JMLR: Workshop and Conference Proceedings 19 (2011) 739–772 24th Annual Conference on Learning Theory

Agnostic KWIK learning and efficient approximate reinforcement learning

István Szita SZITA@UALBERTA.CA
Csaba Szepesvári SZEPESVA@UALBERTA.CA
Department of Computing Science
University of Alberta, Canada

Editor: Sham Kakade, Ulrike von Luxburg

Abstract

A popular approach in reinforcement learning is to use a model-based algorithm, i.e., an algorithm that utilizes a model learner to learn an approximate model to the environment. It has been shown that such a model-based learner is efficient if the model learner is efficient in the so-called "knows what it knows" (KWIK) framework. A major limitation of the standard KWIK framework is that, by its very definition, it covers only the case when the (model) learner can represent the actual environment with no errors. In this paper, we study the agnostic KWIK learning model, where we relax this assumption by allowing nonzero approximation errors. We show that with the new definition an efficient model learner still leads to an efficient reinforcement learning algorithm. At the same time, though, we find that learning within the new framework can be substantially slower as compared to the standard framework, even in the case of simple learning problems.

Keywords: KWIK learning, agnostic learning, reinforcement learning, PAC-MDP

1. Introduction

The *knows what it knows* (KWIK) model of learning (Li et al., 2008) is a framework for online learning against an adversary. Before learning, the KWIK learner chooses a hypothesis class and the adversary selects a function from this hypothesis class, mapping inputs to responses. Then, the learner and the adversary interact in a sequential manner: Given the past interactions, the adversary chooses an input, which is presented to the learner. The learner can either pass, or produce a prediction of the value one would obtain by applying the function selected by the adversary to the selected input. When the learner passed and only in that case, the learner is shown the noise-corrupted true response. *All* predictions produced by the learner must be in a close vicinity to the true response (up to a prespecified tolerance), while the learner's efficiency is measured by the number of times it passes.

The problem with this framework is that if the hypothesis class is small, it unduly limits the power of the adversary, while with a larger hypothesis class efficient learning becomes

problematic. Hence, in this paper we propose an alternative framework that we call the *agnostic KWIK* framework, where we allow the adversary to select functions outside of the hypothesis class, as long the function remains "close" to the hypothesis class, while simultaneously relaxing the accuracy requirement on the predictions.

New models of learning abound in the learning theory literature, and it is not immediately clear why the KWIK framework makes these specific assumptions on the learning process. For the extension investigated in the paper, the *agnostic KWIK* model, even the name seems paradoxical: "agnostic" means "no knowledge is assumed", while KWIK is acronym for "knows what it knows". Therefore, we begin the paper by motivating the framework.

1.1. Motivation

The motivation of the KWIK framework is rooted in reinforcement learning (RL). An RL agent makes sequential decisions in an environment to maximize the long-term cumulated reward it incurs during the interaction (Sutton and Barto, 1998). The environment is initially unknown to the agent, so the agent needs to spend some time exploring it. Exploration, however, is costly as an agent exploring its environment may miss some reward collecting opportunities. Therefore, an efficient RL agent must spend as little time with exploration as possible, while ensuring that the best possible policy is still discovered.

Many efficient RL algorithms (Kearns and Singh, 2002; Brafman and Tennenholtz, 2001; Strehl, 2007; Szita and Lőrincz, 2008; Szita and Szepesvári, 2010) share a common core idea: (1) they keep track of which parts of the environment are known with high accuracy; (2) they strive to get to unknown areas and collect experience; (3) in the known parts of the environment, they are able to plan the path of the agent to go wherever it wants to go, such as the unknown area or a highly rewarding area. The KWIK learning model of Li et al. (2008) abstracts the first point of this core mechanism. This explains the requirements of the framework:

- Accuracy of predictions: a plan based on an approximate model will be usable only if the approximation is accurate. Specifically, a single large error in the model can fatally mislead the planning procedure.
- Adversarial setting: the state of the RL agent (and therefore, the queries about the model) depend on the (unknown) dynamics of the environment in a complex manner. While the assumption that the environment is fully adversarial gives more power to the adversary, this assumption makes the analysis easier (while not preventing it).
- Noisy feedback: the rewards and next states are determined by a stochastic environment, so feedback is necessarily noisy.

The main result of the KWIK framework states that if an algorithm "KWIK-learns" the parameters of an RL environment then it can be augmented to an efficient reinforcement learning algorithm (Li, 2009, Chapter 7.1) and (Li et al., 2011a). The result is significant because it reduces efficient RL to a conceptually simpler problem and unifies a large body of previous works (Li, 2009; Strehl et al., 2007; Diuk et al., 2008; Strehl and Littman, 2007).

For finite horizon learning problems, it is even possible to construct *model-free* efficient RL algorithms using an appropriate KWIK learner, as shown by Li and Littman (2010).

An important limitation of the KWIK framework is that the environment must be exactly representable by the learner. Therefore, to make learning feasible, we must assume that the environment comes from a small class of models (that is, characterized with a small number of parameters), for example, it is a Markov decision process (MDP) with a small, finite state space.

However, such a model of the environment is often just an approximation, and in such cases, not much is known about efficient learning in a KWIK-like framework. The agnostic KWIK learning framework is aimed to fill this gap. In this new framework the learner tries to find a good approximation to the true model with a restricted model class.[1] Of course, we will not be able to predict the model parameters accurately any more (the expressive power of our hypothesis class is insufficient), so the accuracy requirement needs to be relaxed. Our main result is that with this definition the augmentation result of Li et al. (2011a) still holds: an efficient agnostic KWIK-learning algorithm can be used to construct an efficient reinforcement learning algorithm even when the environment is outside of the hypothesis class of the KWIK learner. To our knowledge, this is the first result for reinforcement learning that allows for a nonzero approximation error.

1.2. The organization of the paper

In the next section (Section 2) we introduce the KWIK framework and its agnostic extension. In the two sections following Section 2 we investigate simple agnostic KWIK learning problems. In particular, in Section 3 we investigate learning when the responses are noiseless. Two problems are considered: As a warm-up we consider learning with finite hypothesis classes, followed by the investigation of learning when the hypothesis class contains linear functions with finitely many parameters. In Section 4 we analyze the case when the responses are noisy. Section 5 contains our main result: the connection between agnostic KWIK and efficient approximate RL. Our conclusions are drawn in Section 6. Proofs of technical theorems and lemmas have been moved to the Appendix.

2. From KWIK learning to agnostic KWIK learning

A *problem* is a 5-tuple $G = (\mathcal{X}, \mathcal{Y}, g, Z, \|\cdot\|)$, where $\mathcal{X}$ is the set of inputs, $\mathcal{Y} \subseteq \mathbb{R}^d$ is a measurable set of possible responses, $Z : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ is the noise distribution that is assumed to be zero-mean ($\mathcal{P}(\mathcal{Y})$ denotes the space of probability distributions over $\mathcal{Y}$) and $\|\cdot\| : \mathbb{R}^d \to \mathbb{R}_+$ is a semi-norm on $\mathbb{R}^d$. A *problem class* $\mathcal{G}$ is a set of problems. When each problem in a class shares the same domain $\mathcal{X}$, response set $\mathcal{Y}$ and same semi-norm $\|\cdot\|$, for brevity, the semi-norm will be omitted from the problem specifications. If the noise distribution underlying every $G \in \mathcal{G}$ is a Dirac-measure, we say that the problem class is *deterministic*. For such problem classes, we will also omit to mention the distribution.

1. The real environment is *not known* to belong to the restricted model class, hence the name "agnostic".

The *knows what it knows* (KWIK) framework Li et al. (2011a) is a model of online learning where an (online) learner interacts with an environment.[2] In this context, an *online learner* L is required to be able to perform two operations:

- **predict**: For an input $x \in \mathcal{X}$, L must return an answer $\hat{y} \in \mathcal{Y} \cup \{\perp\}$. The answer $\hat{y} = \perp$ means that the learner *passes*.
- **update**: Upon receiving an input-response pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$, L should update its internal representation.

At the beginning of learning, the environment secretly selects a problem $(\mathcal{X}, \mathcal{Y}, g^*, Z)$ from some class and it also selects the inputs x_t which are presented to the learner in a sequential manner. Given an input x_t, the learner has the option to pass (say "I don't know"), or to make a prediction. An admissible learner is required to make accurate predictions only. When the learner passes (and only in that case), the environment tells it the response (or answer) y_t, which is randomly chosen so that $z_t = y_t - g^*(x_t) \sim Z(x_t)$ and z_t is independent of the past given x_t. The learner's goal is to minimize the number of passes, while staying admissible. The environment is assumed to choose the problem and the inputs adversarially and so we sometimes call the environment the *adversary*. In the case of noisy responses, exact, or near-optimal predictions are in general impossible to achieve with certainty. Correspondingly, we introduce two parameters: the required accuracy $\epsilon \geq 0$ and the maximum permitted failure probability δ.

The KWIK protocol, controlling the interaction between the online learner and the adversary, is shown as Algorithm 1. Note that in the standard KWIK-framework, before

Algorithm 1 *The KWIK protocol* $(\mathcal{G}, \epsilon)$.

$N_0 = 0$ {N_t is the number of times the learner "passed".}
Adversary picks problem $G^* = (\mathcal{X}, \mathcal{Y}, g^*, Z, \|\cdot\|) \in \mathcal{G}$ and $(\mathcal{X}, \mathcal{Y}, \|\cdot\|)$ is told learner
for $t = 1, 2, \ldots$ **do**
  Adversary picks query $x_t \in \mathcal{X}$, which is announced to learner.
  Learner computes answer $\hat{y}_t \in \mathcal{Y} \cup \{\perp\}$ (**predict** is called), which is announced to adversary.
  $N_t = N_{t-1}$
  if $\hat{y}_t = \perp$ **then**
    Adversary tells learner $y_t = g^*(x_t) + z_t$, where $z_t \sim Z(\cdot|x_t)$
    Learner updates itself (**update** is called)
    N_t is incremented by 1
  else if $\|\hat{y}_t - g^*(x_t)\| > \epsilon$ **then**
    return FAIL

learning, both adversary and learner are given $\mathcal{G}$ and ϵ. Further, learner might be given a confidence parameter $0 \leq \delta \leq 1$, whose role will be explained soon. In particular, this means that the learner can adjust its strategy to $(\mathcal{G}, \epsilon, \delta)$. Note also that the environment

2. Our definitions are slightly different from the original ones, mostly for the sake of increased rigour and to make them better fit our results. Specifically, we explicitly include the noise as part of the concept.

is allowed to pick x_t based on any past information available to it up to time t.[3] We note in passing that if in an application, regardless of the decision of the learner, the response y_t is generated and is communicated to the learner at every step, the learning problem can only become easier. We call the so-modified protocol, the *relaxed KWIK protocol*. If the learner is reasonable, the extra information will help it, though, this calls for an explicit proof. All the definitions below extend to the relaxed KWIK protocol.

Definition 2.1 *Fix $\epsilon \geq 0$ and $0 \leq \delta \leq 1$ and a problem class $\mathcal{G}$. A learner L is an* admissible (and bounded) *(ϵ, δ) KWIK-learner for $\mathcal{G}$ if, with probability at least $1 - \delta$, it holds that when L and an arbitrary adversary interact following the KWIK protocol, the protocol does not fail (and the number of passes N_t stays bounded by a finite deterministic quantity $B(\mathcal{G}, \epsilon, \delta)$). We call the quantity $B(\mathcal{G}, \epsilon, \delta)$ the learner's KWIK-bound. The problem class $\mathcal{G}$ is* (ϵ, δ) KWIK-learnable*, if there exists a bounded, admissible (ϵ, δ) KWIK-learner L for $\mathcal{G}$. Further, $\mathcal{G}$ is* KWIK-learnable*, if it is (ϵ, δ) KWIK learnable for any $\epsilon > 0$, $0 < \delta < 1$. If $B(\mathcal{G}, \epsilon, \delta)$ is the learner's KWIK-bound, we say that $\mathcal{G}$ is KWIK-learnable with KWIK-bound $B(\mathcal{G}, \epsilon, \delta)$.*

Note that the learners can be specialized to $\mathcal{G}$, ϵ and δ. However, interesting results concern *general* KWIK-learners which are operate for any $\mathcal{G}$ from a *meta-class* of concept-classes $\mathcal{C}$. For example, the memorization learner of Li (2009) is a bounded, admissible KWIK learner for *any* problem class $\mathcal{G}$ where the problems in $\mathcal{G}$ are deterministic and share the same finite input space $\mathcal{X}$.

In addition to the above concepts, it is also customary to define the notion of KWIK-learnability:

Definition 2.2 *Let $c : \mathcal{G} \to \mathbb{R}_+$ be a real-valued function. The problem class $\mathcal{G}$ is* c-efficiently KWIK-learnable*, if, for any $\epsilon > 0$, $0 \leq \delta \leq 1$, it is (ϵ, δ) KWIK-learnable by a learner L whose KWIK-bound B satisfies $B(\mathcal{G}, \epsilon, \delta) \leq \mathrm{poly}(c(\mathcal{G}), 1/\epsilon, \log(1/\delta))$ for some polynomial* poly*. Further, $\mathcal{G}$ is* c-efficiently deterministically KWIK-learnable *if the above polynomial is independent of δ. Finally, $\mathcal{G}$ is* c-exactly KWIK-learnable *if* $\mathrm{poly}(c, 1/\epsilon, \log(1/\delta))$ *is independent of* $1/\epsilon$.[4]

Examples of KWIK-learnable classes can be found in the thesis by Li (2009).

2.1. Agnostic KWIK learning

From now on, we will assume that $\mathcal{G}$ is such that all problems in it share the same domain and response spaces. A crucial assumption of the KWIK framework is that learner gets to know $\mathcal{G}$ at the beginning of learning – the so-called *realizability* assumption.[5] To illustrate

3. The choice must be measurable to avoid pathologies.
4. In the definition, contrary to previous work, we intentionally use $\log(1/\delta)$ instead of $1/\delta$ because $\log(1/\delta)$ is more natural in a learning context and a $1/\delta$-bound looks unnecessarily weak.
5. If the learner knows $\mathcal{G}$, it can "realize" any problem chosen by the adversary, hence the name.

the importance of this assumption take $\mathcal{X} = \mathbb{N}$, $\mathcal{Y} = \{-1, +1\}$ and let $\mathcal{G}$ have two disjoint deterministic problems (functions) in it. Then, trivially, there is a KWIK-learner which is bounded and admissible *independently of* how the two deterministic functions are chosen. However, for any KWIK-learner who remains *uninformed about the choice of* $\mathcal{G}$ there exists a class $\mathcal{G}$ with two functions that makes the learner fail.

However, in practice the realizibility assumption might be restrictive: The user of a learning algorithm might give the learner a problem class (the hypothesis class), $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$, that may or may not contain the problem to be learned. In this case one still expects performance to degrade in a graceful manner as a function of the "distance" between the problem selected by the adversary and $\mathcal{H}$. In particular, it is reasonable to relax the accuracy requirements in proportion to this distance. Therefore, in our *agnostic KWIK learning framework* we propose to allow prediction errors of size $rD + \epsilon$ (instead of ϵ), where D is the maximum tolerable approximation error.

