JMLR: Workshop and Conference Proceedings vol 49:1–31, 2016

Policy Error Bounds for Model-Based Reinforcement Learning with Factored Linear Models

Bernardo Ávila Pires BPIRES@UALBERTA.CA
Department of Computing Science
University of Alberta

Csaba Szepesvári SZEPESVA@CS.UALBERTA.CA
Department of Computing Science
University of Alberta

Abstract

In this paper we study a model-based approach to calculating approximately optimal policies in Markovian Decision Processes. In particular, we derive novel bounds on the loss of using a policy derived from a factored linear model, a class of models which generalize numerous previous models out of those that come with strong computational guarantees. For the first time in the literature, we derive performance bounds for model-based techniques where the model inaccuracy is measured in weighted norms. Moreover, our bounds show a decreased sensitivity to the discount factor and, unlike similar bounds derived for other approaches, they are insensitive to measure mismatch. Similarly to previous works, our proofs are also based on contraction arguments, but with the main differences that we use carefully constructed norms building on Banach lattices, and the contraction property is only assumed for operators acting on "compressed" spaces, thus weakening previous assumptions, while strengthening previous results.

1. Introduction

The recent years have witnessed a renewed interest in model-based reinforcement learning (MBRL). Barreto et al. (2011); Kveton and Theocharous (2012) and Precup et al. (2012), building on the seminal work of Ormoneit and Sen (2002), studied various approaches to stochastic factorizations of the transition probability kernel, while Grünewälder et al. (2012) proposed to use RKHS embeddings to approximate the transition kernel, with further enhancements proposed recently by Lever et al. (2016). A key common feature of these otherwise distant-looking works is that once the model is set up, it leads to a policy in a computationally efficient way (i.e., in poly-time and space in the size of the model). Having realized that this is not a mere coincidence, Yao et al. (2014) introduced the concept of factored linear models, which keeps the advantageous computational properties, while generalizing all previous works. While efficient computation is a necessity, efficient learning and good performance of the policy are equally important. In this paper we focus on the second of these criteria, namely the performance of the policy derived from the model. The argument for omitting the learning part for the time being is that one should better understand first what errors need to be controlled because this will influence the choice of the learning objective and hence the algorithms (we also note in passing that, in the above-mentioned examples, the statistical analysis of the model learning algorithms is well understood by now).

We are not the first to consider the performance of the policy as a function of the model errors. In fact, most of the previously mentioned works also give bounds on the policy error (we define the policy error to be the performance loss due to using the derived policy instead of an optimal one). However, all these previous works derive bounds that express model errors in a supremum norm. While the supremum norm is a convenient choice when working with Markovian Decision Processes (which give the theoretical foundations in these works), an observation that goes back to at least Whitt (1978), the supremum norm is also known to be a rather unforgiving metric: In learning settings, when data comes from a large cardinality set, and the data may have an uneven distribution, while the objects of interest lack appropriate smoothness, or other helpful structural properties, we expect errors measured in the supremum norm to decrease rather slowly. Furthermore, most learning algorithms aim to reduce some weighted norms, hence deriving bounds for the supremum norm is neither natural, nor desirable. Can existing bounds of the policy error from the MBRL literature be extended to other norms? In the analogue context of approximate dynamic programming methods, Munos (2003) pioneered a technique to allow the use of weighted L^p-norms to bound the policy error, while in the context of approximate linear programming, de Farias and Van Roy (2003) proposed a different technique to allow the use of weighted supremum norms, both leading to substantial further work (Busoniu et al., 2010), (Wiering and van Otterlo, 2012, Chapter 3). While the use of weighted norms is a major advance, these bounds do not come without any caveats. In particular, in ALP, the bounds rely on the similarity of the so-called constraint sampling distribution to the stationary distribution μ^* of the optimal policy, while in ADP they rely on the similarity of the data sampling distribution and the start-state distribution, leading to hard to control error terms. Can this be avoided by model-based approaches?

Contributions. We derive bounds on the policy error of policies derived from factored linear models in MBRL. The policy error is bounded in supremum, weighted supremum and weighted L^p norms (Theorems 8, 10 and 11). The results hold under some conditions: the left factor of the approximate factorization of the transition kernel must satisfy a mild boundedness condition (Assumption 5), the right factor must be a join-homomorphism (Assumption 2), the operator obtained by swapping the left and right factors must satisfy a boundedness condition (Assumption 3 or Assumption 4). This latter condition is not mild as the one on the left factor, but it i) generalizes the conditions used to derive previous policy error bounds; and ii) can be easier to enforce as it constrains the norm of a low-dimensional operator, unlike the analogue constraints in previous works.

We recover results for unfactored linear models that satisfy a contraction assumption, and we recover existing supremum norm bounds for factored linear models that meet Assumption 2. In addition to being able to recover previous results, we also provide a new type of analysis, which has interesting implications. The new analysis shows that MBRL can in fact escape the sensitivities in ALP and ADP (cf. Theorem 11, term ε_1), answering the above major question on the positive. In fact, the new bound also shows the potential for better scaling with the discount factor, which is another surprising result. We attribute this success to the systematic use of the language of Banach lattices, which forced us to discover amongst other things a definition of *mixed* norms for action-value functions which is general, yet makes the so-called value selection operators non-expansions (cf. Proposition 1). For the skeptics who believe that MBRL is "hard" because the derived policy cannot be good before the model approximates "reality" uniformly everywhere, we point out that already the first ever bound derived for policy error in MBRL (due to Whitt (1978)) shows that the model has to be accurate only in an extremely localized way. Our bounds also share this characteristic of previous bounds.

Our analysis builds on techniques borrowed from approximate policy iteration (API) and approximate linear programming (ALP), and provide new insights to existing results for ALP (Proposition 4). However, the MBRL setup we consider is nevertheless different from API and ALP, so the connections in our proofs are not a mere translation of API or ALP results to MBRL, as we will explain in Section 6, which is also attested by the novel features of our bounds.

Other miscellaneous contributions include an example that shows why controlling the deviation between the optimal value function underlying the true and approximate models, a metric often used in some previous works to evaluate model quality, is insufficient to derive a policy error bound (Proposition 19 and Theorem 20). We present a characterization of linear join-homomorphisms (Proposition 3). We show that our supremum norm bounds are tight to arbitrary accuracy (Proposition 16), but that quantifying policy error in supremum norm can be harsh, so it pays off to consider the policy error in $L^p(\mu)$ norm instead (Proposition 17).

The rest of the paper is organized as follows: We start by providing the necessary background on MDPs in Section 2, followed by introducing factored linear models and the questions studied in Section 3. After this, we state our assumptions in Section 4, present our main results in Section 5, and close with placing our work in the context of existing work, and providing an outlook for future work in Section 6. While we include the proofs of our main results in the main body of the paper, proofs of technical results are relegated to the appendix.

2. Markov Decision Processes

We shall describe the agent-environment interaction using the framework of *Markov Decision Processes* (MDPs), with which the reader is assumed to be familiar. The notation used here is perhaps closest to that of Szepesvári (2010), but the reader may also consult the books of Puterman (1994) and Sutton and Barto (1998) on background. Here, we describe only the main concepts so as to clarify our notation. Well-understood technical details (such as measurability) are (mostly) omitted for brevity. The first two paragraphs describe standard notation, while the rest of the section defines less standard notation which is essential to understand the paper.

Markov Decision Processes. An MDP is a tuple $\langle \mathcal{X}, \mathcal{A}, \mathcal{P}, r \rangle$, where $\mathcal{X}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P} = (\mathcal{P}^a)_{a \in \mathcal{A}}$ is the *transition probability kernel* and $r = (r^a)_{a \in \mathcal{A}}$ is the *reward function*. For each state $x \in \mathcal{X}$ and action $a \in \mathcal{A}$, $\mathcal{P}^a(\cdot|x)$ gives a distribution over the states in $\mathcal{X}$, interpreted as the distribution over the next states given that action a is taken in state x. For each action $a \in \mathcal{A}$ and state $x \in \mathcal{X}$, $r^a(x)$ gives a real number, which is interpreted as the reward received when action a is taken in state x.[1]

An MDP describes the interaction of an agent and its environment. The interaction happens in a sequential manner where in each step the agent chooses an action $A_t \in \mathcal{A}$ based on the past information it has, sends the action to the environment, which then moves from the current state X_t to the next one according to the transition kernel: $X_{t+1} \sim \mathcal{P}^a(\cdot|X_t)$. The agent then observes the next state and the reward associated with the transition. In this paper we assume that the agent's

1. The standard MDP definitions would allow stochastic rewards, which may also be correlated with the next state. Our simplified model enhances clarity and extending our results to the case of stochastic rewards is trivial under a suitable set of assumptions.

goal is to maximize the *expected total discounted reward*, $\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r^{A_t}(X_t)\right]$, where $0 \le \gamma < 1$ is the so-called *discount factor*. A rule describing the way an agent acts given its past actions and observations is called a *policy*. The *value* of a policy π in a state x, denoted by $V^\pi(x)$, is the expected total discounted reward when the initial state (X_0) is x assuming the agent follows the policy. An *optimal policy* is one that achieves the maximum possible value amongst all policies in each state $x \in \mathcal{X}$. The *optimal value* for state x is denoted by $V^*(x)$. A *deterministic Markov policy* disregards everything in the history except the last step. Such policies can and will be identified with a map $\pi : \mathcal{X} \to \mathcal{A}$, and the space of measurable deterministic Markov policies will be denoted by Π. We will assume that the action set is finite. When, in addition, the reward function is bounded, which we assume from now on, all the value functions are bounded and one can always find a deterministic Markov policy that is optimal (Puterman, 1994). The suboptimality or *policy error* of a policy π at a state x is the difference $V^*(x) - V^\pi(x)$. Loosely speaking, a policy is near-optimal when this difference is small for the states that one cares about. In this work we are interested in bounding the policy error (for policies described in Section 3) in different norm choices: supremum norm, a weighted supremum norm and an $L^p(\mu)$ norm.

Spaces of value functions. Let $(\mathcal{V}, \|\cdot\|_{\mathcal{V}})$ be a Banach space of real-valued measurable functions over $\mathcal{X}$, equipped with a given norm, and $(\mathcal{V}^{\mathcal{A}}, \|\cdot\|_{\mathcal{V}^{\mathcal{A}}})$ be a Banach space of all measurable functions mapping $\mathcal{A}$ to $\mathcal{V}$. Elements of $\mathcal{V}$ are called *value functions*, while elements of $\mathcal{V}^{\mathcal{A}}$ are called *action-value functions*. Oftentimes, we will choose $\|\cdot\|_{\mathcal{V}}$ to be the norms mentioned before. The choice of $\|\cdot\|_{\mathcal{V}^{\mathcal{A}}}$ will in general depend on that of $\|\cdot\|_{\mathcal{V}}$, but this will be made clear in the actual context. Of course, $\mathcal{V}^{\mathcal{A}}$ can also be identified with the set of real-valued functions with domain $\mathcal{X} \times \mathcal{A}$ (since $\mathcal{A}$ is finite). To avoid too many parentheses, for $V \in \mathcal{V}^{\mathcal{A}}$, we will use V^a as an alternate notation to $V(a)$. Conveniently, $V^a \in \mathcal{V}$. With a slight abuse of notation, we denote by $\mathcal{P}^a$ the $\mathcal{V} \to \mathcal{V}$ right linear operator defined by $(\mathcal{P}^a V)(x) \doteq \int V(x')\mathcal{P}^a(dx'|x)$ (we assume that $V \in \mathcal{V}$ implies integrability, hence the integrals are well defined). We also view $\mathcal{P}^a$ as a left linear operator, acting over the space of probability measures defined over $\mathcal{X}$: $\mathcal{P}^a : \mathcal{M}_1(\mathcal{X}) \to \mathcal{M}_1(\mathcal{X})$, $(\mu\mathcal{P}^a)(dy) = \int \mu(dx)P^a(dy|x)$, $\mu \in \mathcal{M}_1(\mathcal{X})$. In what follows, whenever a norm is uniquely identifiable from its argument, we will drop the index of the norm denoting the underlying space.

Operators. The *Bellman return operator* w.r.t. to $\mathcal{P}$, $T_{\mathcal{P}} : \mathcal{V} \to \mathcal{V}^{\mathcal{A}}$, is defined by $T_{\mathcal{P}}V \doteq r + \gamma\mathcal{P}V$ (the indexing of T with $\mathcal{P}$ will help us to replace $\mathcal{P}$ with some other operator) and the so-called *maximum selection* operator $M : \mathcal{V}^{\mathcal{A}} \to \mathcal{V}$ is defined by $(MV)(x) \doteq \max_a V^a(x)$. Then, $MT_{\mathcal{P}}$, corresponds to the *Bellman optimality operator* (Puterman, 1994). The optimal value function satisfies $V^* = MT_{\mathcal{P}}V^*$ (Puterman, 1994), a non-linear fixed-point equation, which is known as the *Bellman optimality equation*. The *greedy operator* $G : \mathcal{V}^{\mathcal{A}} \to \Pi$, which selects the maximizing actions chosen by M, is defined by $GV(x) \doteq \operatorname{argmax}_a V^a(x)$ ($x \in \mathcal{X}$, with ties broken arbitrarily). Recall that GV^* is an optimal policy (Puterman, 1994).

Planning in MDPs. In the *online planning problem* we wish to compute, at any given state x, an action that a near-optimal policy would take (the attribute "online" signifies that one is allowed some amount of calculation for each state). By collecting all actions at all states, a planning method defines a policy $\hat{\pi}$. Disregarding computation, planning methods are compared by how good the policy they return is, i.e., by the policy error of $\hat{\pi}$. One approach to efficient online planning is to use an abstract model which i) contains relevant information about the MDP, ii) can be efficiently constructed, and iii) allows $\hat{\pi}(x)$ to be computed efficiently at any state x. In this work we are interested in online planning with a special type of abstract models, called *factored linear models*.

3. Factored Linear Models

In this section, we define factored linear models, the core of our MBRL approach. We also show examples of MBRL approaches that use factored linear models.

In a *factored linear model* we approximate the MDP's stochastic kernel $\mathcal{P}$ as the product of two linear operators, $\mathcal{Q}\mathcal{R}$, where $\mathcal{R} : \mathcal{V} \to \mathcal{W}$, $\mathcal{Q} = (\mathcal{Q}^a)_{a \in \mathcal{A}}$ and $\mathcal{Q}^a : \mathcal{W} \to \mathcal{V}^{\mathcal{A}}$ (Yao et al., 2014). Here, $\mathcal{W} = (\mathcal{W}, \| \cdot \|_{\mathcal{W}})$ is a Banach space of functions with (measurable) domain $\mathcal{I}$. We will refer to elements of $\mathcal{W}$ as *compressed value functions* and elements of $\mathcal{W}^{\mathcal{A}}$ as *compressed action value functions* (and, occasionally, the corresponding spaces will also be called compressed, while the spaces $\mathcal{V}$ and $\mathcal{V}^{\mathcal{A}}$ will be called uncompressed). These names come from the fact that often we will want to choose $\mathcal{I}$ to be "small". In fact, for computational reasons one should choose $\mathcal{I}$ to be finite, in which case $\mathcal{W}$ will be a finite-dimensional Euclidean space. We also allow infinite $\mathcal{I}$, so that we can then use $\mathcal{I} = \mathcal{X}$ and compare the tightness of our results to existing results that consider unfactored linear models.

In this work, for simplicity, we assume that the reward function r remains the same in the factored linear model (the extension of our results to the case when the reward function is also approximated is routine). Formally, in this work we will call a tuple of the form $\langle \mathcal{X}, \mathcal{A}, \mathcal{Q}, \mathcal{R}, r \rangle$ a factored linear model, where $\mathcal{Q}$ and $\mathcal{R}$ are as above. Finally, note that we do not require that $\mathcal{Q}\mathcal{R}$ is a stochastic operator. Hence, a factored linear model defines a *pseudo-MDP* (Yao et al., 2014).

