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 $\mathcal{M}$ when we only have access to an approximate model $\hat{\mathcal{M}}$. How well does an optimal policy $\hat{\pi}^\star$ of the approximate model perform when used in the original model $\mathcal{M}$? We answer 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 $\mathcal{M}$. We then extend our results and obtain potentially tighter upper bounds by considering affine transformations of the per-step cost. We further provide upper bounds that explicitly depend on the weighted distance between cost functions and weighted distance between transition kernels of the original and approximate models. We present examples to illustrate our results.**

***Index Terms*—Markov decision processes, model approximation, Bellman operators, integral probability metrics.**

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 the 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 [2], 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 [3]–[5]. Similar results for state discretization were obtained in [6], [7], state and action discretization in [8] and for models with state dependent discounting in [9]. A general framework to view model approximation using the lens of integral probability metrics was presented by Müller [10]. There have been considerable advances on these ideas in recent years [11]–[13], including generalizations to partially observed models [14], [15]. However, these approximation results are restricted to models with bounded per-step cost.

A preliminary version of this paper was presented at CDC 2023 [1].

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 at McGill was supported by IDEaS grant CFPMN2-30, NSERC grant RGPIN-2021-03511, and IVADO MSc Excellence Fellowship. The work at USC was supported by NSF grants ECCS 2025732 and ECCS 1750041.

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 $\mathcal{M}$? One of the earliest work in this direction is that of Fox [16], who studied the continuity of state discretization procedures. Sufficient conditions for continuity of value functions on model parameters were presented in [17]. There is a series of recent papers which significantly generalize these results, including characterizing conditions under which the optimal policy is continuous in model parameters [11]–[13], [18]–[22].

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 [23], [24], bisimulation metrics [25]–[27], state abstraction [28], and approximate latent state models [29], [30].

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 function 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 $P^{\hat{\pi}^\star}$). See Sec. IV-D for exact details. For this model, and for many 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 result, do not provide any insights into the quality of the approximation.

The standard approach to deal with unbounded per-step cost is to use a weighted norm rather than a sup norm [10],

[12], [13], [31], [32]. However, in most of the existing literature a weighted norm is used to establish existence and uniqueness of a dynamic programming solution. As far as we are aware, the only works which use the weighted norm for model approximation in models with unbounded per-step cost are [12] and [13, Chapter 4], where the authors 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.

Our main contributions in this paper are as follows:

- We provide upper bounds on the approximation error in terms of the weighted-norm:
$$\|V^{\hat{\pi}^\star} - V^\star\|_w := \sup_{s\in\mathcal{S}} \frac{\left|V^{\hat{\pi}^\star}(s) - V^\star(s)\right|}{w(s)},$$
where $w\colon \mathcal{S} \to [1,\infty)$ is a weight function. Our bounds are derived using a new functional, which we call the *Bellman mismatch functional*.
- In the literature on the existence of dynamic programming solution for models with unbounded per-step cost, it is assumed that the weight function is such that the dynamics under all policies satisfies a Lyaponov stability-type condition [10], [12], [13], [31], [32]. In contrast, we assume that such a stability condition is satisfied for only a few policies, including the optimal policies of the original and the approximate model, and policies obtained by using optimal and approximate value functions as one-step look ahead value function. (See Assumptions 2–4 for the precise definition).
- We then extend our results and obtain potentially tighter upper bounds by considering affine transformations of the per-step cost. These transformations allow us to trade off between the mismatches in the dynamics with the mismatches in the per-step cost.
- We present examples to illustrate that for suitable choices of the weight functions and affine transformations our bounds are tighter than existing sup-norm bounds, even for models with bounded per-step cost. In addition, we revisit the LQR example mentioned previously and show that the weighted-norm approximation bounds provide meaningful approximation guarantees for such unbounded-cost models.
- We provide further upper bounds that explicitly depend on the weighted distance between cost functions and weighted distance between transition kernels of the original and approximate models. In the special case when $w(s) \equiv 1$, our bounds recover the existing sup-norm bounds [10], [28], [30]

Notation: We use calligraphic letters to denote sets (e.g. $\mathcal{S}$), uppercase letters to denote random variables (e.g. S) and lowercase letters to denote their realizations (e.g. s). The space of probability measures on a set $\mathcal{S}$ is expressed by $\Delta(\mathcal{S})$. Subscripts indicate time, so S_t denotes a random variable at time t. $S_{1:t}$ is a short hand notation for $(S_1, \dots, S_t)$.

We use $\mathbb{R}$ to denote the set of real numbers, $\mathbb{Z}_{\geq 0}$ to denote the set of non-negative integers, $\mathbb{P}(\cdot)$ to denote the probability of an event, $\mathbb{E}[\cdot]$ to denote expectation of a random variable, and $\mathbb{1}\{\cdot\}$ to denote indicator of an event. For functions $v_1, v_2\colon \mathcal{S} \to \mathbb{R}$, the notation $v_1 \leq v_2$ denotes that $v_1(s) \leq v_2(s)$ for all $s \in \mathcal{S}$.

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
$$\begin{aligned}&\mathbb{P}(S_{t+1} \in B \mid S_{1:t} = s_{1:t}, A_{1:t} = a_{1:t})\\ &\quad = \mathbb{P}(S_{t+1} \in B \mid S_t = s_t, A_t = a_t) =: P(B \mid s_t, a_t).\end{aligned}$$
- $c\colon \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ is the per-step cost function which is assumed to be measurable. We further assume that the per-step cost is bounded from below, i.e., there exists a finite constant $c_{\min}$ such that $c(s,a) \geq c_{\min}$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$.
- $\gamma \in (0,1)$ is the discount factor.

A stochastic kernel $\pi\colon \mathcal{S} \to \Delta(\mathcal{A})$ 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
$$V^{\pi^\star}(s) \leq V^\pi(s), \quad \forall s \in \mathcal{S}, \forall \pi \in \Pi. \tag{2}$$
Since we consider Borel state and action spaces with possibly unbounded (from above) per-step cost function, an optimal policy is not guaranteed to exist. If an optimal policy exists, its value function is called the *optimal value function*. We focus on MDPs for which optimal value function exists 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 $\mathcal{S} \to [\frac{c_{\min}}{1-\gamma}, \infty) \cup \{+\infty\}$.

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

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

$\mathcal{V}_w = \{v \in \mathcal{V} : \|v\|_w < \infty\}$ and $\mathcal{P}_w = \{p \in \Delta(\mathcal{S}) : \int w dp < \infty\}$.

It can be easily verified that $\|\cdot\|_w$ is a norm and that $\mathcal{V}_w$ is a Banach space.

Remark 1 When the weight function $w(s) \equiv 1$, then the weighted norm $\|v\|_w$ is equal to the sup-norm $\|v\|_\infty := \sup_{s\in\mathcal{S}} |v(s)|$. In this case, $\mathcal{V}_w$ is the set of all bounded functions in $\mathcal{V}$ and $\mathcal{P}_w$ is the set of all probability measures on $\mathcal{S}$.

Definition 2 (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 $v \in \mathcal{V}$,

$$[\mathcal{B}^\pi v](s) = c_\pi(s) + \gamma \int_{\mathcal{S}} v(s') P_\pi(ds' \mid s),$$

  where $c_\pi(s) = \int_{\mathcal{A}} c(s,a)\pi(da|s)$, and $P_\pi(ds'|s) = \int_{\mathcal{A}} \pi(da|s) P(ds'|s,a)$.
- Define the *Bellman optimality operator* $\mathcal{B}^\star$ 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 3 (One-step greedy policy) We say that a policy $\pi \in \Pi$ is *one-step greedy* with respect to a value function $v \in \mathcal{V}$ if $\mathcal{B}^\pi v = \mathcal{B}^\star v$. We denote the set of all one-step greedy policies with respect to v by $\mathcal{G}(v)$.

Definition 4 (Dynamic programming solvability) An MDP $\mathcal{M}$ is said to be *dynamic programming solvable* (DP-solvable, for short) with respect to a weight function $w \colon \mathcal{S} \to [1,\infty)$ if the following conditions are satisfied:

1) For any $v \in \mathcal{V}_w$, $\mathcal{B}^\star v \in \mathcal{V}_w$.
2) There exists a $V^\star \in \mathcal{V}_w$ such that for all $\pi \in \Pi$,

$$V^\star(s) \le V^\pi(s), \quad \forall s \in \mathcal{S}$$

   with equality at all states for at least one $\pi \in \Pi$.
3) $V^\star$ is a fixed point of $\mathcal{B}^\star$, i.e., it satisfies the dynamic programming equation

$$V^\star = \mathcal{B}^\star V^\star.$$

MDPs with finite state and action spaces are always DP-solvable. For MDPs with general state and action spaces, there are several conditions in the literature which imply DP-solvability. See [31] for an overview.

C. Weighted-norm stability

Definition 5 ((κ, w) stability of a policy) 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

$$\int_{\mathcal{S}} w(s') P_\pi(ds' \mid s) \le \kappa w(s), \quad \forall s \in \mathcal{S}, \tag{4}$$

where $P_\pi(ds'|s) = \int_{\mathcal{A}} \pi(da|s) P(ds'|s,a)$.

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

Remark 2 As stated in Remark 1, when $w(s) \equiv 1$, the weighted norm is the same as the sup-norm. Furthermore, with $w(s) \equiv 1$, inequality (4) of Definition 5 holds with $\kappa = 1$ for any policy π. Thus, for the case of $w(s) \equiv 1$, $\Pi_S(1, w \equiv 1)$ is the set of all policies π for which $\|c_\pi\|_\infty < \infty$.

For the model approximation results that are developed later, we will assume that certain policies are (κ, w) stable. It is worthwhile to contrast (κ, w)-stability of a policy with a stronger assumption that is typically imposed in the literature [10], [12], [31], [32]. To make that comparison, we define the following (which is the same as [32, Assumption 8.3.2]):

Definition 6 ($(\bar\kappa, \bar w)$ stability of the model) Given an MDP $\mathcal{M}$ and a tuple $(\bar\kappa, \bar w)$, where $\bar\kappa$ is a positive constant with $\gamma\bar\kappa < 1$ and $\bar w$ is a measurable function from $\mathcal{S}$ to $[1,\infty)$, we say that $\mathcal{M}$ is $(\bar\kappa, \bar w)$ stable if there exists a $c_{\max} < \infty$ such that

$$\|c(\cdot, a)\|_{\bar w} \le c_{\max}, \quad \forall a \in \mathcal{A} \tag{5}$$

and

$$\int_{\mathcal{S}} \bar w(s') P(ds' \mid s,a) \le \bar\kappa \bar w(s), \quad \forall s \in \mathcal{S}, \forall a \in \mathcal{A}. \tag{6}$$

Remark 3 It is shown in [32] that $(\bar\kappa, \bar w)$ stability of the model is sufficient for DP-solvability. The notion of (κ, w) stability of a policy is weaker. In particular, $(\bar\kappa, \bar w)$ stability of the model implies that any (time-homogeneous) policy is also $(\bar\kappa, \bar w)$ stable. However, (κ, w) stability of a particular policy does not imply (κ, w) stability of the model. For a given weight function, the smallest value of κ that satisfies (4) is given by

$$\kappa_w = \sup_{s\in\mathcal{S}} \frac{\int_{\mathcal{S}} w(s') P_\pi(ds'|s)}{w(s)} \tag{7}$$

while the smallest value of $\bar\kappa$ that satisfies equation (6) is given by

$$\bar\kappa_w = \sup_{s\in\mathcal{S}, a\in\mathcal{A}} \frac{\int_{\mathcal{S}} w(s') P(ds'|s,a)}{w(s)}. \tag{8}$$

It is clear from the definitions that $\kappa_w \le \bar\kappa_w$. We show via an example in Sec. IV-E that using the weaker notion of (κ, w) stability of a policy drastically increases the range of possible choices of the weight function and leads to tighter approximation bounds.

Lemma 1 *Given an MDP $\mathcal{M}$ and a tuple (κ, w), for any policy $\pi \in \Pi_S(\kappa, w)$, we have the following:*

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

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

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	$V^\star$	$\hat{V}^\star$
Optimal policy	$\pi^\star$	$\hat{\pi}^\star$
Bellman operator of policy π	$\mathcal{B}^\pi$	$\hat{\mathcal{B}}^\pi$
Bellman optimality operator	$\mathcal{B}^\star$	$\hat{\mathcal{B}}^\star$
Set of one-step greedy policies w.r.t. v	$\mathcal{G}(v)$	$\hat{\mathcal{G}}(v)$
Set of (κ, w)-stable policies	$\Pi_S(\kappa, w)$	$\hat{\Pi}_S(\kappa, w)$

3) The fixed point equation

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

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

See Appendix A for proof.

III. Problem formulation and approximation bounds

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. As before, we assume that both c and $\hat{c}$ are bounded from below. Thus, there exists a finite constant $c_{\min}$ such that $c(s,a) \geq c_{\min}$ and $\hat{c}(s,a) \geq c_{\min}$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.

We further assume that both models $\mathcal{M}$ and $\hat{\mathcal{M}}$ are well-behaved in the following sense, which we assume to hold in the rest of the paper.

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

We will use the superscript $\hat{}$ (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 $\hat{\mathcal{M}}$. For each start state s, provide a bound for the loss in performance when using $\hat{\pi}^\star$ in the original model $\mathcal{M}$, i.e., bound $V^{\hat{\pi}^\star}(s) - V^\star(s)$.*

B. Modeling Assumptions

In the rest of the paper, we will work with a fixed (κ, w) where κ is a non-negative constant such that $\gamma\kappa < 1$ and $w \colon \mathcal{S} \to [1, \infty)$. Note that $\Pi_S(\kappa, w)$ and $\hat{\Pi}_S(\kappa, w)$ denote the sets of (κ, w)-stable policies for models $\mathcal{M}$ and $\hat{\mathcal{M}}$, respectively. Also, $\mathcal{G}(v)$ and $\hat{\mathcal{G}}(v)$ denote the sets of one-step greedy polices with respect to v for models $\mathcal{M}$ and $\hat{\mathcal{M}}$, respectively. We impose the following additional assumption on the models.