The formal definitions are as follows. Let

$$\Delta(\mathcal{G}, \mathcal{H}) \stackrel{\text{def}}{=} \sup_{(\mathcal{X},\mathcal{Y},g,Z)\in\mathcal{G}} \inf_{h\in\mathcal{H}} \|h - g\|_\infty.$$

denote the error of approximating the functions of $\mathcal{G}$ by elements of $\mathcal{H}$.

Definition 2.3 *Fix a hypothesis class $\mathcal{H}$ over the domain $\mathcal{X}$ and response set $\mathcal{Y}$, an approximation error bound $D > 0$, a competitiveness factor $r > 0$, an accuracy-slack $\epsilon \geq 0$ and a confidence parameter $0 \leq \delta \leq 1$. Let $\mathcal{G}$ be a problem class $\mathcal{G}$ over $(\mathcal{X}, \mathcal{Y})$ that satisfies $\Delta(\mathcal{G}, \mathcal{H}) \leq D$. Then, a learner L is a* (D, r, ϵ, δ) agnostic KWIK-learner *for the pair $(\mathcal{H}, \mathcal{G})$ if L is an $(rD + \epsilon, \delta)$ KWIK-learner for $\mathcal{G}$.*

The other learnability concepts (e.g., (ϵ, δ) learnability, learnability, efficiency, etc.) can also be defined analogously.

3. Learning deterministic problem classes

The results in this section are used to illustrate the definitions, and the role of the various parameters (such as r). The common property of the learning problems studied here is that the responses are noise-free. As a warm-up, learning with finite hypothesis classes is considered. Next, we consider learning with a hypothesis class composed of linear functions. We will see that in this case KWIK-learning is still possible, but can be exponentially slow as a function of the dimension of the input space. Note that our learners are deterministic. Hence, all statements hold either with probability one or probability zero. In particular, a bounded, admissible KWIK-learner is necessarily a bounded and admissible KWIK-learner even with the choice of $\delta = 0$.

3.1. Learning with a finite hypothesis class

For any $d > 0$ and $y \in \mathcal{Y}$, define the d-ball around y as

$$B_d(y) \stackrel{\text{def}}{=} \{y' \in \mathcal{Y} \ : \ \|y' - y\| \leq d\}\,.$$

Algorithm 2 Generic Agnostic Learner for deterministic problem classes.

initialize(D, $\mathcal{H}$)
$\mathcal{F} := \mathcal{H}$ and store D

learn(x, y)
$\mathcal{F} := \mathcal{F} \setminus \{f \in \mathcal{F} \ : \ \|f(x) - y\| > D\}$

predict(x)
$Y := \bigcap_{f \in \mathcal{F}} B_D(f(x))$
if $Y \neq \emptyset$ **then**
  return an arbitrary $\hat{y} \in Y$
else
  return $\perp$

Consider the Generic Agnostic Learner (Algorithm 2) of Littman (the algorithm is published by Li (2009), without analysis and for $\mathcal{Y} \subseteq \mathbb{R}$). Every time a new query x_t is received, the algorithm checks whether there exists some value $\hat{y}_t$ that remaining functions agree at x_t up to the accuracy D. If such a value exists, the learner predicts $\hat{y}_t$, otherwise it passes. When the learner passes it learns the response y_t, based on which it can exclude at least one concept. This results in the following statement:

Theorem 3.1 *Let $(\mathcal{X}, \mathcal{Y})$ be arbitrary sets, $r = 2$, $\epsilon = 0$, $D > 0$, $\mathcal{H}$ a finite hypothesis class over $(\mathcal{X}, \mathcal{Y})$, $\mathcal{G}$ a deterministic problem class over $(\mathcal{X}, \mathcal{Y})$ with $\Delta(\mathcal{G}, \mathcal{H}) \leq D$. Then, the Generic Agnostic Learner is an agnostic (D, r, ϵ) KWIK-learner for $(\mathcal{H}, \mathcal{G})$ with KWIK-bound $|\mathcal{H}| - 1$.*

The factor $r = 2$ in the above theorem is the best possible as long as $\mathcal{X}$ is infinite:

Theorem 3.2 *Fix any $D > 0$ and an infinite domain $\mathcal{X}$. Then, there exists a finite response set $\mathcal{Y} \subset \mathbb{R}$, a two-element hypothesis class $\mathcal{H}$ and a deterministic problem class $\mathcal{G}$, both over $(\mathcal{X}, \mathcal{Y})$, that satisfy $\Delta(\mathcal{G}, \mathcal{H}) \leq D$ such that there is no bounded agnostic $(D, r, 0)$ KWIK-learner for $(\mathcal{H}, \mathcal{G})$ with competitiveness factor $0 \leq r < 2$.*

Because of this result, in what follows, we will restrict out attention to $r = 2$. Note that the KWIK-bound we got is identical to the (worst-case) bound that is available when $D = 0$, that is, seemingly there is no price associated to $D > 0$ in terms of the KWIK-bound. However, the worst-case approach taken here leads to overly conservative bounds as the structure of a hypothesis space may allow much better bounds (this is discussed in Section 5.5.2 of Li (2009)). In the next section we study linear hypothesis classes for which the above bound would be vacuous.

3.2. Learning with linear hypotheses

Sometimes Algorithm 2 is applicable even when the hypothesis class $\mathcal{H}$ is infinite: First, the set of remaining hypotheses $\mathcal{F} = \mathcal{H}(x_1, y_1, \ldots, x_n, y_n)$ after n passes can be "implicitly

represented" by storing the list $(x_1, y_1, \ldots, x_n, y_n)$ of pairs received after the n passes. Then, the algorithm remains applicable as long as there is some procedure for checking whether $Y = \bigcap_{f\in\mathcal{F}} B_D(f(x))$ is empty (and finding an element of Y, if it is non-empty). It remains to see then if the procedure is efficient and if the learner stays bounded (that the procedure stays admissible for $r \geq 2$ follows from its definition).

In this section we consider the case when the functions in the hypothesis class are linear in the inputs. More specifically, let $\mathcal{X} = [-X_{\max}, X_{\max}]^d$ for some $d \in \{1, 2, \ldots\} = \mathbb{N}$, $\mathcal{Y} = \mathbb{R}$, $\|\cdot\| = |\cdot|$, $M > 0$ and choose $\mathcal{H}$ to be the set of bounded-parameter linear functions: Denote by $f_\theta : \mathcal{X} \to \mathbb{R}$ the linear function $f_\theta(x) = \theta^\top x$, $x \in \mathcal{X}$. Then,

$$\mathcal{H}_{\mathrm{lin}(M)} \stackrel{\mathrm{def}}{=} \{f_\theta \,:\, \theta \in \mathbb{R}^d, \|\theta\|_\infty \leq M\}\,.$$

Then, $\mathcal{F} = \mathcal{H}(x_1, y_1, \ldots, x_n, y_n) = \{f_\theta \,:\, \theta \in \mathbb{R}^d, -M \leq \theta_i \leq M, y_j - D \leq f_\theta(x_j) \leq y_j + D, 1 \leq i \leq d, 1 \leq j \leq n\}$ since for any (x_i, y_i) pair the hypotheses not excluded must satisfy $|f_\theta(x) - y| \leq D$. Now, $Y = [y^-, y^+]$, where $y^- = \min_{f_\theta\in\mathcal{F}} f_\theta(x)$ and $y^+ = \max_{f_\theta\in\mathcal{F}} f_\theta(x)$ and both y^- and y^+ can be efficiently computed using linear programming. The resulting algorithm is called the deterministic linear agnostic learner (Algorithm 3). In the algorithm, we also allow for a slack $\epsilon > 0$.

Algorithm 3 Deterministic Linear Agnostic Learner

initialize($X_{\max}$, M, D, ϵ)
$C := \{\theta : -M \leq \theta_i \leq M, \quad \forall i \in \{1, \ldots, d\}\}$

learn(x, y)
$C := C \cap \{\theta : y - D \leq \theta^T x \leq y + D\}$

predict(x)
$y^+ := \max_{\theta\in C} \theta^T x$ {solve LP}
$y^- := \min_{\theta\in C} \theta^T x$ {solve LP}
if $y^+ - y^- \leq 2(D + \epsilon)$ **then**
  return $(y^+ + y^-)/2$
else
  return $\perp$

The following theorem shows that for $\epsilon > 0$ this algorithm is a bounded, admissible agnostic KWIK learner:

Theorem 3.3 *Let $X_{\max} > 0$, $\mathcal{X} = [-X_{\max}, X_{\max}]^d$, $\mathcal{Y} = \mathbb{R}$, $M, D, \epsilon > 0$, $r = 2$, $\mathcal{H} = \mathcal{H}_{lin(M)}$. Then, for any $\mathcal{G}$ deterministic problem class over $(\mathcal{X}, \mathcal{Y})$ with $\Delta(\mathcal{G}, \mathcal{H}) \leq D$, it holds that the deterministic linear agnostic learner is an agnostic (D, r, ϵ) KWIK-learner for $(\mathcal{H}, \mathcal{G})$ with the KWIK-bound $2d! \left(\frac{MX_{\max}}{\epsilon} + 1\right)^d$.*

The theorem cannot hold with $\epsilon = 0$, as shown by the following result:

Theorem 3.4 *Let $\mathcal{X}, \mathcal{Y}, D, r, \mathcal{H}$ be as in Theorem 3.3. Then, there exists a problem class $\mathcal{G}$ that satisfies $\Delta(\mathcal{G}, \mathcal{H}) \leq D$ such that there is no bounded, agnostic $(D, r, 0)$ KWIK-learner for $(\mathcal{H}, \mathcal{G})$.*

In Theorem 3.3 the KWIK-bound scales exponentially with d. The next result shows that this is the best possible scaling behavior:

Theorem 3.5 *Fix $\mathcal{X}, \mathcal{Y}, D, \mathcal{H}, \epsilon$ as in Theorem 3.3 and let $r \geq 2$. Then, there exists some problem class $\mathcal{G}$ so that any algorithm that agnostic (D, r, ϵ) KWIK learns $(\mathcal{H}, \mathcal{G})$ will pass at least 2^{d-1} times.*

Whether the scaling behavior of the bound of Theorem 3.3 as a function of ϵ can be improved remains for future work. Note that for $D = 0$ we get an algorithm close to Algorithm 13 of Li (2009), the difference being that Algorithm 13 never predicts unless the new input lies in the span of past training vectors. Nevertheless, for $D = 0$ and any $\epsilon \geq 0$ (including $\epsilon = 0$), our algorithm is also an ϵ KWIK-learner with KWIK-bound d. Thus, we see that the price of non-realizibility is quite high.

4. Learning in bounded noise

As opposed to the previous section, in this section we consider the case when the responses are noisy. We first consider the case of learning with finite hypothesis classes and then we briefly outline a simple discretization-based approach to the case when the hypothesis class is infinite. We further assume that $\mathcal{Y} \subseteq \mathbb{R}$ and the range of the noise in the responses is bounded by K.

4.1. The case of finite hypothesis classes

Let the finite hypothesis class given to the learner be $\mathcal{H}$ and fix some $D > 0$. Let g^* be the function underlying the problem chosen by the adversary from some class $\mathcal{G}$ which satisfies $\Delta(\mathcal{G}, \mathcal{H}) \leq D$. Assuming that the noise in the responses is bounded to lie in $[-K, K]$ for some $K > 0$, an application of the Hoeffding-Azuma inequality gives that for any $\epsilon > 0$ and any fixed function $f \in \mathcal{H}$ such that $\|f - g^*\| \leq D$ (such functions exists, by our assumption connecting $\mathcal{G}$ and $\mathcal{H}$), $0 < \delta \leq 1$, with probability $1 - \delta$, it holds that

$$\left| \frac{1}{m} \sum_{k=1}^{m} \{f(x_k) - y_k\} \right| \leq D + K \sqrt{\frac{2}{m} \log\left(\frac{2}{\delta}\right)} \leq D + \epsilon \,, \tag{1}$$

where $m = m(\epsilon, \delta) > 0$ is chosen large enough so that the second inequality is satisfied and where $((x_k, y_k); k = 1, 2, \ldots)$ is the list of training examples available to the algorithm (the application of Hoeffding-Azuma is not entirely immediate, for the details see Lemma A.1). One idea then is to eliminate those functions f from $\mathcal{H}$ which fail to satisfy (1) after m examples have been seen. The problem is that this rule is based on an average. A clever adversary, who wants to prevent the elimination of some function $\hat{f} \in \mathcal{H}$ could then provide many examples (x_k, y_k) such that y_k is close to $\hat{f}(x_k)$, thus shifting the average to a small value. Therefore, we propose an alternate strategy which is based on the pairwise comparison of hypothesis.

The idea is that if f, f' are far from each other, say at $x \in \mathcal{X}$ it holds that $|f(x) - f'(x)| > 2(D + \epsilon)$, then if the adversary feeds x enough number of times, we can eliminate at least one of f and f'. The following definition will become handy: for two numbers $y, y' \in \mathbb{R}$, we define

$y \ll y'$ by $y + 2(D+\epsilon) \le y'$. We index the elements of $\mathcal{H}$ from 1 to N: $\mathcal{H} = \{f_1, \dots, f_N\}$. The algorithm that we propose is shown as Algorithm 4.

Algorithm 4 Pairwise Elimination-based Agnostic Learner

initialize(D, $\mathcal{H}$, ϵ)
$N := |\mathcal{H}|$, $I := \{1, \dots, N\}$
$n_{i,j} := 0, s_{i,j} = 0, \forall i, j \in I$
$m := \lceil \frac{2K^2}{\epsilon^2} \log \frac{2(N-1)}{\delta} \rceil$

predict(x)
$Y := \bigcap_{i \in I} B_{D+\epsilon}(f_i(x))$
if $Y \neq \emptyset$ **then**
  return an arbitrary $\hat{y} \in Y$
else
  return $\perp$

learn(x, y)
for all $i, j \in I$ such that $f_i(x) \ll f_j(x)$ **do**
  $n_{i,j} := n_{i,j} + 1$
  $s_{i,j} := s_{i,j} + (f_i(x) + f_j(x))/2 - y$
  if $n_{i,j} = m$ **then**
    if $s_{i,j} < 0$ **then**
      $I := I \setminus \{i\}$
    else
      $I := I \setminus \{j\}$

The following theorem holds true for this algorithm:

Theorem 4.1 *Let $\mathcal{H}$ be a finite hypothesis class over $(\mathcal{X}, \mathbb{R})$, D, $\epsilon > 0$, $0 \le \delta \le 1$, $r = 2$. Then, for any $\mathcal{G}$ problem class such that the noise in the responses lies in $[-K, K]$ and $\Delta(\mathcal{G}, \mathcal{H}) \le D$ it holds that the pairwise elimination based agnostic learner is an agnostic (D, r, ϵ, δ) KWIK-learner for $(\mathcal{H}, \mathcal{G})$ with KWIK-bound $((\lceil \frac{2K^2}{\epsilon^2} \log \frac{2(N-1)}{\delta} \rceil - 1)N + 1)(N - 1) = O\left(\frac{K^2N^2}{\epsilon^2} \log \frac{N}{\delta}\right)$.*

4.2. The case of infinite hypothesis classes

Note that by introducing an appropriate discretization, the algorithm can also be applied to problems when the hypothesis set is infinite. In particular, given $\epsilon > 0$, if there exists $N > 0$ and $\mathcal{H}_N \subset \mathcal{H}$ with n functions such that for any function $f \in \mathcal{H}$ there exists some function $f' \in \mathcal{H}_N$ such that $\|f - f'\|_\infty \le \epsilon/2$ then if we run the above algorithm with $\mathcal{H}_N$ and $\epsilon/2$ (instead of ϵ) then from $\Delta(\mathcal{G}, \mathcal{H}) \le D$ it follows that $\Delta(\mathcal{G}, \mathcal{H}_N) \le D + \epsilon/2$. Therefore, by the above theorem, the pairwise elimination based agnostic learner with $\mathcal{H}_N$ will be $2D + \epsilon$ accurate and will pass at most $O\left(\frac{K^2N^2}{\epsilon^2} \log \frac{N}{\delta}\right)$ times, outside of an event of probability at most δ. Therefore, it is a (D, r, ϵ, δ) agnostic KWIK-learner for $(\mathcal{H}, \mathcal{G})$. In the case of a linear hypothesis class with a d-dimensional input, $N = \Theta((1/\epsilon)^d)$ and thus the number of passes in the bound scales with $(1/\epsilon)^{2d+2}$. We see that as compared to the noise-free case, we lose a factor of two in the exponent. The approach just described is general, but the complexity explodes with the dimension. It remains to be seen if there exists alternative, more efficient algorithms.

