Improved Regret Bounds for Undiscounted Continuous Reinforcement Learning

K.Lakshmanan LKSHMNAN.K@GMAIL.COM
Montanuniversität Leoben, Franz-Josef-Strasse 18, 8700 Leoben, AUSTRIA

Ronald Ortner RORTNER@UNILEOBEN.AC.AT
Montanuniversität Leoben, Franz-Josef-Strasse 18, 8700 Leoben, AUSTRIA

Daniil Ryabko DANIIL@RYABKO.NET
INRIA Lille - Nord Europe, 40 Avenue Halley, 59650 Villeneuve d'Ascq, FRANCE

Abstract

We consider the problem of undiscounted reinforcement learning in continuous state space. Regret bounds in this setting usually hold under various assumptions on the structure of the reward and transition function. Under the assumption that the rewards and transition probabilities are Lipschitz, for 1-dimensional state space a regret bound of $\tilde{O}(T^{\frac{3}{4}})$ after any T steps has been given by Ortner and Ryabko (2012). Here we improve upon this result by using non-parametric kernel density estimation for estimating the transition probability distributions, and obtain regret bounds that depend on the smoothness of the transition probability distributions. In particular, under the assumption that the transition probability functions are smoothly differentiable, the regret bound is shown to be $\tilde{O}(T^{\frac{2}{3}})$ asymptotically for reinforcement learning in 1-dimensional state space. Finally, we also derive improved regret bounds for higher dimensional state space.

1. Introduction

Reinforcement learning (RL) in continuous domains is still a big challenge, from the practical as well as from the theoretical point of view. The setting theoretically best understood is the continuous multi-armed bandit problem. Under Hölder conditions on the reward function, regret bounds have been given by Kleinberg (2005), Auer et al. (2007), Kleinberg et al. (2008), and Bubeck et al. (2010).

In more general RL settings, often strong assumptions on the transition structure are made. Thus, there are theoretical results for RL with deterministic transitions in the discounted setting (Bernstein and Shimkin, 2010), as well as for RL with transition functions that are linear in state and action (Strehl and Littman, 2008; Brunskill et al., 2009; Abbasi-Yadkori and Szepesvári, 2011; Ibrahmi et al., 2012). More generally, the work of Kakade et al. (2003) considers PAC-learning for continuous RL in metric state spaces. Recently, Osband and Van Roy (2014) have derived bounds on the *expected* regret under the assumption that the reward and the transition probability function belong to a given class of functions. The bounds then depend on particular parameters of these function classes, called the *Kolmogorov dimension* and the *eluder dimension*. Unlike that, here we try to assume the most general setting making only smoothness assumptions on rewards and transition probabilities.

Our research is based on the work of Ortner and Ryabko (2012), which has given the most general regret bounds in a continuous state RL setting so far. Under the assumption that reward and transition functions are Hölder continuous, sublinear regret bounds depending on the Hölder parameters have been shown. The suggested algorithm discretizes the state space and employs the UCRL algorithm of Jaksch et al. (2010) on the discretized MDP. We improve upon this algorithm and the respective regret bound by using kernel density estimation instead of histograms for estimating the probability density functions. Kernel-based methods have been employed in RL before, starting with (Ormoneit and Sen, 2002). Here we provide the first regret bounds for a kernel-based algorithm for RL in continuous state space. In order to derive our regret bounds we need concentration bounds for the employed kernel density estimator. Such bounds can be found e.g. in (Devroye, 1987). However, for our particular case, we extend results of

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s).

Ibragimov and Hasminskii (1981) and Vogel and Schettler (2013) to the case where the samples are assumed to be only independent but not necessarily i.i.d.

The regret bounds we obtain improve over known bounds for the UCCRL algorithm (Ortner and Ryabko, 2012), provided that the transition probability functions are sufficiently smooth. While the UCCRL algorithm gives $\tilde{O}(T^{\frac{2+\alpha}{2+2\alpha}})$ regret for MDPs with 1-dimensional state space and Hölder-continuous rewards and transition probabilities with parameter α, the proposed UCCRL-KD algorithm has regret of order $\tilde{O}(T^{\frac{\beta+\alpha\beta+2\alpha}{\beta+2\alpha\beta+2\alpha}})$, where the transition function is assumed to be κ-times smoothly differentiable and $\beta := \kappa+\alpha$. Thus, we obtain improved bounds if $\alpha < \kappa$. For the simple case of Lipschitz continuous densities, i.e. $\alpha = 1$, the regret is $\tilde{O}(T^{\frac{3}{4}})$ for UCCRL, while for UCCRL-KD it asymptotically approaches $\tilde{O}(T^{\frac{2}{3}})$, provided that the transition probability functions are infinitely often smoothly differentiable. For general d-dimensional state space we show that the regret for UCCRL-KD is $\tilde{O}(T^{\frac{1+d\alpha+\alpha}{1+d\alpha+2\alpha}})$, improving over the bound of $\tilde{O}(T^{\frac{2d+\alpha}{2d+2\alpha}})$ for UCCRL.

2. Setting

For the sake of simplicity, we concentrate on the 1-dimensional case. Details for the general d-dimensional setting are given in Section 4.1 below. Thus, consider a Markov decision process (MDP) with state space $[0,1]$ and finite action space of size A. We assume that the random reward in any state s under any action a is bounded in $[0,1]$ with mean $r(s,a)$. The transition probability distribution at state s under action a is denoted by $p(\cdot|s,a)$. We make the following assumptions on the reward and transition probability functions.

Assumption 1. *There are L, $\alpha > 0$ such that for any two states s, s' and all actions a,*

$$\big|r(s,a) - r(s',a)\big| \leq L|s-s'|^{\alpha}.$$

Assumption 2. *There are L, $\alpha > 0$ such that for any two states s, s' and all actions a,*

$$\big\|p(\cdot|s,a) - p(\cdot|s',a)\big\|_1 \leq L|s-s'|^{\alpha}.$$

These two assumptions are the same as in (Ortner and Ryabko, 2012). They guarantee that rewards and transition probabilities are close in close states, but do not make any assumption on the shape of the transition probability densities. Here, we additionally assume that the transition functions are smooth, which allows us to obtain improved regret bounds.

