Onlne Supplement for "Demand-Side Energy Management under Time-Varying Prices" by Liang, Deng, and Shen

Yong Liang
Department of Management Science and Engineering, School of Economics and Management, Tsinghua University, Beijing, China 100084, liangyong@sem.tsinghua.edu.cn

Tianhu Deng
Department of Industrial Engineering, Tsinghua University, Beijing, China 100084, deng13@mail.tsinghua.edu.cn

Zuo-Jun Max Shen
Department of Industrial Engineering and Operations Research and Department of Civil and Environmental Engineering, University of California, Berkeley, CA 94720, maxshen@berkeley.edu

E-Companion

EC.1. The Greedy Algorithm for CVBKP and a VBKP example for which the Greedy Heuristic Fails

We first present the greedy algorithm that solves for optimal solutions of the CVBKP, which is the linear relaxation of VBKP. The only condition for this algorithm to work is that $G_t(z)$ is lower semi-continuous and increasing in z. Define the efficiency ratio of an item as the ratio of its value over weight. The greedy algorithm picks items in the order of decreasing efficiency ratios, and it stops packing if $G'_{t-}(\tilde{z})$, the negative directional derivative at $\tilde{z}$, satisfies $G'_{t-}(\tilde{z}) \leq \frac{w_t(\tilde{z})}{q_t(\tilde{z})}$, where $\tilde{z}$ represents the current total weights of the packed items. $\frac{w_t(\tilde{z})}{q_t(\tilde{z})}$ is named the efficiency ratio of the item being packed.

The heuristic for VBKP stems from the above greedy algorithm for CVBKP. The heuristic also picks items in the order of decreasing efficiency ratios. After picking each item, the algorithm calculates the current total wight z_0, and stops picking If $G'_{t-}(\tilde{z})$ is greater than the efficiency ratio of the next unpicked item, where $\tilde{z}$ represents the current total weights of the packed items. For example, suppose $G_t(z)$ is linear in z with slope p_0, then we can simply select all items whose efficiency ratios are no less than p_0.

Nonetheless, the heuristic may fail for the VBKP even when $G_t(z)$ is convex increasing. As a counterexample, consider the case in which we have two items, with $v_1 = 11, u_1 = 1$ and $v_2 = 21.5, u_2 = 2$. Function $G_t(z)$ is defined as:

$$G_t(z) = \begin{cases} 10z & \text{if } z \leq 2 \\ 11.5z & \text{if } z > 2 \end{cases}$$

By ranking the efficiency ratios we pick item one first. Because picking item two does not change the objective value, we can either pick it or leave it, and the objective function values are 1 in both cases. However, the optimal solution is to pick item two only and the corresponding objective function value is 1.5. This example illustrates that the greedy algorithm that solves the CVBKP fails to work for problems with piecewise linear cost structures and mixed integer or integer decision variables.

EC.2. Proofs

We first provide the following Lemma before proving Proposition 1:

LEMMA EC.1. *Point-wise maximization (resp. minimization) or taking supremum (resp. infimum) preserves monotonicity.*

Proof of Lemma EC.1 Suppose $f:\mathbb{R}^u \times \mathbb{R}^v \to \mathbb{R}$ satisfies that $f(x,y) \geq f(z,y)$, $\forall x \leq z, \forall y$. We can show that function $g:\mathbb{R}^u \to \mathbb{R}$ defined as $g(x) = \sup_y f(x,y)$ is decreasing by contradiction.

Suppose g is not decreasing in x, that is, $\exists\ g(x) < g(z)$ for some $x \leq z$, then:

$$g(x) = f(x,\tilde{y}) = \sup_y f(x,y) < g(z) = \sup_y f(z,y) = f(z,\widehat{y}) \leq f(x,\widehat{y})$$

where the second inequality follows from the property of function $f(\cdot)$. However, $f(x,\tilde{y}) < f(x,\widehat{y})$ contradicts with $\tilde{y}$ being the minimizer of $\sup_y f(x,y) = g(x)$. Therefore, point-wise maximization (or taking supremum) over decreasing functions preserves monotonicity.

Similarly, suppose $f:\mathbb{R}^u \times \mathbb{R}^v \to \mathbb{R}$ satisfies $f(x,y) \leq f(z,y)$, $\forall x \leq z, \forall y$. We can show that function $g:\mathbb{R}^u \to \mathbb{R}$ defined as $g(x) = \sup_y f(x,y)$ is increasing by contradiction.

Suppose g is not increasing, that is, $\exists\ g(x) > g(z)$ for some $x \leq z$, then:

$$f(z,\widehat{y}) \geq f(x,\widehat{y}) = \sup_y f(x,y) = g(x) > g(z)$$

contradicts with $g(z) = \sup_y f(z,y) \geq f(z,\widehat{y})$. Therefore, point-wise maximization (or taking supremum) over increasing functions preserves monotonicity.

The proof for point-wise minimization or taking infimum preserves monotonicity is similar, thus is omitted here. $\square$

Based on Lemma EC.1, we can prove Proposition 1, which states the monotonicity of the value-to-go function, as follows:

Proof of Remark 1

(Part (a) & (b)) Firstly, we show part (a) and (b) by mathematical induction. To simplify the notation, assume we are in period 1, so the last period in the planning horizon is period T.

