Off-policy Q-learning: set-point design for optimizing dual-rate rougher flotation operational processes

DOI:
10.1109/TIE.2017.2760245

Document Version
Accepted author manuscript

To appear in IEEE Transactions on Industrial Electronics (17-TIE-1057)

Off-policy Q-learning: set-point design for optimizing dual-rate rougher flotation operational processes

J. Li, T.Y. Chai, F. Lewis, J. Fan, Z. Ding and J. Ding

***Abstract*—Rougher flotation, composed of unit processes operating at a fast time scale and economic performance measurements known as operational indices measured at a slower time scale, is very basic and the first concentration stage for flotation plants. Optimizing operational process for rougher flotation circuits is extremely important due to high economic profit arising from the optimality of operational indices. This paper presents a novel off-policy Q-learning method to learn the optimal solution to rougher flotation operational processes without the knowledge of dynamics of unit processes and operational indices. To this end, first, the optimal operational control (OOC) for dual-rate rougher flotation processes is formulated. Second, H∞ tracking control problem is developed to optimally prescribe the set-points for the rougher flotation processes. Then, a zero-sum game off-policy Q-learning algorithm is proposed to find the optimal set-points by using measured data. Finally, simulation experiments are employed to show the effectiveness of the proposed method.**

***Index Terms*—Rougher flotation, Operational optimization, Q-learning, Zero-sum game, H∞ tracking control.**

I. Introduction

ROUGHER flotation processes connected by several flotation cells in a serial structure are composed of flotation cell control processes, operating at a fast time scale and economic benefit measurement known as operational indices calculated at a slow time scale. Roughing is the first and basic stage of flotation processes which also include cleaning, scavenger, regrinding and classification circuits, and its primary objective is to get the valuable mineral recovery as much as possible subject to great energy consuming. Optimizing economic benefit by trading off increase of recovery rate and decrease of energy consumption, or keeping it following a desired value by forcing concentrate grades and tail grades to proper trajectories at a fast time scale is a key issue. It is desired to find optimal set-points for rougher flotation circuits to ensure that the economic operational indices stay within their target ranges or at their desired target values like most of industrial operational processes [1-3].

Most existing literature about rougher flotation processes mainly focused on control and optimization of the basic device loop, i.e., concentrate grades and tail grades are controlled or are optimally controlled to desired set-points [4-15]. Optimal control [4], adaptive control [5-7], multivariable control [8], expert control [9] and multivariable predictive control [10, 11] were developed and adopted in rougher flotation circuits. Based on the previous work of [11], [12], and [13], [14] presented multivariable model based predictive control (MPC) strategies with consideration of the intermediate cell grade estimates for a rougher circuit, such that concentrate grades and tail grades can follow the desired values by the optimal approach. [15] implemented dynamic programming to minimize the Cu tailing grade in each cell given a final Cu concentrate grade by considering phenomenological models.

Actually, optimal control for achieving regulation and optimization of the local processes and an optimization procedure used to generate the set-points that maximize the economic performance function are both involved in a general industrial process control scheme [16]. [17, 18] presented real-time optimization (RTO) based set-point compensation method by integrating RTO with MPC technique and applied it into a rougher flotation process. [19] proposed set-point compensation by designing feedback control strategy to achieve desired operational index. Note that the set-point compensation methods [17-19] require the dynamics of the rougher flotation process to be accurately known, which, in general, is very difficult to obtain in practical flotation industrial environment since there exist disturbance, chemical and physical reactions between ore and chemicals [1, 2]. Therefore, the difficulty of knowing exact models of rougher flotation processes motivates us to attempt data-driven optimal control research for achieving optimality of dual-rate rougher flotation processes.

Neural network methods and case-based-reasoning intelligent control methods have been developed to design or correct prescribed optimal set-points for large-scale complex industrial processes without requiring complete knowledge of the system dynamics [1, 2, 20, 21]. Q-learning, one of reinforcement learning (RL) schemes, is also called action-dependent heuristic dynamic programming (ADHDP) and can be used to solve optimal control problems [22-24]. One of the strengths of Q-learning is that it has a capability of evaluating utility and updating control policy without requiring models of the environment [22-24]. Especially off-policy Q-learning including behavior policies and target policies is more practical and efficient technique compared with on-policy Q-learning [22-24] for dealing with optimal control problem as it can generate data of systems for enriching data exploration while the target policies are updated to find the optimal policies but not to be employed to systems [25, 26]. Particularly, adding probing noise into the behavior policy does not produce bias of solution when implementing policy evaluation [25]. Therefore,

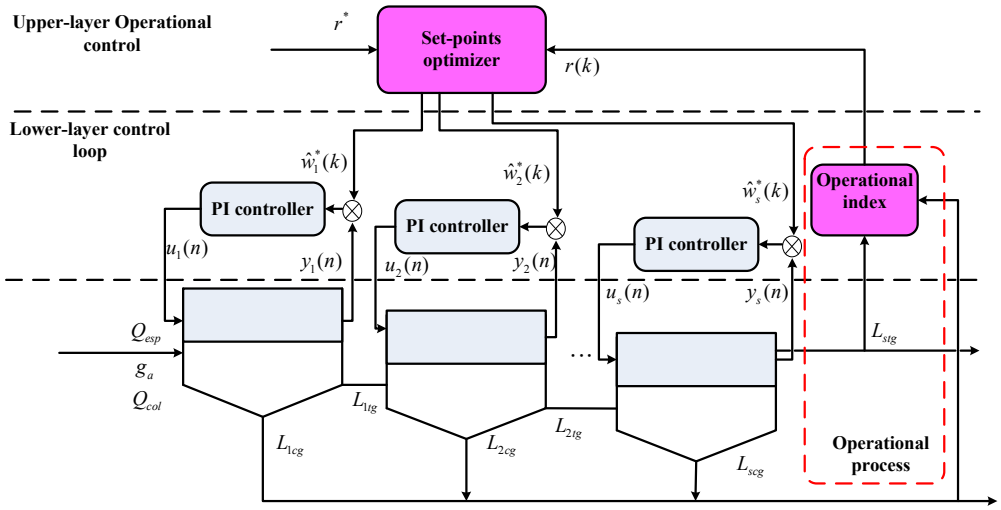

Fig. 1. Two-layer architecture for optimal operational control of rougher flotation processes

the advantages of off-policy Q-learning also motivate us to attempt off-policy Q-learning research for optimal control of rougher flotation processes. However, to our knowledge, how to design off-policy Q-learning algorithm and use it to OOC problem of rougher flotation processes with completely unknown dynamics of controlled unit processes and unknown functional dependence of economic operational index generation have not been developed yet. The off-policy Q-learning algorithm design to achieve OOC for rougher flotation circuits is challenging due to dual rates and nonlinearities existed in complex rougher flotation circuits.

In this paper, an off-policy Q-learning algorithm is presented to learn optimal set-points for rougher flotation processes using only measured data, such that economic benefit tracks the desired value by forcing the concentrate grades and tail grades of the lower-layer control loops to the set-points. The main contributions of this paper are summarized below.

1. Different from model-based OOC methods [17-19], a novel off-policy Q-learning algorithm is presented to approximate the solutions to the discrete-time (DT) game Bellman equation and learn the optimal set-points for the rougher flotation circuits without requiring any knowledge of the system dynamics.

2. By contrast to the neural network methods and intelligent control methods [1, 2, 21, 22], the proposed Q-learning algorithm can approximate the optimal set-point by interacting with the rougher flotation environment, and it is an off-policy RL approach wherein the target policies are evaluated and updated until convergence while they do not need to be applied to systems. In this sense, this off-policy Q-learning is more practical and easy to be realized for complex industrial applications.

The rest of paper is arranged as follows. Section II formulates OOC problem for rougher flotation processes. Section III constructs H∞ tracking problems for the OOC problem. Section IV presents an on-policy Q-learning method for achieving Nash equilibrium by using zero-sum game method and stability of the optimal solution. In Section V, an off-policy Q-learning algorithm is proposed to learn optimal set-points using data generated from rougher flotation processes. Section VI verifies the effectiveness of the proposed method for rougher flotation processes. Conclusions are stated in Section VII.

II. Optimal Operational Control Formulation for Rougher Flotation Processes

In this section, the OOC problem of dual-rate rougher flotation process is formulated, along with lower-layer rougher flotation cells with fast sampling and the upper-layer economic benefit operational control with slow updating as shown in Fig. 1.