Assumption 3. *The transition functions $p(\cdot|s,a)$ are κ-times smoothly differentiable for all states s and all actions a. That is, there are L, $\alpha > 0$ such that for any state s and all actions a,*

$$\big|p^{(\kappa)}(s'|s,a) - p^{(\kappa)}(s''|s,a)\big| \leq L|s'-s''|^{\alpha}.$$

For the sake of simplicity, in the following we assume that L and $\alpha \leq 1$ in Assumptions 1–3 are the same. Note that for $\alpha > 1$ the transition functions would be constant and learning hence trivial.

We assume (for the following assumptions and technical details see Section 2 of Ortner and Ryabko, 2012) the existence of an optimal policy π^* with optimal average reward ρ^* independent of the initial state. Further, we assume that for each measurable policy π the Poisson equation[1]

$$\rho_\pi + \lambda(\pi,s) = r(s,\pi(s)) + \int p(ds'|s,\pi(s))\,\lambda(\pi,s')$$

holds, where ρ_π is the average reward of π and $\lambda(\pi,s)$ is the *bias* of policy π in state s. Note that for any policy π the Poisson equation is satisfied under modest assumptions such as geometric convergence to an invariant probability measure μ_π, cf. Chapter 10 of (Hernández-Lerma and Lasserre, 1999).

We recall from (Ortner and Ryabko, 2012) that under Assumptions 1 and 2 the bias of the optimal policy is bounded. The performance of an algorithm is measured by the *regret* it receives after T time steps, defined as

$$\Delta_T = T\rho^* - \sum_{t=1}^{T} r_t,$$

where r_t is the (random) reward obtained by the algorithm at time step t. Note that (cf. Chapter 10 of Hernández-Lerma and Lasserre, 1999) no policy can obtain higher accumulated reward than $T\rho^* + H$ after any T steps, where

$$H := \sup_s \lambda(\pi^*,s) - \inf_s \lambda(\pi^*,s)$$

is the *bias span* of the optimal policy.

3. Algorithm

As already indicated, our algorithm is based on the UCCRL algorithm of Ortner and Ryabko (2012). In the UCCRL algorithm, in the first step the state space $[0,1]$ is discretized into n intervals I_j of equal size. Thus, the estimates for rewards and transition probabilities are aggregated correspondingly, that is, states contained in the same interval I_j are clubbed together and considered as coming from a single (discrete) state. This gives a discrete-state MDP, to

[1] In the following, we usually skip the range of integration when it is clear from context.

which the UCRL algorithm of (Jaksch et al., 2010) can be applied. The algorithm UCCRL-KD that we propose uses the same aggregation technique for the rewards. However, concerning the estimated transition probability functions, we only aggregate inasmuch as states contained in the same interval I_j will obtain the same estimated transition function and that for computing this estimate we use all samples of states in the same interval. The estimation of this function will be done by a kernel density estimate, that is, without using any kind of discretization.

For the sake of completeness, our UCCRL-KD algorithm is depicted as Algorithm 1. Proceeding in episodes $k = 1, 2, \ldots$ in which the chosen policy $\tilde{\pi}_k$ remains the same, in each episode k a set of plausible MDPs $\mathcal{M}_k$, determined by confidence intervals for rewards and transition probabilities, is considered (cf. line 7 of the algorithm). From this set the algorithm chooses the (so-called *optimistic*) MDP $\tilde{M}_k$ whose optimal policy $\tilde{\pi}_k$ promises the highest possible average reward $\rho^*(\tilde{M}_k)$ (line 8). This policy is then employed in episode k, which is terminated if some action in some interval I_j has been played as often in the episode as before the episode (line 10), so that recomputation of estimates and policy is justified.

Basically, UCCRL-KD looks the same as UCCRL, only that the estimates and confidence intervals for the transition probabilities are different. For these we do not use a histogram based estimator as for UCCRL, but (results for) a kernel density estimator. The confidence intervals employed in line 7 of the algorithm are given by

$$\mathrm{conf}_r(s,a,n,A,\delta,t) := Ln^{-\alpha} + \sqrt{\frac{7\log(2nAt/\delta)}{2N_t(I(s),a)}}, \quad (1)$$

$$\mathrm{conf}_p(s,a,n,A,\delta,t) := C_0 Ln^{-\alpha} + N_t(I(s),a)^{\frac{-\beta}{2\beta+2}} C_1' \log\Big(\sqrt{14\log\left(\tfrac{2nAt}{\delta}\right)}\Big). \quad (2)$$

Here $N_t(I_j,a)$ is the maximum of 1 and the number of times action a has been played in a state contained in interval I_j at step t. Further, $I(s)$ denotes the interval I_j that contains the state s. The constants C_0 and $C_1' := C_1 L + C_2 + \frac{C_3}{2\pi}$ depend on the employed kernel density estimator, cf. Assumption 4 and Section 5.2.3 below. Finally, $\beta := \kappa + \alpha$ depends on the smoothness of the transition functions. The confidence intervals for the transition probabilities come from the tail bounds that we derive in Section 5.1. In the following, we describe the kernel density estimator in detail.

3.1. Kernel Density Estimation

While the estimates $\hat{r}(s,a)$ for the mean rewards are computed as for UCCRL (that is, one takes the average of the rewards observed in all states in $I(s)$), for the estimates $\hat{p}(\cdot|s,a)$ we use a suitable kernel density estimator.

Algorithm 1 UCCRL-Kernel Density Algorithm

```
1: Input: State space $[0,1]$, number of actions $A$, confidence parameter $\delta$, discretization parameter $n \in \mathbb{N}$, upper bound $H$ on the bias span, Lipschitz parameters $L$, $\alpha$, smoothness parameter $\kappa$.
2: Initialization: Let $I_1 := \big[0, \frac{1}{n}\big]$, $I_j := \big(\frac{j-1}{n}, \frac{j}{n}\big]$ for $j = 2, 3, \ldots, n$. Set $t := 1$, and observe the initial state $s_1$.
3: for $k = 1, 2, \ldots$ do
4:     Let $N_k(I_j, a)$ be the maximum of 1 and the number of times action $a$ has been chosen in a state $\in I_j$ prior to episode $k$. Further, let $v_k(I_j, a)$ be the respective counts in episode $k$.
       Initialize episode k:
5:     Set the start time of episode $k$, $t_k := t$.
6:     Compute the estimates $\hat{r}_k(s,a)$ for rewards and the kernel density estimates $\hat{p}_k(\cdot|s,a)$ for transition probabilities (cf. Section 3.1) using all samples from states in the same interval $I_j$ as $s$.
       Compute policy $\tilde{\pi}_k$ :
7:     Let $\mathcal{M}_k$ be the set of plausible MDPs $\tilde{M}$ with $H(\tilde{M}) \leq H$ and rewards $\tilde{r}(s,a)$ and transition probabilities $\tilde{p}(\cdot|s,a)$ satisfying (cf. (1) and (2))
           $|\tilde{r}(s,a) - \hat{r}_k(s,a)| \leq \mathrm{conf}_r(s,a,n,A,\delta,t_k)$,
           $\big\|\tilde{p}(\cdot|s,a) - \hat{p}_k(\cdot|s,a)\big\|_1 \leq \mathrm{conf}_p(s,a,n,A,\delta,t_k)$.
8:     Choose policy $\tilde{\pi}_k$ and $\tilde{M}_k$ such that
           $\rho_{\tilde{\pi}_k}(\tilde{M}_k) = \arg\max\{\rho^*(M) | M \in \mathcal{M}_k\}$.
9:     Execute policy $\tilde{\pi}_k$:
10:    while $v_k(I(s_t), \tilde{\pi}_k(s_t)) < N_k(I(s_t), \tilde{\pi}_k(s_t))$ do
11:    Choose action $a_t = \tilde{\pi}_k(s_t)$, obtain reward $r_t$, and observe $s_{t+1}$. Set $t := t + 1$.
12: end for
```