When we are in period T, recall the boundary condition: $J^*_{T+1}(S_{T+1}) \equiv 0$ for all S_{T+1}, thus, $V^*_{T+1}(S_T, X_T) \equiv 0 = \widetilde{V}^*_{T+1}(S^X_T, X_T) \equiv 0$ for all (S_T, X_T) and for all (S^X_T, X_t), thus we can say that V^*_{T+1} is increasing in $\boldsymbol{q}_T$ and $\boldsymbol{q}^X_T$. As a result, the optimal total cost in period T, given the state in period T being S_T is:

$$
\begin{aligned}
J^*_T(S_T) = \min_{X_T} & \left[C_T(S_T, X_T) + \langle \boldsymbol{q}^X_T, \boldsymbol{\pi}_T \rangle + \mathbb{E}_{\boldsymbol{d}_T} \left[\langle (\Lambda^I \boldsymbol{q}^X_T + \boldsymbol{d}_T - \boldsymbol{u}^c) \vee \boldsymbol{0}, \boldsymbol{\rho_T} \rangle | S_T, X_T \right] \right] \\
s.t. \quad & (1) - (4)
\end{aligned}
$$

Note that for any feasible X_T, the electricity cost term $C_T(S_T, X_T)$ is increasing in x^G_t, which is in turn increasing in $\boldsymbol{q}_T$. Because the second term is linear in $\boldsymbol{q}^X_T$, and $\boldsymbol{\pi}_T \succeq 0$, and $\boldsymbol{q}^X_T$ is increasing in $\boldsymbol{q}_T$, the second term is increasing in $\boldsymbol{q}_T$ as well. At last, because $\mathbb{E}_{\boldsymbol{d}_T} [\langle (\boldsymbol{q}^X_T + \boldsymbol{d}_T - \boldsymbol{u}^c) \vee \boldsymbol{0}, \boldsymbol{\rho_T} \rangle | S_T, X_T]$ is increasing in $\boldsymbol{q}^X_T$, the last term is increasing in $\boldsymbol{q}_T$. Then it follows from Lemma EC.1 that $J^*_T(S_T)$ is increasing in $\boldsymbol{q}_T$. It also follows from the fact of $\boldsymbol{q}_T$ being increasing in $\boldsymbol{q}^X_{T-1}$ that $J^*_T(S_T)$ is increasing in $\boldsymbol{q}^X_{T-1}$.

Then, suppose $J^*_{t+1}(S_{t+1})$ is increasing in $\boldsymbol{q}_{t+1}$. In order to calculate $J^*_t(S_t)$, we need to first calculate the value-to-go function:

$$
\begin{aligned}
\widetilde{V}^*_{t+1}(S^X_t, X_t) &= \mathbb{E}_{W_t} \left[J^*_{t+1}(S_{t+1}) | S_t, X_t \right] \\
&= \mathbb{E}_{W_t} \left[J^*_{t+1}\left(\boldsymbol{q}_{t+1}, \boldsymbol{b}_{t+1}, \mathcal{H}_{t+1} \right) | S_t, X_t \right] \\
&= \mathbb{E}_{W_t} \left[J^*_{t+1}\left(\left((\Lambda^I \boldsymbol{q}^X_t + \boldsymbol{d}_t) \wedge \boldsymbol{u}^c, (\boldsymbol{b}^X_t + \Theta^J(l_t; \boldsymbol{b}^X_t, \boldsymbol{r}^u, \boldsymbol{r}^c)) \wedge \boldsymbol{r}^c \right), \mathcal{H}_{t+1} \right) | \mathcal{H}_t, X_t \right]
\end{aligned}
$$

We first note that $\boldsymbol{q}_{t+1}$ is increasing in $\boldsymbol{q}^X_t$. Because $\boldsymbol{q}^X_t$ is increasing in $\boldsymbol{q}_t$, and taking the expectation of $J^*_{t+1}(\cdot)$ over W_t is essentially taking a convex combination of increasing functions of $\boldsymbol{q}^X_t$ (and $\boldsymbol{q}_t$), which preserves monotonicity, $\widetilde{V}^*_{t+1}(S^X_t, X_t)$ is increasing in $\boldsymbol{q}^X_t$ (and $\boldsymbol{q}_t$),that is, Part (b) holds.

Following from previous proof, $C_t(S_t, X_t)$ is increasing in $\boldsymbol{q}_t$, while $\langle \boldsymbol{q}^X_t, \boldsymbol{\pi}_t \rangle$ as well as $\mathbb{E}_{\boldsymbol{d}_t} [\langle (\Lambda^I \boldsymbol{q}^X_t + \boldsymbol{d}_t - \boldsymbol{u}^c) \vee \boldsymbol{0}, \boldsymbol{\rho_t} \rangle | S_t, X_t]$ are increasing in $\boldsymbol{q}^X_t$ (and $\boldsymbol{q}_t$). Therefore, the sum of the terms in the objective function are increasing in $\boldsymbol{q}_t$. Then, applying Lemma EC.1, we obtain that

$J_t^*(S_t)$ is increasing in $\boldsymbol{q}_t$, that is, Part (a) holds.

(Part (c) & (d)) Let $S_t = (\boldsymbol{q}_t, \boldsymbol{b}_t)$, $S_t' = (\boldsymbol{q}_t, \boldsymbol{b}_t + \boldsymbol{\epsilon})$. In order to prove part (c), we need to show $J_t^*(S_t) \geq J_t^*(S_t')$ for $\boldsymbol{\epsilon} > 0$. It can be verified that:

$$J_t^*(S_t') \leq J_t^*(S_t) - \langle \mathbf{1}, \boldsymbol{\epsilon} \rangle p_t(-\langle \mathbf{1}, \boldsymbol{\epsilon} \rangle) \leq J_t^*(S_t)$$

where the first inequality comes from the fact that selling the difference in storage is feasible, but not necessarily optimal. Note that $\langle \mathbf{1}, \boldsymbol{\epsilon} \rangle p_t(-\langle \mathbf{1}, \boldsymbol{\epsilon} \rangle)$ is the upper bound on the profit from selling the difference. The second inequality follows from the fact that $-\langle \mathbf{1}, \boldsymbol{\epsilon} \rangle p_t(-\langle \mathbf{1}, \boldsymbol{\epsilon} \rangle) \leq 0$. Therefore, Part (c) holds.

