Design of a Resistive Brake Controller for Power System Stability Enhancement Using Reinforcement Learning

Mevludin Glavic

***Abstract*—Computation of the closed-loop control laws, capable to realize multiple switching operations of a resistive brake (RB) aimed to enhance power system stability, is the primary topic of this brief. The problem is formulated as a multistage decision problem and use of a model-based reinforcement learning (RL) method, known as prioritized sweeping, to compute the control law is considered. To illustrate the performances of the proposed approach results obtained using the model of a synthetic four-machine power system are given. Handling measurement transmission delays is discussed and illustrated.**

***Index Terms*—Closed-loop control, multiple switching, power system stability, reinforcement learning (RL), resistive brake (RB).**

I. Introduction

RESISTIVE brake (RB) (the terms braking resistor and dynamic brake are also in use) have been recognized and used as a cost-effective measure for transient stability enhancement for a long time [1]–[5]. The essence of the control is the insertion of a resistance, usually at a generation bus (mainly hydraulic generating station), upon the clearing of a system disturbance.

The key problems in a RB controller design include choosing appropriate input signal and adopting a proper control scheme to decide when to switch ON or OFF the resistor, or so called "switching times," in order to meet the specific stabilization objectives [5], [6]. A comprehensive survey of early considerations and implementations of mechanically switched RB is reported in [5]. The prevailing approach in these early implementations was to apply only one switch of the brake for a prespecified insertion time [3]–[6], and the control initiation is based on the recognition of the prespecified system variable changes. These variables include: generator speed, generator angle, the system frequency, active power at generator terminals, generator voltage magnitude, or some combinations of those.

The coarse switching times control, introduction of thyristor-switched and thyristor-controlled RBs, and advancements in RB technology gave rise to the considerations of advanced methods application in design of a RB controllers. A variety of approaches have been considered and proposed with the aim of full utilization of RB potential to provide transient damping after large disturbances. These approaches include application of dynamic programming method [7], variable structure control [8], [9], optimal control theory [10], [11], rule-based system [12], fuzzy logic control [13], and artificial neural networks [14]. Recently, the potential of using RB to damp power swings have been investigated [15] and the use of multiple switching operations of the RB for a transient stability emergency control have been reported in [16].

The advantage of automatic control strategies capable to realize multiple insertions of RB has been recognized in [4] and [7] and recently confirmed as a control strategy of the choice in the real-life system implementation [16]. A need exists for RB control algorithms, which are robust, closed-loop in nature, and are more systematically designed.

In this brief, the problem of multiple switching of a thyristor-switched resistive brake (TSRB) is formulated as a multistage decision problem. [17] provides a formal framework to solve this problem and in this brief its application to compute a closed-loop control law of a TSRB with the solution of the [17] being approximated by using a reinforcement learning (RL) algorithm [18]–[20], the subject of increasing interest in power system control applications [21]–[24], is considered. A TSRB is aimed to damp electromechanical oscillations and to avoid the system loss of synchronism taking into account limits of the brake.

The organization of the brief is as follows. RL theoretical framework is given in Section II. In Section III, the four-machine power system model is described. The controller design and simulation results are given in Section IV together with the adopted way to handle communication delays. Discussion is provided in Section V while a conclusion is given in Section VI.

II. Reinforcement Learning

RL is a computational approach to learning from interactions with a system or its simulation model (by trial-and-error). In this brief, a problem how to control a TSRB is considered and natural choice of theoretical framework to present RL is to consider it as a way to learning (approximate) solutions of optimal control problems.

A. Theoretical Framework

RL is presented here in the framework of discrete optimal control of a deterministic nonlinear system with constant sampling period. If x_t represents the sampled state vector of the system at instant t, u_t the control action taken at t, then the state vector of the system at instant $t+1$ is given by

$$x_{t+1} = f(x_t, u_t). \tag{1}$$

The RL method used in this brief belongs to the temporal difference type of methods that suppose the existence of a reward $r_t(x_t, u_t) \in \Re$ ($\Re$ is the set of real numbers) associated to the transition from x_t to x_{t+1} while taking action $u_t \in U, \forall t \in$

Manuscript received August 16, 2004. Manuscript received in final form December 20, 2004. Recommended by Associate Editor I. Kolmanovsky.

The author is an independent consultant in Bosnia. He was with the University of Liege, Department of Electrical Engineering and Computer Science, B-4000 Liege, Belgium (e-mail: mglavic@ieee.org).

Digital Object Identifier 10.1109/TCST.2005.847339

$[0, 1, 2, \ldots)$ (U is assumed to be finite) [18]. The discounted return $R(x_0, u_0, u_1, u_2, \ldots)$, which depends on the initial state x_0 and on the sequence of control actions $u_{\{t\}} = (u_0, u_1, u_2, \ldots)$, applied to the system, is defined as

$$R\left(x_0, u_{\{t\}}\right) = \sum_{t=0}^{\infty} \gamma^t r(x_t, u_t). \tag{2}$$

where γ, $(0 \leq \gamma < 1)$, is a parameter called the discount rate.