In general, given i.i.d. samples $X_1, \ldots, X_N$ from a common density f, the generalized density (or Parzen-Rosenblatt) estimator $\hat{f}_N$ is given by (cf. Section 1.2 of Tsybakov, 2009)

$$\hat{f}_N(x) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x - X_i}{h}\right), \quad (3)$$

where $K : \mathbb{R} \to \mathbb{R}$ is an integrable *kernel* function satisfying

$$\int_{-\infty}^{\infty} K(u)du = 1, \quad (4)$$

and h is the bandwidth of the estimator.

In our case we want to estimate the transition probability distribution in each state s. Since in general it may be the case that we visit a state s not more than once (or not at all), we compute the estimate $\hat{p}(\cdot|s,a)$ using all samples from states that are contained in $I(s)$, the interval I_j containing s. Note that therefore our samples will in general only be independent but not i.i.d. Still, we will show in Theorem 1 below that due to Assumptions 2 and 3 on the probability distributions the density estimator in (3) will give a sufficiently good estimation.

Assumptions on the kernel To guarantee that our confidence intervals for the estimated transition functions $\hat{p}(\cdot|s,a)$ hold with high probability (cf. Theorem 1 below), additionally to (4) we need the following assumptions on the employed kernel function K.

Assumption 4. *Let κ be the smoothness parameter defined in Assumption 3. Then the kernel function K satisfies*

$$\int_{-\infty}^{\infty} x^j K(x)\, dx \;=\; 0 \;\; \textit{for } j = 1, 2, \ldots, \kappa, \tag{5}$$

$$\int_{-\infty}^{\infty} |K(x)|\, dx \;=\; C_0 < \infty, \tag{6}$$

$$\int_{-\infty}^{\infty} |x^\kappa K(x)|\, dx \;=\; C_1 < \infty, \tag{7}$$

$$\sup_{x \in \mathbb{R}} |K(x)| \;=\; C_2 < \infty. \tag{8}$$

Finally, for $k(x) := \int_{-\infty}^{\infty} e^{ix^T y} K(y) dy$, it holds that

$$\int_{-\infty}^{\infty} |k(x)|\, dx = C_3 < \infty. \tag{9}$$

Thus, we have to use a kernel that depends on the smoothness of the transition probabilities. While kernels of infinite order and compact support do not exist (see p.101 of Devroye, 1987), there are ways to generate higher order kernels from lower order kernels (Abdous, 1995), which guarantees that for each κ there is a suitable kernel available. In particular, polynomial kernels of arbitrary finite order and compact support exist (Gasser et al., 1985), which by definition satisfy equation (5). It can be easily verified that polynomial kernels with compact support also satisfy equations (6)–(9).

4. Results

In order to establish our improved bounds on the regret, we need high probability bounds for the new confidence intervals (1) and (2). Note that, for the densities $p(\cdot|s,a)$ we are interested in, the observed transitions in general are from different states in $I(s)$ with close but different densities. Thus, these observations are in general not i.i.d., but only independent. Still, the following tail bound for the respective kernel density estimator $\hat{f}_N$ computed from these independent observations can be established.

Theorem 1. *Let $f := p(\cdot|s,a)$ be a transition probability distribution satisfying Assumptions 2 and 3, and let $\hat{f}_N$ be a kernel density estimate of f for which Assumption 4 holds and which is computed from samples $X_1, \ldots, X_n$ of the transition probability distributions $f_1 := p(\cdot|s_1,a), \ldots, f_N := p(\cdot|s_N,a)$ in states $s_1, \ldots, s_N$ that are contained in the same interval I_j as s. Then for all $N \in \mathbb{N}$ and all $u > 0$*

$$\Pr\Big\{ \sup_x |\hat{f}_N(x) - f(x)| \geq \frac{u}{\sqrt{N}h} + \frac{C_3}{2\pi\sqrt{N}h} + C_0 L n^{-\alpha} + C_1 L h^\beta \Big\} \leq 2e^{\frac{-u^2}{2C_2^2}}.$$

Theorem 1 allows us to derive the following regret bound.

Theorem 2. *Consider an MDP with state space $[0,1]$, A actions, rewards and transition probabilities satisfying Assumptions 1–3, and bias span (upper bounded by) H. Then with probability $1-\delta$, the regret of UCCRL-KD (with input parameters $n \leq T$ and δ) after T steps is upper bounded by*

$$c \cdot C_1' H \sqrt{14A \log\left(\tfrac{2nAT}{\delta}\right)}\, n^{\frac{\beta}{2\beta+2}} T^{\frac{\beta+2}{2\beta+2}} + c' \cdot C_0 H L n^{-\alpha} T, \tag{10}$$

where C_0, $C_1' := C_1 L + C_2 + \frac{C_3}{2\pi}$ are constants depending on Assumption 4, and c, c' are independent constants.

Setting $n = T^{\frac{\beta}{\beta+2\alpha\beta+2\alpha}}$ gives an upper bound of

$$c'' H (C_0 L + C_1') \sqrt{14A \log\left(\tfrac{2AT^2}{\delta}\right)}\, T^{\frac{\beta+\alpha\beta+2\alpha}{\beta+2\alpha\beta+2\alpha}}$$

for an independent constant c''.

