functions with polynomials, with an error that is lower the more continuous derivative the function has.

**Theorem 8.** *For every $\nu, d \in \mathbb{N}$, there is a constant $J_{d,\nu}$ such that for every function $f : [-1,1]^d \to \mathbb{R}$ in $\mathcal{C}^{\nu,1}([-1,1]^d)$ it holds, for $N > \nu$,*

$$\exists p_N \in \mathcal{P}_N : \ \|f - p_N\|_{L^\infty} \leqslant 2 J_{d,\nu} \|f\|_{\mathcal{C}^{\nu,1}} N^{-\nu-1},$$

*where $\mathcal{P}_N$ is the space of multivariate polynomials of degree at most $N$. Moreover, $\|p_N\|_{\mathcal{C}^{\nu,1}} \leqslant 3 J_{d,\nu} \|f\|_{\mathcal{C}^{\nu,1}}$.*

*Proof.* This proof will be based on Theorem 1 from (Bagby et al., 2002), which says that, for any function in $\mathcal{C}^{\nu+1}$ with compact support ($[-1,1]^d$ in our case), there is a polynomial $p_N \in \mathcal{P}_N$ such that, for every multi-index $\boldsymbol{\alpha}$ such that $|\boldsymbol{\alpha}| \leqslant \nu + 1$,

$$\|D^{\boldsymbol{\alpha}} f - D^{\boldsymbol{\alpha}} p_N\|_{L^\infty} \leqslant J_{d,\nu} N^{|\boldsymbol{\alpha}|-\nu-1} \omega_{f,\nu}(N^{-1}), \tag{7}$$

where $J_{d,\nu}$ is a constant, which we can impose to be $> 1$ without loss of generality, and $\omega_{f,\nu+1}(\cdot)$ is the $\nu-$modulus of continuity, defined as

$$\omega_{f,\nu+1}(\delta) := \sup_{|\boldsymbol{\alpha}| < \nu+1} \sup_{\|x-y\|_2 \leqslant \delta} |D^{\boldsymbol{\alpha}} f(x) - D^{\boldsymbol{\alpha}} f(y)|.$$

This theorem cannot be applied on $f$, which is not in $\mathcal{C}^{\nu+1}([-1,1]^d)$, therefore we apply it on the mollified function $f_\varepsilon$ obtained from Lemma 1. Note that

$$\omega_{f_\varepsilon,\nu+1}(\delta) \leqslant 2 \sup_{|\boldsymbol{\alpha}| < \nu+1} \|D^{\boldsymbol{\alpha}} f_\varepsilon\|_{L^\infty} \leqslant 2\|f_\varepsilon\|_{\mathcal{C}^{\nu,1}} \leqslant 2\|f\|_{\mathcal{C}^{\nu,1}},$$

where the last inequality is from Lemma 1. This allows us to see that, taking $\boldsymbol{\alpha} = 0$ (the zero multi-index) in Equation 7,

$$\|f_\varepsilon - p_N\|_{L^\infty} \leqslant 2\|f\|_{\mathcal{C}^{\nu,1}} J_{d,\nu} N^{-\nu-1}. \tag{8}$$

Moreover, taking the other values of $\boldsymbol{\alpha}$ results in

$$\forall \boldsymbol{\alpha}: \ |\boldsymbol{\alpha}| \leqslant \nu+1, \qquad \|D^{\boldsymbol{\alpha}} f_\varepsilon - D^{\boldsymbol{\alpha}} p_N\|_{L^\infty} \leqslant 2\|f\|_{\mathcal{C}^{\nu,1}} J_{d,\nu}. \tag{9}$$

These two results allow us to obtain the thesis: applying the triangular inequality to Equation 8, by Lemma 1,

$$\|f - p_N\|_{L^\infty} \leqslant \|f_\varepsilon - f\|_{L^\infty} + \|f_\varepsilon - p_N\|_{L^\infty} \leqslant \varepsilon + 2\|f\|_{\mathcal{C}^{\nu,1}} J_{d,\nu} N^{-\nu-1}.$$

Since this is valid for every $\varepsilon > 0$, we obtain $\|f - p_N\|_{L^\infty} \leqslant 2\|f\|_{\mathcal{C}^{\nu,1}} J_{d,\nu} N^{-\nu-1}$, proving the first statement. As for the second statement, applying the triangular inequality to Equation 9 leads to

$$\begin{aligned}
\|p_N\|_{\mathcal{C}^{\nu,1}} &= \max_{|\boldsymbol{\alpha}| \leqslant \nu+1} \|D^{\boldsymbol{\alpha}} p_N\|_{L^\infty} \\
&\leqslant \max_{|\alpha| \leqslant \nu+1} \|D^{\boldsymbol{\alpha}} f_\varepsilon\|_{L^\infty} + \|D^{\boldsymbol{\alpha}} f_\varepsilon - D^{\boldsymbol{\alpha}} p_N\|_{L^\infty} \\
&\leqslant \max_{|\boldsymbol{\alpha}| \leqslant \nu+1} \|D^{\boldsymbol{\alpha}} f_\varepsilon\|_{L^\infty} + 2\|f\|_{\mathcal{C}^{\nu,1}} J_{d,\nu} \qquad \text{(Equation 9)} \\
&\leqslant 3\|f\|_{\mathcal{C}^{\nu,1}} J_{d,\nu},
\end{aligned}$$

Where we have used lemma 1 in the last passage to bound $\max_{|\boldsymbol{\alpha}| \leqslant \nu+1} \|D^{\boldsymbol{\alpha}} f_\varepsilon\|_{L^\infty}$, and also the fact that $J_{d,\nu} \geqslant 1$. □

The main part of the proof of Theorem 1 revolves around showing that Weakly Smooth MDPs paired with Legendre representation map have low inherent Bellman error with respect to some sequence of sets $\mathcal{B}_h$. Recall the definition of inherent Bellman error:

$$\mathcal{I} := \max_{h \in [H]} \sup_{\theta \in \mathcal{B}_{h+1}} \inf_{\theta' \in \mathcal{B}_h} \|\boldsymbol{\varphi}(s,a)^\top \theta' - \mathcal{T}^* Q(\theta)(s,a)\|_{L^\infty}.$$

where $Q_\theta(s,a)$ is the function $\boldsymbol{\varphi}(s,a)^\top \theta$.

**Theorem 9.** *Let $M$ be a Weakly Smooth MDP. Let us consider the pair $(M, \boldsymbol{\varphi}_{L,N}^d)$ given by the MDP and the Legendre feature map of degree $N$. There is a sequence of compact sets $\mathcal{B}_h \subset \mathbb{R}^{\widetilde{N}}$ such that the inherent Bellman error of $(M, \boldsymbol{\varphi}_{L,N}^d)$ w.r.t. $\{\mathcal{B}_h\}_h$ satisfies, for $N > \nu$,*

$$\mathcal{I} \leqslant 2J_{d,\nu} C_{\mathcal{T}*} \left( \frac{(3J_{d,\nu} C_{\mathcal{T}*})^H - 1}{3J_{d,\nu} C_{\mathcal{T}*} - 1} + 1 \right) N^{-\nu-1},$$

*where $J_{d,\nu}$ is the constant from Theorem 8.*

*Proof.* First thing, we have to define the sequence of compact sets $\mathcal{B}_h \subset \mathbb{R}^{\widetilde{N}}$. We define

$$\mathcal{B}_h := \left\{ \theta \in \mathbb{R}^{\widetilde{N}} : \|\boldsymbol{\varphi}_{L,N}^d(\cdot,\cdot)^\top \theta\|_{\mathcal{C}^{\nu,1}} \leqslant B(h) \right\}, \tag{10}$$

for a constant $B(h)$ to be defined later.