Assumption 2 We assume that

1) The set $\mathcal{G}(V^\star) \cap \Pi_S(\kappa, w)$ is nonempty.
2) The set $\hat{\mathcal{G}}(\hat{V}^\star) \cap \Pi_S(\kappa, w) \cap \hat{\Pi}_S(\kappa, w)$ is nonempty.

When $\mathcal{G}(V^\star) \cap \Pi_S(\kappa, w) \neq \emptyset$, we can show that any policy $\pi \in \mathcal{G}(V^\star) \cap \Pi_S(\kappa, w)$ is optimal, i.e., $V^\pi = V^\star$. From now on, we assume that the optimal policy $\pi^\star$ for model $\mathcal{M}$ belongs to $\mathcal{G}(V^\star) \cap \Pi_S(\kappa, w)$. Similarly, we assume that the optimal policy $\hat{\pi}^\star$ for model $\hat{\mathcal{M}}$ belongs to $\hat{\mathcal{G}}(\hat{V}^\star) \cap \Pi_S(\kappa, w) \cap \hat{\Pi}_S(\kappa, w)$. Since $\pi^\star \in \Pi_S(\kappa, w)$ and $\hat{\pi}^\star \in \hat{\Pi}_S(\kappa, w)$, we have by Lemma 1 that $V^\star, \hat{V}^\star \in \mathcal{V}_w$. The extra intersection in the second part of Assumption 2 ensures that $\hat{\pi}^\star \in \Pi_S(\kappa, w)$ and, therefore, $V^{\hat{\pi}^\star} \in \mathcal{V}_w$.

For some of the results, we impose one of the following assumptions:

Assumption 3 The set $\mathcal{G}(\hat{V}^\star) \cap \Pi_S(\kappa, w)$ is nonempty.

Assumption 4 The set $\hat{\mathcal{G}}(V^\star) \cap \hat{\Pi}_S(\kappa, w)$ is nonempty.

Assumption 3 implies that using $\hat{V}^\star$ as the one-step look ahead value function in the original model produces a stable policy. Similarly, Assumption 4 implies that using $V^\star$ as the one-step look ahead value function in the approximate model produces a stable policy.

Recall that (κ, w)-model stability implies that every time-homogeneous policy is (κ, w) stable. Therefore, if models $\mathcal{M}$ and $\hat{\mathcal{M}}$ are (κ, w)-model stable, then (i) Assumption 2 is equivalent to the non-emptiness of $\mathcal{G}(V^\star)$ and $\hat{\mathcal{G}}(\hat{V}^\star)$; and (ii) Assumption 3 (resp. Assumption 4) is equivalent to non-emptiness of $\mathcal{G}(\hat{V}^\star)$ (resp. $\hat{\mathcal{G}}(V^\star)$).

As the following example illustrates, for specific models it is possible to use the structural properties of value functions and optimal policies to guarantee that Assumptions 2–4 are satisfied without explicitly computing the greedy policies used in these assumptions.

Example 1 (Inventory management) Consider an inventory management problem with state space $\mathcal{S} = \mathbb{R}$ and action space $\mathcal{A} = \mathbb{R}_{\geq 0}$. The dynamics are given by

$$S_{t+1} = S_t + A_t - N_t,$$

where the demand N_t is assumed to be an i.i.d. stochastic process with support $[0, N_{\max}]$. The per-step cost is given as

$$c(s,a) = pa + \bar{c}(s),$$

where $\bar{c}(s) = c_h s \mathbb{1}_{\{s \geq 0\}} - c_s s \mathbb{1}_{\{s < 0\}}$ and p, c_h, c_s are positive constants.

The optimal policy for Example 1 is a base-stock policy [33], [34], i.e, there exists an optimal base-stock level σ such that the optimal policy is $\pi^\star(s) = \max\{0, \sigma - s\}$.

Proposition 1 *In the inventory management model, consider a weight function $w(s) = 1 + \ell \bar{c}(s)$ such that*

$$1 + \ell \max\{c_h, c_s\} N_{\max} < \frac{1}{\gamma}. \tag{9}$$

Pick any $\kappa \in [1 + \ell \max\{c_h, c_s\} N_{\max}, 1/\gamma)$. Then, all base-stock policies with base-stock level $\sigma \in (0, N_{\max}]$ are (κ, w) stable.

See Appendix B for proof.

Now consider two inventory management models $\mathcal{M}$ and $\hat{\mathcal{M}}$ that differ in the demand distribution (with the same support). If the parameters of both models satisfy (9), then, Proposition 1 implies that the set Λ_{base} of all base stock policies with base-stock level in $(0, N_{\max}]$ is a subset of $\Pi_S(\kappa, w)$ and $\hat{\Pi}_S(\kappa, w)$. This fact can be used to obtain simpler sufficient conditions for Assumptions 2–4 by simply replacing the stable policy sets with Λ_{base}. These simpler sufficient conditions can be verified by standard dynamic programming arguments for inventory models [34].

C. Approximation Bounds

Definition 7 (Bellman mismatch functionals) Suppose Assumptions 1 and 2 hold. Define the following functionals:

- For any $\pi \in \Pi_S(\kappa, w)$ and $\hat{\pi} \in \hat{\Pi}_S(\kappa, w)$, define the *Bellman mismatch functional* $\mathcal{D}_w^{\pi,\hat{\pi}} \colon \mathcal{V}_w \to \mathbb{R}_{\geq 0}$ as follows: for any $v \in \mathcal{V}_w$,
$$\mathcal{D}_w^{\pi,\hat{\pi}} v = \|\mathcal{B}^{\pi} v - \hat{\mathcal{B}}^{\hat{\pi}} v\|_w.$$
- For any $\pi \in \Pi_S(\kappa, w) \cap \hat{\Pi}_S(\kappa, w)$, define the *Bellman mismatch functional* $\mathcal{D}_w^{\pi} \colon \mathcal{V}_w \to \mathbb{R}_{\geq 0}$ as follows: for any $v \in \mathcal{V}_w$,
$$\mathcal{D}_w^{\pi} v = \mathcal{D}_w^{\pi,\pi} v = \|\mathcal{B}^{\pi} v - \hat{\mathcal{B}}^{\pi} v\|_w.$$
- Define the *Bellman optimality mismatch functional* $\mathcal{D}_w^{\star} \colon \mathcal{V}_w \to \mathbb{R}_{\geq 0}$ as follows: for any $v \in \mathcal{V}_w$,
$$\mathcal{D}_w^{\star} v = \|\mathcal{B}^{\star} v - \hat{\mathcal{B}}^{\star} v\|_w.$$

In the rest of the paper, we assume that w is fixed. Therefore, we omit the subscript w in the mismatch functionals in the rest of the discussion.

Lemma 2 (Policy error bounds) *For any two policies $\pi \in \Pi_S(\kappa, w)$ and $\hat{\pi} \in \hat{\Pi}_S(\kappa, w)$, we have*

$$\|V^{\pi} - \hat{V}^{\hat{\pi}}\|_w \leq \frac{1}{1-\gamma\kappa} \min\{\mathcal{D}^{\pi,\hat{\pi}} V^{\pi}, \mathcal{D}^{\pi,\hat{\pi}} \hat{V}^{\hat{\pi}}\}. \tag{10}$$

See Appendix C for proof.

Lemma 3 (Value error bounds) *The following hold:*

1) *If Assumptions 1 and 2 hold, we have*
$$\|V^{\star} - \hat{V}^{\star}\|_w \leq \frac{1}{1-\gamma\kappa} \min\{\mathcal{D}^{\pi^\star,\hat{\pi}^\star} V^{\star}, \mathcal{D}^{\pi^\star,\hat{\pi}^\star} \hat{V}^{\star}\}. \tag{11}$$
2) *If Assumptions 1, 2 and 3 hold, we have*
$$\|V^{\star} - \hat{V}^{\star}\|_w \leq \frac{1}{1-\gamma\kappa} \mathcal{D}^{\star} \hat{V}^{\star}. \tag{12}$$
3) *If Assumptions 1, 2 and 4 hold, we have*
$$\|V^{\star} - \hat{V}^{\star}\|_w \leq \frac{1}{1-\gamma\kappa} \mathcal{D}^{\star} V^{\star}. \tag{13}$$

See Appendix D for proof.

We can establish the following theorem by combining policy and value error bounds.

Theorem 1 *We have the following bounds on $V^{\hat{\pi}^\star} - V^{\star}$:*

1) *Under Assumptions 1 and 2, we have*
$$\left\|V^{\hat{\pi}^\star} - V^{\star}\right\|_w \leq \frac{1}{1-\gamma\kappa} \left[\mathcal{D}^{\hat{\pi}^\star} \hat{V}^{\star} + \mathcal{D}^{\pi^\star,\hat{\pi}^\star} \hat{V}^{\star}\right]$$
and
$$\left\|V^{\hat{\pi}^\star} - V^{\star}\right\|_w \leq \frac{1}{1-\gamma\kappa} \mathcal{D}^{\hat{\pi}^\star} V^{\star} + \frac{(1+\gamma\kappa)}{(1-\gamma\kappa)^2} \mathcal{D}^{\pi^\star,\hat{\pi}^\star} V^{\star}.$$
2) *Under Assumptions 1, 2, and 3, we have*
$$\left\|V^{\hat{\pi}^\star} - V^{\star}\right\|_w \leq \frac{1}{1-\gamma\kappa} \left[\mathcal{D}^{\hat{\pi}^\star} \hat{V}^{\star} + \mathcal{D}^{\star} \hat{V}^{\star}\right].$$
3) *Under Assumptions 1, 2, and 4, we have*
$$\left\|V^{\hat{\pi}^\star} - V^{\star}\right\|_w \leq \frac{1}{1-\gamma\kappa} \mathcal{D}^{\hat{\pi}^\star} V^{\star} + \frac{(1+\gamma\kappa)}{(1-\gamma\kappa)^2} \mathcal{D}^{\star} V^{\star}.$$

See Appendix E for proof.

Remark 4 Since $V^{\hat{\pi}^\star}(s) \geq V^{\star}(s)$, we have
$$V^{\hat{\pi}^\star}(s) - V^{\star}(s) \leq \left\|V^{\hat{\pi}^\star} - V^{\star}\right\|_w w(s). \tag{14}$$
Thus, the bounds on $\left\|V^{\hat{\pi}^\star} - V^{\star}\right\|_w$ stated in Theorem 1 provide a bound on the performance loss when $\hat{\pi}^\star$ is used in the original model $\mathcal{M}$ with a start state s.

D. Discussion

Obtaining a solution of Problem 1 requires some knowledge of the model. If we were to obtain an exact expression for $\|V^{\star} - V^{\hat{\pi}^\star}\|_w$, we would need to compute $V^{\star}$ and $V^{\hat{\pi}^\star}$, which are the fixed points of $\mathcal{B}^{\star}$ and $\mathcal{B}^{\hat{\pi}^\star}$, respectively. Computing $V^{\star}$ and $V^{\hat{\pi}^\star}$ requires starting with an initial choice V_0 and then iteratively computing $\{(\mathcal{B}^{\star})^n V_0\}_{n\geq 1}$ and $\{(\mathcal{B}^{\hat{\pi}^\star})^n V_0\}_{n\geq 1}$ until convergence. In contrast, the upper bounds of Theorem 1, part 2, are in terms of the mismatch Bellman operators, which require *one* update of the Bellman operators $\mathcal{B}^{\star}$ and $\mathcal{B}^{\hat{\pi}^\star}$. It is worth highlighting that we do not need to compute $V^{\star}$ or $\pi^{\star}$ in order to use the bounds of Theorem 1, part 2. Thus, our upper bounds provide significant computational savings, especially when computing a Bellman update in the original model is computationally expensive.

Another feature of our results is that they characterize the sensitivity of the optimal performance to model approximation. To make this notion precise, we need to define a notion of *distance* between the original and approximate model. We elaborate on this direction in Sec. V. We first present a few generalizations of the bounds.

E. Bounds under stability of deterministic open loop policies

Let π_a denote the deterministic open loop policy that selects action a with probability 1 in all states, i.e., $\pi_a(s) = a$ for all s. In this section, we assume that all such policies are (κ, w) stable in $\mathcal{M}$ and $\hat{\mathcal{M}}$, and simplify the bounds of Theorem 1.

Assumption 5 For each $a \in \mathcal{A}$, $\pi_a \in \Pi_S(\kappa, w) \cap \hat{\Pi}_S(\kappa, w)$. Moreover $\mathcal{G}(\hat{V}^{\star})$ and $\hat{\mathcal{G}}(V^{\star})$ are nonempty.

Assumption 5 is weaker than (κ, w) stability of the model because Assumption 5 does not imply a *uniform* upper bound on $\|c(\cdot, a)\|_w$ over all $a \in \mathcal{A}$.

Lemma 4 *Assumptions 1 and 5 imply Assumptions 3 and 4.*

See Appendix F for proof.

Definition 8 Suppose Assumption 5 holds. Define the *Bellman maximum mismatch functional* $\mathcal{D}_w^{\max}\colon \mathcal{V}_w \to \mathbb{R}_{\geq 0}$ as follows: for any $v \in \mathcal{V}_w$,

$$\mathcal{D}_w^{\max} v = \sup_{a \in \mathcal{A}} \mathcal{D}_w^{\pi_a} v = \sup_{a \in \mathcal{A}} \|\mathcal{B}^{\pi_a} v - \hat{\mathcal{B}}^{\pi_a} v\|_w.$$

In the sequel, we omit the subscript w from the functional defined above for simplicity.

Lemma 5 *Under Assumptions 1, 2 and 5, the Bellman mismatch functionals satisfy the following for any $v \in \mathcal{V}_w$:*

$$\sup_{\pi \in \Pi_S(\kappa,w) \cap \hat{\Pi}_S(\kappa,w)} \mathcal{D}^{\pi} v = \mathcal{D}^{\max} v \quad \text{and} \quad \mathcal{D}^{\star} v \leq \mathcal{D}^{\max} v. \tag{15}$$

See Appendix G for proof.