We must define some additional operators in order to describe how we use factored linear models to derive policies. The extension of $\mathcal{R}$ to multiple actions, $\mathcal{R}^{\mathcal{A}} : \mathcal{V}^{\mathcal{A}} \to \mathcal{W}^{\mathcal{A}}$, is defined by $(\mathcal{R}^{\mathcal{A}} V)^a \doteq \mathcal{R}V$ ($a \in \mathcal{A}$), where $\mathcal{W}^{\mathcal{A}}$ is a Banach space of $\mathcal{A} \to \mathcal{W}$ functions analogously to $\mathcal{V}^{\mathcal{A}}$. The Bellman return operator for $\mathcal{Q}$, written as $T_{\mathcal{Q}} : \mathcal{W} \to \mathcal{V}^{\mathcal{A}}$, is defined by $T_{\mathcal{Q}} w \doteq r + \gamma \mathcal{Q} w$ ($w \in \mathcal{W}$). We also define the shorthands $T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}} \doteq \mathcal{R}^{\mathcal{A}} T_{\mathcal{Q}} = \mathcal{R}^{\mathcal{A}} r + \gamma \mathcal{R}^{\mathcal{A}} \mathcal{Q}$ (the equality holds by linearity of $\mathcal{R}^{\mathcal{A}}$) and $T_{\mathcal{Q}\mathcal{R}} \doteq T_{\mathcal{Q}} \mathcal{R} = r + \gamma \mathcal{Q}\mathcal{R}$ (by linearity of $\mathcal{R}$). Finally, $M' : \mathcal{W}^{\mathcal{A}} \to \mathcal{W}$, the counterpart of the maximum selection operator M, is defined by $(M'w)(i) = \max_{a \in \mathcal{A}} w^a(i)$ ($i \in \mathcal{I}$). The relationship between these operators is shown on Fig. 1, and we collected the operators defined here in Appendix A of the appendix into a table for easy of reference.

$$
\begin{tikzcd}
\mathcal{V} \arrow[dr, "\mathcal{R}"] & & \\
& \mathcal{W} \arrow[dl, "T_{\mathcal{Q}}"] \arrow[r, dashed, bend left, "T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}}"] & \mathcal{W}^{\mathcal{A}} \arrow[l, dashed, bend left, "M'"] \\
\mathcal{V}^{\mathcal{A}} \arrow[uu, "M"] \arrow[r, dotted, "G"] & \Pi &
\end{tikzcd}
\qquad
\begin{tikzcd}
\mathcal{W} \arrow[loop left, "M'T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}}"] \arrow[r, "\mathcal{Q}"] \arrow[d, bend left, "MT_{\mathcal{Q}}"] \arrow[dr, "\mathcal{R}^{\mathcal{A}}\mathcal{Q}"] & \mathcal{V}^{\mathcal{A}} \arrow[d, "\mathcal{R}^{\mathcal{A}}"] \\
\mathcal{V} \arrow[loop left, "MT_{\mathcal{Q}\mathcal{R}}"] \arrow[u, bend left, "\mathcal{R}"] & \mathcal{W}^{\mathcal{A}}
\end{tikzcd}
$$

Figure 1: Commutative diagrams showing the operators and the spaces that they act on.

The *factored linear model approach to reinforcement learning* is as follows: Given the factored linear model $\langle \mathcal{X}, \mathcal{A}, \mathcal{Q}, \mathcal{R}, r \rangle$, we take the policy

$$\hat{\pi} \doteq GT_{\mathcal{Q}}u^*, \tag{1}$$

where

$$u^* = M'T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}}u^* \,. \tag{2}$$

that is, the policy $\hat{\pi}$ does a *Bellman lookahead* with $T_{\mathcal{Q}}$ from $u^* \in \mathcal{W}$, a function that satisfies a fixed-point equation. Note that even when $\mathcal{X}$ is very large, or infinite, $\mathcal{W}$ can be finite dimensional, in which case a good approximation to u^* can often be found in a computationally efficient manner, for example by iterating $u_{k+1} = M'T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}}u_k$, which can be seen as a form of value iteration (Yao et al., 2014). The dashed lines on the left subfigure on Fig. 1 show that this computation can be done over the compressed spaces $\mathcal{W}$ and $\mathcal{W}^{\mathcal{A}}$. The diagram also shows that once u^* is found, $T_{\mathcal{Q}}$ extends this function to $\mathcal{V}^{\mathcal{A}}$, from where using the greedy operator G one obtains a policy. Note that in the applications the policy itself does not need to be explicitly represented, but the actions that the policy takes in a particular state $x \in \mathcal{X}$ can be computed "on demand" given u^* and the Bellman return operator $T_{\mathcal{Q}}$. (The right-hand side figure shows some more useful relationships between the operators involved.) We will say that this approach is *viable* when u^* is well-defined.

Factored linear models allow one to analyze modeling errors in seemingly distant model-based planning methods in a *unified manner*. This will be illustrated soon by describing how models proposed in numerous previous works can be written in a factored form (this was also shortly mentioned by Yao et al. (2014)). Before describing these previous models, we need some more definitions, to be able to describe the differences and similarities between them. In particular, the models will differ in terms of whether $\mathcal{R}$ is stochastic, or more specifically $\mathcal{R}$ is also a point-evaluator. Recall that the operator $\mathcal{R}$ is stochastic if $\inf_{V \geq 0} \inf_x (\mathcal{R}V)(x) \geq 0$ and $\mathcal{R}\mathbf{1}_{\mathcal{V}} = \mathbf{1}_{\mathcal{W}}$ where $\mathbf{1}_{\mathcal{V}}(x) = 1$ for all $x \in \mathcal{X}$ and $(\mathbf{1}_{\mathcal{W}})_i = 1$ for all $i \in \mathcal{I}$. Here, we started to use the convention of using w_i instead of $w(i)$ to reduce clutter. Also, we say that $\mathcal{R}$ is a *point-evaluator* if $\mathcal{I}$ indexes elements of $\mathcal{X}$ and $(\mathcal{R}V)_i = V(x_i)$ for all $i \in \mathcal{I}, V \in \mathcal{V}$. Note that point evaluators are stochastic. Choosing $\mathcal{I} = \mathcal{X}$ allows us to choose $\mathcal{R}$ to be the identity, which becomes a point evaluator when choosing $x_i = i$, $i \in \mathcal{I}$.

When $\mathcal{R}$ is a point selector, a short direct calculation shows that $\mathcal{R}M = M'\mathcal{R}^{\mathcal{A}}$, which means that on Fig. 1 the solid cycle and the dashed cycle starting from $\mathcal{W}$ are equivalent and we can interweave solid and dashed lines. For example, starting from $\mathcal{V}$: $MT_{\mathcal{Q}}M'T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}}\mathcal{R} = (MT_{\mathcal{Q}\mathcal{R}})^2$. The equivalence $M'T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}} = \mathcal{R}MT_{\mathcal{Q}}$ gives that $U^* \doteq MT_{\mathcal{Q}}u^*$ is a fixed point of $MT_{\mathcal{Q}\mathcal{R}}$, and that the identity $u^* = \mathcal{R}U^*$ also holds (cf. Theorem 5). It also follows that if $M'T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}}$ is a contraction (though $MT_{\mathcal{Q}\mathcal{R}}$ may not be), the factored linear model approach is viable. To the best of our knowledge, this observation has not been made elsewhere: In all previous works, viability was achieved by assuming that $\mathcal{Q}$ and $\mathcal{R}$ are both stochastic, or that $\mathcal{R}$ is a point evaluator and $\mathcal{Q}\mathcal{R}$ is a non-expansion in supremum norm. (In both cases $M'T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}}$ is a contraction, so u^* is well-defined and the factored linear model approach is viable.)

With this, we are ready to present different instances of the factored linear model approach:

Example 1 (Kernel-based reinforcement learning) *In kernel-based reinforcement learning (KBRL), introduced by Ormoneit and Sen (2002), $\mathcal{I}$ is indexing elements of $\mathcal{X}$, and $\mathcal{Q}$ is a stochastic operator constructed from kernel functions at elements of $S \doteq \{x_i : i \in \mathcal{I}\}$. Moreover,* *(a) S is an i.i.d. sample from $\mathcal{X} \doteq \mathbb{R}^d$ and $\mathcal{R}$ is a point evaluator (Ormoneit and Sen, 2002); or*

(b) S is a set of reference states *and $\mathcal{R}$ is stochastic (Barreto et al., 2011; Kveton and Theocharous, 2012; Precup et al., 2012).*

KBRL is viable because $\mathcal{Q}$ and $\mathcal{R}$ are stochastic, so $\mathcal{R}^{\mathcal{A}}\mathcal{Q}$ is also stochastic.

Example 2 (Pseudo-MDPs) *Pseudo-MDPs (Yao et al., 2014) are factored linear models with a point evaluator $\mathcal{R}$. In pseudo-MDPs, $\mathcal{Q}$ is no longer stochastic, but $\mathcal{QR}$ is assumed to be a non-expansion in supremum norm (Grünewälder et al., 2012; Yao et al., 2014; Lever et al., 2016). It can be shown that under this assumption both $MT_{\mathcal{QR}}$ and $M'T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}}$ are contractions. In the approach of these authors, one should take $\tilde{\pi} \doteq GT_{\mathcal{QR}}U^*$, where U^* is the fixed point of $MT_{\mathcal{QR}}$. Our formulation still applies, though, because we can show that $u^* = \mathcal{R}U^*$ is the fixed point of $M'T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}}$ (cf. Theorem 5), so that $\tilde{\pi} = GT_{\mathcal{QR}}U^* = GT_{\mathcal{Q}}u^* = \hat{\pi}$. Here, $\mathcal{Q}$ is essentially learned using a penalized least-squared approach.*

Example 3 (State aggregation) *State aggregation (Whitt, 1978; Bertsekas, 2011) in MBRL generalizes KBRL. Here, too, $\mathcal{I}$ is an index set over $\mathcal{X}$, and $\{x_i : i \in \mathcal{I}\}$ is the set of reference states. In hard aggregation, $\mathcal{R}$ is a point evaluator, while in soft aggregation (Singh et al., 1995) it is stochastic.*

Example 4 (MDP homomorphisms) *MDP homomorphisms (Ravindran, 2004; Sorg and Singh, 2009) can be used for transfer learning in reinforcement learning. Here, $\mathcal{I}$ is not identified with an index set over $\mathcal{X}$. If $\mathcal{R}$ is a point-evaluator, we recover MDP homomorphisms* per se *(Ravindran, 2004), and the more general case of $\mathcal{R}$ stochastic yields soft MRP homomorphisms (Sorg and Singh, 2009).*

Example 5 (Unfactored linear models) *It is possible to recover unfactored linear models as a special case of factored linear models by taking $\mathcal{W} = \mathcal{V}$, and $\mathcal{R}$ to be the identity mapping. For the approach to be viable, it is sufficient for $\mathcal{Q}$ to be stochastic, which is often assumed with unfactored linear models.*

4. Assumptions

The purpose of this section is to state and discuss the assumptions that will be used in our subsequent results.

Our first assumption states that the operators $M : \mathcal{V}^{\mathcal{A}} \to \mathcal{V}$, $M' : \mathcal{W}^{\mathcal{A}} \to \mathcal{W}$, and the related policy based value selector operators $M^\pi : \mathcal{V}^{\mathcal{A}} \to \mathcal{V}$ and $M'^\pi : \mathcal{W}^{\mathcal{A}} \to \mathcal{W}$ to be defined soon are non-expansions. Operator M^π is defined by $(M^\pi V)(x) \doteq V^{\pi(x)}(x)$ ($x \in \mathcal{X}$, $\pi \in \Pi$), while $(M'^\pi w)_i \doteq w_i^{\pi(i)}$ ($i \in \mathcal{I}$, $\pi : \mathcal{I} \to \mathcal{A}$). Now, recall that an operator $J : \mathcal{E} \to \mathcal{F}$ mapping between Banach spaces $\mathcal{E} = (\mathcal{E}, \|\cdot\|_{\mathcal{E}})$, $\mathcal{F} = (\mathcal{F}, \|\cdot\|_{\mathcal{F}})$ is called a non-expansion when its Lipschitz constant does not exceed one. The Lipschitz constant of J is defined by

$$\mathrm{Lip}(J) \doteq \sup_{e,e' \in \mathcal{E}: e \neq e'} \frac{\|Je - Je'\|}{\|e - e'\|},$$

where we follow the convention that the identity of the norm is derived from what space the argument belongs to. Note the dependence of Lip on the norms of $\mathcal{E}$ and $\mathcal{F}$, which we suppressed.

The definition implies that for any e, e', $\|Je - Je'\| \leq \mathrm{Lip}(J)\|e - e'\|$. Useful properties of Lip include that it is submultiplicative ($\mathrm{Lip}(JJ') \leq \mathrm{Lip}(J)\,\mathrm{Lip}(J')$), it is invariant to constant shifts of operators ($\mathrm{Lip}(J + e) = \mathrm{Lip}(J)$, where $J + e$ is defined by $(J + e)e' = e + Je'$) and when J is a linear operator, $\mathrm{Lip}(J) = \|J\|$, the induced operator norm of J, which is defined by

$$\|J\| \doteq \sup_{e \in \mathcal{E}, e \neq 0} \frac{\|Je\|}{\|e\|}.$$

Again, the induced norm depends on the norms that the operator acts between, but we suppress this dependence.

Let us now formally state the aforementioned assumption:

Assumption 1 (Non-expanding selectors) *We have* $\mathrm{Lip}(M) \leq 1$, $\mathrm{Lip}(M') \leq 1$ *and for any* $\pi_1 \in \Pi$, $\pi_2 : \mathcal{I} \to \mathcal{A}$, $\mathrm{Lip}(M^{\pi_1}) \leq 1$ *and* $\mathrm{Lip}(M'^{\pi_2}) \leq 1$.

Note that this assumption constrains what norms can be selected for the spaces $\mathcal{V}^{\mathcal{A}}$, $\mathcal{V}$, $\mathcal{W}^{\mathcal{A}}$ and $\mathcal{W}$. Assumption 1 will be helpful to establish that various operators involving M are Lipschitz with a factor strictly below one, i.e., that they are *contractions*. For example, to establish that $MT_{\mathcal{P}}$ is a contraction, one can use $\mathrm{Lip}(MT_{\mathcal{P}}) \leq \mathrm{Lip}(M)\,\mathrm{Lip}(T_{\mathcal{P}}) \leq \gamma\,\mathrm{Lip}(\mathcal{P}) = \gamma\|\mathcal{P}\|$, reducing the question to showing $\gamma\|\mathcal{P}\| < 1$. Similar arguments work the other operators that will involve M', M^{π}, or M'^{π}.

As it was alluded to earlier, we will use a number of different norms. However, in all cases we choose the norm for $\mathcal{V}^{\mathcal{A}}$ ($\mathcal{W}^{\mathcal{A}}$) based on the norm of $\mathcal{V}$ (respectively, the norm of $\mathcal{W}$) to be a mixed max-norm: In particular, for $\mathcal{U}$ being either $\mathcal{V}$ or $\mathcal{W}$, the norm of $\mathcal{U}^{\mathcal{A}}$ will be defined as $\|U\|_{\mathcal{U}^{\mathcal{A}}} = \|M_{|\cdot|}U\|_{\mathcal{U}}$ where $M_{|\cdot|} : \mathcal{U}^{\mathcal{A}} \to \mathcal{U}$ is defined by $(M_{|\cdot|}U)(\cdot) = \max_a |U^a(\cdot)|$. We call the resulting norm the *mixed max-norm* w.r.t. the norm of $\mathcal{U}$.

