Near-Optimality Bounds for Greedy Periodic Policies with Application to Grid-Level Storage

Yuhai Hu
Department of Industrial & Systems Engineering
Lehigh University, USA
Email: yuh212@lehigh.edu

Boris Defourny
Department of Industrial & Systems Engineering
Lehigh University, USA
Email: defourny@lehigh.edu

Abstract—This paper is concerned with periodic Markov Decision Processes, as a simplified but already rich model for nonstationary infinite-horizon problems involving seasonal effects. Considering the class of policies greedy for periodic approximate value functions, we established improved near-optimality bounds for such policies, and derive a corresponding value-iteration algorithm suitable for periodic problems. The effectiveness of a parallel implementation of the algorithm is demonstrated on a grid-level storage control problem that involves stochastic electricity prices following a daily cycle.

I. INTRODUCTION

For stochastic dynamic programs with seasonality effects, such as inventory or storage problems with daily demand patterns, rolling-horizon look-ahead policies [1] often appear as a well-suited class of policies. A defect of look-ahead policies, however, is that end-of-horizon effects can be detrimental to the optimality of the decisions. For instance, the case opposing JP Morgan Ventures Energy Corporation (JPMVEC) to the Federal Energy Regulatory Commission (FERC), exposed bidding strategies that were designed to exploit flaws in the market clearing algorithm of the California Independent System Operator (CAISO); one flaw was directly related to the truncation of the planning horizon [2].

Two countermeasures are classically considered to mitigate end-of-horizon effects. The first countermeasure is the introduction of a terminal reward function. This essentially amounts to approximate the value function around the state at the terminal stage — and this is an art that requires domain knowledge. The second countermeasure is the extension of the horizon over which the look-ahead is performed. This amounts to assume that end-of-horizon effects die out by the time the backward optimization reaches the first stages. A challenge of this approach is the increased complexity of the look-ahead optimization problem, and in certain contexts, the unavailability of data relative to the extended horizon — for instance, longer term forecasts may not be available; for a multistep bidding problem, market participants may not have been required to submit offers further in the future; etc.

This paper is motivated by the synthesis of these two mitigation strategies. We consider policies that solve a discounted periodic dynamic program, over an infinite horizon, constructed by replicating the look-ahead problem or by appending a steady-state cycle to the look-ahead. The rationale is that the structure of a policy optimal for a finite-horizon Markov decision problem on $p$ stages is the same as the one for an infinite-horizon, discounted $p$-periodic Markov decision problem. However, the cyclo-stationary extension could significantly improve the approximation of the future reward process. Early use of this strategy can be found in the water reservoir operations literature [3], [4].

In this paper, we are concerned with the subproblem to be solved by the proposed class of policies, and we develop effective methods to solve periodic dynamic programs.

Periodic dynamic programs have of course been considered earlier [5], [6], [7], as well as variations thereof [8]. In particular, it is known that an optimal $p$-periodic Markov policy can be derived using $p$ value functions coupled by a Bellman-type recursion. Periodic dynamic programs can be viewed as dynamic programs with stationary reward and state-transition functions over a state space augmented with the position of time in the cycle, and therefore results from abstract dynamic programming are directly available [9].

The contribution of this paper is twofold.

- We study the near-optimality of nonstationary policies greedy for periodic approximate value functions, and provide bounds that are tighter than the general bounds used with stationary value functions on an augmented state space or specialized bounds established for periodic Markov Decision Processes.

- We formulate a periodic Markov Decision Process model for a grid-level energy storage control problem where random electricity prices follow daily patterns. The idea is to recalibrate the model every day and then solve it for operations on the next day in a rolling-horizon fashion. A numerical example with a daily cycle of 24 periods is provided.

The paper is organized as follows. Section II defines the periodic Markov Decision Problem, and recalls the optimality conditions. Section III establishes bounds useful to control the near-optimality of periodic policies based on periodic approximate value functions. Section IV describes a value-iteration algorithm, based on the results of the previous section. Section V formulates a simple model for optimizing grid-level storage operations given day-ahead and historical electricity prices, based on the periodic Markov Decision Process framework. It also reports on numerical work carried out to evaluate the effectiveness of a parallel implementation of the value iteration algorithm. Section VI concludes the present paper.
II. PERIODIC DISCOUNTED PROBLEMS

In this section, we recall the mathematical formulation and optimality conditions of the $p$-periodic discounted Markov Decision Problem ($p$-MDP).

“Periodic” refers to the way the reward and state transition functions vary cyclically over time — this is distinct from the notion of “periodic state” in the theory of Markov chains.

In our formulation, the state space and the action space are periodic. This proves to be useful to adapt the states to the time-dependent characteristics of the problem.

A. Problem Formulation

For some integer $p \geq 1$ referred to as the cycle length, let

$$\{(\mathcal{S}_i, \mathcal{A}_i, P_i, R_i)\}_{i=0, \ldots, p-1}$$

define a $p$-periodic Markov Decision Process ($p$-MDP):