The aim of RL methods in the framework of infinite time horizon with discounted reward is to find, for every possible initial state x_0, a good approximation of the optimal control sequence $u^*_{\{t\}}(x_0)$ that maximizes the discounted return. A policy $u_{\{t\}}$ is said to be better or equal to a policy $u'_{\{t\}}$ if its return is greater than or equal to that of $u'_{\{t\}}$ for all states. There is always at least one policy that is better than or equal to all other policies. This is an optimal policy. In order to determine this policy, one defines the value function $V(x)$

$$V(x) = \max_{u_{\{t\}}} R\left(x, u_{\{t\}}\right). \tag{3}$$

Using the dynamic programming (DP) principle [17], it can be proven that the value function satisfies the condition

$$V(x) = \max_{u \in U} \left(r(x, u) + \gamma V\left(f(x, u)\right)\right) \tag{4}$$

where $r(x, u)$ and $f(x, u)$ are, respectively, the reward observed and the next state reached when taking action u while being in state x. DP computes the value function in order to find the optimal control with a feedback control policy. Indeed, from the value function the following optimal feedback control policy is deduced

$$u^*(x) = \arg\max_{u \in U} \left(r(x, u) + \gamma V\left(f(x, u)\right)\right). \tag{5}$$

Alternatively, one can define so-called Q function as

$$Q(x, u) = r(x, u) + \gamma V\left(f(x, u)\right). \tag{6}$$

Then $V(x)$ can be expressed as a function of $Q(x, u)$

$$V(x) = \max_{u \in U} Q(x, u). \tag{7}$$

Equation (5) can be rewritten as

$$u^*(x) = \arg\max_{u \in U} Q(x, u). \tag{8}$$

Equation (8) provides a straightforward way to determine the optimal control law from the knowledge of the Q.

RL algorithms estimate the Q function by interacting with the system. From the knowledge of the Q function, they can decide by using (8) which value of the control to associate to a state in order to maximize the discounted return (2). Unfortunately, RL in a continuous state-space implies that the Q function has to be approximated [18]. A discretization technique is used in this brief to approximate it because it is easy to implement and revealed numerically stable in the simulations performed.

B. State Space Discretization

A discretization technique consists in dividing the state space into a finite number of regions and then considering that on each region the Q function depends only on u. Then, in the RL algorithms, the notion of state used is not the real state of the system x but rather the region of the state space to which x belongs. The letter s is used rather than x to denote the state of the system in order to stress that refers now not to x itself but to a region of the state space. Moreover, the finite set containing all the discretized states of the system is denoted by S. The discretization of the state space introduces some stochastic aspects. While being in one region of the state space and taking an action, the region of the state space reached at the next sampling instant is not fully determined. The stochastic aspects introduced by the discretization lead to suppose that $Q(s, u)$ does not obey anymore to the deterministic equation (6) but rather to

$$Q(s, u) = r(s, u) + \gamma \sum_{s' \in S} p(s'|s, u) \max_{u \in U} Q(s', u) \tag{9}$$

where $p(s'|s, u)$ represents the probability to reach at the next sampling instant the state s' when being in the state s while taking action u.

Rewards $r(s, u)$ and probabilities $p(s'|s, u)$ describe the model of the discretized system. They associate to each discretized state and to each value of the control u transition probabilities to other states and the value of a reward. Assuming that they describe a Markov decision process (MDP), $Q(s, u)$ can be easily estimated using a classical DP algorithm for solving MDP like the value iteration or the policy iteration [18], [19]. The optimal control to associate to a state is the one that maximizes Q for this state.

RL methods either estimate the transition probabilities and the associated rewards (model based learning methods) and then compute the Q function, or compute directly the Q function without learning any model (nonmodel based learning methods) [18]. In this brief, a model based algorithm is used because these algorithms offer some important advantages in comparison to nonmodel based, and those are: more efficient use of data gathered, they find better policies, and handle changes in the environment more efficiently [20].

C. Generic Model Based RL Algorithm

The algorithm is given in Table I [18], [20]. The ε-greedy method used to choose the action suggests that there is probability ε that the action chosen is not necessary the one which minimizes Q, but an action taken at random. This provides the algorithm with some exploratory behavior such that on average each $1/\varepsilon$ time a random action is taken.

The N function used in this algorithm does not intervene to describe the model as such but is necessary for its updating. The term $\beta (0 \leq \beta \leq 1)$ provides the algorithm with some adaptive behavior by giving more importance (if $\beta < 1$) to the last data acquired.

TABLE I
GENERIC ALGORITHM FOR MODEL BASED LEARNING METHOD

Initialize $Q(s,u)=0, \quad \forall s\in S \ \text{and} \ \forall u\in U$

Initialize parameters of the model:

$N(s'|s,u)=0, \quad \forall s,s'\in S \quad \text{and} \quad \forall u\in U$

$p(s'|s,u)=0, \quad \forall s,s'\in S \quad \text{and} \quad \forall u\in U$

$r(s,u)=0, \quad \forall s\in S \quad \text{and} \quad \forall u\in U$

Do forever:

Observe current state s

Choose action u from s using knowledge of Q (e.g. $\varepsilon-greedy$)

Take action u and observe s' and r.

Update model:

$$N(i|s,u)\leftarrow \beta N(i|s,u), \quad \forall i\in S$$

$$r(s,u)\leftarrow \frac{r(s,u)\sum_{j\in S}N(j|s,u)+r}{\sum_{j\in S}N(j|s,u)+1},$$

$$N(s'|s,u)\leftarrow N(s'|s,u)+1,$$

$$p(i|s,u)\leftarrow \frac{N(i|s,u)}{\sum_{j\in S}N(j|s,u)}, \quad \forall i\in S$$

Compute Q by solving (9)

$s\leftarrow s'$

III. DESCRIPTION OF THE TEST POWER SYSTEM, LEARNING SCENARIOS, AND CONTROL LAW LEARNED

To illustrate capabilities of the proposed control this brief makes use of the four-machine power system, described in Fig. 1. Its characteristics are mainly inspired from [1].

Detailed description of the system model is given in the Appendix. While the system operates in steady-state conditions, the generators G1, G2 (hydro) and G3, G4 (thermal) produce approximately the same active powers (700 MW) and the two loads L7, L10 consume, respectively, 990 and 1790 MW. The TSBR is located at bus 6 and sized as $g = 5.0$ p.u. mhos on a 100-megaVoltAmpere (MVA) base.