The next proposition shows that this choice of the mixed norm makes Assumption 1 hold whenever the underlying spaces are so-called Banach lattices (Meyer-Nieber, 1991). Recall that a lattice is a non-empty set $\mathcal{U}$ with a partial ordering $\leq$ such that every pair $f, g \in \mathcal{U}$ has a supremum (or least upper bound), denoted by $f \vee g$, and an infimum (greatest lower bound), denoted by $f \wedge g$. Spaces of real-valued functions are lattices with the componentwise ordering, our default choice in what follows, when it comes to $\mathcal{V}$ and $\mathcal{W}$. Operator $\vee$ is also called a *join*, a terminology we will adopt. A vector lattice $\mathcal{U}$ is a lattice that is also a vector space. In a vector lattice, for $f \in \mathcal{U}$, $f_+ = f \vee 0$, $f_- = (-f) \vee 0$ and $|f| = f_+ + f_-$ (these generalize the usual definitions of positive part, negative part and absolute value). A Banach lattice $\mathcal{U}$ is a normed vector lattice where $\mathcal{U}$ is also a Banach space and the norm satisfies that for any $f, g \in \mathcal{V}$, $|f| \leq |g| \implies \|f\| \leq \|g\|$. With this we are ready to restate and prove the said statement:

Proposition 1 *Assume that $\mathcal{V}$ and $\mathcal{W}$ are Banach lattices. Then Assumption 1 is satisfied.*

Proof To see why this holds, take for example M. Then for any $U, V \in \mathcal{V}^{\mathcal{A}}$, $MU - MV \leq M_{|\cdot|}(U - V)$ ($\leq$ denotes the componentwise ordering) and by swapping the order of U, V, we also get $|MU - MV| \leq M_{|\cdot|}(U - V)$. Now, since for any $f, g \in \mathcal{V}$, $|f| \leq |g|$ implies $\|f\| \leq \|g\|$, we get $\|MU - MV\| \leq \|M_{|\cdot|}(U - V)\| = \|U - V\|_{\mathcal{V}^{\mathcal{A}}}$. For M^{π}, since it is a linear operator, $\mathrm{Lip}(M^{\pi}) = \|M^{\pi}\|$, and for any $V \in \mathcal{V}^{\mathcal{A}}$, $\left|M^{\pi}V^{\mathcal{A}}\right| \leq M\left|V^{\mathcal{A}}\right|$, so $\mathrm{Lip}(M^{\pi}) \leq \mathrm{Lip}(M) \leq 1$. The statement is proven for the other operators analogously. ■

Let us now define the norms we will use in this paper. The weighted supremum norm of a function $f : \mathcal{Z} \to \mathbb{R}$ with respect to weight $w : \mathcal{Z} \to \mathbb{R}_+$ is defined as $\|f\|_{\infty,w} = \sup_{z\in\mathcal{Z}} |f(z)|/w(z)$. When $w = \mathbf{1}$ (i.e., $w(z) = 1$ for all $z \in \mathcal{Z}$), we drop w from the index and use $\|f\|_\infty$. For $p \ge 1$, the $L^p(\mu)$-norm of f is defined as $\|f\|_{\mu,p}^p \doteq \int_{\mathcal{Z}} |f(z)|^p \, d\mu(z)$. By slightly abusing notation, the mixed norm of space $\mathcal{U}^{\mathcal{A}}$ derived from $\|\cdot\|_{\infty,w}$, or $\|\cdot\|_{p,\mu}$ will be denoted identically (i.e., for $V \in \mathcal{V}^{\mathcal{A}}$, $\|V\|_{\infty,w}$ is a mixed norm defined using $M_{|\cdot|}$). Since these norms make their underlying spaces a Banach lattice, we immediately get the following corollary to Proposition 1:

Corollary 2 *Assume that the norms over $\mathcal{V}$ and $\mathcal{W}$ are supremum norms, weighted supremum norms, or $L^p(\mu)$ and $L^p(\rho)$ norms, and equip the spaces $\mathcal{V}^{\mathcal{A}}$ and $\mathcal{W}^{\mathcal{A}}$ with the respective mixed norms. Then Assumption 1 is satisfied.*

Note that $(\mathcal{V}, \vee)$ is a semi-lattice (a lattice with only a join). For the sake of simplicity, we make the following assumption, which will be assumed to hold until Theorem 12.

Assumption 2 ($\mathcal{R}$ is a join-homorphism) *The operator $\mathcal{R}$ is a join-homomorphism of the semi-lattice $(\mathcal{V}, \vee)$ into the semi-lattice $(\mathcal{W}, \vee)$, i.e., $\mathcal{R}(U \vee V) = (\mathcal{R}U) \vee (\mathcal{R}V)$ for any $U, V \in \mathcal{V}$.*

This assumption ensures that $\mathcal{R}M = M'\mathcal{R}^{\mathcal{A}}$, an identity which can be seen to hold simply by using the definitions and the above assumption, and which will be frequently used in our proofs.

The point evaluator defined in Section 3 is a linear join-homomorphism, and, since the identity operator is a point evaluator, it is also a linear join-homomorphism. However, stochastic operators (also often used in place of $\mathcal{R}$) may not be join-homomorphisms. As it turns out, the class of linear join-homomorphisms is not very diverse. Proposition 3 supports this claim for finite-dimensional $\mathcal{V}$ and $\mathcal{W}$, and the extension to infinite-dimensional spaces can be obtained by projection on finite-dimensional spaces. For a positive integer m, we let $[m] = \{1, \ldots, m\}$.

Proposition 3 *Assume that $\mathcal{V} = \mathbb{R}^m$ and $\mathcal{W} = \mathbb{R}^n$, and let $\mathcal{R}$ be any linear join-homomorphism. Then there exists $a \in \mathbb{R}_+^n$ and $J \in [m]^n$ s.t. $(\mathcal{R}v)_i = a_i \, v_{J_i}$ for all $v \in \mathcal{V}$ and $i \in [n]$.*

Our subsequent assumptions will ensure that certain operators are contractions in appropriate norms. We start with the simplest of these assumptions:

Assumption 3 *The following hold for $\mathcal{Q}$ and $\mathcal{R}^{\mathcal{A}}$: $\|\mathcal{R}^{\mathcal{A}}\mathcal{Q}\| \le 1$.*

Note that $\mathcal{R}^{\mathcal{A}}\mathcal{Q}$ is a $(\mathcal{W}, \|\cdot\|_{\mathcal{W}}) \to (\mathcal{W}^{\mathcal{A}}, \|\cdot\|_{\mathcal{W}^{\mathcal{A}}})$ operator and the norm used in Assumption 3 is the respective operator norm. As mentioned earlier, whenever Assumption 1 holds (which is the case for the norms under which we bound the policy error, cf. Corollary 2), we have that $\mathrm{Lip}(M'T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}}) \le \gamma \left\|\mathcal{R}^{\mathcal{A}}\mathcal{Q}\right\|$, and then Assumption 3 implies that $M'T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}}$ is a γ-contraction (again, for the respective operator norm). That $\mathcal{R}^{\mathcal{A}}\mathcal{Q}$ is a map between the compressed spaces $\mathcal{W}$ and $\mathcal{W}^{\mathcal{A}}$ is significant: When $\mathcal{W}$ is a finite dimensional space, Assumption 3 can be enforced during a learning procedure as done, e.g., by Yao et al. (2014). In fact, Yao et al. (2014) argue by means of some examples that enforcing this constraint as opposed to enforcing $\|\mathcal{Q}\mathcal{R}\| \le 1$ (which may be difficult to enforce as it constrains the norm of an operator between potentially infinite dimensional spaces) can lead to better results in some learning settings.

When the norms are specifically chosen to be weighted supremum norms, the previous assumption can be replaced by a weaker one, to be stated next. To state this assumption, we need to introduce the concept of Lyapunov functions, building on a more specialized definition due to de Farias and Van Roy (2003). As de Farias and Van Roy (2003) showed by means of an example, using weighted supremum norms can greatly reduce the error bounds. Intuitively, one achieves this by assigning large weights to unimportant states, i.e., to states that are infrequently visited by any policy. Indeed, one should not expect much data, or a good behavior at such states, but since they are not visited often, the errors made at such states can be safely discounted.

Given $\mathcal{Z} = (\mathcal{Z}, \|\cdot\|_{\infty,w})$, with $w : \mathcal{Z} \to \mathbb{R}_+$, and an operator $J : \mathcal{Z} \to \mathcal{Z}$, first let us define

$$\beta_{w,J} = \gamma \sup_{f:|f|=w} \|Jf\|_{\infty,w} .$$

Then, we say that the function w is *γ-Lyapunov* with respect to operator J if $\beta_{w,J} < 1$. We also extend the definition for operators of the form $K : \mathcal{Z} \to \mathcal{Z}^{\mathcal{A}}$, i.e., when $K = (K^a)_{a\in\mathcal{A}}$. In this case, we say that w is γ-Lyapunov w.r.t. to K if it is γ-Lyapunov w.r.t. to each operator K^a for any $a \in \mathcal{A}$. If J satisfies $Jf \le J|f|$ for all $f \in \mathcal{Z}$ (*e.g.*, if J is a stochastic operator), then the definition of $\beta_{w,J}$ simplifies to $\gamma \|Jw\|_{\infty,w}$, coinciding with the definition of de Farias and Van Roy (2003).

Lyapunov functions enable us to ensure that $MT_{\mathcal{P}}$, $M^{\pi}T_{\mathcal{P}}$ ($\pi \in \Pi$) and $M'T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}}$ are contractions in the corresponding weighted supremum norms. For this, notice that the following hold:

Proposition 4 *Given ($\mathcal{U}$, $\|\cdot\|_{\infty,\nu}$) with $\nu : \mathcal{U} \to \mathbb{R}_+$, and $J : \mathcal{U} \to \mathcal{U}^{\mathcal{A}}$, if each J^a is a linear operator, then $\gamma \operatorname{Lip}(J) = \beta_{\nu,J}$.*

Now, if ν is γ-Lyapunov w.r.t. the probability kernel $\mathcal{P}$, then we immediately get from Corollary 2 and Proposition 4 that $MT_{\mathcal{P}}$ and $M^{\pi}T_{\mathcal{P}}$ (for any $\pi \in \Pi$) are $\beta_{\nu,\mathcal{P}}$-contractions in ν-weighted supremum norm. Similarly, if η is γ-Lyapunov w.r.t. to $\mathcal{R}^{\mathcal{A}}\mathcal{Q}$, then $M'T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}}$ is a $\beta_{\eta,\mathcal{R}^{\mathcal{A}}\mathcal{Q}}$-contraction in η-weighted supremum norm.

With this, we can state the assumption that we will use to relax Assumption 3 when the norms used the respective function spaces are weighted supremum norms. In what follows we fix two functions, $\nu : \mathcal{V} \to \mathbb{R}_+$ and $\eta : \mathcal{W} \to \mathbb{R}_+$, which will act as weighting functions.

Assumption 4 (Lyapunov weights) *The following hold for $\mathcal{Q}$, $\mathcal{R}^{\mathcal{A}}$, ν, and η:*

(i) ν is γ-Lyapunov w.r.t. $\mathcal{P}$;

(ii) η is γ-Lyapunov w.r.t. to $\mathcal{R}^{\mathcal{A}}\mathcal{Q}$.

Note that choosing the weight function ν to be the constant one function, Assumption 4(i) is automatically satisfied, while choosing η to be the constant one function, Assumption 4(ii) is equivalent to Assumption 3 when the norm used there is the supremum norm.

Some (but not all) of our bounds will have a dependency on $\operatorname{Lip}(T_{\mathcal{Q}}) = \gamma \|\mathcal{Q}\|$. Therefore, we will also make Assumption 5 to avoid vacuous bounds.

Assumption 5 *We have that $B \doteq \|\mathcal{Q}\| < \infty$.*

Note that this assumption is mild: Learning procedures would more often than not guarantee finiteness of the objects they return. In fact, by appropriate normalization, even $\|Q\| \le 1$ can be arranged (if necessary) as done, for example, by Grünewälder et al. (2012).

5. Results

In this section we present our main results. We start with a viability result (explaining why our minimal assumptions are sufficient for the existence of the policy whose performance we are interested in), followed by a short review of previous bounds on the policy error. These previous bounds provide the context for our new results, which we present afterwards. After each result we discuss their relative merits and present their proofs. We reiterate that for all the results in this section Assumption 2 is assumed to hold, *i.e.*, $\mathcal{R}$ is assumed to be a join-homomorphism.

5.1. A viability result

Theorem 5 formalizes that u^* is well-defined (the MBRL approach with factored linear models is viable) under Assumption 3 or Assumption 4 (ii), provided that the norm over $\mathcal{W}^{\mathcal{A}}$ is a mixed max-norm w.r.t. the norm over $\mathcal{W}$. Theorem 5 shows that $M'T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}}$ is a contraction (in $\|\cdot\|_{\mathcal{W}}$) and we can compute u^* by value iteration. Therefore, as remarked in Section 2, if the compressed space $\mathcal{W}$ is finite dimensional, we are able to evaluate $M'T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}}$ and thus also approximate u^* efficiently (up to the desired accuracy). Evaluating $\hat{\pi}(x)$ can be done by computing $(T_{\mathcal{Q}}u^*)(x)$ for each x as needed. Theorem 5 also shows that $MT_{\mathcal{Q}\mathcal{R}}$ has a unique fixed point $U^* = MT_{\mathcal{Q}}u^*$, and it is not hard to see that U^* is a fixed point of $M^{\hat{\pi}}T_{\mathcal{Q}\mathcal{R}}$ as well. The fixed points U^* and u^*, as well the contraction $M'T_{\mathcal{Q}\mathcal{R}}$, will play pivotal roles in our bounds.

Theorem 5 *Assume that the norm over $\mathcal{W}^{\mathcal{A}}$ is the mixed max-norm w.r.t. the norm over $\mathcal{W}$, and let Assumption 3 or Assumption 4 (ii) hold. Assume also that $\mathcal{R}$ satisfies Assumption 2. Then $M'T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}}$ is a contraction w.r.t. to the norm underlying $\mathcal{W}$, $M'T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}}$ has a unique fixed point u^*, and the iteration $u_{k+1} = M'T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}}u_k$ converges geometrically to u^*, for any $u_0 \in \mathcal{W}$. Moreover, $U^* \doteq MT_{\mathcal{Q}}u^*$ is the unique fixed point of $MT_{\mathcal{Q}\mathcal{R}}$, and the identity $u^* = \mathcal{R}U^*$ holds.*

Before this work, it was not known that $M'T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}}$ being a contraction is sufficient for $MT_{\mathcal{Q}\mathcal{R}}$ to have a unique fixed point. As pointed out in Section 2, to the best of our knowledge, all previous works either assumed or imposed a contraction property on $MT_{\mathcal{Q}\mathcal{R}}$. In fact, with the exception of Yao et al. (2014), all previous works required $\mathcal{Q}\mathcal{R}$ to be stochastic.

In the proof of Theorem 5, which is presented ahead, we will use the following more general result:

Lemma 6 *Let $(\mathcal{V}, \|\cdot\|_{\mathcal{V}})$ and $(\mathcal{W}, \|\cdot\|_{\mathcal{W}})$ be two Banach spaces. Let $T : \mathcal{W} \to \mathcal{V}$ and $H : \mathcal{V} \to \mathcal{W}$ be two operators such that $\mathrm{Lip}((HT)^m) < 1$ for some $m > 0$. Then HT has a unique fixed point W^*, and $V^* \doteq TW^*$ is the unique fixed point of TH.*

The proof of Lemma 6 can be found in Appendix B. The argument we use is intuitive when $m = 1$: If HT is a contraction, it has a fixed point W^*, so defining $V^* \doteq TW^*$ gives $V^* = TW^* = THTW^* = THV^*$, so V^* is a fixed point of TH (and we also have the identity $W^* = HV^*$). The operator TH need not be a contraction for V^* to be its fixed point; indeed, we can even have $\mathrm{Lip}(TH) = \infty$ and still have $V^* = THV^*$ (cf. Proposition 23). The argument for $m > 1$ and for ensuring uniqueness relies largely on Banach's fixed point theorem.

Proof (of Theorem 5). To prove Theorem 5, we can apply Lemma 6 with $m = 1$, $T = MT_{\mathcal{Q}}$ and $H = \mathcal{R}$, but we have to ensure that $\mathrm{Lip}(M'T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}}) < 1$. We can use submultiplicativity of Lip and affinity of $T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}}$ to get that $\mathrm{Lip}(M'T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}}) \le \gamma\,\mathrm{Lip}(M')\,\mathrm{Lip}(\mathcal{R}^{\mathcal{A}}\mathcal{Q})$. By the choice of norm over $\mathcal{W}^{A}$, $\mathrm{Lip}(M') \le 1$, and by assumption $\gamma\,\mathrm{Lip}(\mathcal{R}^{\mathcal{A}}\mathcal{Q}) < 1$, so, indeed, $\mathrm{Lip}(M'T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}}) < 1$.

So far we have established that $M'T_{\mathcal{R}\mathcal{A}\mathcal{Q}}$ is a contraction, and Lemma 6 gives us that u^* is the fixed point of $M'T_{\mathcal{R}\mathcal{A}\mathcal{Q}}$, that U^* is the fixed point of $MT_{\mathcal{Q}\mathcal{R}}$, and that the two fixed points satisfy $u^* = \mathcal{R}U^*$. Because $M'T_{\mathcal{R}\mathcal{A}\mathcal{Q}}$ is a contraction, the iteration $u_{k+1} = M'T_{\mathcal{R}\mathcal{A}\mathcal{Q}}u_k$ converges geometrically to u^*, for any $u_0 \in \mathcal{W}$, by Banach's fixed-point theorem. ■

5.2. Previous results on the policy error

The typical MBRL performance bound is a supremum-norm bound on the policy error of $\tilde{\pi} \doteq GT_{\widetilde{\mathcal{P}}}\widetilde{V}$, where $\widetilde{\mathcal{P}}$ is stochastic and $\widetilde{V}$ is the fixed point of $MT_{\widetilde{\mathcal{P}}}$.

Theorem 7 (Baseline bound on MBRL policy error) *Consider some transition probability kernel $\widetilde{\mathcal{P}}$ for the state and action spaces $\mathcal{X}$ and $\mathcal{A}$. Let $\widetilde{V}$ be the fixed point of $MT_{\widetilde{\mathcal{P}}}$, and $\tilde{\pi} = GT_{\widetilde{\mathcal{P}}}\widetilde{V}$. Then*

$$\left\|V^* - V^{\tilde{\pi}}\right\|_\infty \leq \frac{2\gamma}{1-\gamma}\left\|(\mathcal{P} - \widetilde{\mathcal{P}})\widetilde{V}\right\|_\infty .$$

This result is essentially contained in the works of Whitt (1978, Corollary to Theorem 3.1), Singh and Yee (1994, Corollary 2)[2], Bertsekas (2012, Proposition 3.1), and Grünewälder et al. (2011, Lemma 1.1).