- $\{\mathcal{S}_i\}_{i=0, \ldots, p-1}$ form a base collection of finite state spaces, such that the state $S_t$ at time $t \geq 0$ is in $\mathcal{S}_i$ where $i \equiv \text{mod}(t, p)$, or equivalently, $t = i + kp$ for some integer $k \geq 0$;
- $\{\mathcal{A}_i\}_{i=0, \ldots, p-1}$ form a base collection of finite action spaces, such that the action $A_t$ at time $t \geq 0$ is in $\mathcal{A}_i$ where $i \equiv \text{mod}(t, p)$.
- $P_i : \mathcal{S}_i \times \mathcal{A}_i \times \mathcal{S}_{i+1} \mapsto [0, 1]$ for $i = 0, \ldots, p-2$ and $P_{p-1} : \mathcal{S}_{p-1} \times \mathcal{A}_{p-1} \times \mathcal{S}_0 \mapsto [0, 1]$, form a base collection of state transition probability functions, such that $\text{Prob}(S_{t+1} = s' \mid S_t = s, A_t = a) = P_i(s, a, s')$ where $i \equiv \text{mod}(t, p)$.
- $R_i : \mathcal{S}_i \times \mathcal{A}_i \mapsto \mathbb{R}$ for $i = 0, \ldots, p-1$ form a base collection of bounded reward functions, such that the reward at time $t$ given $S_t = s$, $A_t = a$, is $r_t = R_i(s, a)$ where $i \equiv \text{mod}(t, p)$.

We then define $i(t) = \text{mod}(t, p)$ and define $(\mathcal{S}_t, \mathcal{A}_t, P_t, R_t)$ for $t \geq p$, where $\mathcal{S}_t = \mathcal{S}_{i(t)}$, $\mathcal{A}_t = \mathcal{A}_{i(t)}$, $P_t = P_{i(t)}$, and $R_t = R_{i(t)}$.

When $p = 1$, the problem of course reduces to a stationary MDP $(\mathcal{S}, \mathcal{A}, P, R)$.

Consider the class $\Pi$ of admissible nonstationary Markov policies

$$\pi = \{A^\pi_i\}_{i \geq 0},$$

where $A^\pi_i : \mathcal{S}_i \mapsto \mathcal{A}_i$ is the decision rule at time $t$, that maps the current state $s$ to an action $a = A^\pi_i(s)$ selected from a subset $A_i(s) \subseteq \mathcal{A}_i$ that represents a set of admissible actions given $s$. To streamline the notation, we just write $a \in A_t$ instead of $a \in A_i(s)$ in the sequel.

Let $\gamma \in (0, 1)$ be a discount factor, and $s \in \mathcal{S}_0$ an initial state. We consider the $p$-periodic Markov Decision Problem consisting in maximizing the expected discounted total return by the choice of an admissible nonstationary policy $\pi$:

$$V^\pi_0(s) = \max_{\pi \in \Pi} \mathbb{E}^\pi_0[\sum_{t=0}^{\infty} \gamma^t R_t(S_t, A^\pi_t(S_t)) \mid S_0 = s],$$

where $\mathbb{E}^\pi$ emphasizes that the probability distribution of $S_t$ depends on $\pi$.

B. Optimality Conditions

For brevity, we use the short-hand $P^i_{ss'}(a) = P_i(s, a, s')$. For all $t$, and for a fixed nonstationary policy $\pi$, the expected discounted cumulative reward-to-go at time $t$ when being in state $s$ and following policy $\pi$ is given by

$$V^\pi_t(s) = R_t(s, A^\pi_t(s)) + \gamma \sum_{s' \in \mathcal{S}_{i+1}} P^i_{ss'}(A^\pi_t(s)) V^\pi_{t+1}(s').$$

(3)

$V^\pi_0(s)$ is the value of policy $\pi$ when starting from state $s$.

Due to the periodic structure of the problem, Bellman’s principle of optimality leads to a system of equations involving $p$ value functions only,

$$V^\pi_i(s) = \max_{a \in \mathcal{A}_i} \{R_i(s, a) + \gamma \sum_{s' \in \mathcal{S}_0} P^i_{ss'}(a) V^\pi_{i+1}(s')\}$$

for $i = 0, 1, \ldots, p-2$,

$$V^\pi_{p-1}(s) = \max_{a \in \mathcal{A}_{p-1}} \{R_{p-1}(s, a) + \gamma \sum_{s' \in \mathcal{S}_0} P^{p-1}_{ss'}(a) V^\pi_0(s')\}.$$  (4)

These equations are written more compactly as

$$V_i = T_i V_{i+1} \quad \text{for } i = 0, 1, \ldots, p-2,$$

$$V_i = T_i V_0 \quad \text{for } i = p-1,$$

where the operators $T_i$ are defined from (4). By induction,

$$V_0 = T_0 T_1 \ldots T_{p-1} V_0,$$

$$V_1 = T_1 \ldots T_{p-1} T_0 V_1,$$

$$\ldots$$

$$V_{p-1} = T_{p-1} T_{p-2} \ldots T_0 V_{p-1},$$

showing that $V_i$ is a fixed point of the operator $T_i = (T_i T_{i+1} \ldots T_{p-1} T_0 \ldots T_{i-1})$.

(5)

The operator $T_i$ inherits the contraction mapping property of the operators $T_i$ (Section IV-B provides more details), and therefore, the system (4) admits a unique solution

$$V^* = (V^*_0, V^*_1, \ldots, V^*_{p-1}),$$

which we refer to as the optimal periodic value function.

We say that a policy $\pi = \{A^\pi_t\}_{t \geq 0}$ is greedy for a periodic value function $V = (V^*_0, \ldots, V^*_{p-1})$ when

$$A^\pi_t(s) \in \arg \max_{a \in \mathcal{A}_t} \{R_t(s, a) + \gamma \sum_{s' \in \mathcal{S}_{i+1}} P^i_{ss'}(a) V_{t+1}(s')\}$$

for $t = i + kp$ with $i \in \{0, 1, \ldots, p-2\}$,

$$A^\pi_t(s) \in \arg \max_{a \in \mathcal{A}_t} \{R_t(s, a) + \gamma \sum_{s' \in \mathcal{S}_0} P^i_{ss'}(a) V_0(s')\}$$

for $t = i + kp$ with $i = p-1$.

