Weighted-norm bounds on model approximation in MDPs with unbounded per-step cost

Berk Bozkurt, Aditya Mahajan, Ashutosh Nayyar, and Yi Ouyang

***Abstract*— We consider the problem of designing a control policy for an infinite-horizon discounted cost Markov Decision Process (MDP) $\mathcal{M}$ when we only have access to an approximate model $\hat{\mathcal{M}}$. If we design an optimal policy $\hat{\pi}^\star$ for the approximate model, how well does it perform when used in the true model $\mathcal{M}$? We provide an answer to this question by bounding a weighted norm of the difference between the value function of $\hat{\pi}^\star$ when used in $\mathcal{M}$ and the optimal value function of $\mathcal{M}$. The use of weighted norm allows us to obtain meaningful bounds for performance loss even when the per-step cost function is unbounded. This is in contrast to much of the prior literature which has largely focused only on the case of bounded per-step cost. We illustrate our results for two specific instances — a finite MDP model for an inventory control problem and the discounted linear quadratic regulator problem.**

I. Introduction

We consider the problem of model approximation in Markov decision processes (MDPs), i.e., the problem of designing an optimal controller for an MDP using an approximate model (e.g. designing gait controller of a robot using a simulation model). Let $\mathcal{M}$ denote the true model of the system and let $\hat{\mathcal{M}}$ denote an approximate model. Suppose we solve the approximate model $\hat{\mathcal{M}}$ to identify a policy $\hat{\pi}^\star$ which is optimal for $\hat{\mathcal{M}}$. How well does $\hat{\pi}^\star$ perform in the original model $\mathcal{M}$?

Several variations of this question have been studied in the MDP literature. Perhaps the earliest work investigating this is that of Fox [1], who investigated approximating MDPs by a finite state approximation. In a series of papers, Whitt generalized these results to approximating general MDPs via state aggregation [2]–[4]. Similar results for state discretization were obtained in [5], [6], state and action discretization in [7] and for models with state dependent discounting in [8]. A general framework to view model approximation using the lens of integral probability metrics was presented by Müller [9]. There has been considerable recent advances on these ideas in recent years [10], [11], including generalizations to partially observed models [12], [13].

A related question is that of continuity of optimal policy in model approximation. In particular, if $\{\hat{\mathcal{M}}_n\}_{n\geq 1}$ is a sequence of models that converge to $\mathcal{M}$ in some sense, do the corresponding optimal policies $\{\hat{\pi}^\star_n\}_{n\geq 1}$, where $\hat{\pi}^\star_n$ is optimal for $\hat{\mathcal{M}}_n$, converge to an optimal policy for $\mathcal{M}$? Perhaps the earliest work in this direction is that of Fox [14], who studied the continuity of state discretization procedures. Sufficient conditions for continuity of value function on model parameters were presented in [15]. There are series of recent papers which significantly generalize these results, include characterizing conditions under which the optimal policy is continuous in model parameters [10], [16]–[21].

The question of model approximation is also relevant for learning optimal policies when the system model is unknown. Therefore, several notions related to model approximation have been studied in the reinforcement learning literature including approximate homeomorphisms [22], [23], bisimulation metrics [24]–[26], state abstraction [27], and approximate latent state models [28], [29].

The basic results of model approximation may be characterized as follows. Let $\mathcal{M}$ and $\hat{\mathcal{M}}$ be two MDP models with the same state space $\mathcal{S}$ and action space $\mathcal{A}$. Let $\hat{\pi}^\star \colon \mathcal{S} \to \mathcal{A}$ be an optimal policy for model $\hat{\mathcal{M}}$. Let $V^{\hat{\pi}^\star} \colon \mathcal{S} \to \mathbb{R}$ denote the performance of policy $\hat{\pi}^\star$ in model $\mathcal{M}$ and let $V^\star \colon \mathcal{S} \to \mathbb{R}$ denote the optimal value of model $\mathcal{M}$. Most of the existing literature on model approximation provides bounds on $\|V^{\hat{\pi}^\star} - V^\star\|_\infty \coloneqq \sup_{s\in\mathcal{S}} \left|V^{\hat{\pi}^\star}(s) - V^\star(s)\right|$ in terms of the parameters of the models $\mathcal{M}$ and $\hat{\mathcal{M}}$.

However, such bounds are not appropriate for models with non-compact state spaces and unbounded per-step cost. To illustrate this limitation, consider the linear quadratic regulation (LQR) problem in which the objective is to minimize the infinite-horizon expected discounted total cost. Let $\mathcal{M}$ and $\hat{\mathcal{M}}$ be two such LQR models and $\hat{\pi}^\star$ be the optimal policy of $\hat{\mathcal{M}}$. It is well known that

$$V^\star(s) = s^\intercal P s + q \quad \text{and} \quad V^{\hat{\pi}^\star}(s) = s^\intercal P^{\hat{\pi}^\star} s + q^{\hat{\pi}^\star},$$

where $s \in \mathbb{R}^{n_s}$ is the state, P is the solution of an appropriate Riccati equation, $P^{\hat{\pi}^\star}$ is a solution of an appropriate Lyapunov equation (which depends on the gain of policy $\hat{\pi}^\star$) and q and $q^{\hat{\pi}^\star}$ are constants (where $q^{\hat{\pi}^\star}$ depends on the gain of policy $\hat{\pi}^\star$). See Sec. IV-C for exact details.

Note that for this model and, in general for models with unbounded per-step cost, $\|V^\star - V^{\hat{\pi}^\star}\|_\infty = \infty$. Therefore, the approximation bounds on $\|V^\star - V^{\hat{\pi}^\star}\|_\infty$ provided by the existing literature will also evaluate to ∞ and, as a

Berk Bozkurt and Aditya Mahajan are with the Department of Electrical and Computer Engineering, McGill University, Montreal, QC, Canada. (email: berk.bozkurt@mail.mcgill.ca, aditya.mahajan@mcgill.ca)

Ashutosh Nayyar is with the Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA, USA. (email: ashutoshn@usc.edu)

Yi Ouyang is with Preferred Networks America, Burlingame, CA, USA (email: ouyangyi@preferred-america.com)

The work of BB and AM was supported in part by the Innovation for Defence Excellence and Security (IDEaS) Program of the Canadian Department of National Defence through grant CFPMN2-30 and the work of BB was additionally supported by IVADO MSc Excellence Fellowship. The research at USC was supported by NSF grants ECCS 2025732 and ECCS 1750041.

result, do not provide any insights into the quality of the approximation.

Our main contribution in this paper is to provide an alternative characterization of the modeling error in terms of the weighted norm:

$$\|V^{\hat{\pi}^\star}(s) - V^\star(s)\|_w := \sup_{s\in\mathcal{S}} \frac{\left|V^{\hat{\pi}^\star}(s) - V^\star(s)\right|}{w(s)},$$

where $w\colon \mathcal{S} \to [1,\infty)$ is a weight function which satisfies some technical conditions. Our bounds are derived using what we call the *Bellman mismatch functional*. In the special case when $w(s) \equiv 1$, our bounds recover the existing sup-norm bounds. For general weight functions, we present some illustrative examples to compare our weighted-norm approximation bounds with existing sup-norm approximation bounds. Finally, we revisit the LQR example illustrated above and show that the weighted-norm approximation bounds provide meaningful approximation guarantees for such unbounded-cost models.

Perhaps the closest result to ours in the literature is [17] which considers finite approximation of MDP models in general state spaces. For unbounded per-step cost functions, they establish sufficient conditions under which $\hat{V}_n^\star \to V^\star$ and $V^{\hat{\pi}_n^\star} \to V^\star$, where $\hat{V}_n^\star$ and $\hat{\pi}_n^\star$ are value function and optimal policy of a discretized model with grid cells of size less than $1/n$. However, they do not establish the approximation error when a specific approximate model is used. The results of [17] are also derived using weighted norms, but there are subtle differences in the way we use weighted norms. See Remark 1 and Sec. IV-D for details.