5. Reinforcement learning with KWIK-learning

In reinforcement learning an agent is interacting with its environment by executing actions and observing states and rewards of the environment, the goal being to collect as much reward as possible. Here we consider the case when the state transitions are Markovian. An agent unfamiliar with its environment must spend some time exploring the environment, or it may miss essential information and loose a lot reward in the long run. However, with time the agent must reduce exploration, or it will fail to collect reward. The basic question is how to find the right balance between exploration and exploitation.

The KWIK-Rmax construction of Li et al. (2011a) shows that if efficient KWIK-learning is possible for some environment models then efficient reinforcement learning is possible for the same class. The purpose of this section is to show that this result readily extends to the agnostic case in a sensible manner, justifying the choice of the agnostic learning model proposed.

5.1. Markovian Decision Processes and efficient learning agents

In this section we introduce a minimal formal framework for studying the efficiency of reinforcement learning algorithms when they are used to learn to control Markovian Decision Processes (MDPs). For further information on learning in MDPs the reader is referred to the book of Szepesvári (2010) and the references therein.

Technically, an MDP M is a triple $(S, A, \mathcal{P})$, where S is the set of states, A is the set of actions, both are non-empty, Borel-spaces; and $\mathcal{P}$, determining the evolution of the decision process, is a transition probability map from $S \times A$ to $S \times \mathbb{R}$.[6] In particular, an agent interacting with an environment described by an MDP receives at time step the state $s_t \in S$ of the environment, decides about the action $a_t \in A$ to take based on the information available to it and then executes the action in the environment. As a result the environment moves the state and generates the reward associated with the transition: $(r_t, s_{t+1}) \sim P(\cdot|s_t, a_t)$. The process then repeats. An algorithm (which may use randomness) for computing an action based on past information is called a (non-stationary) policy. The value of a policy π in state s is the expected total discounted reward, or expected return when the interaction starts at state s. Formally,

$$V^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, (r_t, s_{t+1}) \sim P(\cdot|s_t, a_t), a_t \sim \pi(\cdot|s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_t), t \geq 0, s_0 = s\right].$$

An *optimal policy* π^* is such that in every state $s \in S$, $V^{\pi^*}(s)$ is the best possible value: $V^{\pi^*}(s) = V^*(s) \stackrel{\text{def}}{=} \sup_\pi V^\pi(s)$. Here, V^* is the so-called *optimal value function* and V^π is

6. We call $\mathcal{P}$ a transition probability map between measurable spaces $(E_1, \mathcal{E}_1)$ and $(E_2, \mathcal{E}_2)$ if *(i)* $\mathcal{P}(\cdot|e_2)$ is a probability measure on $(E_1, \mathcal{E}_1)$ for any $e_2 \in E_2$ and *(ii)* the function $\mathcal{P}(B|\cdot)$ is measurable on E_2 for any $B \in \mathcal{E}_1$. In what follows, to minimize clutter, we omit technical assumptions needed to establish e.g. the existence of measurable optimal stationary policies. See e.g. Theorem 6.11.11 in the book by Puterman (2005) for a compact result and the references in this book. All results presented hold for finite MDPs without any further assumptions.

called the *value function* of policy π. In finite discounted MDPs an optimal policy always exists. We also need the *optimal action-value function* $Q^* : S \times A \to \mathbb{R}$; $Q^*(s,a)$ is defined as the total expected discounted reward assuming that the interaction starts at state s, the first action is a and in the subsequent timesteps an optimal policy is followed. We will write V_M^π, V_M^*, Q_M^* when there is a need to emphasize that these object are specific to the MDP M.

A learning agent's goal is to act "near-optimally" in every finite MDP while having little *a priori* information about the MDP. In particular, given a finite MDP $M = (S, A, \mathcal{P})$, the learning agent $\mathcal{A}$ is told S, A, a bound on the rewards and their expected value, a discount factor $0 < \gamma < 1$. Then, $\mathcal{A}$ starts interacting with an environment described by M. Note that a learning agent is slightly more general then a policy: A policy is MDP-specific (by definition), while a learning agent must be able to act in any (finite) MDP. We can also view learning as the learning agent *choosing* an appropriate (non-stationary) policy to follow based on the *a priori* information received. We also identify an agent $\mathcal{A}$ with its "learning algorithm" and will also say "algorithm $\mathcal{A}$".

One possible goal for a learning agent is to minimize the number of timesteps when the future value collected from the state just visited is worse than the optimal value less some value $\epsilon > 0$. Following Kakade (2003) and Strehl and Littman (2005), we formalize this as follows:

Definition 5.1 (ϵ-mistake count) *Let $\epsilon > 0$ be a prescribed accuracy. Assume that a learning agent $\mathcal{A}$ interacts with an MDP M and let $(s_t, a_t, r_t)_{t \geq 0}$ be the resulting $S \times A \times \mathbb{R}$-valued stochastic process. Define the expected future return of $\mathcal{A}$ at time step $t \geq 0$ as*

$$V_{t,M}^{\mathcal{A}} = \mathbb{E}\left[\sum_{s=0}^{\infty} \gamma^s r_{t+s} \,\middle|\, s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_t \right].$$

Agent $\mathcal{A}$ is said to make an ϵ-mistake *at time step if $V_{t,M}^{\mathcal{A}} < V_M^*(s_t) - \epsilon$ and we will use $N_{M,\epsilon}^{\mathcal{A}}$ to denote the number of* ϵ-mistakes *agent $\mathcal{A}$ makes in M:*

$$N_{M,\epsilon}^{\mathcal{A}} = \sum_{t=0}^{\infty} \mathbb{I}_{\{V_{t,M}^{\mathcal{A}} < V_M^*(s_t) - \epsilon\}}.$$

The competence of an algorithm $\mathcal{A}$ is measured by $N_{M,\epsilon}^{\mathcal{A}}$.[7] Formally, an algorithm $\mathcal{A}$ is called *PAC-MDP* if for any $\epsilon > 0$, $0 < \delta \leq 1$, MDP M, $N_{M,\epsilon}^{\mathcal{A}}$ can be bounded with probability $1 - \delta$ with a polynomial of the form $\mathrm{poly}(|S|, |A|, 1/\epsilon, \log(1/\delta), 1/(1-\gamma))$, assuming that the rewards belong to $[0,1]$ interval. For MDPs with infinite state and action spaces, $|S|$ and $|A|$ should be replaced by an appropriate "complexity" measure.

7. An alternative, closely related way for measuring competence is to count the number of timesteps when $Q_M^*(x_t, a_t) < V_M^*(x_t) - \epsilon$: $\hat{N}_{M,\epsilon}^{\mathcal{A}} = \sum_{t=0}^{\infty} \mathbb{I}_{\{Q_M^*(s_t,a_t) < V_M^*(s_t) - \epsilon\}}$. If the learning agent does not randomize, $Q_M^*(s_t, a_t) > V_{t,M}^{\mathcal{A}}$ follows from the definitions. It follows then that $N_{M,\epsilon}^{\mathcal{A}} \leq \hat{N}_{M,\epsilon}^{\mathcal{A}}$ holds almost surely. For further, alternative notions of efficient learning consult Fiechter (1994); Auer et al. (2008).

From a static, non-learning, computational viewpoint, the stochasticity of the rewards does not play a role. Hence, by slightly abusing notation, we will also call a 4-tuple (S, A, P, R) an MDP, where P is a transition probability map from $S \times A$ to S, and $R : S \times A \to \mathbb{R}$ is the immediate expected reward function underlying $\mathcal{P}$. A policy π which chooses the actions based on the last state only in a fixed manner is called a *stationary (Markov) policy*.[8] Such a policy can and will be identified with a transition probability map from S to A. In particular, we will use $\pi(\cdot|s)$ to denote the probability distribution over A designated by π at $s \in S$.

5.2. A general theorem on efficient RL

In this section we show a general result on constructing efficient reinforcement learning algorithms. The result states that if a policy *i*) keeps track whether state-action pairs are "known", *ii*) ensures that the model parameters corresponding to "known" pairs are indeed known with reasonable accuracy, *iii*) makes an effort to get to unknown areas, then it will be efficient as soon as the number of times it happens that the currently visited state-action pair is not "known" is small.

Theorems of this style have appeared in Strehl et al. (2006, 2009); Li (2009); Li et al. (2011b,a). These results in fact generalize the proof of Kakade (2003) which shows that Rmax is efficient. The closest in spirit to the result below is Theorem 10 by Strehl et al. 2009 (originally appeared as Proposition 1 of Strehl et al. (2006) and then repeated as Theorem 4 by Li et al. 2011b). The differences between the result below and this theorem are mostly at the cosmetic level, although the particular form below makes our theorem particularly easy to apply to the model-based setting of the next section and it also allows imperfect (even stochastic) planners.[9] However, the main reason we included this result is to give a fully rigorous proof, which avoids the inaccuracies and ambiguities of previous proofs.[10] Also, we slightly improve the previous bounds: we remove a $1/(1-\gamma)$-factor.

Let $\mathcal{Y} = M(S) \times \mathbb{R}$, where $M(S)$ is the space of finite signed-measures. We define the norm $\|\cdot\|_{\mathcal{Y}} : \mathcal{Y} \to \mathbb{R}_+$ as follows: For any $P \in M(S)$, $r \in \mathbb{R}$, let $\|(P, r)\|_{\mathcal{Y}} \stackrel{\text{def}}{=} |r| + \gamma \|P\|_{TV} V_{\max}$, where $V_{\max} > 0$ will be chosen to be a common upper bound on the value functions in a class of MDPs of interest and $\|\cdot\|_{TV}$ is the total-variation norm of finite signed measures ($\|P\|_{TV} = |P|(S)$). Define the distance of MDPs $M_1 = (S, A, P_1, R_1)$, $M_2 = (S, A, P_2, R_2)$ sharing S and A by

$$d(M_1, M_2) \stackrel{\text{def}}{=} \sup_{(s,a) \in S \times A} \|(P_1(\cdot|s,a) - P_2(\cdot|s,a), r_1(s,a) - r_2(s,a))\|_{\mathcal{Y}} .$$

8. A policy is Markov if the distribution assigned to a history depends only on the last state of the history. In what follows, by a stationary policy we will always mean a stationary Markov policy.
9. Imperfect planners are also allowed in Walsh et al. (2010).
10. Some of the differences in our proof follow from our slightly changed assumptions. However, the proof of Lemma 5.3 is new. In place of our argument, earlier works either did not present a proof or used an argument which did not allow serial correlations and thus, strictly speaking, cannot be applied in the setting considered here.

Theorem 5.1 *Consider agent $\mathcal{A}$ interacting with an MDP $M = (S, A, \mathcal{P})$. Let $(s_1, a_1, r_1, s_2, a_2, r_2, \ldots)$ be the trajectory that results when $\mathcal{A}$ interacts with M and let $\mathcal{F}_t = \sigma(s_1, a_1, r_1, \ldots, s_t)$ be the σ-field corresponding to the history at time t. Let $G \subset \Omega$ be a measurable event of the probability space that holds the random variables $(s_t, a_t, r_t)_{t\geq 1}$. Assume that there exist a sequence of state-action sets $(K_t)_{t\geq 1}$, $K_t \subseteq S \times A$, a sequence of models $(\hat{M}_t)_{t\geq 1}$, and a sequence of stationary policies $(\pi_t)_{t\geq 1}$ such that for some $e_{\text{plan}}, e_{\text{model}}, V_{\max} > 0$ and for any $t \geq 1$ the following hold:*

(a) the expected immediate rewards underlying M and $(\hat{M}_t)_{t\geq 1}$ are bounded by $(1-\gamma)V_{\max}$;

(b) K_t, $\hat{M}_t$, π_t are $\mathcal{F}_t$-measurable;

(c) $a_t = \pi_t(s_t)$ (the action at time t is selected by π_t);

(d) $V^{\pi_t}_{\hat{M}_t}(s_t) \geq V^*_{\hat{M}_t}(s_t) - e_{\text{plan}}$ (the policy π_t is at least e_{plan}-optimal at s_t in model $\hat{M}_t$);

(e) for any $(s,a) \in K_t$, $\left\|\hat{M}_t(s,a) - M(s,a)\right\|_{\mathcal{Y}} \leq e_{\text{model}}$ holds true on G (the model is e_{model}-accurate for "known" state-action pairs);

(f) for any $(s,a) \notin K_t$, any stationary policy π, $V_{\max} \leq Q^{\pi}_{\hat{M}_t}(s,a)$ (in "unknown" states, the model is optimistic, but not overly so); and

(g) if $(s_t, a_t) \in K_t$ then $\pi_{t+1}(s_{t+1}) = \pi_t(s_{t+1})$ (old policy used as long as visiting known state-action pairs).

Let B be a deterministic *upper bound on the number of times it happens that $(s_t, a_t) \notin K_t$ on G: $\sum_{t=1}^{\infty} \mathbb{I}_{\{(s_t,a_t)\notin K_t, G\}} \leq B\mathbb{I}_{\{G\}}$. Then, for any $0 < \delta \leq 1$ there exists an event $F = F_\delta$ such that $\mathbb{P}(F_\delta) \geq 1 - \delta$ and on $F \cap G$, the number of $5e_{\text{model}}/(1-\gamma) + e_{\text{plan}}$-mistakes is bounded by $\frac{2V_{\max}(1-\gamma)L}{e_{\text{model}}}\left\{B + (\sqrt{2B}+3)\sqrt{\log\left(\frac{L}{\delta}\right)} + 6\log\left(\frac{L}{\delta}\right)\right\}$, where $L = \max(1, \lceil (1-\gamma)^{-1}\log(V_{\max}(1-\gamma)/e_{\text{model}})\rceil)$.*

We will need two lemmas for the proof, which we state first. The first lemma is a standard result which follows from simple contraction arguments:

Lemma 5.2 (Simulation Lemma) *For any two MDPs M_1 and M_2 sharing the same state-action set $S\times A$ and any stationary policy π over (S, A), it holds that $\left\|V^{\pi}_{M_1} - V^{\pi}_{M_2}\right\|_\infty \leq d(M_1, M_2)/(1-\gamma)$.*

The next lemma compares the number of times the bias underlying an infinite sequence of coin flips can exceed a certain threshold ϵ given that the biases are sequentially chosen and that the number of heads in the infinite coin flip sequence assumes some bound m. The proof is based on the idea underlying Markov's inequality and a stopping time construction used in conjunction with a Bernstein-like inequality due to Freedman (1975).