(7)

Let $\pi^*$ be a policy greedy for $V^*$ as defined by (6). Then $\pi^*$ is nonstationary but periodic with cycle length $p$, and by definition of $T_i$, it is optimal for the problem (2).

Without loss of optimality, the search over the class $\Pi$ of nonstationary policies is thus reduced to a search over the class $\Pi_p$ of $p$-periodic admissible policies $\pi$, such that $A^\pi_t = A^\pi_0$ with $i = i(t)$:

$$V^\pi_0(s) = \max_{\pi \in \Pi_p} \mathbb{E}^\pi_0[\sum_{t=0}^{\infty} \gamma^t R_t(S_t, A^\pi_t(S_t)) \mid S_0 = s].$$

(8)

A policy $\pi \in \Pi_p$ is uniquely defined by $(A^\pi_0, \ldots, A^\pi_{p-1})$. 
III. NEAR-OPTIMALITY BOUNDS FOR GREEDY POLICIES

This section establishes upper error bounds for the difference between the optimal return and the return of a policy greedy with respect to a periodic approximate value function. This situation covers the case of a policy greedy with respect to a periodic approximate value function obtained by value iteration.

The error bounds can be evaluated numerically under the assumption that the periodic optimal value function is known. As this assumption is not met in practice, we establish upper bounds to be used when the optimal value function is unknown.

A. Error bound for difference between optimal and approximate value functions

Let \( V^* = (V_0^*, \ldots, V_{p-1}^*) \) be the optimal periodic value function (6), and \( \pi^* \in \Pi_p \) the optimal \( p \)-periodic policy greedy for \( V^* \).

In many cases, the value functions \( V_i^* \) are difficult or even impossible to evaluate. In order to overcome such situations, it is common to use approximate value functions.

Let \( \tilde{V} = (\tilde{V}_0, \ldots, \tilde{V}_{p-1}) \) denote a periodic approximate value function, and let \( \tilde{\pi} \in \Pi_p \) be a \( p \)-periodic policy greedy for \( \tilde{V} \), that is,

\[
\tilde{\pi}_i \in \text{argmax}_{a \in A_i} R_i(s, a) + \gamma \sum_{s' \in S} P_{ss'}(a) \tilde{V}_{i+1}(s')
\]

for \( i = 0, \ldots, p-2 \),

\[
\tilde{\pi}_{p-1} \in \text{argmax}_{a \in A_{p-1}} R_{p-1}(s, a) + \gamma \sum_{s' \in S} P_{ss'}^{p-1}(a) \tilde{V}_0(s').
\]

Let \( V_i^\pi \) denote the “value” of policy \( \pi^* \) at time \( i \), and let \( V_i^\tilde{\pi} \) denote the “value” of policy \( \tilde{\pi} \) at time \( i \), as defined by (3), where “value” at time \( i \) means the expected cumulative reward-to-go obtained by following the policy from time \( i \) onwards. By definition of \( \pi^* \) and \( V^* \), we have \( V_i^\pi = V_i^* \), but in the case of \( \tilde{\pi} \), in general we have

\[
V_i^\tilde{\pi} \neq \tilde{V}_i.
\]

Definition 1. Given an optimal policy \( \pi^* \) associated to \( V^* \) and a policy \( \tilde{\pi} \) greedy for \( \tilde{V} \), the function \( \overline{L}_i : S_i \rightarrow \mathbb{R} \) is defined as the difference between the expected reward-to-go at time \( i \) of those two policies: For all \( s \in S_i \),

\[
\overline{L}_i(s) = V_i^*(s) - V_i^\tilde{\pi}(s).
\]

In particular, \( \overline{L}_0(s) \) quantifies the suboptimality of policy \( \tilde{\pi} \) for the periodic MDP started from initial state \( s \).

Assumption. In each time period \( i \), the value function \( V_i^* \) is approximated by \( \tilde{V}_i \), and for all \( s \in S_i \), the difference between those value functions is bounded by \( \epsilon_i \):

\[
|V_i^*(s) - \tilde{V}_i(s)| \leq \epsilon_i.
\]

Under the assumption above, we provide the following proposition. The mechanism of the proof is based on \([10]\).

Proposition 1. Let \( V^* = (V_0^*, \ldots, V_{p-1}^*) \) be the optimal periodic value function (6), let \( \pi^* \) be the associated optimal policy, and let \( \tilde{\pi} \) be a policy greedy for \( \tilde{V} = (\tilde{V}_0, \ldots, \tilde{V}_{p-1}) \), where \( \tilde{V} \) satisfies the assumption (11) for \( i = 0, \ldots, p - 1 \). Then for all states \( s \),

\[
\overline{L}_i(s) \leq \sum_{k=0}^{i} \gamma^{p+k-i}(2\epsilon_k) + \sum_{k=i+1}^{p-1} \gamma^{k-i}(2\epsilon_k) / (1 - \gamma^p).
\]

Before establishing Proposition 1, we note that each \( \overline{L}_i \) in (12) depends on all the \( \epsilon_k \), but the weighting differs among each \( i \). In the stationary case (\( p = 1 \)), the bound reduces to

\[
\overline{L}_0 \leq 2\epsilon_0 \gamma / (1 - \gamma).
\]

With \( \epsilon = \max_i \epsilon_i \), the bound also reduces to

\[
\overline{L}_i \leq 2\epsilon_0 \gamma / (1 - \gamma).
\]