Equation (10) gives a bound on the regret that resembles that for UCCRL: The second term corresponds to the discretization error (and is the same as for UCCRL), while the first term corresponds to the error in the discrete MDP (and is improved compared to the respective regret of UCCRL).

Remark 1. *Compared to the regret bound of $\tilde{O}(HL\sqrt{A} \cdot T^{\frac{2+\alpha}{2+2\alpha}})$ for UCCRL (Ortner and Ryabko, 2012), the bound for UCCRL-KD has improved dependence on T for all α as soon as $\kappa > \alpha$. For the Lipschitz case $\alpha = 1$ the bound for UCCRL-KD approaches $\tilde{O}(T^{\frac{2}{3}})$ when $\kappa \to \infty$, while the respective bound for UCCRL is $\tilde{O}(T^{\frac{3}{4}})$.*

Remark 2. *As for the UCCRL algorithm, if the horizon T is unknown then the doubling trick can be used to give the same bound with slightly worse constants.*

4.1. d-dimensional State Space

Under the following additional assumptions UCCRL-KD (with modified confidence intervals) gives also improved regret bounds in MDPs with state space of dimension d.

Assumption 5. *The transition probability functions are in* $C^2(\mathbb{R})$ *and their partial derivatives of order 1 and 2 are bounded by a constant* C_4.

Assumption 6. $\int |x|^2 K(x)\,dx = C_5 < \infty$.

Under Assumptions 5 and 6 one can replace Theorem 1 by the following result.

Theorem 3. *Let* $f := p(\cdot|s,a)$ *be a transition probability distribution satisfying Assumptions 2, 3, and 5, and let* $\hat{f}_N$ *be a kernel density estimate of* f *as in Theorem 1 that additionally satisfies Assumption 6. Then for all* $N \in \mathbb{N}$ *and all* $u > 0$

$$\Pr\Big\{ \sup_x |\hat{f}_N(x) - f(x)| \geq \\ \tfrac{u}{\sqrt{Nh}} + \tfrac{C_3}{2\pi\sqrt{Nh}} + C_0 L n^{-\alpha} + \tfrac{1}{2} C_4 C_5 \cdot h^{\frac{2}{d}} \Big\} \leq 2e^{\frac{-u^2}{2C_2^2}}.$$

Choosing $h = N^{\frac{-d}{2(2+d)}}$, one can use Theorem 3 to obtain confidence intervals that allow us to derive the following regret bound for UCCRL-KD in d-dimensional state space.

Theorem 4. *Consider an MDP with state space* $[0,1]^d$, A *actions, rewards and transition probabilities satisfying Assumptions 1, 2, 3 and 5, and bias span* $\leq H$. *Then with probability* $1-\delta$, *the regret of UCCRL-KD (with modified confidence intervals according to Theorem 3 and input parameters* $n \leq T$ *and* δ*) after* T *steps is upper bounded by*

$$c \cdot C_2' H \sqrt{14 A \log\left(\tfrac{2nAT}{\delta}\right)} n^{\frac{1}{d+2}} T^{\frac{d+1}{d+2}} + c' \cdot C_0 H L n^{-\alpha} T$$

for independent constants c, c' *and* C_0, $C_2' := C_2 + \frac{C_3}{2\pi} + \frac{C_4 C_5}{2}$ *depending on Assumptions 4 and 5.*

Setting $n = T^{\frac{1}{1+d\alpha+2\alpha}}$ *gives a bound of order* $\tilde{O}(T^{\frac{1+d\alpha+\alpha}{1+d\alpha+2\alpha}})$.

Remark 3. *This bound is an improvement over the bound of* $\tilde{O}(T^{\frac{2d+\alpha}{2d+2\alpha}})$ *of Ortner and Ryabko (2012) for all* α *and all dimensions* d *except for the Lipschitz case (*$\alpha = 1$*) in dimension* $d = 1$*, where the two bounds coincide. In particular, also for* $d = 1$ *and* $\alpha < 1$ *the bound of Theorem 4 improves over the 1-dimensional bound for UCCRL. However, when* $\beta > 2$ *Theorem 2 provides a better bound for* $d = 1$ *than Theorem 4.*

5. Proofs

In the following, we give detailed proofs only of Theorems 1 and 2. The proof of Theorem 3 is similar to that of Theorem 1, only that Lemma 1 is replaced by an analogue based on the Lemma on p.7 of Vogel and Schettler (2013). Theorem 4 is then shown as Theorem 2 with Theorem 1 replaced by Theorem 3.

5.1. Proof of Theorem 1

Lemma 1. *Let* f, $\hat{f}_N$, *and* $f_1, \dots, f_N$ *be as given in Theorem 1. Then for all* $x \in [0,1]$

$$\big|\mathbb{E}[\hat{f}_N(x)] - f(x)\big| \leq C_0 L n^{-\alpha} + C_1 L h^{\beta}.$$

Proof. Using that, by Assumption 2,

$$|f_i(y) - f(y)| \leq L n^{-\alpha},$$

and that due to $\int K(u)du = 1$ we have

$$\int K\Big(\frac{x-y}{h}\Big) dy = h \int K(u)\,du = h, \tag{11}$$

we can rewrite

$$\begin{aligned}
&\big|\mathbb{E}[\hat{f}_N(x)] - f(x)\big| \\
&= \Big|\mathbb{E}\Big[\frac{1}{Nh}\sum_{i=1}^N K\Big(\frac{x-X_i}{h}\Big)\Big] - f(x)\Big| \\
&= \Big|\frac{1}{Nh}\sum_{i=1}^N \mathbb{E}\Big[K\Big(\frac{x-X_i}{h}\Big)\Big] - f(x)\Big| \\
&= \Big|\frac{1}{Nh}\sum_{i=1}^N \int K\Big(\frac{x-y}{h}\Big) f_i(y)\,dy - f(x)\Big| \\
&\leq \Big|\frac{1}{Nh}\sum_{i=1}^N \int K\Big(\frac{x-y}{h}\Big)\big(f_i(y) - f(y)\big)\,dy\Big| \\
&\quad + \Big|\frac{1}{Nh}\sum_{i=1}^N \int K\Big(\frac{x-y}{h}\Big) f(y)\,dy - f(x)\Big| \\
&\leq \frac{Ln^{-\alpha}}{h} \int \Big|K\Big(\frac{x-y}{h}\Big)\Big| dy \\
&\quad + \Big|\frac{1}{h}\int K\Big(\frac{x-y}{h}\Big)\big(f(y) - f(x)\big)\,dy\Big|.
\end{aligned} \tag{12}$$