A. Technological Underpinnings

RBs currently in use [3], [5], [7], [8] are large size brakes. The sizing of the RB is a point of great interest. From a control point of view, the bigger the brake the better. From an engineering point of view, increasing the size of the brake increases its cost and maintenance. To be viable for stability enhancement the RB must be economical to fabricate and has low maintenance cost.

Controller design introduced in this brief permits the use of smaller brakes with lower cost. The stability improvement comes from multiple switching of the RB. Several technological solutions for RB are available: bulk metallic, bulk nonmetallic, and thick film technologies. Of the three types, thick film resistors are felt to be most appropriate. They are low cost, have low inductance, and are virtually maintenance free. Although heat transfer can be an issue and a longer OFF time can be necessary it is possible to use a bank of smaller RB of the same size that can individually be switchedON during a short time period.

Advancements in communications technology (most appealing is the use of GPS technology in conjecture with phasor measurement units [22]–[24]) allow fast and accurate collection of the synchronized measurements across wide geographical areas. This is the reason centralized control scheme is considered in proposed design methodology.

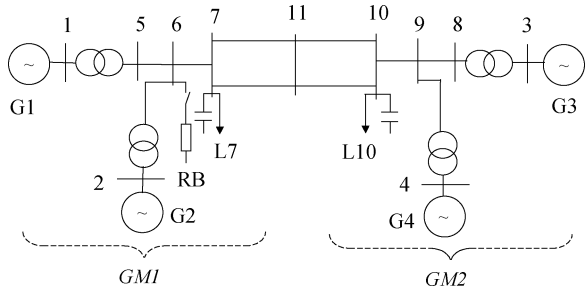


Fig. 1. Four-machine power system.

B. Learning Scenarios Description

The learning period is partitioned into different scenarios. Each scenario starts with the power system being at rest and is such that at 10 s a short circuit (self-clearing) at bus 10 occurs. The fault duration is chosen at random in the interval [0,350 ms]. The scenario stops either when the instability is reached or when t is greater than 60 s. No learning is realized during the fault period.

C. Control Law Learned

The RL algorithm is used to learn an approximation of the optimal closed-loop control law (strictly speaking, the closed-loop control law learned will be different from the optimal one due to the facts that the input signal of the RL algorithm is discretized and represents something else than the system real state). But to each power system configuration corresponds an optimal control law. The strategy proposed here is to realize the learning by using always the same configuration and to assess the control law robustness to justify the use of the control law in configurations that do not correspond to the one in which the learning has been done.

IV. CONTROLLER DESIGN AND SIMULATION RESULTS

The aim is to design the controller in order to be able to control particular system mode. The mode considered is relative motion of one group of the machines (identified by $GM1$ in Fig. 1) with respect to another $(GM2)$. The one machine-infinite bus (OMIB) transformation [25] is applied to this one. The transformation reduces dimensionality and resolve curse of dimensionality problem, the problem inherent to the most of RL algorithms. Having identified the two groups of machines the transformation proceeds as follows (let A denotes the set of machines in the group $GM1$, machines 1 and 2 in Fig. 1).

- Transform the two groups into two equivalent machines, using their corresponding partial center of angle. For group $GM1$ this results in

$$\delta_{GM1t}=M_{GM1}^{-1}\sum_{k\in A}M_k\delta_{kt} \tag{10}$$

$$\omega_{GM1t}=M_{GM1}^{-1}\sum_{k\in A}M_k\omega_{kt}; \quad M_{GM1}=\sum_{k\in A}M_k \tag{11}$$

where δ_{kt} and ω_{kt} denote the machines internal angle and the rotor speed deviation, and M_k represents the machines inertia. Similar expressions hold for group $GM2$.

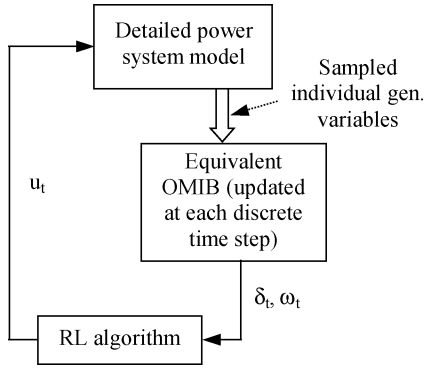


Fig. 2. RL algorithm interaction with the system model.

- Reduce the two-machine system into an equivalent OMIB system whose machine angle and speed deviation are defined by

$$\delta_t = \delta_{GM1t} - \delta_{GM2t}; \quad \omega_t = \omega_{GM1t} - \omega_{GM2t}. \tag{12}$$

The system is described by 30 variables (16 state and 14 algebraic, see the Appendix) that are, using OMIB transformation, *a priori* reduced to a two-dimensional (2-D) signal composed of relative angle and relative speed of the two groups of machines. Of course, the amount of information in these two variables is less than in the 30 variables but will be sufficient according to simulations performed to obtain, after the learning, a good quality closed-loop control law. An RL algorithm interaction with the system simulation model is illustrated in Fig. 2.

At each discrete time step the RL algorithm receives a representation of the system state (equivalent OMIB angle and speed deviation, and these define the state for the RL algorithm). RL algorithm selects an action from the set of actions available in the state. As a consequence of taking action the algorithm receives a numerical reward that the algorithm tries to maximize over time.

A. State Definition

It is assumed that the angle, speed deviation, electrical and mechanical power (individual generator variables in Fig. 2, see the Appendix) of each generator is available (they can be either measured directly or estimated). The transmission delays and measurement errors are neglected (this issue is discussed in Section IV-I). The state for RL algorithm (to be differentiated from the system state) at time t is, thus, represented as

$$s_t = (\delta_t, \omega_t) \tag{13}$$

where δ_t, ω_t are equivalent OMIB angle and speed deviation.