Now, for every $\theta \in \mathcal{B}_{h+1}$, we have

$$\mathcal{T}^* Q(\theta)(s,a) = \mathcal{T}^* \boldsymbol{\varphi}_{L,N}^d(s,a)^\top \theta.$$

Note that, being $\boldsymbol{\varphi}_{L,N}^d(s,a)^\top \theta$ the scalar product between a constant and a vector of polynomials, it is also $\mathcal{C}^\infty$. Moreover, by definition of $\mathcal{B}_{h+1}$, we have, for all $\theta \in \mathcal{B}_{h+1}$,

$$\|\boldsymbol{\varphi}_{L,N}^d(\cdot,\cdot)^\top \theta\|_{\mathcal{C}^{\nu,1}} \leqslant B(h+1).$$

So, having assumed that the process is Weakly Smooth of order $\nu$,

$$\|\mathcal{T}^* \boldsymbol{\varphi}_{L,N}^d(\cdot,\cdot)^\top \theta\|_{\mathcal{C}^{\nu,1}} \leqslant C_{\mathcal{T}*} \left( \|\boldsymbol{\varphi}_{L,N}^d(\cdot,\cdot)^\top \theta\|_{\mathcal{C}^{\nu,1}} + 1 \right) \leqslant C_{\mathcal{T}*} \left( B(h+1) + 1 \right). \tag{11}$$

Applying Theorem 8, this entails the existence of a polynomial $p_N \in \mathcal{P}_N$ such that

$$\|\mathcal{T}^* \boldsymbol{\varphi}_{L,N}^d(\cdot,\cdot)^\top \theta - p_N(\cdot)\|_{L^\infty} \leqslant 2J_{d,\nu} C_{\mathcal{T}*} \left( B(h+1) + 1 \right) N^{-\nu-1}. \tag{12}$$

If we prove that there is $\theta' \in \mathcal{B}_h$ such that $p_N(\cdot) = \boldsymbol{\varphi}_{L,N}^d(\cdot,\cdot)^\top \theta'$, then we have proved that the inherent Bellman error is bounded by the right hand side of equation (12). The fact that this $\theta' \in \mathcal{B}_h$ exists follows from two considerations:

1. As the set $\{\boldsymbol{\varphi}_{L,N}^d(\cdot,\cdot)_i\}_{i=1}^{\widetilde{N}}$ is a basis for the vector space of $d-$variate polynomials of degree $N$, there is $\theta' \in \mathbb{R}^{\widetilde{N}}$ such that $p_N(\cdot) = \boldsymbol{\varphi}_{L,N}^d(\cdot,\cdot)^\top \theta'$.
2. From Theorem 8 we also have $\|p_N\|_{\mathcal{C}^{\nu,1}} \leqslant 3J_{d,\nu} C_{\mathcal{T}*} \left( B(h+1) + 1 \right)$.

Therefore, from the second point, it is sufficient that the value of $B(h)$ in the definition of $\mathcal{B}_h$ satisfies

$$B(h) \geqslant 3J_{d,\nu} C_{\mathcal{T}*} \left( B(h+1) + 1 \right),$$

which is in particular satisfied by the choice

$$B(h) = \sum_{\tau=1}^{H-h} (3J_{d,\nu} C_{\mathcal{T}*})^\tau = \frac{(3J_{d,\nu} C_{\mathcal{T}*})^{H-h+1} - 1}{3J_{d,\nu} C_{\mathcal{T}*} - 1} - 1 \tag{13}$$

substituting this value into equation (12), we get

$$\mathcal{I} \leqslant 2J_{d,\nu} C_{\mathcal{T}*} \left( \frac{(3J_{d,\nu} C_{\mathcal{T}*})^H - 1}{3J_{d,\nu} C_{\mathcal{T}*} - 1} \right) N^{-\nu-1},$$

which ends the proof. □

Now that we have proved Theorem 9, it is sufficient to apply the results of the literature for the case of MDPs with low inherent Bellman error (Zanette et al., 2020) to achieve a regret bound. For our convinience, we report here this result

**Theorem 10.** *(Assumption 1 and Theorem 1 from (Zanette et al., 2020)) Let $(M, \boldsymbol{\varphi})$ be a pair MDP-feature map that satisfy the low-inherent Bellmann error assumption with respect to a sequence of sets $\{\mathcal{B}_h\}_{h=1}^H$ (see (5)). Assume that,*

1. *$|Q_h^\pi(s,a)| \leqslant 1$ for every $h, s, a$ and every policy $\pi$.*
2. *$\|\boldsymbol{\varphi}(s,a)\|_2 \leqslant 1$ for every $s, a$.*
3. *The reward noise is $1-$subgaussian.*
4. *The sets $\{\mathcal{B}_h\}_{h=1}^H$ are all compact and define $N_h := \sup_{\theta \in \mathcal{B}_h} \|\theta\|_2^2$.*

*Then, the regret of* ELEANOR *applied on $(M, \boldsymbol{\varphi})$ satisfies, with probability at least $1 - \delta$*

$$R_K \leqslant \widetilde{\mathcal{O}} \left( \sum_{h=1}^H N_h \sqrt{K} + \sum_{h=1}^H \sqrt{N}_h \mathcal{I} K \right)$$

We can pass trough this result to achieve a regret bound for every Weakly Smooth MDP.

**Theorem 11.** *Let us consider a Weakly Smooth MDP $M$ with state action space $[-1,1]^d$. Under the condition that $\nu > d/2-1$,* LEGENDRE-ELEANOR*, with probability at least $1-\delta$, suffers a regret of order at most:*

$$R_K \leqslant \widetilde{\mathcal{O}}\left(C_{\text{ELE}}^H\left(\widetilde{N}\sqrt{K} + N^{-\nu-1}\sqrt{\widetilde{N}}K\right)\right),$$

*where the constant depends only on $d$ and $\nu$ and the $\widetilde{\mathcal{O}}$ hides logarithmic functions of $K$, $\delta$.*

*Proof.* By design of the algorithm, to prove the regret bound we have to show that the couple given by the MDP $(M, \boldsymbol{\varphi}_{L,N}^d)$ satisfies the assumptions of theorem 10 and then apply its regret bound. Here, we report, point by point, why every assumption is verified.

1. The fact that $|Q_h^\pi(s,a)| \leqslant 1$ is assumed.

2. The feature map satisfies, for every $s \in \mathcal{S}, a \in \mathcal{A}$,

$$\begin{aligned}\|\boldsymbol{\varphi}_{L,N}^d(s,a)\|_2 &= \widetilde{N}^{-1/2}\left\|\left\{\varphi_{L,N_1}(x_1)\times\varphi_{L,N_2}(x_2)\ldots\varphi_{L,N_d}(x_d) : \sum_{i=1}^d N_i \leqslant N\right\}\right\|_2 \\ &\leqslant \widetilde{N}^{-1/2}\sqrt{\sum_{i=1}^{\widetilde{N}} 1} = 1.\end{aligned}$$

The last inequality being valid due to the fact that Legendre polynomials are bounded in $[-1,1]$.

3. Follows from the sub-Gaussianity of the noise and the fact that the state-action value function of every policy is bounded.

4. This is the only difficult point. We start proving that the sets $\mathcal{B}_h$, defined as in Equation (10) are compact. Let $B(h)$ be defined as in Equation (13). By the Heine-Borel theorem (Rudin, 1974), a subset of $\mathbb{R}^{\widetilde{N}}$ is compact is and only if it is closed and bounded. We start proving the closure, as we get boundedness from free by the norm inequality proved later. Let $\{\theta_n\}_n \subset \mathcal{B}_h$ such that $\theta_n \to \theta$. Then,

$$\begin{aligned}\|\boldsymbol{\varphi}_{L,N}^d(\cdot,\cdot)^\top(\theta-\theta_n)\|_{\mathcal{C}^{\nu,1}} &= \max_{|\boldsymbol{\alpha}|\leqslant\nu+1}\|D^{\boldsymbol{\alpha}}\varphi_{L,N}^d(\cdot,\cdot)^\top(\theta-\theta_n)\|_{L^\infty} \\ &\leqslant \max_{|\boldsymbol{\alpha}|\leqslant\nu+1}\sup_{s,a}|D^{\boldsymbol{\alpha}}\varphi_{L,N}^d(\cdot,\cdot)^\top(\theta-\theta_n)| \\ &\leqslant \max_{|\boldsymbol{\alpha}|\leqslant\nu+1}\sup_{s,a}\|D^{\boldsymbol{\alpha}}\varphi_{L,N}^d(\cdot,\cdot)^\top\|_2\|\theta-\theta_n\|_2 \\ &= \max_{|\boldsymbol{\alpha}|\leqslant\nu+1}\|\|D^{\boldsymbol{\alpha}}\varphi_{L,N}^d(\cdot,\cdot)^\top\|_2\|_{L^\infty}\|\theta-\theta_n\|_2\end{aligned}$$