A. Notations

R^n n dimensional Euclidean space
$\otimes$ Kronecker product
$\mathrm{vec}(L)$ Turning any matrix L into a single column vector
$\mathrm{diag}(*)$ Diagonal matrix of *
$r(k)$ Economic benefit at time instant k, and k is some non-negative integer
r^* Desired operational index
l Cell l ($l=1,2,\cdots,s$), s is some positive integer
M_{lp}^i, M_{le}^i Pulp mass and froth mass
i Mineralogical class 1, 2.
h_{lp}, q_{la} Pulp level and feed flow rate
L_{lcg}, L_{ltg} Concentrate grade and Tail grade
g_a Feed mineral grade
Q_{esp} Frother
Q_{col} Collector
$x_l=[M_{lp}^i\ M_{le}^i]^{\mathrm{T}}$ State
$u_l=[h_{lp}\ q_{la}]^{\mathrm{T}}$ Control input
$y_l=[L_{lcg}\ L_{ltg}]^{\mathrm{T}}$ Control output
n Sampling time instant of the lower-layer control loop, $n=0,1,\cdots$

K_{lp} and K_{Il} Proportional coefficient and integral coefficient controller gains of proportional–integral (PI) controller

$e_l(n) = w_l(n) - y_l(n)$ Tracking errors

$w_l(n)$ Set-point

w_l^*, $\hat{w}_l^*$ Optimal set-point and approximate optimal set-point

$\frac{\partial f_l}{\partial x}$, $\frac{\partial f_l}{\partial u}$, $\frac{\partial g_l}{\partial r}$, $\frac{\partial g_l}{\partial y}$ Partial derivatives with respect to x, u, r and y, respectively

$p_l^{'}$, $p_l^{''}$ First-order and second-order derivatives, respectively

$\frac{\partial^2}{\partial x} f_l$, $\frac{\partial^2}{\partial u} f_l$, $\frac{\partial^2}{\partial r} g_l$, $\frac{\partial^2}{\partial y} g_l$ Second-order partial derivatives with respect to x, u, r and y, respectively

x_{le}, u_{le}, r_e, y_{le}, $\hat{y}_e$ Equilibrium points

γ Attenuation factor, $\gamma > 0$

β Discount factor, $0 < \beta \le 1$

B. Lower-layer Rougher Flotation Process Dynamics

A rougher flotation process consists of several cells, whose nonlinear dynamics can be expressed as

$$\begin{aligned} \dot{x}_l(t) &= f_l(x_l(t), u_l(t)) \\ y_l(t) &= p_l(x_l(t)) \end{aligned} \tag{1}$$

where $f_l(x_l(t), u_l(t))$ and $p_l(x_l(t))$ are assumed to be second-order continuously differentiable.

The objective of the lower-layer loop control is to ensure the control output to steadily track the set-points input by the operational control layer. To this end, a digital output feedback PI controller is employed as follows:

$$u_l(n) = K_{lP} e_l(n) + K_{Il} E_l(n) \tag{2}$$

where $E_l(n) = \sum_{i=1}^{n-1} e_l(i)$ is the summation of the tracking errors, and its dynamic can be expressed in the following form:

$$E_l(n+1) = E_l(n) - y_l(n) + w_l(n) \tag{3}$$

C. Economic Benefit Optimization

Since economic benefit is closely related to the concentrate grade, the tail grade and the control input of rougher flotation process, and their relationship usually shows nonlinear feature, then the following nonlinear function is employed to show the dynamics of the economic benefit:

$$r(k+1) = g(r(k), L_{lcg}(k), L_{stg}(k), u(k)) \tag{4}$$

where the nonlinear function $g(r(k), L_{lcg}(k), L_{stg}(k), u(k))$ is usually second-order continuously differentiable.

In order to render the economic benefit to a desired value by the optimal approach, the goal of this paper is to design optimal set-points for achieving the optimal control of the economic benefit by minimizing the following performance index

$$J = \sum_{k=0}^{\infty} \beta^k ((r(k) - r^*)^{\mathrm{T}} Q (r(k) - r^*) + w^{\mathrm{T}}(k) R w(k)) \tag{5}$$

where Q and R are positive semi-definite matrix and positive definite matrix, respectively.

Remark 1: Naturally the discount factor β is introduced into (5) since the set-point $w(k)$ usually depends on the desired operational index r^* and $r^* \neq 0$.

Problem 1 is presented to clearly formulate the OOC problem for rougher flotation circuits.

***Problem 1*:**

Control objective:

$$\min_{w(k)} \sum_{k=0}^{\infty} \beta^k ((r(k) - r^*)^{\mathrm{T}} Q (r(k) - r^*) + w^{\mathrm{T}}(k) R w(k)) \tag{6}$$

Subject to (1)-(4)

Remark 2: It is quite hard to solve Problem 1 due to: 1) nonlinear dynamics constraints of the rougher flotation cells and the economic benefit index. 2) The dual rates, i.e., fast sampling rate in the lower-layer control loops and slow operational velocity in the upper-layer economic benefit operational control.

III. H∞ TRACKING CONTROL FOR ROUGHER FLOTATION PROCESSES

This section focuses on converting Problem 1 into H∞ tracking control problem for rougher flotation operational processes. First, the dynamics of the rougher flotation process operating at steady state are linearized using Taylor Series Expansion. Second, the lifting technique [17-19] is employed for dealing with the dual-rate sampling, and further solving OOC problem in Problem 1 is transformed as finding the solution of H∞ tracking control problem.

A. Linearization of Rougher Flotation Process

Since the dynamics of rougher flotation operational process are approximately linear near steady-state points. Thus systems (1) and (4) are rewritten using Taylor Series Expansion as [14, 17-19]

$$\begin{aligned} \dot{x}_l(t) &= G_l x_l(t) + N_l u_l(t) + d_{l1}(t) \\ y_l(t) &= C_l x_l(t) + d_{l2}(t) \end{aligned} \tag{7}$$

$$r(k+1) = L r(k) + M \hat{y}(k) + S u(k) + \gamma(k) \tag{8}$$

where

$$\hat{y}(k) = \begin{bmatrix} L_{1cg}(k) & L_{2cg}(k) & \cdots & L_{scg}(k) & L_{stg}(k) \end{bmatrix}^{\mathrm{T}},$$

$$G_l = \frac{\partial f_l(x_{le}, u_{le})}{\partial x},\ N_l = \frac{\partial f_l(x_{le}, u_{le})}{\partial u},\ C_l = p_l^{'}(x_{le}),$$

$$d_{l1}(t) = x_{le} - u_{le} \frac{\partial f_l(x_{le}, u_{le})}{\partial u} - x_{le} \frac{\partial f_l(x_{le}, u_{le})}{\partial x} + R_{l1}(t),$$

$$d_{l2}(t) = y_{le} - x_{le} \frac{\partial p_l(x_{le})}{\partial x} + R_{l2}(t),$$

$$L = \frac{\partial g(r_e, \hat{y}_e, u_e)}{\partial r},\ M = \frac{\partial g(r_e, \hat{y}_e, u_e)}{\partial \hat{y}},\ S = \frac{\partial g(r_e, \hat{y}_e, u_e)}{\partial u}.$$

$R_{l1}(t) = \frac{1}{2}((x_l(t) - x_{le}) \frac{\partial}{\partial x} + ((u_l(t) - u_{le}) \frac{\partial}{\partial u})^2 f_l(\hbar_l, \tilde{\lambda}_l)$ and $R_{l2}(t) = \frac{1}{2} p_l^{''}(\varsigma_l)(x_l(t) - x_{le})^2$ are the bounded residual errors.

$u_e = [u_{1e}^{\mathrm{T}}\ u_{2e}^{\mathrm{T}}\ \cdots\ u_{se}^{\mathrm{T}}]^{\mathrm{T}}$, $\hbar_l = x_{le} + \theta_{l1}(x_l(t) - x_{le})$, $\tilde{\lambda}_l = u_{le} + \theta_{l2}(u_l(t) - u_{le})$, $\varsigma_l = x_{le} + \theta_{l3}(x_l(t) - x_{le})$ and $0 \le \theta_{li} \le 1$ $(i = 1, 2, 3)$.

$\gamma(k)=r_e-y_eM-r_eL+\rho(k)-u_eS$, $\rho(k)=\frac{1}{2}((r(k)-r_e)\frac{\partial}{\partial r}+(\hat{y}(k)-\hat{y}_e)\frac{\partial}{\partial \hat{y}}+(u(k)-u_e)\frac{\partial}{\partial u})^2 g(\kappa,\upsilon,\tau)$. κ,υ,τ are real numbers between r_e and $r(k)$, $\hat{y}_e$ and $\hat{y}(k)$, u_e and $u(k)$, respectively.

Remark 3: Since systems (1) and (4) are linearized in terms of Taylor Series Expansion due to their steady operation at steady state, the solution of the focused OOC problem in this paper is essentially suboptimal.

B. Lifting Technique for Dual-rate Rougher Flotation Processes

Since digital controller (2) is employed, then system (7) is in fact the following discrete-time system

$$\begin{aligned} x_l(n+1)&=\hat{G}_l x_l(n)+\hat{N}_l u_l(n)+\hat{d}_{l1}(n)\\ y_l(n)&=C_l x_l(n)+d_{l2}(n)\end{aligned} \tag{9}$$

where $\hat{G}_l=e^{G_lT_0}$, $\hat{N}_l=\int_0^{T_0}e^{G_l\tau}d\tau N_l$, $\hat{d}_{l1}(n)=\int_0^{T_0}e^{G_l\tau}d_{l1}(\tau)d\tau$.