B. Reward Definition

It is critical that the rewards truly indicate what is wanted to be accomplished, not how it is wanted to be achieved [19]. For the particular problem considered the aim of the RL controller is threefold: to improve damping of particular system mode, to avoid the loss of synchronism between the generators when a severe incident occurs, and to limit the time the RB is switched ON. The oscillations are observable in the magnitude of the equivalent speed deviation, and the aim of the controller is to limit its magnitude. All these can be accomplished by defining the reward as

$$r(s_t, u_t) = \begin{cases} -|\omega_t| - c \cdot u_t, & \text{if } |\delta_t| \leq \pi \text{ rad} \\ -1000, & \text{if } |\delta_t| > \pi \text{ rad} \end{cases} \tag{14}$$

where the $u_t \in \{0, 1\}$ (0 meaning that the brake is switched OFF and 1 ON) and where c determines how much the fact that the brake is ON is penalized. To strongly penalize unstable operation, a very large negative reward (-1000) is obtained when the system has lost synchronism. A widely used heuristic criterion for detecting system loss of synchronism is the value of angular separation between the two groups of machines. Based on preliminary simulations, for particular system this is considered to be the case when $|\delta_t| > \pi$ rad(appreciable increase in speed deviations does not necessarily imply that the synchronism is lost).

Observe that the equivalent OMIB angle is not included into reward definition to avoid problems with estimating its value in post-fault system equilibrium. For the OMIB speed deviation there is no need to estimate its value in post-fault equilibrium because it is *a priori* known that individual generator speed deviations are 0 if the system is in equilibrium.

C. Values of Parameters

The period between two samplings is chosen equal to 50 ms. Large value of γ implies the algorithm will take long-term benefit control actions. However, a too large value (a value close to 1) can lead to convergence problems. Simulations carried out have shown that $\gamma = 0.95$ represents a reasonable tradeoff. The value of parameter c in (14) is chosen as $c = 2.0$. ε-greedy factor is set to 0.1 which means that a random action will be taken at each tenth sampling on average. The factor ε is set to rather high value to encourage the RL algorithm exploration. The equivalent angle and speed are uniformly discretized in 100 values within the intervals [−3.15,3.15] rad and [−10,10] rad/s.

D. Learned Control Policy

Fig. 3(a) shows the control laws obtained in the (δ, ω) plane after 100 scenarios have been presented to the RL algorithm. In this figure, each tile corresponds to a discretized state. The black tiles correspond to states where the control value is 1 and the light ones to the opposite case. Observe that after 100 iterations, the control law still seems rather erratic, which is due to the fact the RL algorithm has not yet converged. After 1000 scenarios [Fig. 3(b)], on the other hand, one can observe that organized structure has appeared in the way the tiles are distributed. At this stage, additional learning can only bring minor changes to the learned control law. The total number of scenarios generated is 1000 out of which 163 were unstable.

E. Enlarging of the Stability Domain

For the 350 ms duration self-clearing fault, the uncontrolled system loses stability 1.75 s after the fault clearance (the maximum fault duration it can withstand without losing stability is 215 ms), but using learned control law the controller stabilizes the system. The evolution of the equivalent angle, speed deviation, and control actions taken are represented in Fig. 4(a).

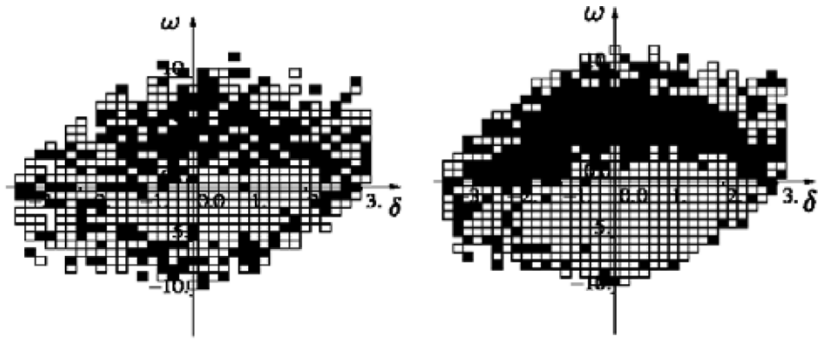


Fig. 3. Learned control policy (δ is expressed in rad and ω in rad/s).