Proof of Proposition 1: For each period \( i \), there exists a state, say \( z_i \), that achieves the maximal loss \( \overline{L}_i \) at this period:

\[
\overline{L}_i(z_i) = \overline{L}_i(s) \text{ for all } s \in S_i.
\]

To this state \( z_i \) corresponds the optimal action

\[
a = A_i^z(z_i),
\]

and the action of the policy greedy for \( \tilde{V} \),

\[
b = A_i^\tilde{z}(z_i).
\]

Momentarily let us assume \( i \leq p - 2 \). Since \( \tilde{\pi} \) is greedy for \( \tilde{V} \), we have

\[
R_i(z_i, a) + \gamma \sum_{s' \in S_{i+1}} P_{ss'}(a) \tilde{V}_{i+1}(s') \leq R_i(z_i, b) + \gamma \sum_{s' \in S_{i+1}} P_{ss'}(b) \tilde{V}_{i+1}(s')
\]

From the assumption \( |V_i^*(s) - \tilde{V}_i(s)| \leq \epsilon_i \) we have

\[
R_i(z_i, a) + \gamma \sum_{s' \in S_{i+1}} P_{ss'}(a)(V_i^*(s') - \epsilon_{i+1}) \leq R_i(z_i, b) + \gamma \sum_{s' \in S_{i+1}} P_{ss'}(b)(V_i^*(s') + \epsilon_{i+1}),
\]

which is equivalent to

\[
R_i(z_i, a) - R_i(z_i, b) \leq 2\gamma \epsilon_{i+1} + \gamma \sum_{s' \in S_{i+1}} [P_{ss'}(b) - P_{ss'}(a)] V_{i+1}(s') \leq 2\gamma \epsilon_{i+1} + \gamma \sum_{s' \in S_{i+1}} [P_{ss'}(b) - P_{ss'}(a)] V_{i+1}(s').
\]
On the other hand, we have, by definition of $\bar{L}_i$,
\[ \bar{L}_i(z_i) = V^*(z_i) - V_i^+(z_i) = R_i(z_i, a) - R_i(z_i, b) + \gamma \sum_{s' \in \mathcal{S}_{i+1}} [P^i_{z_i, s'}(a) V_{i+1}(s') - P^i_{z_i, s'}(b) V_i^+(s')] . \tag{15} \]

Combining (14) and (15), we obtain (for $i = 0, \ldots, p - 2$)
\[ \bar{L}_i(z_i) \leq 2\gamma \epsilon_{i+1} + \gamma \sum_{s'} P^i_{z_i, s'}(b) [V^*_i(s') - V_i^+(s')] = 2\gamma \epsilon_{i+1} + \gamma \sum_{s'} P^i_{z_i, s'}(b) \bar{L}_{i+1}(s') \leq 2\gamma \epsilon_{i+1} + \gamma \sum_{s'} P^i_{z_i, s'}(b) \bar{L}_{i+1}(z_{i+1}) = 2\gamma \epsilon_{i+1} + \gamma \bar{L}_{i+1}(z_{i+1}) , \tag{16} \]
where the second inequality results from the definition of $z_{i+1}$.

Similarly, for $i = p - 1$, we obtain
\[ \bar{L}_{p-1}(z_{p-1}) \leq 2\gamma \epsilon_0 + \gamma \bar{L}_0(z_0) . \tag{17} \]

By induction,
\[ \bar{L}_i(z_i) \leq 2\gamma \epsilon_{i+1} + 2\gamma^2 \epsilon_{i+2} + \cdots + 2\gamma^{p-1} \epsilon_{p-1} + 2\gamma^p \epsilon_i + \gamma^p \bar{L}_i(z_i) = 2 \sum_{k=i+1}^{p-1} \gamma^{k-i} \epsilon_k + 2 \sum_{k=0}^{p-1} \gamma^{p-i+k} \epsilon_k + \gamma^p \bar{L}_i(z_i) , \]
and finally,
\[ \bar{L}_i(s) \leq \bar{L}_i(z_i) \leq 2 \sum_{k=0}^{p-1} \gamma^{p-i+k} \epsilon_k + 2 \sum_{k=i+1}^{p-1} \gamma^{k-i} \epsilon_k . \]

B. Bounding $\epsilon_k$

From the analysis above, we may obtain an upper bound for the suboptimality of a periodic policy greedy for the periodic approximate value function $V$. However, in practice, the optimal periodic value function $V^*$ is unknown, and therefore we cannot compute the $\epsilon_k$’s of the assumption. Fortunately, we may bound $\epsilon_k$ using quantities obtained in the course of one iteration of the value iteration algorithm. The mechanism of the proof is based on results from the theory of value iteration presented for instance in [9].

Definition 2. Given $\bar{V}^t = (\bar{V}^t_0, \ldots, \bar{V}^t_{p-1})$, let $\bar{V}^{t+1}$ be defined by one value-iteration performed as follows:
\[ \bar{V}^{t+1}_i(s) = \max_{a \in \mathcal{A}_i} [R_i(s, a) + \gamma \sum_{s' \in \mathcal{S}_{i+1}} P^i_{s, s'}(a) \bar{V}^t_{i+1}(s')] = (T_i \bar{V}^t_i)(s) \text{ for } i = 0, \ldots, p - 2, \]
\[ \bar{V}^{t+1}_i(s) = \max_{a \in \mathcal{A}_i} [R_i(s, a) + \gamma \sum_{s' \in \mathcal{S}_0} P^i_{s, s'}(a) \bar{V}^t_0(s')] = (T_i \bar{V}^t_i)(s) \text{ for } i = p - 1. \]