T_0 is the sampling period of the rougher flotation cells.

An augmented system is constructed by defining a compact form $\xi_l(n)=[x_l^{\mathrm{T}}(n)\ \ E_l^{\mathrm{T}}(n)]^{\mathrm{T}}$ as

$$\xi_l(n+1)=\overline{A}_l\xi_l(n)+\overline{B}_l w_l(n)+\overline{I}_l\tilde{d}_l(n) \tag{10}$$

where

$$\overline{A}_l=\begin{bmatrix}\hat{G}_l-\hat{N}_lK_PC & \hat{N}_lK_I\\ -C_l & I\end{bmatrix},\ \overline{B}_l=\begin{bmatrix}\hat{N}_lK_P\\ I\end{bmatrix},$$

$$\overline{I}_l=\begin{bmatrix}I & -\hat{N}_lK_P\\ 0 & -I\end{bmatrix},\ \tilde{d}_l(n)=\begin{bmatrix}\hat{d}_{l1}(n)\\ d_{l2}(n)\end{bmatrix}$$

Further, defining $\xi(n)=[\xi_1^{\mathrm{T}}(n)\ \ \xi_2^{\mathrm{T}}(n)\cdots\xi_s^{\mathrm{T}}(n)]^{\mathrm{T}}$ yields the following compact form for the rougher flotation process

$$\xi(n+1)=\overline{A}\xi(n)+\overline{B}w(n)+\overline{I}\vec{d}(n) \tag{11}$$

where $w(n)=[w_1^{\mathrm{T}}(n)\ \ w_2^{\mathrm{T}}(n)\cdots w_s^{\mathrm{T}}(n)]^{\mathrm{T}}$, $\tilde{d}(n)=[\tilde{d}_1^{\mathrm{T}}(n)\ \ \tilde{d}_2^{\mathrm{T}}(n)\cdots\tilde{d}_s^{\mathrm{T}}(n)]^{\mathrm{T}}$, $\overline{A}=\mathrm{diag}(\overline{A}_1,\overline{A}_2,\cdots,\overline{A}_s)$, $\overline{B}=\mathrm{diag}(\overline{B}_1,\overline{B}_2,\cdots,\overline{B}_s)$, $\overline{I}=\mathrm{diag}(\overline{I}_1,\overline{I}_2,\cdots,\overline{I}_s)$. And (8) can be rewritten as

$$\begin{aligned} r(k+1)&=Lr(k)+(\overline{M}-\overline{S}+S\overline{K}_I)\xi(k)+S\overline{K}_p w(k)\\ &\quad+\gamma(k)+(\hat{M}-\hat{S})\tilde{d}(k)\end{aligned} \tag{12}$$

where

$\overline{M}=M\hat{I}\overline{C}$, $\overline{C}=\mathrm{diag}(\overline{C}_1,\overline{C}_2,\cdots,\overline{C}_s)$, $\overline{C}_j=[C_j\ 0]$,

$\hat{M}=M\hat{I}\overline{\Pi}$, $\overline{\Pi}=\mathrm{diag}(\Pi,\Pi,\cdots,\Pi)$, $\Pi=[0\ I]$,

$\overline{S}=S\overline{K}_p\overline{C}$, $\overline{K}_P=\mathrm{diag}(K_{1P},K_{2P},\cdots,K_{sP})$, $\hat{S}=S\overline{K}_p\overline{\Pi}$,

$\overline{K}_I=\mathrm{diag}([0\ K_{1I}],[0\ K_{2I}],\cdots,[0\ K_{sI}])$, $\hat{I}=\begin{bmatrix}I & 0\\ 0 & 1\end{bmatrix}$.

It is well known that data are measured at the time instant n in the lower-layer flotation cell, while the economic benefit value is updated at the time instant $k=N_0n$ in the upper-layer operational control loop. Then the following form holds

$$w(k)=w(nN_0)=w(nN_0+1)=\ldots=w(nN_0+N_0-1) \tag{13}$$

and (11) is rewritten in terms of slow-time scale as

$$\begin{aligned} \xi((k+1))&=\xi((n+1)N_0)=\xi(nN_0+N_0)\\ &=\tilde{A}\xi(k)+\tilde{B}w(k)+D\vartheta(k)\end{aligned} \tag{14}$$

where $\tilde{A}=\overline{A}^{N_0}$, $\tilde{B}=\sum_{i=0}^{N_0-1}\overline{A}^i\overline{B}$, $D=[\overline{A}^0\overline{I},\overline{A}^1\overline{I},\cdots,\overline{A}^{N_0-1}\overline{I}]$,

$\vartheta(k)=[\tilde{d}^{\mathrm{T}}(nN_0+(N_0-1))\ \tilde{d}^{\mathrm{T}}(nN_0+(N_0-2))\cdots\tilde{d}^{\mathrm{T}}(nN_0)]^{\mathrm{T}}$.

Let $X(k)=[\xi^{\mathrm{T}}(k)\ r^{\mathrm{T}}(k)\ r^{*\mathrm{T}}(k)]^{\mathrm{T}}$, one has

$$X(k+1)=\tilde{G}X(k)+\tilde{N}w(k)+\tilde{H}\chi(k) \tag{15}$$

where

$$\tilde{G}=\begin{bmatrix}\tilde{A} & 0 & 0\\ \overline{M}-\overline{S}+S\overline{K}_I & L & 0\\ 0 & 0 & I\end{bmatrix},\ \tilde{N}=\begin{bmatrix}\tilde{B}\\ S\overline{K}_p\\ 0\end{bmatrix},\ \tilde{H}=\begin{bmatrix}D & 0\\ \hat{D} & I\\ 0 & 0\end{bmatrix},$$

$$\hat{D}=[0\ 0\cdots\hat{M}-\hat{S}],\ \chi(k)=\begin{bmatrix}\vartheta(k)\\ \gamma(k)\end{bmatrix}$$

Remark 4: Here, the desired economic benefit is constant, i.e., $r^*(k+1)=r^*(k)$.

Note that the residual error $\chi(k)$ is dependent of the state $X(k)$ of augmented system (15) by the definition of $\chi(k)$. For solving Problem 1, $\chi(k)$ should be considered as the disturbance corresponding to the state $X(k)$. Thus, the OOC problem shown in Problem 1 can be formulated as the H∞ tracking control problem below.

Problem 2:

Find a set-point $w(k)=w(X_k)$ satisfying

1) the attenuation condition below for a specific attenuation factor $\gamma>0$:

$$\sum_{k=0}^{\infty}\beta^k(X(k)^{\mathrm{T}}\tilde{Q}X(k)+w^{\mathrm{T}}(k)Rw(k))\le\gamma^2\sum_{k=0}^{\infty}\beta^k\left\|\chi(k)\right\|^2 \tag{16}$$

2) The tracking error $e_r(k)$ $(e_r(k)=r(k)-r^*)$ converges to zero.

IV. ON-POLICY Q-LEARNING FOR SOLVING H∞ CONTROL BASED ON ZERO-SUM GAME

This section shows how to find the solution of Problem 2 by using zero-sum game based Q-learning method.

Define a new performance index in terms of (16):

$$\begin{aligned} &J(w(k),\chi(k))\\ &=\sum_{k=0}^{\infty}\beta^k(X(k)^{\mathrm{T}}\tilde{Q}X(k)+w^{\mathrm{T}}(k)Rw(k)-\gamma^2\left\|\chi(k)\right\|^2)\end{aligned} \tag{17}$$

Actually solving H∞ tracking control problem is equivalent to maximizing and minimizing the cost function (17) by using zero-sum game approach [24-29], that is,

$$\begin{aligned} J(w^*(k),\chi^*(k))&=\min_{w(k)}\max_{\chi(k)}J(w(k),\chi(k))\\ &=\min_{w(k)}\max_{\chi(k)}\Bigg(\sum_{k=0}^{\infty}\beta^k(X(k)^{\mathrm{T}}\tilde{Q}X(k)\\ &\quad+w^{\mathrm{T}}(k)Rw(k)-\gamma^2\left\|\chi(k)\right\|^2)\Bigg)\end{aligned} \tag{18}$$

Define the value function as

$$V^*(X(k)) = \min_{w_k}\max_{\chi_k}\sum_{k=0}^{\infty}\beta^k (X(k)^{\mathrm{T}}\tilde{Q}X(k) + w^{\mathrm{T}}(k)Rw(k) - \gamma^2\left\|\chi(k)\right\|^2) \tag{19}$$

Combining with $w^*(k) = -K_1^* X(k)$ and $\chi^*(k) = -K_2^* X(k)$ yields $V^*(X_k) = X^{\mathrm{T}}(k)PX(k)$ $(P>0)$ by referring to [24, 30]. By (19), the action-dependent optimal Q-function is defined below:

$$Q^*(X(k), w(k), \chi(k)) = X(k)^{\mathrm{T}}\tilde{Q}X(k) + w^{\mathrm{T}}(k)Rw(k) - \gamma^2\chi(k)^{\mathrm{T}}\chi(k) + \beta V^*(X(k+1)) \tag{20}$$

and one has

$$V^*(X(k)) = \min_{w_k}\max_{\chi_k} Q^*(X(k), w(k), \chi(k)) = Q^*(X(k), w^*(k), \chi^*(k)) \tag{21}$$

Thus Proposition 1 is naturally derived by referring to [24].