Lemma 5.3 *Let $0 < \epsilon < 1$, $m \in \mathbb{N}$ be deterministic constants, $(\mathcal{F}_t)_{t\geq 1}$ be some filtration and let $(A_t)_{t\geq 1}$ be an $(\mathcal{F}_{t+1})_{t\geq 1}$-adapted sequence of indicator variables. Let*

$$a_t = \mathbb{E}\left[A_t|\mathcal{F}_t\right]$$

and let G be an event such that on G the inequality $\sum_{t=1}^{\infty} A_t \leq m$ holds almost surely. Then, for any $0 < \delta \leq 1$ with probability $1-\delta$, either G^c holds, or

$$\sum_{t=1}^{\infty} \mathbb{I}_{\{a_t \geq \epsilon\}} \leq \frac{1}{\epsilon}\left\{ m + \sqrt{2m\log\left(\tfrac{1}{\delta}\right)} + 3\sqrt{\log\left(\tfrac{1}{\delta}\right)} + 6\log\left(\tfrac{1}{\delta}\right)\right\}.$$

With this, we are ready to give the proof of Theorem 5.1:

Proof We need to show that with an appropriate choice of F, and for $\epsilon \stackrel{\text{def}}{=} 5e_{\text{model}}/(1-\gamma) + e_{\text{plan}}$, the inequality $N^{\mathcal{A}}_{M,\epsilon} \leq \frac{2V_{\max}(1-\gamma)L}{e_{\text{model}}}\left\{B + (\sqrt{2B}+3)\sqrt{\log\left(\frac{L}{\delta}\right)} + 6\log\left(\frac{L}{\delta}\right)\right\}$ holds on $F \cap G$.

Let $e_{\text{trunc}} \stackrel{\text{def}}{=} e_{\text{model}}/(1-\gamma)$ be the allowed truncation-error. Note that with this notation, $L = \max(1, \lceil \frac{1}{1-\gamma}\log\frac{V_{\max}}{e_{\text{trunc}}}\rceil)$. The quantity L is known as the so-called e_{trunc}-horizon: if $V^{\pi}_M(s;L)$ denotes the expected L-step return of π in M when the decision process starts at s then $|V^{\pi}_M(s) - V^{\pi}_M(s;L)| \leq e_{\text{trunc}}$.

Fix $t \geq 1$ and let E_t be the event that in the next L steps, the agent "escapes" the known set:

$$E_t = \bigcup_{i=0}^{L-1} \left\{(s_{t+i}, a_{t+i}) \notin K_{t+i}\right\}. \tag{2}$$

The plan for the proof is as follows: We first show that whenever $p_t = \mathbb{P}\left(E_t|\mathcal{F}_t\right)$ is small, in particular, when

$$p_t \leq \frac{e_{\text{trunc}}}{2V_{\max}} \tag{3}$$

then, on G, the agent does not make an ϵ-mistake at time t. Then, based on Lemma 5.3 we will give a high-probability bound on the number of timesteps when (3) fails to hold.

Turning to the first step, assume that (3) holds. We want to show that the agent does not make an ϵ-mistake on G, i.e., on this event, $V^{\mathcal{A}}_{t,M} \geq V^*_M(s_t) - \epsilon$. Let $\tilde{\pi}_t$ be the non-stationary policy of M induced by $\mathcal{A}$ at time step t. Then, $V^{\mathcal{A}}_{t,M} = V^{\tilde{\pi}_t}_M(s_t) \geq V^{\tilde{\pi}_t}_M(s_t;L) - e_{\text{trunc}}$ by the choice of L. Let $\bar{M}_t$ be the model which agrees with M on K_t, while it agrees with $\hat{M}_t$ outside of K_t.

Claim 5.4 *We have $V^{\tilde{\pi}_t}_M(s_t;L) \geq V^{\pi_t}_{\bar{M}_t}(s_t;L) - 2V_{\max}p_t$, with probability one.*

The proof, which is given in the appendix, uses that the immediate rewards for both M and $\hat{M}_t$ are bounded by $(1-\gamma)V_{\max}$ (i.e., condition (a)), the measurability condition (b) and the condition that there is no policy update while visiting known states (i.e., condition (g)).

Now, by the definition of L, $V^{\pi_t}_{\tilde{M}_t}(s_t;L) \geq V^{\pi_t}_{\tilde{M}_t}(s_t) - e_{\text{trunc}}$, while by the Simulation Lemma (Lemma 5.2) and (e), on G, it holds that $V^{\pi_t}_{\tilde{M}_t}(s_t) \geq V^{\pi_t}_{\hat{M}_t}(s_t) - e_{\text{model}}/(1-\gamma)$. By (d), $V^{\pi_t}_{\hat{M}_t}(s_t) \geq V^{*}_{\hat{M}_t}(s_t) - e_{\text{plan}}$. Let π^* be an optimal stationary policy *in* M. We have $V^{*}_{\hat{M}_t}(s_t) \geq V^{\pi^*}_{\hat{M}_t}(s_t)$.

Now, let $\tilde{M}_t$ be the MDP which is identical to $\hat{M}_t$ for state-action pairs in K_t, while outside of K_t it is identical to M. We claim that the following holds true:

Claim 5.5 *On G, it holds that $V^{\pi^*}_{\hat{M}_t}(s_t) \geq V^{\pi^*}_{\tilde{M}_t}(s_t)$.*

The proof, which is again given in the appendix, uses the optimism condition (condition (f)), condition (a).

By the Simulation Lemma and (e), on G, it holds that $V^{\pi^*}_{\tilde{M}_t}(s_t) \geq V^{\pi^*}_{M}(s_t) - e_{\text{model}}/(1-\gamma)$. Chaining the inequalities obtained, we get that, on G, the inequality $V^{\mathcal{A}}_{t,M} \geq V^{\pi^*}_{M}(s_t) - (2p_t V_{\max} + 4e_{\text{trunc}} + e_{\text{plan}})$ holds. Thus, when (3) holds, on G, we also have $V^{\mathcal{A}}_{t,M} \geq V^{\pi^*}_{M}(s_t) - (5e_{\text{trunc}} + e_{\text{plan}}) = V^{\pi^*}_{M}(s_t) - \epsilon$, which concludes the proof of the first step.

Let us now turn to the second step of the proof. By the first step, on G, $\sum_{t=1}^{\infty} \mathbb{I}_{\{V^{\mathcal{A}}_{t,M} < V^{*}_{M} - \epsilon\}} \leq \sum_{t=1}^{\infty} \mathbb{I}_{\{p_t < \frac{e_{\text{trunc}}}{1-\gamma}\}}$. Let $T_{\text{non-opt}} = \sum_{t=1}^{\infty} \mathbb{I}_{\{p_t > e_{\text{trunc}}/(2V_{\max})\}}$. In order to bound $T_{\text{non-opt}}$, we write it as the sum of L terms as follows

$$T_{\text{non-opt}} = \sum_{i=0}^{L-1} \underbrace{\sum_{j=0}^{\infty} \mathbb{I}_{\{p_{jL+i+1} > e_{\text{trunc}}/(2V_{\max})\}}}_{T^{(i)}_{\text{non-opt}}} . \tag{4}$$

We will apply Lemma 5.3 to each of the resulting L terms separately. To do so, fix $0 \leq i \leq L-1$ and choose the sequence of random variables $(A_t)_{t\geq 0}$ to be $(\mathbb{I}_{\{E_{tL+i+1}\}})_{t\geq 0}$, while let the corresponding sequence of σ-fields be $(\mathcal{F}_{tL+i+1})_{t\geq 0}$. Further, choose ϵ of Lemma 5.3 as $\epsilon = e_{\text{trunc}}/(2V_{\max})$. By condition (b), A_t is $\mathcal{F}_{(t+1)L+i+1}$-measurable, since $E_{tL+i+1} \in \mathcal{F}_{(t+1)L+i+1}$. The upper bound m on the sum $\sum_t A_t$ is obtained from

$$\sum_{j=0}^{\infty} \mathbb{I}_{\{E_{jL+i+1}\}} \leq \sum_{t=1}^{\infty} \mathbb{I}_{\{E_t\}} \leq \sum_{t=1}^{\infty} \mathbb{I}_{\{(x_t,a_t)\notin K_t\}} ,$$

where the last inequality follows from the definition of E_t (cf. (2)). By assumption, on G, the last expression is bounded by B. Therefore, on G, $\sum_{j=0}^{\infty} \mathbb{I}_{\{E_{jL+i+1}\}} \leq B$ also holds and Lemma 5.3 gives that with probability $1-\delta/L$, either G^c holds or

$$T^{(i)}_{\text{non-opt}} \leq \frac{2V_{\max}}{e_{\text{trunc}}} \left\{ B + \sqrt{2B \log\left(\tfrac{L}{\delta}\right)} + 3\sqrt{\log\left(\tfrac{L}{\delta}\right)} + 6\log\left(\tfrac{L}{\delta}\right) \right\} .$$

Combining this with (4) gives that, with probability $1-\delta$, either G^c holds or

$$T_{\text{non-opt}} \leq \frac{2V_{\max} L}{e_{\text{trunc}}} \left\{ B + (\sqrt{2B}+3)\sqrt{\log\left(\tfrac{L}{\delta}\right)} + 6\log\left(\tfrac{L}{\delta}\right) \right\} ,$$

thus, finishing the proof. ■

5.3. The KWIK-Rmax construction

In this section we consider the KWIK-Rmax algorithm of Li et al. (2011a) (see, also Li et al. (2011b)). This algorithm is identical to the Rmax algorithm of Brafman and Tennenholtz (2000), except that the model learning and planning components of Rmax are replaced by general components. This way one gets a whole family of algorithms, depending on what model learner and planner is used. In addition to unifying a large number of previous works which considered different model learners and planners (for a list of these, see the introduction of this article), this allowed Li et al. to improve the previously known efficiency bounds, too (see, Chapter 7 of Li (2009) and Li et al. (2011a)).

Here, the exact same algorithm is considered, but we derive a more general result: We show that if the model learning algorithm is an *agnostic* KWIK-learner enjoying some KWIK-bound and a "good" planner is used then the resulting instance of KWIK-Rmax will be efficient even when the MDP considered is outside of the hypothesis class that the KWIK-learner uses. In essence, our analysis shows how approximation errors propagate in a reinforcement learning context. In the special case of realizable learning, our result reproduces the result of Li et al. (2011a) (our bound is slightly better in terms of its dependence on the discount factor γ).

The KWIK-Rmax algorithm is shown as Algorithm 5. The algorithm takes as input two "objects", a learner (MDPLearner) and a planner (Planner). The learner's job is to learn an approximation to the MDP that KWIK-Rmax interacts with. The learned model is fed to the planner. The planner is assumed to interact with models by querying next state distributions and immediate rewards at select certain state-action pairs. The **predict** method of a model is assumed to return the returned values for the planner. The planner itself could use these in many ways – the details of the planning mechanism are not of our concern here (our result allows for both deterministic and stochastic planners). An important aspect of the algorithm is that the model learned is not actually fed directly to the planner, but it is fed to a wrapper. In fact, since a KWIK learner might pass in any round, it is the job of the wrapper to produce a next-state distribution, reward pair in all cases. This can be done in many different ways. However, the main idea here is to return a next-state distribution and a reward, which makes an "unknown" state-action pair highly desirable. One implementation of this is shown in the right-hand side of algorithm listing 5. The KWIK-Rmax algorithm itself repeatedly calls the planner (with the optimistically wrapped model learned by the KWIK learner), executes the returned action and upon observing the next state and reward, if the KWIK learner passed, it feeds the learner with the observed values. Note that for the analysis it is critical that the learner is not fed with information when it did not pass, contradicting one's intuition.

Our main result, stated below, says that if MDPLearner is an agnostic KWIK-learner and the planner is near-optimal then KWIK-Rmax will be an efficient RL algorithm. In

Algorithm 5 The KWIK-Rmax Algorithm and the Optimistic Wrapper

KWIK-Rmax(MDPLearner, Planner)
MDPLearner.initialize(...)
Planner.initialize(...)
Observe s_1
for $t := 1, 2, \ldots$ **do**
  a_t = Planner.plan(Opt(MDPLearner), s_t)
  Execute a_t and observe s_{t+1}, r_t
  if MDPLearner.predict$(s_t, a_t) = \perp$ **then**
    MDPLearner.learn$((s_t, a_t), (\delta_{s_{t+1}}, r_t))$

{Optimistic Wrapper}
Opt(MDPLearner)**.predict**(s, a)
if MDPLearner.predict$(s, a) = \perp$ **then**
  return $(\delta_s(\cdot), (1-\gamma)V_{\max})$
else
  return MDPLearner.predict(s, a)

order to state this result formally, first we need to define what we mean by KWIK-learning in the context of MDPs.

As before, we fix the set of states (S) and actions (A) and $V_{\max} > 0$, an upper bound on the value functions for the MDPs of interest. Remember that $\mathcal{Y} = M(S) \times \mathbb{R}$ and the norm $\|\cdot\|_{\mathcal{Y}}$ defined by $\|(P, r)\|_{\mathcal{Y}} = |r| + \gamma \|P\|_{TV} V_{\max}$ ($P \in M(S)$, $r \in \mathbb{R}$). The space $\mathcal{Y}$ will be the output space for the predictors. Learning an MDP model means learning the immediate expected rewards and transition probabilities when the inputs are state-action pairs. That is, we let $\mathcal{X} = S \times A$ and encode an MDP as a mapping $g : \mathcal{X} \to \mathcal{Y}$, where for $x = (s, a) \in \mathcal{X}$, $g(s, a) = (P, r)$, where P is the next-state distribution over S and $r \in \mathbb{R}$ is the associated expected immediate reward. What is left in specifying an *MDP problem instance* $(\mathcal{X}, \mathcal{Y}, g, Z, \|\cdot\|)$ is the noise component Z. In the $\mathbb{R}^S$ component of $\mathcal{Y}$, the noise is completely determined by g, while, for the reward component the noise distribution is arbitrary, except that the noisy reward is restricted to be bounded by $(1-\gamma)V_{\max}$. A subset of all $g : \mathcal{X} \to \mathcal{Y}$ functions will be called an MDP hypothesis space. Now, we are ready to state our main result.

Theorem 5.6 *Fix a state space S and an action space A, which are assumed to be non-empty Borel spaces. Let $\mathcal{X}, \mathcal{Y}$ be as described above, $\mathcal{H}$ be an MDP hypothesis set, $\mathcal{G}$ be a set of MDP problem instances, both over $\mathcal{X}, \mathcal{Y}$. Assume that $\Delta(\mathcal{G}, \mathcal{H}) \le D$. Assume that $V_{\max} > 0$ is such that $(1-\gamma)V_{\max}$ is an upper bound on the immediate rewards of the MDPs determined by members of $\mathcal{H}$ and $\mathcal{G}$. Fix $\epsilon > 0, r \ge 1$, $0 < \delta \le 1/2$. Assume that* MDPLearner *is an agnostic (D, r, ϵ) KWIK-learner for $(\mathcal{H}, \mathcal{G})$ with KWIK-bound $B(\delta)$. Assume further that we are given a* Planner *which is $e_{planner}$-accurate. Consider the instance of the* KWIK-Rmax *algorithm which uses* MDPLearner *and* Planner*, interacting with some MDP M from $\mathcal{G}$. Let $\epsilon' = \frac{5(rD+\epsilon)}{1-\gamma} + e_{planner}$. Then, with probability $1 - 2\delta$, the number of ϵ'-mistakes, $N_{M,\epsilon'}$, made by* KWIK-Rmax *is bounded by* $\frac{2V_{\max}(1-\gamma)L}{rD+\epsilon}\left\{B(\delta) + (\sqrt{2B(\delta)}+3)\sqrt{\log\left(\frac{L}{\delta}\right)} + 6\log\left(\frac{L}{\delta}\right)\right\}$, *where* $L = \max(1, \lceil (1-\gamma)^{-1}\log(V_{\max}(1-\gamma)/(rD+\epsilon))\rceil)$.