Part (d) can be verified via the following equation:

$$\widetilde{V}_{t+1}^*(S_t^X, X_t) = \mathbb{E}_{W_t}\left[J_{t+1}^*\left(\left((\Lambda^I \boldsymbol{q}_t^X + \boldsymbol{d}_t) \wedge \boldsymbol{u}^c, (\boldsymbol{b}_t^X + \Theta^J(l_t; \boldsymbol{b}_t^X, \boldsymbol{r}^u, \boldsymbol{r}^c)) \wedge \boldsymbol{r}^c \right), \mathcal{H}_{t+1} \right) | \mathcal{H}_t, X_t \right].$$

Since $J_{t+1}^*(S_{t+1})$ is decreasing in $\boldsymbol{b}_{t+1}$, which is increasing in $\boldsymbol{b}_t^X$, which is in turn increasing in $\boldsymbol{b}_t$, part (d) holds. $\square$

Proof of Proposition 1

Firstly, we argue that the complexity can be further reduced through a reduction in the dimension of the decision space. To start with, we first show that there is no incentive to reshape the storage levels of storage devices. In other words, when operating under normal conditions, the EMS chooses to charge (discharge) one storage device if and only if it decides not to discharge (charge) any other storage device. The incentive of reshaping the storage level profile only exists under the following condition: we expect to have massive charging (discharging) in the future that requires higher charging (discharging) rate, but those storage devices with high charging (discharging) rate are near full capacity (empty), hence we want to reshape the storage levels so that the near full capacity (empty) ones will have sufficient capacity remaining (energy available) to be charged (discharged). However, it is never optimal to reshape the storage level profile, simply because charging and discharging will cause loss of energy due to charging/discharging efficiency. For example, if the current solution is to charge device 1 with δ_1 ($\delta_1 < 0$ representing charging) and discharge storage 2 with δ_2 ($\delta_2 > 0$ representing discharging), then charge both with a total of $|\delta_1 + \delta_2|$ when $\delta_1 + \delta_2 < 0$ (or discharge both with a total of $\delta_1 + \delta_2$ when $\delta_1 + \delta_2 > 0$) is a strictly dominant solution.

In fact, the above discussed condition, under which reshaping is desired, can be avoided. Following from the same reasoning as presented in subsection 3.4, we can first determine the total amount of charge or discharge, then solve for the allocation problem in alignment with the operating goals of the storage devices. In other words, we can peel the allocation of the total charge or discharge among storage devices off the optimization problem, and the allocation is chosen by either a pre-determined function or a separated optimization module. For the ease of exposition, we perform change of variables by defining $y_t = \langle \mathbf{1}, \boldsymbol{x}_t^J \rangle$, and $\psi_t = \frac{1}{y_t} \langle \boldsymbol{x}_t^{J^*}, \boldsymbol{\zeta}_t \rangle$, where $\boldsymbol{x}_t^{J^*}$ is the optimal allocation given the total is y_t. Furthermore, it is not hard to check against Proposition 1 that $\psi_t \leq 0$.

We prove the first half of Proposition 1 by contradiction. Suppose $\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle_1 < \langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle_2$, and $G_t(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle_1) > G_t(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle_2)$. Denote the optimal discharging decisions for $G_t(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle_1)$ and $G_t(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle_2)$ as $(\boldsymbol{x}_{t1}^{J*}, x_{t1}^{N*})$ and $(\boldsymbol{x}_{t2}^{J*}, x_{t2}^{N*})$, respectively. Next, construct a feasible solution $(\boldsymbol{x}_{t1}^J, x_{t1}^N)$ for input $\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle_1$ as follows:

$$\begin{cases} \boldsymbol{x}_{t1}^J &= \boldsymbol{x}_{t2}^{J*} \\ x_{t1}^N &= x_{t2}^{N*} - \langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle_2 + \langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle_1 \end{cases}$$

then clearly:

$$\begin{aligned} G_t(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle_1) &\leq x_{t1}^N p_t\left(x_{t1}^N\right) + \langle \boldsymbol{\zeta}_t, \boldsymbol{b}_t - \boldsymbol{x}_{t1}^J \rangle \\ &\leq x_{t2}^{N*} p_t\left(x_{t2}^{N*}\right) + \langle \boldsymbol{\zeta}_t, \boldsymbol{b}_t - \boldsymbol{x}_{t2}^{J*} \rangle \\ &= G_t(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle_2) \\ &< G_t(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle_1) \qquad \text{(contradiction)} \end{aligned}$$

where the first inequality holds because $(\boldsymbol{x}_{t1}^J, x_{t1}^N)$ is feasible but not necessarily optimal, and the second inequality holds because p_t is increasing and $x_{t1}^N < x_{t2}^{N*}$. The result conflicts with the assumption, hence G_t must be increasing when p_t is increasing.