By (6) and an analogue of (11) we can bound the first term of (12) as

$$\frac{Ln^{-\alpha}}{h}\int \Big|K\Big(\frac{x-y}{h}\Big)\Big| dy \;\leq\; C_0 L n^{-\alpha}. \tag{13}$$

Concerning the second term of (12), we substitute $\frac{x-y}{h} = u$ and note that $|z| = |-z|$ to get

$$\begin{aligned}
&\Big|\frac{1}{h}\int K\Big(\frac{x-y}{h}\Big)\big(f(y) - f(x)\big)dy\Big| \\
&\quad = \Big|\int K(u)\big(f(x) - f(x-hu)\big)du\Big|.
\end{aligned} \tag{14}$$

Now Taylor's theorem applied to f shows that there is a $\xi \in (x - hu, x)$ such that

$$f(x) = f(x-hu) + \sum_{j=1}^{\kappa-1} \frac{f^{(j)}(x-hu)}{j!}(hu)^j + \frac{f^{(\kappa)}(\xi)}{\kappa!}(hu)^{\kappa}.$$

Plugging this into (14), by (5) all terms in the Taylor series except the last one vanish, and we get using (5) once more (in the third line) and by Assumption 3 and (7) that

$$\begin{aligned}
&\left|\frac{1}{h}\int K\Big(\frac{x-y}{h}\Big)\big(f(y)-f(x)\big)dy\right| \\
&= \frac{h^\kappa}{\kappa!}\left|\int u^\kappa K(u) f^{(\kappa)}(\xi)du\right| \\
&= \frac{h^\kappa}{\kappa!}\left|\int u^\kappa K(u)\big(f^{(\kappa)}(\xi)-f^{(\kappa)}(x)\big)du\right| \\
&\leq \frac{h^\kappa}{\kappa!}\left|f^{(\kappa)}(\xi)-f^{(\kappa)}(x)\right|\cdot\int\left|u^\kappa K(u)du\right| \\
&< \frac{h^\kappa}{\kappa!}L|\xi-x|^\alpha C_1 \ \leq\ C_1 L h^\beta. \qquad (15)
\end{aligned}$$

This latter argument is similar to the one in Theorem 4.1 of Ibragimov and Hasminskii (1981). Combining (12), (13), and (15) proves the lemma. □

Proof of Theorem 1. We first split

$$\begin{aligned}
\sup_x|\hat{f}_N(x)-f(x)| &\leq \mathbb{E}\Big[\sup_x|\hat{f}_N(x)-f(x)|\Big] \\
&+\Big|\sup_x|\hat{f}_N(x)-f(x)|-\mathbb{E}\Big[\sup_x|\hat{f}_N(x)-f(x)|\Big]\Big|. \quad (16)
\end{aligned}$$

Concerning the first term in (16), we first bound it by

$$\begin{aligned}
\mathbb{E}\Big[\sup_x|\hat{f}_N(x)-f(x)|\Big] &\leq \sup_x|\mathbb{E}[\hat{f}_N(x)]-f(x)| \\
&+\mathbb{E}\Big[\sup_x|\hat{f}_N(x)-\mathbb{E}[\hat{f}_N(x)]|\Big]. \quad (17)
\end{aligned}$$

The second term of (17) can be bounded by

$$\mathbb{E}\Big[\sup_x|\hat{f}_N(x)-\mathbb{E}[\hat{f}_N(x)]|\Big] \leq \frac{C_3}{2\pi\sqrt{N}h} \qquad (18)$$

as shown in the Lemma on p.6 of Vogel and Schettler (2013). It is straightforward to check that the given proof works also for independent samples and actually does not make use of the i.i.d. assumption. For the first term of (17) we can use Lemma 1, so that we obtain

$$\mathbb{E}\Big[\sup_x|\hat{f}_N(x)-f(x)|\Big] \leq \frac{C_3}{2\pi\sqrt{N}h}+C_0Ln^{-\alpha}+C_1Lh^\beta. \qquad (19)$$

The second term of (16) can be bounded as in Theorem 1 of Vogel and Schettler (2013). In particular, the i.i.d. assumption is not used in the arguments and independence is sufficient to obtain

$$\begin{aligned}
\Pr\Big\{\Big|\sup_x|\hat{f}_N(x)-f(x)|-\mathbb{E}\Big[\sup_x|\hat{f}_N(x)-f(x)|\Big]\Big| \\
\geq \frac{u}{\sqrt{N}h}\Big\} \leq 2e^{\frac{-u^2}{2C_1^2}}. \quad (20)
\end{aligned}$$

From (16), (19), and (20), we finally get the claim of the theorem. □

5.2. Proof of Theorem 2

The proof structure follows that of Ortner and Ryabko (2012) so that we can take some equations directly from there. However, some arguments have to be changed and adapted.

5.2.1. A Lemma

The main regret term in the discretized MDP comes from a sum over all confidence intervals in the visited state-action pairs. In order to bound this term we use the following lemma. This lemma is a generalization of Lemma 19 of Jaksch et al. (2010), which showed the result for the case $\alpha = \frac{1}{2}$.

Lemma 2. *For any sequence of positive numbers* $z_1,\ldots,z_n$ *with* $0 \leq z_k \leq Z_{k-1} := \max\{1,\sum_{i=1}^{k-1} z_i\}$ *and any* $\alpha\in[0,1]$,

$$\sum_{k=1}^n \frac{z_k}{Z_{k-1}^{1-\alpha}} \leq \frac{Z_n^\alpha}{2^\alpha-1}.$$

Proof. For $n=1$ the lemma is easy to verify. Proceeding by induction on n, note that for $x\in[0,1]$ and $\alpha\in[0,1]$ it holds that $1+(2^\alpha-1)x \leq (1+x)^\alpha$. Thus, choosing $x=\frac{z_n}{Z_{n-1}}$ and multiplying with $\frac{Z_{n-1}^\alpha}{2^\alpha-1}$ yields

$$\frac{Z_{n-1}^\alpha}{2^\alpha-1}+\frac{z_n}{Z_{n-1}^{1-\alpha}} \leq \frac{1}{2^\alpha-1}\big(Z_{n-1}+z_n\big)^\alpha.$$