Proposition 1: If we set $z(k) = [X^{\mathrm{T}}(k)\ w^{\mathrm{T}}(k)\ \chi^{\mathrm{T}}(k)]^{\mathrm{T}}$, $w(k) = -K_1 X(k)$, $\chi(k) = -K_2 X(k)$, thus the following form holds

$$Q^*(X(k), w(k), \chi(k)) = z^{\mathrm{T}}(k)Hz(k) \tag{22}$$

H is denoted as

$$\begin{bmatrix} H_{XX} & H_{Xw} & H_{X\chi} \\ H_{Xw}^{\mathrm{T}} & H_{ww} & H_{w\chi} \\ H_{X\chi}^{\mathrm{T}} & H_{w\chi}^{\mathrm{T}} & H_{\chi\chi} \end{bmatrix} = \begin{bmatrix} \beta\tilde{G}^{\mathrm{T}}P\tilde{G} + \tilde{Q} & \beta\tilde{G}^{\mathrm{T}}P\tilde{N} & \beta\tilde{G}^{\mathrm{T}}P\tilde{H} \\ \beta(\tilde{G}^{\mathrm{T}}P\tilde{N})^{\mathrm{T}} & \beta\tilde{N}^{\mathrm{T}}P\tilde{N} + R & \beta\tilde{N}^{\mathrm{T}}P\tilde{H} \\ \beta(\tilde{G}^{\mathrm{T}}P\tilde{H})^{\mathrm{T}} & \beta(\tilde{N}^{\mathrm{T}}P\tilde{H})^{\mathrm{T}} & -\gamma^2 I + \beta\tilde{H}^{\mathrm{T}}P\tilde{H} \end{bmatrix} \tag{23}$$

and

$$P = (M^*)^{\mathrm{T}} H M^* \tag{24}$$

where $M^* = \begin{bmatrix} I & -(K_1^*)^{\mathrm{T}} & -(K_2^*)^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$.

By Proposition 1, one has the Q-function based game Bellman equation

$$z^{\mathrm{T}}(k)Hz(k) = X^{\mathrm{T}}(k)\tilde{Q}X(k) + w^{\mathrm{T}}(k)Rw(k) - \gamma^2\chi^{\mathrm{T}}(k)\chi(k) + \beta z^{\mathrm{T}}(k+1)Hz(k+1) \tag{25}$$

Implementing $\partial Q^*(X(k), w(k), \chi(k))/\partial w(k) = 0$ and $\partial Q^*(X(k), w(k), \chi(k))/\partial\chi(k) = 0$ yields the optimal set-point and the worst disturbance as

$$\begin{aligned} w^*(k) &= -K_1^* X(k) = -(H_{ww} - H_{w\chi}H_{\chi\chi}^{-1}(H_{w\chi})^{\mathrm{T}})^{-1} \cdot ((H_{Xw}^{\mathrm{T}} - H_{w\chi}(H_{\chi\chi}^{-1})(H_{X\chi}^{\mathrm{T}}))X(k) \\ \chi^*(k) &= -K_2^* X(k) = -(H_{\chi\chi} - H_{w\chi}^{\mathrm{T}}H_{ww}^{-1}H_{w\chi})^{-1} \cdot (H_{X\chi}^{\mathrm{T}} - H_{w\chi}^{\mathrm{T}}H_{ww}^{-1}H_{Xw}^{\mathrm{T}})X(k) \end{aligned} \tag{26}$$

where the matrix H satisfies (25). The policy iteration (PI) is used to learn the optimal set-point and the worst disturbance.

Algorithm 1: Model-free Q-learning algorithm

1. Initialization: Given stabilizing set-point and disturbance policy gains K_1^0 and K_2^0, $0 < \beta \le 1$ and $\gamma > 0$. Let $j = 0$, where j denotes iteration index;

2. Policy evaluation by solving Q-function matrix H^{j+1}:

$$z^{\mathrm{T}}(k)H^{j+1}z(k) = X^{\mathrm{T}}(k)\tilde{Q}X(k) + (w^j(k))^{\mathrm{T}}Rw^j(k) - \gamma^2(\chi^j(k))^{\mathrm{T}}\chi^j(k) + \beta z^{\mathrm{T}}(k+1)H^{j+1}z(k+1) \tag{27}$$

3. Policy update:

$$w^{j+1}(k) = -K_1^{j+1}X(k),\ \chi^{j+1}(k) = -K_2^{j+1}X(k)$$

where

$$\begin{aligned} K_1^{j+1} &= (H_{ww}^{j+1} - H_{w\chi}^{j+1}(H_{\chi\chi}^{j+1})^{-1}(H_{w\chi}^{j+1})^{\mathrm{T}})^{-1} \cdot ((H_{Xw}^{j+1})^{\mathrm{T}} - H_{w\chi}^{j+1}(H_{\chi\chi}^{j+1})^{-1}(H_{X\chi}^{j+1})^{\mathrm{T}}) \\ K_2^{j+1} &= (H_{\chi\chi}^{j+1} - (H_{w\chi}^{j+1})^{\mathrm{T}}(H_{ww}^{j+1})^{-1}H_{w\chi}^{j+1})^{-1} \cdot ((H_{X\chi}^{j+1})^{\mathrm{T}} - (H_{w\chi}^{j+1})^{\mathrm{T}}(H_{ww}^{j+1})^{-1}(H_{Xw}^{j+1})^{\mathrm{T}}) \end{aligned} \tag{28}$$

$$K_2^{j+1} = (H_{\chi\chi}^{j+1} - (H_{w\chi}^{j+1})^{\mathrm{T}}(H_{ww}^{j+1})^{-1}H_{w\chi}^{j+1})^{-1} \cdot ((H_{X\chi}^{j+1})^{\mathrm{T}} - (H_{w\chi}^{j+1})^{\mathrm{T}}(H_{ww}^{j+1})^{-1}(H_{Xw}^{j+1})^{\mathrm{T}}) \tag{29}$$

4. Stop when $\left\|H^j - H^{j+1}\right\| \le \varepsilon$ with a small constant ε $(\varepsilon > 0)$.

Remark 5: Note that Algorithm 1 is in fact an on-policy Q-learning approach wherein the disturbance $\chi(k)$ needs to be updated and act the rougher flotation cells using $\chi_k^{j+1} = -K_2^{j+1}X_k$, while it essentially is the residual error generated by linearizing the nonlinear dynamics and cannot be specified. Moreover, as pointed out in [25], adding probing noise to the set-point results in a bias in solving the real value of H^{j+1}. Compared with on-policy Q-learning, off-policy Q-learning including behavior policy and target policy is more practical and efficient technique for dealing with optimal control problem as it can overcome the two shortcomings generated by on-policy Q-learning. Hence the sequels will devote to designing off-policy Q-learning algorithm for achieving optimal operation of rougher flotation processes.

V. Off-policy Q-Learning Algorithm

This section first presents an off-policy Q-learning algorithm for learning the optimal set-point and the worst disturbance using only data, and then rules used for selecting the optimum set-point are proposed as shown in Fig. 2.

A. Derivation of Off-policy Q-learning Algorithm

Q-function based Lyapunov equation is given as follows by using (27)

$$(M^j)^{\mathrm{T}}H^{j+1}M^j = (M^j)^{\mathrm{T}}\Pi M^j + \beta\tilde{G}_c^{\ T}(M^j)^{\mathrm{T}}H^{j+1}M^j\tilde{G}_c \tag{30}$$

where $M^j = \begin{bmatrix} I & -(K_1^j)^{\mathrm{T}} & -(K_2^j)^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$, $\Pi = \mathrm{diag}(\tilde{Q}, R, -\gamma^2 I)$.