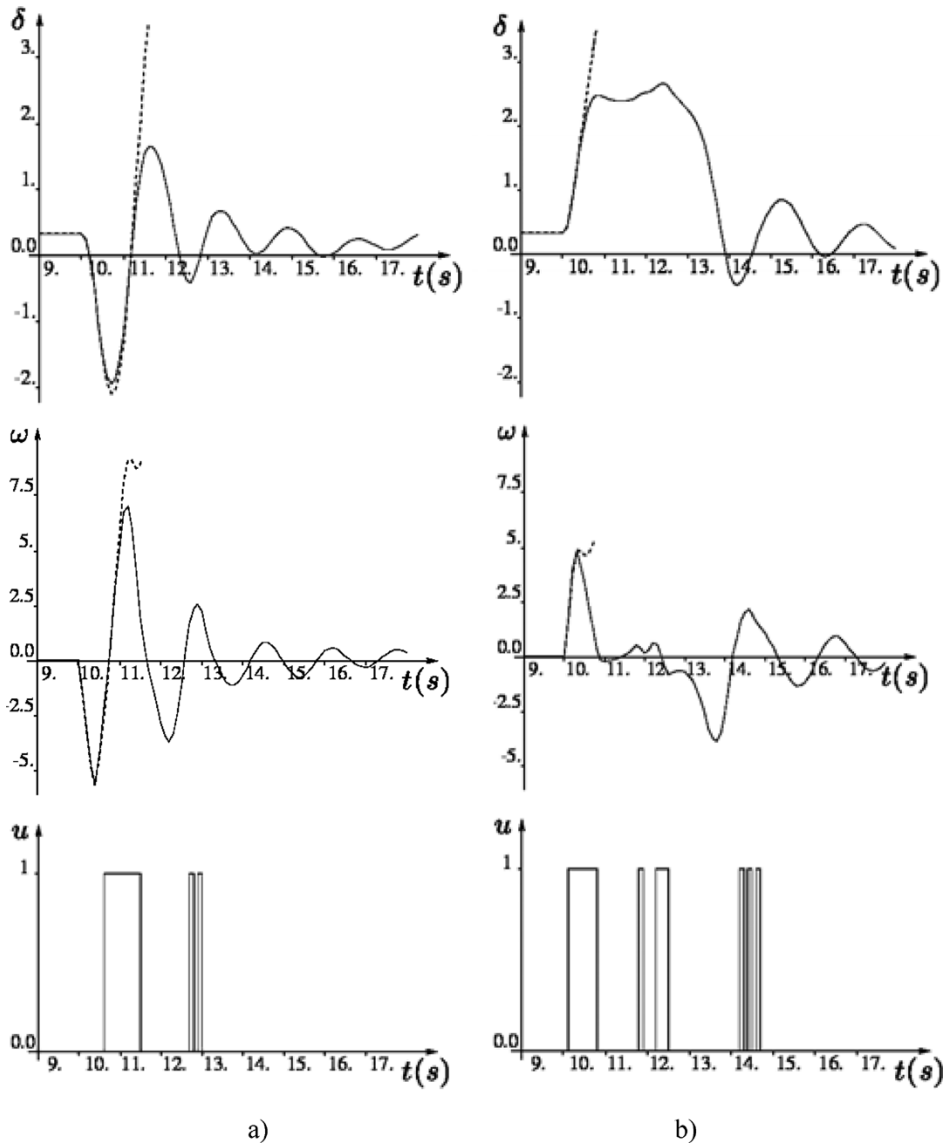


Fig. 4. Evolution of δ (rad), ω (rad/s), and u for two different fault scenarios. The dashed curves represent the evolution in the case of uncontrolled system.

F. Control Law Robustness

To assess robustness of the proposed control, the learned control law is used to control the system when subjected to different fault scenarios than those used in the learning. The faults considered include the following.

1) Short circuit applied near bus 7 and cleared by opening one of the two lines connecting bus 7 to bus 10 (change of the fault location + change in the system configuration).
2) Self-clearing short circuit applied near bus 7 (change of the fault location) with modified system prefault con-

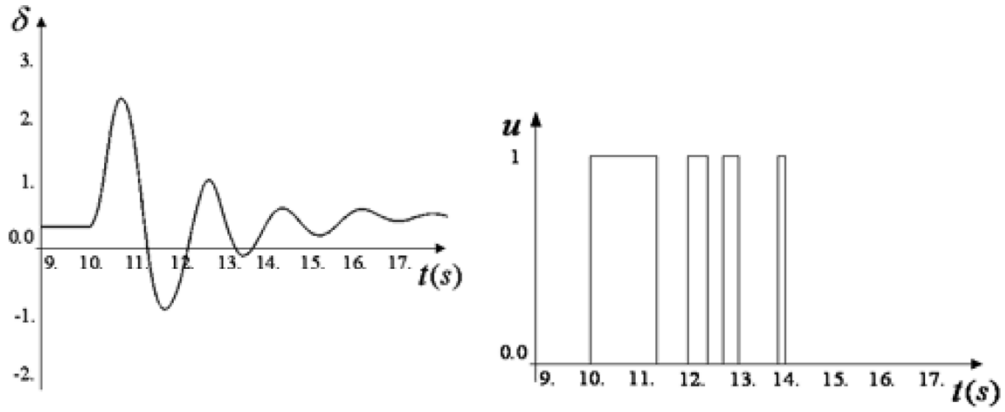


Fig. 5. Evolution of δ (rad) and u for prefault conditions different than during the learning.

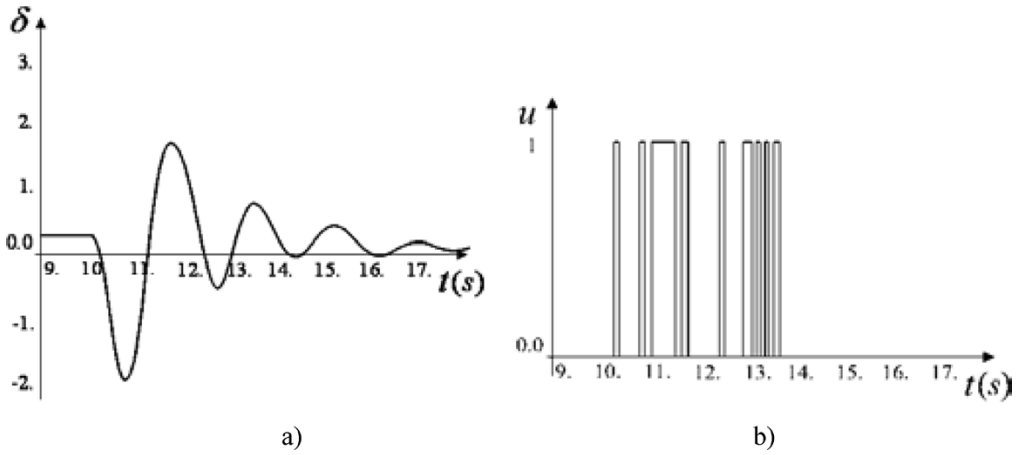


Fig. 6. Evolution of δ (rad) and u for the 350-ms self-clearing fault and $c = 0.5$.