II. Preliminaries

A. Markov decision processes

A discrete-time infinite-horizon discounted cost Markov decision process (MDP) is a tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, c, \gamma\rangle$ where

- $\mathcal{S}$ is the state space, which is assumed to be a Borel space. The state at time t is denoted by $S_t \in \mathcal{S}$.
- $\mathcal{A}$ is the action space, which is assumed to be a Borel space. The action at time t is denoted by $A_t \in \mathcal{A}$.
- $P\colon \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$ is a controlled stochastic kernel, which specifies the system dynamics. In particular, for any time t and any $s_{1:t} \in \mathcal{S}^t$, $a_{1:t} \in \mathcal{A}^t$ and any Borel set $B \subset \mathcal{S}$, we have

$$\begin{aligned}\mathbb{P}(S_{t+1} \in B \mid S_{1:t} = s_{1:t}, A_{1:t} = a_{1:t}) \\ = \mathbb{P}(S_{t+1} \in B \mid S_t = s_t, A_t = a_t) \\ =: P(B \mid s_t, a_t).\end{aligned}$$

- $c\colon \mathcal{S}\times\mathcal{A} \to \mathbb{R}_{\geq 0}$ is the per-step cost function which is assumed to be measurable.
- $\gamma \in (0,1)$ is the discount factor.

A mapping $\pi\colon \mathcal{S} \to \Delta(\mathcal{A})$ is called a (time-homogeneous) policy. Let Π denote the space of all time-homogeneous (and possibly randomized) policies. The performance of any policy $\pi \in \Pi$ starting from an initial state $s \in \mathcal{S}$ is given by

$$V^\pi(s) = \mathbb{E}^\pi\left[\sum_{t=1}^{\infty} \gamma^{t-1} c(S_t, A_t) \;\middle|\; S_1 = s\right] \tag{1}$$

where $\mathbb{E}^\pi$ denotes the expectation with respect to the probability measure on all system variables induced by the choice of policy π. The function V^π is called the *value function* of policy π.

A policy $\pi^\star \in \Pi$ is called an *optimal policy* if

$$V^{\pi^\star}(s) \leq V^\pi(s), \quad \forall s \in \mathcal{S}, \forall \pi \in \Pi. \tag{2}$$

Note that since we consider general Borel state and action spaces with possibly unbounded per-step cost function, an optimal policy is not guaranteed to exist or be unique. If an optimal policy exists, its value function is called the optimal value function. We focus on the case when such value functions exist and can be obtained via dynamic programming. We formally define this as dynamic programming solvability in the next section.

B. Dynamic programming solvability

Let $\mathcal{V}$ denote the space of measurable functions from $\mathcal{S} \to \mathbb{R}_{\geq 0} \cup \{+\infty\}$.

Definition 1 (Bellman operators) Define the following two operators:

- For any $\pi \in \Pi$, define the *Bellman operator* $\mathcal{B}^\pi\colon \mathcal{V} \to \mathcal{V}$ as follows: for any $v \in \mathcal{V}$,

$$[\mathcal{B}^\pi v](s) = \int_{\mathcal{A}} \pi(da \mid s)\Big[c(s,a) + \gamma \int_{\mathcal{S}} v(s') P(ds' \mid s, a)\Big].$$

- Define the *Bellman optimality operator* $\mathcal{B}^\star\colon \mathcal{V} \to \mathcal{V}$ as follows: for any $v \in \mathcal{V}$,

$$[\mathcal{B}^\star v](s) = \inf_{a\in\mathcal{A}} \left[c(s,a) + \gamma \int_{\mathcal{S}} v(s') P(ds' \mid s, a)\right].$$

Definition 2 (Dynamic programming solvability) An MDP $\mathcal{M}$ is said to be *dynamic programming solvable* (DP-solvable, for short) if it satisfies the following properties:

1) There exists a unique fixed point $V^\star \in \mathcal{V}$ of the dynamic programming equation
$$V = \mathcal{B}^\star V.$$
2) There exists an optimal policy $\pi^\star \in \Pi$ such that
$$V^{\pi^\star} = V^\star \quad \text{and} \quad \mathcal{B}^{\pi^\star} V^\star = \mathcal{B}^\star V^\star.$$

Models with finite state and action spaces are always DP-solvable. For models with general state and action spaces, there are several conditions in the literature which imply DP-solvability. See [30] for an overview.

C. Weighted-norm stability

Definition 3 (Weighted norm) Given a weight function $w\colon \mathcal{S} \to [1,\infty)$, we define the weighted norm $\|\cdot\|_w$ on $\mathcal{V}$ as follows: for any $v \in \mathcal{V}$,

$$\|v\|_w = \sup_{s\in\mathcal{S}} \frac{|v(s)|}{w(s)}.$$

Note that when the weight function $w(s) \equiv 1$, then the weighted norm $\|v\|_w$ is equivalent to the sup-norm $\|v\|_\infty := \sup_{s\in\mathcal{S}} |v(s)|$.

Definition 4 ((κ, w)-stability) Given an MDP $\mathcal{M}$ and a tuple (κ, w), where κ is a positive constant with $\gamma\kappa < 1$ and w is a function from $\mathcal{S}$ to $[1,\infty)$, we say a policy $\pi \in \Pi$ is (κ, w)-stable if

$$\|c_\pi\|_w < \infty \tag{3}$$

where $c_\pi(s) = \int_{\mathcal{A}} c(s,a)\pi(da|s)$ and

$$\int_{\mathcal{A}} \pi(da \mid s) \int_{\mathcal{S}} w(s')P(ds' \mid s,a) \le \kappa w(s), \quad \forall s \in \mathcal{S}. \tag{4}$$

Let $\Pi_S(\kappa, w)$ denote the set of all (κ, w)-stable policies for model $\mathcal{M}$. Note that depending on the choice of (κ, w), the set $\Pi_S(\kappa, w)$ might be empty.

Remark 1 The definition of (κ, w)-stability in Definition 4 is similar to but weaker than the notion of stability typically used in the literature (e.g. [17], [30], [31]). For example, in Assumption 8.3.2 of [31], it is assumed that there exists a tuple $(\bar\kappa, \bar w)$ where $\bar\kappa$ is a positive constant with $\bar\kappa\gamma < 1$ and $\bar w$ is a function from $\mathcal{S}$ to $[1,\infty)$ such that $\|c(\cdot,a)\|_{\bar w} < \infty$ for all actions and

$$\int_{\mathcal{S}} \bar w(s')P(ds' \mid s,a) \le \bar\kappa \bar w(s), \quad \forall s \in \mathcal{S}, \forall a \in \mathcal{A}. \tag{5}$$

It is shown in [31] that this assumption is sufficient for DP-solvability. Note that the notion of (κ, w)-stability is weaker. We only require the two inequalities to hold for a given policy rather than for all actions. As we show via an example in Sec. IV-D, using the weaker notion of (κ, w)-stability drastically increases the possible choices of the weight function and lead to tighter upper bounds on performance.

Lemma 1 *Given an MDP $\mathcal{M}$ and the tuple (κ, w), define*

$$\bar{\mathcal{V}}_w = \{v \in \mathcal{V} : \|v\|_w < \infty\}.$$

Then, for any policy $\pi \in \Pi_S(\kappa, w)$, we have the following:

1) *If $v \in \bar{\mathcal{V}}_w$, then $\mathcal{B}^\pi v \in \bar{\mathcal{V}}_w$.*
2) *$\mathcal{B}^\pi$ is a $\|\cdot\|_w$-norm contraction with contraction factor $\gamma\kappa$, i.e., for any $v_1, v_2 \in \bar{\mathcal{V}}_w$, we have*

$$\|\mathcal{B}^\pi v_1 - \mathcal{B}^\pi v_2\|_w \le \gamma\kappa \|v_1 - v_2\|_w.$$

3) *The fixed point equation*

$$V = \mathcal{B}^\pi V$$

has a unique solution in $\bar{\mathcal{V}}_w$ and that solution is equal to V^π.

TABLE I: Notation for the variables used for the two models

Variable	Model $\mathcal{M}$	Model $\hat{\mathcal{M}}$
Dynamics	P	$\hat P$
per-step cost	c	$\hat c$
Value function of policy π	V^π	$\hat V^\pi$
Optimal value function	$V^\star$	$\hat V^\star$
Bellman operator of policy π	$\mathcal{B}^\pi$	$\hat{\mathcal{B}}^\pi$
Bellman optimality operator	$\mathcal{B}^\star$	$\hat{\mathcal{B}}^\star$
Set of (κ, w)-stable policies	$\Pi_S(\kappa, w)$	$\hat\Pi_S(\kappa, w)$

III. Problem formulation and main results

A. Model approximation in MDPs

We are interested in the problem of model approximation in MDPs. In particular, suppose there is an MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, c, \gamma\rangle$ of interest, but the system designer has access to only an approximate model $\hat{\mathcal{M}} = \langle \mathcal{S}, \mathcal{A}, \hat P, \hat c, \gamma\rangle$. Note that both models $\mathcal{M}$ and $\hat{\mathcal{M}}$ have the same state and action spaces, but have different transition dynamics and per-step cost.