Definition 3. Define $\delta_i$ as the maximal change of the value function relative to period $i$ using the update described in Definition 2, over all states $s \in \mathcal{S}$:
\[ \delta_i = \max_{s \in \mathcal{S}} |\bar{V}_i^{t+1}(s) - \bar{V}_i^t(s)| = \max_{s \in \mathcal{S}} |(T_i \bar{V}^t_i)(s) - \bar{V}_i^t(s)| \text{ for } i = 0, \ldots, p - 2 \]
\[ \max_{s \in \mathcal{S}_0} |(T_i \bar{V}^t_0)(s) - \bar{V}_i^t(s)| \text{ for } i = p - 1. \]

Proposition 2. Let $\bar{V}^t$ be an approximation to the optimal periodic value function $V^*$, and let $\tilde{\delta}_i$ be defined as above. In this situation,
\[ \epsilon_i = \max_{s \in \mathcal{S}_i} |V^*_i(s) - \bar{V}_i^t(s)| \]
admits for $i = 0, \ldots, p - 1$ the upper bound
\[ \epsilon_i \leq \sum_{k=0}^{i-1} \gamma^{p+k-i} \delta_k + \sum_{k=i+1}^{p-1} \gamma^{k-i} \delta_k \frac{1}{1 - \gamma^p} . \]

Before establishing Proposition 2, we note that in the stationary case ($p = 1$), the expression reduces to
\[ \epsilon_0 \leq \delta_0/(1 - \gamma) . \]

With $\delta = \max_i \delta_i$, the bound also reduces to
\[ \epsilon_i \leq \delta \frac{1}{1 - \gamma^p} \sum_{k=0}^{p-1} \gamma^k = \frac{1 - \gamma^p}{1 - \gamma} = \frac{\delta}{1 - \gamma} , \]
which is a bound known in the literature. The bound in Proposition 2 is tighter, since it does not replace $\delta_i$ by $\delta$. In fact, if instead of $\delta_i$ from Definition 3, we define
\[ \bar{\delta}_i = \max_{s \in \mathcal{S}_i} (\bar{V}_i^{t+1}(s) - \bar{V}_i^t(s)), \tag{18} \]
and consider $\bar{\epsilon}_i$, $\epsilon_i$ as defined in (13), then the result of Proposition 2 translates to
\[ \bar{\epsilon}_i \leq \sum_{k=0}^{i-1} \gamma^{p+k-i} \bar{\delta}_k + \sum_{k=i+1}^{p-1} \gamma^{k-i} \bar{\delta}_k \frac{1 - \gamma^p}{1 - \gamma^p} . \]

Corollary. The value for $2\epsilon_k$ in Proposition 1 can be set to
\[ 2\epsilon_k = \sum_{k=0}^{i-1} \gamma^{p+k-i} (\bar{\delta}_k - \delta_k) + \sum_{k=i+1}^{p-1} \gamma^{k-i} (\delta_k - \bar{\delta}_k) \frac{1 - \gamma^p}{1 - \gamma^p} . \tag{20} \]

As a preliminary to the proof of Proposition 2, we recall the following properties of the operators $T_i$.

1) Monotonicity: $V_i^{t+1} \geq V_{i+1}^{t+1}$ implies $T_i V_i^{t+1} \geq T_i V_{i+1}^{t+1}$, in the sense that $V_i^{t+1}(s) \geq V_{i+1}^{t+1}(s)$ for all $s \in \mathcal{S}_{i+1}$ implies $(T_i V_i^{t+1})(s) \geq (T_i V_{i+1}^{t+1})(s)$ for all $s \in \mathcal{S}_{i+1}$.

2) Uniform-shift: Let $I_i : \mathcal{S}_i \mapsto \mathbb{R}$ denote the constant-valued function defined on $\mathcal{S}_i$ with value one. Then for any $c \in \mathbb{R}$, it holds that that
\[ T_i (V_{i+1} + c I_{i+1}) = T_i V_{i+1} + c I_i , \]
since for all \( s \in \mathcal{S}_i \),
\[
\max_{a \in \mathcal{A}_i} [R_i(s, a) + \gamma \sum_{s' \in \mathcal{S}_{i+1}} P_{ss'}^i (s')(V_{i+1}(s) + c)]
\]
\[
= \max_{a \in \mathcal{A}_i} [R_i(s, a) + \gamma \sum_{s' \in \mathcal{S}_{i+1}} P_{ss'}^i (s')(V_{i+1}(s))] + \gamma c.
\]

**Proof of Proposition 2:** By definition of \( \delta_{i-1} \), we have
\[
\tilde{V}_{i-1}^\ell + \delta_{i-1} 1_{i-1} \geq T_{i-1} \tilde{V}_{i}^\ell.
\]
(21)

Applying \( T_{i-2} \) to both sides of the inequality, and using the uniform-shift and monotonicity properties of \( T_{i-2} \), we get
\[
T_{i-2} \tilde{V}_{i-1}^\ell + \gamma \delta_{i-1} 1_{i-2} \geq T_{i-2} T_{i-1} \tilde{V}_{i}^\ell.
\]

By definition of \( \delta_{i-2} \), we deduce
\[
\tilde{V}_{i-2}^\ell + \delta_{i-2} 1_{i-2} + \gamma \delta_{i-1} 1_{i-2} \geq T_{i-2} T_{i-1} \tilde{V}_{i}^\ell
\]
(22)