TABLE II
SUMMARY OF THE SIMULATIONS FOR ROBUSTNES CHECK

Case	Number of simulations	Fault duration drawn at random from interval	Maximum fault duration without loss of stability	
			Uncontrolled	Controlled
1	10	150-250 ms	141 ms	252 ms
2	10	100-320 ms	200 ms	330 ms

ditions (active power production of generators within the group $GM1$ have been increased by 20 MW).

Summary of the simulations performed is given in Table II.

The system response and actions taken are illustrated in Fig. 4(b) for the first case with fault duration 225 ms and in Fig. 5 for the case with changed system prefault conditions and the fault duration 300 ms.

In spite of the change in system configuration and the system prefault conditions, the controller succeeds to control efficiently the system being subjected to the "unseen" scenario in all 20 simulations. Thus, the learned control law is robust to these changes (changes usually considered in checking robustness of power system controllers).

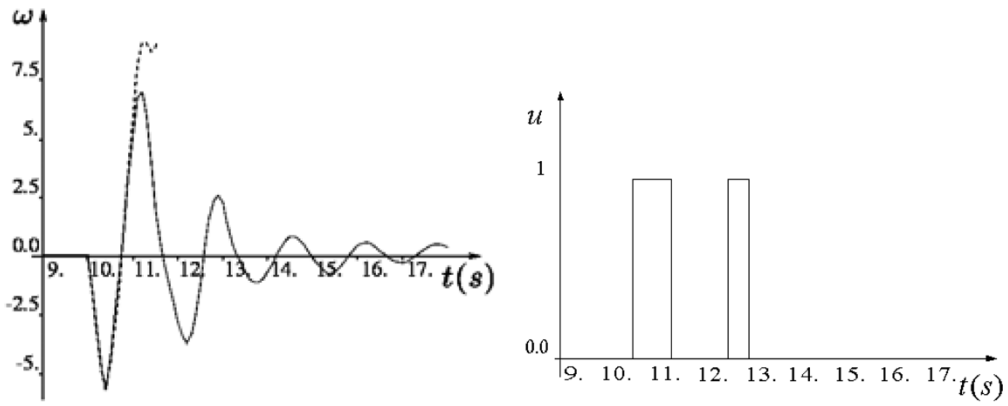
G. Influence of Penalizing Control Efforts

To assess the influence of parameter c on the approximation of the optimal control law new 1000 scenarios were generated, out of which 115 unstable, (different number of unstable scenarios is mainly due to random choice of the fault duration) with parameter c set to 0.5. The response of the controlled system subjected to the 350 ms duration self-clearing fault is represented in Fig. 6(a) and control actions taken in Fig. 6(b).

Note that, due to lower penalization of the fact that the brake is ON, the controller stabilizes the system using 9 brake switches while in the case when c is set to 2.0 (Fig. 4) only three switches were sufficient. This illustrates the main purpose of including parameter c in reward definition, i.e., by careful choice of the value of this parameter it is possible to accommodate different technological constraints inherent to different types of RB (maximum insertion time, maximum number of consecutive switches, etc.).

H. Further Increase in Control Flexibility

One of the attractions of RL approach is the flexibility this approach provides while designing controllers for a given problem. In previous subsections, it is demonstrated how some limitations can be accommodated through the proper choice of parameter c. Another limitation can be the fact that technological solution for the brake is such that after each ON period when the brake goes OFF must stay OFF for some time. Assume that the time the brake must stay OFF is $t_{\min}$. This limitation



a) b)

Fig. 7. Evolution of ω (rad/s) and u for the 350-ms self-clearing fault: $c = 2.0$, transmission delay 50 ms and $t_{min} = 0.2$ s.

can be easily incorporated into the control scheme by proper reward definition and one possibility is to define the reward as

$$r(s_t, u_t) = \begin{cases} -|\omega_t| - c \cdot u_t, & \text{if } |\delta_t| \leq \pi \text{ rad} \\ -1000, & \text{if } \begin{matrix} u_{t-1}=0 \text{ and} \\ u_t=1 \text{ and } t_{\text{off}}<t_{\min} \end{matrix} \\ -1000, & \text{if } |\delta_t| > \pi \text{ rad} \end{cases}. \quad (15)$$

This reward definition requires change of the state definition in respect to one given by (13), and the state must be extended to include a short history of actions taken

$$s_t = (\delta_t, \omega_t, u_{t-1}, u_t). \quad (16)$$

Variable t_{off} has to be reset to 0 whenever the control goes from ON to OFF.

I. Handling Transmission Delays

The purpose of the simulation results presented in previous subsections is to highlight the potential and flexibility in applying RL to design the RB controller. However, the transmission delays, not taken into account, are important issue that may have detrimental effect on final controller performances. Very recent theoretical results about MDP with delays and asynchronous cost collection [26] offer a sound solution to this problem. The main result of [26] is that an MDP with delays may be reformulated as an MDP without delays with augmented state space. The approach is similar to the one used in partially observable MDP and consist in defining the controller state from the history of measurements and actions taken. In the presence of measurement transmission delays, the controller cannot observe the system's current state and, by extension, it cannot observe the immediate reward. Instead, it observes the state the system was in τ (a constant transmission delay) stages before. The history of the controller's observations and actions up to time step t is

$$H_t = (s_0, u_0, \ldots, s_{t-\tau}, u_{t-\tau}, \ldots, u_{t-1}) \quad (17)$$

and defines a probability distribution over possible current states $s_t \in S$. From the Markov property of the system's state transitions, it follows that [26]:

$$\begin{aligned} p(s_t|H_t) &= p(s_t|s_0, u_0, \ldots, s_{t-\tau}, u_{t-\tau}, \ldots, u_{t-1}) \\ &= p(s_t|s_{t-\tau}, u_{t-\tau}, \ldots, u_{t-1}). \end{aligned} \quad (18)$$

Therefore, the truncated history

$$H'_t = (s_{t-\tau}, u_{t-\tau}, \ldots, u_{t-1}) \quad (19)$$

constitutes a sufficient statistic for the decision process and can be considered as the new state of the controller. Further considerations in [26] revealed that the reward should be defined based on $s_{t-\tau}$ and $u_{t-\tau}$.