We assume that both models $\mathcal{M}$ and $\hat{\mathcal{M}}$ are well behaved in the following sense.

Assumption 1 Models $\mathcal{M}$ and $\hat{\mathcal{M}}$ are DP-solvable.

We will use the superscript hat to denote variables/operators corresponding to the approximate model, as summarized in Table I. We are interested in the following approximation problem.

Problem 1 *Let $\hat\pi^\star$ be an optimal policy for the approximate model $\hat{\mathcal{M}}$. Given a start state s, bound the loss in performance when using $\hat\pi^\star$ in the original model $\mathcal{M}$ (compared to the optimal performance in the original model), i.e., bound $V^{\hat\pi^\star}(s) - V^\star(s)$.*

B. Main results

We impose the following additional assumption on the models.

Assumption 2 There exists a tuple (κ, w), where κ is a positive constant such that $\gamma\kappa < 1$ and $w\colon \mathcal{S} \to [1,\infty)$ such that

- there exists an optimal policy $\pi^\star$ of the original model $\mathcal{M}$ such that $\pi^\star \in \Pi_S(\kappa, w)$,
- there exists an optimal policy $\hat\pi^\star$ of the approximate model $\hat{\mathcal{M}}$ such that $\hat\pi^\star \in \Pi_S(\kappa, w) \cap \hat\Pi_S(\kappa, w)$,

where $\Pi_S(\kappa, w)$ and $\hat\Pi_S(\kappa, w)$ denote the set of (κ, w)-stable policies for models $\mathcal{M}$ and $\hat{\mathcal{M}}$, respectively.

For most problems, the optimal policy of a model is (κ, w)-stable for an appropriate choice of (κ, w). Therefore, if $\hat{\mathcal{M}}$ is a reasonably good approximation of $\mathcal{M}$, we expect $\hat\pi^\star$ to be close to the optimal policy of $\mathcal{M}$ and be (κ, w)-stable in $\mathcal{M}$. So, Assumption 2 holds whenever $\hat{\mathcal{M}}$ is close to $\mathcal{M}$.

Definition 5 (Bellman mismatch functionals) Given a weight function $w\colon \mathcal{S} \to [1,\infty)$, define the following two functionals:

- For any $\pi \in \Pi$, define the *Bellman mismatch functional* $\mathcal{D}_w^\pi \colon \bar{\mathcal{V}}_w \to \mathbb{R}_{\geq 0}$ as follows: for any $v \in \bar{\mathcal{V}}_w$,

$$\mathcal{D}_w^\pi v = \|\mathcal{B}^\pi v - \hat{\mathcal{B}}^\pi v\|_w.$$

- Define the *Bellman optimality mismatch functional* $\mathcal{D}_w^\star \colon \bar{\mathcal{V}}_w \to \mathbb{R}_{\geq 0}$ as follows: for any $v \in \bar{\mathcal{V}}_w$,

$$\mathcal{D}_w^\star v = \|\mathcal{B}^\star v - \hat{\mathcal{B}}^\star v\|_w.$$

Theorem 1 *Under Assumptions 1 and 2, we have the following two bounds on $V^{\hat{\pi}^\star} - V^\star$:*

1) *Bound in terms of properties of $\hat{V}^\star$:*

$$\left\|V^{\hat{\pi}^\star} - V^\star\right\|_w \leq \frac{1}{1-\gamma\kappa}\left[\mathcal{D}_w^{\hat{\pi}^\star}\hat{V}^\star + \mathcal{D}_w^\star \hat{V}^\star\right].$$

2) *Bound in terms of properties of $V^\star$:*

$$\left\|V^{\hat{\pi}^\star} - V^\star\right\|_w \leq \frac{1}{1-\gamma\kappa}\mathcal{D}_w^{\hat{\pi}^\star}V^\star + \frac{(1+\gamma\kappa)}{(1-\gamma\kappa)^2}\mathcal{D}_w^\star V^\star.$$

Remark 2 The bounds on $\left\|V^{\hat{\pi}^\star} - V^\star\right\|_w$ stated in Theorem 1 can be used to bound the performance loss when using $\hat{\pi}^\star$ in the original model $\mathcal{M}$ with a start state s. This is because, by the definition of $\|\cdot\|_w$,

$$V^{\hat{\pi}^\star}(s) - V^\star(s) \leq \left\|V^{\hat{\pi}^\star} - V^\star\right\|_w w(s), \tag{6}$$

where we have used the fact that $V^{\hat{\pi}^\star}(s) - V^\star(s)$ is non-negative.

Remark 3 The weight function w is assumed to be greater than or equal to one. This assumption is not necessary as long as $\inf_{s\in\mathcal{S}} w(s) > 0$. The definition of (κ, w)-stability and the bounds on performance loss obtained using (6) and Theorem 1 are invariant under positive scaling of the weight function.

Remark 4 Suppose there is a family $\mathcal{W}$ of weight functions, such that for every $w \in \mathcal{W}$, there exists a $\kappa_w < 1/\gamma$ such that (κ_w, w) satisfies Assumption 2. Then, we can strengthen the result of (6) as follows:

$$V^{\hat{\pi}^\star}(s) - V^\star(s) \leq \inf_{w\in\mathcal{W}}\left\{\left\|V^{\hat{\pi}^\star} - V^\star\right\|_w w(s)\right\}. \tag{7}$$

Thus, the choice of weight function that gives the tightest bound can depend on the start state s. We illustrate the benefit of such a state-dependent choice of weight function in Sec. IV-B.

C. A simpler upper bound

In this section, we present a simpler upper bound on the result of Theorem 1. For that purpose, we define the *Bellman maximum mismatch functional* $\mathcal{D}_w^{\max} \colon \bar{\mathcal{V}}_w \to \mathbb{R}_{\geq 0}$ as follows: for any $v \in \bar{\mathcal{V}}_w$,

$$\mathcal{D}_w^{\max} v = \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\left|\Xi^{(s,a)}v\right|}{w(s)}$$

where

$$\Xi^{(s,a)}v = c(s,a) - \hat{c}(s,a) + \gamma\int_{\mathcal{S}} v(s')[P(ds'|s,a) - \hat{P}(ds'|s,a)].$$

Proposition 1 *The following properties hold:*

$$\sup_{\pi\in\Pi}\mathcal{D}_w^\pi v = \mathcal{D}_w^{\max} v \quad \textit{and} \quad \mathcal{D}_w^\star v \leq \mathcal{D}_w^{\max} v. \tag{8}$$

Therefore, under Assumptions 1 and 2, we have:

1) *Bound in terms of properties of $\hat{V}^\star$:*

$$\left\|V^{\hat{\pi}^\star} - V^\star\right\|_w \leq \frac{2}{1-\gamma\kappa}\mathcal{D}_w^{\max}(\hat{V}^*).$$

2) *Bound in terms of properties of $V^\star$:*

$$\left\|V^{\hat{\pi}^\star} - V^\star\right\|_w \leq \frac{2}{(1-\gamma\kappa)^2}\mathcal{D}_w^{\max}(V^*).$$

D. The special case of sup-norm

For models where the per-step cost is bounded, the results of Theorem 1 also provide a bound on $\|V^{\hat{\pi}^\star} - V^\star\|_\infty$ by taking the weight function $w(s) \equiv 1$. As mentioned earlier, when $w(s) \equiv 1$, $\|v\|_w = \|v\|_\infty$. Also observe that when $w(s) \equiv 1$, Eq. (4) is always satisfied with $\kappa = 1$. Thus, for a model with bounded per-step cost, any policy π is (κ, w)-stable with $\kappa = 1$ and $w(s) \equiv 1$. Therefore, if we take two models $\mathcal{M}$ and $\hat{\mathcal{M}}$ with bounded per-step cost, then Assumption 2 is always satisfied.