Theorem 2 *Under Assumptions 1, 2 and 5, we have the following two bounds on $V^{\hat{\pi}^\star} - V^\star$:*

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

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

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

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

PROOF The result follows from Theorem 1 (parts 2 and 3), Lemma 4, and Lemma 5. ∎

F. Generalized bounds based on affine transformations of the cost

Given an MDP $\mathcal{M}$ and a tuple $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ of real numbers where $\alpha_1 > 0$, define a new MDP $\mathcal{M}_{\boldsymbol{\alpha}}$ with the same dynamics as $\mathcal{M}$ but with the cost function modified to $\alpha_1 c(s,a) + \alpha_2$. For any policy π, let $V_{\boldsymbol{\alpha}}^\pi$ denote the value function of π in $\mathcal{M}_{\boldsymbol{\alpha}}$. Similarly, let $V_{\boldsymbol{\alpha}}^\star$ denote the optimal value function for $\mathcal{M}_{\boldsymbol{\alpha}}$.

Lemma 6 *The following properties hold for any $s \in \mathcal{S}$:*

1) *For any policy π, $V_{\boldsymbol{\alpha}}^\pi(s) = \alpha_1 V^\pi(s) + \alpha_2/(1-\gamma)$.*
2) *If $\pi^\star$ is optimal for $\mathcal{M}$, then it is also optimal for $\mathcal{M}_{\boldsymbol{\alpha}}$ and $V_{\boldsymbol{\alpha}}^\star(s) = \alpha_1 V^\star(s) + \alpha_2/(1-\gamma)$.*
3) *For any policy π and weight function $w\colon \mathcal{S} \to [1,\infty)$, $\|V_{\boldsymbol{\alpha}}^\pi - V_{\boldsymbol{\alpha}}^\star\|_w = \alpha_1 \|V^\pi - V^\star\|_w$.*

PROOF Properties 1 and 2 are immediate consequences of the definitions. Property 3 follows from properties 1 and 2. ∎

Lemma 6 provides an alternative way of bounding the performance loss when the optimal policy $\hat{\pi}^\star$ for the approximate model $\hat{\mathcal{M}}$ is used in the true model $\mathcal{M}$. We can first view $\hat{\mathcal{M}}$ as an approximation for $\mathcal{M}_{\boldsymbol{\alpha}}$ and bound the approximation error $\|V_{\boldsymbol{\alpha}}^{\hat{\pi}^\star} - V_{\boldsymbol{\alpha}}^\star\|_w$ in $\mathcal{M}_{\boldsymbol{\alpha}}$. Part 3 of Lemma 6 implies that the approximation error in $\mathcal{M}$ is simply $1/\alpha_1$ times the approximation error in $\mathcal{M}_{\boldsymbol{\alpha}}$.

To bound the approximation error in $\mathcal{M}_{\boldsymbol{\alpha}}$, let $\mathcal{B}_{\boldsymbol{\alpha}}^\pi$ and $\mathcal{B}_{\boldsymbol{\alpha}}^\star$ denote the Bellman operator and Bellman optimality operator for $\mathcal{M}_{\boldsymbol{\alpha}}$. Let $\mathcal{G}_{\boldsymbol{\alpha}}(v)$ denote the set of one-step greedy policies with respect to v in model $\mathcal{M}_{\boldsymbol{\alpha}}$.

Note that for any (α_1, α_2) with $\alpha_1 > 0$, we have that: (i) If $\mathcal{M}$ is DP-solvable, then so is $\mathcal{M}_{\boldsymbol{\alpha}}$; (ii) the set of (κ, w)-stable policies is the same for $\mathcal{M}$ and $\mathcal{M}_{\boldsymbol{\alpha}}$. Consequently, if any of Assumptions 1, 2 or 5 holds for $\mathcal{M}$ and $\hat{\mathcal{M}}$, then it also holds for $\mathcal{M}_{\boldsymbol{\alpha}}$ and $\hat{\mathcal{M}}$. Instead of Assumptions 3 and 4, we need the following alternative assumptions.

Assumption 6 The set $\mathcal{G}_{\boldsymbol{\alpha}}(\hat{V}^\star) \cap \Pi_S(\kappa, w)$ is nonempty.

Assumption 7 The set $\hat{\mathcal{G}}(V_{\boldsymbol{\alpha}}^\star) \cap \hat{\Pi}_S(\kappa, w)$ is nonempty.

We can now define mismatch functionals (analogous to those defined in Definitions 7 and 8) using $\mathcal{M}_{\boldsymbol{\alpha}}$ and $\hat{\mathcal{M}}$.

Definition 9 Suppose Assumptions 1 and 2 hold. Define the following functionals:

- For any $\pi \in \Pi_S(\kappa, w)$ and $\hat{\pi} \in \hat{\Pi}_S(\kappa, w)$, define the *Bellman mismatch functional* $\mathcal{D}_{\boldsymbol{\alpha}}^{\pi,\hat{\pi}}\colon \mathcal{V}_w \to \mathbb{R}_{\geq 0}$ as follows: for any $v \in \mathcal{V}_w$,
$$\mathcal{D}_{\boldsymbol{\alpha}}^{\pi,\hat{\pi}} v = \|\mathcal{B}_{\boldsymbol{\alpha}}^\pi v - \hat{\mathcal{B}}^{\hat{\pi}} v\|_w.$$
- For any $\pi \in \Pi_S(\kappa, w) \cap \hat{\Pi}_S(\kappa, w)$, define the *Bellman mismatch functional* $\mathcal{D}_{\boldsymbol{\alpha}}^{\pi}\colon \mathcal{V}_w \to \mathbb{R}_{\geq 0}$ as follows: for any $v \in \mathcal{V}_w$
$$\mathcal{D}_{\boldsymbol{\alpha}}^{\pi} v = \mathcal{D}_{\boldsymbol{\alpha}}^{\pi,\pi} v = \|\mathcal{B}_{\boldsymbol{\alpha}}^\pi v - \hat{\mathcal{B}}^{\pi} v\|_w.$$
- Define the *Bellman optimality mismatch functional* $\mathcal{D}_{\boldsymbol{\alpha}}^{\star}\colon \mathcal{V}_w \to \mathbb{R}_{\geq 0}$ as follows: for any $v \in \mathcal{V}_w$,
$$\mathcal{D}_{\boldsymbol{\alpha}}^{\star} v = \|\mathcal{B}_{\boldsymbol{\alpha}}^\star v - \hat{\mathcal{B}}^{\star} v\|_w.$$

Definition 10 Suppose Assumption 5 holds. Define the *Bellman maximum mismatch functional* $\mathcal{D}_{\boldsymbol{\alpha}}^{\max}\colon \mathcal{V}_w \to \mathbb{R}_{\geq 0}$ as follows: for any $v \in \mathcal{V}_w$,

$$\mathcal{D}_{\boldsymbol{\alpha}}^{\max} v = \sup_{a \in \mathcal{A}} \mathcal{D}_{\boldsymbol{\alpha}}^{\pi_a} v = \sup_{a \in \mathcal{A}} \|\mathcal{B}_{\boldsymbol{\alpha}}^{\pi_a} v - \hat{\mathcal{B}}^{\pi_a} v\|_w.$$

We can now use Lemma 6 to present variants of Theorem 1 and Theorem 2.

Theorem 3 *For any $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ with $\alpha_1 > 0$, we have the following bounds on $V^{\hat{\pi}^\star} - V^\star$:*

1) *Under Assumptions 1 and 2, we have*

$$\left\|V^{\hat{\pi}^\star} - V^\star\right\|_w \leq \frac{1}{\alpha_1(1-\gamma\kappa)} \left[\mathcal{D}_{\boldsymbol{\alpha}}^{\hat{\pi}^\star} \hat{V}^\star + \mathcal{D}_{\boldsymbol{\alpha}}^{\pi^\star,\hat{\pi}^\star} \hat{V}^\star\right]$$

and

$$\left\|V^{\hat{\pi}^\star} - V^\star\right\|_w \leq \frac{1}{\alpha_1(1-\gamma\kappa)} \mathcal{D}_{\boldsymbol{\alpha}}^{\hat{\pi}^\star}(\alpha_1 V^\star) + \frac{(1+\gamma\kappa)}{\alpha_1(1-\gamma\kappa)^2} \mathcal{D}_{\boldsymbol{\alpha}}^{\pi^\star,\hat{\pi}^\star}(\alpha_1 V^\star).$$

2) *Under Assumptions 1, 2 and 6, we have*

$$\left\|V^{\hat{\pi}^\star} - V^\star\right\|_w \leq \frac{1}{\alpha_1(1-\gamma\kappa)} \left[\mathcal{D}_{\boldsymbol{\alpha}}^{\hat{\pi}^\star} \hat{V}^\star + \mathcal{D}_{\boldsymbol{\alpha}}^{\star} \hat{V}^\star\right].$$

3) *Under Assumptions 1, 2 and 7, we have*

$$\left\|V^{\hat{\pi}^\star} - V^\star\right\|_w \leq \frac{1}{\alpha_1(1-\gamma\kappa)} \mathcal{D}_{\boldsymbol{\alpha}}^{\hat{\pi}^\star}(\alpha_1 V^\star) + \frac{(1+\gamma\kappa)}{\alpha_1(1-\gamma\kappa)^2} \mathcal{D}_{\boldsymbol{\alpha}}^{\star}(\alpha_1 V^\star).$$

See Appendix H for the proof.

Theorem 4 *Under Assumptions 1, 2 and 5, we have the following two bounds on* $V^{\hat{\pi}^\star} - V^\star$*:*

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

$$\left\|V^{\hat{\pi}^\star} - V^\star\right\|_w \le \frac{2}{\alpha_1(1-\gamma\kappa)} \mathcal{D}^{\max}_{\boldsymbol{\alpha}} \hat{V}^\star.$$

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

$$\left\|V^{\hat{\pi}^\star} - V^\star\right\|_w \le \frac{2}{\alpha_1(1-\gamma\kappa)^2} \mathcal{D}^{\max}_{\boldsymbol{\alpha}}(\alpha_1 V^\star).$$

PROOF The result follows from Theorem 3 (parts 2 and 3), Lemma 5 and the fact that Assumption 5 implies Assumptions 6 and 7 (using the same argument as Lemma 4). ∎

Some remarks:

- It is possible to optimize the bounds in Theorems 3 and 4 by optimizing over the choice of $\boldsymbol{\alpha}$. For Theorem 3, parts 2 and 3, we need to ensure that the choice of α satisfies Assumption 6 or 7, as appropriate.
- For $(\alpha_1, \alpha_2) = (1, 0)$, $\mathcal{M}_{\boldsymbol{\alpha}} = \mathcal{M}$ and hence, the bounds in Theorems 3 and 4 are identical to those in Theorems 1 and 2, respectively.
- Thus, if we optimize over appropriate (α_1, α_2), then the bounds of Theorems 3 and 4 are tighter than those of Theorems 1 and 2. For instance, if $\hat{\mathcal{M}} = \mathcal{M}_{(2,1)}$, then the bounds of Theorems 3 and 4 are zero for $\boldsymbol{\alpha} = (2, 1)$, while the bounds of Theorems 1 and 2 may be positive.

IV. Some instances of the main results

A. Inventory management

In this section, we illustrate the results of Theorem 1 for the inventory management problem described in Example 1 in Sec. III-B, with state space $\mathcal{S} = \{-S_{\max}, -S_{\max}+1, \dots S_{\max}\}$, action space $\mathcal{A} = \{0, 1, \dots, S_{\max}\}$, and dynamics

$$S_{t+1} = \left[S_t + A_t - N_t\right]^{S_{\max}}_{-S_{\max}}$$

where $[\cdot]^{S_{\max}}_{-S_{\max}}$ denotes a function which clips its value between $-S_{\max}$ and $S_{\max}$. For this example, we assume the demand N_t is an i.i.d. Binomial(n, q) process. We denote such a model by $\mathcal{M} = (S_{\max}, \gamma, n, q, c_h, c_s, p)$.

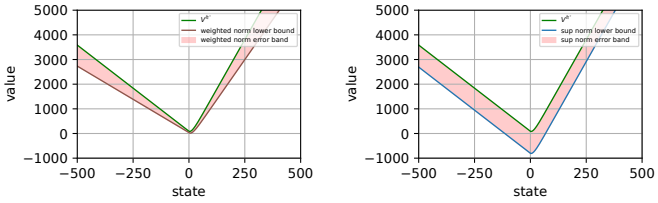
We consider two models:

- 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 (with any choice of weight function). We choose the weight function to have a similar shape as the per-step cost. In particular, we take $w(s) = 1 + (1.5 \cdot 10^{-2})\left[\hat{c}_h s \mathbb{1}_{\{s \ge 0\}} - \hat{c}_s s \mathbb{1}_{\{s<0\}}\right]$, where $\hat{c}_h$ and $\hat{c}_s$ denote the per-unit holding and shortage costs of the approximate model, respectively. We verify that Assumptions 2 and 3 are satisfied with $\kappa = 1.07$.

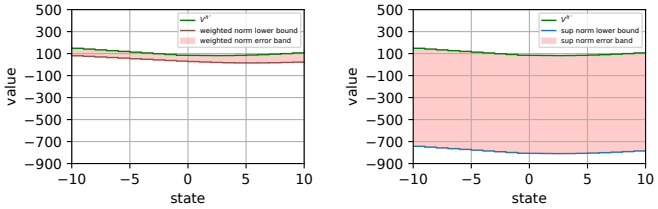
The weighted-norm bound of Theorem 1, part 2 implies that

$$V^{\hat{\pi}^\star}(s) - \frac{1}{1-\gamma\kappa}\left[\mathcal{D}^{\hat{\pi}^\star}\hat{V}^\star + \mathcal{D}^\star \hat{V}^\star\right] w(s) \le V^\star(s) \le V^{\hat{\pi}^\star}(s). \tag{16}$$



(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 these bounds with the sup-norm bounds obtained by taking $w \equiv 1$.

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 upper bound $V^{\hat{\pi}^\star}$, and the weighted- and sup-norm lower bounds on $V^\star(s)$ given by the left hand side of (16) in Fig. 1.