Using this and the induction assumption, we get

$$\begin{aligned}
\sum_{k=1}^n \frac{z_k}{Z_{k-1}^{1-\alpha}} &= \sum_{k=1}^{n-1}\frac{z_k}{Z_{k-1}^{1-\alpha}}+\frac{z_n}{Z_{n-1}^{1-\alpha}} \leq \frac{Z_{n-1}^\alpha}{2^\alpha-1}+\frac{z_n}{Z_{n-1}^{1-\alpha}} \\
&\leq \frac{(Z_{n-1}+z_n)^\alpha}{2^\alpha-1} = \frac{Z_n^\alpha}{2^\alpha-1}.
\end{aligned}$$

□

5.2.2. Splitting into Episodes

Let $v_k(s,a)$ be the number of times action a has been chosen in episode k when being in state s. Define the regret in episode k to be

$$\Delta_k := \sum_{s,a} v_k(s,a)\big(\rho^*-r(s,a)\big). \qquad (21)$$

Then, as in Section 5.1 of Ortner and Ryabko (2012) (cf. also Section 4.1 of Jaksch et al., 2010), with probability at least $1-\frac{\delta}{12T^{5/4}}$ the regret of UCCRL-KD is upper bounded by

$$\sqrt{\tfrac{5}{8}T\log\left(\tfrac{8T}{\delta}\right)}+\sum_k\Delta_k. \qquad (22)$$

5.2.3. Failing Confidence Intervals

We continue by considering the regret when the true MDP is not contained in the set of plausible MDPs.

Rewards For the rewards we have the same Assumption 1 as Ortner and Ryabko (2012), so this case can be handled in the same way. Thus, the estimated rewards $\hat{r}(s,a)$ are computed from the observed rewards in states s_i that are in the same interval as state s. Assume that at step t there have been $N > 0$ samples of action a in such states. Then we obtain as in (Ortner and Ryabko, 2012) from Hoeffding inequality that

$$\Pr\Big\{\Big|\hat{r}(s,a)-\mathbb{E}[\hat{r}(s,a)]\Big| \geq \sqrt{\tfrac{7}{2N}\log\big(\tfrac{2nAt}{\delta}\big)}\Big\} \leq \frac{\delta}{60nAt^7}.$$

Further, we have $\mathbb{E}[\hat{r}(s,a)] = \frac{1}{N}\sum_{i=1}^{N} r(s_i,a)$. Since the s_i are assumed to be in the same interval $I(s)$ as s, it follows that $|\mathbb{E}[\hat{r}(s,a)] - r(s,a)| < Ln^{-\alpha}$.

A union bound over all actions, all n intervals I_j and all t possible values of N then shows that with probability at least $1 - \frac{\delta}{15t^6}$ it holds that

$$\big|\hat{r}(s,a) - r(s,a)\big| < Ln^{-\alpha} + \sqrt{\frac{7\log(2nAt/\delta)}{2N_t(I(s),a)}}. \qquad (23)$$

Transition Probabilities Now for the estimates of the transition probabilities we apply Theorem 1 to obtain confidence intervals that hold with high probability. At step t, for each state s in which we want to estimate $p(\cdot|s,a)$ the samples will only come from nearby states $s_1,\dots,s_N$ that are in the interval $I(s)$ also containing s. Thus, the samples will be independent and, according to Assumption 2, from close (but not necessarily identical) distributions.

We apply Theorem 1 to obtain confidence intervals for the transition probability estimates. Choosing $h = N^{\frac{-1}{2\beta+2}}$ in Theorem 1 gives

$$\Pr\Big\{\big\|\hat{p}_N(\cdot|s,a) - p(\cdot|s,a)\big\|_1 \geq N^{\frac{-\beta}{2\beta+2}}\cdot(u+C') + C_0Ln^{-\alpha}\Big\} < 2e^{\frac{-u^2}{2C_2^2}},$$

where $\hat{p}_N(\cdot|s,a)$ is the kernel density estimate for $p(\cdot|s,a)$ computed from N samples, and $C' := \frac{C_3}{2\pi} + C_1L$. Hence, with probability at least $1 - \frac{\delta}{15nAt^7}$

$$\big\|\hat{p}_N(\cdot|s,a) - p(\cdot|s,a)\big\|_1 \leq C_0Ln^{-\alpha} + N^{\frac{-\beta}{2\beta+2}}\Big(C_2\sqrt{2\log\big(\tfrac{30nAt^7}{\delta}\big)} + C'\Big).$$

A union bound over all n intervals, all actions, and all t possible values for N then shows that with probability at least $1 - \frac{\delta}{15t^6}$

$$\big\|\hat{p}(\cdot|s,a) - p(\cdot|s,a)\big\|_1 \leq C_0Ln^{-\alpha} + N_t(I(s),a)^{\frac{-\beta}{2\beta+2}}C_1'\sqrt{14\log\big(\tfrac{2nAt}{\delta}\big)} \quad (24)$$

for the actual value $N_t(I(s),a)$ and all state-action pairs (s,a), where we choose $C_1' := C_2 + C'$.

Regret when Confidence Intervals Fail In (23) and (24) we have shown that the confidence intervals conf_r and conf_p for rewards and transition probabilities as given in (1) and (2) hold with error probability $\frac{\delta}{15t^6}$ each. These error probabilities are the same as in (Ortner and Ryabko, 2012). Therefore, we obtain (cf. also Section 4.2 of Jaksch et al., 2010) the same regret bound for the case when the true MDP is not contained in the set of plausible MDPs, that is, with probability at least $1 - \frac{\delta}{12T^{5/4}}$,

$$\sum_k \Delta_k \mathbb{I}_{M\notin\mathcal{M}_k} \leq \sqrt{T}. \qquad (25)$$

5.2.4. Regret in Episodes with $M \in \mathcal{M}_k$

Now let us finally turn to the regret in episodes where the true MDP M is contained in the set of plausible MDPs $\mathcal{M}_k$. Note that in this case by the optimistic choice of $\tilde{\pi}_k$ it holds that $\tilde{\rho}_k^* := \rho^*(\tilde{M}_k) \geq \rho^*$. Therefore,

$$\rho^* - r(s,a) \leq (\tilde{\rho}_k^* - \tilde{r}_k(s,a)) + (\tilde{r}_k(s,a) - r(s,a)),$$

and we can bound the regret Δ_k of episode k as defined in (21) by (23) and the definition (1) of the confidence intervals conf_r as

$$\Delta_k \leq \sum_s v_k(s,\tilde{\pi}_k(s))\big(\tilde{\rho}_k^* - \tilde{r}_k(s,\tilde{\pi}_k(s))\big) + 2Ln^{-\alpha}\tau_k + \sqrt{14\log\big(\tfrac{2nAt}{\delta}\big)}\sum_{j=1}^{n}\sum_{a\in A}\frac{v_k(I_j,a)}{\sqrt{N_k(I_j,a)}}, \qquad (26)$$

where $\tau_k := t_{k+1} - t_k$ denotes the length of episode k.