We use the notation $\mathcal{D}_\infty^\pi$, $\mathcal{D}_\infty^\star$, $\mathcal{D}_\infty^{\max}$ to denote the Bellman mismatch operators when the weight function $w(s) \equiv 1$. Then, an immediate consequence of Theorem 1 and Proposition 1 is the following:

Corollary 1 *Under Assumptions 1, if the per-step cost is bounded, we have the following two bounds on $V^{\hat{\pi}^\star} - V^\star$:*

1) *Bound in terms of properties of $\hat{V}^\star$:*

$$\begin{aligned}\left\|V^{\hat{\pi}^\star} - V^\star\right\|_\infty &\leq \frac{1}{1-\gamma}\left[\mathcal{D}_\infty^{\hat{\pi}^\star}\hat{V}^\star + \mathcal{D}_\infty^\star\hat{V}^\star\right] \\ &\leq \frac{2}{1-\gamma}\mathcal{D}_\infty^{\max}\hat{V}^\star.\end{aligned}$$

2) *Bound in terms of properties of $V^\star$:*

$$\begin{aligned}\left\|V^{\hat{\pi}^\star} - V^\star\right\|_\infty &\leq \frac{1}{1-\gamma}\mathcal{D}_\infty^{\hat{\pi}^\star}V^\star + \frac{(1+\gamma)}{(1-\gamma)^2}\mathcal{D}_\infty^\star V^\star \\ &\leq \frac{2}{(1-\gamma)^2}\mathcal{D}_\infty^{\max}V^\star.\end{aligned}$$

The result of Corollary 1 can be further simplified as follows.

Proposition 2 *Under Assumption 1, if the per-step cost is bounded, we have the following bounds on $V^{\hat{\pi}^\star} - V^\star$:*

- *Bound in terms of total-variation distance:*

$$\left\|V^{\hat{\pi}^\star} - V^\star\right\|_\infty \leq \frac{2}{1-\gamma}\left[\|c - \hat{c}\|_\infty + \gamma\sup_{\substack{s\in\mathcal{S}\\a\in\mathcal{A}}}\mathrm{TV}(P(\cdot|s,a), \hat{P}(\cdot|s,a))\left\|\hat{V}^\star\right\|_\infty\right]$$

where $\mathrm{TV}(\mu,\nu)$ denotes the total variation distance between two measures.

- *Bound in terms of Wasserstein distance:*

$$\left\|V^{\hat{\pi}^\star} - V^\star\right\|_\infty \le \frac{2}{1-\gamma}\Big[\|c-\hat{c}\|_\infty + \gamma \sup_{\substack{s\in\mathcal{S}\\ a\in\mathcal{A}}} \mathrm{Was}(P(\cdot|s,a), \hat{P}(\cdot|s,a))\,\mathrm{Lip}(\hat{V}^\star)\Big]$$

where $\mathrm{Was}(\mu,\nu)$ *denotes the Wasserstein distance between two measures and* $\mathrm{Lip}(\cdot)$ *denote the Lipschitz constant of a function.*

PROOF (OUTLINE) First note that, from triangle inequality, we have

$$\mathcal{D}^\star_\infty v \le \|c-\hat{c}\|_\infty + \gamma \sup_{\substack{s\in\mathcal{S}\\ a\in\mathcal{A}}} \left| \int_{\mathcal{S}} v(s')[P(ds'|s,a) - \hat{P}(ds'|s,a)] \right|$$

and a similar bound holds for $\mathcal{D}^{\hat{\pi}^\star}_\infty v$. The result then follows by observing that for any two measures μ and ν and any function v, we have $\left|\int v d\mu - \int v d\nu\right| \le \|v\|_\infty \mathrm{TV}(\mu,\nu)$ and also $\left|\int v d\mu - \int v d\nu\right| \le \mathrm{Lip}(v)\mathrm{Was}(\mu,\nu)$. See [13] for general discussion of such bounds. ■

The results of Prop 2 are similar to the approximation results presented in [9], [27], [29] (some of those results assumed that the approximate model has a smaller state space than the original model).

IV. SOME INSTANCES OF THE MAIN RESULTS

A. Inventory management

In this section, we illustrate the results of Theorem 1 for an inventory management problem with state space $\mathcal{S} = \{-S_{\max}, -S_{\max}+1, \dots S_{\max}\}$ and action space $\mathcal{A} = \{0, 1, \dots, S_{\max}\}$. Let $S_t \in \mathcal{S}$ denote the amount of stock at the beginning of day t, $A_t \in \mathcal{A}$ denote the stock ordered at the beginning of day t, and $W_t \in \mathbb{Z}_{\ge 0}$ denote the demand during day t. The dynamics are given by

$$S_{t+1} = \big[S_t + A_t - W_t\big]^{S_{\max}}_{-S_{\max}}$$

where $[\cdot]^{S_{\max}}_{-S_{\max}}$ denotes a function which clips its value between $-S_{\max}$ and $S_{\max}$. The demand W_t is assumed to be an i.i.d. Binomial(n,q) process. The per-step cost is given by

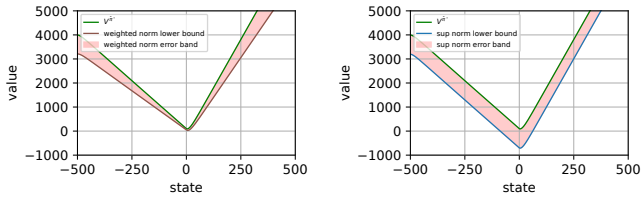
$$c(s,a) = pa + c_h s \mathbb{1}_{\{s\ge 0\}} - c_s s \mathbb{1}_{\{s<0\}}$$

c_h is the per-unit holding cost, c_s is the per-unit shortage cost, and p is the per-unit procurement cost. We denote the above model by $\mathcal{M} = (S_{\max}, \gamma, n, q, c_h, c_s, p)$.

We consider two models:

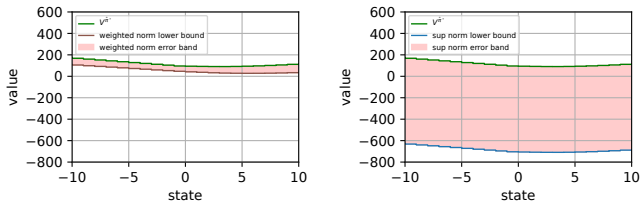
- True model $\mathcal{M} = (500, 0.75, 10, 0.4, 4.0, 2, 5)$.
- Approx. model $\hat{\mathcal{M}} = (500, 0.75, 10, 0.5, 3.8, 2, 5)$.

Since both models have finite state and action spaces, Assumption 1 is satisfied. We take the weight function to be $w(s) = 1 + (1.5\cdot 10^{-2})\big[\hat{c}_h s\mathbb{1}_{\{s\ge 0\}} - \hat{c}_s s \mathbb{1}_{\{s<0\}}\big]$, where $\hat{c}_h$ and $\hat{c}_s$ denote the per-unit holding and shortage costs of the approximate model, respectively. We verify that Assumption 2 is satisfied with $\kappa = 1.15$,



(a) weighted-norm bound (b) sup-norm bound

Fig. 1: Comparison of the bounds on $V^\star(s)$ based on weighted-norm and sup-norm.



(a) weighted-norm bound (b) sup-norm bound

Fig. 2: Zoomed in versions of the bounds of Fig. 1

We compare the bounds obtained for the weighted norm (Theorem 1) with the sup-norm bounds (Corollary 1). In particular, the weighted-norm bound of Theorem 1 states that

$$V^{\hat{\pi}^\star}(s) - V^\star(s) \le \frac{1}{1-\gamma\kappa}\big[\mathcal{D}^{\hat{\pi}^\star}_w \hat{V}^\star + \mathcal{D}^\star_w \hat{V}^\star\big] w(s) \quad (9)$$

while the sup-norm bound of Corollary 1 states that

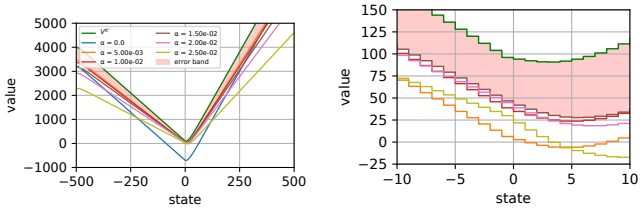
$$V^{\hat{\pi}^\star}(s) - V^\star(s) \le \frac{1}{1-\gamma}\big[\mathcal{D}^{\hat{\pi}^\star}_\infty \hat{V}^\star + \mathcal{D}^\star_\infty \hat{V}^\star\big] \quad (10)$$

For the models $\mathcal{M}$ and $\hat{\mathcal{M}}$ described above, we compute the policy $\hat{\pi}^\star$ using value iteration, compute $V^{\hat{\pi}^\star}$ using policy evaluation, and then plot the weighted- and sup-norm bounds of (9) and (10) in Fig. 1. To better compare the error bounds, we zoom into the region of $\bar{\mathcal{S}} := \{-10, -9, \dots, 10\}$ in Fig. 2, which shows that the weighted-norm is significantly better than the sup-norm for small values of start state.