Next, we proceed to prove that $G_t(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle)$ is convex increasing in $\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle$ under the two conditions described in Proposition 1:

[Part (a)]: If $p_t(\cdot)$ is an increasing stepwise function, then there exists some $j \in \{1, 2, \ldots, l\}$, such that $p_t^{j-1} \leq -\psi_t < p_t^j$. Then:

(1) if $\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - y_t < b^j$, then there exists $0 < \delta_y \leq b_t^j + y_t - \langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle$, such that lowering y_t by δ_y (equivalently, charging more or discharging less) reduces the objective of (7) by:

$$\begin{aligned}
&\psi_t \delta_y + \left(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - y_t + \delta_y \right) p_t \left(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - y_t + \delta_y \right) \\
&\qquad - \left(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - y_t \right) p_t \left(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - y_t \right) \leq \psi_t \delta_y + \delta_y p_t^{j-1} \leq 0 \ ,
\end{aligned}$$

where the first inequality results from $p_t(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - y_t) \leq p_t^{j-1}$, and the second inequality follows from $p_t^{j-1} \leq -\psi_t$. Thus, reducing y_t whenever $\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - y_t < b^j$ decreases (7);

(2) if $\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - y_t \geq b^j$, then there exists $0 < \delta_y \leq (\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - b_t^j - y_t)$, such that increasing y_t by $\delta_y > 0$ (equivalently, discharging more or charging less) decreases the objective of (7) by:

$$\begin{aligned}
&-\psi_t \delta_y + \left(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - y_t - \delta_y \right) p_t \left(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - y_t - \delta_y \right) \\
&\qquad - \left(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - y_t \right) p_t \left(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - y_t \right) \leq -\psi_t \delta_y - \delta_y p_t^j < 0
\end{aligned}$$

where the first inequality holds because $p_t(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - y_t) \geq p_t^j$, and the second inequality comes from $-\psi_t \leq p_t^j$. Thus, increasing y_t whenever $\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - y_t \geq b^j$ decreases (7).

Therefore, setting y_t so as to let $\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - y_t$ be as close to b_t^j as possible minimizes the objective of (7). Note that it may not be possible to have $\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - y_t$ equals to b_t^j.

In addition, increasing $\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle$ increases $G_t(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle)$. Plugging in the above results, the one-dimensional directional derivative of $G_t(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle)$ satisfies:

$$\begin{aligned}
&\lim_{\epsilon \to 0} \frac{G_t(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle + \epsilon) - G_t(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle)}{\epsilon} \\
&= \begin{cases}
p_t^k, & \text{if } \langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - \min\limits_{y \in \mathcal{Y}_t} y < b_t^j, \text{ for } k \text{ such that } b_t^k \leq \langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - \min\limits_{y \in \mathcal{Y}_t} y \leq b_t^{k+1} \\
-\psi_t, & \text{if } \langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - \max\limits_{y \in \mathcal{Y}_t} y \leq b_t^j \leq \langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - \min\limits_{y \in \mathcal{Y}_t} y \\
p_t^k, & \text{if } \langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - \max\limits_{y \in \mathcal{Y}_t} y > b_t^j, \text{ for } k \text{ such that } b_t^k \leq \langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - \max\limits_{y \in \mathcal{Y}_t} y \leq b_t^{k+1}
\end{cases}
\end{aligned}$$

where it can be verified that the directional derivative is increasing. Since $G_t(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle)$ is continuous and is a real function, it is convex in $\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle$. Moreover, since $p_t(\cdot)$ is increasing, $G_t(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle)$ is increasing in $\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle$.

[Part (b)]: if $p_t(\cdot)$ is a twice-differentiable convex increasing function, and $p_t''(x_t^G) = 0$ for all $x_t^G < 0$, then first of all:

$$\begin{aligned}
&\frac{d^2}{dy_t^2} \left(p_t \left(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - y_t \right) \cdot \left(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - y_t \right) \right) \\
&\qquad = 2p_t' \left(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - y_t \right) + \left(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - y_t \right) p_t'' \left(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - y_t \right)
\end{aligned}$$

Then the objective of (7) is convex, because $p_t(x_t^G)$ is increasing convex, and $p_t''(x_t^G) = 0$ when $x_t^G < 0$. From the KKT conditions, we can verify that the optimal solution y_t^*, which we denote by $y_t^*(\boldsymbol{x}_t^I)$, satisfies:

$$y_t^*(\boldsymbol{x}_t^I) = \begin{cases} \min\limits_{y \in \mathcal{Y}_t} y & \text{if } \min\limits_{y \in \mathcal{Y}_t} y > y_t^0 \left(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle \right) \\ y_t^0 \left(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle \right) & \text{if } \min\limits_{y \in \mathcal{Y}_t} y \le y_t^0 \left(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle \right) \le \max\limits_{y \in \mathcal{Y}_t} y \\ \max\limits_{y \in \mathcal{Y}_t} y & \text{if } \max\limits_{y \in \mathcal{Y}_t} y < y_t^0 \left(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle \right) \end{cases},$$

where $y_t^0 \left(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle \right)$ is the solution of equation

$$\left(y - \langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle \right) p_t' \left(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - y \right) - p_t \left(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - y \right) = \psi_t.$$

Denote $\min\limits_{y \in \mathcal{Y}_t} y$ as $\underline{y}$ and $\max\limits_{y \in \mathcal{Y}_t} y$ as $\overline{y}$, Then, applying the Envelope theorem, we can calculate the derivative of $G_t(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle)$ over $\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle$ as follows:

$$\begin{aligned}
&\frac{\partial}{\partial \langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle} G_t(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle) \\
&\quad = \begin{cases} (\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - \underline{y}) p_t'(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - \underline{y}) + p_t(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - \underline{y}) & \text{if } \underline{y} > y_t^0 \left(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle \right) \\ -\psi_t & \text{if } \underline{y} \le y_t^0 \left(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle \right) \le \overline{y} \\ (\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - \overline{y}) p_t'(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - \overline{y}) + p_t(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - \overline{y}) & \text{if } \overline{y} < y_t^0 \left(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle \right) \end{cases}
\end{aligned}$$

Since in all cases the derivative is non-negative, $G_t(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle)$ is increasing in $\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle$. Furthermore, based on this result, one can verify that the second derivative is non-negative. Therefore, since

$G_t(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle)$ is a real function, it is convex, given $p_t(\cdot)$ is a twice-differentiable convex increasing function with $p_t''(x) = 0$ for all $x \leq 0$.