where we have used the Cauchy-Schwartz inequality. Now, note that $\varphi_{L,N}^d(\cdot,\cdot)^\top$ is a vector valued function with any component being a Legendre polynomial, so $\mathcal{C}^\infty$ in particular. Thus, $\max_{|\boldsymbol{\alpha}|\leqslant\nu+1}\|\|D^{\boldsymbol{\alpha}}\varphi_{L,N}^d(\cdot,\cdot)^\top\|_2\|_{L^\infty}$ is bounded and we have

$$\|\boldsymbol{\varphi}_{L,N}^d(\cdot,\cdot)^\top(\theta-\theta_n)\|_{\mathcal{C}^{\nu,1}} \leqslant \underbrace{\max_{|\boldsymbol{\alpha}|\leqslant\nu+1}\|\|D^{\boldsymbol{\alpha}}\varphi_{L,N}^d(\cdot,\cdot)^\top\|_2\|_{L^\infty}}_{<+\infty}\underbrace{\|\theta-\theta_n\|_2}_{\to 0} \to 0.$$

Moreover, we have, by reverse triangular inequality,

$$\|\boldsymbol{\varphi}_{L,N}^d(\cdot,\cdot)^\top\theta\|_{\mathcal{C}^{\nu,1}} \leqslant \inf_n\|\boldsymbol{\varphi}_{L,N}^d(\cdot,\cdot)^\top(\theta-\theta_n)\|_{\mathcal{C}^{\nu,1}} + \|\boldsymbol{\varphi}_{L,N}^d(\cdot,\cdot)^\top\theta_n\|_{\mathcal{C}^{\nu,1}}$$

$$\begin{aligned}&\leqslant \inf_n \|\boldsymbol{\varphi}_{L,N}^d(\cdot,\cdot)^\top(\theta-\theta_n)\|_{\mathcal{C}^{\nu,1}} + B(h)\\ &= B(h).\end{aligned}$$

Where we have used the fact that $\theta_n \in \mathcal{B}_h$ to bound $\|\boldsymbol{\varphi}_{L,N}^d(\cdot,\cdot)^\top\theta_n\|_{\mathcal{C}^{\nu,1}}$, and the fact that $\|\boldsymbol{\varphi}_{L,N}^d(\cdot,\cdot)^\top(\theta-\theta_n)\|_{\mathcal{C}^{\nu,1}} \to 0$ to ensure that $\inf_n \|\boldsymbol{\varphi}_{L,N}^d(\cdot,\cdot)^\top(\theta-\theta_n)\|_{\mathcal{C}^{\nu,1}} = 0$. This proves that $\theta \in \mathcal{B}_h$, which means that the set is closed.

The norm inequality follows from the fact that the Legendre polynomials form an orthogonal basis of $L^2(\mathcal{S}\times\mathcal{A})$. Indeed we have, by definition, that $\forall\theta\in\mathcal{B}_h, \|\boldsymbol{\varphi}_{L,N}^d(\cdot,\cdot)^\top\theta\|_{L^\infty} \leqslant \|\boldsymbol{\varphi}_{L,N}^d(\cdot,\cdot)^\top\theta\|_{\mathcal{C}^{\nu,1}} \leqslant B(h)$.

Being in a bounded domain $\mathcal{S}\times\mathcal{A}\subset[-1,1]^d$, the $L^\infty$ norm is stronger that the $L^2$ one, and precisely we have $\|\cdot\|_{L^2} \leqslant \sqrt{Vol(\mathcal{S}\times\mathcal{A})}\|\cdot\|_{L^\infty}$. Therefore, we have

$$\forall\theta\in\mathcal{B}_h, \|\boldsymbol{\varphi}_{L,N}^d(\cdot,\cdot)^\top\theta\|_{L^2} \leqslant \sqrt{Vol(\mathcal{S}\times\mathcal{A})}B(h). \tag{14}$$

Here the definition Legendre polynomials plays a crucial role: as $\{\boldsymbol{\varphi}_{L,N}^d(\cdot,\cdot)_i\}_{i=1}^{\widetilde{N}}$ is an orthogonal sequence normalized in $L^2$ to $\widetilde{N}^{-1/2}$, it follows from Parseval's theorem (Rudin, 1974) on the Hilbert space $L^2(\mathcal{S}\times\mathcal{A})$ that we can bound $\|\boldsymbol{\varphi}_{L,N}^d(\cdot,\cdot)^\top\theta\|_{L^2}$. Indeed,

$$\begin{aligned}\|\boldsymbol{\varphi}_{L,N}^d(\cdot,\cdot)^\top\theta\|_{L^2} &= \left\|\sum_{i=1}^{\widetilde{N}}[\boldsymbol{\varphi}_{L,N}^d]_i(\cdot,\cdot)\theta_i\right\|_{L^2}\\ &\overset{(Par)}{=} \sqrt{\sum_{i=1}^{\widetilde{N}}\|[\boldsymbol{\varphi}_{L,N}^d]_i(\cdot,\cdot)\|_{L^2}^2\theta_i^2}\\ &= \widetilde{N}^{-1/2}\sqrt{\sum_{i=1}^{\widetilde{N}}\theta_i^2}\\ &= \widetilde{N}^{-1/2}\|\theta\|_2.\end{aligned}$$

Where at passage $(Par)$ we have used Parseval's theorem, exploiting the fact that the $\widetilde{N}$ components of $\boldsymbol{\varphi}_{L,N}^d$ are all orthogonal in $L^2(\mathcal{S}\times\mathcal{A})$, by definition of Legendre polynomials, and have been normalized to $\widetilde{N}^{-1/2}$. Note that Parseval theorem works also for infinite components, but here we are considering a function $\varphi_{L,N}^d(\cdot,\cdot)^\top\theta$ which is a linear combination only of the first $\widetilde{N}$ elements of the Legendre basis, so that all the following components are identically zero.

Substituting this result into Equation 14, we have

$$\|\theta\|_2 \leqslant \sqrt{Vol(\mathcal{S}\times\mathcal{A})}B(h)\widetilde{N}^{1/2}.$$

Having the additional term $\sqrt{Vol(\mathcal{S}\times\mathcal{A})}B(h)$ multiplying $\widetilde{N}^{1/2}$ has the effect of enlarging the regret of the same quantity, which still does not depend on $N$.

For this reason, Theorem 1 from Zanette et al. (2020) results in the following regret bound in high probability

$$R_K \leqslant \widetilde{O}\left(\sqrt{Vol(\mathcal{S}\times\mathcal{A})}B(1)\left[H\widetilde{N}\sqrt{K} + \mathcal{I}\sqrt{\widetilde{N}}K\right]\right),$$

which, once Theorem 9 is applied to bound the inherent Bellman error, leads to

$$R_K \leqslant \widetilde{O}\left(\sqrt{Vol(\mathcal{S}\times\mathcal{A})}B(1)\left[H\widetilde{N}\sqrt{K} + 2J_{d,\nu}C_{\mathcal{T}*}\left(\frac{(3J_{d,\nu}C_{\mathcal{T}*})^H-1}{3J_{d,\nu}C_{\mathcal{T}*}-1}+1\right)N^{-\nu-1}\sqrt{\widetilde{N}}K\right]\right).$$

Grouping all the constants independent of $N$ and $K$, we get the factor

$$2\sqrt{Vol(\mathcal{S}\times\mathcal{A})}B(1)HJ_{d,\nu}C_{\mathcal{T}*}\left(\frac{(3J_{d,\nu}C_{\mathcal{T}*})^H-1}{3J_{d,\nu}C_{\mathcal{T}*}-1}+1\right)\leqslant C_{\text{ELE}}^H,$$

for a suitably large constant $C_{\text{ELE}}$ only depending on $d$ and $\nu$. □

At this point, it is easy to achieve the regret bound in the form of Theorem 1: from a regret bound in the form of theorem 11, we can substitute the value of $N=\lceil K^\beta\rceil$ to achieve

$$R_K\leqslant\widetilde{\mathcal{O}}\left(K^{d\beta}\sqrt{K}+K^{-\beta(\nu+1)}K^{d\beta/2}K\right).$$

Imposing that the exponents are equal leads to $d\beta+\frac{1}{2}=1+\beta(d/2-\nu-1)\implies\beta(d/2+\nu+1)=\frac{1}{2}\implies\beta=\frac{1}{d+2(\nu+1)}$. Substituting in the regret bound, we get precisely

$$R_K\leqslant\widetilde{\mathcal{O}}\left(K^{\frac{3d/2+\nu+1}{d+2(\nu+1)}}\right),$$

which is the statement of Theorem 1.