The optimal policy for an inventory management model described above is a base-stock policy [32]: $\pi^\star(s) = \max(0, s^\star - s)$. For the model $\hat{\mathcal{M}}$, the base-stock level $s^\star = 2$. Since the demand has finite support of $\{0, 1, \dots, 10\}$, after an initial transient period, the inventory level always remains between $\{-8, -7, \dots, 2\}$. Thus, we care about the performance of an approximate policy in this region and, here, the weighted-norm bounds are substantially tighter than the sup-norm bounds. ***These results show that even for finite state and action spaces, weighted-norm bounds can be better than sup-norm bounds.***

B. Initial state dependent weight function

The bounds shown in Fig. 1 show that for smaller values of s, the weighted norm bound is tighter but for significantly larger values of s, the sup-norm bound becomes tighter. This is a general feature of our bounds: the best choice of weight function depends on the value of state. As discussed



(a) entire state space (b) zoomed in

Fig. 3: Upper bounds obtained by different weight functions. Note that the curve corresponding to $\alpha = 0$ is not visible in the zoomed in plot (b).

in Remark 4, we can exploit this feature to come up with a tighter bound.

To illustrate this result, we reconsider the inventory management problem of the previous section and consider a family of weight functions:

$$\mathcal{W} = \{1+\alpha\bar{c}(s) : \alpha \in \{0, 0.5\cdot 10^{-2}, 10^{-2}, \dots, 2.5\cdot 10^{-2}\}\},$$

where $\bar{c}(s) = \hat{c}_h s\mathbb{1}_{\{s\geq 0\}} - \hat{c}_s s\mathbb{1}_{\{s<0\}}$. Note that for $\alpha = 0$, $w(s) = 1$ and, therefore, this corresponds to the sup-norm. For each $w \in \mathcal{W}$, we compute the smallest κ_w such that Assumption 2 is satisfied. We plot the corresponding upper bound given in (9) in Fig. 3. As can be seen from the figure, the best choice of weight function depends on the state. As per Remark 4, we can get a tighter bound by minimizing over all the upper bounds by (7). This tighter upper bound is highlighted in Fig. 3 using shaded areas shown in red.

C. Linear quadratic regulator

In this section, we use the linear quadratic regulator (LQR) to show that weighted norm bounds of Theorem 1 provide meaningful results for models with unbounded per-step cost. Consider a LQR problem with state space $\mathcal{S} = \mathbb{R}^{n_s}$ and action space $\mathcal{A} = \mathbb{R}^{n_a}$. The dynamics are given by

$$s_{t+1} = As_t + Ba_t + w_t,$$

where A and B are system matrices of appropriate dimensions and $\{w_t\}_{t\geq 1}$ is an i.i.d. zero-mean noise process with covariance Σ_W. The per-step cost is given by

$$c(s_t, a_t) = s_t^\intercal Q s_t + a_t^\intercal R a_t,$$

where Q and R are positive semidefinite and positive definite matrices of appropriate dimensions. We will denote this model by $\mathcal{M} = (A, B, Q, R, \Sigma_W, \gamma)$ where γ is the discount factor.

Under standard assumptions of stabilizability and detectability, it is known that the optimal value function is

$$V^\star(s_t) = s_t^\intercal P s_t + q,$$

where P is the solution of the discounted Riccati equation [33]

$$P = Q+\gamma A^\intercal PA-\gamma^2 A^\intercal PB(R+\gamma B^\intercal PB)^{-1}B^\intercal PA, \quad (11)$$

and $q = \gamma \operatorname{Tr}(\Sigma_W P)/(1-\gamma)$.

We consider two models, a true model $\mathcal{M} = (A, B, Q, R, \Sigma_W, \gamma)$ and an approximate model $\hat{\mathcal{M}} = (\hat{A}, \hat{B}, \hat{Q}, \hat{R}, \hat{\Sigma}_W, \gamma)$. Under standard conditions of stabilizability and detectability, both models $\mathcal{M}$ and $\hat{\mathcal{M}}$ satisfy Assumption 1. Let P and $\hat{P}$ denote the solution of the Riccati equations corresponding to models $\mathcal{M}$ and $\hat{\mathcal{M}}$.

We take the weight function to be $w(s) = 1 + s^\intercal s$ and assume that models $\mathcal{M}$ and $\hat{\mathcal{M}}$ are close enough that Assumption 2 is satisfied for some $\kappa < 1/\gamma$. We follow the same notation as before and let $\hat{\pi}^\star$ denote the optimal policy of model $\hat{\mathcal{M}}$ and use $V^{\hat{\pi}^\star}$ and $V^\star$ to denote the value function of policy $\hat{\pi}^\star$ and the optimal value function for model $\mathcal{M}$, respectively.

Then, the result of Theorem 1 simplifies as follows:

Proposition 3 *Under Assumptions 1 and 2, we have*

$$\begin{aligned}&\left\|V^{\hat{\pi}^\star} - V^\star\right\|_w \\ &\leq \frac{1}{1-\gamma\kappa}\left[\max(\rho(D^\star), d_\Sigma) + \max(\rho(D^{\hat{\pi}^\star}), d_\Sigma)\right], \quad (12)\end{aligned}$$

where $\rho(\cdot)$ denotes the spectral radius of a matrix and

$$\begin{aligned}D^\star &= Q - \hat{Q} + \gamma(A^\intercal \hat{P} A - \hat{A}^\intercal \hat{P}\hat{A}) \\ &\quad - \gamma^2(A^\intercal \hat{P}B(R+\gamma B^\intercal \hat{P}B)^{-1}B^\intercal \hat{P}A \\ &\quad - \hat{A}^\intercal \hat{P}\hat{B}(\hat{R}+\gamma \hat{B}^\intercal \hat{P}\hat{B})^{-1}\hat{B}^\intercal \hat{P}\hat{A}), \quad (13)\\ D^{\hat{\pi}^\star} &= Q - \hat{Q} + \gamma(A^\intercal \hat{P} A - \hat{A}^\intercal \hat{P}\hat{A}) \\ &\quad - \gamma\left(A^\intercal \hat{P}B\hat{K} + \hat{K}^\intercal B^\intercal \hat{P}A\right) \\ &\quad + \hat{K}^\intercal(R+\hat{R}+\gamma B^\intercal \hat{P}B + \gamma \hat{B}^\intercal \hat{P}\hat{B})\hat{K}, \quad (14)\\ \hat{K} &= \gamma(\hat{R}+\gamma \hat{B}^\intercal \hat{P}\hat{B})^{-1}\hat{B}^\intercal \hat{P}\hat{A}, \quad (15)\end{aligned}$$

and

$$d_\Sigma = \gamma \operatorname{Tr}((\Sigma_W - \hat{\Sigma}_W)\hat{P}). \quad (16)$$

Remark 5 An immediate implication of Proposition 3 is that for $s = 0$,

$$\begin{aligned}V^{\hat{\pi}^\star}(0) - V^\star(0) &\leq \left\|V^{\hat{\pi}^\star} - V^\star\right\|_w w(0) \\ &\leq \frac{1}{1-\gamma\kappa}\left[\max(\rho(D^\star), d_\Sigma) + \max(\rho(D^{\hat{\pi}^\star}), d_\Sigma)\right]. \quad (17)\end{aligned}$$

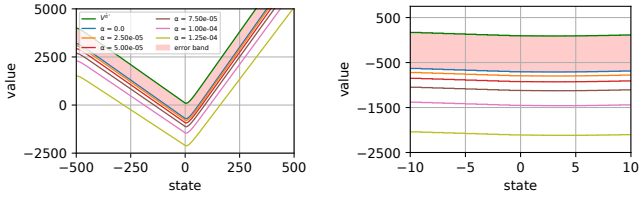
A salient feature of this upper bound is that it does not depend on Riccati gain P or the control gain K of the true model.