For the particular problem considered in this brief, the state and reward are redefined to accommodate a constant transmission delay (delay in applying control action is not considered here) as

$$s_t = (\delta_{t-\tau}, \omega_{t-\tau}, u_{t-\tau}, \ldots, u_{t-1}) \quad (20)$$

$$r(s_{t-\tau}, u_{t-\tau}) = \begin{cases} -|\omega_{t-\tau}| - c \cdot u_{t-\tau}, & \text{if } |\delta_{t-\tau}| \leq \pi \text{ rad} \\ -1000, & \text{if } \begin{matrix} u_{t-\tau-1}=0 \text{ and} \\ u_{t-\tau}=1 \text{ and } t_{\text{off}}<t_{\min} \end{matrix} \\ -1000, & \text{if } |\delta_{t-\tau}| > \pi \text{ rad} \end{cases}. \quad (21)$$

Note that the reward used to increase flexibility of the control (15) is used here. The increase in the controller state dimension requires more learning scenarios to be performed.

For particular case considered 2000 scenarios (out of which 225 were unstable) were necessary to achieve the controlled system response [Fig. 7(a)] comparable to the one presented in Fig. 4(a). The values of additional parameters are chosen as $t_{\min} = 0.2$ s and τ as an integer multiple of time step (in particular case it is equal to 50 ms). The control actions taken are illustrated in Fig. 7(b). Two brake switches are sufficient to approximately optimally stabilize the system.

V. Discussion

The use of OMIB transformation resolved the curse of dimensionality problem by reducing system state representation to 2-D. Consequently, one may wonder why a classical DP is not used because the DP law itself appears to be numerically computable. However, OMIB transformation brings a key obstacle for classical DP methods to be applied to this problem: a mathematical model of equivalent OMIB system dynamics is not completely available, as described in the following.

Reasoning, intuitive but with numerous practical evidences [25], behind the concept of OMIB transformation is as follows.

1) The most important variables to analyze and assess power system angle stability are angles and speed deviations of individual generators.
2) The loss of synchronism of a multimachine power system originates from the irrevocable separation of its generators into two groups, which can be replaced by a two-machine system and then by an OMIB equivalent.

Accuracy of this transformation, from the rigorous theory stance, is questionable. However, numerous tests (including real-world power systems [25]) revealed if detailed power system model is considered (as in this brief) and the OMIB parameters: angle and speed deviation (12) as well as equivalent OMIB mechanical and electrical power

$$P_{mt} = M\left(M_{GM1}^{-1}\sum_{k\in GM1} P_{mkt} M_{GM2}^{-1}\sum_{j\in GM2} P_{mjt}\right) \quad (22)$$

$$P_{et} = M\left(M_{GM1}^{-1}\sum_{k\in GM1} P_{ekt} M_{GM2}^{-1}\sum_{j\in GM2} P_{ejt}\right) \quad (23)$$

are updated sufficiently often (as in this brief) then OMIB (also referred as generalized OMIB [25]) offers a good quality image of multimachine system time evolution. In expressions (22) and (23), $M = (M_{GM1}M_{GM2})/(M_{GM1} + M_{GM2})$ denotes the equivalent OMIB inertia coefficient. The equivalent OMIBs dynamics, in general, is expressed by

$$\dot{\delta} = \omega \quad (24)$$

$$M\cdot\dot{\omega} = P_m - P_e = f(\delta,\omega). \quad (25)$$

The generalized OMIB transformation does not make any assumption about analytical relation of P_m and P_e on δ and ω and, thus, preventing use of classical DP methods to solve the problem. When a mathematical model of the system or its approximate model dynamics is not, or is just partially, known a possible resort, as argued in this brief, is to use RL to solve problem posed as MDP (the fact that equivalent OMIB angle and speed deviation contain all relevant information for decision making imply Markov property).

The four-machine power system is a synthetic system purposely designed for studying power system inter-area electromechanical oscillations (the system mode to be controlled) [1] and this motivated use of the system in the brief. In the case of a real (large) power system, the major difference, with respect to the four-machine system, is identification of particular system mode to be controlled and two groups of generators that swing against each other. Extensive simulations, on detailed power system model, of different possible system contingencies, can do this. Having identified these groups and with the help of detailed system model one exactly proceeds as described in this brief. Successful applications of OMIB transformation on French, Hydro-Quebec, Brazilian, and Mexican systems [25], give support to the claim that the methodology advocated in the brief is applicable to real power systems. Moreover, communication delays are more pronounced in real (geographically spread) systems and handling these delays is included in the brief with the aim of strengthening methodology practicality.

VI. Conclusion

Recent advancements in RB and power electronics technology seems give a new momentum in RB application in enhancing power system stability. The main advantage of automatic control strategies capable to realize multiple insertions of RB is that they permit use of less capacity RB and alleviate not only the first swing but also the subsequent ones. In this brief, a design of RB controller based on RL is considered with the aim of enhancing damping of first and subsequent swings in the system after large disturbances through multiple switching operations of the RB. The control law flexibility, robustness, and enlarging of the stability domain are demonstrated using a synthetic four-machine power system. The results observed qualify the proposed control as effective to handle the problem considered.

Only constant transmission delays are considered. The future work will include random delays and further improvements along the theoretical results from [26].