An important implication of this result, which we feel is often overlooked, is that the approximation $\widetilde{\mathcal{P}}$ to $\mathcal{P}$ does not have to be precise everywhere (at all functions $V \in \mathcal{V}$), but only at $\widetilde{V}$, the fixed point of the approximate model – a self-fulfilling prophecy, prone to failure? To understand why this works, consider the case when $\widetilde{\mathcal{P}}\widetilde{V}$ perfectly matches $\mathcal{P}\widetilde{V}$, *i.e.*, when the bound on the right-hand side is zero. In this case $\widetilde{V} = MT_{\widetilde{\mathcal{P}}}\widetilde{V} = MT_{\mathcal{P}}\widetilde{V}$, which implies that $\widetilde{V} = V^*$ and, $\tilde{\pi} = GT_{\widetilde{\mathcal{P}}}\widetilde{V} = GT_{\mathcal{P}}V^*$ is optimal. *The moral is that models do not have to be precise everywhere; if $\mathcal{P}\tilde{V}$ can be estimated, the above inequality can be used to derive a posteriori bounds on the policy error and even form the basis of improving the model.* This can be viewed as a major, unexpected win for model-based RL.

Ormoneit and Sen (2002); Barreto et al. (2011); Barreto and Fragoso (2011); Precup et al. (2012); Barreto et al. (2014b,a) bound $\|V^* - \widetilde{V}\|_\infty$ rather than the policy error. We emphasize (cf. Appendix D) that $\|V^* - \widetilde{V}\|_\infty$ is not the correct quantity to bound in order to understand the quality of $\hat{\pi}$, and that the policy error should be bounded. As we also discuss in Appendix D, this contrasts to ADP bounds, where, in order to understand the policy error in supremum norm, it is sufficient to bound the deviation between the optimal value function and the value estimate that generates the policy.

5.3. Bounds on the policy error in factored linear models

Our first novel result is a supremum-norm bound for policy error when we use factored linear models: Theorem 8. Because we can recover results for unfactored linear models by taking $\mathcal{R}$ to be the identity mapping over $\mathcal{X}$, we can use Theorem 8 to get a bound that is tighter than Theorem 7. Strictly speaking, taking $\mathcal{Q}$ stochastic, $\mathcal{R}$ as the identity mapping, and upper-bounding the right-hand side of Theorem 8 by $2\varepsilon_2$ gives us Theorem 7.

2. Singh and Yee (1994) correctly bound $\left\|V^* - V^{\tilde{\pi}}\right\|_\infty$, but their statement of Corollary 2 suggests that they are bounding a different quantity.

Theorem 8 (Supremum-norm bound) *Let $\hat{\pi}$ be the policy derived from the factored linear model defined using* (1) *and* (2)*. If Assumptions 3 and 5 hold, then*

$$\left\| V^* - V^{\hat{\pi}} \right\|_\infty \le \varepsilon(V^*) + \varepsilon(V^{\hat{\pi}}), \tag{3}$$

where $\varepsilon(V) = \min(\varepsilon_1(V), \varepsilon_2)$*, and*

$$\varepsilon_1(V) = \gamma \left\| (\mathcal{P} - \mathcal{Q}\mathcal{R})V \right\|_\infty + \frac{B\gamma^2}{1-\gamma} \left\| \mathcal{R}(\mathcal{P} - \mathcal{Q}\mathcal{R})V \right\|_\infty,$$
$$\varepsilon_2 = \frac{\gamma}{1-\gamma} \left\| (\mathcal{P} - \mathcal{Q}\mathcal{R})U^* \right\|_\infty.$$

The comments after Theorem 7 apply to Theorem 8: Curiously, it is enough if the model is "good" at its own fixed point. However, what is most striking about Theorem 8 is the $\varepsilon_1(V)$ term. It means that if B is not too big, and if the error of the model at V^* and $V^{\hat{\pi}}$ *in the compressed space* $\mathcal{W}^{\mathcal{A}}$ is small, then the term that depends on $\frac{1}{1-\gamma}$ is small. Moreover, we can expect this term to be easier to control than $\|(\mathcal{P} - \mathcal{Q}\mathcal{R})V\|_\infty$, though while the term with ε_2 may lead to *a posteriori* bounds, due to the presence of V^* and $V^{\hat{\pi}}$, objects in the true MDP, the ε_1 terms are better treated as *a priori* bounds.

The proof of Theorem 8 (presented below) uses the triangle inequality

$$\|V^* - V^{\hat{\pi}}\| \le \|V^* - U^*\| + \|U^* - V^{\hat{\pi}}\|, \tag{4}$$

combined with Lemma 9 stated next (Lemma 9 is a technical lemma and its proof is in Appendix C):

Lemma 9 *Let Assumptions 1 and 5 hold, and assume that* $\gamma \operatorname{Lip}(\mathcal{R}^{\mathcal{A}}\mathcal{Q}) \le \alpha < 1$*. For* $V \in \left\{V^*, V^{\hat{\pi}}\right\}$ *we have that*

$$\|V - U^*\| \le \gamma \left\| (\mathcal{P} - \mathcal{Q}\mathcal{R})V \right\| + \frac{B\gamma^2}{1-\alpha} \left\| \mathcal{R}^{\mathcal{A}}(\mathcal{P} - \mathcal{Q}\mathcal{R})V \right\|. \tag{5}$$

Additionally, if $\gamma \operatorname{Lip}(\mathcal{P}) \le \beta < 1$ *(or, alternatively,* $\gamma \operatorname{Lip}(M^{\hat{\pi}}\mathcal{P}) \le \beta < 1$*), we also have for* $V = V^*$ *(respectively,* $V = V^{\hat{\pi}}$*) that*

$$\|V - U^*\| \le \frac{\gamma}{1-\beta} \left\| (\mathcal{P} - \mathcal{Q}\mathcal{R})U^* \right\|. \tag{6}$$

Lemma 9 (5) can be interpreted as the bound we get by doing a Bellman lookahead with $MT_{\mathcal{Q}}$, followed by application of the well-known bound for an α-contraction T with fixed point $\widetilde{V}$ (Bertsekas, 1995):

$$\left\| V - \widetilde{V} \right\| \le \frac{1}{1-\alpha} \left\| V - TV \right\| \tag{7}$$

(with $T = M' T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}}$ in the case of Lemma 9). Similarly, taking $T = MT_{\mathcal{P}}$ ($T = M^{\hat{\pi}} T_{\mathcal{P}}$) in (7) combined with $\gamma \operatorname{Lip}(\mathcal{P}) \le \beta < 1$ ($\gamma \operatorname{Lip}(M^{\hat{\pi}}\mathcal{P}) \le \beta < 1$), allows us to see that $MT_{\mathcal{P}}$ ($M^{\hat{\pi}} T_{\mathcal{P}}$) is a β-contraction, so (7) gives us Lemma 9 (6) for V^* ($V^{\hat{\pi}}$). Lemma 9 (5) is also interesting in the special case of unfactored linear models (when $\mathcal{R}$ is the identity mapping) with $\mathcal{Q}$ as a non-expansion (*e.g.*, $\mathcal{Q}$ stochastic): Because $B = 1$ and $\alpha = \gamma$, the bound becomes

$$\|V - U^*\| \le \frac{\gamma}{1-\gamma} \left\| (\mathcal{P} - \mathcal{Q}\mathcal{R})V \right\|,$$

and in this case no looseness was introduced by doing a Bellman lookahead and then applying (7), relative to applying (7) directly. This will allow us to recover results for unfactored linear models from the bounds we derive from Lemma 9.

Proof (of Theorem 8) We will verify the assumptions of Lemma 9, so that we can bound the terms on the right-hand side (RHS) of (4) with the help of this lemma. Lemma 9 needs: Assumption 1, Assumption 5, $\gamma \operatorname{Lip}(\mathcal{R}^{\mathcal{A}}\mathcal{Q}) < 1$, $\gamma \operatorname{Lip}(\mathcal{P}) < 1$ and $\gamma \operatorname{Lip}(M^{\hat{\pi}}\mathcal{P}) < 1$. Assumption 1 holds by Corollary 2, whose assumptions are satisfied because Theorem 8 uses supremum norms. Assumption 5 holds by assumption. Next, Assumption 3 implies that $\gamma \operatorname{Lip}(\mathcal{R}^{\mathcal{A}}\mathcal{Q}) \leq \gamma < 1$. Because $\operatorname{Lip}(\mathcal{P}) = 1$ in supremum norm, we get $\gamma \operatorname{Lip}(P) \leq \gamma < 1$. Finally, $\operatorname{Lip}(\mathcal{P}) = 1$ and Assumption 1 imply together that $\operatorname{Lip}(M^{\hat{\pi}}\mathcal{P}) \leq \gamma < 1$. The result is obtained by using Lemma 9 (with $\alpha = \beta = \gamma$) to bound the terms on the RHS of (4). ■

Theorem 8 is tight, as shown by Proposition 16 (cf. Appendix C). Trivially, we can use Theorem 8 to crudely upper-bound the policy error in $L^p(\mu)$ norm, but the bound we obtain this way is not very interesting. This is because supremum norm bounds, though easy to prove, can be too harsh: V^* and $V^{\hat{\pi}}$ can be close in other meaningful norms, while not being close in supremum norm, in which case the right-hand side of the bound in Theorem 8 can be large even if the left-hand side is small (cf. Proposition 17, Appendix C).

De Farias and Van Roy (2003) show that the harshness of the supremum norm can be mitigated by considering the policy error in weighted supremum norm. Intuitively, the error in states that are unlikely to be visited by π^* should be underweighted, as we discussed earlier. Thus, one alternative to supremum norm bounds is to use a generalization of Theorem 8 for the weighted supremum norm:

Theorem 10 (Weighted supremum norm bound) *Let $\hat{\pi}$ be the policy derived from the factored linear model defined using* (1) *and* (2). *If Assumptions 4 and 5 hold, then*

$$\left\| V^* - V^{\hat{\pi}} \right\|_{\infty,\nu} \leq \varepsilon(V^*) + \varepsilon(V^{\hat{\pi}}),$$

where $\varepsilon(V) = \min(\varepsilon_1(V), \varepsilon_2)$, and

$$\begin{aligned}
\varepsilon_1(V) &= \gamma \left\| (\mathcal{P} - \mathcal{Q}\mathcal{R})V \right\|_{\infty,\nu} + \frac{B\gamma^2}{1 - \beta_{\eta,\mathcal{R}^{\mathcal{A}}\mathcal{Q}}} \left\| \mathcal{R}(\mathcal{P} - \mathcal{Q}\mathcal{R})V \right\|_{\infty,\eta} \\
\varepsilon_2 &= \frac{\gamma}{1 - \beta_{\nu,\mathcal{P}}} \left\| (\mathcal{P} - \mathcal{Q}\mathcal{R})U^* \right\|_{\infty,\nu}.
\end{aligned}$$

Under Assumption 3 and Assumption 4 (i), Theorem 10 holds with $\beta_{\eta,\mathcal{R}^{\mathcal{A}}\mathcal{Q}} = \gamma$. The comments about $\varepsilon_1(V)$ and ε_2 in Theorems 7 and 8 are also valid for Theorem 10, but the dependencies are, evidently, expressed in different norms. Moreover, by taking $\nu = x \mapsto 1$ and $\eta = i \mapsto 1$, and by realizing that ν is γ-Lyapunov w.r.t. to $\mathcal{P}$ and, under Assumption 3, η is γ-Lyapunov w.r.t. to $\mathcal{R}^{\mathcal{A}}\mathcal{Q}$, we recover Theorem 8 from Theorem 10. Previously, weighted-supremum norm bounds were derived for ALP. However, the weakness of these bounds is that they are sensitive to the measure-change between the "ideal constraint sampling distribution" (which depends on unknown quantities whose knowledge basically implies the knowledge of the optimal policy) and the actual one used in the algorithm (de Farias and Van Roy, 2003).

Proof (of Theorem 10) We start with the triangle inequality in (4). To obtain $\varepsilon_1(V)$ we use Lemma 9 (5) with $\alpha = \beta_{\eta,\mathcal{R}^\mathcal{A}\mathcal{Q}}$. The conditions of Lemma 9 (5) are fulfilled by Corollary 2 and Assumption 5, and because η is γ-Lyapunov w.r.t. $\mathcal{R}^\mathcal{A}\mathcal{Q}$ (via Assumption 4 (ii)).

Lemma 9 (5) gives ε_2 after we realize that $\mathrm{Lip}(MT_\mathcal{P}) \le \gamma \,\mathrm{Lip}(\mathcal{P}) = \gamma\beta_\nu < 1$ and that $\mathrm{Lip}(M^{\hat\pi}T_\mathcal{P}) \le \gamma \,\mathrm{Lip}(\mathcal{P}) = \gamma\beta_\nu < 1$, since ν is γ-Lyapunov w.r.t. $\mathcal{P}$ by Assumption 4 (i). ■

Normally, we are interested in the policy error w.r.t. to an initial state distribution, or a stationary distribution of a policy (*e.g.*, a stationary distribution of π^*), and we can naturally consider the policy error in $L^1(\mu)$ norm, where μ is a measure over $\mathcal{X}$ that we are interested in. We can get an immediate bound for the more general $L^p(\mu)$ norm (for any $p \ge 1$) of the policy error, using Theorem 10 (cf. Theorem 18, Appendix C). However, we can also bound the policy error in $L^p(\mu)$ "directly", *i.e.*, in terms of model errors in $L^p(\mu)$ norm, as Theorem 11, to be stated next, shows.

In order to state Theorem 11, we need to use a *concentrability coefficient* $C_{\hat\pi,\mathcal{P},\mu,\xi}$ (although part of our bound will be free of this coefficient). Consider a measure ξ over $\mathcal{X}$, and the operator $I - \gamma M^{\hat\pi}\mathcal{P} : (\mathcal{V}, \|\cdot\|_{\xi,p}) \to (\mathcal{V}, \|\cdot\|_{\mu,p})$. If $I - \gamma M^{\hat\pi}\mathcal{P}$ has no inverse (as an operator acting between the above two spaces), define $C_{\gamma,\hat\pi,\mathcal{P},\mu,\xi} \doteq \infty$, otherwise let the concentrability coefficient be

$$C_{\gamma,\hat\pi,\mathcal{P},\mu,\xi} \doteq (1-\gamma)\,\mathrm{Lip}((I - \gamma M^{\hat\pi}\mathcal{P})^{-1}) = (1-\gamma)\left\|(I - \gamma M^{\hat\pi}\mathcal{P})^{-1}\right\| . \tag{8}$$

(Note that here both $\mathrm{Lip}(\cdot)$ and $\|\cdot\|$ hide a dependence on ξ, π and p.) As opposed to previous uses of concentrability coefficients (Munos, 2003; Farahmand et al., 2010), our coefficient depends only on the policy computed, which makes it more suitable for the estimation of our bound. In case the $C_{\gamma,\hat\pi,\mathcal{P},\mu,\xi}$ is not very large, we can get meaningful bounds from Theorem 11 from ε_2, but even if $C_{\gamma,\hat\pi,\mathcal{P},\mu,\xi} = \infty$ and ε_2 is vacuous, we can still get a priori bounds with a dependence on $\varepsilon_1(V^{\hat\pi})$, in addition to the dependence on $\varepsilon_1(V^*)$. The $\varepsilon_1(V)$ term can be analyzed similarly to its analogues in Theorems 8 and 10, modulo the norm differences. We are flexible about the choice of $\|\cdot\|_\mathcal{W}$ (which nonetheless affects Assumptions 3 and 5). One may think of choosing $\|\cdot\|_\mathcal{W} = \|\cdot\|_{\rho,p}$ for some ρ, however with this norm choice, Assumption 3 becomes restrictive. When it comes to satisfying Assumption 3, a weighted supremum norm is reasonable, as discussed earlier, so we choose this norm as the norm over the compressed space $\mathcal{W}$ in Theorem 11.