D. Advantage of using (κ, w)-stability

As mentioned in Remark 1, a condition stronger than (κ, w)-stability is typically imposed in the literature. In this section, we show that if we restrict attention to weight functions which satisfy this stronger condition (described in Remark 1), then the upper bound is looser or even non-applicable.

For the LQR problem, (5) in Remark 1 cannot be satisfied for $w(s) = 1 + s^\intercal s$ with any finite constant $\bar{\kappa}$ due to the unbounded action space. For the inventory management problem, considering a family of weight functions

$$\bar{\mathcal{W}} = \{1 + \alpha\bar{c}(s) : \alpha \in \{0, 0.25\cdot 10^{-4}, \dots, 2.00\cdot 10^{-4}\}\},$$



(a) entire state space (b) zoomed in

Fig. 4: Upper bounds obtained using stability for all actions. Note that the curves corresponding to $\alpha = 1.50 \cdot 10^{-4}$ and $\alpha = 1.75 \cdot 10^{-4}$ are not visible in both plots.

for each $\bar{w} \in \bar{\mathcal{W}}$, we compute the smallest $\bar{\kappa}_{\bar{w}}$ such that (5) is satisfied. Greatest possible α that yields the corresponding $\bar{\kappa}_{\bar{w}} < 1/\gamma$ is found as $\alpha = 1.75 \cdot 10^{-4}$.

We plot the corresponding upper bound given in (9) in Fig. 4. As can be seen from the plot, in this case the weight function $\bar{w}(s) \equiv 1$ (equivalent to the sup-norm) gives the tightest upper bound. But, as was seen by the bounds of Fig. 3, the bounds obtained by weighted functions in class $\mathcal{W}$ were significantly tighter. This highlights the importance of working with weight functions which satisfy Assumption 2 rather than weight functions which satisfy the conditions of Remark 1.

V. Conclusion

We considered the problem of designing a control policy for an infinite-horizon discounted cost Markov Decision Process (MDP) $\mathcal{M}$ when we only have access to an approximate model $\hat{\mathcal{M}}$. We provided an upper bound on the performance loss when an optimal policy $\hat{\pi}^*$ for the approximate model is used in the true model and the start state is s. Our bounds are in terms of the weighted norm of the difference between the value function of $\hat{\pi}^\star$ when used in $\mathcal{M}$ and the optimal value function of $\mathcal{M}$. The use of weighted norm allows us to obtain meaningful bounds for performance loss even when the per-step cost function is unbounded. While we focused on MDPs in this paper, our approach may prove to be useful for partially observed and multi-agent systems as well.

References

[1] B. L. Fox, "Finite-state approximations to denumerable-state dynamic programs," *J. Math. Analy. Appl.*, vol. 34, no. 3, pp. 665–670, 1971.

[2] W. Whitt, "Approximations of dynamic programs, I," *Math. Oper. Res.*, vol. 3, no. 3, pp. 231–243, 1978.

[3] ——, "Approximations of dynamic programs, II," *Math. Oper. Res.*, vol. 4, no. 2, pp. 179–185, 1979.

[4] ——, "Representation and approximation of noncooperative sequential games," *SIAM J. Contr. Optim.*, vol. 18, no. 1, pp. 33–48, 1980.

[5] D. Bertsekas, "Convergence of discretization procedures in dynamic programming," *IEEE Trans. Autom. Control*, vol. 20, no. 3, pp. 415–419, 1975.

[6] C.-S. Chow and J. N. Tsitsiklis, "An optimal one-way multigrid algorithm for discrete-time stochastic control," *IEEE transactions on automatic control*, vol. 36, no. 8, pp. 898–914, 1991.

[7] F. Dufour and T. Prieto-Rumeau, "Approximation of Markov decision processes with general state space," *J. Math. Analy. Appl.*, vol. 388, no. 2, pp. 1254–1267, 2012.

[8] A. Haurie and P. L'ecuyer, "Approximation and bounds in discrete event dynamic programming," *IEEE Trans. Autom. Control*, vol. 31, no. 3, pp. 227–235, 1986.

[9] A. Müller, "How does the value function of a Markov decision process depend on the transition probabilities?" *Math. Oper. Res.*, vol. 22, no. 4, pp. 872–885, 1997.

[10] N. Saldi, T. Linder, and S. Yüksel, "Asymptotic optimality and rates of convergence of quantized stationary policies in stochastic control," *IEEE Trans. Autom. Control*, vol. 60, no. 2, pp. 553–558, 2014.

[11] ——, *Finite Approximations in discrete-time stochastic control*. Springer, 2018.

[12] A. D. Kara, "Near optimality of finite memory feedback policies in partially observed markov decision processes," *J. Mach. Learn. Res.*, vol. 23, no. 1, pp. 437–482, 2022.

[13] J. Subramanian, A. Sinha, R. Seraj, and A. Mahajan, "Approximate information state for approximate planning and reinforcement learning in partially observed systems." *J. Mach. Learn. Res.*, vol. 23, no. 12, pp. 1–83, 2022.

[14] B. L. Fox, "Discretizing dynamic programs," *J. Opt. Theory and Appl.*, vol. 11, pp. 228–234, 1973.

[15] P. K. Dutta, M. K. Majumdar, and R. K. Sundaram, "Parametric continuity in dynamic programming problems," *J. Economic Dynamics and Control*, vol. 18, no. 6, pp. 1069–1092, 1994.

[16] N. Saldi, S. Yüksel, and T. Linder, "Near optimality of quantized policies in stochastic control under weak continuity conditions," *J. Math. Analy. Appl.*, vol. 435, no. 1, pp. 321–337, 2016.

[17] ——, "On the asymptotic optimality of finite approximations to Markov decision processes with borel spaces," *Math. Oper. Res.*, vol. 42, no. 4, pp. 945–978, 2017.

[18] ——, "Asymptotic optimality of finite model approximations for partially observed markov decision processes with discounted cost," *IEEE Trans. Autom. Control*, vol. 65, no. 1, pp. 130–142, 2019.

[19] A. D. Kara and S. Yuksel, "Robustness to incorrect priors in partially observed stochastic control," *SIAM J. Contr. Optim.*, vol. 57, no. 3, pp. 1929–1964, 2019.

[20] ——, "Robustness to incorrect system models in stochastic control," *SIAM J. Cont. Optim.*, vol. 58, no. 2, pp. 1144–1182, 2020.

[21] A. D. Kara, M. Raginsky, and S. Yüksel, "Robustness to incorrect models and data-driven learning in average-cost optimal stochastic control," *Automatica*, vol. 139, p. 110179, 2022.

[22] B. Ravindran and A. G. Barto, "Approximate homomorphisms: A framework for non-exact minimization in Markov decision processes," in *KBCS*, 2004.

[23] E. van der Pol, T. Kipf, F. A. Oliehoek, and M. Welling, "Plannable approximations to mdp homomorphisms: Equivariance under actions," *arXiv preprint arXiv:2002.11963*, 2020.

[24] N. Ferns, P. Panangaden, and D. Precup, "Metrics for finite Markov decision processes." in *UAI*, vol. 4, 2004, pp. 162–169.

[25] ——, "Bisimulation metrics for continuous Markov decision processes," *SIAM J. Comp.*, vol. 40, no. 6, pp. 1662–1714, 2011.

[26] P. S. Castro, P. Panangaden, and D. Precup, "Equivalence relations in fully and partially observable Markov decision processes." in *IJCAI*, vol. 9, 2009, pp. 1653–1658.

[27] D. Abel, D. Hershkowitz, and M. Littman, "Near optimal behavior via approximate state abstraction," in *ICML*. PMLR, 2016, pp. 2915–2923.

[28] V. François-Lavet, G. Rabusseau, J. Pineau, D. Ernst, and R. Fonteneau, "On overfitting and asymptotic bias in batch reinforcement learning with partial observability," *J. Artif. Intel. Res.*, vol. 65, pp. 1–30, 2019.

[29] C. Gelada, S. Kumar, J. Buckman, O. Nachum, and M. G. Bellemare, "Deepmdp: Learning continuous latent space models for representation learning," in *ICML*. PMLR, 2019, pp. 2170–2179.

[30] O. Hernández-Lerma and J. B. Lasserre, *Discrete-time Markov control processes: basic optimality criteria*. Springer Science & Business Media, 2012.

[31] ——, *Further topics on discrete-time Markov control processes*. Springer Science & Business Media, 2012.

[32] K. J. Arrow, T. Harris, and J. Marschak, "Optimal inventory policy," *Econometrica: Journal of the Econometric Society*, pp. 250–272, 1951.