6. Discussion

In the first part of the paper we formalized and explored the agnostic KWIK framework (first mentioned in Li (2009)), and presented several simple agnostic KWIK learning algorithms for finite hypothesis classes with and without noise, and for deterministic linear hypothesis classes. In the second part of the paper we showed that an agnostic KWIK-learner leads to an efficient reinforcement learning, even when the environment is outside of the hypothesis class that the KWIK-learner uses. To our knowledge, this is the first result that proves any kind of efficiency for an RL algorithm in the agnostic setting. Our bound also saves a factor of $1/(1-\gamma)$ compared to previous bounds. Unfortunately, our (limited) exploration of agnostic KWIK-learning indicated that efficient agnostic KWIK-learning might be impossible for some of the most interesting (simple) hypothesis classes. These negative results do not imply that efficient agnostic reinforcement learning is impossible, but indicate that the problem itself requires further work.

Acknowledgments

This work was supported in part by AICML, AITF (formerly iCore and AIF), NSERC and the PASCAL2 Network of Excellence under EC grant no. 216886. We thank Yaoliang Yu for his careful reading of some parts of this paper, in particular for his suggestions that led to a simpler proof of Lemma A.1.

Appendix A. Proofs

A.1. Proofs for Section 3

Theorem 3.1 *Let $(\mathcal{X},\mathcal{Y})$ be arbitrary sets, $r=2$, $\epsilon=0$, $D>0$, $\mathcal{H}$ a finite hypothesis class over $(\mathcal{X},\mathcal{Y})$, $\mathcal{G}$ a deterministic problem class over $(\mathcal{X},\mathcal{Y})$ with $\Delta(\mathcal{G},\mathcal{H})\le D$. Then, the Generic Agnostic Learner is an agnostic (D,r,ϵ) KWIK-learner for $(\mathcal{H},\mathcal{G})$ with KWIK-bound $|\mathcal{H}|-1$.*

Proof Suppose that the adversary chose problem $G=(\mathcal{X},\mathcal{Y},g,0)$. Since $\Delta(\mathcal{G},\mathcal{H})\le D$ and $\mathcal{H}$ is finite, there exists a function $f^*\in\mathcal{H}$ such that $\|f^*-g\|_\infty=\Delta(G,\mathcal{H})\le D$. Consequently, f^* will be never excluded from $\mathcal{F}$: for any x and $y=g(x)$, $\|f^*(x)-y\|\le D$.

Let us now show that every time the learner makes a prediction, the prediction is $2D$-accurate. Indeed, if $\hat{y}$ is the prediction of the learner, $\|\hat{y}-y\|\le\|\hat{y}-f^*(x)\|+\|f^*(x)-y\|$. Now, by the definition of $\hat{y}$ and because f^* is never excluded, $\|\hat{y}-f^*(x)\|\le D$. Further, by the choice of f^* and because $y=g(x)$, we also have $\|f^*(x)-y\|\le D$, altogether showing that the prediction error is upper bounded by $2D$.

It remains to show that the learner cannot pass more than $|\mathcal{H}|-1$ times. This follows, because after each pass, the learner excludes *at least one* hypothesis. To see why note that $Y=\emptyset$ means that for every $y'\in\mathcal{Y}$ there is some $f\in\mathcal{F}$ such that $y'\notin B_D(f(x))$.

Specifically, this also holds for $y = g(x)$ and thus there is a function $f \in \mathcal{F}$ such that $\|f(x) - y\| > D$. By definition, this function will be eliminated in the update. As the learner eliminates at least one hypothesis from $\mathcal{F}$ when passing, and f^* is never eliminated from $\mathcal{F}$, the number of passes is at most $|\mathcal{H}| - 1$. ■

Theorem 3.2 *Fix any $D > 0$ and an infinite domain $\mathcal{X}$. Then, there exists a finite response set $\mathcal{Y} \subset \mathbb{R}$, a two-element hypothesis class $\mathcal{H}$ and a deterministic problem class $\mathcal{G}$, both over $(\mathcal{X}, \mathcal{Y})$, that satisfy $\Delta(\mathcal{G}, \mathcal{H}) \le D$ such that there is no bounded agnostic $(D, r, 0)$ KWIK-learner for $(\mathcal{H}, \mathcal{G})$ with competitiveness factor $0 \le r < 2$.*

Proof Fix $D > 0$. Let $\mathcal{X}$ be an infinite set, $x_1, x_2, \dots$ be a sequence of distinct elements in $\mathcal{X}$, $\mathcal{Y} = \{-2D, -D, 0, D, 2D\}$ and $\mathcal{H}$ be a two-element set containing $f_{+D} \equiv D$ and $f_{-D} \equiv -D$. For $n \in \mathbb{N}$, define the functions

$$g_{n,\pm D}(x) = \begin{cases} \pm 2D, & \text{if } x = x_n\,; \\ 0, & \text{otherwise} \end{cases}$$

and let $\mathcal{G}$ be the set of these functions. Clearly, $\Delta(\mathcal{G}, \mathcal{H}) = D$. Before picking a hypothesis, the adversary simulates its interaction with the learner. At step t of the simulation, the adversary asks query x_t. If the learner passes with probability 1, then A answers 0. Suppose now that there is some t when the learner makes some prediction $\hat{y}_t$ with probability $p > 0$. Without loss of generality we may assume that $\mathbb{P}(\hat{y}_t \ge 0) \ge p/2$. At this point, the adversary stops the simulation and picks $g_{t,-D}$. During the learning process, L passes to any x_i with $i < t$ and gets feedback 0. At time step t, however, with probability $p/2$,

$$\hat{y}_t - g_{t,-D}(x_t) \ge 0 + 2D,$$

so the learner fails. On the other hand, if the learner always passes then it is not bounded. ■

Theorem 3.3 *Let $X_{\max} > 0$, $\mathcal{X} = [-X_{\max}, X_{\max}]^d$, $\mathcal{Y} = \mathbb{R}$, $M, D, \epsilon > 0$, $r = 2$, $\mathcal{H} = \mathcal{H}_{lin(M)}$. Then, for any $\mathcal{G}$ deterministic problem class over $(\mathcal{X}, \mathcal{Y})$ with $\Delta(\mathcal{G}, \mathcal{H}) \le D$, it holds that the deterministic linear agnostic learner is an agnostic (D, r, ϵ) KWIK-learner for $(\mathcal{H}, \mathcal{G})$ with the KWIK-bound $2d!\left(\frac{MX_{\max}}{\epsilon} + 1\right)^d$.*

Proof First of all, we need to show that the calculations of the algorithm are meaningful. Specifically, solutions y^+ and y^- to the linear programs need to be finite. This will hold because $\Theta(C)$ is bounded and non-empty during any point of the learning. Boundedness holds because $\Theta(C) \subseteq \{\theta : \|\theta\|_\infty \le M\}$. We assumed that there is a hypothesis f_{θ^*} so that $\|f_{\theta^*} - g^*\|_\infty \le D$, and the KWIK protocol sends training samples (x, y) such that $y = g^*(x)$. Therefore, $D \ge |f_{\theta^*}(x) - g^*(x)| = |(\theta^*)^T x - y|$, so θ^* satisfies all constraints in C, making $\Theta(C)$ nonempty.

Secondly, the following calculation shows that if a prediction is made, it is correct:

$$\begin{aligned}\hat{y}_t - y_t &= y_t^+/2 + y_t^-/2 - g(x_t) \\ &\leq D + \epsilon + y_t^-/2 + y_t^-/2 - g(x_t) \\ &= D + \epsilon + (y_t^- - f^*(x_t)) + (f^*(x_t) - g(x_t)) \\ &\leq D + \epsilon + 0 + D\,,\end{aligned}$$

and similarly,

$$\begin{aligned}\hat{y}_t - y_t &= y_t^+/2 + y_t^-/2 - g(x_t) \\ &\geq y_t^+/2 + y_t^+/2 - (D+\epsilon) - g(x_t) \\ &= -D - \epsilon + (y_t^+ - f^*(x_t)) + (f^*(x_t) - g(x_t)) \\ &\geq -D - \epsilon + 0 - D\,,\end{aligned}$$

so $|\hat{y}_t - y_t| \leq 2D + \epsilon$, as required.

Finally, we prove the upper-bound on the number of $\perp$s. Let $\mathcal{X}' \stackrel{\text{def}}{=} [-X_{\max} - \epsilon/M, X_{\max} + \epsilon/M]^d$ be a slightly increased version of $\mathcal{X}$. Let H_k be the set of $x \in \mathcal{X}'$ for which the learner makes a prediction, that is, $y^+(x) - y^-(x) \leq 2D(1+\epsilon)$. The set H_k is convex, following from the convexity of the constraints and the convexity of the max operator (and the concavity of min). If the adversary asks some $x \notin H_k$, then x gets added to the known set, together with some of its neighborhood $B(x) \stackrel{\text{def}}{=} \{x' \in \mathcal{X}' : \|x' - x\|_1 \leq \epsilon/M\}$. This follows from the calculation below: let $x' \in B(x)$, then for any $\theta \in \Theta(C_{k+1})$,

$$\begin{aligned}&\theta^T x' \leq \theta^T x + \|\theta\|_\infty \|x' - x\|_1 \leq \theta^T x + \epsilon\,, \text{ so} \\ &\max_{\theta \in \Theta(C_{k+1})} \theta^T x' \leq \max_{\theta \in \Theta(C_{k+1})} \theta^T x + \epsilon\,, \text{ that is,} \\ &y^+(x') \leq y^+(x) + \epsilon\,.\end{aligned}$$

With similar reasoning, $y^-(x') \geq y^-(x) - \epsilon$. In step $k+1$, the constraint $y - D \leq \theta^T x \leq y + D$ was added for x, so $y^+(x) - y^-(x) \leq 2D$. Consequently, $y^+(x') - y^-(x') \leq 2(D+\epsilon)$, so the learner knows x'.

The set H_k is convex, $x \notin H_k$, and $B(x)$ is symmetric to x. So for any $x' \in B(x)$, at most one of x' and $x + (x - x')$ can be $\in B(x)$, that is, at least half of the volume of $B(x)$ is outside H_k. $H_k \cup B(x) \subseteq H_{k+1}$, so $\text{Vol}(H_{k+1}) \geq \text{Vol}(H_k) + \text{Vol}(B(x))/2$. The volume of $B(x)$ is $(2\epsilon/M)^d/d!$, so $\text{Vol}(H_k) \geq k\frac{(2\epsilon/M)^d}{2d!}$. On the other hand, $H_k \subseteq \mathcal{X}'$, so $\text{Vol}(H_k) \leq (2X_{\max} + 2\epsilon/M)^d$, which yields an upper bound on k:

$$k \leq 2d! \left(\frac{MX_{\max}}{\epsilon} + 1\right)^d .$$

■

Theorem 3.4 *Let $\mathcal{X}, \mathcal{Y}, D, r, \mathcal{H}$ be as in Theorem 3.3. Then, there exists a problem class $\mathcal{G}$ that satisfies $\Delta(\mathcal{G}, \mathcal{H}) \leq D$ such that there is no bounded, agnostic $(D, r, 0)$ KWIK-learner for $(\mathcal{H}, \mathcal{G})$.*

Proof For any learner L, we will construct a problem class $\mathcal{G}$, a strategy for the adversary so that L will either make a mistake larger than $2D$ with nonzero probability, or passes infinitely many times. The construction will be similar to the previous one.

Without loss of generality, let $\mathcal{X} = [0, 2]$ and let $\mathcal{Y} = \mathbb{R}$ with the absolute loss norm. Let $(x_1, x_2, \ldots)$ be a strictly increasing sequence of numbers with $1 < x_t < 2$, $t \in \mathbb{N}$. For convenience, define $x_0 = 1$. For $n \in \mathbb{N}$, let

$$g_{n,\pm}(x) \stackrel{\text{def}}{=} \begin{cases} \pm D(1 + \frac{x}{x_{n-1}}) & \text{if } x > x_{n-1} \ , \\ 0 & \text{if } x \le x_{n-1} \end{cases}$$

and let $\mathcal{G}$ be the set of these functions. For this set of problems, $\Delta(\mathcal{G}, \mathcal{H}) \le D$: for any concept, there exists a function $f \in \mathcal{H}_{\text{lin}(M)}$ that is at most at distance D from the chosen concept. Specifically, for $g_{n,+}$, f_θ with $\theta = D/x_{n-1}$ is satisfactory:

$$\begin{aligned} \left\| g_{n,+} - f_{D/x_{n-1}} \right\|_\infty &= \max \left\{ \max_{x \le x_{n-1}} |g_{n,+}(x) - f_{D/x_{n-1}}(x)|; \max_{x > x_{n-1}} |g_{n,+}(x) - f_{D/x_{n-1}}(x)| \right\} \\ &= \max \left\{ \max_{x \le x_{n-1}} |0 - \frac{D}{x_{n-1}} x|; \max_{x > x_{n-1}} |D(1 + \frac{x}{x_{n-1}}) - \frac{D}{x_{n-1}} x| \right\} = D \, . \end{aligned}$$

We prove the statement by contradiction. Assume that there exists a bounded, agnostic learner for the above problem. The adversary proceeds similarly to the adversary of Theorem 3.2: it finds out the first index t where the learner would make a prediction with nonzero probability (provided that it receives feedback 0 only). Unless the learner passes infinitely many times, such a t exists. If the first prediction is nonnegative with at least $1/2$ chance, the adversary picks $g_{t,-}$, otherwise it picks $g_{t,+}$. In the first case,

$$\hat{y}_t - g_{t,-}(x_t) \ge 0 + D(1 + \frac{x_n}{x_{n-1}}) > 2D,$$

with nonzero probability, and similarly, $\hat{y}_t - g_{t,-}(x_t) < -2D$ for the second case, showing that a bounded learner will make a mistake larger than $2D$ with positive probability. ■

Theorem 3.5 *Fix $\mathcal{X}, \mathcal{Y}, D, \mathcal{H}, \epsilon$ as in Theorem 3.3 and let $r \ge 2$. Then, there exists some problem class $\mathcal{G}$ so that any algorithm that agnostic (D, r, ϵ) KWIK learns $(\mathcal{H}, \mathcal{G})$ will pass at least 2^{d-1} times.*

Proof $\mathcal{X} = [-2, 2]$, $\mathcal{Y} = \mathbb{R}$ with the absolute loss norm, fix some $1 > D > 0$. The adversary asks the 2^{d-1} vertices of the hypercube $\{-1, +1\}^d$ that have positive first coordinates. It is easy to see that the adversary can pick the values on the vertices independently, and if the learner predicts anything with nonzero probability then the adversary can make the protocol fail. The proof is completely analogous to the previous one. ■

A.2. Proofs for Section 4

The following lemma, which follows from an application of the Hoeffding-Azuma inequality and a careful argumentation with "skipping processes", will be our basic tool. The novelty of the lemma is that we allow for the possibility of unbounded stopping times, otherwise the lemma would directly follow from Theorem 2.3 in Chapter VII of Doob (1953) and the Hoeffding-Azuma inequality. (It is very well possible that the lemma exists in the literature, however, we could not find it.)