Theorem 11 ($L^p(\mu)$ norm bound) *Let $\hat\pi$ be the policy derived from the factored linear model defined using* (1) *and* (2)*. Choose the norms so that $\|\cdot\|_\mathcal{V} = \|\cdot\|_{\mu,p}$ and $\|\cdot\|_\mathcal{W} = \|\cdot\|_{\infty,\eta}$. If Assumptions 3 and 5 hold, then*

$$\left\|V^* - V^{\hat\pi}\right\|_{\mu,p} \le \varepsilon_1(V^*) + \min\left\{\varepsilon_1(V^{\hat\pi}), \varepsilon_2\right\},$$

where

$$\varepsilon_1(V) = \gamma\left\|(\mathcal{P} - \mathcal{Q}\mathcal{R})V\right\|_{\mu,p} + \frac{B\gamma^2}{1-\gamma}\left\|\mathcal{R}^\mathcal{A}(\mathcal{P} - \mathcal{Q}\mathcal{R})V\right\|_{\infty,\eta},$$

$$\varepsilon_2 = C_{\gamma,\hat\pi,\mathcal{P},\mu,\xi}\,\frac{\gamma}{1-\gamma}\left\|(\mathcal{P} - \mathcal{Q}\mathcal{R})U^*\right\|_{\xi,p},$$

where $C_{\gamma,\hat\pi,\mathcal{P},\mu,\xi}$ is defined in (8).

Before the proof, let us point out that ε_1 is independent of the concentrability coefficient. Further, as remarked beforehand, its dependence on the discount factor can be quite mild (if the second term in the definition of ε_1 is small).

Proof The first step is to use (4). Then we see that Corollary 2 ensures that Assumption 1 is satisfied, and Assumption 3 guarantees that $\left\|\mathcal{R}^{\mathcal{A}}\mathcal{Q}\right\| \leq 1$. Thus, Lemma 9 (5) with $\alpha = \gamma$ gives us $\|U^* - V\|_{\mu,p} \leq \varepsilon_1(V)$ for $V \in \left\{V^*, V^{\hat{\pi}}\right\}$.

To bound $\left\|U^* - V^{\hat{\pi}}\right\|_{\mu,p} \leq \varepsilon_2(V^{\hat{\pi}})$ we proceed as follows. If $(I - \gamma M^{\hat{\pi}}\mathcal{P})$ is not invertible, then $C_{\gamma,\hat{\pi},\mathcal{P},\mu,\xi} = \infty$ and the result holds vacuously, so assume otherwise. Since $V^{\hat{\pi}} = M^{\hat{\pi}} T_{\mathcal{P}} V^{\hat{\pi}}$,

$$(I - \gamma M^{\hat{\pi}}\mathcal{P})V^{\hat{\pi}} = M^{\hat{\pi}} r.$$

Moreover,

$$U^* - \gamma M^{\hat{\pi}}\mathcal{P}U^* - M^{\hat{\pi}} r = U^* - M^{\hat{\pi}} T_{\mathcal{P}} U^*.$$

Hence,

$$\begin{aligned}
\left\|U^* - V^{\hat{\pi}}\right\|_{\mu,p} &= \left\|(I - \gamma M^{\hat{\pi}}\mathcal{P})^{-1}(I - \gamma M^{\hat{\pi}}\mathcal{P})(U^* - V^{\hat{\pi}})\right\|_{\mu,p} \\
&\leq \operatorname{Lip}((I - \gamma M^{\hat{\pi}}\mathcal{P})^{-1}) \left\|(I - \gamma M^{\hat{\pi}}\mathcal{P})(U^* - V^{\hat{\pi}})\right\|_{\xi,p} \\
&= C_{\gamma,\hat{\pi},\mathcal{P},\mu,\xi} \frac{1}{1-\gamma} \left\|U^* - M^{\hat{\pi}} T_{\mathcal{P}} U^*\right\|_{\xi,p} \\
&\leq C_{\gamma,\hat{\pi},\mathcal{P},\mu,\xi} \operatorname{Lip}(M^{\hat{\pi}}) \frac{\gamma}{1-\gamma} \left\|(\mathcal{P} - \mathcal{Q}\mathcal{R})U^*\right\|_{\xi,p}
\end{aligned}$$

To conclude, we use that $\operatorname{Lip}(M^{\hat{\pi}}) \leq 1$ by Corollary 2. ■

6. Discussion and summary

Our results unify, strengthen and extend previous works. The unifying framework of factored linear models was introduced by Yao et al. (2014). The focus of the present work is the derivation of policy error bounds, while putting issues of designing and analyzing algorithms to learn models aside. We believe that in fact this should be the preferred approach to developing theories for reinforcement learning: By first figuring out what quantities control the policy error in a given error, one is in a better position to design learning algorithms which then control the said quantities (this is distantly reminiscent to choosing surrogate losses in supervised learning).

Previous work that derives policy error bounds goes back to at least Whitt (1978). In fact, looking at the literature we see that the results of Whitt (1978) have been independently rederived in part or as a whole multiple times (often confounded with the issue of statistical questions), e.g., in the works mentioned in Section 3. Compared to the work of Whitt (1978), main advances in deriving policy error bounds have been the introduction of norms other than the supremum norm, though this happened in different contexts (e.g., de Farias and Van Roy 2003; Munos 2003), and breaking down the bound of Whitt (1978) to more specialized models (e.g., Ormoneit and Sen 2002; Ravindran 2004; Barreto et al. 2011; Sorg and Singh 2009).

One of the main novelties of the present work is that we are importing previous techniques to model-based RL to obtain policy error bounds in norms other than (unweighted) supremum norms.

In particular, to derive policy error bounds that use weighted supremum norms, we are building on the work of de Farias and Van Roy (2003), and we bring Lyapunov analysis from the approximate linear programming (ALP) methodology to model-based RL. At the same time, to derive policy error bounds that use weighted L^p-norms we import ideas from Munos (2003), who analyzed approximate dynamic programming (ADP) algorithms. During this process we streamlined the definitions from these works by sticking to the language of operator algebras (specifically, Banach lattices). The use of this language has two main benefits: It allowed us to present shorter and rather direct proofs, while it also shed light on the algebraic and geometric assumptions that were key in the proofs. We believe that our operator algebra approach could also improve previous results in either ALP or ADP. An interesting avenue for further work is to investigate the minimum set of assumptions under which our calculations remain valid: At present it appears that we use very little of the rich structure of the function spaces involved. We speculate that the results can also be proven in certain max-plus (a.k.a. tropical) algebras, leading to results that may hold, *e.g.*, for various versions of sequential games.

Another major novel aspect of the present work is that we tightened previous bounds. In particular, our bounds come in two forms: One form (the "ε_1" term) tells us how model errors should be controlled in the spaces of compressed value functions, while the other form (the "ε_2" term) tells us that it is enough if the model operator approximates the true model operator at only the (uncompressed) value function derived *from the model*.

While we shorten and improve previous results, we also managed to relax the key condition of previous works that required that the Bellman operator acting on uncompressed value functions and underlying the model needs to be a contraction. While we are still relying on contraction-type arguments, the contraction arguments are used with the compressed space, as previously suggested (but not analyzed) by Yao et al. (2014). We feel that it is more natural to require that the Bellman operator for the compressed space is a contraction than to require the same for the respective operator acting on the uncompressed space. Indeed, our bounds show that this second assumption is entirely superfluous (cf. the "ε_2" terms).

One limitation of the results presented so far is that we assumed that $\mathcal{R}$ was a join-homomorphism. In many models, such as state-aggregation (soft or not) or stochastic factorization Van Roy (2006); Barreto et al. (2011), $\mathcal{R}$ is linear (and stochastic) but is *not* a join-homomorphism. Investigating our proofs reveals that we can allow $\mathcal{R}$ to be a linear operator (and $\mathcal{R}^{\mathcal{A}}$ to be a linear operator s.t. $(\mathcal{R}^{\mathcal{A}})^a \neq \mathcal{R}$) at the price of introducing additional error terms. For the sake of illustration, in Theorem 12 below we present a version of the $L^p(\mu)$-norm bounds (and a sketch of proof) that can be obtained for such operators.

For presenting Theorem 12, we will use the greedy action selector in the compressed space as well, *i.e.* G' mapping compressed action value functions to policies in $\mathcal{W}$ (*i.e.*, $M'^{G'w}w = M'w$ for $w \in \mathcal{W}^{\mathcal{A}}$). It is important to recall the definition of U^* for Theorem 12: $U^* = MT_{\mathcal{Q}}u^*$. Note that if $\mathcal{R}M = M'\mathcal{R}^{\mathcal{A}}$, then we also have $U^* = MT_{\mathcal{Q}\mathcal{R}}U^*$, and we can recover Theorem 11 from Theorem 12. However, U^* is not a fixed point of $MT_{\mathcal{Q}\mathcal{R}}$ in general when $\mathcal{R}$ is not a join-homomorphism, a fact that will be important in our discussion below.

Theorem 12 ($L^p(\mu)$ norm bound for linear $\mathcal{R}, \mathcal{R}^{\mathcal{A}}$) *Let $\hat{\pi}$ be the policy derived from the factored linear model defined using (1) and (2). Choose the norms so that $\|\cdot\|_{\mathcal{V}} = \|\cdot\|_{\mu,p}$ and $\|\cdot\|_{\mathcal{W}} = \|\cdot\|_{\infty,\eta}$. Assume that $\mathcal{R}, \mathcal{R}^{\mathcal{A}}$ are linear (but not necessarily join-homomorphisms, and $(\mathcal{R}^{\mathcal{A}})^a$ not* *necessarily equal to $\mathcal{R}$). If Assumptions 3 and 5 hold, then*

$$\left\|V^* - V^{\hat{\pi}}\right\|_{\mu,p} \leq \varepsilon_1(V^*, M') + \min\left\{\varepsilon_1(V^{\hat{\pi}}, M'^{G' T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}} u^*}), \varepsilon_2\right\}, \tag{9}$$

where

$$\begin{aligned}
\varepsilon_1(V, N') &= \gamma \left\|(\mathcal{P} - \mathcal{Q}\mathcal{R})V\right\|_{\mu,p} \\
&\qquad + \frac{B\gamma}{1-\gamma}\left(\left\|\mathcal{R}V - N'\mathcal{R}^{\mathcal{A}} T_{\mathcal{P}} V\right\|_{\infty,\eta} + \gamma\left\|\mathcal{R}^{\mathcal{A}}(\mathcal{P} - \mathcal{Q}\mathcal{R})V\right\|_{\infty,\eta}\right), \\
\varepsilon_2 &= C_{\gamma,\hat{\pi},\mathcal{P},\mu,\xi}\,\frac{1}{1-\gamma}\left\|\mathcal{P}U^* - \mathcal{Q}u^*\right\|_{\xi,p},
\end{aligned}$$

and where $C_{\gamma,\hat{\pi},\mathcal{P},\mu,\xi}$ is defined in (8).

Proof (Sketch) The ε_1 terms are obtained by appropriately modifying Lemma 15 (which is an intermediate result, presented in the appendix, that is used in the proof of Lemma 9), as we describe below. We will take $V = V^*$ ($V = V^{\hat{\pi}}$), $N = M$ (resp. $N = M^{\hat{\pi}}$) and $N' = M'$ (resp. $N' = M'^{G' T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}} u^*}$. Then the identity $u^* = N' T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}} u^*$ holds.

Because we cannot use the identity $\mathcal{R}M = M'\mathcal{R}^{\mathcal{A}}$, we need to use the following chain of inequalities:

$$\begin{aligned}
\left\|\mathcal{R}(V - U^*)\right\| &= \inf_{k\geq 1} \frac{1}{1-\alpha^k}\left\|\mathcal{R}V - (N' T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}})^k \mathcal{R}V\right\| \\
&\leq \frac{1}{1-\alpha}\left\|\mathcal{R}N T_{\mathcal{P}} V - N'\mathcal{R}^{\mathcal{A}} T_{\mathcal{Q}\mathcal{R}} V\right\| \\
&\leq \frac{1}{1-\alpha}\left(\left\|\mathcal{R}N T_{\mathcal{P}} V - N'\mathcal{R}^{\mathcal{A}} T_{\mathcal{P}} V\right\| + \left\|N'\mathcal{R}^{\mathcal{A}} T_{\mathcal{P}} V - N'\mathcal{R}^{\mathcal{A}} T_{\mathcal{Q}\mathcal{R}} V\right\|\right) \\
&\leq \frac{1}{1-\alpha}\left(\left\|\mathcal{R}N T_{\mathcal{P}} V - N'\mathcal{R}^{\mathcal{A}} T_{\mathcal{P}} V\right\| + \gamma\left\|\mathcal{R}^{\mathcal{A}}(\mathcal{P} - \mathcal{Q}\mathcal{R})V\right\|\right).
\end{aligned}$$

To obtain ε_2, we cannot use that $U^* = M T_{\mathcal{Q}\mathcal{R}} U^*$, so we simply write

$$\begin{aligned}
\left\|U^* - M^{\hat{\pi}} T_{\mathcal{P}} U^*\right\| &= \left\|M^{\hat{\pi}} T_{\mathcal{Q}} u^* - M^{\hat{\pi}} T_{\mathcal{P}} U^*\right\| \\
&\leq \left\|\mathcal{Q}u^* - \mathcal{P}U^*\right\|.
\end{aligned}$$

The above can be used to modify Lemma 9 as well, leading to analogues of Theorems 8 and 10 where $\mathcal{R}$ is linear but not a join-homomorphism. ■

Note that both this result and Theorem 11 show a curious scaling as a function of $1/(1-\gamma)$. In fact, the astute reader may recall that policy error bounds typically scale with $1/(1-\gamma)^2$. A little thinking reveals that our result may be subject to the same scaling: Just like in Theorem 7, where $\widetilde{V}$ hides $1/(1-\gamma)$, in the above bounds the value functions themselves bring in another $1/(1-\gamma)$, too. Is the scaling with $1/(1-\gamma)^2$ necessary? The answer is no: Theorem 4.1 of Van Roy (2006) shows that in some version of state-aggregation the policy error can scale with $1/(1-\gamma)$ only (as a sidenote, the only result so far with this property). Thus, it may be worthwhile to look at the differences between Theorem 4.1 and the above result. First, recall that in his Theorem 4.1 Van Roy (2006) bounds the error of the policy $\tilde{\pi}$ that is greedy with respect to the fixed point $\widetilde{U}^*$

of $MT_{\mathcal{QR}}$, where $\mathcal{R} = \mathcal{R}_{\tilde{\pi}}$ is chosen to depend on the policy (for some policy π, $\mathcal{R}_\pi$ is a weighted Euclidean projection to the compressed space induced by the aggregation, where the weights depend on the stationary distribution of π). Formally, the policy is defined by $\tilde{\pi} = GT_{\mathcal{QR}_{\tilde{\pi}}}\tilde{U}^*$ where $\tilde{U}^* = MT_{\mathcal{QR}_{\tilde{\pi}}}\tilde{U}^*$. Thus, the policy whose error he bounds is different from ours in two respects: As pointed out above, $U^* = MT_{\mathcal{QR}}u^*$ (that our $\hat{\pi}$ is greedy with respect to) is not necessarily the fixed point of $MT_{\mathcal{QR}}$. Further, our result is proven for general $\mathcal{R}$. At this time it is not clear whether with a specific choice of $\mathcal{R}$ (like $\mathcal{R}_{\hat{\pi}}$) the terms involved in the definition of ε_1 would cancel the additional $1/(1-\gamma)$ factor. For what it is worth, we note that for the "counterexample" that Van Roy (2006) presents, when $\mathcal{R} = \mathcal{R}_{\hat{\pi}}$, ε_1 scales with $1/(1-\gamma)$ only (as opposed to scaling with $1/(1-\gamma)^2$), showing that our bound has the ability to exploit the benefits of a "good" choice of $\mathcal{R}$. However, it remains to be seen whether this or some other systematic way of choosing $\mathcal{R}$ will always cancel the extra $1/(1-\gamma)$ factor.

To summarize, this paper advances our understanding of model errors on policy error in reinforcement learning. We do this by improving previous bounds by using a versatile set of norms and introducing a completely new bound which has the potential of better scaling with the discount factor, while at the same time we extend the range of the models by relaxing previous assumptions. We also showed that (some) of our bounds are unimprovable. By effectively using the language of Banach lattices, our proofs are shorter, while at the same time hold the promise of being generalizable beyond MDPs. We believe that our approach may lead to advances in the analysis and design of alternate approaches to reinforcement learning, namely both in approximate linear programming and approximate dynamic programming.

Acknowledgements

This work was supported by Alberta Innovates Technology Futures and NSERC.

References

André Barreto, Doina Precup, and Joelle Pineau. Practical kernel-based reinforcement learning. *arXiv preprint arXiv:1407.5358*, 2014a.

André M. S. Barreto and Marcelo D Fragoso. Computing the stationary distribution of a finite Markov chain through stochastic factorization. *SIAM Journal on Matrix Analysis and Applications*, 32(4):1513–1523, 2011.

André M. S. Barreto, Joelle Pineau, and Doina Precup. Policy iteration based on stochastic factorization. *Journal of Artificial Intelligence Research*, 50:763–803, 2014b.