Dealing with the Transition Functions The remaining term $\sum_s v_k(s,\tilde{\pi}_k(s))\big(\tilde{\rho}_k^* - \tilde{r}_k(s,\tilde{\pi}_k(s))\big)$ can be analysed similar to Section 5.4 of Ortner and Ryabko (2012). That is, let $\tilde{\lambda}_k := \lambda(\tilde{\pi}_k,\cdot)$ be the bias function of policy $\tilde{\pi}_k$ in the optimistic MDP $\tilde{M}_k$. Then by the Poisson equation,

$$\begin{aligned}\tilde{\rho}_k^* &- \tilde{r}_k(s,\tilde{\pi}_k(s)) \\ &= \int \tilde{p}_k(ds'|s,\tilde{\pi}_k(s))\cdot\tilde{\lambda}_k(s') - \tilde{\lambda}_k(s) \\ &= \int p(ds'|s,\tilde{\pi}_k(s))\cdot\tilde{\lambda}_k(s') - \tilde{\lambda}_k(s) \\ &+ \int \Big(\tilde{p}_k(ds'|s,\tilde{\pi}_k(s)) - p(ds'|s,\tilde{\pi}_k(s))\Big)\cdot\tilde{\lambda}_k(s'). \end{aligned} \qquad (27)$$

The last term in (27) can be bounded by splitting up

$$\tilde{p}_k(\cdot|s,a) - p(\cdot|s,a) = \big(\tilde{p}_k(\cdot|s,a) - \hat{p}_k(\cdot|s,a)\big) + \big(\hat{p}_k(\cdot|s,a) - p(\cdot|s,a)\big)$$

and applying (24) and the definition (2) of the confidence intervals conf_p. Noting that by definition of the algorithm

$\|\tilde{\lambda}_k\|_\infty \leq H$, this gives

$$\sum_s v_k(s,\tilde{\pi}_k(s)) \int \Big(\tilde{p}_k(ds'|s,\tilde{\pi}_k(s)) - p(ds'|s,\tilde{\pi}_k(s))\Big) \tilde{\lambda}_k(s')$$
$$\leq 2H \sum_{j=1}^{n} \sum_{a\in A} v_k(I_j,a) N_k(I_j,a)^{\frac{-\beta}{2\beta+2}} C_1' \sqrt{14 \log\left(\tfrac{2nAT}{\delta}\right)}$$
$$+ 2HC_0 L n^{-\alpha} \tau_k. \quad (28)$$

For the first term in (27), the same martingale argument as given in (Ortner and Ryabko, 2012) yields that with probability at least $1 - \frac{\delta}{12T^{5/4}}$

$$\sum_k \sum_s v_k(s,\tilde{\pi}_k(s)) \Big(\int p(ds'|s,\tilde{\pi}_k(s)) \cdot \tilde{\lambda}_k(s') - \tilde{\lambda}_k(s) \Big)$$
$$\leq H\sqrt{\tfrac{5}{2} T \log\left(\tfrac{8T}{\delta}\right)} + HnA \log_2\left(\tfrac{8T}{nA}\right). \quad (29)$$

5.2.5. Total Regret

Summing Δ_k over all episodes with $M \in \mathcal{M}_k$ we obtain from (26), (27), (28), and (29) that with probability at least $1 - \frac{\delta}{12T^{5/4}}$

$$\sum_k \Delta_k \mathbb{I}_{M\in\mathcal{M}_k} \leq 2HC_0 L n^{-\alpha} T$$
$$+ 2H \sum_k \sum_{j=1}^{n} \sum_{a\in A} v_k(I_j,a) N_k(I_j,a)^{\frac{-\beta}{2\beta+2}}$$
$$\times C_1' \sqrt{14 \log\left(\tfrac{2nAT}{\delta}\right)} + H\sqrt{\tfrac{5}{2} T \log\left(\tfrac{8T}{\delta}\right)}$$
$$+ HnA \log_2\left(\tfrac{8T}{nA}\right) + 2Ln^{-\alpha} T$$
$$+ \sqrt{14 \log\left(\tfrac{2nAT}{\delta}\right)} \sum_k \sum_{j=1}^{n} \sum_{a\in A} \frac{v_k(I_j,a)}{\sqrt{N_k(I_j,a)}}. \quad (30)$$

Writing $N(I_j,a)$ for the total number of times a has been played in a state in I_j after T steps, we have $\sum_j \sum_a N(I_j,a) = T$, and application of Lemma 2 and Jensen's inequality yields

$$\sum_k \sum_{j=1}^{n} \sum_{a\in A} v_k(I_j,a) N_k(I_j,a)^{\frac{-\beta}{2\beta+2}}$$
$$\leq \frac{1}{2^{1-\frac{\beta}{2\beta+1}} - 1} \sum_{j=1}^{n} \sum_{a\in A} N(I_j,a)^{1-\frac{\beta}{2\beta+2}}$$
$$\leq \frac{1}{2^{\frac{\beta+1}{2\beta+1}} - 1} (nA)^{\frac{\beta}{2\beta+2}} T^{\frac{\beta+2}{2\beta+2}}. \quad (31)$$

As for UCCRL (cf. also Appendix C.3 of Jaksch et al., 2010) we also have that (provided that $n \leq T$)

$$\sum_k \sum_{j=1}^{n} \sum_{a\in A} \frac{v_k(I_j,a)}{\sqrt{N_k(I_j,a)}} \leq (\sqrt{2}+1)\sqrt{nAT}$$
$$\leq (\sqrt{2}+1)\sqrt{A} \cdot n^{\frac{\beta}{2\beta+2}} T^{\frac{\beta+2}{2\beta+2}}. \quad (32)$$