[33] D. P. Bertsekas, "Dynamic programming and optimal control," *Athena Scientific*, 2015.

APPENDIX I
PROOF OF LEMMA 1

We prove each part of Lemma 1 separately.

Proof of part 1)

Fix a state $s \in \mathcal{S}$. For a policy $\pi \in \Pi_S(\kappa, M, w)$ and a value function $v \in \bar{\mathcal{V}}_w$, we have

$$\begin{aligned}
&\left|\frac{\mathcal{B}^\pi v(s)}{w(s)}\right| \\
&\overset{(a)}{\leq} \left|\frac{c_{\pi(s)}}{w(s)}\right| + \gamma \left|\int_{\mathcal{A}} \pi(da \mid s) \int_{\mathcal{S}} P(ds' \mid s, a) \frac{v(s')}{w(s')} \frac{w(s')}{w(s)}\right| \\
&\overset{(b)}{\leq} \|c_\pi\|_w + \gamma \|v\|_w \left|\int_{\mathcal{A}} \pi(da \mid s) \int_{\mathcal{S}} P(ds' \mid s, a) \frac{w(s')}{w(s)}\right| \\
&\overset{(c)}{\leq} \|c_\pi\|_w + \gamma \|v\|_w \kappa < \infty,
\end{aligned}$$

where (a) follows from the triangle inequality, (b) follows from the definition of $\|\cdot\|_w$ and (c) follows from the fact that π is (κ, w) stable.

Proof of part 2)

Fix a state $s \in \mathcal{S}$. We have

$$\begin{aligned}
&\left|\frac{[\mathcal{B}^\pi v_1 - \mathcal{B}^\pi v_2](s)}{w(s)}\right| \\
&= \gamma \left|\int_{\mathcal{A}} \pi(da \mid s) \int_{\mathcal{S}} P(ds' \mid s, a) \left[\frac{v_1(s') - v_2(s')}{w(s')}\right] \frac{w(s')}{w(s)}\right| \\
&\overset{(a)}{\leq} \gamma \|v_1 - v_2\|_w \left|\int_{\mathcal{A}} \pi(da \mid s) \int_{\mathcal{S}} P(ds' \mid s, a) w(s') \frac{1}{w(s)}\right| \\
&\overset{(b)}{\leq} \gamma\kappa \|v_1 - v_2\|_w
\end{aligned}$$

where (a) holds from the definition of $\|\cdot\|_w$ and (b) holds because π is (κ, w) stable.

Proof of part 3)

From parts 1) and 2) of Lemma 1, we know that $\mathcal{B}^\pi : \bar{\mathcal{V}}_w \mapsto \bar{\mathcal{V}}_w$ is a contraction. Since $\bar{\mathcal{V}}_w$ is a complete metric space (under the $\|\cdot\|_w$ metric), it follows from Banach fixed point theorem that $\mathcal{B}^\pi$ has a unique fixed point F in $\bar{\mathcal{V}}_w$. If V_n^π denotes the $n-$step discounted cost for policy π, then it can be shown that $V_{n+1}^\pi = \mathcal{B}^\pi V_n^\pi$ and that $V_n^\pi \in \bar{\mathcal{V}}_w$ for all n. Thus, by Banach fixed point theorem, V_n^π converges to the fixed point F of $\mathcal{B}^\pi$ in the $\|\cdot\|_w$ metric. Since convergence in $\|\cdot\|_w$ metric implies pointwise convergence, we have $F(s) = \lim_{n\to\infty} V_n^\pi(s) = V^\pi(s)$ for all $s \in \mathcal{S}$.

APPENDIX II
PROOF OF THEOREM 1

We prove each part separately.

A. Proof of part 1

By triangle inequality, we have

$$\left\|V^{\hat{\pi}^\star} - V^\star\right\|_w \leq \left\|V^{\hat{\pi}^\star} - \hat{V}^\star\right\|_w + \left\|\hat{V}^\star - V^\star\right\|_w \tag{18}$$

We will bound each term separately. For the first term, using $\hat{V}^\star = \hat{V}^{\hat{\pi}^\star}$, we can write

$$\begin{aligned}
\left\|V^{\hat{\pi}^\star} - \hat{V}^\star\right\|_w &= \left\|\mathcal{B}^{\hat{\pi}^\star} V^{\hat{\pi}^\star} - \hat{\mathcal{B}}^{\hat{\pi}^\star} \hat{V}^\star\right\|_w \\
&\leq \left\|\mathcal{B}^{\hat{\pi}^\star} V^{\hat{\pi}^\star} - \mathcal{B}^{\hat{\pi}^\star} \hat{V}^\star\right\|_w + \left\|\mathcal{B}^{\hat{\pi}^\star} \hat{V}^\star - \hat{\mathcal{B}}^{\hat{\pi}^\star} \hat{V}^\star\right\|_w \\
&\leq \gamma\kappa \left\|V^{\hat{\pi}^\star} - \hat{V}^\star\right\|_w + \mathcal{D}_w^{\hat{\pi}^\star} \hat{V}^{\hat{\pi}^\star}
\end{aligned} \tag{19}$$

where the first inequality follows from triangle inequality, and the last from Lemma 1 as $\hat{\pi}^\star \in \Pi_S(\kappa, w)$ and from the definition of Bellman mismatch functional. Re-arranging the terms in (19), we obtain

$$\left\|V^{\hat{\pi}^\star} - \hat{V}^\star\right\|_w \leq \frac{1}{1-\gamma\kappa} \mathcal{D}_w^{\hat{\pi}^\star} \hat{V}^\star. \tag{20}$$

By a similar argument, we can show that

$$\left\|\hat{V}^\star - V^\star\right\|_w \leq \frac{1}{1-\gamma\kappa} \mathcal{D}_w^\star \hat{V}^\star. \tag{21}$$

Combining (20) and (21), we have

$$\left\|V^{\hat{\pi}^\star} - V^\star\right\|_w \leq \frac{1}{1-\gamma\kappa} \left[\mathcal{D}_w^{\hat{\pi}^\star} \hat{V}^\star + \mathcal{D}_w^\star \hat{V}^\star\right]. \tag{22}$$

B. Proof of part 2

We have

$$\begin{aligned}
\left\|V^{\hat{\pi}^\star} - V^\star\right\|_w &= \left\|\mathcal{B}^{\hat{\pi}^\star} V^{\hat{\pi}^\star} - \mathcal{B}^\star V^\star\right\|_w \\
&\leq \left\|\mathcal{B}^{\hat{\pi}^\star} V^{\hat{\pi}^\star} - \mathcal{B}^{\hat{\pi}^\star} V^\star\right\|_w + \left\|\mathcal{B}^{\hat{\pi}^\star} V^\star - \hat{\mathcal{B}}^{\hat{\pi}^\star} V^\star\right\|_w \\
&\quad + \left\|\hat{\mathcal{B}}^{\hat{\pi}^\star} V^\star - \hat{\mathcal{B}}^{\hat{\pi}^\star} \hat{V}^\star\right\|_w + \left\|\hat{\mathcal{B}}^{\hat{\pi}^\star} \hat{V}^\star - \mathcal{B}^\star V^\star\right\|_w \\
&\leq \gamma\kappa \left\|V^{\hat{\pi}^\star} - V^\star\right\|_w + \mathcal{D}_w^{\hat{\pi}^\star} V^\star \\
&\quad + \gamma\kappa \left\|V^\star - \hat{V}^\star\right\|_w + \left\|\hat{V}^\star - V^\star\right\|_w
\end{aligned} \tag{23}$$

where the first inequality holds from triangle inequality and the last from the definition of Bellman mismatch functional and from Lemma 1 as $\hat{\pi}^\star \in \Pi_S(\kappa, w) \cap \hat{\Pi}_S(\kappa, w)$. Re-arranging the terms in (23), we obtain

$$\left\|V^{\hat{\pi}^\star} - V^\star\right\|_w \leq \frac{1}{1-\gamma\kappa} \left[\mathcal{D}_w^{\hat{\pi}^\star} V^\star + (1+\gamma\kappa) \left\|\hat{V}^\star - V^\star\right\|_w\right]. \tag{24}$$