### B.6. Lipschitz MDPs have regret bound exponential in $H$

In this section we prove that every regret bound for algorithms in the Lipschitz MDP setting must grow exponentially with the time horizon $H$. The proof strategy is the following: we start from an instance of a Lipschitz bandit problem with a Lipschitz constant that is exponential in $H$, and show that this can be reduced to a standard Lipschtz MDP (where all Lipschitz constants are independent on $H$). Since it has been shown that the regret bound in a Lipschitz bandit problem is proportional to the Lipschitz constant, this shows that the regret of the Lipschitz MDP is also exponential in $H$.

**Theorem 12.** *The regret in a Lipschtz MDP is at least of order* $R_T=\Omega(L_p^{\frac{d(H-2)}{d+2}}K^{\frac{d+1}{d+2}})$.

*Proof.* Let $f:[-1,1]^d\to[-1,1]$ be an $2L_p^{H-2}-$Lipschtz function and $\eta$ a noise bounded in $[-1,1]$.

Define $\widetilde{f}:=(L_p^{-H+2}/2)f,\widetilde{\eta}:=(L_p^{-H+2}/2)\eta$, so that $\widetilde{f}:[-1,1]^d\to[-L_p^{-H+2}/2,L_p^{-H+2}/2]$ is a $1/2-$Lipschitz function and $\widetilde{\eta}$ a noise bounded in $[-L_p^{-H+2}/2,L_p^{-H+2}/2]$. Define the following MDP:

- The state and action space coincide: $\mathcal{S}=\mathcal{A}=[-1,1]^{d/2}$. In this way, $\mathcal{S}\times\mathcal{A}=[-1,1]^d$
- The starting state is $[0,\dots 0]$ almost surely.
- The transition function is defined in the following way:
  - For $h=1$, $p_1(s'|s,a)=\delta(s'=a)$, so that the first action becomes the second state.
  - For $h=2$, $p_2(s'|s,a)=\delta(s'^{(1)}=\widetilde{f}(s,a)+\widetilde{\eta})\prod_{i=2}^{d/2}\delta_0(s'^{(i)})$, meaning that the next state has the first coordinate equal to $\widetilde{f}(s,a)$ plus the noise $\eta$, and all the other ones set to zero. Note that this is coherent with the definition of $f$, which goes $[-1,1]^{d/2}=\mathcal{S}\times\mathcal{A}\to\mathbb{R}$.
  - For $h=2,\dots H$ we have $p_h(s'|s,a)=\delta(s'=L_ps)$, so that the next state is the previous one times a constant (note that, by the bounds on $\widetilde{f},\widetilde{\eta}$, the state never hits the boundary).
- The reward function $r_h$ is zero for the first $H-1$ time steps, and $r_H(s,a)=s^{(1)}$, the first component of the state.

By definition, it is easy to check that the MDP is Lipschtz with $L_p=L_p$ and $L_r=1$.

In this very peculiar MDP, where only the first two actions $a_1$ and $a_2$ matter, the return can be expresses as a function of them. Precisely, since the reward is only given at the last time step, we have

$$\text{Return}(a_1, a_2) = L_p^{H-2}(\widetilde{f}(a_1, a_2) + \widetilde{\eta}) = \frac{1}{2}(f(a_1, a_2) + \eta).$$

In this way, we have shown that the return for this Lipschitz MDP corresponds exact to the feedback in the Lipschitz bandit problem with reward function $\widetilde{f}/2$ (which is $L_p^{H-2}$-LC) and noise $\eta/2$.

This shows that any Lipschitz bandit problem with Lipschtz constant $L_P^{H-2}$ can be reduced to to a Lipschtz MDP with constants bounded independently of $H$. Therefore, the regret on the latter problem is as most as high as the one of the former one. As the regret of the latter is well-known to be of order $\Omega(L^{\frac{d}{d+2}}K^{\frac{d+1}{d+2}}) = \Omega(L_p^{(H-2)\frac{d}{d+2}}K^{\frac{d+1}{d+2}})$, the proof is complete. □

### B.7. Proof of the regret bound for LEGENDRE-LSVI (Theorem 6.3)

For our convenience, we recall here the main result about Linear MDPs that we are going to use to prove our regret bound.

**Theorem 13.** *(Assumption B + thm. 3.2 from (Jin et al., 2020)). Let $(M, \boldsymbol{\varphi})$ a pair MDP-feature map with $\boldsymbol{\varphi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{\widetilde{N}}$, $\boldsymbol{\mu}(\cdot) : \mathcal{S} \to \mathbb{R}^{\widetilde{N}}$ a vector of signed measures and $\zeta$ a positive number that satisfy, for any $s, a, s', h$,*

1. $\|\boldsymbol{\varphi}(s, a)\|_2 \leqslant 1$ *and* $\|\boldsymbol{\mu}(s')\|_2 \leqslant \sqrt{\widetilde{N}}$
2. $\|\boldsymbol{\theta}_h\|_2 \leqslant \sqrt{\widetilde{N}}$
3. *TV*$(p_h(\cdot|s, a) - \langle \boldsymbol{\varphi}(s, a), \boldsymbol{\mu}_h(\cdot)\rangle) \leqslant \zeta$
4. $|r_h(s, a) - \langle \boldsymbol{\varphi}(s, a), \boldsymbol{\theta}_h\rangle| \leqslant \zeta$

*Then, algorithm LSVI-UCB satisfies, for every $\delta > 0$, with probability at least $1 - \delta$*

$$R_K \leqslant \widetilde{\mathcal{O}}(H^{3/2}\widetilde{N}^{3/2}\sqrt{K} + \zeta N H K),$$

*where $\widetilde{\mathcal{O}}$ hides quantities that are logarithmic in $H, K, N, \delta$.*

We now have to prove that any Strongly Smooth MDP equipped with a Legendre feature map becomes a LinearMDP.

**Theorem 14.** *Let us consider a Strongly Smooth MDP $M$ with state action space $[-1, 1]^d$. Under the condition that $d \leqslant \nu + 1$,* LEGENDRE-LSVI, *with probability at least $1 - \delta$, suffers a regret of order at most:*

$$R_K \leqslant \widetilde{\mathcal{O}}\left(H^{3/2}\widetilde{N}^{3/2}\sqrt{K} + H^{3/2}N^{-\nu-1}\widetilde{N}K\right).$$

*where $\widetilde{\mathcal{O}}$ hides logarithmic functions of $K$, $\delta$, and $H$.*

*Proof.* By design of the algorithm, to prove the regret bound we have to show that the couple given by the MDP and the Legendre feature map forms a $\zeta-$approximate LinearMDP, so that LSVI-UCB is guaranteed to work. To prove that the MDP is a $\zeta-$approximate LinearMDP we have to satisfy the assumptions to apply theorem 13. The first step is to show what are the two components $\boldsymbol{\mu}_h$ and $\theta_h$ in our setting. Indeed, by assuming that the MDP is Strongly Smooth,

$$\forall h \in [H]\ \forall s' \in \mathcal{S}, \qquad r_h(\cdot, \cdot), p_h(s'|\cdot, \cdot) \in \mathcal{C}^{\nu,1}(\mathcal{S} \times \mathcal{A}),$$

with $\sup_{h,s'} \|p(s'|\cdot, \cdot)\|_{\mathcal{C}^{\nu,1}} := C_p < \infty$ and $\sup_{h,s'} \|r(\cdot, \cdot)\|_{\mathcal{C}^{\nu,1}} := C_r < +\infty$. Theorem 8 ensures that

$$\forall s' \in \mathcal{S},\ \forall h \in [H],\ \exists p_N \in \mathcal{P}_N :\ \|p_h(s'|\cdot, \cdot) - p_N(\cdot)\|_{L^\infty} \leqslant 2J_{d,\nu}C_pN^{-\nu-1},$$
$$\forall h \in [H],\ \exists p_N \in \mathcal{P}_N :\ \|r_h(\cdot, \cdot) - p_N(\cdot)\|_{L^\infty} \leqslant 2J_{d,\nu}C_rN^{-\nu-1}.$$