Lemma A.1 *Let $\mathcal{F} = (\mathcal{F}_t)_{t\geq 1}$ be a filtration and let $(\epsilon_t, Z_t)_{t\geq 1}$ be a sequence of $\{0,1\}\times\mathbb{R}$-valued random variables such that ϵ_t is $\mathcal{F}_{t-1}$-measurable, Z_t is $\mathcal{F}_t$-measurable, $\mathbb{E}[Z_t|\mathcal{F}_{t-1}] = 0$ and $Z_t \in [A, A+K]$ for some deterministic quantities $A, K \in \mathbb{R}$. Let $m > 0$ and let $\tau = \min\{t \geq 1 : \sum_{s=1}^t \epsilon_s = m\}$, where we take $\tau = \infty$ when $\sum_{s=1}^\infty \epsilon_s < m$. Then, for any $0 < \delta \leq 1$, with probability $1-\delta$,*

$$\sum_{t=1}^{\tau} \epsilon_t Z_t \leq K\sqrt{\frac{m}{2}\log\left(\frac{1}{\delta}\right)}. \tag{5}$$

Remark A.1 (Analysis of the sum in (5)) *Let $(\Omega, \mathcal{A})$ be the probability space holding the random variables and the filtration $\mathcal{F}$, and let $S_n = \sum_{t=1}^n \epsilon_t Z_t$ ($n = 0, 1, \ldots$, the empty sum being zero). The sum S on the left-hand side of (5) is well-defined almost everywhere on Ω as it has at most m terms no matter whether $\tau(\omega) < \infty$ or $\tau(\omega) = \infty$. It also holds true that the sum S is an integrable random variable. To see why, consider $S'_\infty \stackrel{\text{def}}{=} \lim_{n\to\infty} S_{\tau\wedge n}$. We claim that the random variable S'_∞ is well-defined, integrable and $S'_\infty = S$ holds almost surely. First, notice that $(S_{\tau\wedge n})_{n\geq 1}$ is a martingale (this follows, e.g., from the Corollary on p.341 to Theorem 9.3.4 of Chung 2001). Next, note that $(S_{\tau\wedge n})_{n\geq 1}$ is L^1-bounded (by the condition on Z_t and because $S_{\tau\wedge n}$ has at most m terms, we have $\mathbb{E}[|S_{\tau\wedge n}|] \leq m(|A|+K)$) and therefore it is also uniformly integrable. Hence, Theorem 9.4.6 of Chung 2001 gives that S'_∞ is a well-defined, integrable random variable. Finally, a simple case analysis shows that $S'_\infty = S$ holds almost surely.*

Proof Let $S'_n = S_{\tau\wedge n}$, where $S_n = \sum_{t=1}^n \epsilon_t Z_t$, $n \geq 0$. Define $\mathbb{N}_\infty = \mathbb{N}\cup\{\infty\}$. By Remark A.1, $S = S'_\infty$ a.s., where S is the sum on the right-hand side of (5) and $S'_\infty = \lim_{n\to\infty} S_{\tau\wedge n}$, Theorem 9.4.6 of Chung 2001 mentioned in the remark not only gives that S'_∞ is integrable, but it also gives that $(S'_n, \mathcal{F}_n)_{n\in\mathbb{N}_\infty}$ is a martingale, i.e., S'_∞ is a "closure" of $(S'_n, \mathcal{F}_n)_{n\in\mathbb{N}}$. Let $\tau_i = \min\{k \geq 1 : \sum_{t=1}^k \epsilon_t = i\}$, $i = 1, \ldots, m$ (as before, $\min\emptyset = \infty$). Note that $\tau_m = \tau$ and $\tau_i \leq \tau_{i+1}$, $i = 1, \ldots, m-1$. Note that just like S_τ, S_{τ_i} is also well-defined by Remark A.1. Consider the process $(S'_{\tau_i}, \mathcal{F}_{\tau_i})_{i=1,\ldots,m}$, where $S'_{\tau_i}(\omega) \stackrel{\text{def}}{=} S'_{\tau_i(\omega)}(\omega)$ and $\mathcal{F}_{\tau_i}$ is the σ-algebra of pre-τ_i events. By the optional sampling theorem of Doob (see, e.g., Theorem 9.3.5 of Chung 2001), $(S'_{\tau_i}, \mathcal{F}_{\tau_i})_{i=1,\ldots,m}$ is a martingale. Let us now apply the Hoeffding-Azuma inequality to this martingale. In order to be able to do this, we need to show that the increments, $X_i = S'_{\tau_{i+1}} - S'_{\tau_i}$ lie in some bounded set for $i = 0, \ldots, m-1$, where we define $S'_0 = 0$. When $\tau_{i+1} = \infty$, $S'_{\tau_{i+1}} = S'_\infty = S$. Now, if $\tau_i = \infty$, we also

have $S'_{\tau_i} = S'_\infty = S$, while if $\tau_i < \infty$, we have $S = S_{\tau_i} = S'_{\tau_i}$. Thus, in both cases, $X_i = 0 \in [A, A+K]$ (that zero is in this interval follows because (Z_t) is a martingale increment). When $\tau_{i+1} < \infty$, we also have $\tau_i < \infty$ and thus $S'_{\tau_{i+1}} = S_{\tau_{i+1}}$ and also $S'_{\tau_i} = S_{\tau_i}$ and so $S'_{\tau_{i+1}} - S'_{\tau_i} = \epsilon_{\tau_{i+1}} Z_{\tau_{i+1}}$ and so $X_i \in [A, A+K]$ by our assumption on (Z_t). Thus, we have shown that $X_i \in [A, A+K]$ almost surely. The application of the Hoeffding-Azuma inequality to $(S'_{\tau_i}, \mathcal{F}_{\tau_i})_{i=1,\dots,m}$ gives then the desired result. ■

Remark A.2 *Note that the proof does not carry through for the case when we replace the assumption on the range of Z_t by the assumptions that $|Z_t| \le B$ a.s. and $Z_t \in [A_t, A_t + K]$ for some A_t, $\mathcal{F}_{t-1}$-measurable random variable and a deterministic constant $K > 0$. The problem is twofold: $(A_{\tau_i})_{i=1,\dots,m}$ might not be well-defined and even if it is, all we can say is that $\epsilon_{\tau_{i+1}} Z_{\tau_{i+1}} \in [A_{\tau_{i+1}}, A_{\tau_{i+1}} + K]$, but $A_{\tau_{i+1}}$ is not necessarily $\mathcal{F}_{\tau_i}$-measurable.*

In the proof of Theorem 4.1 we will need the one-dimensional version of Helly's theorem, which we nevertheless state for the d-dimensional Euclidean spaces:

Theorem A.2 (Helly's Theorem) *Let $d, N \in \mathbb{N}$, $N > d$, $C_1, \dots, C_N \subseteq \mathbb{R}^d$ be convex subsets of $\mathbb{R}^d$. If the intersection of any $d+1$ of these sets is nonempty, then $\cap_{i=1}^N C_i \neq \emptyset$.*

With these preparations, we are ready to prove Theorem 4.1.

Theorem 4.1 *Let $\mathcal{H}$ be a finite hypothesis class over $(\mathcal{X}, \mathbb{R})$, D, $\epsilon > 0$, $0 \le \delta \le 1$, $r = 2$. Then, for any $\mathcal{G}$ problem class such that the noise in the responses lies in $[-K, K]$ and $\Delta(\mathcal{G}, \mathcal{H}) \le D$ it holds that the pairwise elimination based agnostic learner is an agnostic (D, r, ϵ, δ) KWIK-learner for $(\mathcal{H}, \mathcal{G})$ with KWIK-bound $((\lceil \frac{2K^2}{\epsilon^2} \log \frac{2(N-1)}{\delta} \rceil - 1)N + 1)(N-1) = O\left(\frac{K^2 N^2}{\epsilon^2} \log \frac{N}{\delta}\right)$.*

Proof Fix $\mathcal{H}$, $\mathcal{G}$, D, ϵ, δ as in the theorem statement. Let $g^* : \mathcal{X} \to \mathbb{R}$ be the function underlying the problem chosen by the adversary. Let i^* be the index of a function $f_i \in \mathcal{H}$ that satisfies $\|f_{i^*} - g^*\|_\infty \le D$. By our assumption on $\mathcal{G}$ and $\mathcal{H}$, i^* is well-defined.

Let E be the (error) event when i^* is eliminated by the algorithm. We will show that $\mathbb{P}(E) \le \delta$ as from this, the rest follows easily: Indeed, on the complementer of E, i.e., on E^c, by the definition of **predict**, the algorithm makes $2D + \epsilon$-accurate predictions since if for some input $x \in \mathcal{X}$, $Y = \bigcap_{i \in I} B_{D+\epsilon}(f_i(x)) \neq \emptyset$ then for any $\hat{y} \in Y$, $|\hat{y} - f_{i^*}(x)| \le D + \epsilon$ (since $i^* \in I$) and thus

$$|\hat{y} - g(x)| \le |\hat{y} - f_{i^*}(x)| + |f_{i^*}(x) - g(x)| \le 2D + \epsilon\,.$$

Also, every time the algorithm passes, at least one of the counters $n_{i,j}$ is incremented. Indeed, if upon receiving input x the algorithm did not pass then $\bigcap_{i \in I} B_{D+\epsilon}(f_i(x)) = \emptyset$. Therefore, $|I| > 1$ and it follows from Helly's theorem (Theorem A.2) that there exists two distinct indices $i, j \in I$ such that $B_{D+\epsilon}(f_i(x)) \cap B_{D+\epsilon}(f_j(x)) = \emptyset$. For this pair (i, j), either

$n_{i,j}$ or $n_{j,i}$ is incremented. When some counter $n_{i,j}$ reaches the value m, at least one of i and j is excluded. Therefore, there can be at most $(m-1)N(N-1)$ calls to **learn** with no exclusions. Further, there can only be $N-1$ exclusions (since on E^c the index i^* does not get excluded). Thus, on E^c, there can be no more than $(m-1)N(N-1)+(N-1)$ calls to **learn**. Plugging in the value of m gives the KWIK-bound.

Thus, it remains to show that the probability of E is small, i.e., that $\mathbb{P}(E) \le \delta$. To prove this we need some more notation. Let $G = (\mathcal{X}, \mathcal{Y}, g^*, Z)$ be the problem and let $(x_t)_{t\ge 1}$ be the sequence of covariates ($x_t \in \mathcal{X}$) chosen by the adversary. To simplify the presentation, we introduce for *each* $t \ge 1$ a response, $y_t = g(x_t) + z_t$. Let $\mathcal{F}_t = \sigma(x_1, y_1, \dots, x_t, y_t)$. By assumption, the noise satisfies $z_t \sim Z(x_t)$ and, in particular, z_t lies in $[-K, K]$, and $\mathbb{E}[z_t | \mathcal{F}_{t-1}, x_t] = 0$. Let π_t be the indicator of whether the learner has passed when presented with the input x_t: $\pi_t = 1$ if the learner passed (and thus y_t is available for learning) and $\pi_t = 0$, otherwise. For $i, j \in \{1, \dots, N\}$, $t \ge 1$ let $\epsilon_t^{(i,j)} = \mathbb{I}_{\{f_i(x_t) \ll f_j(x_t)\}}$ and let $\tau^{(i,j)} = \min\{n \ge 1 : \sum_{t=1}^n \pi_t \epsilon_t^{(i,j)} = m\}$ be the time when the counter $n_{i,j}$ reaches m and thus either i or j gets eliminated by the algorithm (if it was not eliminated before). Here we let $\tau^{(i,j)} = \infty$ when $\sum_{t=1}^\infty \pi_t \epsilon_t^{(i,j)} < m$. Note that i^* gets eliminated only if one of $\tau^{(i^*,j)}$ or $\tau^{(j,i^*)}$ is finite for some $j \ne i^*$, $1 \le j \le N$. Thus,

$$E = \bigcup_{j \ne i^*}^* (E \cap \{\tau^{(i^*,j)} < \infty\}) \cup^* \bigcup_{j \ne i^*}^* (E \cap \{\tau^{(j,i^*)} < \infty\}). \tag{6}$$

We show that for $j \ne i^*$, both $E \cap \{\tau^{(i^*,j)} < \infty\}$ and $E \cap \{\tau^{(j,i^*)} < \infty\}$ happen with small probability.

Fix $j \ne i^*$ and consider $E \cap \{\tau^{(i^*,j)} < \infty\}$. To simplify the notation introduce $\tau = \tau^{(i^*,j)}$ and $\epsilon_t = \epsilon_t^{(i^*,j)}$. Let F be the event $F = \{\sum_{t=1}^\tau \pi_t \epsilon_t (f_{i^*}(x_t) + f_j(x_t)) < 2 \sum_{t=1}^\tau \pi_t \epsilon_t y_t\}$. Then $E \cap \{\tau^{(i^*,j)} < \infty\} = E \cap \{\tau < \infty\} \subset F \cap \{\tau < \infty\}$ holds because by the definition of the **learn** procedure, if i^* gets eliminated at time τ due to $n_{i^*,j}$ reaching m then it must hold that

$$\sum_{t=1}^\tau \pi_t \epsilon_t (f_{i^*}(x_t) + f_j(x_t)) < 2 \sum_{t=1}^\tau \pi_t \epsilon_t y_t. \tag{7}$$

Define $G = \{\sum_{t=1}^\tau \pi_t \epsilon_t z_t > K\sqrt{2m \log(2(N-1)/\delta)}\}$. We claim that

$$F \cap \{\tau < \infty\} \subset G \cap \{\tau < \infty\}. \tag{8}$$

To prove this, assume that (7) holds and $\tau < \infty$. Then,

$$\begin{aligned}
\sum_{t=1}^{\tau} \pi_t \epsilon_t z_t &= \sum_{t=1}^{\tau} \pi_t \epsilon_t (y_t - g^*(x_t)) \\
&> \sum_{t=1}^{\tau} \pi_t \epsilon_t \left\{ \frac{f_{i^*}(x_t) + f_j(x_t)}{2} - g^*(x_t) \right\} && \text{(because of (7))} \\
&\geq \sum_{t=1}^{\tau} \pi_t \epsilon_t \left\{ f_{i^*}(x_t) + (D+\epsilon) - g^*(x_t) \right\} && \text{(definition of } \epsilon_t = \epsilon_t^{(i^*,j)}\text{)} \\
&\geq \left\{ \sum_{t=1}^{\tau} \pi_t \epsilon_t \right\} \epsilon && \text{(because } f_{i^*}(x_t) \geq g^*(x_t) - D\text{)} \\
&= m\epsilon && \text{(definition of } \tau \text{ and } \tau < \infty\text{)} \\
&\geq K \sqrt{2m \log\left(\frac{2(N-1)}{\delta} \right)} && \text{(definition of } m\text{)},
\end{aligned}$$

finishing the proof of (8). By Lemma A.1, $\mathbb{P}(G) \leq \delta/(2(N-1))$ and thus we also have

$$\mathbb{P}\left(E \cap \{\tau^{(i^*,j)} < \infty\} \right) \leq \frac{\delta}{2(N-1)}.$$

With an entirely similar argument we can show that $\mathbb{P}\left(E \cap \{\tau^{(j,i^*)} < \infty\} \right) \leq \delta/(2(N-1))$ holds, too. Therefore, by the decomposition (6), $\mathbb{P}(E) \leq \delta$, finishing the proof. ■

Appendix B. Proofs for Section 5

Before turning to the proof of Lemma 5.3, we state Freedman's version of Bernstein's inequality (see, Freedman 1975, Theorem 1.6).