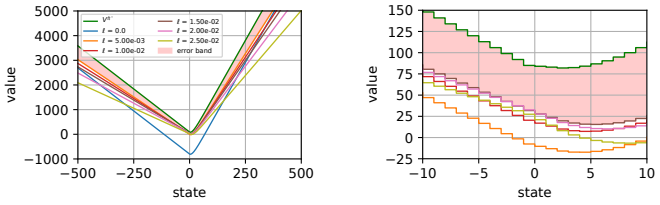
Fig. 1 shows that the weighted-norm bound is slightly tighter than the sup-norm bound for most states. To better compare the error bounds, we zoom into the region of $\bar{\mathcal{S}} := \{-10, -9, \dots, 10\}$ in Fig. 2, where the weighted-norm bound is significantly better than the sup-norm bound.

The optimal policy for the inventory management model described above is a base-stock policy [33]: $\pi^\star(s) = \max(0, \sigma - s)$, where there is an optimal base-stock level σ and whenever the inventory is less than σ, the optimal action is to order goods so that the inventory becomes σ. For the model $\hat{\mathcal{M}}$, the base-stock level $\sigma = 2$. 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

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 (14) as follows:

$$V^{\hat{\pi}^\star}(s) - \inf_{w \in \mathcal{W}} \left\{\left\|V^{\hat{\pi}^\star} - V^\star\right\|_w w(s)\right\} \le V^\star(s) \le V^{\hat{\pi}^\star}(s). \tag{17}$$



(a) entire state space (b) zoomed in

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

Note that the choice of weight function that gives the tightest bound can vary with the start state s. We illustrate the benefit of such a state dependent choice of weight function for the inventory management model of the previous section.

We consider the following family of weight functions for the inventory management model:

$$\mathcal{W} = \{1 + \ell \bar{c}(s) : \ell \in \{0, 0.5 \cdot 10^{-2}, 10^{-2}, \dots, 2.5 \cdot 10^{-2}\}\}, \tag{18}$$

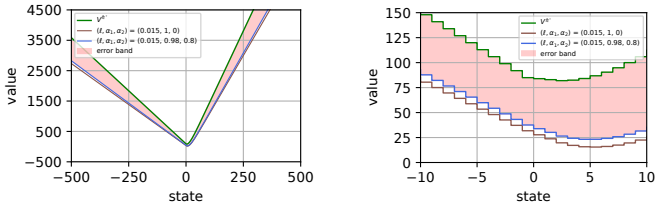
where $\bar{c}(s) = \hat{c}_h s \mathbb{1}_{\{s \geq 0\}} - \hat{c}_s s \mathbb{1}_{\{s<0\}}$. Note that for $\ell = 0$, $w(s) = 1$ which corresponds to the sup-norm. For each $w \in \mathcal{W}$, we compute the smallest κ_w such that Assumption 2 is satisfied as per (7) and further verify that this value satisfies Assumption 3. We plot the lower bounds on $V^\star(s)$ corresponding to each $w \in \mathcal{W}$ in Fig. 3. As can be seen from the figure, the best choice of weight function depends on the state. Minimizing over all $w \in \mathcal{W}$ as per (17) gives a tighter bound. This tighter lower bound is highlighted in Fig. 3 using the shaded area shown in red.

C. Generalized bounds based on cost transformation

The generalized bounds of Theorem 3, part 2, imply that

$$V^{\hat{\pi}^\star}(s) - \frac{1}{\alpha_1(1-\gamma\kappa)}\left[\mathcal{D}_{\boldsymbol{\alpha}}^{\hat{\pi}^\star}\hat{V}^\star + \mathcal{D}_{\boldsymbol{\alpha}}^{\star}\hat{V}^\star\right] w(s) \leq V^\star(s) \leq V^{\hat{\pi}^\star}(s) \tag{19}$$

To show that this bound can be better than that of (16) obtained from Theorem 1, part 2, we consider the setup of Sec. IV-B with $\ell = 1.5 \times 10^{-2}$ and compare $\boldsymbol{\alpha} = (0.98, 0.8)$ with $\boldsymbol{\alpha} = (1, 0)$. We verify that the appropriate assumptions are satisfied and plot the two bounds in Fig. 4. As can be seen from the plots, the bound corresponding to Theorem 3 is tighter.



(a) entire state space (b) zoomed in

Fig. 4: Lower bounds obtained by different choices of $\boldsymbol{\alpha}$.

D. 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 finite covariance matrix 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, respectively, 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

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

where P is the unique positive semidefinite solution of the discounted Riccati equation

$$P = Q + \gamma A^\intercal P A - \gamma^2 A^\intercal P B (R + \gamma B^\intercal P B)^{-1} B^\intercal P A, \tag{20}$$

and $q = \gamma \operatorname{Tr}(\Sigma_W P)/(1-\gamma)$. Furthermore, the optimal policy is given as $\pi^\star(s) = -K^\star s$ where $K^\star = \gamma(R + \gamma B^\intercal P B)^{-1} B^\intercal P A$ is the optimal gain matrix [34].

We consider two models, a true model $\mathcal{M} = (A, B, Q, R, \Sigma_W, \gamma)$ and an approximate model $\hat{\mathcal{M}} = (\hat{A}, \hat{B}, \hat{Q}, \hat{R}, \hat{\Sigma}_W, \gamma)$. We take the weight function to be $w(s) = 1 + \ell s^\intercal s$, where $\ell > 0$ is a parameter. Under standard conditions of stabilizability and detectability (see [34]), both models $\mathcal{M}$ and $\hat{\mathcal{M}}$ satisfy Assumption 1. Let P and $\hat{P}$ denote the solutions of the Riccati equations corresponding to models $\mathcal{M}$ and $\hat{\mathcal{M}}$, and let $\pi^\star(s) = -K^\star s$ and $\hat{\pi}^\star(s) = -\hat{K}^\star s$ denote the optimal policies of models $\mathcal{M}$ and $\hat{\mathcal{M}}$.

For any linear policy $\pi(s) = -Ks$, we use the notation $A_K = A - BK$ and $\hat{A}_K = \hat{A} - \hat{B}K$. We further use $K_{\mu^\star}$ to denote the gain matrix of the (unique) policy $\mu^\star \in \mathcal{G}(\hat{V}^\star)$. We impose the following assumption.

Assumption 8 The models $\mathcal{M}$ and $\hat{\mathcal{M}}$ are such that

$$b_\Sigma := \max\{1 + \ell \operatorname{Tr}(\Sigma_W), 1 + \ell \operatorname{Tr}(\hat{\Sigma}_W)\} \leq \frac{1}{\gamma}$$

and

$$b_\sigma := \max\{\sigma_1^2(A_{K^\star}), \sigma_1^2(A_{\hat{K}^\star}), \sigma_1^2(\hat{A}_{\hat{K}^\star}), \sigma_1^2(A_{K_{\mu^\star}})\} \leq \frac{1}{\gamma}$$

where $\sigma_1(A)$ is the operator norm of A (i.e., the largest singular value of A).

Lemma 7 *Assumption 8 implies Assumptions 2 and 3.*

PROOF Fix a policy $\pi(s) = Ks$. Eq. (3) is always satisfied because

$$\|c_\pi\|_w = \sup_{s\in\mathcal{S}} \frac{s^\intercal(Q + K^\intercal R K)s}{1 + \ell s^\intercal s} \leq \frac{1}{\ell}\rho(Q + K^\intercal R K) < \infty$$

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

Moreover largest value of κ for which (4) is satisfied is given by (7), which simplifies to

$$\begin{aligned}\kappa_w &= \sup_{s\in\mathcal{S}} \frac{\mathbb{E}[w(s_{t+1})|s_t = s]}{w(s)} \\ &= \sup_{s\in\mathcal{S}} \frac{1 + \ell \operatorname{Tr}(\Sigma_W) + \ell s^\intercal A_K^\intercal A_K s}{1 + \ell s^\intercal s} \\ &\le \max(1 + \ell \operatorname{Tr}(\Sigma_W), \sigma_1^2(A_K)).\end{aligned}$$

Thus, if Assumption 8 holds, then Assumptions 2 and 3 hold with $\kappa := \max\{b_\Sigma, b_\sigma\}$. ∎

Then, the result of Theorem 3, part 2 simplifies as follows:

Proposition 2 *Under Assumptions 1 and 8, we have for $\alpha_1 = 1$ and any α_2,*

$$\begin{aligned}\left\|V^{\hat\pi^\star} - V^\star\right\|_w \le &\frac{1}{1-\gamma\kappa}\big[\max\{\rho(D^\star)/\ell, |d_\Sigma + \alpha_2|\} \\ &+ \max\{\rho(D^{\hat\pi^\star})/\ell, |d_\Sigma + \alpha_2|\}\big], \quad (21)\end{aligned}$$

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

$$\begin{aligned}D^\star &= \big(Q + \gamma A^\intercal \hat P A - \gamma^2 A^\intercal \hat P B (R + \gamma B^\intercal \hat P B)^{-1} B^\intercal \hat P A\big) \\ &\quad - \hat P \quad (22)\\ D^{\hat\pi^\star} &= \big(Q + (\hat K^\star)^\intercal R \hat K^\star + \gamma A_{\hat K^\star}^\intercal \hat P A_{\hat K^\star}\big) - \hat P \quad (23)\\ \hat K^\star &= \gamma(\hat R + \gamma \hat B^\intercal \hat P \hat B)^{-1} \hat B^\intercal \hat P \hat A, \quad (24)\end{aligned}$$

and

$$d_\Sigma = \gamma \operatorname{Tr}((\Sigma_W - \hat\Sigma_W)\hat P). \quad (25)$$

By taking $\alpha_2 = -d_\Sigma$ we obtain

$$\left\|V^{\hat\pi^\star} - V^\star\right\|_w \le \frac{1}{\ell(1-\gamma\kappa)}\big[\rho(D^\star) + \rho(D^{\hat\pi^\star})\big]. \quad (26)$$

See Appendix I for proof.

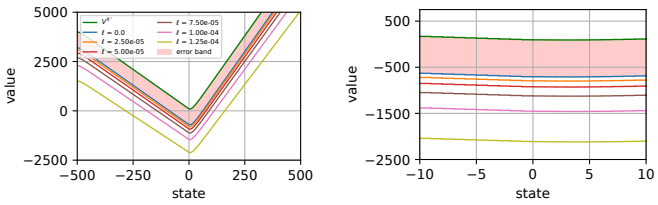
Remark 5 Under Assumptions 1 and 8, the bound obtained in (26) does not depend on the Riccati solution (P, K) of the true model $\mathcal{M}$.

Remark 6 Consider the case when $\hat{\mathcal{M}}$ is the same model as $\mathcal{M}$ except $\hat\Sigma_W = 0$. In this case, $D^\star = 0$ and $D^{\hat\pi^\star} = 0$. Therefore, from (26), we get that $\|V^{\hat\pi^\star} - V^\star\|_w = 0$, which corresponds to the classical certainty equivalence principle of LQR control.

E. Advantage of using (κ, w) stability of policy over (κ, w) stability of model

As mentioned in Remark 3, it is typically assumed in the literature that the model is $(\bar\kappa, \bar w)$ stable, while we impose a weaker assumption that certain policies are (κ, w) stable. In this section, we illustrate two advantages of imposing the weaker assumption.

First, when the per-step cost is unbounded in the actions, as is the case for the LQR problem considered in Sec. IV-D, the model $\mathcal{M}$ is not $(\bar\kappa, \bar w)$ stable for any choice of weight function $\bar w$. However, as illustrated in Sec. IV-D, specific policies may be (κ, w) stable for $w(s) = 1 + s^\intercal s$. Thus, imposing a weaker assumption of stability allows us to derive approximation bounds for a larger class of models.



(a) entire state space (b) zoomed in

Fig. 5: Lower bounds obtained using stability of the model. Note that the curves corresponding to $\ell = 1.50 \cdot 10^{-4}$ and $\ell = 1.75 \cdot 10^{-4}$ are not visible in both plots.

Second, imposing a weak assumption of stability allows us to derive tighter approximation bounds. To illustrate this, we reconsider the inventory management problem in the setting of Sec. IV-B. In Sec. IV-B, we had computed the lower bounds of (16) when using (κ, w) that satisfy Definition 5. Now, we consider $(\bar\kappa, \bar w)$ that satisfy Definition 6 instead. In particular, consider a family of weight functions $\mathcal{W}$ as defined in (18). For each $\bar w \in \mathcal{W}$, we compute the smallest $\bar\kappa_{\bar w}$ such that (6) is satisfied as per (8). The largest value of ℓ for which $\bar\kappa_{\bar w} < 1/\gamma$ is $\ell = 1.75 \cdot 10^{-4}$.

We plot the corresponding lower bound given in (16) in Fig. 5. 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 lower 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 imposing the weaker assumption of (κ, w)-stability of policy rather than the $(\bar\kappa, \bar w)$-stability of the model.

V. Integral probability metrics (IPM) and bounds based on distance between models

In this section, we provide upper bounds for the results of Sec. III that can be computed in terms of the *distance* between models $\mathcal{M}$ and $\hat{\mathcal{M}}$. To define such a distance, we first start with the definition of integral probability metrics (IPM) [10].

A. Integral probability metrics (IPM)

Definition 11 Let $\mathfrak{F}$ be a convex and balanced subset of $\mathcal{V}_w$. Then, the IPM distance (w.r.t. $\mathfrak{F}$) between two probability laws $\nu_1, \nu_2 \in \mathcal{P}_w$ is given by[1]

$$d_{\mathfrak{F}}(\nu_1, \nu_2) = \sup_{f\in\mathfrak{F}} \left| \int f d\nu_1 - \int f d\nu_2 \right|.$$

Definition 12 In the setting of Definition 11, the Minkowski functional of any measurable function $f \in \mathcal{V}_w$ is defined as

$$\rho_{\mathfrak{F}}(f) = \inf\left\{\rho \in \mathbb{R}_{>0} : \frac{f}{\rho} \in \mathfrak{F}\right\}.$$

Note that if for every positive ρ, $f/\rho \notin \mathfrak{F}$, then $\rho_{\mathfrak{F}}(f) = \infty$.

[1]Since $\nu_i \in \mathcal{P}_w$, $i \in \{1, 2\}$, we have $\int f d\nu_i < \infty$ for any $f \in \mathcal{V}_w$.