From equations (30), (31), and (32) we obtain in combination with (22) and (25) that the regret with probability at least $1 - \frac{\delta}{4T^{5/4}}$ is upper bounded as

$$\sqrt{\tfrac{5}{8} \log\left(\tfrac{8T}{\delta}\right)} + \sum_k \Delta_k \mathbb{I}_{M\notin\mathcal{M}_k} + \sum_k \Delta_k \mathbb{I}_{M\in\mathcal{M}_k}$$
$$\leq \sqrt{\tfrac{5}{8} \log\left(\tfrac{8T}{\delta}\right)} + \sqrt{T} + H\sqrt{\tfrac{5}{2} \log\left(\tfrac{8T}{\delta}\right)}$$
$$+ HnA \log_2\left(\tfrac{8T}{nA}\right) + 2(HC_0+1)Ln^{-\alpha} T$$
$$+ \left(\frac{2HC_1'}{2^{\frac{\beta+1}{2\beta+1}} - 1} + \sqrt{2} + 1\right) n^{\frac{\beta}{2\beta+2}} T^{\frac{\beta+2}{2\beta+2}} \sqrt{14A \log\left(\tfrac{2nAT}{\delta}\right)}.$$

A final union bound over all possible values of T shows after a few simplifications (cf. Appendix C.4 of Jaksch et al., 2010) that with probability at least $1-\delta$ the regret after any T steps is bounded by

$$c \cdot C_1' H \sqrt{14A \log\left(\tfrac{2nAT}{\delta}\right)} n^{\frac{\beta}{2\beta+2}} T^{\frac{\beta+2}{2\beta+2}} + c' \cdot C_0 H L n^{-\alpha} T.$$

□

6. Open Questions

The main missing link is respective lower bounds. Some preliminary bounds have been given by Ortner and Ryabko (2012), but they appear to be not optimal. On the other hand, the construction of lower bounds in our setting (taking into account the assumptions on the transition probabilities) seems not easy. In general, we believe that getting improved lower bounds in the continuous state setting is closely related to the still open problem of closing the gap between known upper and lower bound for the regret in finite state MDPs, cf. (Jaksch et al., 2010).

Concerning computational issues, already for UCCRL it is not clear whether there is an efficient method to compute the optimistic plausible MDP and the respective optimal policy in line 8 of the algorithm. This issue has not been resolved for UCCRL-KD and remains an open question. Also, the necessary input of an upper bound on the bias deteriorates the bounds (by a big additive constant) for UCCRL-KD just like for UCCRL when this bound has to be guessed.

With respect to the need of knowledge of the smoothness parameters, as suggested by Ortner and Ryabko (2012), one can use the model-selection technique introduced in (Maillard et al., 2012) and refined by Maillard et al. (2013) to obtain regret bounds also without explicit knowledge of κ, L, and α. However, these bounds have worse dependence on T. Still, as our bounds are an improvement over the bounds of Ortner and Ryabko (2012), we expect to get an improvement in this case as well. However, the respective technical details still have to be worked out.

Acknowledgments

This work was supported by the Austrian Science Fund (FWF): P 26219-N15, the French Ministry of Higher Education and Research, and the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 270327 (CompLACS).

References

Y. Abbasi-Yadkori and C. Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Learning Theory, 24th Annual Conference on Learning Theory COLT, JMLR Proceedings Track*, volume 19, pages 1–26, 2011.

B. Abdous. Computationally efficient classes of higher-order kernel functions. *Candian Journal Statistics*, 3(1): 21–27, 1995.

P. Auer, R. Ortner, and C. Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In *Learning Theory, 20th Annual Conference on Learning Theory COLT*, pages 454–468, 2007.

A. Bernstein and N. Shimkin. Adaptive-resolution reinforcement learning with polynomial exploration in deterministic domains. *Machine Learning*, 81(3):359–397, 2010.

E. Brunskill, B. R. Leffler, L. Li, M. L. Littman, and N. Roy. Provably efficient learning with typed parametric models. *Journal of Machine Learning Research*, 10: 1955–1988, 2009.

S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvári. Online optimization of χ-armed bandits. In *Advances in Neural Information Processing Systems NIPS*, volume 22, pages 201–208, 2010.

L. Devroye. *A course in density estimation*. Birkhäuser, 1987.

T. Gasser, H. Müller, and V. Mammitzsch. Kernels for nonparametric curve estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2):238–252, 1985.

O. Hernández-Lerma and J. B. Lasserre. *Further topics on discrete-time Markov control processes*, volume 42 of *Applications of mathematics*. Springer, 1999.

I. A. Ibragimov and R. Z. Hasminskii. *Statistical Estimation: Asymptotic Theory*. Springer-Verlag, New York, 1981.

M. Ibrahmi, A. Javanmard, and B. V. Roy. Efficient reinforcement learning for high dimensional linear quadratic systems. In *Advances Neural Information Processing Systems NIPS*, volume 25, pages 2645–2653, 2012.

T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

S. Kakade, M. J. Kearns, and J. Langford. Exploration in metric state spaces. In *Proceedings of 20th International Conference on Machine Learning ICML*, pages 306–312, 2003.

R. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Advances Neural Information Processing Systems NIPS*, volume 17, pages 697–704, 2005.

R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandits in metric spaces. In *Proceedings of 40th Annual ACM Symposium on Theory of Computing STOC*, pages 681–690, 2008.

O.-A. Maillard, P. Nguyen, R. Ortner, and D. Ryabko. Optimal regret bounds for selecting the state representation in reinforcement learning. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Proceedings*, pages 543–551, 2013.

O.-A. Maillard, R. Munos, and D. Ryabko. Selecting the state-representation in reinforcement learning. In *Advances Neural Processing Systems NIPS*, volume 24, pages 2627–2635, 2012.

D. Ormoneit and S. Sen. Kernel-based reinforcement learning. *Machine Learning*, 49(2-3):161–178, 2002.

R. Ortner and D. Ryabko. Online regret bounds for undiscounted continuous reinforcement learning. In *Advances in Neural Information Processing Systems NIPS*, volume 25, pages 1763–1772, 2012.

I. Osband and B. Van Roy. Model-based reinforcement learning and the eluder dimension. In *Advances in Neural Information Processing Systems NIPS*, volume 27, pages 1466–1474, 2014.

A. L. Strehl and M. L. Littman. Online linear regression and its application to model-based reinforcement learning. In *Advances Neural Information Processing Systems NIPS*, volume 20, pages 1417–1424, 2008.

A. Tsybakov. *Introduction to nonparametric estimation*. Springer, 2009.

S. Vogel and A. Schettler. A uniform concentration-of-measure inequality for multivariate kernel density estimators. Technical Report M13/09, Technische Universität Ilmenau, Institut für Mathematik, 2013.