By repeating this process to cover a single cycle, we obtain the inequalities
\[
\tilde{V}_{i}^\ell + (\gamma^0 \delta_i + \gamma^1 \delta_{i+1} + \cdots + \gamma^{p-1} \delta_{p-1} +
\gamma^{p-1} \delta_0 + \cdots + \gamma^{p-1} \delta_{i-1}) 1_i
\]
\[
\geq T_{i} T_{i+1} \cdots T_{p-1} T_0 \cdots T_{i-1} \tilde{V}_{i}^\ell
\]

By induction over an infinite number of cycles, we obtain
\[
\tilde{V}_{i}^\ell + (\gamma^0 \delta_i + \gamma^1 \delta_{i+1} + \cdots + \gamma^{p-1} \delta_{p-1} +
\gamma^{p-1} \delta_0 + \cdots + \gamma^{p-1} \delta_{i-1}) 1_i
\]
\[
\geq \lim_{N \to \infty} (T_i T_{i+1} \cdots T_{p-1} T_0 \cdots T_{i-1})^N \tilde{V}_{i}^\ell
\]
\[
= V_i^* \equiv V_i^\ell
\]
where the last equality comes from the convergence of the value iteration algorithm for finding the fixed point \( V_i^\ell \) of the operator \( T_i \) as defined in (5).

Rearranging the terms of the inequality above, we obtain
\[
\frac{1}{1 - \gamma^p} \left( \sum_{k=0}^{i-1} \gamma^{p+k-i} \delta_k + \sum_{k=i}^{p-1} \gamma^{k-i} \delta_k \right) 1_i \geq V_i^* - \tilde{V}_{i}^\ell
\]

Over all states, this implies
\[
\sum_{k=0}^{i-1} \gamma^{p+k-i} \delta_k + \sum_{k=i}^{p-1} \gamma^{k-i} \delta_k \geq \max_{s \in \mathcal{S}_i} (V_i^*(s) - \tilde{V}_{i}^\ell(s))
\]
(23)

A similar reasoning starting from the inequality
\[
\tilde{V}_{i-1}^\ell - \delta_{i-1} 1_{i-1} \leq T_{i-1} \tilde{V}_{i}^\ell
\]
leads to
\[
\frac{1}{1 - \gamma^p} \left( \sum_{k=0}^{i-1} \gamma^{p+k-i} \delta_k + \sum_{k=i}^{p-1} \gamma^{k-i} \delta_k \right) \leq \min_{s \in \mathcal{S}_i} (V_i^*(s) - \tilde{V}_{i}^\ell(s))
\]

which together with (23) implies
\[
\frac{\sum_{k=0}^{i-1} \gamma^{p+k-i} \delta_k + \sum_{k=i}^{p-1} \gamma^{k-i} \delta_k}{1 - \gamma^p} \geq \max_{s \in \mathcal{S}_i} |V_i^*(s) - \tilde{V}_{i}^\ell(s)|
\]
\[
= \epsilon_i
\]

**IV. VALUE-ITERATION**

In this section, we describe a value-iteration algorithm suitable for periodic Markov Decision Processes in finite state-action spaces. The algorithm outputs a periodic value function \( \bar{V} \) such that a periodic policy greedy for \( \bar{V} \) is guaranteed to be (at least) \( \eta \)-optimal.

**A. Iteration mechanism**

The optimality condition \( V_0 = (T_0 \cdots T_{p-1}) V_0 \) suggests a value iteration algorithm for solving the \( p \)-periodic Markov Decision Problem.

Given \( \eta > 0 \), the algorithm returns a periodic value function \( \bar{V} = (V_0, \cdots \bar{V}_{p-1}) \) such that a policy greedy for \( \bar{V} \) is guaranteed to be (at least) \( \eta \)-optimal.

1) Initialization: Guess an initial value function \( \bar{V}_0^\ell \) for \( \ell = 0 \) (for instance, \( \bar{V}_0^0 \equiv 0 \)).

2) Value iteration: Compute successively
\[
\bar{V}_{p-1}^\ell = T_{p-1} \bar{V}_0^\ell \]
\[
\bar{V}_{p-2}^\ell = T_{p-2} \bar{V}_{p-1}^\ell
\]
\[
\cdots
\]
\[
\bar{V}_0^\ell + p = T_0 \bar{V}_0^\ell + 1.
\]

For \( i = 0, \ldots, p - 1 \), compute \( \delta_i, \bar{V}_i \) from (18),(19), and \( 2 \epsilon_i \) from (20).

Compute \( \bar{L}_0 \) from (12).

3) Set \( \ell \leftarrow \ell + 1 \) and repeat Step 2 until the stopping criterion \( \bar{L}_0 \leq \eta \) is met.

The value-iteration algorithm utilizes the bounds of Section III which are adapted to each value function in the cycle.

As it can be seen from (10), \( \bar{L}_0(s) \leq \eta \) indicates that a policy greedy with respect to the current value function \( \bar{V}^\ell \) is \( \eta \)-optimal. This does not imply that \( \bar{V}_0^\ell \) represent the value of the policy; that is, that \( \bar{V}_0^\ell \) has converged to \( V_0^* \). To see this, consider the example of a policy greedy with respect to \( V^* + cI \) where \( cI \) is a constant-valued function; this policy is optimal and its value is \( V_0^* \) where \( s \) is the initial state.

**B. Convergence rate**

For a function \( V_i : \mathcal{S}_i \rightarrow \mathbb{R} \), we consider the sup-norm
\[
||V_i||_{\infty,i} = \max_{s \in \mathcal{S}_i} |V_i(s)|,
\]
where we write max instead of sup because \( \mathcal{S}_i \) is finite.