As the set $\{\varphi_{L,N}^d(s, a)_i\}_{i=1}^{\widetilde{N}}$ is a basis for the vector space $\mathcal{P}_N$ of $d-$variate polynomials of degree $N$, we can define $\boldsymbol{\mu}_h(s') \in \mathbb{R}^{\widetilde{N}}$ and $\theta_h \in \mathbb{R}^{\widetilde{N}}$ to be the coefficients such that

$$\forall s' \in \mathcal{S},\ \forall h \in [H]:\ \|p_h(s'|\cdot,\cdot) - \boldsymbol{\varphi}_{L,N}^d(\cdot,\cdot)^\top \boldsymbol{\mu}_h(s')\|_{L^\infty} \leqslant 2J_{d,\nu}C_p N^{-\nu-1}, \tag{15}$$

$$\forall h \in [H]:\ \|r_h(\cdot) - \boldsymbol{\varphi}_{L,N}^d(\cdot,\cdot)^\top \theta_h\|_{L^\infty} \leqslant 2J_{d,\nu}C_r N^{-\nu-1}. \tag{16}$$

Now we just have to prove that the four assumtpions of theorem 13 hold,

1. The norm bound on the feature map follows exactly as in the proof of Theorem 11. For every $s \in \mathcal{S}, a \in \mathcal{A}$,

$$\begin{aligned}\|\boldsymbol{\varphi}_{L,N}^d(s,a)\|_2 &= \widetilde{N}^{-1/2}\left\|\left\{\varphi_{L,N_1}(x_1) \times \varphi_{L,N_2}(x_2) \dots \varphi_{L,N_d}(x_d) : \sum_{i=1}^d N_i \leqslant N\right\}\right\|_2 \\ &\leqslant \widetilde{N}^{-1/2}\sqrt{\sum_{i=1}^{\widetilde{N}} 1} = 1.\end{aligned}$$

The inequality being valid due to the fact that Legendre polynomials are bounded in $[-1,1]$.

2. Since $\{\boldsymbol{\varphi}_{L,N}^d(s,a)_i\}_{i=1}^{\widetilde{N}}$ is an orthogonal sequence normalized in $L^2$ to $\widetilde{N}^{-1/2}$, by Parseval's theorem (Rudin, 1974), for all $s' \in \mathcal{S}, h \in [H]$,

$$\begin{aligned}\|\boldsymbol{\mu}_h(s')\|_2 &= \widetilde{N}^{1/2}\|\boldsymbol{\varphi}_{L,N}^d(\cdot,\cdot)^\top \boldsymbol{\mu}_h(s')\|_{L^2} \\ &\leqslant \widetilde{N}^{1/2}\sqrt{Vol(\mathcal{S}\times\mathcal{A})}\|\boldsymbol{\varphi}_{L,N}^d(\cdot,\cdot)^\top \boldsymbol{\mu}_h(s')\|_{L^\infty} \\ &\leqslant 2\widetilde{N}^{1/2}\sqrt{Vol(\mathcal{S}\times\mathcal{A})}\|p_h(s'|\cdot,\cdot)\|_{L^\infty} \leqslant 2\widetilde{N}^{1/2}C_p\sqrt{Vol(\mathcal{S}\times\mathcal{A})}.\end{aligned}$$

Where equality is by Parseval's theorem (cf. proof of Theorem 11), the first inequality from the $L^2 - L^\infty$ norm inequality since $\mathcal{S}\times\mathcal{A}$ is bounded and has finite measure, the second inequality from the second part of Theorem 8, and the last one by definition of $C_p$. Analogous steps show that

$$\|\boldsymbol{\theta}_h\|_2 \leqslant 2\widetilde{N}^{1/2}C_r\sqrt{Vol(\mathcal{S}\times\mathcal{A})}.$$

Having proved that $\max_{h\in[H]}\sup_{s'\in\mathcal{S}}\{\|\boldsymbol{\mu}_h(s')\|_2,\ \|\boldsymbol{\theta}_h\|_2\} \leqslant 2\max\{C_r, C_p\}\sqrt{Vol(\mathcal{S}\times\mathcal{A})}\widetilde{N}^{1/2}$, we have that the regret of LSVI-UCB given by theorem 13 will just be multiplied by the constant $2\max\{C_r, C_p\}\sqrt{Vol(\mathcal{S}\times\mathcal{A})}$.

3. To prove the bound on the total variation difference with the transition function, we just need to apply Equation (15):

$$\begin{aligned}\mathrm{TV}\left(p_h(\cdot|s,a), \boldsymbol{\varphi}_{L,N}^d(s,a)^\top \boldsymbol{\mu}_h(\cdot)\right) &\leqslant Vol(\mathcal{S})\sup_{s'\in\mathcal{S}}|p_h(s'|s,a) - \boldsymbol{\varphi}_{L,N}^d(s,a)^\top \boldsymbol{\mu}_h(\cdot)| \\ &\leqslant Vol(\mathcal{S})\sup_{s'\in\mathcal{S}}\|p_h(s'|\cdot,\cdot) - \boldsymbol{\varphi}_{L,N}^d(\cdot,\cdot)^\top \boldsymbol{\mu}_h(s')\|_{L^\infty} \\ &\leqslant 2Vol(\mathcal{S})J_{d,\nu}C_p N^{-\nu-1}.\end{aligned}$$

As usual, the first inequality comes from the $L^1 - L^\infty$ norm inequality for finite measure spaces.

4. Lastly, the condition on the difference in norm of reward function and its approximation is proved analogously to the previous point. Indeed, by Equation 16,

$$\|r_h - \boldsymbol{\varphi}_{L,N}^d(s,a)^\top \theta_h\|_{L^\infty} \leqslant 2J_{d,\nu}C_r N^{-\nu-1}.$$

All assumptions have been satisfied. Therefore, LSVI-UCB enjoys the regret bound provided by theorem 13, except for the constant factor mentioned in the second point. □

As before, the statement of Theorem 6.3 follows trivially from the last result. Indeed, we can just replace $N = \lceil K^{\frac{1}{d+2(\nu+1)}}\rceil$ in

$$R_K \leqslant \widetilde{\mathcal{O}}\left(H^{3/2}\widetilde{N}^{3/2}\sqrt{K} + H^{3/2}N^{-\nu-1}\widetilde{N}K\right),$$

to get

$$R_K \leqslant \widetilde{\mathcal{O}}\left(H^{3/2}K^{\frac{2d+\nu+1}{d+2(\nu+1)}}\right).$$

### B.8. Proofs from Section 6.5

The algorithms described in Section 6.5 assume that a function class $\mathcal{F}$ (composed of possible feature mappings or Hypotheses on $Q^*$) is known. All the regret bounds presented there depend on the covering number of $\mathcal{F}$ in $L^\infty$ norm (denoted $\mathcal{N}_\infty$), the Eluder dimension (Russo & Van Roy, 2013) of $\mathcal{F}$ (denoted $\dim_E$), or both. In order to compare our regret bounds with the ones there obtained, we have to assume that $\mathcal{F} = \mathcal{F}(B) = \{f \in \mathcal{C}^{\nu,1}(\mathcal{S}\times\mathcal{A}) : \|f\|_{\mathcal{C}^{\nu,1}} \leqslant B\}$. In this way, the algorithms have access to an upper bound $B$ on "how smooth the MDP is", which is a small advantage over the standard setting. We present here a two results bounding the Eluder dimension and the covering number of this function class.

The first result, about the Eluder dimension, is of independent interest, as it shows that the regret bounds for Thompson Sampling for continuous spaces, which were proved by Grant & Leslie (2020) only for $d = 1$, can be extended to arbitrary dimension. Before coming to the actual theorem about the Eluder dimension, we state a simple result which slightly generalizes Proposition 6 from (Russo & Van Roy, 2013).