Introducing the auxiliary variables $w^j(k) = -K_1^j X(k)$ and $\chi^j(k) = -K_2^j X(k)$ into augmented system (15) yields

$$X(k+1) = \tilde{G}_c X(k) + \tilde{N}(K_1^j X(k) + w(k)) + \tilde{H}(K_2^j X(k) + \chi(k)) \tag{31}$$

where $\tilde{G}_c = \tilde{G} - \tilde{N}K_1^j - \tilde{H}K_2^j$. Along the trajectory of (31), one has

$$\begin{aligned} &Q^{*,j+1}(X(k), w^j(k), \chi^j(k)) - \beta X^{\mathrm{T}}(k)\tilde{G}_c^{\ T}(M^j)^{\mathrm{T}}H^{j+1}M^j\tilde{G}_c X(k) \\ &= X^{\mathrm{T}}(k)(M^j)^{\mathrm{T}}H^{j+1}M^j X(k) - \beta(X(k+1) - \tilde{N}(K_1^j X(k) + w(k)) - \tilde{H}(K_2^j X(k) + \chi(k)))^{\mathrm{T}} \end{aligned}$$

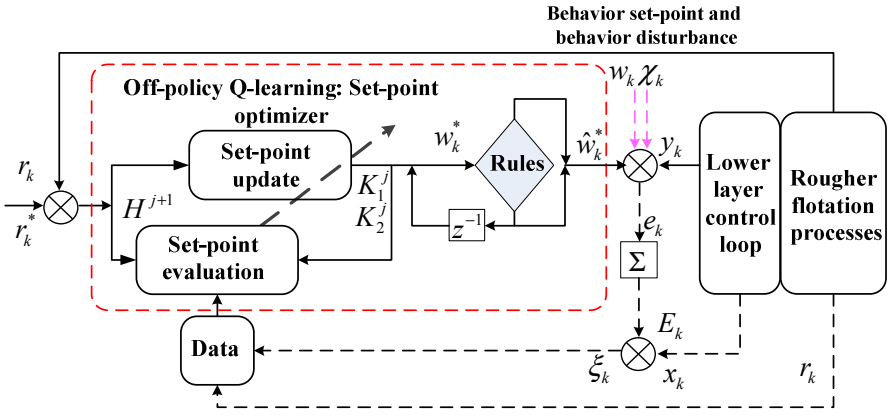

Fig. 2. Off-policy Q-learning scheme by using zero-sum games

$$\begin{aligned}&\cdot(M^j)^{\mathrm{T}}H^{j+1}M^j(X(k+1)-\tilde{N}(K_1^jX(k)+w(k))\\&-\tilde{H}(K_2^jX(k)+\chi(k)))=X^{\mathrm{T}}(k)(M^j)^{\mathrm{T}}\prod M^jX(k)\end{aligned}\tag{32}$$

Since P^{j+1} and H^{j+1} have the same relationship as that shown in (24), then the off-policy Q-function game Bellman equation can be obtained

$$\begin{aligned}&X^{\mathrm{T}}(k)(M^j)^{\mathrm{T}}H^{j+1}M^jX(k)\\&-\beta X^{\mathrm{T}}(k+1)(M^j)^{\mathrm{T}}H^{j+1}M^jX(k+1)\\&+2\beta X^{\mathrm{T}}(k+1)P^{j+1}\tilde{N}(K_1^jX(k)+w(k))\\&+2\beta X^{\mathrm{T}}(k+1)P^{j+1}\tilde{H}(K_2^jX(k)+\chi(k))\\&-\beta(K_1^jX(k)+w(k))^{\mathrm{T}}\tilde{N}^{\mathrm{T}}P^{j+1}\tilde{N}(K_1^jX(k)+w(k))\\&-2\beta(K_1^jX(k)+w(k))^{\mathrm{T}}\tilde{N}^{\mathrm{T}}P^{j+1}\tilde{H}(K_2^jX(k)+\chi(k))\\&-\beta(K_2^jX(k)+\chi(k))^{\mathrm{T}}\tilde{H}^{\mathrm{T}}P^{j+1}\tilde{H}(K_2^jX(k)+\chi(k))\\&=X^{\mathrm{T}}(k)(M^j)^{\mathrm{T}}\Pi M^jX(k)\end{aligned}\tag{33}$$

Consider that the relationship between H^{j+1} and P^{j+1} is same as that shown in (23), (33) can be rewritten as

$$H^j(k)L^{j+1}=\rho_k^j\tag{34}$$

where

$$\rho_k^j=X^{\mathrm{T}}(k)\tilde{Q}X(k)+w^{\mathrm{T}}(k)Rw(k)-\gamma^2\chi^{\mathrm{T}}(k)\chi(k),$$

$$L_1^{j+1}=H_{XX}^{j+1},\ L_2^{j+1}=H_{Xw}^{j+1},\ L_3^{j+1}=H_{X\chi}^{j+1},$$

$$L_4^{j+1}=H_{ww}^{j+1},\ L_5^{j+1}=H_{w\chi}^{j+1},\ L_6^{j+1}=H_{\chi\chi}^{j+1},$$

$$L^{j+1}=\left[(\mathrm{vec}(L_1^{j+1}))^{\mathrm{T}}\quad(\mathrm{vec}(L_2^{j+1}))^{\mathrm{T}}\quad\cdots\quad(\mathrm{vec}(L_6^{j+1}))^{\mathrm{T}}\right]^{\mathrm{T}},$$

$$H^j(k)=\left[H_1^j\quad H_2^j\quad\cdots\quad H_6^j\right],$$

$$H_1^j=(X^{\mathrm{T}}(k)\otimes X^{\mathrm{T}}(k))-\beta(X^{\mathrm{T}}(k+1)\otimes X^{\mathrm{T}}(k+1)),$$

$$H_2^j=2X^{\mathrm{T}}(k)\otimes w^{\mathrm{T}}(k)+2\beta X^{\mathrm{T}}(k+1)\otimes(K_1^jX(k+1))^{\mathrm{T}},$$

$$H_3^j=2X^{\mathrm{T}}(k)\otimes\chi^{\mathrm{T}}(k)+2\beta X^{\mathrm{T}}(k+1)\otimes(K_2^jX(k+1))^{\mathrm{T}},$$

$$\begin{aligned}H_4^j&=-\beta(K_1^jX(k+1)))^{\mathrm{T}}\otimes(K_1^j(X(k+1))^{\mathrm{T}}\\&+w^{\mathrm{T}}(k)\otimes w^{\mathrm{T}}(k),\end{aligned}$$

$$\begin{aligned}H_5^j&=-2\beta(K_1^jX(k+1))^{\mathrm{T}}\otimes(K_2^j(X(k+1))^{\mathrm{T}}\\&+2\chi^{\mathrm{T}}(k)\otimes w^{\mathrm{T}}(k),\end{aligned}$$

$$\begin{aligned}H_6^j&=-\beta(K_2^jX(k+1))^{\mathrm{T}}\otimes(K_2^jX(k+1))^{\mathrm{T}}\\&+\chi^{\mathrm{T}}(k)\otimes\chi^{\mathrm{T}}(k)\end{aligned}$$

If L^{j+1} is estimated, then K_1^{j+1} and K_2^{j+1} can be calculated as

$$\begin{aligned}K_1^{j+1}&=(L_4^{j+1}-L_5^{j+1}(L_6^{j+1})^{-1}(L_5^{j+1})^{\mathrm{T}})^{-1}\\&\cdot((L_2^{j+1})^{\mathrm{T}}-L_5^{j+1}(L_6^{j+1})^{-1}(L_3^{j+1})^{\mathrm{T}})\end{aligned}\tag{35}$$

$$\begin{aligned}K_2^{j+1}&=(L_6^{j+1}-(L_5^{j+1})^{\mathrm{T}}(L_4^{j+1})^{-1}L_5^{j+1})^{-1}\\&\cdot((L_3^{j+1})^{\mathrm{T}}-(L_5^{j+1})^{\mathrm{T}}(L_4^{j+1})^{-1}(L_2^{j+1})^{\mathrm{T}})\end{aligned}\tag{36}$$

B. Off-policy Q-learning Algorithm

Algorithm 2: Off-policy Q-learning algorithm

1. Data collection: Collect system data $(x(n), E(n), r(k))$ from the rougher flotation operational process using a behavior set-point $w(k)$ and a behavior disturbance $\chi(k)$ and store them in the sample sets H_i^j and ρ^j. Given $\beta>0$ and $\gamma>0$;

2. Initiation: Choose the initial gains K_1^0 and K_2^0, such that system (15) can be stabilized. Let $j=0$;

3. Implementing Q-learning: L_i^{j+1} $(i=1,2,\cdots,6)$ are estimated in terms of (34) using the collected data in Step 1, and then K_1^j and K_2^j are updated in terms of (35) and (36);

4. If $\left\|K_1^j-K_1^{j-1}\right\|\le l_1$ and $\left\|K_2^j-K_2^{j-1}\right\|\le l_2$ (l_1 and l_2 are small positive numbers), then stop the iteration and the optimal set-points have been obtained. Otherwise, let $j=j+1$ and go back to Step 3.

Remark 6: In Algorithm 2, a specific behavior set-point and a specific behavior disturbance are applied to generate data of rougher flotation operational process while the behavior set-point is tracked by the output of lower-layer control loops, thus the data $(x_l(n), E_l(n), r(k))$ can be firstly collected. Then the set-point $w^j(k)=-K_1^jX(k)$ and the disturbance policy $\chi^j(k)=-K_2^jX(k)$ are evaluated and updated using the collected data, but they are not applied to the rougher flotation operational processes, which is different from the on-policy RL [22-24, 30] where the target set-point and target disturbance under evaluation are applied to systems to collect data.