**Proposition 3.** The rate of convergence of the value-iteration algorithm is governed by
\[
||T_0^k V_0 - V_0^*||_{\infty,0} \leq \gamma^k ||\bar{V}_0 - V_0^*||_{\infty,0}
\]

computed using $\ell = kp$ iterations.

**Proof:** The mapping $T_i = (T_i T_{i+1} \ldots T_p T_0 \ldots T_{i-1})$ is contractive with modulus $\gamma^p$, in the sense that for all functions $V_i, V'_i$ from $\mathcal{X}_i$ to $\mathbb{R}$,

$$||T_i V_i - T_i V'_i||_{\infty,i} \leq \gamma^p ||V_i - V'_i||_{\infty,i}.$$  \hfill (24)

To see this, note first that the mappings $T_i$ are contractions with modulus $\gamma$, in the sense that, for $i = 0, \ldots, p-2$, the following property holds for all functions $V_i, V'_i$ from $\mathcal{X}_{i+1}$ to $\mathbb{R}$:

$$||T_i V_{i+1} - T_i V'_{i+1}||_{\infty,i} \leq \gamma ||V_{i+1} - V'_{i+1}||_{\infty,i+1},$$

and for $i = p-1$, the following property holds for all functions $V_0, V'_0$ from $\mathcal{X}_0$ to $\mathbb{R}$:

$$||T_{p-1} V_0 - T_{p-1} V'_0||_{\infty,p-1} \leq \gamma ||V_0 - V'_0||_{\infty,0}.$$  

Then, by using the contractive property of the operators $T_k$ successively for $k = i, i+1, \ldots, p-1, 0, \ldots, i-1$, one gets

$$||T_i V_i - T_i V'_i||_{\infty,i} = ||(T_i T_{i+1} \ldots T_{i-1}) V_i - (T_i T_{i+1} \ldots T_{i-1}) V'_i||_{\infty,i} \leq \gamma ||(T_{i+1} \ldots T_{i-1}) V_i - (T_{i+1} \ldots T_{i-1}) V'_i||_{\infty,i+1}.$$

$$\leq \cdots \leq \gamma^p ||V_i - V'_i||_{\infty,i}.$$  

In particular, for $V_i = V_i^\ast$, we have $T_i V_i^\ast = V_i^\ast$ and therefore

$$||T_i V_i - V_i^\ast||_{\infty,i} \leq \gamma^p ||V_i - V_i^\ast||_{\infty,i}.$$  

It remains to set $i = 0$ and iterate $k$ times the mapping $T_0$ to get the result. \hfill \Box

V. APPLICATION TO GRID-LEVEL STORAGE OPERATIONS

We consider a grid-level storage control problem where the goal is to operate a battery to exchange electricity with the power grid at the current hourly spot price. The problem is formulated as $p$-periodic Markov Decision Process where the goal is to maximize the expected net proceeds from the purchase and selling of electricity over an infinite horizon.

A very appealing feature of the periodic Markov Decision Process model proposed in this paper is that its computational tractability will not be affected by adopting shorter time periods, for instance periods of 5 minutes or less, making it suitable for various storage devices with different physical characteristics (power and energy capacities) or operating at different time scales.

In our numerical implementation, we use C as our programming language, with OpenMP for parallel computations.

A. Model Description

The state $S_t$ is the current price $S_t^{\text{price}}$ and the battery energy charge level $S_t^{\text{battery}}$. The decision $A_t$ is the power at which we charge ($A_t < 0$) or discharge the battery ($A_t > 0$). The instantaneous reward $R_t$ is the revenue over the time period: price $\times$ energy injected to the grid,

$$R_t(S_t, A_t) = S_t^{\text{price}} A_t \Delta t,$$

where $\Delta t$ is the time period duration (1 hour).

The discount factor $\gamma$ is set to 0.99. Hence the weight of the reward of tomorrow’s hour-1 is $\gamma^{24} = 0.78$, and the weight of the reward of next week’s hour-1 is $\gamma^{168} = 0.18$.

The stochastic hourly price process is modeled as a cyclo-stationary process with a cycle length of $p = 24$ hours. The means are chosen to match the day-ahead prices posted by the independent system operator on the day preceding the exploitation of the policy. For the price volatility, we use the historical volatility of prices on similar days, although using implied volatility of options on forward contracts should yield better predictive distributions. Inter-hour correlations are neglected, but as the current price is in the state, specifying an order-1 Markov model for the price would simply change the state transition probabilities without increasing the complexity of the periodic MDP.

The hourly prices are assumed to follow a lognormal distribution $LN(\mu_i, \sigma_i^2)$. Other distributions could easily be accommodated. We formulate the problem on parameters estimated from PJM price data for one day of 2013, with historical volatilities estimated from prices of the corresponding month. The estimated parameters $\mu_i, \sigma_i$ are given in Table I, along with the corresponding mean prices $\exp(\mu_i + \sigma_i^2/2)$.

For the battery, we use the parameters of a GM Chevy Volt battery pack repurposed for energy storage by ABB, having $C^{\text{battery}} = 10 \text{ kWh}$ of usable capacity [11]. We assume a power rating of $P^{\text{battery}} = 5 \text{ kW}$, such that a full charge over the 10 kWh range can be done in 2 hours if desired. Intermediate charge and discharge rates are allowed, including the null injection $A_t = 0$ (pure storage).

The battery state transition function is given by

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

with a state-dependent action space defined by the constraints

$$-P^{\text{battery}} \leq A_t \leq P^{\text{battery}} \quad \text{(power capacity)},$$

$$S_t^{\text{battery}} - C^{\text{battery}} \leq A_t \Delta t \leq S_t^{\text{battery}} \quad \text{(energy capacity)}.$$  

A more detailed battery model could easily be accommodated.