Theorem B.1 *Let $\mathcal{F} = (\mathcal{F}_k)_{k\geq 0}$ be a filtration and consider a sequence of $\mathcal{F}$-adapted random variables $(X_k)_{k\geq 1}$. Assume that $\mathbb{E}[X_k|\mathcal{F}_{k-1}] \leq 0$ and $X_k \leq R$ a.s. for $k = 1, 2, 3, \ldots$. Let the kth partial sum of $(X_k)_{k\geq 1}$ be S_k and let V_k^2 be the total conditional variance up to time k: $S_k = \sum_{i=1}^k X_i$, $V_k^2 = \sum_{i=1}^k \mathrm{Var}\left[X_i^2|\mathcal{F}_{i-1}\right]$. Let τ be a (not necessarily finite) stopping time w.r.t. $\mathcal{F}$. Then, for all $t \geq 0$, $v \in \mathbb{R}$,*

$$\mathbb{P}\left(S_\tau \geq t,\, V_\tau^2 \leq v^2 \;\; \textit{and } \tau < \infty \right) \leq \exp\left\{ -\frac{t^2/2}{v^2 + Rt/3} \right\}.$$

In the literature the above theorem is sometimes stated for *finite* stopping times only (or for the specific case when $\tau = k$ for some k). In fact, inequality 1.5(a) in Freedman's paper, from which the above theorem follows, is presented for finite stopping times only. However, the form of Theorem 1.6 of Freedman (1975) is actually equivalent to Theorem B.1. The next result follows from Theorem B.1 by a simple "inversion" argument and is given here because it will suit our needs better:

Corollary B.2 *Let $\mathcal{F}$, $(X_k)_{k\geq 1}$, $(S_k)_{k\geq 1}$, $(V_k^2)_{k\geq 1}$, R and τ be as in Theorem B.1. Then, for any $v \in \mathbb{R}$, $0 < \delta \leq 1$, it holds that*

$$\mathbb{P}\left(S_\tau \geq \sqrt{2v^2 \ln\left(\frac{1}{\delta}\right)} + \frac{2R}{3}\ln\left(\frac{1}{\delta}\right) \;,\; V_\tau^2 \leq v^2 \;\; \textit{and } \tau < \infty \right) \leq \delta \,.$$

Let us now turn to the proof of Lemma 5.3:

Lemma 5.3 *Let $0 < \epsilon < 1$, $m \in \mathbb{N}$ be deterministic constants, $(\mathcal{F}_t)_{t\geq 1}$ be some filtration and let $(A_t)_{t\geq 1}$ be an $(\mathcal{F}_{t+1})_{t\geq 1}$-adapted sequence of indicator variables. Let*

$$a_t = \mathbb{E}\left[A_t | \mathcal{F}_t\right]$$

and let G be an event such that on G the inequality $\sum_{t=1}^{\infty} A_t \leq m$ holds almost surely. Then, for any $0 < \delta \leq 1$ with probability $1-\delta$, either G^c holds, or

$$\sum_{t=1}^{\infty} \mathbb{I}_{\{a_t \geq \epsilon\}} \leq \frac{1}{\epsilon}\left\{ m + \sqrt{2m\log\left(\tfrac{1}{\delta}\right)} + 3\sqrt{\log\left(\tfrac{1}{\delta}\right)} + 6\log\left(\tfrac{1}{\delta}\right) \right\} .$$

It is interesting to compare the result of this lemma to what happens when $(a_t)_{t\geq 1}$ is a deterministic sequence, and A_t is Bernoulli with parameter a_t, independently chosen of all the other random variables. Clearly, in this case if $\sum_t A_t \leq m$ holds almost surely, we will also have $\sum_{t=1}^{\infty} a_t \leq m$. In contrast to this, in the sequential setting of the above lemma there exists (A_t, a_t) satisfying the conditions of the lemma such that for any $B > 0$, with positive probability, $\sum_{t=1}^{\infty} a_t > B$ (note that in both cases, $\mathbb{E}\left[\sum_{t=1}^{\infty} a_t\right] = \mathbb{E}\left[\sum_{t=1}^{\infty} A_t\right]$). Since $\sum_t a_t \geq \sum_t \mathbb{I}_{\{a_t \geq \epsilon\}} a_t \geq \epsilon \sum_t \mathbb{I}_{\{a_t \geq \epsilon\}}$, in the setting of independent Bernoulli trials, we get that almost surely, $\sum_t \mathbb{I}_{\{a_t \geq \epsilon\}} \leq \frac{1}{\epsilon}\sum_t A_t \leq m/\epsilon$. Thus, we see that the dependent and independent cases are quite different and the above lemma can be seen as quantifying the price of choosing a_t in a sequential manner.

Proof Define $S_n = \sum_{t=1}^{n}(a_t - A_t)$, $s_n = \sum_{t=1}^{n} a_t$, $V_n^2 = \sum_{t=1}^{n} \mathrm{Var}\left[a_t - A_t | \mathcal{F}_t\right] = \sum_{t=1}^{n} a_t(1-a_t)$, $\hat{s}_n = \sum_{t=1}^{n} \mathbb{I}_{\{a_t \geq \epsilon\}}$, for $n = 1, 2, \ldots, \infty$. Note that $V_n^2 \leq s_n$ holds for any n and V_n^2 is the total conditional variance associated with the martingale $(S_n, \mathcal{F}_{n+1})_{n\geq 0}$ (the empty sum is defined to be zero).

Fix $0 < \delta \leq 1$. Let $f = f(m, \delta)$ be a real number to be chosen later. We will define this number such that on some event F_δ, whose probability is at least $1-\delta$, we will have that

$$s_\infty \geq f \text{ implies that } \sum_{t=1}^{\infty} A_t > m \,. \tag{9}$$

Once we prove this, it follows by our assumption on $\sum_{t=1}^{\infty} A_t$ that on $F_\delta \cap G$, we must have $s_\infty < f$. Now, using $\mathbb{I}_{\{a_t \geq \epsilon\}}\epsilon \leq \mathbb{I}_{\{a_t \geq \epsilon\}} a_t$, we get that $\epsilon \hat{s}_n \leq \sum_{t=1}^{n} \mathbb{I}_{\{a_t \geq \epsilon\}} a_t \leq s_n$. Therefore, on $F_\delta \cap G$, $\hat{s}_\infty \leq \epsilon^{-1} s_\infty \leq \epsilon^{-1} f$. Plugging in the value of f will then finish the proof.

The event F_δ is chosen as follows: Let τ be the first index n when $s_n \geq f$ holds and let $\tau = \infty$ when there is no such index. Using Corollary B.2, we get that the probability of the event

$$E = \left\{ S_\tau \geq \sqrt{2(f+1)\log\left(\tfrac{1}{\delta}\right)} + \tfrac{2}{3}\log\left(\tfrac{1}{\delta}\right),\ V_\tau^2 \leq f+1\,, \tau < \infty \right\}$$

is at most δ: $\mathbb{P}(E) \leq \delta$. Note that $V_\tau^2 \leq s_\tau \leq f+1$ holds almost surely, where the last inequality follows from the definition of τ and because $a_t \in [0,1]$. Therefore, the second condition can be dropped in the definition of E without changing it: $E = \left\{ S_\tau \geq \sqrt{2(f+1)\log\left(\tfrac{1}{\delta}\right)} + \tfrac{2}{3}\log\left(\tfrac{1}{\delta}\right)\,, \tau < \infty \right\}$. Take $F_\delta = E^c$. Thus, $\mathbb{P}(F_\delta) \geq 1-\delta$ and on F_δ, we have

$$S_\tau < \sqrt{2(f+1)\log\left(\tfrac{1}{\delta}\right)} + \tfrac{2}{3}\log\left(\tfrac{1}{\delta}\right) \text{ or } \tau = \infty$$

which is equivalent to

$$\sum_{t=1}^{\tau} A_t > s_\tau - \sqrt{2(f+1)\log\left(\tfrac{1}{\delta}\right)} - \tfrac{2}{3}\log\left(\tfrac{1}{\delta}\right) \text{ or } \tau = \infty\,. \tag{10}$$

Let us now show that on F_δ, (9) holds. Consider an outcome $\omega \in F_\delta$ and assume that we also have $s_\infty(\omega) \geq f$ (to avoid clutter, we suppress ω in what follows). Because of $s_\infty \geq f$, it follows that $\tau < \infty$ and $s_\tau \geq f$. Therefore, from (10) and from $\sum_{t=1}^{\tau} A_t \leq \sum_{t=1}^{\infty} A_t$, we get that

$$\sum_{t=1}^{\infty} A_t > (f+1) - \sqrt{2(f+1)\log\left(\tfrac{1}{\delta}\right)} - \tfrac{2}{3}\log\left(\tfrac{1}{\delta}\right) - 1\,. \tag{11}$$

Now, define $f = f(m,\delta)$ to be a number such that

$$(f+1) - \sqrt{2(f+1)\log\left(\tfrac{1}{\delta}\right)} - \tfrac{2}{3}\log\left(\tfrac{1}{\delta}\right) - 1 \geq m\,. \tag{12}$$

Such a number exists because the left hand side, as a function of f, is unbounded. In fact, a simple calculation shows that, with the definitions $c = \sqrt{2\log(1/\delta)}$ and $L = 2/3\log(1/\delta)+1$, choosing f so that $(f+1)^{1/2}$ is larger than $(c + \sqrt{c + 4(m+L)})/2$ makes (12) hold true. Some calculation shows that $f \leq m + \sqrt{2m\log(1/\delta)} + 3\sqrt{\log(1/\delta)} + 6\log(1/\delta)$. Chaining the inequality (11) with the inequality (12), we get that on F_δ, (9) indeed holds, thus, finishing the proof.

■

Claim 5.4 *We have $V_M^{\tilde{\pi}_t}(s_t; L) \geq V_{\bar{M}_t}^{\pi_t}(s_t; L) - 2V_{\max} p_t$, with probability one.*

Proof Let $\mathcal{Z} = (S \times A)^L$ be the space of trajectories of length L which is viewed as a measurable space with the product σ-algebra (for S, A finite, this is just the discrete σ-algebra). Let π_t° be an arbitrary $\mathcal{F}_t$-measurable policy, $M_t = (S, A, P_{M_t}, R_{M_t})$ be an MDP,

where P_{M_t} and R_{M_t} are $\mathcal{F}_t$-measurable. Let F_{t,M_t,π_t° be the measure induced by M_t and π_t° on the space of L-step trajectories $\mathcal{Z}$, and the initial state distribution that is concentrated at the single state s_t (i.e., the initial state distribution used in the definition of F_{t,M_t,π_t° is the Dirac-measure $\delta_{s_t}(\cdot)$). Note that F_{t,M_t,π_t° is a random measure, which is itself $\mathcal{F}_t$-measurable. Let $\mathcal{R}_{M_t} : \mathcal{Z} \to \mathbb{R}$ be the mapping that assigns M_t-returns to the trajectories in $\mathcal{Z}$:

$$\mathcal{R}_{M_t}(s_0, a_0, \ldots, s_{L-1}, a_{L-1}) = \sum_{i=0}^{L-1} \gamma^i R_{M_t}(s_i, a_i)\,. \tag{13}$$

Now, consider the measures $F_{t,M,\tilde{\pi}_t}$ and $F_{t,\bar{M}_t,\tilde{\pi}_t}$ (these are $\mathcal{F}_t$-measurable thanks to condition (b)). An important property of these measures is that they agree when restricted to Z_{K_t}:

$$F_{t,M,\tilde{\pi}_t}\big|_{Z_{K_t}} = F_{t,\bar{M}_t,\tilde{\pi}_t}\big|_{Z_{K_t}}\,. \tag{14}$$

This property will play a crucial role in proving the desired inequality. Two further identities that we will need are the following: Let $Z_{K_t} = K_t^L$ be the set of L-step trajectories that stay within K_t. Then, we have

$$p_t = \int \mathbb{I}_{\{z \notin Z_{K_t}\}} dF_{t,M,\tilde{\pi}_t}(z) \tag{15}$$

$$= \int \mathbb{I}_{\{z \notin Z_{K_t}\}} dF_{t,M,\pi_t}(z)\,. \tag{16}$$

Clearly, these two equations are equivalent to the following ones:

$$1 - p_t = \int \mathbb{I}_{\{z \in Z_{K_t}\}} dF_{t,M,\tilde{\pi}_t}(z) \tag{17}$$

$$= \int \mathbb{I}_{\{z \in Z_{K_t}\}} dF_{t,M,\pi_t}(z)\,. \tag{18}$$

Therefore, it will suffice to prove that these latter equations hold true.

To show (17), first notice that $1 - p_t = \mathbb{E}\left[\mathbb{I}_{\{E_t^c\}} | \mathcal{F}_t\right]$ and $\mathbb{I}_{\{E_t^c\}} = \prod_{i=0}^{L-1} \mathbb{I}_{\{(s_{t+i}, a_{t+i}) \in K_t\}}$. Therefore, $\mathbb{E}\left[\mathbb{I}_{\{E_t^c\}} | \mathcal{F}_t\right] = \int \mathbb{I}_{\{z \in Z_{K_t}\}} dF_{t,M,\tilde{\pi}_t}(z)$, which shows that (17) indeed holds.

Let us now turn to the proof of (18). By (14), $\int \mathbb{I}_{\{z \in Z_{K_t}\}} dF_{t,M,\tilde{\pi}_t}(z) = \int \mathbb{I}_{\{z \in Z_{K_t}\}} dF_{t,\bar{M}_t,\tilde{\pi}_t}(z)$. Thanks to condition (g), along those trajectories that stay in K_t, the policy followed does not change. This implies that

$$dF_{t,\bar{M}_t,\tilde{\pi}_t}\big|_{Z_{K_t}} = dF_{t,\bar{M}_t,\pi_t}\big|_{Z_{K_t}}\,. \tag{19}$$

Therefore, we also have $\int \mathbb{I}_{\{z \in Z_{K_t}\}} dF_{t,\bar{M}_t,\tilde{\pi}_t}(z) = \int \mathbb{I}_{\{z \in Z_{K_t}\}} dF_{t,\bar{M}_t,\pi_t}(z)$, finishing the proof of (18).