It is worth noting that, increasing condition of $p_t(x_t^G)$ alone does not suffice to provide convex $G_t(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle)$. For example, assuming that $p_t(x_t^G)$ is twice-differentiable, and let y^* be the optimal solution to (7), then by the Envelope theorem, the second derivative of $G_t(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle)$ can be written as:

$$\frac{\partial^2}{\partial \langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle^2} G_t(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle) = 2p_t'(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - y^*) + (\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - y^*) p_t''(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - y^*),$$

which implies that the convexity of $G_t(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle)$ depends strongly on the shape of $p_t(x_t^G)$. As a final remark, it is noteworthy that part (b) cannot be readily obtained using the preservation result of convexity under minimization as $p_t(x_t^G) x_t^G$ is not necessarily convex. □

Proof of Corollary 1

Let $p_t^{-1} : \mathbb{R} \to 2^{\mathbb{R}}$ denote the inverse mapping of the average-price function $p_t(\cdot)$. Since $p_t(\cdot)$ is continuously increasing, $p_t^{-1}(p)$ is closed and convex, that is, $p_t^{-1}(p)$ defines a closed interval on $\mathbb{R}$ for any given average price. Without loss of generality, let $p_t^{-1}(-\zeta_t) = [\underline{h}, \overline{h}]$, and let the discharging decision be x.

If $\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - x < \underline{h}$, lowering x by $0 < \delta_x \leq (\underline{h} + x - \langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle)$ changes the objective function (7) by:

$$\begin{aligned}
&\zeta_t \delta_x + (\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - x + \delta_x) p_t(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - x + \delta_x) - (\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - x) p_t(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - x + \delta_x) \\
&\quad \leq \zeta_t \delta_x + (-\zeta_t) \delta_x = 0
\end{aligned}$$

Similarly, if $\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - x > \overline{h}$, then increasing x by $0 < \delta_x \leq \langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - x - \overline{h}$ changes the objective function (7) by:

$$\begin{aligned}
&-\zeta_t \delta_x + (\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - x - \delta_x) p_t(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - x - \delta_x) - (\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - x) p_t(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - x + \delta_x) \\
&\quad \leq -\zeta_t \delta_x - \zeta_t(-\delta_x) = 0
\end{aligned}$$

At last, when $(\langle \boldsymbol{\iota}^I, \boldsymbol{x}_t^I \rangle - x) \in [\underline{h}, \overline{h}]$, x obviously minimizes the objective function (7), and that completes the proof. □

Proof of Remark 2

We first show that the decision problem of the variable budget knapsack problem is NP-complete by proving that it is in the class of NP, and by showing that the knapsack problem (KP) reduces to variable budget knapsack problems in polynomial time.

a) We first show that the problem is in the class of NP. Consider the decision version of the problem: whether a certain payoff, defined as the total value of picked items subtracting the cost of budget, can be achieve while satisfying budget constraint.

Under the certificate-based (or verifier-based) definition of NP, because the certificate consists of a realization of the decision on picking the items x^I and a realization of the budget decision z, which less than the sum of weights of all items, the certificate is polynomial in the size of input, which is determined by the items and their weights. Since (1) the certificate checking algorithm that verifies the sum of electricity demands of task is with $x^I_{i,t} = 1$ being less than or equal to the budget takes $O(|\mathcal{I}|)$ operations, and (2) the algorithm that calculates the sum of the valuations of these items (which takes $O(|\mathcal{I}|)$) subtracting the cost of the budget takes polynomial time because the cost of budget can also be evaluated in polynomial time, the whole verification takes polynomial time, hence the problem is in the class of NP.

b) To show the decision form of VBKP is NP-complete, we show that the KP reduces to the VBKP in polynomial time. Consider an arbitrary KP:

$$
\begin{aligned}
(\text{P1}): \quad \max \quad & \sum_{i=1}^{\mathcal{I}} w_i x_i \\
s.t. \quad & \sum_{i=1}^{\mathcal{I}} q_i x_i \leq Z \\
& x_i \in \{0,1\}, \qquad \forall i \in \mathcal{I}
\end{aligned}
$$

where $\mathcal{I}$ is the set of candidate items. $\{w\}_{i\in\mathcal{I}}$ and $\{q\}_{i\in\mathcal{I}}$ are the values and weights of the items, respectively. Z is the capacity of the "knapsack". We construct a corresponding instance of the VBKP as follows. Let the convex cost function $G(z)$ take the ensuing form:

$$
G(z) = \begin{cases} 0 & \text{if } z \leq Z \\ \infty & \text{if } z > Z \end{cases}
$$

Next, setting the weights and benefits of the items to be the same in this problem as in KP, we have the following VBKP:

$$
\begin{aligned}
(\mathrm{P2}): \quad \max \quad & \sum_{i=1}^{\mathcal{I}} w_i x_i - G(z) \\
s.t. \quad & \sum_{i=1}^{\mathcal{I}} q_i x_i \le z \\
& x_i \in \{0,1\}, \qquad \forall i \in \mathcal{I} \\
& z \ge 0
\end{aligned}
$$

It remains to show that problem (P2) is equivalent to problem (P1). In (P2), the budget z never exceeds Z, because if $z > Z$, and whenever the budget constraint $\sum_{i=1}^{\mathcal{I}} q_i x_i \le Z$ of problem (P1) is violated, the objective of (P2) goes negative infinity. When the budget is less than or equal to Z, $G(z)$ equals to zero, hence the objective functions of problem (P2) and problem (P1) are the same. Therefore, it follows that the optimal solution of problem (P2) solves problem (P1).