**Lemma 2.** *Define the $(\varepsilon_{left}, \varepsilon_{right})-$Eluder dimension of a function class $\mathcal{F}$ as the maximum $n \in \mathbb{N}$ such that there is a sequence $\{x_i\}_{i=1}^n \subset \mathcal{X}$ such that*

$$\forall n_0 \leqslant n, \exists f_1, f_2 \in \mathcal{F},\ g = f_1 - f_2 : \qquad \sum_{i=1}^{n_0-1} g(x_i)^2 \leqslant \varepsilon_{left}^2,\ g(x_{n_0}) > \varepsilon_{right}.$$

*Note that, for $\varepsilon_{left} = \varepsilon_{right}$, this dimension corresponds to the standard Eluder. Then, if $\mathcal{F}$ is a linear class of dimension $N$ (in the linear sense), we have that its $(\varepsilon_{left}, \varepsilon_{right})-$Eluder dimension is bounded by*

$$C_0 N \frac{2\varepsilon_{left}^2 + \varepsilon_{right}^2}{{\varepsilon_{right}}^2}\left[\log\left(\frac{2\varepsilon_{left}^2 + \varepsilon_{right}^2}{{\varepsilon_{right}}^2}\right) + \log(1+\varepsilon_{left}^{-2}) + C_1\right]$$

*for some constant $C$.*

*Proof.* We follow the proof of Proposition 6 from (Russo & Van Roy, 2013). The only change occurs in the last step where, in the fraction $\frac{1+x}{x}$ we set $x = \frac{\varepsilon_{\text{right}}^2}{2\varepsilon_{\text{left}}^2}$ instead of $x = \frac{1}{2}$. □

**Theorem 15.** *Let $\mathcal{F}(B) = \{f \in \mathcal{C}^{\nu,1}(\mathcal{S}\times\mathcal{A}) : \|f\|_{\mathcal{C}^{\nu,1}} \leqslant B\}$. Then, for every $\varepsilon > 0$, $\dim_E(\mathcal{F}, \varepsilon) = \widetilde{\mathcal{O}}(B^{\frac{d}{\nu+1}}\varepsilon^{-d/(\nu+1)})$.*

*Proof.* Let $\mathcal{G} = \mathcal{F}(B) - \mathcal{F}(B) \subset \mathcal{F}(2B)$. By definition, to prove that $\mathcal{F}$ has an $\varepsilon-$Eluder dimension bounded by $n$ corresponds to proving that there are no points $x_1, \dots x_n \in [-1,1]^d$ such that

$$\forall n_1 \leqslant n\ \exists g \in \mathcal{G} \qquad \sum_{i=1}^{n_1-1} g(x_i)^2 \leqslant \varepsilon^2, \qquad g(x_{D_1}) \geqslant \varepsilon.$$

We can reason as follows. Let us divide $[-1,1]^d$ into disjoint hypercubes of side $1/\ell$, for $\ell \in \mathbb{N}$. This can be done with exactly $\ell^d$ hypercubes, that we are going to call $\{C_j\}_{j=1}^{\ell^d}$. We start by proving that, under each of the hypercubes, each function of $\mathcal{G}$ is almost linear in some features.

To prove it, note that letting $g \in \mathcal{G}$, Definition 1 in (Liu et al., 2021) (originally from Tsybakov, 2008) ensures that for any fixed $y \in [-1,1]^d$,

$$\forall x \in [-1,1]^d, \qquad |g(x) - T_y(x)| \leqslant 2B\|x-y\|_\infty^{\nu+1}, \qquad T_y[g](x) = \sum_{|\boldsymbol{\alpha}|\leqslant\nu} \frac{D^{\boldsymbol{\alpha}} g(y)}{\boldsymbol{\alpha}!}(x-y)^{\boldsymbol{\alpha}},$$

where the term $T_y(\cdot)$ corresponds to the Taylor polynomial of order $\nu$ centered in $y$. If we apply this to $y = y_{j^\star}$, the middle points of $C_{j^\star}$, we have that, for every $x \in C_{j^\star}$

$$\begin{aligned}|g(x) - T_{y_{j^\star}}[g](x)| &\leqslant 2B\|x - y_{j^\star}\|_\infty^{\nu+1}\\ &\leqslant 2B(2\ell)^{-\nu-1},\end{aligned}$$

where the last inequality comes from the fact that any point of $C_{j^\star}$ cannot have an $\ell_\infty$ distance more than $1/(2\ell)$ form the center of the hypercube. Being valid for every $x \in C_{j^\star}$, this result entails that

$$\|g(\cdot) - T_{y_{j^\star}}[g](\cdot)\|_{L^\infty(C_{j^\star})} \leqslant 2B(2\ell)^{-\nu-1}. \tag{17}$$

At this point, assuming that there are $n_{\text{ind}}$ $\varepsilon-$independent points in $C_{j^\star}$ corresponds to assume that there is a sub-sequence $\{x_{i_k}\}_{k=1}^{n_{\text{ind}}} \subset \{x_i\}_{i=1}^{n} \cap C_{j^\star}$ such that

$$\forall k_0 \leqslant n_{\text{ind}} \qquad \exists g \in \mathcal{G} : \sum_{k=1}^{k_0-1} g(x_{i_k})^2 \leqslant \varepsilon^2, \qquad g(x_{k_0}) \geqslant \varepsilon. \tag{18}$$

Still, applying Equation (17), we have that, whichever the choice of $\{x_{i_k}\}_{k=1}^{n_{\text{ind}}}$, denoting with $\|g\|_{p,x}$ the norm $p-$of $g$ under the set $\{x_{i_k}\}_{k=1}^{n_{\text{ind}}}$,

$$\begin{aligned}\sqrt{\sum_{k=1}^{k_0-1} g(x_{i_k})^2} - \sqrt{\sum_{k=1}^{k_0-1} T_{y_{j^\star}}[g](x_{i_k})^2} &= \|g(\cdot)\|_{2,x} - \|T_{y_{j^\star}}[g](\cdot)\|_{2,x} \\ &\leqslant \|g(\cdot) - T_{y_{j^\star}}[g](\cdot)\|_{2,x} \\ &\leqslant \sqrt{n_{\text{ind}}}\|g(\cdot) - T_{y_{j^\star}}[g](\cdot)\|_{\infty,x} \\ &\leqslant 2B\sqrt{n_{\text{ind}}}(2\ell)^{-\nu-1},\end{aligned}$$

where the first step is by the definition of 2$-$norm, the second from the triangular inequality, the third from bounding the 2-norm with the $\infty-$norm, and the last one from Equation (17). From this follows that any choice $\{x_{i_k}\}_{k=1}^{n_{\text{ind}}}$ satisfying Equation (18) must also satisfy

$$\forall k_0 \leqslant n_{\text{ind}} \qquad \exists g \in \mathcal{G} : \sum_{k=1}^{k_0-1} T_{y_{j^\star}}[g](x_{i_k})^2 \leqslant (\underbrace{\varepsilon + 2B\sqrt{n_{\text{ind}}}(2\ell)^{-\nu-1}}_{\varepsilon_{\text{left}}})^2, \qquad T_{y_{j^\star}}[g](x_{k_0}) \geqslant \underbrace{\varepsilon - 2B(2\ell)^{-\nu-1}}_{\varepsilon_{\text{right}}}. \tag{19}$$

Equation (19) corresponds to the $(\varepsilon_{\text{left}}, \varepsilon_{\text{right}})-$Eluder dimension (defined in lemma 2) of the function class $T_{y_{j^\star}}[g](\cdot) : g \in \mathcal{G}$, with $\varepsilon_{\text{left}} = \varepsilon + 2B\sqrt{n_{\text{ind}}}(2\ell)^{-\nu-1}$ and $\varepsilon_{\text{right}} = \varepsilon - 2B(2\ell)^{-\nu-1}$. For the rest, note that, by definition, $\{T_{y_{j^\star}}[g](\cdot) : g \in \mathcal{G}\}$ is a subset of a vector space spanned by $\{(\cdot - y)^{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}}$ for every multi-index $|\boldsymbol{\alpha}| \leqslant \nu$. The dimension (in the linear sense) of this vector space corresponds to $\binom{\nu+d}{d}$.