Let us continue with the lower bound on $V^{\tilde{\pi}_t}(s_t; L)$. Then, $V_M^{\tilde{\pi}_t}(s_t; L) = \int \mathcal{R}_M(z) dF_{t,M,\tilde{\pi}_t}(z)$, where $\mathcal{R}_M(z)$ is the return assigned by model M to trajectory $z \in \mathcal{Z}$ (cf. (13)). Now, break

the integral into two parts using the decomposition $\mathcal{Z} = Z_{K_t} \cup^* Z_{K_t}^c$. For the integral over $Z_{K_t}^c$ use that $|\mathcal{R}_M(z)| \le V_{\max}$ (which holds by condition (a)) and then (15) to get $V_M^{\tilde{\pi}_t}(s_t; L) \ge \int_{Z_{K_t}} \mathcal{R}_M(z) dF_{t,M,\tilde{\pi}_t}(z) - V_{\max} p_t$. By (14) and (19),

$$\int_{Z_{K_t}} \mathcal{R}_M(z) dF_{t,M,\tilde{\pi}_t}(z) = \int_{Z_{K_t}} \mathcal{R}_{\bar{M}_t}(z) dF_{t,\bar{M}_t,\pi_t}(z) = V_{\bar{M}_t}^{\pi_t}(s_t; L) - \int_{Z_{K_t}^c} \mathcal{R}_{\bar{M}_t}(z) dF_{t,\bar{M}_t,\pi_t}(z)\,.$$

Using again $\mathcal{R}_{\bar{M}_t}(z) \le V_{\max}$ (which follows from condition (a)) and then (16) and chaining the previous equalities and inequalities, we get $V_M^{\tilde{\pi}_t}(s_t; L) \ge V_{\bar{M}_t}^{\pi_t}(s_t; L) - 2p_t V_{\max}$, which is the inequality that was to be proven. ■

We need some preparations before we give the proof of Claim 5.5. The next lemma also follows from a simple contraction argument (the proof is omitted). The lemma uses the partial ordering of functions: $f_1 \le f_2$ if $f_1(x) \le f_2(x)$ holds for all $x \in \mathrm{Dom}(f_1) = \mathrm{Dom}(f_2)$. Also, an operator is *isotone* if it preserves the ordering of its arguments.

Lemma B.3 (Comparison Lemma) *Let B be a Banach-space of real-valued functions over some domain D. Let $T_1, T_2 : B \to B$ be contractions and let $f_1^*, f_2^* \in B$ be their respective (unique) fixed-points. Assume that T_1 is isotone. Then, if $T_1 f_2^* \le T_2 f_2^* = f_2^*$ then $f_1^* \le f_2^*$.*

In the proof below, we also need the concept of Bellman operators. Let $M = (S, A, P, R)$ be an MDP and let π be a stationary policy over (S, A). The Bellman operator $T_M^\pi : \mathbb{R}^{S\times A} \to \mathbb{R}^{S\times A}$ underlying M and π is defined by

$$(T_M^\pi Q)(s,a) = R(s,a) + \gamma \int Q(s',a)\, d\pi(a|s') dP(s'|s,a), \qquad (s,a) \in S \times A\,.$$

As it is well known, T_M^π is a contraction with respect to the maximum norm and if Q_M^π denotes the unique fixed point of T_M^π, $\int Q_M^\pi(s,a) d\pi(a|s) = V_M^\pi(s)$ holds for all $s \in S$ (in fact, $Q_M^\pi(s,a)$ is the so-called action-value function underlying π, i.e., $Q_M^\pi(s,a)$ is the expected total discounted return if the decision process is started at state s, the first action is a and the subsequent actions are chosen by π).

Claim 5.5 *On G, it holds that $V_{\hat{M}_t}^{\pi^*}(s_t) \ge V_{\bar{M}_t}^{\pi^*}(s_t)$.*

Proof Instead of the claimed inequality, we prove the stronger inequality $Q_{\hat{M}_t}^{\pi^*} \ge Q_{\bar{M}_t}^{\pi^*}$. To prove this, we apply the Comparison Lemma (Lemma B.3). Choose B as the Banach-space of bounded, real-valued functions over $S \times A$ with the supremum norm $\|\cdot\|_\infty$, and consider two operators $T_{\hat{M}_t}^{\pi^*}, T_{\bar{M}_t}^{\pi^*} : B \to B$. Operator $T_{\hat{M}_t}^{\pi^*}$ is the policy evaluation operator corresponding to π^* on $\hat{M}_t$: $T_{\hat{M}_t}^{\pi^*} Q(s,a) = R_{\hat{M}_t}(s,a) + \gamma \int Q(s',a)\, d\pi^*(a|s') dP_{\hat{M}_t}(s'|s,a)$, while $T_{\bar{M}_t}^{\pi^*}$ is the $V_{\max}$-*truncated* policy evaluation operator corresponding to π^* on $\hat{M}_t$:

$T^{\pi^*}_{\hat{M}_t} Q(s,a) = \Pi\left[R_{\tilde{M}_t}(s,a) + \gamma \int Q(s',a)\, d\pi(a|s') dP_{\tilde{M}_t}(s'|s,a)\right]$, where $\Pi : \mathbb{R} \to \mathbb{R}$ is the projection to $[-V_{\max}, V_{\max}]$, i.e., $\Pi(x) = \max(\min(x, V_{\max}), -V_{\max})$. Clearly, $Q^{\pi^*}_{\hat{M}_t}$ is the fixed point of $Q^{\pi^*}_{\hat{M}_t}$, while $Q^{\pi^*}_{\tilde{M}_t}$ is the fixed point of $Q^{\pi^*}_{\tilde{M}_t}$, the latter of which follows because $\left\|Q^{\pi^*}_{\tilde{M}_t}\right\|_\infty \le V_{\max}$, thanks to condition (a). It is also clear that both operators are contractions. Take any $(s,a) \in S \times A$. We claim that $T^{\pi^*}_{\hat{M}_t} Q^{\pi^*}_{\hat{M}_t}(s,a) \ge T^{\pi^*}_{\tilde{M}_t} Q^{\pi^*}_{\hat{M}_t}(s,a)$. Let us first show that this inequality holds when $(s,a) \in K_t$. In this case, $|T^{\pi^*}_{\tilde{M}_t} Q^{\pi^*}_{\hat{M}_t}(s,a)| \le |r_{\tilde{M}_t}(s,a)| + \gamma V_{\max} \le V_{\max}$, because, by construction $r_{\tilde{M}_t}(s,a) = r_{\hat{M}_t}(s,a)$ and by condition (a) $|r_{\hat{M}_t}(s,a)| \le (1-\gamma)V_{\max}$. Therefore, the projection has no effect. Using that on the set K_t the models $\hat{M}_t$ and $\tilde{M}_t$ coincide, we get that $T^{\pi^*}_{\hat{M}_t} Q^{\pi^*}_{\hat{M}_t}(s,a) = T^{\pi^*}_{\tilde{M}_t} Q^{\pi^*}_{\hat{M}_t}(s,a)$. Now, consider the case when $(s,a) \notin K_t$. In this case, $T^{\pi^*}_{\hat{M}_t} Q^{\pi^*}_{\hat{M}_t}(s,a) = Q^{\pi^*}_{\hat{M}_t}(s,a) \ge V_{\max} \ge T^{\pi^*}_{\tilde{M}_t} Q^{\pi^*}_{\hat{M}_t}(s,a)$, where the first inequality follows from condition (f), while the second follows because $T^{\pi^*}_{\tilde{M}_t} Q(s,a)$ is restricted to the interval $[-V_{\max}, V_{\max}]$. This finishes the verification of the conditions of the Comparison Lemma. Therefore, the lemma gives that $Q^{\pi^*}_{\hat{M}_t} \ge Q^{\pi^*}_{\tilde{M}_t}$, which is the inequality that we wished to prove. ■

Theorem 5.6 *Fix a state space S and an action space A, which are assumed to be non-empty Borel spaces. Let $\mathcal{X}, \mathcal{Y}$ be as described above, $\mathcal{H}$ be an MDP hypothesis set, $\mathcal{G}$ be a set of MDP problem instances, both over $\mathcal{X}, \mathcal{Y}$. Assume that $\Delta(\mathcal{G}, \mathcal{H}) \le D$. Assume that $V_{\max} > 0$ is such that $(1-\gamma)V_{\max}$ is an upper bound on the immediate rewards of the MDPs determined by members of $\mathcal{H}$ and $\mathcal{G}$. Fix $\epsilon > 0, r \ge 1$, $0 < \delta \le 1/2$. Assume that* MDPLearner *is an agnostic (D, r, ϵ) KWIK-learner for $(\mathcal{H}, \mathcal{G})$ with KWIK-bound $B(\delta)$. Assume further that we are given a* Planner *which is $e_{planner}$-accurate. Consider the instance of the* KWIK-Rmax *algorithm which uses* MDPLearner *and* Planner*, interacting with some MDP M from $\mathcal{G}$. Let $\epsilon' = \frac{5(rD+\epsilon)}{1-\gamma} + e_{planner}$. Then, with probability $1-2\delta$, the number of ϵ'-mistakes, $N_{M,\epsilon'}$, made by* KWIK-Rmax *is bounded by $\frac{2V_{\max}(1-\gamma)L}{rD+\epsilon}\left\{B(\delta) + (\sqrt{2B(\delta)}+3)\sqrt{\log\left(\frac{L}{\delta}\right)} + 6\log\left(\frac{L}{\delta}\right)\right\}$, where $L = \max(1, \lceil (1-\gamma)^{-1}\log(V_{\max}(1-\gamma)/(rD+\epsilon))\rceil)$.*

Proof We apply Theorem 5.1 to the agent KWIK-Rmax that uses MDPLearner and Planner. Fix $0 < \delta \le 1$. The event G is constructed as follows: MDPLearner interacts with an "environment" according to the KWIK protocol. Consider the event on which it holds that the number of timesteps when MDPLearner passes is bounded by $B(\delta)$, while the learner's prediction errors (on the same event) is always below $rD + \epsilon$. By assumption, this event has probability at least $1-\delta$. Call this event G. Now, the sequence $(K_t)_{t \ge 1}$ is simply determined as follows. Let $g_t : S \times A \to \mathcal{Y} \cup \{\perp\}$ be the function underlying the predictions made by MDPLearner in step t. Then, $K_t = \{(s,a) \in S \times A : g_t(s,a) \neq \perp\}$. Further, let M_t be the model returned by the optimistic wrapper and let the policy π_t be the policy that Planner would "compute" at time t (i.e., $\pi_t(\cdot|s)$ is the distribution of actions returned by Planner if state s is fed to it). Let us verify the conditions of Theorem 5.1. The

bound on the expected immediate rewards (condition (a)) holds by assumption, just like the measurability condition (b) and that the action selected at time t is sampled from $\pi_t(\cdot|s_t)$ (condition (c)). The condition on the accuracy of the planner (condition (d)) was assumed as a condition of this theorem. The accuracy condition (e) holds with $e_{\text{model}} = rD + \epsilon$ on G by the choice of G, while the optimism condition (f) is met because of the use of the optimistic wrapper (in fact, because of this wrapper, $Q^{\pi}_{\hat{M}_t}(s,a) = V_{\max}$ holds for any $(s,a) \notin K_t$). Also, condition (g) is met because the learn method of MDPLearner is not called when $(s_t, a_t) \in K_t$, hence in that case $g_{t+1} = g_t$ and thus $\pi_{t+1} = \pi_t$. Finally, on G, $B = B(\delta)$ bounds the number of times $(s_t, a_t) \notin K_t$ happens. Therefore, by the conclusion of Theorem 5.1, with probability at least $1 - 2\delta$, the number of $5e_{\text{model}}/(1-\gamma) + e_{\text{plan}}$-mistakes is bounded by $\frac{2V_{\max}(1-\gamma)L}{e_{\text{model}}} \left\{ B + (\sqrt{2B}+3)\sqrt{\log\left(\frac{L}{\delta}\right)} + 6\log\left(\frac{L}{\delta}\right) \right\}$, where $L = \max(1, \lceil (1-\gamma)^{-1} \log(V_{\max}(1-\gamma)/e_{\text{model}}) \rceil)$. Plugging in the value $e_{\text{model}} = rD + \epsilon$ gives the final bound. ■

References

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In *NIPS*, pages 89–96, 2008.

Ronen I. Brafman and Moshe Tennenholtz. A near-optimal polynomial time algorithm for learning in certain classes of stochastic games. *Artificial Intelligence*, 121(1-2):31–47, 2000.

Ronen I. Brafman and Moshe Tennenholtz. R-MAX - a general polynomial time algorithm for near-optimal reinforcement learning. In *IJCAI*, pages 953–958, 2001.

Kai Lai Chung. *A course in probability theory.* Academic Press, 3 edition, 2001.

Carlos Diuk, Andre Cohen, and Michael L. Littman. An object-oriented representation for efficient reinforcement learning. In *ICML*, pages 240–247, 2008.

Joseph L. Doob. *Stochastic processes.* John Wiley & Sons, 1953.

Claude-Nicolas Fiechter. Efficient reinforcement learning. In *COLT*, pages 88–97, 1994.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1): 100–118, 1975.

Sham M. Kakade. *On the Sample Complexity of Reinforcement Learning.* PhD thesis, University College London, 2003.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2-3):209–232, 2002.

Lihong Li. *A Unifying Framework for Computational Reinforcement Learning Theory*. PhD thesis, Department of Computer Science, Rutgers University, New Brunswick, NJ, USA, 2009.

Lihong Li and Michael L. Littman. Reducing reinforcement learning to kwik online regression. *Annals of Mathematics and Artificial Intelligence*, 58(3-4):217–237, 2010.

Lihong Li, Michael L. Littman, and Thomas J. Walsh. Knows what it knows: A framework for self-aware learning. In *ICML*, pages 568–575, 2008.

Lihong Li, Michael L. Littman, Thomas J. Walsh, and Alexander L. Strehl. Knows what it knows: A framework for self-aware learning. *Machine Learning*, 82(3):399–443, 2011a.

Lihong Li, Michael L. Littman, Thomas J. Walsh, and Alexander L. Strehl. Knows what it knows: a framework for self-aware learning. *Machine learning*, 82:399–443, 2011b.

Martin .L. Puterman. *Markov Decision Processes — Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, 2005.

Alexander L. Strehl. Model-based reinforcement learning in factored-state MDPs. In *IEEE ADPRL*, pages 103–110, 2007.

Alexander L. Strehl and Michael L. Littman. A theoretical analysis of model-based interval estimation. In *Proceedings of the 22nd international conference on Machine learning*, pages 856–863, 2005.

Alexander L. Strehl and Michael L. Littman. Online linear regression and its application to model-based reinforcement learning. In *NIPS*, 2007.

Alexander L. Strehl, Lihong Li, and Michael L. Littman. Incremental model-based learners with formal learning-time guarantees. In *UAI*, pages 485–493, 2006.

Alexander L. Strehl, Carlos Diuk, and Michael L. Littman. Efficient structure learning in factored-state MDPs. In *AAAI*, pages 645–650, 2007.

Alexander L. Strehl, Lihong Li, and Michael L. Littman. Reinforcement learning in finite MDPs: PAC analysis. *The Journal of Machine Learning Research*, 10:2413–2444, 2009.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, 1998.

Csaba Szepesvári. *Algorithms for Reinforcement Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2010.

István Szita and András Lőrincz. The many faces of optimism: a unifying approach. In *ICML*, pages 1048–1055, 2008.

István Szita and Csaba Szepesvári. Model-based reinforcement learning with nearly tight exploration complexity bounds. In *ICML*, pages 1031–1038, June 2010.

Thomas J. Walsh, Sergiu Goschin, and Michael Littman. Integrating sample-based planning and model-based reinforcement learning. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, 2010.