André S. M. Barreto, Doina Precup, and Joelle Pineau. Reinforcement learning using kernel-based stochastic factorization. In J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 720–728. Curran Associates, Inc., 2011.

Dimitri P. Bertsekas. *Dynamic programming and optimal control*, volume 2. Athena Scientific, 1995.

Dimitri P. Bertsekas. Approximate policy iteration: A survey and some new methods. *Journal of Control Theory and Applications*, 9(3):310–335, 2011.

Dimitri P. Bertsekas. Weighted sup-norm contractions in dynamic programming: A review and some new applications. *Dept. Elect. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, Tech. Rep. LIDS-P-2884*, 2012.

Lucian Busoniu, Robert Babuska, Bart De Schutter, and Damien Ernst. *Reinforcement Learning and Dynamic Programming Using Function Approximators*. CRC Press, Inc., Boca Raton, FL, USA, 1st edition, 2010. ISBN 1439821089, 9781439821084.

Daniela Pucci de Farias and Benjamin Van Roy. The linear programming approach to approximate dynamic programming. *Operations Research*, 51(6):850–865, 2003.

Amir Massoud Farahmand, Ré mi Munos, and Csaba Szepesvá ri. Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems 23*, pages 568–576, 2010.

Steffen Grünewälder, Luca Baldassarre, Massimiliano Pontil, Arthur Gretton, and Guy Lever. Modeling transition dynamics in MDPs with RKHS embeddings of conditional distributions. *CoRR*, abs/1112.4722, 2011.

Steffen Grünewälder, Guy Lever, Luca Baldassarre, Massimilano Pontil, and Arthur Gretton. Modelling transition dynamics in MDPs with RKHS embeddings. In John Langford and Joelle Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning*, ICML'12, pages 535–542. Omnipress, 2012. ISBN 978-1-4503-1285-1.

Branislav Kveton and Georgios Theocharous. Kernel-based reinforcement learning on representative states. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, AAAI'12, pages 977–983. AAAI Press, 2012.

Guy Lever, John Shawe-Taylor, Ronnie Stafford, and Csaba Szepesvári. Compressed conditional mean embeddings for model-based reinforcement learning. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16. AAAI Press, 2016.

Peter Meyer-Nieber. *Banach Lattices*. Springer, 1991.

Rémi Munos. Error bounds for approximate policy iteration. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 560–567, 2003.

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

Doina Precup, Joelle Pineau, and André S Barreto. On-line reinforcement learning using incremental kernel-based stochastic factorization. In *Advances in Neural Information Processing Systems 25*, pages 1484–1492, 2012.

Martin L. Puterman. *Markov Decision Processes — Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 1994.

Balaraman Ravindran. *An algebraic approach to abstraction in reinforcement learning*. PhD thesis, University of Massachusetts Amherst, 2004.

S Singh, T Jaakkola, and M Jordan. Reinforcement learning with soft state aggregation. In *NIPS-7*, pages 361–368, 1995.

Satinder P Singh and Richard C Yee. An upper bound on the loss from approximate optimal-value functions. *Machine Learning*, 16(3):227–233, 1994.

Jonathan Sorg and Satinder Singh. Transfer via soft homomorphisms. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems – Volume 2*, pages 741–748. International Foundation for Autonomous Agents and Multiagent Systems, 2009.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning)*. The MIT Press, 1998. ISBN 0262193981.

Csaba Szepesvári. *Algorithms for Reinforcement Learning*. Morgan & Claypool, 2010.

Benjamin Van Roy. Performance loss bounds for approximate value iteration with state aggregation. *Mathematics of Operations Research*, 31(2):234–244, 2006.

Ward Whitt. Approximations of dynamic programs, I. *Mathematics of Operations Research*, 3(3): 231–243, 1978.

Marco Wiering and Martijn van Otterlo, editors. *Reinforcement Learning: State-of-the-Art*. Springer, 2012.

Hengshuai Yao, Csaba Szepesvári, Bernardo Avila Pires, and Xinhua Zhang. Pseudo-MDPs and factored linear action models. In *IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL), 2014*, pages 1–9. IEEE, 2014.

Appendix A. List of operators

For ease of reference, we present Table 1, which gives a summary of the operators we define and use.

Operator	Between	Definition
$\mathcal{P}^a$	$\mathcal{V} \to \mathcal{V}$	$(\mathcal{P}^a V)(x) \doteq \int V(x')\mathcal{P}^a(dx'\vert x)$
$\mathcal{Q}^a$	$\mathcal{W} \to \mathcal{V}$	$(\mathcal{Q}^a w)_i \doteq \int w_{i'} \mathcal{Q}^a(di'\vert i)$
$\mathcal{P}$	$\mathcal{V} \to \mathcal{V}^{\mathcal{A}}$	$(\mathcal{P}V)(a)(= (\mathcal{P}V)^a) \doteq \mathcal{P}^a V$
$\mathcal{Q}$	$\mathcal{W} \to \mathcal{V}^{\mathcal{A}}$	$(\mathcal{Q}V)(a)(= (\mathcal{Q}V)^a) \doteq \mathcal{Q}^a V$
$\mathcal{R}$	$\mathcal{V} \to \mathcal{W}$	almost always a join-homomorphism
$\mathcal{R}^{\mathcal{A}}$	$\mathcal{V}^{\mathcal{A}} \to \mathcal{W}^{\mathcal{A}}$	$(\mathcal{R}^{\mathcal{A}} V)^a = \mathcal{R}V, \ \forall a \in \mathcal{A}$
$\mathcal{R}^{\mathcal{A}}\mathcal{Q}$	$\mathcal{W} \to \mathcal{W}^{\mathcal{A}}$	$(\mathcal{R}^{\mathcal{A}}\mathcal{Q}w)(a)(= (\mathcal{R}^{\mathcal{A}}\mathcal{Q}w)^a) \doteq \mathcal{R}\mathcal{Q}^a w$
M	$\mathcal{V}^{\mathcal{A}} \to \mathcal{V}$	$(MV)(x) \doteq \max_a V^a(x)$
M'	$\mathcal{W}^{\mathcal{A}} \to \mathcal{W}$	$(M'w)_i \doteq \max_a w_i^a$
M^π	$\mathcal{V}^{\mathcal{A}} \to \mathcal{V}$	$(M^\pi V)(x) \doteq V^{\pi(x)}(x)$
M'^π	$\mathcal{W}^{\mathcal{A}} \to \mathcal{W}$	$(M'^\pi w)_i \doteq w_i^{\pi(i)}$
G	$\mathcal{V}^{\mathcal{A}} \to \Pi$	$GV(x) \doteq \operatorname{argmax}_a V^a(x)$
$T_{\mathcal{P}}$	$\mathcal{V} \to \mathcal{V}^{\mathcal{A}}$	$T_{\mathcal{P}} V \doteq r + \gamma \mathcal{P} V$
$T_{\mathcal{Q}}$	$\mathcal{W} \to \mathcal{V}^{\mathcal{A}}$	$T_{\mathcal{Q}} w \doteq r + \gamma \mathcal{Q} w$
$T_{\mathcal{Q}\mathcal{R}}$	$\mathcal{V} \to \mathcal{V}^{\mathcal{A}}$	$T_{\mathcal{Q}\mathcal{R}} V \doteq T_{\mathcal{Q}}\mathcal{R}V = r + \gamma \mathcal{Q}\mathcal{R}V$
$T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}}$	$\mathcal{W} \to \mathcal{W}^{\mathcal{A}}$	$T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}} w \doteq \mathcal{R}^{\mathcal{A}} T_{\mathcal{Q}} w = \mathcal{R}^{\mathcal{A}} r + \gamma \mathcal{R}^{\mathcal{A}}\mathcal{Q} w$

Table 1: Definitions of operators used in the paper.

Appendix B. General results

In this section, we present some general technical results.

Proposition 13 *Consider an operator $T : \mathcal{V} \to \mathcal{V}$ mapping a normed space $(\mathcal{V}, \|\cdot\|_{\mathcal{V}})$ to itself. If $\mathrm{Lip}(T) < \infty$ and T^m is a contraction for some $m > 0$, then T has a unique fixed point.*

Proof Banach's fixed point theorem ensures that T^m has a unique fixed point V, which must also be the unique fixed point of T^{m^2} and $T^{m(m+1)}$, so $V = T^{m(m+1)} = TT^{m^2}V = TV$, so V is a fixed point of T. Since every fixed point of T is also a fixed point of T^m, it follows that V is the unique fixed point of T. ■

Lemma 6 *Let $(\mathcal{V}, \|\cdot\|_{\mathcal{V}})$ and $(\mathcal{W}, \|\cdot\|_{\mathcal{W}})$ be two Banach spaces. Let $T : \mathcal{W} \to \mathcal{V}$ and $H : \mathcal{V} \to \mathcal{W}$ be two operators such that $\mathrm{Lip}((HT)^m) < 1$ for some $m > 0$. Then HT has a unique fixed point W^*, and $V^* \doteq TW^*$ is the unique fixed point of TH.*

Proof Since $(HT)^m$ is a contraction, Proposition 13 ensures that $(HT)^m$ has a unique fixed point W^*, which is also the unique fixed point of HT. Defining $V^* \doteq TW^*$, we can see that $THV^* = THTW^* = TW^* = V^*$. It remains to show that V^* is the unique fixed point of TH, so let us assume that there exists $V' \neq V^*$ s.t. $V' = THV'$. Then with $W' \doteq HV'$ we have $TW' = V'$. Now, $HTW' = HV' = W'$, so W' is a fixed point of HT, which implies $W' = W^*$, since the fixed point of HT is unique, but then $V' = TW' = TW^* = V^*$, which is a contradiction. ■

Lemma 14 *Let $(\mathcal{V}, \|\cdot\|)$ be a Banach space and $T : \mathcal{V} \to \mathcal{V}$ be an operator. Assume that there exists $V^* \in \mathcal{V}$ such that $TV^* = V^*$, and that there exist constants $a < 1$ and b such that for all $m \geq 0$ we have $\operatorname{Lip}(T^{m+1}) \leq ba^m$. Then for all $V \in \mathcal{V}$ and $m \geq 0$ such that $ba^m < 1$,*

$$\|V - V^*\| \leq \frac{1}{1 - ba^m} \left\|V - T^{m+1}V\right\|.$$

Further, if we take the infimum of both sides for m such that $ba^m < 1$, we get an equality.

Proof We have that for all $m \geq 0$,

$$\begin{aligned}
\|V - V^*\| &= \left\|V - T^{m+1}V^*\right\| \\
&= \left\|V - T^{m+1}V + T^{m+1}V - T^{m+1}V^*\right\| \\
&\leq \left\|V - T^{m+1}V\right\| + \left\|T^{m+1}V - T^{m+1}V^*\right\| \\
&\leq \left\|V - T^{m+1}V\right\| + ba^m \|V - V^*\|.
\end{aligned}$$

To arrive at an upper-bound, we need to move the third term to the right-hand side and divide the inequality by $1 - ba^m$. The inequality is preserved after division only for those m when $ba^m < 1$, giving the result.

To see why we get the equality, note that $T^\infty V = V^*$. Hence,

$$\inf_{m:ba^m<1} \frac{1}{1 - ba^m} \left\|V - T^{m+1}V\right\| \leq \|V - V^*\|.$$

■

Proposition 3 *Assume that $\mathcal{V} = \mathbb{R}^m$ and $\mathcal{W} = \mathbb{R}^n$, and let $\mathcal{R}$ be any linear join-homomorphism. Then there exists $a \in \mathbb{R}^n_+$ and $J \in [m]^n$ s.t. $(\mathcal{R}v)_i = a_i\, v_{J_i}$ for all $v \in \mathcal{V}$ and $i \in [n]$.*

Proof Consider $v \geq 0$. We can write $v = \sum_{j=1}^m v_j e_j$, where $(e_j)_{j=1}^m$ is the Euclidean basis. Because $v \geq 0$, we can also write $v = \bigvee_{j=1}^m v_j e_j$. By linearity of $\mathcal{R}$, we have that $\mathcal{R}v = \sum_{j=1}^m v_j \mathcal{R}e_j$, and since $\mathcal{R}$ is a join-homomorphism and linear, we also have $\mathcal{R}v = \bigvee_{j=1}^m \mathcal{R}(v_j e_j) = \bigvee_{j=1}^m v_j \mathcal{R}e_j$.

Next, we show that for all i, $(\mathcal{R}e_j)_i \neq 0$ for at most one $j \in [m]$. Taking v s.t. $v_i = 1$ for all i, we have that for all $i \in [n]$

$$\left(\sum_{j=1}^m \mathcal{R}e_j\right)_i = (\mathcal{R}v)_i = \left(\bigvee_{j=1}^m \mathcal{R}e_j\right)_i,$$

which implies that for all $i \in [n]$ there is at most one $j \in [m]$ s.t. $(\mathcal{R}e_j)_i > 0$, and J_i is defined as such j if it exists, otherwise arbitrary. Defining $a_i \doteq (\mathcal{R}e_{J_i})_i$ for $(i \in [n])$ gives the result. ■

Proposition 4 *Given $(\mathcal{U}, \|\cdot\|_{\infty,\nu})$ with $\nu : \mathcal{U} \to \mathbb{R}_+$, and $J : \mathcal{U} \to \mathcal{U}^{\mathcal{A}}$, if each J^a is a linear operator, then $\gamma \operatorname{Lip}(J) = \beta_{\nu,J}$.*

Proof Define $A(U) \doteq \{U' \in \mathcal{U} : |U'| = |U|\}$ for $U \in \mathcal{U}$. Since J is linear, $\operatorname{Lip}(J) = \|J\|$. Since $\|\cdot\| \doteq \|\cdot\|_{\infty,\nu}$ is a type of supremum norm, $\|J\| = \max_a \|J^a\|$ (the maximum over the actions and states commute). Thus, we have that

$$\begin{aligned}
\operatorname{Lip}(J) &= \sup_{U:\|U\|=1} \max_a \|J^a U\| \\
&= \sup_{U:\|U\|=1} \max_a \sup_x \frac{|(J^a U)(x)|}{\nu(x)} \\
&= \sup_x \max_a \sup_{U>0:\|U\|=1} \sup_{U' \in A(U)} \frac{|(J^a U')(x)|}{\nu(x)}.
\end{aligned}$$

Note that equality still holds in the last line by equivalence of the suprema with the supremum on the previous line. The term $\sup_{U' \in A(U)} \frac{|(J^a U')(x)|}{\nu(x)}$ can be maximized w.r.t. U by maximizing U subject to $\frac{U(x)}{\nu(x)} \leq 1$, for all $x \in \mathcal{X}, a \in \mathcal{A}$. Therefore the term is maximized by $U = \nu$, and, since $A(\nu) = \{U' \in \mathcal{U} : |U'| = \nu\}$, we get

$$\gamma \operatorname{Lip}(J) = \gamma \sup_{U:|U|=\nu} \|JU\| = \beta_{\nu,J}.$$

■

Appendix C. MDP-specific results

In this section, we present accessory results and proofs omitted from the main text. Lemma 15 is an intermediate result for Lemma 9. The proof of Lemma 9 is also presented here. Moreover, we present the proof of three omitted results: Propositions 16 and 17 and Theorem 18, respectively a tightness example for Theorem 8, an example showing that the Theorem 8 can be harsh, and a weighted supremum norm bound for the policy error in $L^p(\mu)$ norm.