An immediate consequence of the above two definitions is that for any measurable function $f \in \mathcal{V}_w$,

$$\left| \int f d\nu_1 - \int f d\nu_2 \right| \le \rho_{\mathfrak{F}}(f) d_{\mathfrak{F}}(\nu_1, \nu_2). \tag{27}$$

Many of the commonly used metrics on probability spaces are IPMs. For example

- **Weighted total variation distance**, denoted by $d_{\mathrm{TV},w}$, corresponds to $\mathfrak{F} = \mathfrak{F}_{\mathrm{TV},w} := \{f \in \mathcal{V}_w : \mathrm{osc}_w(f) \le 1\}$, where $\mathrm{osc}_w(f) = \sup_{s,s' \in \mathcal{S}\times\mathcal{S}} |f(s) - f(s')|/(w(s) + w(s'))$ [35]–[37]. For this case, $\rho_{\mathfrak{F}}(f) = \mathrm{osc}_w(f)$. In the special case when $w(s) = 1$ for all $s \in \mathcal{S}$, the weighted total variation distance reduces to total variation distance [10], [38], which we denote by d_{TV}. In this case, $\mathfrak{F} = \mathfrak{F}_{\mathrm{TV}} := \{f \in \mathcal{V}_{w\equiv 1} : \frac{1}{2}\mathrm{span}(f) \le 1\}$ and $\rho_{\mathfrak{F}}(f) = \frac{1}{2}\mathrm{span}(f)$, where $\mathrm{span}(f) = \sup(f) - \inf(f)$.
- **Wasserstein-1 distance**. Suppose $(\mathcal{S}, d_{\mathcal{S}})$ is a metric space Define $\mathfrak{F}_{\mathrm{Was},w} := \{f \in \mathcal{V}_w : \mathrm{Lip}(f) \le 1\}$ where $\mathrm{Lip}(f)$ denotes the Lipschitz constant of a function f. Then, for $\mathfrak{F} = \mathfrak{F}_{\mathrm{Was},w}$, Eq (27) holds for $\rho_{\mathfrak{F}}(f) = \mathrm{Lip}(f)$. Moreover $d_{\mathfrak{F}}(\nu_1, \nu_2) \le d_{\mathrm{Was}}(\nu_1, \nu_2)$, where d_{Was} is the Wasserstein-1 distance [38], [39].

B. Weighted distance between two MDP models

Note that if a policy π is (κ, w) stable, then Eq. (4) implies that $P_\pi(\cdot|s) \in \mathcal{P}_w$ for every $s \in \mathcal{S}$. Therefore, Assumption 2 implies that for all $s \in \mathcal{S}$, $P_{\pi^\star}(\cdot|s), P_{\hat\pi^\star}(\cdot|s), \hat P_{\hat\pi^\star}(\cdot|s) \in \mathcal{P}_w$ and Assumption 5 implies that for all $(s,a) \in \mathcal{S}\times\mathcal{A}$, we have $P(\cdot|s,a), \hat P(\cdot|s,a) \in \mathcal{P}_w$.

We now define two notions of weighted distance between two MDP models.

Definition 13 (Distance between MDP models) Given two MDP models $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, c, \gamma\rangle$ and $\hat{\mathcal{M}} = \langle \mathcal{S}, \mathcal{A}, \hat P, \hat c, \gamma\rangle$, a weight function $w : \mathcal{S} \to [1, \infty)$ and an IPM $d_{\mathfrak{F}}$ as defined in Definition 11, we define the following

1) **Distance between models for given policies:** Given deterministic policies π for model $\mathcal{M}$ and $\hat\pi$ for model $\hat{\mathcal{M}}$ such that $P_\pi(\cdot|s), \hat P_{\hat\pi}(\cdot|s) \in \mathcal{P}_w$ for all $s \in \mathcal{S}$, define

$$\varepsilon_{\boldsymbol{\alpha}}(\pi, \hat\pi) := \sup_{s\in\mathcal{S}} \frac{\left|\alpha_1 c_\pi(s) + \alpha_2 - \hat c_{\hat\pi}(s)\right|}{w(s)},$$
$$\delta_{\mathfrak{F}}(\pi, \hat\pi) := \sup_{s\in\mathcal{S}} \frac{d_{\mathfrak{F}}\big(P_\pi(\cdot|s), \hat P_{\hat\pi}(\cdot|s)\big)}{w(s)}.$$

2) **Maximal distance between models**: Under Assumption 5, define

$$\varepsilon^{\max}_{\boldsymbol{\alpha}} := \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\left|\alpha_1 c(s,a) + \alpha_2 - \hat c(s,a)\right|}{w(s)},$$
$$\delta^{\max}_{\mathfrak{F}} := \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{d_{\mathfrak{F}}\big(P(\cdot|s,a), \hat P(\cdot|s,a)\big)}{w(s)}.$$

Note that the distances defined above depend on the weight function w, but we don't explicitly capture that dependence in the notation.

C. IPM based approximation bounds

Lemma 8 *We have the following bounds for different mismatch functionals:*

1) If policies π and $\hat\pi$ are such that for all $s \in \mathcal{S}$, $P_\pi(\cdot|s)$, $\hat P_{\hat\pi}(\cdot|s) \in \mathcal{P}_w$, then for all $v \in \mathcal{V}_w$,

$$\mathcal{D}^{\pi,\hat\pi}_{\boldsymbol{\alpha}} v \le \varepsilon_{\boldsymbol{\alpha}}(\pi, \hat\pi) + \gamma\rho_{\mathfrak{F}}(v)\delta_{\mathfrak{F}}(\pi, \hat\pi).$$

2) If Assumption 6 is satisfied,

$$\mathcal{D}^\star_{\boldsymbol{\alpha}} \hat V^\star \le \varepsilon_{\boldsymbol{\alpha}}(\mu^\star, \hat\pi^\star) + \gamma\rho_{\mathfrak{F}}(\hat V^\star)\delta_{\mathfrak{F}}(\mu^\star, \hat\pi^\star),$$

for all $\mu^\star \in \mathcal{G}_{\boldsymbol{\alpha}}(\hat V^\star) \cap \Pi_S(\kappa, w)$ (see Assumption 6).

3) If Assumption 7 is satisfied,

$$\mathcal{D}^\star_{\boldsymbol{\alpha}}(\alpha_1 V^\star) \le \varepsilon_{\boldsymbol{\alpha}}(\pi^\star, \hat\mu^\star) + \alpha_1\gamma\rho_{\mathfrak{F}}(V^\star)\delta_{\mathfrak{F}}(\pi^\star, \hat\mu^\star),$$

for all $\hat\mu^\star \in \hat{\mathcal{G}}(V^\star_{\boldsymbol{\alpha}}) \cap \hat\Pi_S(\kappa, w)$ (see Assumption 7).

4) If Assumption 5 is satisfied, then for all $v \in \mathcal{V}_w$,

$$\mathcal{D}^{\max}_{\boldsymbol{\alpha}} v \le \varepsilon^{\max}_{\boldsymbol{\alpha}} + \gamma\rho_{\mathfrak{F}}(v)\delta^{\max}_{\mathfrak{F}}.$$

See Appendix J for proof.

Substituting the results of Lemma 8 in Theorem 3 and Theorem 4, we obtain the following:

Theorem 5 *We have the following bounds on $V^{\hat\pi^\star} - V^\star$:*

1) Under Assumptions 1 and 2, we have

$$\left\|V^{\hat\pi^\star} - V^\star\right\|_w \le \frac{1}{\alpha_1(1-\gamma\kappa)}\Big[\varepsilon_{\boldsymbol{\alpha}}(\hat\pi^\star, \hat\pi^\star) + \varepsilon_{\boldsymbol{\alpha}}(\pi^\star, \hat\pi^\star) + \gamma\rho_{\mathfrak{F}}(\hat V^\star)\big(\delta_{\mathfrak{F}}(\hat\pi^\star, \hat\pi^\star) + \delta_{\mathfrak{F}}(\pi^\star, \hat\pi^\star)\big)\Big]$$

and

$$\left\|V^{\hat\pi^\star} - V^\star\right\|_w \le \frac{1}{\alpha_1(1-\gamma\kappa)}\big[\varepsilon_{\boldsymbol{\alpha}}(\hat\pi^\star, \hat\pi^\star) + \alpha_1\gamma\rho_{\mathfrak{F}}(V^\star)\delta_{\mathfrak{F}}(\hat\pi^\star, \hat\pi^\star)\big] + \frac{1+\gamma\kappa}{\alpha_1(1-\gamma\kappa)^2}\big[\varepsilon_{\boldsymbol{\alpha}}(\pi^\star, \hat\pi^\star) + \alpha_1\gamma\rho_{\mathfrak{F}}(V^\star)\delta_{\mathfrak{F}}(\pi^\star, \hat\pi^\star)\big].$$

2) Under Assumptions 1, 2 and 6, we have

$$\left\|V^{\hat\pi^\star} - V^\star\right\|_w \le \frac{1}{\alpha_1(1-\gamma\kappa)}\Big[\varepsilon_{\boldsymbol{\alpha}}(\hat\pi^\star, \hat\pi^\star) + \varepsilon_{\boldsymbol{\alpha}}(\mu^\star, \hat\pi^\star) + \gamma\rho_{\mathfrak{F}}(\hat V^\star)\big(\delta_{\mathfrak{F}}(\hat\pi^\star, \hat\pi^\star) + \delta_{\mathfrak{F}}(\mu^\star, \hat\pi^\star)\big)\Big]$$

for all $\mu^\star \in \mathcal{G}_{\boldsymbol{\alpha}}(\hat V^\star) \cap \Pi_S(\kappa, w)$ (see Assumption 6).

3) Under Assumptions 1, 2 and 7, we have

$$\left\|V^{\hat\pi^\star} - V^\star\right\|_w \le \frac{1}{\alpha_1(1-\gamma\kappa)}\big[\varepsilon_{\boldsymbol{\alpha}}(\hat\pi^\star, \hat\pi^\star) + \alpha_1\gamma\rho_{\mathfrak{F}}(V^\star)\delta_{\mathfrak{F}}(\hat\pi^\star, \hat\pi^\star)\big] + \frac{1+\gamma\kappa}{\alpha_1(1-\gamma\kappa)^2}\big[\varepsilon_{\boldsymbol{\alpha}}(\pi^\star, \hat\mu^\star) + \alpha_1\gamma\rho_{\mathfrak{F}}(V^\star)\delta_{\mathfrak{F}}(\pi^\star, \hat\mu^\star)\big].$$

for all $\hat\mu^\star \in \hat{\mathcal{G}}(V^\star_{\boldsymbol{\alpha}}) \cap \hat\Pi_S(\kappa, w)$ (see Assumption 7).

4) Under Assumptions 1, 2 and 5, we have

$$\left\|V^{\hat\pi^\star} - V^\star\right\|_w \le \frac{2}{\alpha_1(1-\gamma\kappa)}\Big[\varepsilon^{\max}_{\boldsymbol{\alpha}} + \gamma\rho_{\mathfrak{F}}(\hat V^\star)\delta^{\max}_{\mathfrak{F}}\Big]$$

and

$$\left\|V^{\hat{\pi}^\star} - V^\star\right\|_w \leq \frac{2}{\alpha_1(1-\gamma\kappa)^2}\left[\varepsilon^{\max}_{\boldsymbol{\alpha}} + \alpha_1\gamma\rho_{\mathfrak{F}}(V^\star)\delta^{\max}_{\mathfrak{F}}\right].$$

The bounds of Theorem 5 may be specialized for specific choices of IPMs. We present these bounds in terms of $(\varepsilon^{\max}_{\boldsymbol{\alpha}}, \delta^{\max}_{\mathfrak{F}})$ and $\hat{V}^\star$. The bounds in terms of $(\varepsilon_{\boldsymbol{\alpha}}(\pi^\star, \hat{\pi}^\star), \delta_{\mathfrak{F}}(\pi^\star, \hat{\pi}^\star))$ etc. and/or $V^\star$ can be expressed in a similar manner.

Corollary 1 *Under Assumptions 1, 2 and 5, we have the following bounds on $V^{\hat{\pi}^\star} - V^\star$:*

1) Bound in terms of total-variation distance:

$$\left\|V^{\hat{\pi}^\star} - V^\star\right\|_w \leq \frac{2}{\alpha_1(1-\gamma\kappa)}\left[\varepsilon^{\max}_{\boldsymbol{\alpha}} + \gamma \sup_{\substack{s\in\mathcal{S}\\ a\in\mathcal{A}}} \frac{d_{\mathrm{TV}}(P(\cdot|s,a), \hat{P}(\cdot|s,a))}{w(s)} \frac{\operatorname{span}(\hat{V}^\star)}{2}\right].$$

2) Bound in terms of Wasserstein-1 distance:

$$\left\|V^{\hat{\pi}^\star} - V^\star\right\|_w \leq \frac{2}{\alpha_1(1-\gamma\kappa)}\left[\varepsilon^{\max}_{\boldsymbol{\alpha}} + \gamma \sup_{\substack{s\in\mathcal{S}\\ a\in\mathcal{A}}} \frac{d_{\mathrm{Was}}(P(\cdot|s,a), \hat{P}(\cdot|s,a))}{w(s)} \operatorname{Lip}(\hat{V}^\star)\right].$$

3) Bound in terms of weighted total variation distance:

$$\left\|V^{\hat{\pi}^\star} - V^\star\right\|_w \leq \frac{2}{\alpha_1(1-\gamma\kappa)}\left[\varepsilon^{\max}_{\boldsymbol{\alpha}} + \gamma \sup_{\substack{s\in\mathcal{S}\\ a\in\mathcal{A}}} \frac{d_{\mathrm{TV},w}(P(\cdot|s,a), \hat{P}(\cdot|s,a))}{w(s)} \operatorname{osc}_w(\hat{V}^\star)\right].$$

Parts 1 and 2 of Corollary 1 may be viewed as weighted generalization of approximation results presented in [10], [28], [30] (some of those results assumed that the approximate model has a smaller state space than the original model).

D. Performance loss in using certainty equivalent control

Certainty equivalence refers to the following design methodology to determine a control policy for a stochastic control problem. Replace the random variables in the stochastic control problem by their (conditional) expectations, solve the resulting deterministic control problem to determine a feedback control policy, and use the resulting *certainty equivalent control policy* in the original stochastic system [40], [41]. It is well known that for systems with linear dynamics and quadratic cost (LQ problems), certainty equivalent control policies are optimal. But this is not the case in general. In this section, we use the results of Theorem 5 to characterize the performance loss when using certainty equivalence for general dynamics with additive noise.