Remark 7: Practical rougher flotation processes are usually steadily operated near the equilibrium points under the control

inputs obtained based on operator's rich experience. Thus the initial stabilizing gains K_1^0 and K_2^0 can be obtained easily. Besides, H∞ control method can be used to yield an initial stabilizing control policies for systems [31].

***Remark 8*:** To accurately estimate the real value of L^{j+1} in (34) by using the recursive least squares (RLS) or batch least squares (BLS), the behavior set-point should be in the form of the $w(k)+e(k)$ ($e(k)$ is a probing noise) to guarantee the persistence of excitation [22-27, 29-31]. Lemma 1 is presented to show that there is no bias as a result of adding probing noise in Algorithm 2.

Lemma 1: Let H^{j+1} be the solution to (32) with the behavior set-point $w(k)$, then it is equivalent to the solution to (32) with the behavior set-point $w(k)+e(k)$ $(e(k)\neq 0)$.

Proof: If the behavior set-point is $w(k)+e(k)$, then the collected data $X(k+1)$ is in fact $\hat{X}(k+1)$ with the form of

$$\begin{aligned}\hat{X}(k+1)=&\tilde{G}_c X(k)+\tilde{N}(K_1^j X(k)+w(k)+e(k))\\&+\tilde{H}(K_2^j X(k)+\chi(k))\end{aligned}\tag{37}$$

Substituting (37) and $w(k)+e(k)$ into (32) yields

$$\begin{aligned}&X^{\mathrm{T}}(k)(M^j)^{\mathrm{T}}H^{j+1}M^j X(k)-\beta(\hat{X}(k+1)\\&-\tilde{N}(K_1^j X(k)+w(k)+e(k))-\tilde{H}(K_2^j X(k)+\chi(k)))^{\mathrm{T}}\\&\cdot(M^j)^{\mathrm{T}}H^{j+1}M^j(\hat{X}(k+1)-\tilde{N}(K_1^j X(k)+w(k)+e(k))\\&=X^{\mathrm{T}}(k)(M^j)^{\mathrm{T}}\Pi M^j X(k)\end{aligned}\tag{38}$$

By (37) and (31), (38) becomes (32). Hence, adding the probing noise during learning in the proposed off-policy Q-learning algorithm cannot produce bias. This completes the proof. ■

Theorem 1: (H^{j+1}, K_1^j, K_2^j) is the solution of (27)-(29) if and only if it is the solution of (34)-(36).

Proof : It is easily concluded that if (H^{j+1}, K_1^j, K_2^j) is the solution of (27)-(29), then it can satisfy (34)-(36) from the above derivation. Next the fact that the solution of (34)-(36) is also the solution of (27)-(29) will be shown.

Note that (34) is equivalent to (33) by checking them. Thus the solution of (34) can make (33) hold. Subtracting (33) from (32), one has

$$\begin{aligned}&X^{\mathrm{T}}(k)(M^j)^{\mathrm{T}}H^{j+1}M^j X(k)\\&-\beta X^{\mathrm{T}}(k)\tilde{G}_c{}^{\mathrm{T}}(M^j)^{\mathrm{T}}H^{j+1}M^j\tilde{G}_c X(k)\\&=X^{\mathrm{T}}(k)(M^j)^{\mathrm{T}}\Pi M^j X(k)\end{aligned}\tag{39}$$

Due to $X(k+1)=\tilde{G}_c X(k)$ by using $w^{j+1}(k)=-K_1^{j+1}X(k)$, and $\chi^{j+1}(k)=-K_2^{j+1}X(k)$ and the definition of $z(k)$, the solution of (39) is equivalent to that of (27). Moreover, (35) and (36) are the same as (28) and (29). This completes the proof. ■

Remark 9: From Theorem 1, it is easy to know that $w^j(k)=-K_1^j X(k)$ and $\chi^{j+1}(k)=-K_2^{j+1}X(k)$ learned by Algorithm 2 can converge to the Nash equilibrium solution $w^*(k)$ and $\chi^*(k)$ as $j\to\infty$, i.e. $\lim_{j\to\infty}w^j(k)=w^*(k)$, $\lim_{j\to\infty}\chi^j(k)=\chi^*(k)$ since the convergence of solutions of Algorithm 1 has been proven in [27, 29].

Remark 10: The existing model-free OOC methods [1, 2, 20, 21] tried to find the optimal set-points, while one can notice that the correction of set-points depends on the operator's experience in [1, 2, 21], and [20] presented the neural-network based set-points design on the premise that the optimal performance indices are known a priori. The proposed off-policy RL method in this paper can learn the optimal set-points for achieving OOC of operational processes with completely unknown dynamics of controlled unit processes and unknown functional dependence of operational index.

Remark 11: In contrast to the on-policy Q-learning method [22-24, 30], the arbitrary behavior policy is introduced when implementing the off-policy Q-learning method, which is used to generate data of systems for enriching data exploration while the target policies, especially disturbance policy, are updated to find the optimal policy but not to be employed to systems. Particularly, adding probing noise into the behavior policy would not produce bias of solution when implementing policy evaluation [25]. Actually, Algorithm 2 can also be applied into other stages of flotation processes such as cleaning, scavenger, regrinding and classification circuits, and therefore the optimal operation of the whole flotation processes can be achieved.

C. Optimum Set-point Selector

In practical rougher flotation operational processes, the set-points followed by the concentrate grades and the tail grades are usually bounded, that is $w_l\in[w_{l\min}, w_{l\max}]$, where $w_{l\max}$ and $w_{l\min}$ are non-negative real vectors with $w_{l\min}\le w_{l\max}$. To satisfy the constraint of set-points, the following rules are presented to get the optimum set-point $\hat{w}_k^*$.

Rule 1. If $w_{\min}\le w_k^*\le w_{\max}$ ($w_{\max}=[w_{1\max}^{\mathrm{T}}\ w_{2\max}^{\mathrm{T}}\cdots w_{s\max}^{\mathrm{T}}]^{\mathrm{T}}$, $w_{\min}=[w_{1\min}^{\mathrm{T}}\ w_{2\min}^{\mathrm{T}}\cdots w_{s\min}^{\mathrm{T}}]^{\mathrm{T}}$), then $\hat{w}_k^*=w_k^*$;

Rule 2. If $w_k^*<w_{\min}$ or $w_k^*>w_{\max}$, then $\hat{w}_k^*=w_{k-1}^*$.

Theorem 2: The set-point derived by the above rules can guarantee $\hat{w}_k^{j+1}$ to converge to the optimal set-points, that is,

$$\lim_{k\to\infty}\hat{w}_k^*=w_k^*\tag{40}$$

Proof: By Rule 1 and Rule 2, one has

$$\hat{w}_k^*=w_k^*+\alpha(w_{k-1}^*-w_k^*)\tag{41}$$

where α=0 or 1. Thus,

$$\left\|\hat{w}_k^*-w_k^*\right\|\le\left\|w_{k-1}^*-w_k^*\right\|\tag{42}$$

Since w_k^* can guarantee the stability of the whole operational process (15), then $w_{k-1}^*-w_k^*\to 0$ as $k\to\infty$. By (42), (40) holds. This completes the proof. ■

Remark 12: Refined Ziegler-Nichols (RZN) method can design the PI controller parameters for the lower-layer rougher flotation control loops [1, 2, 17-19].

VI. Simulation Results

In this section, the proposed off-policy Q-learning algorithm is verified in the rougher flotation operational process with two cells. Moreover, the dual rates operation is analyzed and proper comparisons are made to show the contributions of this paper.

The flotation models according to mass balance in froth and pulp phases are given below [14, 17-19]

$$\begin{aligned}\frac{dM_{lp}^{i}}{dt}=f_{lp}^{i}&=-\left(k_{p}^{i}+\frac{q_{lT}}{\left(1-\varepsilon_{g}\right)A_{l}h_{lp}}\right)M_{lp}^{i}+k_{e}^{i}M_{le}^{i}+q_{la}X_{la}^{i}\\ \frac{dM_{le}^{i}}{dt}=f_{le}^{i}&=-\left(k_{e}^{i}+\frac{q_{lc}}{\left(1-\varepsilon_{g}\right)A_{l}\left(H_{l}-h_{lp}\right)}\right)M_{le}^{i}+k_{p}^{i}M_{lp}^{i}\end{aligned}\tag{43}$$

The concentrate grade and the tail grade in cell l respectively are

$$L_{lcg}=\frac{M_{le}^{1}g_{lcp}^{1}+M_{le}^{2}g_{lcp}^{2}}{M_{le}^{1}+M_{le}^{2}}L_{cu},\ L_{ltg}=\frac{M_{lp}^{1}g_{lcp}^{1}+M_{lp}^{2}g_{lcp}^{2}}{M_{lp}^{1}+M_{lp}^{2}}L_{cu}\tag{44}$$

The general concentrate grade and the recovery respectively are

$$L_{cg}=\frac{\sum_{i=1}^{2}\sum_{l=1}^{2}M_{le}^{i}g_{lcp}^{i}}{\sum_{i=1}^{2}\sum_{l=1}^{2}M_{le}^{i}}L_{cu},\ L_{cg}=\frac{(g_{a}-L_{tg})L_{cg}}{(L_{cg}-L_{tg})g_{a}}\cdot 100\tag{45}$$

where $l=1,2$. The sizes of two cells are both $53.2\times3.2\ \mathrm{m}^3$ and values of other parameters are listed in Table 1 [14]. The economic benefit has the form of $r_k=M_r\hat{y}_k+N_ru_k$, where $M_r=[100\ \ 100\ 10]^{\mathrm{T}}$, $N_r=[0.01\ \ 0.1\ \ 0.01\ \ 0.1]$. The economic benefit objective is 10, $Q=800$, $R=\mathrm{diag}(50,\ 50,50,50)$, $\beta=0.9$, $\gamma=5$, $w_{l\max}=[1\ \ 1]^{\mathrm{T}}$ and $\mathrm{w}_{1\min}=[0\ \ 0]^{\mathrm{T}}$.