Lemma 15 *Let Assumptions 1 and 5 hold, and assume that $\gamma \operatorname{Lip}(\mathcal{R}^{\mathcal{A}}\mathcal{Q}) \leq \alpha < 1$. Then, for $V \in \left\{V^*, V^{\hat{\pi}}\right\}$*

$$\left\|V - U^*\right\| \leq \gamma \left\|(\mathcal{P} - \mathcal{Q}\mathcal{R})V\right\| + \frac{B\gamma^2}{1-\alpha} \left\|\mathcal{R}^{\mathcal{A}}(\mathcal{P} - \mathcal{Q}\mathcal{R})V\right\|,$$

Proof Using that $V^* = MT_{\mathcal{P}}V^*$, $V^{\hat{\pi}} = M^{\hat{\pi}}T_{\mathcal{P}}V^{\hat{\pi}}$ and $M^{\hat{\pi}}T_{\mathcal{Q}\mathcal{R}}U^* = U^* = MT_{\mathcal{Q}\mathcal{R}}U^*$, we first upper-bound, with $N = M$ ($N = M^{\hat{\pi}}$) and $V = V^*$ ($V = V^{\hat{\pi}}$),

$$\begin{aligned}
\|V - U^*\| &= \|NT_{\mathcal{P}}V - NT_{\mathcal{Q}\mathcal{R}}U^*\| \\
&\leq \|NT_{\mathcal{P}}V - NT_{\mathcal{Q}\mathcal{R}}V\| + \|NT_{\mathcal{Q}\mathcal{R}}V - NT_{\mathcal{Q}\mathcal{R}}U^*\| \\
&\leq \gamma \operatorname{Lip}(N) \|(\mathcal{P} - \mathcal{Q}\mathcal{R})V\| + \operatorname{Lip}(NT_{\mathcal{Q}}) \|\mathcal{R}(V - U^*)\| \\
&\leq \gamma \|(\mathcal{P} - \mathcal{Q}\mathcal{R})V\| + B\gamma \|\mathcal{R}(V - U^*)\|.
\end{aligned}$$

Given $N \in \{M, M^{\hat{\pi}}\}$, we define N' as the operator satisfying $\mathcal{R}N = N'\mathcal{R}^{\mathcal{A}}$. In particular, if $N = M$, then $N' = M'$, otherwise $N' = M'^{\pi_2}$ for some $\pi_2 : \mathcal{I} \to \mathcal{A}$ (in either case N' is well-defined because $\mathcal{R}$ is a join-homomorphism, cf. Assumption 2). Since $\mathrm{Lip}(N') \leq 1$ by Assumption 1, we get $\mathrm{Lip}(N'T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}}) \leq \gamma \,\mathrm{Lip}(\mathcal{R}^{\mathcal{A}}\mathcal{Q}) \leq \alpha < 1$. Lemma 14 with $T = N'T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}}$ and $a = b = \alpha < 1$, combined gives for $N = M$ ($N = M^{\hat{\pi}}$) and $V = V^*$ ($V = V^{\hat{\pi}}$),

$$\begin{aligned}
\|\mathcal{R}(V - U^*)\| &= \inf_{k\geq 1} \frac{1}{1-\alpha^k} \left\|\mathcal{R}V - (N'T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}})^k \mathcal{R}V\right\| \\
&\leq \frac{1}{1-\alpha}\left\|N'\mathcal{R}^{\mathcal{A}}T_{\mathcal{P}}V - N'\mathcal{R}^{\mathcal{A}}T_{\mathcal{Q}\mathcal{R}}V\right\| \\
&\leq \frac{\gamma}{1-\alpha}\,\mathrm{Lip}(N')\left\|\mathcal{R}^{\mathcal{A}}(\mathcal{P} - \mathcal{Q}\mathcal{R})V\right\| \\
&\leq \frac{\gamma}{1-\alpha}\left\|\mathcal{R}^{\mathcal{A}}(\mathcal{P} - \mathcal{Q}\mathcal{R})V\right\|,
\end{aligned}$$

where we have also used that $V = NT_{\mathcal{P}}V$ and that $\mathcal{R}N = N'\mathcal{R}^{\mathcal{A}}$. Combining the above gives,

$$\begin{aligned}
\|V - U^*\| &\leq \gamma \left\|(\mathcal{P} - \mathcal{Q}\mathcal{R})V\right\| + B\gamma \left\|\mathcal{R}(V - U^*)\right\| \\
&\leq \gamma \left\|(\mathcal{P} - \mathcal{Q}\mathcal{R})V\right\| + \frac{B\gamma^2}{1-\alpha}\left\|\mathcal{R}^{\mathcal{A}}(\mathcal{P} - \mathcal{Q}\mathcal{R})V\right\|.
\end{aligned}$$

■

Lemma 9 *Let Assumptions 1 and 5 hold, and assume that $\gamma \,\mathrm{Lip}(\mathcal{R}^{\mathcal{A}}\mathcal{Q}) \leq \alpha < 1$. For $V \in \{V^*, V^{\hat{\pi}}\}$ we have that*

$$\|V - U^*\| \leq \gamma \left\|(\mathcal{P} - \mathcal{Q}\mathcal{R})V\right\| + \frac{B\gamma^2}{1-\alpha}\left\|\mathcal{R}^{\mathcal{A}}(\mathcal{P} - \mathcal{Q}\mathcal{R})V\right\|. \tag{5}$$

Additionally, if $\gamma \,\mathrm{Lip}(\mathcal{P}) \leq \beta < 1$ (or, alternatively, $\gamma \,\mathrm{Lip}(M^{\hat{\pi}}\mathcal{P}) \leq \beta < 1$), we also have for $V = V^*$ (respectively, $V = V^{\hat{\pi}}$) that

$$\|V - U^*\| \leq \frac{\gamma}{1-\beta}\left\|(\mathcal{P} - \mathcal{Q}\mathcal{R})U^*\right\|. \tag{6}$$

Proof Recall that V^* is the optimal value function, *i.e.*, the fixed point of the Bellman optimality equation $V^* = MT_{\mathcal{P}}V^*$. Recall also that $V^{\hat{\pi}}$ is the value function of $\hat{\pi}$ and the fixed point of the Bellman equation $V^{\hat{\pi}} = M^{\hat{\pi}}T_{\mathcal{P}}V^{\hat{\pi}}$. Lemma 15 gives us (5) directly.

To prove (6) for $V = V^*$, we use Lemma 14 with $T = MT_{\mathcal{P}}$ and $a = b = \beta$, which gives

$$\begin{aligned}
\|V^* - U^*\| &= \inf_{k\geq 1} \frac{1}{1-\beta^k}\left\|U^* - (MT_{\mathcal{P}})^k U^*\right\| \\
&\leq \frac{1}{1-\beta}\left\|MT_{\mathcal{Q}\mathcal{R}}U^* - MT_{\mathcal{P}}U^*\right\| \\
&\leq \frac{\gamma}{1-\beta}\,\mathrm{Lip}(M)\left\|(\mathcal{P} - \mathcal{Q}\mathcal{R})U^*\right\|,
\end{aligned}$$

and then we plug in $\mathrm{Lip}(M) \leq 1$. For (6) for $V = V^{\hat\pi}$, we observe that $\mathrm{Lip}(M^{\hat\pi} T_{\mathcal{P}}) = \gamma \, \mathrm{Lip}(M^{\hat\pi}\mathcal{P})$, then we follow a similar approach:

$$\begin{aligned}
\left\| V^{\hat\pi} - U^* \right\| &= \inf_{k \geq 1} \frac{1}{1-\beta^k} \left\| U^* - (M^{\hat\pi} T_{\mathcal{P}})^k U^* \right\| \\
&\leq \frac{1}{1-\beta} \left\| M^{\hat\pi} T_{\mathcal{Q}\mathcal{R}} U^* - M^{\hat\pi} T_{\mathcal{P}} U^* \right\| \\
&\leq \frac{\gamma}{1-\beta} \mathrm{Lip}(M^{\hat\pi}) \left\| (\mathcal{P} - \mathcal{Q}\mathcal{R}) U^* \right\|,
\end{aligned}$$

and plug in $\mathrm{Lip}(M^{\hat\pi}) \leq 1$. ■

The recipe for constructing the example proving Proposition 16 is simple: i) create a three-state, two-action MDP with a "fork" state s_1 leading to a high-value terminal state s_2 with action a_1 and a low-value terminal state s_3 with action a_2; ii) choose the rewards so that the immediate reward $r^{a_2}(s_1) > r^{a_1}(s_1)$, while the value of (s_1, a_1) is higher than (s_1, a_2); iii) make a poor model for the fork state s_1, so that $\hat\pi$ becomes nearsighted, picking a_2 rather than a_1. We can also perturb the model for s_3, in order to have a desired value for $\|V^* - U^*\|_\infty$. The rest of the effort pertains to choosing the rewards and the model carefully in order to have the correct value for $\|(\mathcal{P} - \mathcal{Q}\mathcal{R})U^*\|_\infty$. There is factor of $\frac{1}{\gamma}$ in the scaling of the rewards, as a result of requiring the return from s_1 after the first action to dominate the immediate reward at s_1, and the rewards also scale with $\max \tau$ for the bound to scale. The example underlying Proposition 16 is also well-defined for $\varepsilon = 0$, but then $G T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}} u^*$ is no longer unique: It can yield an optimal policy or a policy that is τ-suboptimal in $\|\cdot\|_\infty$, depending on how ties are broken.

Proposition 16 (Theorem 8 is tight) *There exist $\mathcal{P}, \mathcal{Q}$ and $\mathcal{R}$ s.t. for every $\gamma \in (0,1)$, $\tau \geq 0$ and $\varepsilon \in (0,1)$ there exists $r \in \mathcal{V}^{\mathcal{A}}$ (the rewards scale with $\frac{1-\gamma^2}{\gamma}\tau$) s.t. $\mathrm{Lip}(\mathcal{Q}\mathcal{R}) < \infty$, $\mathrm{Lip}(\mathcal{R}^{\mathcal{A}}\mathcal{Q}) \leq 1$, $\frac{2\gamma}{1-\gamma} \|(\mathcal{P} - \mathcal{Q}\mathcal{R})U^*\|_\infty = \tau$, and $\left\|V^* - V^{\hat\pi}\right\|_\infty = (1-\varepsilon)\tau$. Thus, Theorem 8 can be made arbitrarily tight.*

Proof The set of states is $\mathcal{X} = \{x_1, \ldots, x_3\}$, the set of actions is $\mathcal{A} = \{a_1, a_2\}$ and the transition probability kernel is specified by $\mathcal{P}$ as follows:

$$\mathcal{P}^{a_1} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \mathcal{P}^{a_2} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

We let $\mathcal{W} = \mathbb{R}^2$ and $\mathcal{R}V \doteq (V(x_2), V(x_3))^\top$.

The model is

$$\mathcal{Q}^{a_1} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \mathcal{Q}^{a_2} = \begin{pmatrix} 0 & -1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Given $\gamma \in (0,1)$, $\tau \geq 0$ and $\varepsilon \in (0,1)$, define

$$r^{a_1} = \begin{pmatrix} -\frac{\tau}{4}(2\varepsilon + \gamma - 1) \\ \frac{\tau(1-\gamma^2)}{4\gamma} \\ -\frac{\tau(1-\gamma^2)}{4\gamma} \end{pmatrix}, \qquad r^{a_2} = \begin{pmatrix} \frac{\tau}{4}(2\varepsilon + \gamma - 1) \\ \frac{\tau(1-\gamma^2)}{4\gamma} \\ -\frac{\tau(1-\gamma^2)}{4\gamma} \end{pmatrix},$$

which gives $V^* = \frac{\tau}{4}\left(2(1-\varepsilon), \frac{1+\gamma}{\gamma}, -\frac{1+\gamma}{\gamma}\right)^\top$ and $U^* = \frac{\tau}{4}\left(2\varepsilon, \frac{1-\gamma}{\gamma}, -\frac{1-\gamma}{\gamma}\right)^\top$. Given that

$$(\mathcal{P}^{a_1} - \mathcal{Q}^{a_1}\mathcal{R})U^* = \begin{pmatrix} U_2^* + U_2^* \\ U_2^* - U_3^* \\ U_3^* - U_2^* \end{pmatrix}, \qquad (\mathcal{P}^{a_2} - \mathcal{Q}^{a_2}\mathcal{R})U^* = \begin{pmatrix} U_3^* + U_3^* \\ U_2^* - U_3^* \\ U_3^* - U_2^* \end{pmatrix},$$

and that $U_2^* = -U_3^*$, we have $\|(\mathcal{P} - \mathcal{Q}\mathcal{R})U^*\| = 2U_2^* = \frac{1-\gamma}{2\gamma}\tau$, which gives $\frac{2\gamma}{1-\gamma}\|(\mathcal{P} - \mathcal{Q}\mathcal{R})U^*\| = \tau$. It can be seen also that $\hat{\pi}(x_1) = a_2$ (and that the policy in x_2 and x_3 is irrelevant), so $\left\|V^* - V^{\hat{\pi}}\right\| = (1-\varepsilon)\tau$, since $r^{a_2}(x_1) + \gamma V_3^* = -V_1^* = -(1-\varepsilon)\frac{\tau}{2}$. We note in passing that $\|V^* - U^*\| = \left|\frac{\tau}{2} - \tau\varepsilon\right|$ and that $\left\|V^{\hat{\pi}} - U^*\right\| = \frac{\tau}{2}$. ■

Proposition 17 is based on the natural argument that the model does not need to be good in states that are not visited by an optimal policy: i) we can extend the example in Proposition 16 with an initial state with two actions: "stay", or "go to the fork state"; ii) we pick the value of staying to be higher than the value of going to the fork state; iii) we pick an accurate model at the initial state, so that both $\hat{\pi}$ and π^* choose to stay there (rather than go to the fork state). The policy error is zero when we take μ that puts measure one on the initial state, however $\hat{\pi}$ is still near-sighted in the fork state, and it suffers the supremum norm error outlined in Proposition 16.

Proposition 17 (The supremum norm is harsh) *There exist $\mathcal{P}$, $\mathcal{Q}$ and $\mathcal{R}$ s.t. for every $\gamma \in (0,1)$ and $\tau > 0$, there exists $r \in \mathcal{V}^{\mathcal{A}}$ (the rewards scale with $\frac{1-\gamma^2}{\gamma}\tau$) s.t. $\mathrm{Lip}(\mathcal{Q}\mathcal{R}) < \infty$, $\mathrm{Lip}(\mathcal{R}^{\mathcal{A}}\mathcal{Q}) \leq 1$, $\left\|V - V^{\hat{\pi}}\right\|_\infty = \tau$ and $\left\|V^* - V^{\hat{\pi}}\right\|_{\mu,p} = \left\|V^* - V^{\hat{\pi}}\right\|_{\xi,p} = 0$ where μ and ξ are stationary w.r.t. to π^* and $\hat{\pi}$, respectively.*

Proof Pick any $\tau' > 0$. Consider $\mathcal{P}, \mathcal{Q}, r$ as in Proposition 16, for the choice of $\varepsilon = \frac{1}{2}$ and $\tau' = \frac{\tau}{2}$. Add a state, x_4, to $\mathcal{X}$, redefine $\mathcal{R}V \doteq (V(x_2), V(x_3), V(x_4))^\top$, let $\mathcal{P}^{a_1}_{4,4} = 1$, $\mathcal{P}^{a_2}_{4,1} = 1$, $\mathcal{Q}^a_{i,4} = 0$ for all a and $i \neq 4$, and let also $\mathcal{Q}^a_{4,i} = \mathcal{P}^a_{4,i}$ for all a, i. Finally, let $r^{a_1}(x_4) = 2(1-\gamma)\tau'$ and $r^{a_2}(x_4) = 0$

Thus, $V_4^* = 2\tau'$, $\pi^*(x_4) = a_1$, $U_1^* = V_1^* = \tau'$, $\hat{\pi}(x_4) = a_1$ and $U_4^* = 2\tau'$. Moreover, the distribution μ (ξ) defined by $\mu(x_4) \doteq 1$ ($\xi(x_4) \doteq 1$) is stationary w.r.t. π^* ($\hat{\pi}$). This gives $\left\|V^* - V^{\hat{\pi}}\right\|_{\mu,p} = \left\|V^* - V^{\hat{\pi}}\right\|_{\xi,p} = 0$ as desired, and $\left\|V^* - V^{\hat{\pi}}\right\|_\infty = \tau'$, which implies the result. ■

To conclude, we present Theorem 18.

Theorem 18 (Weighted supremum norm bound for the policy error in $L^p(\mu)$ norm) *Let $\hat{\pi}$ be the policy derived from the factored linear model defined using (1) and (2). If Assumptions 4 and 5 holds for the weighted supremum norm over $\mathcal{V}^{\mathcal{A}}$ and $\mathcal{W}^{\mathcal{A}}$, then*

$$\left\|V^* - V^{\hat{\pi}}\right\|_{\mu,p} \leq \|\nu\|_{\mu,p}\left(\varepsilon(V^*) + \varepsilon(V^{\hat{\pi}})\right),$$

where $\varepsilon(V) = \min(\varepsilon_1(V), \varepsilon_2)$, and

$$\varepsilon_1(V) = \gamma\|(\mathcal{P} - \mathcal{Q}\mathcal{R})V\|_{\infty,\nu} + \frac{B\gamma^2}{1 - \beta_{\eta,\mathcal{R}^{\mathcal{A}}\mathcal{Q}}}\|\mathcal{R}(\mathcal{P} - \mathcal{Q}\mathcal{R})V\|_{\infty,\eta}$$
$$\varepsilon_2 = \frac{\gamma}{1 - \beta_{\nu,\mathcal{P}}}\left\|(\mathcal{P} - \mathcal{Q}\mathcal{R})U^*\right\|_{\infty,\nu}.$$

Proof (of Theorem 18) Since

$$\|V\|_\mu \le (\mu(\nu^p))^{\frac{1}{p}} \|\,|V|^p\,\|_{\infty,\nu^p} = \|\nu\|_{\mu,p} \|V\|_{\infty,\nu},$$

we can apply Theorem 10 to obtain the result. ■

Appendix D. Issues with bounding $\|U^* - V^*\|_\infty$ instead of the policy error

As we indicated in Section 5.2, Ormoneit and Sen (2002); Barreto et al. (2011); Barreto and Fragoso (2011); Precup et al. (2012); Barreto et al. (2014b,a) [3] bound $\|V^* - \widetilde{V}\|_\infty$ (not the policy error). We can show by counterexample that this is not is not the correct quantity to bound in order to understand the quality of $\hat{\pi}$, and that the policy error should be bounded instead. The recipe for constructing the counterexample proving this Proposition 19 is similar to the one used in Proposition 16.