Consider a system with state space $\mathbb{R}^n$, action space $\mathbb{R}^m$, and dynamics

$$S_{t+1} = f(S_t, A_t) + N_t \tag{28}$$

where f is a measurable function and $\{N_t\}_{t\geq 1}$ is a zero-mean i.i.d. noise sequence with control law ν_N. The per-step cost is given by $c(S_t, A_t)$.

Now consider a deterministic model obtained by assuming that the noise sequence in (28) takes its expected value, i.e., the dynamics are

$$S_{t+1} = f(S_t, A_t). \tag{29}$$

The per-step cost is the same as before.

Let $\mathcal{M}$ denote the stochastic model and $\hat{\mathcal{M}}$ denote the deterministic model. Then, the certainty equivalent design is to use the control policy $\hat{\pi}^\star$ in original stochastic model $\mathcal{M}$. Suppose Assumptions 1, 2, and 5 are satisfied for some (κ, w). We use the Wasserstein-1 distance based bounds in Corollary 1 to bound $\|V^{\hat{\pi}^\star} - V^\star\|_w$, where we take $\boldsymbol{\alpha} = (1, 0)$ for simplicity. We assume that there is some norm $\|\cdot\|$ on $\mathbb{R}^n$ and the Wasserstein-1 distance and Lipschitz constant are computed with respect to this norm.

Since the costs are the same for both models, $\varepsilon^{\max}_{\boldsymbol{\alpha}} = 0$. We now characterize $\delta^{\max}$. By Kantorovich-Rubinstein duality [38], Wasserstein-1 distance is equivalent to

$$d_{\mathrm{Was}}(\nu_X, \nu_Y) = \inf_{(\tilde{X},\tilde{Y}):\ \tilde{X}\sim\nu_X, \tilde{Y}\sim\nu_Y} \mathbb{E}[\|\tilde{X} - \tilde{Y}\|]. \tag{30}$$

Due to the additive nature of the dynamics and (30), we have that for a fixed (s, a), the Wasserstein-1 distance between $P(\cdot|s, a)$ and $\hat{P}(\cdot|s, a)$ is equal to $\mathbb{E}[\|N\|]$. Thus,

$$\delta^{\max}_{\mathfrak{F}_{\mathrm{Was}}} = \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\mathbb{E}[\|N\|]}{w(s)} \leq \mathbb{E}[\|N\|]$$

Thus, by Corollary 1, part 2, we get

$$\|V^{\hat{\pi}^\star} - V^\star\|_w \leq \frac{2\gamma}{1-\gamma\kappa}\mathbb{E}[\|N\|] \operatorname{Lip}(\hat{V}^\star). \tag{31}$$

This bound precisely quantifies the engineering intuition that certainty equivalent control laws are good when the noise is "small".

Remark 7 The right hand side of (31) does not depend on the weight function (provided the weight function satisfies Assumption 2). Suppose the per-step cost is such that $c_{\min} \geq 0$ and $w = 1 + V^\star$ satisfies Assumption 2 for some $\kappa < 1/\gamma$. Then, Eq. (31) implies that

$$V^\star(s) \leq V^{\hat{\pi}^\star}(s) \leq \left(1 + \frac{2\gamma}{1-\gamma\kappa}\mathbb{E}[\|N\|] \operatorname{Lip}(\hat{V}^\star)\right)\big(1 + V^\star(s)\big).$$

This inequality may be viewed as a generalization of the approximation bounds of [42] to dynamical systems.

VI. Conclusion

In this paper, we present a series of bounds on the weighted approximation error when using the optimal policy of an approximate model $\hat{\mathcal{M}}$ in the original model $\mathcal{M}$. For each bound, we have two types of bounds: one which depends on the value function $\hat{V}^\star$ of the approximate model $\hat{\mathcal{M}}$ and the other which depends on the value function $V^\star$ of the original model $\mathcal{M}$. The first type of bound is more useful in practice because one would obtain $\hat{V}^\star$ when computing the optimal policy of the approximate model $\hat{\mathcal{M}}$. However, the second type of bound is a theoretical upper bound that may be useful for obtaining bounds for reinforcement learning algorithms, e.g., in obtaining sample complexity bounds.

Our results rely on using an appropriate (κ, w) such that certain policies are (κ, w) stable. The choice of the weight function w impacts the tightness of the bounds. Understanding how to choose weight functions is an interesting research direction.

In this paper, we assumed that the approximate model $\hat{\mathcal{M}}$ was given. However, often the approximate model is a design choice. For example, when solving continuous state models, we may decide how to quantize the state space. The approximation bounds obtained in this paper may be useful in guiding the design of such approximate models. They may also be useful in generalizing the convergence guarantees and regret bounds of reinforcement learning algorithms to models with unbounded per-step cost.

REFERENCES

[1] B. Bozkurt, A. Mahajan, A. Nayyar, and Y. Ouyang, "Weighted norm bounds in mdps with unbounded per-step cost," in *Conference on Decision and Control*. Singapore: IEEE, Dec. 2023.

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

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

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

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

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] N. Saldi, S. Yüksel, and T. Linder, "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.

[13] N. Saldi, T. Linder, and S. Yüksel, *Finite Approximations in discrete-time stochastic control*. Springer, 2018.

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

[19] ——, "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.

[20] 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.

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

[22] 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.

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

[24] E. van der Pol, T. Kipf, F. A. Oliehoek, and M. Welling, "Plannable approximations to MDP homomorphisms: Equivariance under actions," in *AAMAS*, Auckland, New Zealand, May 2020.

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

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

[27] 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.

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

[29] 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.

[30] 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.

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

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

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

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

[35] R. Douc, E. Moulines, P. Priouret, and P. Soulier, *Markov chains*. Springer, 2018.

[36] M. Hairer and J. C. Mattingly, "Yet another look at Harris' ergodic theorem for Markov chains," in *Seminar on Stochastic Analysis, Random Fields and Applications V*. Springer, 2011, pp. 109–117.

[37] S. P. Meyn and R. L. Tweedie, "Stability of Markovian processes I: Criteria for discrete-time chains," *Advances in Applied Probability*, vol. 24, no. 3, pp. 542–574, 1992.

[38] C. Villani *et al.*, *Optimal transport: old and new*. Springer, 2008, vol. 338.

[39] L. N. Vaserstein, "Markov processes over denumerable products of spaces, describing large systems of automata," *Problemy Peredachi Informatsii*, vol. 5, no. 3, pp. 64–72, 1969.

[40] H. A. Simon, "Dynamic programming under uncertainty with a quadratic criterion function," *Econometrica: Journal of the Econometric Society*, pp. 74–81, 1956.

[41] H. Theil, "A note on certainty equivalence in dynamic planning," *Econometrica: Journal of the Econometric Society*, pp. 346–349, 1957.

[42] H. Witsenhausen, "Inequalities for the performance of suboptimal uncertain systems," *Automatica*, vol. 5, no. 4, pp. 507–512, Jul. 1969.

APPENDIX A
PROOF OF LEMMA 1

Proof of part 1: Fix a state $s \in \mathcal{S}$. For a policy $\pi \in \Pi_S(\kappa, w)$ and a value function $v \in \mathcal{V}_w$, we have

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

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

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

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 $\mathcal{B}^\pi : \mathcal{V}_w \to \mathcal{V}_w$ is a contraction. Since $\mathcal{V}_w$ is a complete metric space (under the $\|\cdot\|_w$ norm), it follows from Banach fixed point theorem that $\mathcal{B}^\pi$ has a unique fixed point F in $\mathcal{V}_w$. 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 $V_n^\pi \in \mathcal{V}_w$ for all n. Thus, by Banach fixed point theorem, V_n^π converges to the fixed point F of $\mathcal{B}^\pi$ in the $\|\cdot\|_w$ norm. Since convergence in $\|\cdot\|_w$ norm implies pointwise convergence, we have $F(s) = \lim_{n\to\infty} V_n^\pi(s)$ for all $s \in \mathcal{S}$. Furthermore, since per-step costs are bounded from below, we have that for all $s \in \mathcal{S}$,

$$\begin{aligned}
\lim_{n\to\infty} V_n^\pi(s) &= \lim_{n\to\infty} \mathbb{E}^\pi\left[\sum_{t=1}^{n} \gamma^{t-1} c(S_t, A_t)\right] \\
&= \mathbb{E}^\pi\left[\lim_{n\to\infty} \sum_{t=1}^{n} \gamma^{t-1} c(S_t, A_t)\right] = V^\pi(s)
\end{aligned}$$

where the second equality follows from the monotone convergence theorem when $c_{\min} \geq 0$ (the case of $c_{\min} < 0$ follows from a similar argument by shifting $\{V_n^\pi\}_{n\geq 0}$ to make it non-negative and monotone).

Appendix B
Proof of Proposition 1

Consider a base-stock policy π with threshold $\sigma \in (0, N_{\max}]$. For the cost, simple algebra implies that

$$c_\pi(s) \leq p \max\{\sigma - s, 0\} + \max\{c_h, c_s\}|s|$$

Therefore, $\|c_\pi\|_w < \infty$.

For the dynamics, we can exploit the piecewise linear nature of the weight function to show that

$$\sup_{s\in\mathcal{S}} \frac{\int_{\mathcal{S}} w(s') P\big(ds' | s, \pi(s)\big)}{w(s)} \leq 1 + \ell \max\{c_h, c_s\} N_{\max}.$$

Hence, any policy with base-stock level $\sigma \in (0, N_{\max}]$ is (κ, w) stable for any choice of $\kappa \in [1 + \ell \max\{c_h, c_s\} N_{\max}, 1/\gamma)$.

Appendix C
Proof of Lemma 2

Consider

$$\begin{aligned}
\left\|V^\pi - \hat{V}^{\hat{\pi}}\right\|_w &= \left\|\mathcal{B}^\pi V^\pi - \hat{\mathcal{B}}^{\hat{\pi}} \hat{V}^{\hat{\pi}}\right\|_w \\
&\leq \left\|\mathcal{B}^\pi V^\pi - \hat{\mathcal{B}}^{\hat{\pi}} V^\pi\right\|_w + \left\|\hat{\mathcal{B}}^{\hat{\pi}} V^\pi - \hat{\mathcal{B}}^{\hat{\pi}} \hat{V}^{\hat{\pi}}\right\|_w \\
&\leq \mathcal{D}^{\pi,\hat{\pi}} V^\pi + \gamma\kappa \left\|V^\pi - \hat{V}^{\hat{\pi}}\right\|_w \qquad (32)
\end{aligned}$$

where the first inequality follows from triangle inequality, and the last from the definition of Bellman mismatch functional and Lemma 1 as $\hat{\pi} \in \hat{\Pi}_S(\kappa, w)$. Re-arranging the terms in (32), we obtain

$$\left\|V^\pi - \hat{V}^{\hat{\pi}}\right\|_w \leq \frac{1}{1-\gamma\kappa} \mathcal{D}^{\pi,\hat{\pi}} V^\pi. \qquad (33)$$

Next consider

$$\begin{aligned}
\left\|V^\pi - \hat{V}^{\hat{\pi}}\right\|_w &= \left\|\mathcal{B}^\pi V^\pi - \hat{\mathcal{B}}^{\hat{\pi}} \hat{V}^{\hat{\pi}}\right\|_w \\
&\leq \left\|\mathcal{B}^\pi V^\pi - \mathcal{B}^\pi \hat{V}^{\hat{\pi}}\right\|_w + \left\|\mathcal{B}^\pi \hat{V}^{\hat{\pi}} - \hat{\mathcal{B}}^{\hat{\pi}} \hat{V}^{\hat{\pi}}\right\|_w \\
&\leq \gamma\kappa \left\|V^\pi - \hat{V}^{\hat{\pi}}\right\|_w + \mathcal{D}^{\pi,\hat{\pi}} \hat{V}^{\hat{\pi}} \qquad (34)
\end{aligned}$$

where the first inequality follows from triangle inequality, and the last from Lemma 1 as $\pi \in \Pi_S(\kappa, w)$ and from the definition of Bellman mismatch functional. Re-arranging the terms in (34), we obtain

$$\left\|V^\pi - \hat{V}^{\hat{\pi}}\right\|_w \leq \frac{1}{1-\gamma\kappa} \mathcal{D}^{\pi,\hat{\pi}} \hat{V}^{\hat{\pi}}. \qquad (35)$$

Combining (33) and (35) establishes (10).

Appendix D
Proof of Lemma 3

Proof of part 1: The result follows by using Lemma 2 with policies $\pi = \pi^\star$ and $\hat{\pi} = \hat{\pi}^\star$, in which case $V^{\pi^\star} = V^\star$ and $\hat{V}^{\hat{\pi}^\star} = \hat{V}^\star$.

Proof of part 2: If $\mathcal{B}^\star$ were a $\|\cdot\|_w$-norm contraction, then we could have used the exact same proof argument as in proof of part 1. However, we have not established that $\mathcal{B}^\star$ is a $\|\cdot\|_w$-norm contraction under Assumptions 1, 2 and 3.[2] So, we need a different proof argument. We use the shorthand notation $[v]_w$ to denote $\sup_{s\in\mathcal{S}} v(s)/w(s)$ (note that there is no absolute value sign around $v(s)$).