Lastly, since the construction of problem (P2) takes $O(|\mathcal{I}|)$ time, we conclude that the KP reduces in polynomial time to VBKP. As a result, the decision problem of the VBKP is NP-complete.

Then, since it is obvious that the VBKP is reducible to VBPCKP, the VBPCKP is NP-hard, and that completes the proof. □

Proof of Proposition 2

The proof of the first half of Proposition 2 follows directly from the *Principle of Optimality*, while the proof of the complexity of the algorithm being pseudo polynomial when weights can be scaled into integers is rooted in the proof of the pseudopolynomial time algorithm for knapsack problems. Due to its straightforwardness, the detailed proof is omitted. □

Proof of Proposition 3

According to the definition of policies μ^* and μ, decisions $X_t^{\mu^*}(S_t)$ and $X_t^{\mu}(S_t)$ are the one-step optimal solutions to problem (**P**) and the approximate problem (**P′**) when we are at state S_t. Moreover, we have:

$$E_t\big(S_t, X_t^{\mu}(S_t)\big) + \overline{\Gamma}_{t+1}\big(S_t, X_t^{\mu}(S_t)\big) \leq E_t\big(S_t, X_t^{\mu^*}(S_t)\big) + \overline{\Gamma}_{t+1}\big(S_t, X_t^{\mu^*}(S_t)\big), \tag{EC.1}$$

because policy μ is optimal for the approximate problem (**P′**). And similarly:

$$E_t\big(S_t, X_t^{\mu}(S_t)\big) + \Gamma_{t+1}\big(S_t, X_t^{\mu}(S_t)\big) \geq E_t\big(S_t, X_t^{\mu^*}(S_t)\big) + \Gamma_{t+1}\big(S_t, X_t^{\mu^*}(S_t)\big). \tag{EC.2}$$

Re-organizing inequalities (EC.1) and add to both sides the term $\Gamma_{t+1}\big(S_t, X_t^{\mu}(S_t)\big) - \Gamma_{t+1}\big(S_t, X_t^{\mu^*}(S_t)\big)$, we obtain:

$$\begin{aligned}
&E_t\big(S_t, X_t^{\mu}(S_t)\big) + \Gamma_{t+1}\big(S_t, X_t^{\mu}(S_t)\big) - \Big(E_t\big(S_t, X_t^{\mu^*}(S_t)\big) + \Gamma_{t+1}\big(S_t, X_t^{\mu^*}(S_t)\big)\Big)\\
&= J^{\mu}\big(S_t\big) - J_t^{\mu^*}\big(S_t\big)\\
&\leq \overline{\Gamma}_{t+1}\big(S_t, X_t^{\mu^*}(S_t)\big) - \Gamma_{t+1}\big(S_t, X_t^{\mu^*}(S_t)\big) - \overline{\Gamma}_{t+1}\big(S_t, X_t^{\mu}(S_t)\big) + \Gamma_{t+1}\big(S_t, X_t^{\mu}(S_t)\big)\\
&\leq 2\|\Gamma_{t+1}\big(S_t, X_t\big) - \overline{\Gamma}_{t+1}\big(S_t, X_t\big)\|_\infty
\end{aligned} \tag{EC.3}$$

Similarly, subtracting inequality (EC.1) from (EC.2) and reorganizing the terms yields the following inequality:

$$\overline{\Gamma}_{t+1}\big(S_t, X_t^{\mu^*}(S_t)\big) - \overline{\Gamma}_{t+1}\big(S_t, X_t^{\mu}(S_t)\big) \geq \Gamma_{t+1}\big(S_t, X_t^{\mu^*}(S_t)\big) - \Gamma_{t+1}\big(S_t, X_t^{\mu}(S_t)\big) \tag{EC.4}$$

Then, we can tighten the bounds obtained from inequality (EC.3) based on the signs of both sides of inequality (EC.4):

- if $\Gamma_{t+1}\big(S_t, X_t^{\mu^*}(S_t)\big) - \Gamma_{t+1}\big(S_t, X_t^{\mu}(S_t)\big) \geq 0$, then the upper bound of $J^{\mu}\big(S_t\big) - J_t^{\mu^*}\big(S_t\big)$ can be tightened by the following inequality:

$$J^{\mu}\big(S_t\big) - J_t^{\mu^*}\big(S_t\big) \leq \|\overline{\Gamma}_{t+1}\big(S_t, X_t^{\mu^*}(S_t)\big) - \overline{\Gamma}_{t+1}\big(S_t, X_t^{\mu}(S_t)\big)\|_\infty \quad ;$$

- if $\overline{\Gamma}_{t+1}\big(S_t, X_t^{\mu^*}(S_t)\big) - \overline{\Gamma}_{t+1}\big(S_t, X_t^{\mu}(S_t)\big) \leq 0$, then the upper bound of $J^{\mu}\big(S_t\big) - J_t^{\mu^*}\big(S_t\big)$ can be tightened by the following inequality:

$$J^{\mu}\big(S_t\big) - J_t^{\mu^*}\big(S_t\big) \leq \|\Gamma_{t+1}\big(S_t, X_t^{\mu^*}(S_t)\big) - \Gamma_{t+1}\big(S_t, X_t^{\mu}(S_t)\big)\big)\|_\infty \quad ;$$