Proposition 19 (Controlling only $\|U^* - V^*\|_\infty$ is not enough) *There exist $\mathcal{P}$, $\mathcal{Q}$ and $\mathcal{R}$ s.t. satisfying Assumptions 3 and 5 such that for every $\gamma \in (0,1)$, $\tau_1 \ge 0$ and $\tau_2 \ge 0$ there exists a reward function $r \in \mathcal{V}^{\mathcal{A}}$ with $\|r\|_\infty \le 2(\tau_1 \vee \tau_2)/\gamma$ s.t. $\|V^* - U^*\| = \tau_1$, $\left\|V^{\hat{\pi}} - U^*\right\| = \tau_2$ and $\left\|V^* - V^{\hat{\pi}}\right\|_\infty = \tau_1 + \tau_2$. The rewards scale proportionally to $\frac{1-\gamma}{\gamma}\max\{\tau_1, \tau_2\}$.*

Proof The set of states is $\mathcal{X} = \{x_1, \ldots, x_3\}$, the set of actions is $\mathcal{A} = \{a_1, a_2\}$ and the transition probability kernel is specified by $\mathcal{P}$ as follows:

$$\mathcal{P}^{a_1} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \mathcal{P}^{a_2} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

We let $\mathcal{W} = \mathbb{R}$ and $\mathcal{R}V \doteq (V(x_2))^\top$.

Given $\gamma \in (0,1)$, $\tau_1 \ge 0$ and $\tau_2 \ge 0$, define $\tau_{\max} = \max\{\tau_1, \tau_2\}$ and

$$r^{a_1} = \begin{pmatrix} \tau_1 \\ \frac{1-\gamma}{\gamma}(\tau_1 + \tau_{\max}) \\ -\frac{1-\gamma}{\gamma}\tau_2 \end{pmatrix}, \qquad r^{a_2} = \begin{pmatrix} \tau_1 + \tau_{\max} \\ \frac{1-\gamma}{\gamma}(\tau_1 + \tau_{\max}) \\ -\frac{1-\gamma}{\gamma}\tau_2 \end{pmatrix},$$

which gives $V^* = \left(2\tau_1 + \tau_{\max}, \frac{\tau_1 + \tau_{\max}}{\gamma}, -\frac{\tau_2}{\gamma}\right)^\top$.

Next, we construct $\mathcal{Q}$. First, we set $\mathcal{Q}^{a_1}_{1,1} = \mathcal{Q}^{a_2}_{1,1} = 0$. We want $u^* = (V^*_2)$ (which means $\mathcal{R}V \doteq V(x_2)$ for $V \in \mathcal{V}$), so we set $\mathcal{Q}^{a_1}_{2,2} = \mathcal{Q}^{a_2}_{2,2} = 1$. Since $U^* = MT_{\mathcal{Q}}u^*$, we have $U^*_2 = u^*_2$ and $U^*_1 = \max_a r^a(x_1) = \tau_1 + \tau_{\max} = V^*_1 - \tau_1$. We choose $\mathcal{Q}^a_{3,1} = -\frac{\tau_2}{\tau_1 + \tau_{\max} + \mathbb{I}\{\tau_{\max} = 0\}}$ (that is, if $\tau_{\max} = 0$, we set $\mathcal{Q}^a_{3,1} = 0$), so that

$$U^*_3 = \max_a r^a(x_3) + \gamma \mathcal{Q}^a_{3,2} u^*_2 = V^*_3.$$

3. To be precise, the proof of Ormoneit and Sen (2002)'s Theorem 2 implies that this quantity converges to zero as the model error converges to zero (their analysis confounds the estimation and approximation errors). Their Theorem 3, using an additional argument, is concerned with the probability of choosing a suboptimal action when using the approximate model.

To summarize,

$$\mathcal{Q}^{a_1} = \mathcal{Q}^{a_2} = \begin{pmatrix} 0 \\ 1 \\ -\frac{\tau_2}{\tau_1+\tau_{\max}+\mathbb{I}\{\tau_{\max}=0\}} \end{pmatrix}.$$

At this point, we can see that $\mathcal{Q}$ does not depend on γ, that $\mathrm{Lip}(\mathcal{Q}\mathcal{R}) < \infty$ and that $\mathrm{Lip}(\mathcal{R}^{\mathcal{A}}\mathcal{Q}) = 1$.

The policies obtained are given by $\pi^*(x_1) = 1$ and $\hat{\pi}(x_1) = 2$, while the choices for other states are irrelevant. This gives $V^{\pi} = (r^{a_2}(x_1) + \gamma V_3^*, V_2^*, V_3^*)^\top$, so that

$$\|V^* - V^{\hat{\pi}}\| = V_1^* - r^{a_2}(x_1) - \gamma V_3^* = \tau_1 + \tau_2.$$

Moreover,

$$\|V^* - U^*\| = |V_1^* - (V_1^* - \tau_1)| = \tau_1,$$

and

$$\|V^{\hat{\pi}} - U^*\| = V_1^* - \tau_1 - (r^{a_2}(x_1) + \gamma V_3^*) = \tau_2,$$

which concludes the proof. ■

Comparison to ADP. When a simulator of the true MDP is available (a case studied in the so-called simulation optimization literature), one can imagine to be able to compute a policy that is greedy in the true MDP with respect to some fixed value function up to an arbitrary accuracy at any given state. Singh and Yee (1994, Theorem 1), de Farias and Van Roy (2003, Theorem 4.1), Bertsekas (2012, Proposition 3.1) and Grünewälder et al. (2011) bound the suboptimality of the resulting policy.

A potentially more useful result is to bound the suboptimality of a policy derived from an action value-function (derived from a model). Although we were unable to locate such a result in the literature, it can be derived using the techniques in the above-mentioned works. These two results are summarized as follows:

Theorem 20 (ADP policy error bounds) *For any $\widetilde{V} \in \mathcal{V}^{\mathcal{A}}$, if $\tilde{\pi} \doteq G\widetilde{V}$, then*

$$\left\|V^* - V^{\tilde{\pi}}\right\|_\infty \le \frac{2(1+\gamma)}{1-\gamma} \left\|\widetilde{V} - T_{\mathcal{P}}V^*\right\|_\infty. \tag{10}$$

Alternatively, for any $\widetilde{V} \in \mathcal{V}$, if $\tilde{\pi} \doteq GT_{\mathcal{P}}\widetilde{V}$, then

$$\left\|V^* - V^{\tilde{\pi}}\right\|_\infty \le \frac{2\gamma}{1-\gamma} \left\|\widetilde{V} - V^*\right\|_\infty. \tag{11}$$

To finish the discussion of the relevance of bounding the deviation $\|U^* - V^*\|_\infty$ in a model-based setting, from (11) (by choosing $\widetilde{V}' = U^*$) we see that controlling this deviation would suffice if the policy was derived using the true model. When this is not an option, one needs to fall back to (10), calling for bounding the difference between the action-value fixed point of a model and the action-value fixed point of the true model. To that end, we could use Theorem 8, but the resulting bound would scale with $\frac{\gamma}{(1-\gamma)^2}$, while both Theorem 7 and our later results scale with $\frac{\gamma}{1-\gamma}$ only. Therefore, it is better to use Theorem 8 directly to bound the policy error.

Appendix E. Additional remarks about $\mathcal{R}$

In this section, we carry out a brief discussion about the case when $\mathcal{R}$ is a point evaluator (in which case $\mathrm{Lip}(\mathcal{R}) = \mathrm{Lip}(\mathcal{R}^{\mathcal{A}})$).

In supremum norm, we were able to use that $\mathrm{Lip}(\mathcal{R}) \leq 1$ to get from Proposition 21 that if $M'T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}}$ is a contraction and $\mathrm{Lip}(\mathcal{Q}) < \infty$, then some power of $MT_{\mathcal{Q}\mathcal{R}}$ is a contraction. In weighted supremum norm, $\mathrm{Lip}(\mathcal{R}) = \max_i \frac{\eta_i}{\nu(x_i)}$, and, if this quantity is finite, some power of $MT_{\mathcal{Q}\mathcal{R}}$ is a contraction as well.

Proposition 21 *If* $\mathrm{Lip}(M)\,\mathrm{Lip}(\mathcal{Q})\,\mathrm{Lip}(\mathcal{R}) < \infty$*,* $\mathrm{Lip}(M') \leq 1$*, and* $\mathrm{Lip}(\mathcal{R}^{\mathcal{A}}\mathcal{Q}) \leq 1$*,* $(MT_{\mathcal{Q}\mathcal{R}})^m$ *is a contraction for all* m *large enough.*

Proof We have that $\mathrm{Lip}(MT_{\mathcal{Q}}) = \mathrm{Lip}(M)\gamma\,\mathrm{Lip}(\mathcal{Q}) \doteq B' < \infty$, and $\mathrm{Lip}(M'T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}}) \leq \mathrm{Lip}(M')\gamma \leq \gamma$. For $m \geq 0$, $(MT_{\mathcal{Q}\mathcal{R}})^{m+1} = MT_{\mathcal{Q}}(M'T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}})^m\mathcal{R}$, so $\mathrm{Lip}((MT_{\mathcal{Q}\mathcal{R}})^{m+1}) \leq B'\gamma^{m+1}\,\mathrm{Lip}(\mathcal{R})$. Given m s.t. $B'\gamma^m < 1$, thus $(MT_{\mathcal{Q}\mathcal{R}})^{m'}$ is a contraction for all $m' \geq m$. ∎

In the case of $L^p(\mu)$ norms, Proposition 22 gives us the form for $\mathrm{Lip}(\mathcal{R})$. Having noted that $\mathcal{I}$ indexes a measurable subset of $\mathcal{X}$ (since $\mathcal{R}$ is a point evaluator), we extend ρ to $\mathcal{X}$ by $\rho(X) \doteq \rho(i \in \mathcal{I} : x_i \in X)$ for measurable $X \subseteq \mathcal{X}$. We denote absolute continuity of (the extension of) ρ w.r.t. to μ by $\rho \ll \mu$.

Proposition 22 *Assume that* $\mathcal{R}$ *is a point evaluator, and that the norm overs* $\mathcal{V}$ *and* $\mathcal{W}$ *are respectively an* $L^p(\mu)$ *and an* $L^p(\rho)$ *norm. If* $\rho \ll \mu$*, then* $\mathrm{Lip}(\mathcal{R}) = \left\|\frac{d\rho}{d\mu}\right\|_\infty^{\frac{1}{p}}$*, otherwise* $\mathrm{Lip}(\mathcal{R}) = \infty$*.*

Proof Thanks to the linearity of $\mathcal{R}$, we have

$$\mathrm{Lip}(\mathcal{R}) = \sup_{V \neq 0} \frac{\|\mathcal{R}V\|_{\rho,p}}{\|V\|}.$$

From absolute continuity we get that $\int |V(x)|^p\, d\rho(x) = \int |V(x)|^p \left(\frac{d\rho(x)}{d\mu(x)}\right) d\mu(x)$, and from Hölder's inequality we get

$$\int |V(x)|^p \left(\frac{d\rho(x)}{d\mu(x)}\right) d\mu(x) \leq \left\|\frac{d\rho}{d\mu}\right\|_\infty \cdot \int |V(x)|^p\, d\mu(x),$$

which implies that $\mathrm{Lip}(\mathcal{R}) \leq \left\|\frac{d\rho}{d\mu}\right\|_\infty^{\frac{1}{p}}$.

To show that the upper-bound above is tight, we can see that

$$\begin{aligned}\mathrm{Lip}(\mathcal{R}) &= \sup_{V \neq 0} \frac{\|\mathcal{R}V\|_{\rho,p}}{\|V\|} \\ &\geq \sup_{\substack{X \subseteq \mathcal{X} \\ \mu(X) > 0}} \frac{\int_X d\rho(x)}{\int_X d\mu(x)} = \sup_{\substack{X \subseteq \mathcal{X} \\ \mu(X) > 0}} \frac{\rho(X)}{\mu(X)},\end{aligned}$$

because we can restrict V to the indicator function of an $X \subseteq \mathcal{X}$. If ρ is not absolutely continuous w.r.t. to μ, then there exists X s.t. $\mu(X) = 0$ and $\rho(X) > 0$, which implies that $\mathrm{Lip}(\mathcal{R}) = \infty$.

Otherwise, $\mu(X) = 0 \Rightarrow \rho(X) = 0$ for all $X \subseteq \mathcal{X}$, and

$$\sup_{\substack{X\subseteq\mathcal{X}\\ \mu(X)>0}} \frac{\rho(X)}{\mu(X)} = \sup_{X\subseteq\mathcal{X}} \frac{\rho(X)}{\mu(X)} = \left\| \frac{d\rho}{d\mu} \right\|_\infty ,$$

which concludes the proof. ■

Interestingly, an unbounded $\mathrm{Lip}(\mathcal{R})$ can lead to $\mathrm{Lip}(MT_{\mathcal{QR}}) = \infty$, as stated by Proposition 23, and, yet, if Assumption 3 $MT_{\mathcal{QR}}$ still has a fixed point, and, provided that Assumption 5 is met in addition, we can still obtain performance bounds for the policy error of $\hat{\pi}$.

Proposition 23 *Assume $\mathcal{R}$ is a point evaluator, that the norms over $\mathcal{V}$ and $\mathcal{W}$ are, respectively, an $L^p(\mu)$ and an $L^p(\rho)$ norm, and that the norms over $\mathcal{V}^\mathcal{A}$ and $\mathcal{W}^\mathcal{A}$ are the corresponding mixed norms defined using $M_{|\cdot|}$. If $\mu(\{x_i : i \in \mathcal{I}\}) = 0$, then for all $m \geq 0$, the following holds: If* $\mathrm{Lip}(MT_\mathcal{Q}(M'T_{\mathcal{R}^\mathcal{A}\mathcal{Q}})^m) > 0$ *then* $\mathrm{Lip}((MT_{\mathcal{QR}})^{m+1}) = \infty$.

Proof Let $S \doteq \{x_i : i \in \mathcal{I}\}$. When $\mu(S) = 0$, we have $\rho \not\ll \mu$ and $\mathrm{Lip}(\mathcal{R}) = \infty$. We will show that for any $m \geq 0$ either $\mathrm{Lip}((MT_{\mathcal{QR}})^{m+1}) = 0$ or $\mathrm{Lip}((MT_{\mathcal{QR}})^{m+1}) = \infty$. To that end, define $(Zu)(x_i) \doteq u_i$ for $i \in \mathcal{I}$ (for simplicity, assume that, for all $x_i, x_j \in S$, $x_i = x_j \Rightarrow i = j$), and let $(Zu)(x) \doteq 0$ for $x \notin S$. Then $\sup_{u\in\mathcal{W}} \|Zu\| = 0$ and $\mathcal{R}Zu = u$ for all $u \in \mathcal{W}$. The definitions then give:

$$\begin{aligned}
\mathrm{Lip}((MT_{\mathcal{QR}})^{m+1}) &= \mathrm{Lip}((MT_{\mathcal{QR}})^m T_\mathcal{Q}\mathcal{R}) \\
&= \sup_{\substack{V,V'\in\mathcal{V}:\\ V\neq V'}} \frac{\|(MT_{\mathcal{QR}})^m MT_\mathcal{Q}\mathcal{R}V - (MT_{\mathcal{QR}})^m MT_\mathcal{Q}\mathcal{R}V'\|}{\|V - V'\|} \\
&\geq \sup_{\substack{u,u'\in\mathcal{W}:\\ u\neq u'}} \frac{\|(MT_{\mathcal{QR}})^m MT_\mathcal{Q}u - (MT_{\mathcal{QR}})^m MT_\mathcal{Q}u'\|}{\|Zu - Zu'\|},
\end{aligned}$$

which is unbounded unless $\mathrm{Lip}((MT_{\mathcal{QR}})^m MT_\mathcal{Q}) = 0$. To conclude, we observe that $(MT_{\mathcal{QR}})^m MT_\mathcal{Q} = MT_\mathcal{Q}(M'T_{\mathcal{R}^\mathcal{A}\mathcal{Q}})^m$, since $\mathcal{R}$ is a point evaluator. ■