From Assumption 3, we know that $\mathcal{G}(\hat{V}^\star) \cap \Pi_S(\kappa, w) \neq \emptyset$. Let $\mu^\star \in \mathcal{G}(\hat{V}^\star) \cap \Pi_S(\kappa, w)$. Now, consider

$$\begin{aligned}
[V^\star - \hat{V}^\star]_w &= [\mathcal{B}^\star V^\star - \hat{\mathcal{B}}^\star \hat{V}^\star]_w \\
&\overset{(a)}{\leq} [\mathcal{B}^\star V^\star - \mathcal{B}^\star \hat{V}^\star]_w + [\mathcal{B}^\star \hat{V}^\star - \hat{\mathcal{B}}^\star \hat{V}^\star]_w \\
&\overset{(b)}{\leq} [\mathcal{B}^{\mu^\star} V^\star - \mathcal{B}^{\mu^\star} \hat{V}^\star]_w + \mathcal{D}^\star \hat{V}^\star \\
&\overset{(c)}{\leq} \gamma\kappa \|V^\star - \hat{V}^\star\|_w + \mathcal{D}^\star \hat{V}^\star \qquad (36)
\end{aligned}$$

where (a) follows from the definition supremum, (b) follows from $\mu^\star \in \mathcal{G}(\hat{V}^\star)$ and the fact that $\mathcal{B}^\star V^\star \leq \mathcal{B}^{\mu^\star} V^\star$, and

[2] Under Assumption 2, $\mathcal{B}^{\pi^\star}$ is a contraction. Even though $\mathcal{B}^\star$ and $\mathcal{B}^{\pi^\star}$ have the same fixed point, contractivity of $\mathcal{B}^{\pi^\star}$ does not imply contractivity of $\mathcal{B}^\star$ because they are different operators.

(c) follows from contraction of the Bellman operator $\mathcal{B}^{\mu^\star}$ (since $\mu^\star \in \Pi_S(\kappa, w)$).

Now we consider the inequality in the other direction.

$$\begin{aligned}[\hat{V}^\star - V^\star]_w &= [\hat{\mathcal{B}}^\star \hat{V}^\star - \mathcal{B}^\star V^\star]_w \\ &\overset{(d)}{\leq} [\hat{\mathcal{B}}^\star \hat{V}^\star - \mathcal{B}^\star \hat{V}^\star]_w + [\mathcal{B}^\star \hat{V}^\star - \mathcal{B}^\star V^\star]_w \\ &\overset{(e)}{\leq} \mathcal{D}^\star \hat{V}^\star + [\mathcal{B}^{\pi^\star} \hat{V}^\star - \mathcal{B}^{\pi^\star} V^\star]_w \\ &\overset{(f)}{\leq} \mathcal{D}^\star \hat{V}^\star + \gamma\kappa \|\hat{V}^\star - V^\star\|_w \end{aligned} \tag{37}$$

where (d) follows from the definition of supremum, (e) follows from $\pi^\star \in \mathcal{G}(V^\star)$ and the fact that $\mathcal{B}^\star \hat{V}^\star \leq \mathcal{B}^{\pi^\star} \hat{V}^\star$, and (f) follows from contraction of the Bellman operator $\mathcal{B}^{\pi^\star}$ (since $\pi^\star \in \Pi_S(\kappa, w)$).

Combining (36) and (37) and rearranging terms, we get (12).

Proof of part 3: The proof argument is similar to that of part 2. Consider

$$\begin{aligned}[V^\star - \hat{V}^\star]_w &= [\mathcal{B}^\star V^\star - \hat{\mathcal{B}}^\star \hat{V}^\star]_w \\ &\overset{(a)}{\leq} [\mathcal{B}^\star V^\star - \hat{\mathcal{B}}^\star V^\star]_w + [\hat{\mathcal{B}}^\star V^\star - \hat{\mathcal{B}}^\star \hat{V}^\star]_w \\ &\overset{(b)}{\leq} \mathcal{D}^\star V^\star + [\hat{\mathcal{B}}^{\hat{\pi}^\star} V^\star - \hat{\mathcal{B}}^{\hat{\pi}^\star} \hat{V}^\star]_w \\ &\overset{(c)}{\leq} \mathcal{D}^\star V^\star + \gamma\kappa \|V^\star - \hat{V}^\star\|_w \end{aligned} \tag{38}$$

where (a) follows from the definition of supremum, (b) follows the definition of $\hat{\pi}^\star$ and the fact that $\hat{\mathcal{B}}^\star V^\star \leq \hat{\mathcal{B}}^{\hat{\pi}^\star} V^\star$, and (c) follows from contraction of the Bellman operator $\hat{\mathcal{B}}^{\hat{\pi}^\star}$.

From Assumption 4, we know that $\hat{\mathcal{G}}(V^\star) \cap \hat{\Pi}_S(\kappa, w) \neq \emptyset$. Let $\hat{\mu}^\star \in \hat{\mathcal{G}}(V^\star) \cap \hat{\Pi}_S(\kappa, w)$. Now, we consider the inequality in the other direction.

$$\begin{aligned}[\hat{V}^\star - V^\star]_w &= [\hat{\mathcal{B}}^\star \hat{V}^\star - \mathcal{B}^\star V^\star]_w \\ &\overset{(d)}{\leq} [\hat{\mathcal{B}}^\star \hat{V}^\star - \hat{\mathcal{B}}^\star V^\star]_w + [\hat{\mathcal{B}}^\star V^\star - \mathcal{B}^\star V^\star]_w \\ &\overset{(e)}{\leq} [\hat{\mathcal{B}}^{\hat{\mu}^\star} \hat{V}^\star - \hat{\mathcal{B}}^{\hat{\mu}^\star} V^\star]_w + \mathcal{D}^\star V^\star \\ &\overset{(f)}{\leq} \gamma\kappa \|\hat{V}^\star - V^\star\|_w + \mathcal{D}^\star V^\star \end{aligned} \tag{39}$$

where (d) follows from the definition of supremum, (e) follows the definition of $\hat{\mu}^\star$ and the fact that $\hat{\mathcal{B}}^\star \hat{V}^\star \leq \hat{\mathcal{B}}^{\hat{\mu}^\star} \hat{V}^\star$, and (f) follows from contraction of the Bellman operator $\hat{\mathcal{B}}^{\hat{\mu}^\star}$.

Combining (38) and (39) and rearranging terms, we get (13).

Appendix E
Proof of Theorem 1

Proof of part 1: For the bound in terms of $\hat{V}^\star$, by triangle inequality, we have

$$\left\|V^{\hat{\pi}^\star} - V^\star\right\|_w \leq \left\|V^{\hat{\pi}^\star} - \hat{V}^\star\right\|_w + \left\|\hat{V}^\star - V^\star\right\|_w \tag{40}$$

Recall that $\hat{V}^\star = \hat{V}^{\hat{\pi}^\star}$. Since $\hat{\pi}^\star \in \Pi_S(\kappa, w) \cap \hat{\Pi}_S(\kappa, w)$, we can use Lemma 2 to bound the first term of (40) by

$$\begin{aligned}\left\|V^{\hat{\pi}^\star} - \hat{V}^\star\right\|_w &\leq \frac{1}{1-\gamma\kappa} \mathcal{D}^{\hat{\pi}^\star, \hat{\pi}^\star} \hat{V}^{\hat{\pi}^\star} = \frac{1}{1-\gamma\kappa} \mathcal{D}^{\hat{\pi}^\star} \hat{V}^{\hat{\pi}^\star} \\ &= \frac{1}{1-\gamma\kappa} \mathcal{D}^{\hat{\pi}^\star} \hat{V}^\star. \end{aligned} \tag{41}$$

We bound the second term in (40) using Lemma 3, part 1 by

$$\|\hat{V}^\star - V^\star\|_w \leq \frac{1}{(1-\gamma\kappa)} \mathcal{D}^{\pi^\star, \hat{\pi}^\star} \hat{V}^\star. \tag{42}$$

The result is obtained by combining (41) and (42).

For the bound in terms of $V^\star$, we can write

$$\begin{aligned}\left\|V^{\hat{\pi}^\star} - V^\star\right\|_w &= \left\|\mathcal{B}^{\hat{\pi}^\star} V^{\hat{\pi}^\star} - \mathcal{B}^\star V^\star\right\|_w \\ &\leq \left\|\mathcal{B}^{\hat{\pi}^\star} V^{\hat{\pi}^\star} - \mathcal{B}^{\hat{\pi}^\star} V^\star\right\|_w + \left\|\mathcal{B}^{\hat{\pi}^\star} V^\star - \hat{\mathcal{B}}^{\hat{\pi}^\star} V^\star\right\|_w \\ &\quad + \left\|\hat{\mathcal{B}}^{\hat{\pi}^\star} V^\star - \hat{\mathcal{B}}^{\hat{\pi}^\star} \hat{V}^\star\right\|_w + \left\|\hat{\mathcal{B}}^{\hat{\pi}^\star} \hat{V}^\star - \mathcal{B}^\star V^\star\right\|_w \\ &\leq \gamma\kappa \left\|V^{\hat{\pi}^\star} - V^\star\right\|_w + \mathcal{D}^{\hat{\pi}^\star} V^\star \\ &\quad + \gamma\kappa \left\|V^\star - \hat{V}^\star\right\|_w + \left\|\hat{V}^\star - V^\star\right\|_w \end{aligned} \tag{43}$$

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 (43), we obtain

$$\left\|V^{\hat{\pi}^\star} - V^\star\right\|_w \leq \frac{1}{1-\gamma\kappa} \left[\mathcal{D}^{\hat{\pi}^\star} V^\star + (1+\gamma\kappa) \left\|\hat{V}^\star - V^\star\right\|_w\right]. \tag{44}$$

We use Lemma 3, part 1, to bound the last term of (44) by

$$\|\hat{V}^\star - V^\star\|_w \leq \frac{1}{(1-\gamma\kappa)} \mathcal{D}^{\pi^\star, \hat{\pi}^\star} V^\star. \tag{45}$$

The result is obtained by combining (44) and (45).

Proof of part 2: Since Assumption 3 holds, we can use Lemma 3 part 2 to bound the second term of (40) by

$$\left\|\hat{V}^\star - V^\star\right\|_w \leq \frac{1}{1-\gamma\kappa} \mathcal{D}^\star \hat{V}^\star. \tag{46}$$

Combining (41) and (46) completes the proof.

Proof of part 3: Since Assumption 4 holds, we can use Lemma 3 part 3 to bound the last term of (44) by

$$\left\|\hat{V}^\star - V^\star\right\|_w \leq \frac{1}{1-\gamma\kappa} \mathcal{D}^\star V^\star. \tag{47}$$

Combining (44) and (47) completes the proof.

Appendix F
Proof of Lemma 4

Assumption 5 implies that, for each $a \in \mathcal{A}$,

$$\int_{\mathcal{S}} w(s') P(ds' \mid s, a) \leq \kappa w(s), \quad \forall s \in \mathcal{S}. \tag{48}$$

For any $v \in \mathcal{V}_w$ such that $\mathcal{G}(v)$ is nonempty, let π_v denote a policy in $\mathcal{G}(v)$. We will first show that $\pi_v \in \Pi_S(\kappa, w)$ and then use this to prove Assumption 3.

For policy π_v, we have that

$$\begin{aligned}\int_{\mathcal{S}} w(s') P_{\pi_v}(ds'|s) &= \int_{\mathcal{S}} \int_{\mathcal{A}} w(s') \pi_v(da|s) P(ds'|s,a) \\ &= \int_{\mathcal{A}} \int_{\mathcal{S}} w(s') P(ds'|s,a) \pi_v(da|s) \\ &\leq \int_{\mathcal{A}} \kappa w(s) \pi_v(da|s) = \kappa w(s), \quad \forall s \in \mathcal{S}. \end{aligned}$$

Thus, π_v satisfies (4).

Next, from the definition of the Bellman operator, we have

$$\begin{aligned} c_{\pi_v}(s) =& [\mathcal{B}^{\pi_v} v](s) - \gamma \int_{\mathcal{A}} \pi_v(da \mid s) \int_{\mathcal{S}} v(s') P(ds' \mid s, a) \\ \leq& [\mathcal{B}^{\pi_v} v](s) + \gamma |c_{\min}/(1-\gamma)|, \end{aligned}$$

where the inequality above is due to the fact that $-v(s) \leq |c_{\min}/(1-\gamma)|$ for all $v \in \mathcal{V}_w$. Therefore,

$$\begin{aligned} \|c_{\pi_v}\|_w =& \sup_{s \in \mathcal{S}} \frac{|c_{\pi_v}(s)|}{w(s)} \\ \leq& \sup_{s \in \mathcal{S}} \frac{[\mathcal{B}^{\pi_v} v](s) + \gamma |c_{\min}/(1-\gamma)|}{w(s)} \\ \leq& \|\mathcal{B}^{\pi_v} v\|_w + \gamma |c_{\min}/(1-\gamma)| \\ =& \|\mathcal{B}^{\star} v\|_w + \gamma |c_{\min}/(1-\gamma)| < \infty. \end{aligned}$$

Thus, π_v satisfies (3). Hence, $\pi_v \in \Pi_S(\kappa, w)$.

With $v = \hat{V}^\star$ in the above argument, it follows that $\pi_{\hat{V}^\star} \in \Pi_S(\kappa, w)$. Hence, Assumption 3 is satisfied.

Similarly, for any $v \in \mathcal{V}_w$, let $\hat{\pi}_v$ denote a policy in $\hat{\mathcal{G}}(v)$. Using arguments identical to the ones used above, we can show that $\hat{\pi}_v \in \hat{\Pi}_S(\kappa, w)$. Setting $v = V^\star$ then implies that $\hat{\pi}_{V^\star} \in \hat{\Pi}_S(\kappa, w)$. Hence, Assumption 4 is satisfied.