<table>
<thead>
<tr>
<th>Hour i</th>
<th>$\mu_i$</th>
<th>$\sigma_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1–6</td>
<td>3.24</td>
<td>0.78</td>
</tr>
<tr>
<td>7–12</td>
<td>2.82</td>
<td>0.34</td>
</tr>
<tr>
<td>13–18</td>
<td>3.82</td>
<td>0.28</td>
</tr>
<tr>
<td>19–24</td>
<td>4.85</td>
<td>0.34</td>
</tr>
</tbody>
</table>
B. Finite State Approximation

For each hour i, the support of the price distribution is partitioned into N = 20 cells \( [c_{i,j-1}, c_{i,j}) \) of probability \( 1/N \), and discrete price levels \( s_{\text{price}}^{i,j} \) are determined by computing the conditional expectation of the price given the cell:

\[
c_{i,j} = F_i^{-1}(j/N) = \exp(\mu_i + \sigma_i \Phi^{-1}(j/N)),
\]

\[
j = 0, \ldots, N,
\]

\[
s_{\text{price}}^{i,j} = \mathbb{E}[s_{\text{price}}^i \mid c_{i,j-1} \leq s_{\text{price}}^i \leq c_{i,j}]
\]

\[
= \int_{c_{i,j-1}}^{c_{i,j}} x f_i(x) dx / \int_{c_{i,j-1}}^{c_{i,j}} f_i(x) dx
\]

\[
= \frac{e^{\mu_i + \sigma_i^2/2} (\Phi(\Phi^{-1}(\frac{j}{N}) - \sigma_i) - \Phi(\Phi^{-1}(\frac{j-1}{N}) - \sigma_i))}{1/N},
\]

where \( F_i^{-1} \) and \( f_i \) denote the inverse cumulative distribution function (inverse cdf) and probability density function (pdf) of the price, and \( \Phi \) and \( \Phi^{-1} \) denote the cdf and inverse cdf of the standard normal distribution, respectively.

By so doing, the rewards associated to the discrete prices will give the correct expected rewards for the original continuous distribution conditionally to being in the cell.

As for the battery state, we partition the operating range into a uniform grid of 50 discrete levels.

C. Results

Figure 1 shows the near-optimal periodic policy returned by the value iteration algorithm, corresponding to the problem data of Table I. The periodic policy and the price cells \( c_{i,j} \) should be loaded into the battery controller and recomputed periodically, typically every day. Actions in real-time would be selected according to the charge level and the price-cell index hit by the spot price.

Figure 2a shows the evolution of the bounds \( \delta_i \) \((i = 0, \ldots, p-1)\) as a function of the iterations of the value-iteration algorithm. The spread of values among the \( \delta_i \)'s shows that the bounds established in the present paper are tighter than the bound \( V_0 - V_0^\pi \leq L_0 = 2\delta/(1-\gamma)^2 \), which our result would reduce to by setting \( \delta = \max_i \delta_i \).

Finally, Figure 2b depict the running times for computing a near-optimal periodic policy, as a function of the number of cores used in our parallel implementation. The results of our experiments are consistent with parallel computing theory, which predicts that the more cores we use, the less efficiency gains we should get [12].

VI. CONCLUSION

In this paper, we revisit the framework of periodic Markov Decision Processes, motivated by the use of cyclo-stationary models to approximate the expected return of reward processes in non-stationary environments subject to seasonal effects. We apply the approach to a grid-level storage control problem to obtain a near-optimal periodic policy, computed efficiently by combining various techniques proposed in the paper.

Although the numerical example demonstrates the effectiveness of the approach for a cycle of 24 periods of 1 hour, cycles defined on a much larger number of periods can be accommodated without losing tractability, for instance 1440 periods of 5 minutes for a daily cycle. This favorable property of the approach proposed in this paper comes from the choice of considering policies based on the greedy optimization of value functions.

The ability to accommodate short duration periods is especially important for battery storage control problems, for two related reasons. First, in contrast to hydro storage, the capacity of batteries is tiny. Profitable operations can thus come from increasing the frequency of profitable charge-discharge cycles during the day, in addition to providing regulation services for the system operator. Second, if wholesale electricity spot prices are updated every five minutes, it makes sense to have a control policy adapted to this time resolution. The spot
price fed into the battery storage control problem can then be interpreted as the expected average spot price over the next 5 minutes. The charging action determined by the model should be implemented at a uniform rate over the next 5 minutes. As the 5-min spot price is much more volatile than the hourly spot price, operating the battery at the 5-minute time scale is much more effective than operating it at the one-hour time scale.

The present work can be extended in several directions. The theoretical analysis could be extended to handle the case of an approximate evaluation of the Bellman iterations, and to handle approximate periodic value functions of a given approximation architecture. The value of a periodic policy used in a rolling-horizon fashion for appropriate classes of nonstationary problems could be studied in theory and numerically. The concept of using a cyclo-stationary reward process to approximate value functions at a terminal stage could be adapted to other approaches to stochastic optimization besides the Markov Decision Process framework.

At the application level, a more realistic Markov model for the price dynamics and for the battery dynamics of the grid-level storage control problem should be considered in a practical implementation. Better price predictions should be incorporated by calibrating the price model on the forward prices posted day-ahead by the system operator. A more realistic model for the battery should be implemented, to include efficiency losses and long-term capacity degradation. These refinements could be implemented according to the specification of the battery constructor, thanks to the versatility and tractability of the periodic Markov Decision Process framework.

**REFERENCES**