The equilibrium points of the rougher flotation process under the operational parameters listed in Table 1 are $[h_{1p}\ \ q_{1a}]=[2.8\ \ 17]$, $[h_{2p}\ \ q_{2a}]=[2.5\ \ 12]$, $[M_{1p}^{1}\ M_{1p}^{2}\ M_{1e}^{1}\ M_{1e}^{2}]=[16.8\ \ 824.266\ \ 4.56\ \ 0.104]$, $[M_{2p}^{1}\ \ M_{2p}^{2}\ \ M_{2e}^{1}\ \ M_{2e}^{2}]=[20.84\ \ 1300\ \ 5.67\ \ 0.1645]$.

TABLE I:
PARAMETERS AND VALUES

parameter	Physical meaning	Value
k_p^i	Flotation rate	$k_p^1=17.9$, $k_p^2=0.04\ (\mathrm{min}^{-1})$
k_e^i	Drainage rate	$k_e^1=65.6$, $k_p^2=316\ (\mathrm{min}^{-1})$
q_{1T}, q_{2T}	Wake	9.3, 6.3 $(\mathrm{m}^3/\mathrm{min})$
ε_g	Stagnation constant	0
$A_1=A_2$	Cross-section area	$53.2\ \mathrm{m}^2$
X_a^i	Mineral species concentration	$X_a^1=0.1549$ $X_a^2=\frac{g_{cp}^1-g_a}{g_a-g_{cp}^2}X_a^1=3.0484$
g_{1cp}^1, g_{2cp}^1 g_{1cp}^2, g_{2cp}^2	Brass grade in brass pulp and gangue pulp	0.417, 0.62 0.0034, 0.0034
L_{cu}	Chalcopyrite brass grade	0.421
q_{1c}, q_{2c}	Concentrate pulp flow	7.392, 8.103 $(\mathrm{m}^3/\mathrm{min})$
$H_1=H_2$	Total height	3.2 m
g_a	Feed mineral grade	0.0234

First, we implement Algorithm 2 to learn the optimal set-point. Set the sampling interval of the lower-layer device loops and the updating period of the upper-layer operational control loop to be 1min and 30 min, respectively. Fig. 3 demonstrates the convergence results of H^{j+1}, K_1^j and K_2^j. The learned optimal set-point gain K_1^* is presented in (46). The optimal set-point can be further approximated by combining with Rule 1 and Rule 2.

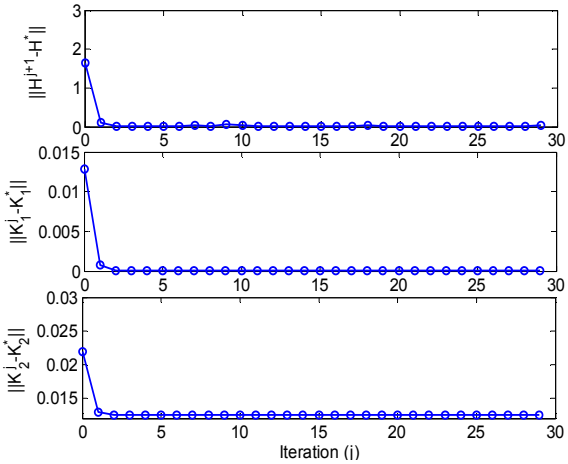

Fig. 3. Convergence results of parameters

$$K_1^*=[K_{11}^*\ K_{12}^*\ K_{13}^*]\tag{46}$$

where

$$K_{11}^*=\begin{bmatrix}-0.0002 & 0.0000 & 0.0038 & -0.1671\\ 0.0002 & -0.0000 & -0.0044 & 0.1919\\ -0.0001 & 0.0000 & 0.0024 & -0.1048\\ 0.0002 & -0.0000 & -0.0043 & 0.1869\end{bmatrix},$$

$$K_{12}^*=\begin{bmatrix}-0.0114 & 0.0127 & -0.0001 & 0.0000 & 0.0054\\ 0.0131 & -0.0146 & 0.0001 & -0.0000 & -0.0062\\ -0.0071 & 0.0080 & -0.0001 & 0.0000 & 0.0034\\ 0.0127 & -0.0142 & 0.0001 & -0.0000 & -0.0060\end{bmatrix},$$

$$K_{13}^*=\begin{bmatrix}-0.1858 & -0.0167 & 0.0296 & 0 & -0.0004\\ 0.2134 & 0.0192 & -0.0339 & 0 & 0.0004\\ -0.1164 & -0.0104 & 0.0185 & 0 & -0.0002\\ 0.2076 & 0.0186 & -0.0329 & 0 & 0.0004\end{bmatrix},$$

In the real operation of rougher flotation process, it is hard to avoid external disturbance and measurement errors. Hence, the external disturbance signal and measurement noise are assumed as $0.05e^{-0.001t}\cos(100v(t))$ $(v(t)=[v_1(t)\ v_2(t)\ v_3(t)\ v_4(t)]^{\mathrm{T}}$ is a vector with $v_i(t)\in[0,1], i=1,2,3,4)$ and put into cell 1 and cell 2, respectively.

Then, three cases of the updating periods of the upper-layer operational control loop (a): 10 min, (b): 30 min and (c): 120 min are taken into account for showing the tracking results of the set-points and the economic operational index. From Fig. 5(a), one can find that the longer it takes to follow the desired economic benefit, the slower updating period of upper-layer

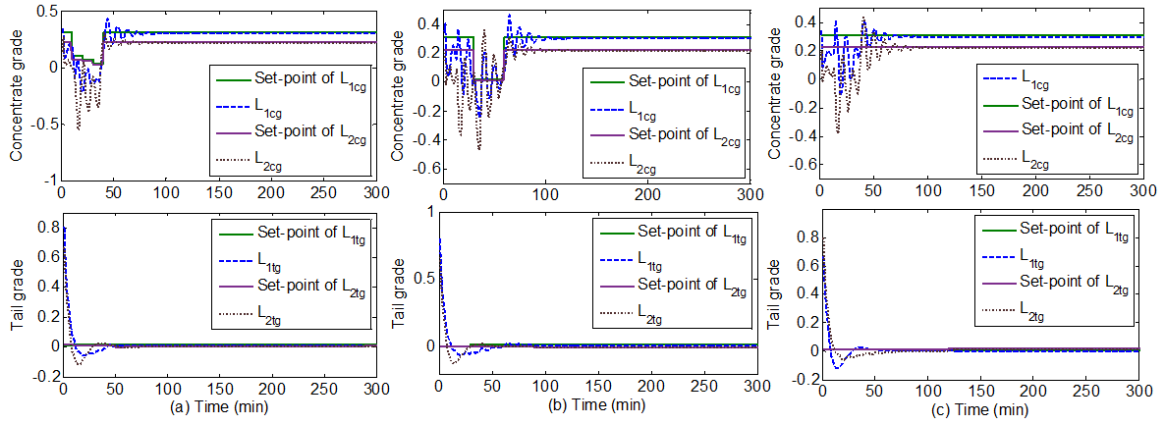

Fig. 4. Tracking performance of the optimal set-points
(a): Case a; (b): Case b; (c): Case c

operational process is. While the fast updating rate of upper-layer operational control loop also can bring increasing computational load. In general, proper updating rate of the upper-layer operational process should be chosen according to the requirements of practical industrial operations. Moreover, the results shown in Fig. 4 and Fig. 5 also indicate that the proposed off-policy Q-learning method has good performance and robustness to both disturbance and measurement errors.

The comparisons with the fixed set-point method and the intelligent control method [1, 2] are made under the same initial operational scenario listed in Table 1. The fixed set-points of cell 1 and cell 2 are respectively chosen as [0.0117 0.01] and [0.1274 0.01]. The set-points in the intelligent control method are adjusted according to the error between the desired economic benefit and the real value. The tracking results of economic benefit and calculated recovery are respectively shown in Fig. 5(b) and Fig. 5(c), where one can see that the better tracking result and high recovery rate are obtained by using the proposed off-policy RL algorithm, since the operators' experience based fixed set-points maybe not the optimal set-points. Moreover, the increments of set-points in the intelligent control method are chosen also based on the operators' experience, then the improper increments could produce negative effects on tracking the desired economic benefit and mineral recovery.