Appendix G
Proof of Lemma 5

Proof of part 1: By definition, $\mathcal{D}^{\max} v = \sup_{a \in \mathcal{A}} \mathcal{D}^{\pi_a} v$. By Assumption 5, $\pi_a \in \Pi_S(\kappa, w) \cap \hat{\Pi}_S(\kappa, w)$ for all $a \in \mathcal{A}$. Therefore,

$$\mathcal{D}^{\max} v \leq \sup_{\pi \in \Pi_S(\kappa, w) \cap \hat{\Pi}_S(\kappa, w)} \mathcal{D}^{\pi} v. \tag{49}$$

Define

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

Then, we have that

$$\mathcal{D}^{\max} v = \sup_{a \in \mathcal{A}} \|\mathcal{B}^{\pi_a} v - \hat{\mathcal{B}}^{\pi_a} v\|_w = \sup_{a \in \mathcal{A}} \sup_{s \in \mathcal{S}} \frac{|\Xi^{(s,a)} v|}{w(s)}$$

and for every $\pi \in \Pi_S(\kappa, w) \cap \hat{\Pi}_S(\kappa, w)$,

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

Therefore,

$$\begin{aligned} \mathcal{D}^{\pi} v =& \sup_{s \in \mathcal{S}} \frac{\left| \int_{\mathcal{A}} \pi(da \mid s) \Xi^{(s,a)} v \right|}{w(s)} \\ \leq& \sup_{s \in \mathcal{S}} \sup_{a \in \mathcal{A}} \frac{|\Xi^{(s,a)} v|}{w(s)} = \mathcal{D}^{\max} v. \end{aligned} \tag{50}$$

Combining (49) and (50), we get the first part of (15)

Proof of part 2: Note that for any set $\mathcal{X}$, $|\inf_{x \in \mathcal{X}} f(x) - \inf_{x \in \mathcal{X}} g(x)| \leq \sup_{x \in \mathcal{X}} |f(x) - g(x)|$. Therefore,

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

Using the above inequality in the definition of $\mathcal{D}^\star$, we get

$$\begin{aligned} \mathcal{D}^\star v =& \sup_{s \in \mathcal{S}} \frac{\left| [\mathcal{B}^\star v](s) - [\hat{\mathcal{B}}^\star v](s) \right|}{w(s)} \\ \leq& \sup_{s \in \mathcal{S}} \frac{\sup_{a \in \mathcal{A}} \left| \Xi^{(s,a)} v \right|}{w(s)} = \mathcal{D}^{\max} v, \end{aligned} \tag{51}$$

which establishes the second part of (15).

Appendix H
Proof of Theorem 3

Proof of part 1: The bound in terms of $\hat{V}^\star$ follows from Theorem 1, part 1 (the bound in terms of $\hat{V}^\star$) and Lemma 6, part 3.

For the bound in terms of $V^\star$, we can use Lemma 6, part 3 and Theorem 1, part 1 to write

$$\begin{aligned} \left\|V^{\hat{\pi}^\star} - V^\star\right\|_w =& \frac{1}{\alpha_1} \left\|V^{\hat{\pi}^\star}_{\boldsymbol{\alpha}} - V^\star_{\boldsymbol{\alpha}}\right\|_w \\ \leq& \frac{1}{\alpha_1(1-\gamma\kappa)} \mathcal{D}^{\hat{\pi}^\star}_{\boldsymbol{\alpha}} V^\star_{\boldsymbol{\alpha}} + \frac{(1+\gamma\kappa)}{\alpha_1(1-\gamma\kappa)^2} \mathcal{D}^{\pi^\star, \hat{\pi}^\star} V^\star_{\boldsymbol{\alpha}} \\ =& \frac{1}{\alpha_1(1-\gamma\kappa)} \mathcal{D}^{\hat{\pi}^\star}_{\boldsymbol{\alpha}} (\alpha_1 V^\star) + \frac{(1+\gamma\kappa)}{\alpha_1(1-\gamma\kappa)^2} \mathcal{D}^{\pi^\star, \hat{\pi}^\star} (\alpha_1 V^\star), \end{aligned}$$

where, the last step uses part 2 of Lemma 6 to conclude that

$$\mathcal{D}^{\hat{\pi}^\star}_{\boldsymbol{\alpha}} V^\star_{\boldsymbol{\alpha}} = \mathcal{D}^{\hat{\pi}^\star}_{\boldsymbol{\alpha}} (\alpha_1 V^\star) \quad \text{and} \quad \mathcal{D}^{\pi^\star, \hat{\pi}^\star}_{\boldsymbol{\alpha}} V^\star_{\boldsymbol{\alpha}} = \mathcal{D}^{\pi^\star, \hat{\pi}^\star}_{\boldsymbol{\alpha}} (\alpha_1 V^\star).$$

Proof of part 2: Assumption 6 is the same as Assumption 3 for model $\mathcal{M}_{\boldsymbol{\alpha}}$. Hence, the result follows immediately from Theorem 1, part 2 and Lemma 6, part 3.

Proof of part 3: Assumption 7 is the same as Assumption 4 for model $\mathcal{M}_{\boldsymbol{\alpha}}$. Hence, the result follows from Theorem 1, part 3 together with Lemma 6, part 3 and the fact that part 2 of Lemma 6 implies

$$\mathcal{D}^\star_{\boldsymbol{\alpha}} V^\star_{\boldsymbol{\alpha}} = \mathcal{D}^\star_{\boldsymbol{\alpha}} (\alpha_1 V^\star).$$

Appendix I
Proof of Proposition 2

When $\alpha_1 = 1$, Assumptions 3 and 6 are equivalent. To apply Theorem 3, we consider $\alpha_1 = 1$ and an arbitrary α_2. Then the Bellman updates for $\hat{V}^\star(s)$ can be calculated as

$$\begin{aligned} &\mathcal{B}^\star_{(1,\alpha_2)} \hat{V}^\star(s) \\ &= s^\intercal \left(Q + \gamma A^\intercal \hat{P} A - \gamma^2 A^\intercal \hat{P} B (R + \gamma B^\intercal \hat{P} B)^{-1} B^\intercal \hat{P} A \right) s \\ &\quad + \gamma(\hat{q} + \operatorname{Tr}(\Sigma_W \hat{P})) + \alpha_2, \end{aligned}$$

and

$$\begin{aligned} \hat{\mathcal{B}}^\star \hat{V}^\star(s) = \hat{V}^\star(s) &= s^\intercal \hat{P} s + \hat{q}. \\ &= s^\intercal \hat{P} s + \gamma(\hat{q} + \operatorname{Tr}(\hat{\Sigma}_W \hat{P})) \end{aligned}$$

where the last term uses the fact that $\hat{q} = \gamma \operatorname{Tr}(\hat{\Sigma}_W \hat{P})/(1-\gamma)$.

Therefore, we have

$$\begin{aligned} &|\mathcal{B}^\star_{(1,\alpha_2)} \hat{V}^\star(s) - \hat{\mathcal{B}}^\star \hat{V}^\star(s)| \\ &= \left| s^\intercal D^\star s + \gamma \operatorname{Tr}((\Sigma_W - \hat{\Sigma}_W) \hat{P}) + \alpha_2 \right|, \end{aligned}$$

where $D^\star$ is given by (22).

Note that $\hat{\pi}^\star(s) = -\hat{K}^\star s = -\gamma(\hat{R} + \gamma \hat{B}^\intercal \hat{P} \hat{B})^{-1} \hat{B}^\intercal \hat{P} \hat{A} s$. As a result, for any α_2, $\mathcal{B}^{\hat{\pi}^\star}_{(1,\alpha_2)} \hat{V}^\star(s)$ is given by

$$\begin{aligned}\mathcal{B}^{\hat{\pi}^\star}_{(1,\alpha_2)} \hat{V}^\star(s) &= s^\intercal \big(Q + (\hat{K}^\star)^\intercal R \hat{K}^\star + \gamma A_{\hat{K}^\star}^\intercal \hat{P} A_{\hat{K}^\star}\big) s \\ &\quad + \gamma(\hat{q} + \operatorname{Tr}(\Sigma_W \hat{P})) + \alpha_2,\end{aligned}$$

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

$$\begin{aligned}&|\mathcal{B}^{\hat{\pi}^\star}_{(1,\alpha_2)} \hat{V}^\star(s) - \hat{\mathcal{B}}^{\hat{\pi}^\star} \hat{V}^\star(s)| \\ &\quad = \Big| s^\intercal D^{\hat{\pi}^\star} s + \gamma \operatorname{Tr}((\Sigma_W - \hat{\Sigma}_W)\hat{P}) + \alpha_2 \Big|,\end{aligned}$$

where $D^{\hat{\pi}^\star}$ is given by (23). Then, the Bellman mismatches functionals of Section III-F for $\hat{V}$ with $(1, \alpha_2)$ can be calculated as follows:

$$\begin{aligned}\mathcal{D}^{\hat{\pi}^\star}_{(1,\alpha_2)} \hat{V}^\star &= \sup_{s \in \mathcal{S}} \frac{|s^\intercal D^{\hat{\pi}^\star} s + d_\Sigma + \alpha_2|}{w(s)}, \\ \mathcal{D}^{\star}_{(1,\alpha_2)} \hat{V}^\star &= \sup_{s \in \mathcal{S}} \frac{|s^\intercal D^{\star} s + d_\Sigma + \alpha_2|}{w(s)},\end{aligned}$$

where d_Σ is given by (25).

Eq. (21) then then follows from Theorem 3 part 2 by observing that for any symmetric matrix D

$$\sup_{s \in \mathcal{S}} \frac{|s^\intercal D s + d_\Sigma + \alpha_2|}{1 + \ell s^\intercal s} \le \max\left\{ \frac{\rho(D)}{\ell}, |d_\Sigma + \alpha_2| \right\}.$$

Appendix J
Proof of Lemma 8

Proof of part 1: For any (α_1, α_2) with $\alpha_2 > 0$, we have

$$\begin{aligned}\mathcal{D}^{\pi,\hat{\pi}}_{\boldsymbol{\alpha}} v =& \sup_{s \in \mathcal{S}} \frac{\big|\mathcal{B}^{\pi}_{\boldsymbol{\alpha}} v(s) - \hat{\mathcal{B}}^{\hat{\pi}} v(s)\big|}{w(s)} \\ \le& \sup_{s \in \mathcal{S}} \frac{\big|\alpha_1 c_\pi(s) + \alpha_2 - \hat{c}_{\hat{\pi}}(s)\big|}{w(s)} \\ &+ \gamma \sup_{s \in \mathcal{S}} \frac{\Big|\int_{\mathcal{S}} v(s') \Big[P_\pi\big(ds'|s)\big) - \hat{P}_{\hat{\pi}}\big(ds'|s)\big)\Big]\Big|}{w(s)} \\ \le& \varepsilon_{\boldsymbol{\alpha}}(\pi, \hat{\pi}) \\ &+ \gamma \rho_{\mathfrak{F}}(v) \sup_{s \in \mathcal{S}} \frac{d_{\mathfrak{F}}\Big(P_\pi(\cdot|s), \hat{P}_{\hat{\pi}}(\cdot|s)\Big)}{w(s)} \\ =& \varepsilon_{\boldsymbol{\alpha}}(\pi, \hat{\pi}) + \gamma \rho_{\mathfrak{F}}(v) \delta_{\mathfrak{F}}(\pi, \hat{\pi}) \qquad (52)\end{aligned}$$

Proofs of part 2 and 3: Part 2 follows because $\mathcal{D}^\star_{\boldsymbol{\alpha}} \hat{V}^\star = \mathcal{D}^{\mu^\star, \hat{\pi}^\star} \hat{V}^\star$. Similarly, part 3 follows because $\mathcal{D}^\star_{\boldsymbol{\alpha}}(\alpha_1 V^\star) = \mathcal{D}^{\pi^\star, \hat{\mu}^\star}(\alpha_1 V^\star)$.

Proof of part 4: Define

$$\begin{aligned}\Xi^{(s,a)}_{\boldsymbol{\alpha}} v =& \alpha_1 c(s,a) + \alpha_2 - \hat{c}(s,a) \\ &+ \gamma \int_{\mathcal{S}} v(s') P(ds'|s,a) - \gamma \int_{\mathcal{S}} v(s') \hat{P}(ds'|s,a).\end{aligned}$$

Then, we have

$$\begin{aligned}\mathcal{D}^{\max}_{\boldsymbol{\alpha}} v =& \sup_{a \in \mathcal{A}} \|\mathcal{B}^{\pi_a}_{\boldsymbol{\alpha}} v - \hat{\mathcal{B}}^{\pi_a} v\|_w \\ =& \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} \frac{|\Xi^{(s,a)}_{\boldsymbol{\alpha}} v|}{w(s)} \\ \le& \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} \frac{|\alpha_1 c(s,a) + \alpha_2 - \hat{c}(s,a)|}{w(s)} \\ &+ \gamma \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} \frac{\big|\int_{\mathcal{S}} v(s')[P(ds'|s,a) - \hat{P}(ds'|s,a)]\big|}{w(s)} \\ \le& \varepsilon^{\max}_{\boldsymbol{\alpha}} + \gamma \rho_{\mathfrak{F}}(v) \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} \frac{d_{\mathfrak{F}}\big(P(\cdot|s,a), \hat{P}(\cdot|s,a)\big)}{w(s)} \\ =& \varepsilon^{\max}_{\boldsymbol{\alpha}} + \gamma \rho_{\mathfrak{F}}(v) \delta^{\max}_{\mathfrak{F}}. \qquad (53)\end{aligned}$$

Berk Bozkurt (Student Member, IEEE) received his BSc degree in Electrical and Electronics Engineering from Bilkent University, Ankara, Turkey, in 2021. He is currently a MSc student in Electrical and Computer Engineering Department at McGill University, Montreal, Canada. His research interests include reinforcement learning, game theory, stochastic control and Markov decision theory.

Aditya Mahajan (Senior Member, IEEE) is Professor of Electrical and Computer Engineering at McGill University, Montreal, Canada. He received the B.Tech degree in Electrical Engineering from the Indian Institute of Technology, Kanpur, India in 2003 and the MS and PhD degrees in Electrical Engineering and Computer Science from the University of Michigan, Ann Arbor, USA in 2006 and 2008, respectively. He serves or has served as Associate Editor of Transactions on Automatic Control, Control Systems Letters, and Math. of Control, Signal, and Systems.

Ashutosh Nayyar (Senior Member, IEEE) is an Associate Professor of Electrical and Computer Engineering at the University of Southern California. He received a M.S. degree in electrical engineering and computer science, a M.S. degree in applied mathematics, and a Ph.D. degree in electrical engineering and computer science from the University of Michigan, Ann Arbor, MI, USA, in 2008, 2011, and 2011, respectively. His research interests are in decentralized stochastic control, decentralized decision-making in sensing and communication systems, reinforcement learning, game theory and mechanism design.

Yi Ouyang received the B.S. degree in Electrical Engineering from the National Taiwan University, Taipei, Taiwan in 2009, and the M.Sc and Ph.D. in Electrical Engineering and Computer Science at the University of Michigan, in 2012 and 2015, respectively. He is currently a researcher at Preferred Networks, Burlingame, CA. His research interests include reinforcement learning, stochastic control, and stochastic dynamic games.