Therefore, Lemma 2 (a slight modification of Proposition 6 from (Russo & Van Roy, 2013)) ensures that Equation (19) is only possible when

$$\begin{aligned}n_{\text{ind}} &\leqslant C_0 N \frac{2\varepsilon_{\text{left}}^2 + \varepsilon_{\text{right}}^2}{\varepsilon_{\text{right}}{}^2}\left[\log\left(\frac{2\varepsilon_{\text{left}}^2 + \varepsilon_{\text{right}}^2}{\varepsilon_{\text{right}}{}^2}\right) + \log(1+\varepsilon_{\text{left}}^{-2}) + C_1\right] \\ &\leqslant C_0 N f_{\log}\left(\frac{2\varepsilon_{\text{left}}^2 + \varepsilon_{\text{right}}^2}{\varepsilon_{\text{right}}{}^2} + \log(1+\varepsilon_{\text{left}}^{-2}) + C_1\right)\end{aligned}$$

Where, for readability, we have called $f_{log}(\cdot) = \cdot \log(\cdot)$ and $C_2 = C\binom{\nu+d}{d}$. Letting $\rho := 2B(2\ell)^{-\nu-1}$, the last quantity is bounded by

$$n_{\text{ind}} \leqslant C_2 N f_{\log}\left(\frac{3(\varepsilon + \sqrt{n_{\text{ind}}}\rho)^2}{(\varepsilon-\rho)^2} + \log(1+\varepsilon^{-2}) + C_1\right).$$

If we take $\rho = \varepsilon / \max\{2, K\}$, for a constant $K$ to be decided later, we get that

$$n_{\text{ind}} \leqslant C_2 f_{\log}\left(\frac{3(\varepsilon + \sqrt{n_{\text{ind}}}\rho)^2}{(\varepsilon-\rho)^2} + \log(1+\varepsilon^{-2}) + C_1\right)$$

$$\begin{aligned}
&\leqslant f_{\log}\left(C_2\frac{3\varepsilon^2(1+\sqrt{n_{\text{ind}}}/K)^2}{(\varepsilon/2)^2}+\log(1+\varepsilon^{-2})+C_1\right)\\
&\leqslant f_{\log}\left(12C_2(1+\sqrt{n_{\text{ind}}}/K)^2+\log(1+\varepsilon^{-2})+C_1\right).
\end{aligned}$$

If we take $K=\lceil\sqrt{1+f_{\log}\left(48C_2+\log(1+\varepsilon^{-2})+C_1\right)}\rceil$, we can see that the previous inequality does not hold for $n_{\text{ind}}=K^2$. Indeed,

$$\begin{aligned}
n_{\text{ind}}&\leqslant f_{\log}\left(12C_2(1+\sqrt{n_{\text{ind}}}/K)^2+\log(1+\varepsilon^{-2})+C_1\right)\\
&\leqslant f_{\log}\left(12C_2(1+1)^2+\log(1+\varepsilon^{-2})+C_1\right)\\
&=f_{\log}\left(48C_2\log(1+\varepsilon^{-2})+C_1\right)\\
&<K^2=n_{\text{ind}}.
\end{aligned}$$

Therefore, this tells us that, for $\rho=\varepsilon/\max\{2,K\}$ and $K=\lceil\sqrt{1+f_{\log}\left(48C_2+\log(1+\varepsilon^{-2})+C_1\right)}\rceil$ the number of independent points cannot be (exactly) $K^2$. This entails, by definition of independence, that $n_{\text{ind}}$ also cannot be any number higher than $K^2$, as a longer sequence would entail that the first $K^2$ points are also independent.

With the previous reasoning, we have shown that, for $\rho=\varepsilon/\max\{2,K\}$, each set $C_j$ cannot contain more than $K^2$ $\varepsilon-$independent points. Therefore, the same holds for any subset of $C_j$, corresponding to $\rho\leqslant\varepsilon/\max\{2,K\}$, which translates in the following bound on $\ell$

$$\varepsilon/\max\{2,K\}\geqslant 2B(2\ell)^{-\nu-1}\implies \ell\geqslant\frac{1}{2}\left(\frac{2B\max\{2,K\}}{\varepsilon}\right)^{\frac{1}{\nu+1}}.$$

We can just take

$$\ell=\left\lceil\frac{1}{2}\left(\frac{2B\max\{2,K\}}{\varepsilon}\right)^{\frac{1}{\nu+1}}\right\rceil,$$

which implies a total number of hypercubes given by $\ell^d$:

$$\ell^d=\left\lceil\frac{1}{2}\left(\frac{2B\max\{2,K\}}{\varepsilon}\right)^{\frac{1}{\nu+1}}\right\rceil^d\leqslant\left(\frac{2B\max\{2,K\}}{\varepsilon}\right)^{\frac{d}{\nu+1}}.$$

In the end, we have proved that, dividing the space $[-1,1]^d$ into this number of hypercubes, no hypercube can contain more than $n_{\text{ind}}$ points that form a $\varepsilon-$independent sequence. Thus, the full length $n$ of $\{x_i\}_{i=1}^n$ is bounded by

$$n\leqslant n_{\text{ind}}\left(\frac{2B\max\{2,K\}}{\varepsilon}\right)^{\frac{d}{\nu+1}}\leqslant K^2\left(\frac{2B\max\{2,K\}}{\varepsilon}\right)^{\frac{d}{\nu+1}},$$

with $C_2=C\binom{\nu+d}{d}$ and $K=\lceil\sqrt{1+f_{\log}\left(48C_2+\log(1+\varepsilon^{-2})+C_1\right)}\rceil$. As all terms $B,C,\nu$ are constants not depending on $\varepsilon$, while $K$ depends on it logathimically, the proof is complete.

□

As side effect of this theorem, we generalize to the multidimensional case Theorem 3 from (Grant & Leslie, 2020). Thus, Thompson Sampling (as described in their section 1.2) can now be shown to have a regret guarantee in dimension higher than one.

**Theorem 16.** *Let $\mathcal{F}(B)=\{f\in\mathcal{C}^{\nu,1}(\mathcal{S}\times\mathcal{A}):\|f\|_{\mathcal{C}^{\nu,1}}\leqslant B\}$. Then, for every $\varepsilon>0$, $\log(\mathcal{N}_\infty(\mathcal{F},\varepsilon))=\mathcal{O}(B^{\frac{d}{\nu+1}}\varepsilon^{-d/(\nu+1)})$.*

*Proof.* Example 4 of (Russo & Van Roy, 2013) shows that, if $\mathcal{G}$ is a vector space of dimension $\widetilde{N}$,

$$\log(\mathcal{N}_\infty(\mathcal{G}, \varepsilon)) = \mathcal{O}(\widetilde{N} \log(\varepsilon^{-1})).$$

Unfortunately, the dimension of $\mathcal{F}(B)$, viewed as a subset of a vector space, is $+\infty$. Nonetheless, we can rely on our Theorem 8 to achieve a non-vacuous bound. The latter ensures that, for any $f \in \mathcal{F}(B)$,

$$\exists p_N \in \mathcal{P}_N : \ \|f - p_N\|_{L^\infty} \leqslant 2 J_{d,\nu} B N^{-\nu-1},$$

where $\mathcal{P}_N$ is the space of multivariate polynomials of degree at most $N$. To ensure $\varepsilon/2 = \|f - p_N\|_{L^\infty}$, we need

$$N \geqslant \left(\frac{4 J_{d,\nu} B}{\varepsilon}\right)^{1/(\nu+1)}.$$

Under this condition, we every $\varepsilon/2-$cover for $\mathcal{P}_N$ corresponds to a $\varepsilon-$cover for $\mathcal{F}$, so that

$$\dim_E(\mathcal{F}, \varepsilon) \leqslant \dim_E(\mathcal{P}_N, \varepsilon/2).$$

Therefore, as the space $\mathcal{P}_N$ is in fact a vector space of dimension $\tilde{N} = \binom{N+d}{N} \approx N^d$, we have

$$\log(\mathcal{N}_\infty(\mathcal{F}, \varepsilon)) \leqslant \log(\mathcal{N}_\infty(\mathcal{P}_N, \varepsilon/2)) = \mathcal{O}(N^d \log(\varepsilon^{-1})) = \widetilde{\mathcal{O}}(B^{\frac{d}{\nu+1}} \varepsilon^{-d/(\nu+1)}).$$

□

With this theorem, we can show that some very recent algorithms achieve a nontrivial regret bound for our setting.