- otherwise, the upper bound of $J^{\mu}\big(S_t\big) - J_t^{\mu^*}\big(S_t\big)$ can be rewritten as:

$$\begin{aligned}
J^{\mu}\big(S_t\big) - J_t^{\mu^*}\big(S_t\big) \leq\ & \|\Gamma_{t+1}\big(S_t, X_t^{\mu^*}(S_t)\big) - \Gamma_{t+1}\big(S_t, X_t^{\mu}(S_t)\big)\big)\|_\infty\\
&+ \|\overline{\Gamma}_{t+1}\big(S_t, X_t^{\mu^*}(S_t)\big) - \overline{\Gamma}_{t+1}\big(S_t, X_t^{\mu}(S_t)\big)\|_\infty \quad .\square
\end{aligned}$$

EC.3. Updating Rule and Exploration Rule

In this section we illustrate the updating rule used in the proposed solution approach. Suppose we are in iteration m and period τ with $t < \tau \leq t+T-1$. The latest coefficient tuple associated with the value-to-go approximation is then $\Theta_t^{(m)}$. We need to first solve problem ($\mathbf{P}'$) with $\Theta_\tau^{(m)}$. Next, we update the coefficient tuple, denoted by $\Theta_{\tau-1}^{(m+1)}$. At last, we proceed to the next period according to a sample path indexed by m, $\overline{W}_\tau^{(m)}$. We first discuss the coefficients updating rule.

Essentially, the sample-path-based approach for estimating coefficients $\Theta_t^{(m)}$ associated with the approximate problem is rooted in stochastic approximation theory (Robbins and Monro 1951). The theoretical foundation is later enriched by Kiefer and Wolfowitz (1952), Blum (1954), Dvoretzky (1956). In this paper, the coefficient updating procedure follows standard stochastic gradient literature, see for example Kushner and Yin (1997). Recall that $\Theta_t^{(m)} \stackrel{\text{def}}{=} (\boldsymbol{\theta}_t^{(m)}, \boldsymbol{\zeta}_t^{(m)}, \eta_t^{(m)})$. Define the loss function $\mathcal{L}(\Theta_t^{(m)})$ as follows:

$$\mathcal{L}(\Theta_t^{(m)}) \stackrel{\text{def}}{=} \frac{1}{2}\mathbb{E}\left[\overline{\Gamma}_{t+1}(S_t^X;\Theta_t^{(m)}) - \widehat{\Gamma}_{t+1}\right]^2 ,$$

where $\overline{\Gamma}_{t+1}(S_t^X;\Theta_t^{(m)})$ is the sum of the approximate discomfort from lost arrivals $\overline{L}_t(\boldsymbol{q}_t^X;\boldsymbol{\theta}''_t,\eta''_t)$ and the value-to-go approximation $\overline{V}_{t+1}(\boldsymbol{q}_t^X,\boldsymbol{b}_t^X;\boldsymbol{\theta}'_t,\boldsymbol{\zeta}_t,\eta'_t)$, while $\widehat{\Gamma}_{t+1}$ is the sum of the realized discomfort from lost arrival $\widehat{L}_t$ and the realized value-to-go $\widehat{V}_{t+1}$ that result from $X_t^{(m)}$ following a sample path. The goal of updating the coefficients is to minimize $\mathcal{L}(\Theta_t^{(m)})$, and we rely on stochastic gradient methods.

There exists a relatively strong temporal correlation in the context of demand-side energy management. To propagate more efficiently the effect of taking a specific decision at later periods back to the value of being at a particular post decision state, we apply *Temporal Difference* learning (Sutton and Barto 1998, Powell 2007). Specifically, for all $t \in \mathcal{T}$:

$$\begin{cases} \boldsymbol{\theta}_t^{(m+1)} = \max\left\{0, \boldsymbol{\theta}_t^{(m)} - \gamma_m \sum\limits_{\{\tau\geq t:\ \tau\in\mathcal{T}\}} \lambda^{\tau-t} D_\tau \boldsymbol{q}_t^X\right\} \\ \boldsymbol{\zeta}_t^{(m+1)} = \min\left\{0, \boldsymbol{\zeta}_t^{(m)} - \gamma_m \sum\limits_{\{\tau\geq t:\ \tau\in\mathcal{T}\}} \lambda^{\tau-t} D_\tau \boldsymbol{b}_t^X\right\} \\ \eta_t^{(m+1)} = \eta_t^{(m)} - \gamma_m \sum\limits_{\{\tau\geq t:\ \tau\in\mathcal{T}\}} \lambda^{\tau-t} D_\tau \end{cases} , \qquad \text{(EC.5)}$$

Run	Price ($)		Generation (kWh)		Arrival Rates		Storage (kWh)		Discomfort ($/kWh)		
	$\bar{p}_{avg}$	$\bar{p}_{spd}$	$\bar{g}_{avg}$	$\bar{g}_{spd}$	d_{avg}	d_{spd}	r_c	$r_d(=r_u)$	π'	ρ	π''
1	0.3	0.25	0	0	0.55	0.45	4	2	0.05	0.15	50
2	0.3	0.25	0	0	0.55	0.45	4	2	0.05	0.15	**0.75**
3	0.3	0.25	0	0	0.55	0.45	4	2	0.05	**0.4**	0.75
4	0.3	0.25	0	0	0.55	0.45	4	2	**0.25**	**0.5**	0.75
5	0.3	0.25	**0.75**	**0.5**	0.55	0.45	4	2	0.05	0.4	0.75
6	0.3	0.25	**1.5**	**0.5**	0.55	0.45	4	2	0.05	0.4	0.75
7	0.3	0.25	0	0	0.55	**0.15**	4	2	0.05	0.4	0.75
8	0.3	0.25	0	0	**0.3**	0.15	4	2	0.05	0.4	0.75
9	0.3	0.25	0	0	0.55	0.45	**8**	**4**	0.05	0.4	0.75