We can bound the last term of (24) as

$$\begin{aligned}
\left\|\hat{V}^\star - V^\star\right\|_w &= \left\|\mathcal{B}^\star V^\star - \hat{\mathcal{B}}^\star \hat{V}^\star\right\|_w \\
&\leq \left\|\mathcal{B}^\star V^\star - \hat{\mathcal{B}}^\star V^\star\right\|_w + \left\|\hat{\mathcal{B}}^\star V^\star - \hat{\mathcal{B}}^\star \hat{V}^\star\right\|_w \\
&\leq \mathcal{D}_w^\star V^\star + \gamma\kappa \left\|V^\star - \hat{V}^\star\right\|_w
\end{aligned} \tag{25}$$

Re-arranging the terms in (25), we obtain

$$\left\|\hat{V}^\star - V^\star\right\|_w \leq \frac{1}{1-\gamma\kappa} \mathcal{D}_w^\star V^\star. \tag{26}$$

Combining (24) and (26), we have

$$\left\|\hat{V}^\star - V^\star\right\|_w \leq \frac{1}{1-\gamma\kappa} \mathcal{D}_w^{\hat{\pi}^\star} V^\star + \frac{1+\gamma\kappa}{(1-\gamma\kappa)^2} \mathcal{D}_w^\star V^\star. \tag{27}$$

Appendix III
Proof of Proposition 1

Note that $\Xi^{(s,a)}v$ may be written as

$$\Xi^{(s,a)}v = [\mathcal{B}^{\pi}v](s) - [\hat{\mathcal{B}}^{\pi}v](s)$$

where π is such that $\pi(s) = a$ for all s. Let Π_O denote all such deterministic open loop policies, i.e., policies where $\pi(s)$ is a constant action. Then,

$$\mathcal{D}_w^{\max}v = \sup_{\pi\in\Pi_O} \mathcal{D}_w^{\pi}v.$$

Since $\Pi_O \subset \Pi$, the above equation implies that

$$\mathcal{D}_w^{\max}v \leq \sup_{\pi\in\Pi} \mathcal{D}_w^{\pi}v. \tag{28}$$

Now consider any $\pi \in \Pi$. Then,

$$[\mathcal{B}^{\pi}v](s) - [\hat{\mathcal{B}}^{\pi}v](s) = \int_{\mathcal{A}} \pi(da \mid s)\Xi^{(s,a)}v$$

Therefore,

$$\begin{aligned}\mathcal{D}_w^{\pi}v &= \sup_{s\in\mathcal{S}} \frac{\left|\int_{\mathcal{A}} \pi(da \mid s)\Xi^{(s,a)}v\right|}{w(s)} \\ &\leq \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{|\Xi^{(s,a)}v|}{w(s)} = \mathcal{D}_w^{\max}v. \end{aligned}\tag{29}$$

Combining (28) and (29), we get the first part of (8).

For the second part, note that for any set $\mathcal{X}$, $|\sup_{x\in\mathcal{X}} f(x) - \sup_{x\in\mathcal{X}} g(x)| \leq \sup_{x\in\mathcal{X}} |f(x) - g(x)|$. Therefore,

$$[\mathcal{B}^{\star}v](s) - [\hat{\mathcal{B}}^{\star}v](s) \leq \sup_{a\in\mathcal{A}} |\Xi^{(s,a)}v|.$$

Hence, we have

$$\mathcal{D}_w^{\star}v = \sup_{s\in\mathcal{S}} \frac{\sup_{a\in\mathcal{A}}\left|\Xi^{(s,a)}v\right|}{w(s)} = \mathcal{D}_w^{\max}v, \tag{30}$$

which establishes the second part of (8).

Appendix IV
Proof of Proposition 3

We can calculate the Bellman updates for $\hat{V}^{\star}(s)$ as

$$\begin{aligned}&\mathcal{B}^{\star}\hat{V}^{\star}(s) \\ &= s^{\intercal}\left(Q + \gamma A^{\intercal}\hat{P}A - \gamma^2 A^{\intercal}\hat{P}B(R+\gamma B^{\intercal}\hat{P}B)^{-1}B^{\intercal}\hat{P}A\right)s \\ &\quad + \gamma(\hat{q} + \operatorname{Tr}(\Sigma_W\hat{P})).\end{aligned}$$

The approximate Bellman update for $\hat{V}^{\star}(s)$ is given by

$$\begin{aligned}&\hat{\mathcal{B}}^{\star}\hat{V}^{\star}(s) \\ &= s^{\intercal}\left(\hat{Q} + \gamma \hat{A}^{\intercal}\hat{P}\hat{A} - \gamma^2 \hat{A}^{\intercal}\hat{P}\hat{B}(\hat{R}+\gamma\hat{B}^{\intercal}\hat{P}\hat{B})^{-1}\hat{B}^{\intercal}\hat{P}\hat{A}\right)s \\ &\quad + \gamma(\hat{q} + \operatorname{Tr}(\hat{\Sigma}_W\hat{P})).\end{aligned}$$

Therefore, we have

$$|\mathcal{B}^{\star}\hat{V}^{\star}(s) - \hat{\mathcal{B}}^{\star}\hat{V}^{\star}(s)| = \left|s^{\intercal}D^{\star}s + \gamma\operatorname{Tr}((\Sigma_W - \hat{\Sigma}_W)\hat{P})\right|,$$

where $D^{\star}$ is given by (13).

Note that $\hat{\pi}^{\star}(s) = -\hat{K}s = -\gamma(\hat{R}+\gamma\hat{B}^{\intercal}\hat{P}\hat{B})^{-1}\hat{B}^{\intercal}\hat{P}\hat{A}s$. As a result, $\mathcal{B}^{\hat{\pi}^{\star}}$ is given by

$$\begin{aligned}&\mathcal{B}^{\hat{\pi}^{\star}}\hat{V}^{\star}(s) \\ &= s^{\intercal}\Big(Q + \gamma A^{\intercal}\hat{P}A - \gamma\big(A^{\intercal}\hat{P}B\hat{K} + \hat{K}^{\intercal}B^{\intercal}\hat{P}A\big) \\ &\qquad + \hat{K}^{\intercal}(R+\gamma B^{\intercal}\hat{P}B)\hat{K}\big)\Big)s \\ &\quad + \gamma(\hat{q} + \operatorname{Tr}(\Sigma_W\hat{P})),\end{aligned}$$

and $\hat{\mathcal{B}}^{\hat{\pi}^{\star}}\hat{V}^{\star}(s) = \hat{\mathcal{B}}^{\star}\hat{V}^{\star}(s)$. Therefore, we have

$$|\mathcal{B}^{\hat{\pi}^{\star}}\hat{V}^{\star}(s) - \hat{\mathcal{B}}^{\hat{\pi}^{\star}}\hat{V}^{\star}(s)| = \left|s^{\intercal}D^{\hat{\pi}^{\star}}s + \gamma\operatorname{Tr}((\Sigma_W - \hat{\Sigma}_W)\hat{P})\right|,$$

where $D^{\hat{\pi}^{\star}}$ is given by (14). Then, the Bellman mismatches in Definition 5 for $\hat{V}$ can be calculated as follows:

$$\mathcal{D}_w^{\pi}\hat{V}^{\star} = \sup_{s\in\mathcal{S}} \frac{\left|s^{\intercal}D^{\hat{\pi}^{\star}}s + d_{\Sigma}\right|}{w(s)}, \quad \mathcal{D}_w^{\star}\hat{V}^{\star} = \sup_{s\in\mathcal{S}} \frac{|s^{\intercal}D^{\star}s + d_{\Sigma}|}{w(s)},$$

where d_{Σ} is given by (16). Since we take the weight function to be $w(s) = 1 + s^{\intercal}s$, Theorem 1 gives

$$\begin{aligned}&\left\|V^{\hat{\pi}^{\star}} - V^{\star}\right\|_w \\ &\leq \frac{1}{1-\gamma\kappa}\left[\sup_{s\in\mathcal{S}} \frac{\left|s^{\intercal}D^{\hat{\pi}^{\star}}s + d_{\Sigma}\right|}{1+s^{\intercal}s} + \sup_{s\in\mathcal{S}} \frac{\left|s^{\intercal}D^{\hat{\pi}^{\star}}s + d_{\Sigma}\right|}{1+s^{\intercal}s}\right] \\ &= \frac{1}{1-\gamma\kappa}\left[\max(\rho(D^{\star}), d_{\Sigma}) + \max(\rho(D^{\hat{\pi}^{\star}}), d_{\Sigma})\right].\end{aligned}$$