**Proposition 17** (Restatement of Theorem 3)**.** *Let us assume we run Algorithm 1 from (Ren et al., 2022) on a Strongly Smooth MDP such that the transition dynamics is Gaussian, in the sense that at any time step $h$ we have $s_{h+1} = f(s_h, a_h) + \epsilon_h$, where $\epsilon_h \sim \mathcal{N}(0, \Sigma)$. Then, provided that the algorithm knows an upper bound $B$ on $\|f\|_{\mathcal{C}^{\nu,1}}$, and $d < \nu + 1$, its regret is bounded, with probability at least $1 - \delta$, by*

$$R_K \leqslant \widetilde{\mathcal{O}}\left(B^{\frac{d}{\nu+1}} H^{\frac{3\nu+d+3}{2\nu+2}} K^{\frac{\nu+d+1}{2\nu+2}}\right),$$

*where the $\widetilde{\mathcal{O}}$ hides a quantity that is logarithmic in $K, H, B, \delta$.*

*Proof.* Under our assumptions we can apply Theorem 5 from (Ren et al., 2022) for $\mathcal{F} = \{f \in \mathcal{C}^{\nu,1}(\mathcal{S} \times \mathcal{A}) : \|f\|_{\mathcal{C}^{\nu,1}} \leqslant B\}$. This gives

$$R_K \leqslant \widetilde{\mathcal{O}}\left(H^{3/2}\sqrt{K \dim_E(\mathcal{F}, (HK)^{-1/2}) \log(\mathcal{N}_2(\mathcal{F}, (HK)^{-1/2}))}\right).$$

Then, by our Theorems 15 and 16,

$$R_K \leqslant \widetilde{\mathcal{O}}\left(B^{\frac{d}{\nu+1}} H^{3/2} \sqrt{K^{1+\frac{d}{\nu+1}} H^{\frac{d}{\nu+1}}}\right) = \widetilde{\mathcal{O}}\left(B^{\frac{d}{\nu+1}} H^{\frac{3\nu+d+3}{2\nu+2}} K^{\frac{\nu+d+1}{2\nu+2}}\right).$$

□

**Proposition 18** (Restatement of Theorem 4)**.** *Let us assume we run algorithm* GOLF *on a Strongly Smooth MDP. Then, provided that the algorithm knows the constant $C_{\mathcal{T}*}$ and $d \leqslant \frac{2}{3}\nu + \frac{2}{3}$, its regret is bounded, with probability at least $1 - \delta$, by*

$$R_K \leqslant \widetilde{\mathcal{O}}\left(C_{\mathcal{T}*}^{\frac{dH}{\nu+1}} K^{\frac{2\nu+3d+2}{4\nu+4}}\right),$$

*where the $\widetilde{\mathcal{O}}$ hides a quantity that is logarithmic in $K, C_{\mathcal{T}*}, \delta$.*

*Proof.* We briefly recall the assumptions of the regret bound for GOLF given by (Jin et al., 2021)

1. Assumption 1: the algorithm knows a set $\mathcal{F} = \{\mathcal{F}_1, \dots \mathcal{F}_H\}$ containing one function class for each time step such that $\forall h, Q^*_h \in \mathcal{F}_h$.
2. Assumption 2: The class $\mathcal{F}$ is closed under Bellmann optimality operator: $\forall f \in \mathcal{F}_{h+1},\ \mathcal{T}^* f \in \mathcal{F}_h$.

Knowing $C_{\mathcal{T}*}$, it is possible to define a function class $\mathcal{F}$ containing all possible candidates for the $Q^*$ function of the MDP. In fact, by the definition of Bellman optimality operator, we can see that, for every $h$,

$$\|Q^*_h\|_{\mathcal{C}^{\nu,1}} = \|\mathcal{T}^* Q^*_{h+1}\|_{\mathcal{C}^{\nu,1}} \leqslant C_{\mathcal{T}*}(1 + \|Q^*_{h+1}\|_{\mathcal{C}^{\nu,1}}) \leqslant \frac{C_{\mathcal{T}*}^{H-h+1} - 1}{C_{\mathcal{T}*} - 1}.$$

Therefore, using the function classes $\mathcal{F}_h = \{f \in \mathcal{C}^{\nu,1}(\mathcal{S} \times \mathcal{A}) : \|f\|_{\mathcal{C}^{\nu,1}} \leqslant \frac{C_{\mathcal{T}*}^{H-h+1}-1}{C_{\mathcal{T}*}-1}\}$ allows to satisfy both assumptions 1 and 2 from (Jin et al., 2021).

Theorem 15 from the same paper ensures that

$$R_K \leqslant \widetilde{\mathcal{O}}\left(H\sqrt{K \dim_{BE}(\mathcal{F}, K^{-1/2}) \log(\mathcal{N}_\infty(\mathcal{F}, K^{-1}))}\right),$$

a result which is again bounded by Proposition 12 (again from the same paper)

$$R_K \leqslant \widetilde{\mathcal{O}}\left(H\sqrt{K \max_{h=1,\dots H} \dim_E(\mathcal{F}_h, K^{-1/2}) \log(\mathcal{N}_\infty(\mathcal{F}, K^{-1}))}\right).$$

Then, by our Theorems 15 and 16, we can bound $\max_{h=1,\dots H} \dim_E(\mathcal{F}_h, K^{-1/2}) \leqslant \dim_E(\mathcal{F}_H, K^{-1/2}) \leqslant \widetilde{\mathcal{O}}(B^{\frac{d}{\nu+1}} K^{d/(2\nu+2)})$ and $\log(\mathcal{N}_\infty(\mathcal{F}, K^{-1})) \leqslant \widetilde{\mathcal{O}}(HB^{\frac{d}{\nu+1}} K^{d/(\nu+1)})$ where $B = \frac{C_{\mathcal{T}*}^{H-h+1}-1}{C_{\mathcal{T}*}-1}$. Substituting this result, we get

$$R_K \leqslant \widetilde{\mathcal{O}}\left(C_{\mathcal{T}*}^{\frac{dH}{\nu+1}} H\sqrt{K^{1+\frac{3d/2}{\nu+1}}}\right) = \widetilde{\mathcal{O}}\left(C_{\mathcal{T}*}^{\frac{dH}{\nu+1}} K^{\frac{2\nu+3d+2}{4\nu+4}}\right),$$

which ends the proof. □

## C. Details of the Numerical Simulation

In section 4.3, we have performed a numerical simulation on a modified version of the Linear Quadratic Regulator (LQR). Both environments took the form

$$s_{h+1} = g(As_h + Ba_h + \xi_h),$$
$$r_h = -s_h^\top Q s_h - a_h^\top R a_h,$$

where $g(x) := \frac{x}{1+\|x\|_2}$. Moreover, in both cases the dimension of the state space corresponds to 2, and the one of the action space to 1. Also, we have in both cases

$$B = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad R = [0.2].$$

what changes is the matrix $A$, which determines most of the dynamics of the system. For this matrix, we have

$$\text{Left experiment: } A = \begin{bmatrix} 0.7 & 0.7 \\ -0.7 & 0.7 \end{bmatrix} \qquad \text{Right experiment: } A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

### C.1. Practical details

Finally, we report some the details on how the computation was performed in the paper. These are important to ensure the truthfulness of the results and the claims based on empirical validation.

**Training Details.** The algorithms were implemented in PYTHON3.7. Each experiment was executed using five random seeds (corresponding to the first five natural numbers), and the computations were distributed across five parallel processes using the JOBLIB library. The total computational time for the first experiment was of 189935 seconds, more than two days and four hours.

**Compute.** We used a server with the following specifications:

- CPU: 88 INTEL(R) XEON(R) CPU E7-8880 V4 @ 2.20GHZ CPUS,
- RAM: 94,0 GB.

As mentioned, we parallelized the computing for the five different random seeds, therefore only five of the 88 cores were actually used.