Table EC.1 Summary of the Settings of the Representative Runs

where D_τ is the typical temporal difference term, defined as follows:

$$D_\tau \equiv \overline{\Gamma}_\tau(S_\tau^X;\Theta_\tau^{(m)}) - [\widehat{L}_{\tau-1} + C_\tau(S_\tau, X_\tau) + U_\tau(S_\tau, X_\tau) - \overline{L}_\tau(\boldsymbol{q}_\tau^X;\boldsymbol{\theta}_\tau''^{(m)}, \eta_\tau''^{(m)}) + \overline{\Gamma}_{\tau+1}(S_{\tau+1}^X;\Theta_{\tau+1}^{(m)})].$$

We use the *Harmonic stepsize* rule, $\gamma_m = \gamma/(\gamma + m - 1)$. Appropriate γ is chosen to ensure convergence. $\lambda \in (0,1)$ is an artificial factor that discounts temporal differences further along sample paths. The stepsize satisfies the basic conditions for convergence of the stochastic gradient method (Kushner and Yin 1997). Moreover, a modified mixed exploration strategy is applied. The rate of exploration is a piece-wise linear function $\rho(m)$, where m is the iteration index. We set $\rho(m)$ in a way such that the algorithm explores more states at early iterations, then exploits the collected information at later iterations.

EC.4. Numerical Study Settings and Results: Tables and Figures

Table EC.1 summarizes some representative parameter settings, while Table EC.2 summarizes the estimated total costs of the four policies on the these runs. Figure EC.1 provides the mean absolute percentage error of ADP, MYO and TRD against EXDP.

References

Blum, Julius R. 1954. Multidimensional stochastic approximation methods. *Annals of Mathematical Statistics*. 737–744.

Run	EXDP			ADP			MYO			TRD	
	Cost	Discomfort	Total	Cost	Discomfort	Total	Cost	Discomfort	Total	Cost	Total
1	-0.6	5.0	4.3	0.3	4.1	4.4	0.2	4.8	5.0	13.3	13.3
2	1.1	3.4	4.6	1.3	3.3	4.6	0.8	4.5	5.3	13.2	13.2
3	1.8	6.2	8.0	2.3	5.9	8.1	1.6	7.7	9.3	12.9	12.9
4	9.3	2.2	11.5	10.0	1.7	11.6	5.0	10.0	15.0	13.0	13.0
5	0.2	6.4	6.6	1.1	6.0	7.1	-0.2	7.9	7.7	11.4	11.4
6	-2.2	6.6	4.4	-0.9	6.0	5.0	-2.7	8.0	5.4	9.0	9.0
7	2.0	6.2	8.2	3.1	5.3	8.4	1.1	8.0	9.2	12.4	12.4
8	1.6	3.2	4.8	3.6	1.8	5.5	0.9	5.2	6.1	7.7	7.7
9	0.8	6.3	7.2	1.5	5.8	7.3	2.0	7.7	9.6	13.1	13.1

Table EC.2 Summary of Results of the Representative Runs (Units: $)

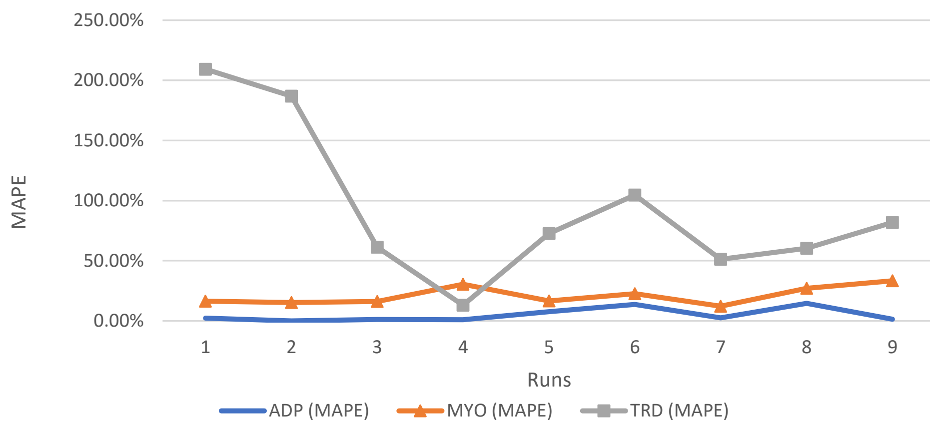

Figure EC.1 Mean Absolute Percentage Error (MAPE) of ADP, MYO and TRD

Dvoretzky, Aryeh. 1956. On stochastic approximation. *I . Berkeley and Los Angeles: University of California*. 39–55.

Kiefer, J., J. Wolfowitz. 1952. Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics* **23**(3) 462–466.

Kushner, Harold J., G. George Yin. 1997. *Stochastic approximation and recursive algorithm and applications*. Springer.

Powell, Warren B. 2007. *Approximate Dynamic Programming: Solving the curses of dimensionality*, vol. 703. John Wiley & Sons.

Robbins, Herbert, Sutton Monro. 1951. A stochastic approximation method. *The annals of mathematical statistics* 400–407.

Sutton, Richard S, Andrew G Barto. 1998. *Reinforcement learning: An introduction*, vol. 1. MIT press Cambridge.