VII. Conclusion

The data-driven OOC problem is addressed for dual-rate rougher flotation operational processes by presenting an off-policy Q-learning algorithm without requiring the knowledge of dynamics of lower-layer rougher flotation cells and economic benefit operation. Combined with zero-sum games method, Q-function based game Bellman equation is derived for solving the optimal set-point and the worst disturbance. Further, an off-policy game Bellman equation is presented to find improved target set-point by evaluating the target set-point and the target disturbance using collected data. Simulations are implemented to show the effectiveness of the proposed method.

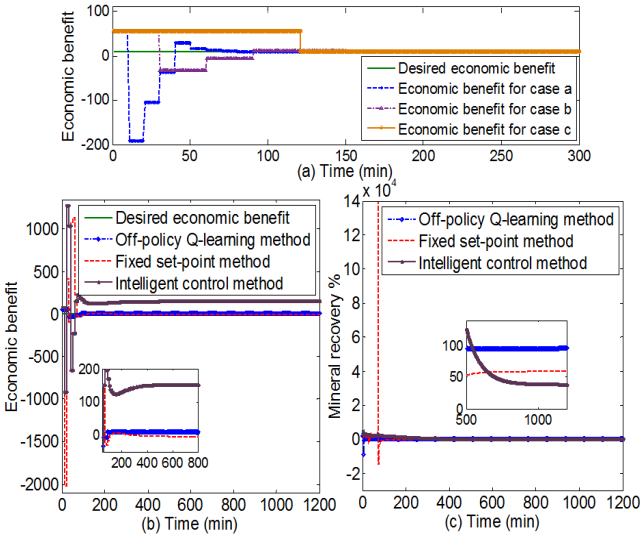

Fig. 5. Tracking performance of the economic benefit and calculated recovery

References

[1] T. Y. Chai, S. J. Qin, and H. Wang, "Optimal operational control for complex industrial processes," *Annual Reviews in Control*, vol. 38, no. 1, pp. 81-92, Feb. 2014.

[2] T. Y. Chai, J. L. Ding, and F. Wu, "Hybrid intelligent control for optimal operation of shaft furnace roasting process," *Control Engineering Practice*, vol. 19, no. 3, pp. 264-275, Mar. 2011.

[3] J. L. Ding, H. Modares, T. Y. Chai, and F. L. Lewis, "Data-Based Multi objective Plant-Wide Performance Optimization of Industrial Processes Under Dynamic Environments," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 2, pp. 454-465, Apr. 2016.

[4] R. Zaragoza and J. A. Herbst, "Model based feedforward control scheme for flotation plants," *Minerals & Metallurgical Processing*, vol. 5, no. 4, pp: 177-85, Jan. 1988.

[5] D. Hodouin and K. Najim, "Adaptive control in mineral processing," *CIM bulletin*, vol. 85, no. 965, pp. 70-78, 1992.

[6] A. J. Thornton, "Cautious adaptive control of an industrial flotation circuit," *Minerals Engineering*, vol. 4, no. 12, pp. 1227-1242, Dec. 1991.

[7] Y. Xie, J. Wu, D. Xu, C. Yang, and W. Gui, "Reagent addition control for stibium rougher flotation based on Sensitive Froth Image Features," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 5, pp: 4199-4206, May 2017.

[8] D. G. Hullbert, "Multivariable control of pulp levels in flotation circuits," Preprints IFAC MMM. Sun City, 71-76, 1995.

[9] K. Pulkkinen, R. Ylinen, S. L. Jamaa-Joumela, and M. Jarvensive, "Integrated expert control system for grinding and flotation," IMM Congress. Parkville, 325-334, 1995.

[10] A. Desbiens, D. Hodouin, K. Najim, and F. Flament, "Long-range predictive control of a rougher flotation unit," *Minerals Engineering*, vol. 7, no. 1, pp. 21-37, Jan. 1994.
[11] R. Pérez-Correa, G. González, A. Casali, A. Cipriano, R. Barrera, and E. Zavala, "Dynamic modelling and advanced multivariable control of conventional flotation circuits," *Minerals Engineering*, vol. 11, no. 4, pp. 333-346, Apr. 1998.
[12] C. Duchesnt, A. Bouajila, G. Bartolacci, P. Pelletier, Y. Breau, J. Fournier, and D. Girade, "Application of multivariate image analysis (MIA) to predict concentrate grade in froth flotation processes," in Proc. 35th Annual Meeting of the Canadian Mineral Processors, Ottawa, 511-526, 2003.
[13] J. P. Ferreira and B. K. Loveday, "Improved model for simulation of flotation circuits," *Minerals Engineering*, vol. 13, no. 14-15, pp. 1441-1453, Dec. 2000.
[14] D. Rojas and A. Cipriano, "Model based predictive control of a rougher flotation circuit considering grade estimation in intermediate cells," *Dyna*, vol. 78, no. 166, pp. 29-37, Apr. 2011.
[15] M. Maldonado, D. Sbarbaro, and E. Lizama, "Optimal control of a rougher flotation process based on dynamic programming," *Minerals Engineering*, vol. 20, no. 3, pp. 221-32, Oct. 2007.
[16] S. J. Qin and T. A. Badgwell, "A survey of industrial model predictive control technology," *Control Engineering Practice*, vol. 11, no. 7, pp. 733-764, Jul. 2003.
[17] T. Y. Chai, L. Zhao, J. Qiu, and F. Liu, "Integrated network-based model predictive control for setpoints compensation in industrial processes," *IEEE Transactions on Industrial Informatics*, vol. 9, no. 1, pp. 417-426, Feb. 2013.
[18] F. Liu, H. Gao, J. Qiu and S. Yin, "Networked multirate output feedback control for setpoints compensation and its application to rougher flotation process," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 1, pp. 460-468, Jan. 2014.
[19] J. L Fan, Y. Jiang and T. Y. Chai, "Operational Feedback Control of Industrial Processes in a Wireless Network Environment," *Acta Automatica Sinica*, vol. 42, no. 8,pp. 1166-1174, Aug. 2016.
[20] W. Dai, T. Y. Chai, and S. X. Yang, "Data-driven optimization control for safety operation of hematite grinding process," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 5, pp. 2930-2941, Oct. 2015.
[21] Z. Wu, Y. Wu, T. Y. Chai, and J. Sun, "Data-driven abnormal condition identification and self-healing control system for fused magnesium furnace," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 3, pp. 1703-1715, Aug. 2015.
[22] Q. L. Wei, D. R. Liu, and G. Shi, "A novel dual iterative Q-learning method for optimal battery management in smart residential environments," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 4, pp. 2509-2518, Apr. 2015.
[23] J. Y. Lee, J. B. Park, and Y. H. Choi, "Integral Q-learning and explorized policy iteration for adaptive optimal control of continuous-time linear systems," *Automatica,* vol. 48, no. 11, pp. 2850-2859, Nov. 2012.
[24] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, "Model-free Q-learning designs for discrete-time zero-sum games with application to H-infinity control," *Automatica*, vol. 43, pp. 473-481, March 2007.
[25] B. Kiumarsi, F. L. Lewis, and Z. P. Jiang, "H∞ control of linear discrete-time systems: Off-policy reinforcement learning," *Automatica*, vol. 37, no. 1, pp. 144-152, Apr. 2017.
[26] H. Modares, F. L. Lewis, and Z. P. Jiang, "H∞ tracking control of completely unknown continuous-time systems via off-policy reinforcement learning," *IEEE Transactions on Neural Networks & Learning Systems,* vol. 26, no. 10, pp. 2550-2562, Oct. 2015.
[27] B. Luo and H. N. Wu, "Computationally efficient simultaneous policy update algorithm for nonlinear H∞ state feedback control with Galerkin's method," *International Journal of Robust & Nonlinear Control*, vol. 23, no. 9, pp. 991–1012, Apr. 2013.
[28] T. Başar and P. Bernard, H∞ optimal control and related minimax design problems : a dynamic game approach. Basel: Birkhäuser, 1995.
[29] A. Al-Tamimi, M. Abu-Khalaf and F. L. Lewis, "Adaptive critic designs for discrete-time zero-sum games with application to control," *IEEE Transactions on Systems Man & Cybernetics Part B Cybernetics,* vol. 37, no. 1, pp. 240-247, Feb. 2007.
[30] B. Kiumarsi, F. L. Lewis, H. Modares, A. Karimpour, and M. Naghibi-Sistani, "Reinforcement Q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics," *Automatica*, vol. 50, no. 4, pp. 1167-1175, Jan. 2014.
[31] H. Modares and F. L. Lewis, "Linear quadratic tracking control of partially unknown continuous-time systems using reinforcement learning," *IEEE Transactions on Automatic Control,* vol. 59, no. 11, pp. 3051-3056, Nov. 2014.