Appendix

Power system dynamics in general is governed by the set of differential-algebraic equations. For four-machine power system the dynamics of the kth ($k = 1,2,3,4$) generator is described by the following differential equations (2 equations describe generator motion, 1 exciter, and 1 automatic voltage regulator) [1]

$$\begin{aligned}
\dot{\delta}_k &= \omega_k \\
M_k\dot{\omega}_k &= P_{mk} - D_k\omega_k - P_{Gk} \\
T'_{dk}\dot{E}'_{qk} &= \frac{x_{dk} - x'_{dk}}{x'_{dk}} V_k\cos(\delta_k - \theta_k) + E_{fdk} - E'_{qk}\frac{x_{dk}}{x'_{dk}} \\
T_A\dot{E}_{fdk} &= -E_{fdk} + K_A V_k
\end{aligned} \quad (26)$$

where

$$P_{Gk} = \frac{1}{x'_{dk}} E'_{qk} V_k \sin(\delta_k - \theta_k) - \frac{x'_{dk} - x_{qk}}{2x'_{dk}x_{qk}} V_k^2 \sin(2(\delta_k - \theta_k))$$

x_{dk}, x_{qk} — d axis and the q axis synchronous reactances;
x'_{dk}, x'_{qk} — d axis and the q axis transient reactances;
E'_{qk} — q axis internal voltage behind transient reactance;
T'_{dk} — d axis transient open-circuit time constant;
E_{fdk} — exciter voltage;
P_{mk} — mechanical power input to the generator;
D_k — damping constant;
V_k — voltage magnitude at the generator terminals;
θ_k — voltage angle at the generator terminals.

For kth generator ($k = 1,2,3,4$) the following algebraic equations (power balance equations or stator equations derived from basic Kirchhoff's laws) can be written:

$$P_k = \sum_{j=1}^{11} B_{kj} V_k V_j \sin(\theta_k - \theta_j) + \frac{E'_{qk} V_k \sin(\theta_k - \delta_k)}{x'_{dk}} + \frac{x'_{dk} - x_{qk}}{2x'_{dk} x_{qk}} V_k^2 \sin(2(\theta_k - \delta_k) \quad (27)$$

$$Q_k = -\sum_{j=1}^{11} B_{kj} V_k V_j \cos(\theta_k - \theta_j) + \frac{V_k^2 - E'_{qk} V_k \cos(\theta_k - \delta_k)}{x'_{dk}} + \frac{x'_{dk} - x_{qk}}{2x'_{dk} x_{qk}} V_k^2 [\cos(2(\theta_k - \delta_k) - 1] \quad (28)$$

and for other system buses ($i = 4,5,\ldots,11$)

$$P_i = \sum_{j=1}^{11} B_{ij} V_i V_j \sin(\theta_i - \theta_j) \quad (29)$$

$$Q_i = -\sum_{j=1}^{11} B_{ij} V_i V_j \cos(\theta_i - \theta_j) \quad (30)$$

where P_i is the active power and Q_i is the reactive power injected into the system from bus i, while B_{ij} depicts ij element of the system admittance matrix (due to the high ratio of reactance to resistance, the transmission line resistances are neglected in forming the admittance matrix).

Four variables are describing dynamics of each generator and for four generators it gives 16 state variables. Power balance or network equations are described by two variables per system bus (voltage magnitude and angle) and for seven buses it gives 14 algebraic variables.

ACKNOWLEDGMENT

The author would like to thank Prof. L. Wehenkel and Dr. D. Ernst for their support. He would also like to thank the four anonymous reviewers for their useful comments that resulted in manuscript improvement.

REFERENCES

[1] P. Kundur, *Power System Stability and Control*. New York: McGraw Hill, 1994.

[2] W. H. Croft and R. H. Hertley, "Improving transient stability by use of dynamic braking," *IEEE Trans. Power App. Syst.*, vol. PAS-59, no. 1, pp. 17–26, Jan.-Feb. 1962.

[3] M. L. Shelton, R. F. Winkelman, W. A. Mittelstadt, and W. L. Bellerby, "Bonneville power administration 1400 MW braking resistor," *IEEE Trans. Power App. Syst.*, vol. PAS-94, no. 2, pp. 602–611, Mar.-Apr. 1975.

[4] S. S. Joshi and D. G. Tamaskar, "Augmentation of transient stability limit of a power system by automatic multiple application of dynamic braking," *IEEE Trans. Power App. Syst.*, vol. PAS-104, no. 11, pp. 3004–3012, Nov. 1985.

[5] "State of the art in non classical means to improve power system stability," *Electra*, no. 118, pp. 87–113, 1988.

[6] H. Jiang, D. T. Habelter, and K. V. Eckroth, "A cost effective generator brake for improved generator transient response," *IEEE Trans. Power Syst.*, vol. 9, no. 4, pp. 1840–1846, Nov. 1994.

[7] D. L. Lubkeman and G. T. Heydt, "The application of dynamic programming in a discrete supplementary control for transient stability enhancement of multimachine power system," *IEEE Trans. Power App. Syst.*, vol. PAS-104, no. 9, pp. 2342–2348, Sep. 1985.

[8] Y. Wang, R. R. Mohler, R. Spee, and W. Mittelstadt, "Variable structure FACTS controllers for power system transient stability," *IEEE Trans. Power Syst.*, vol. 7, no. 1, pp. 307–313, Feb. 1992.

[9] Y. Wang, W. Mittelstadt, and D. J. Maratukulam, "Variable structure braking resistor control in a multimachine power system," *IEEE Trans. Power Syst.*, vol. 9, no. 3, pp. 1557–1562, Aug. 1994.

[10] A. H. M. A. Rahim and D. A. H. Alamgir, "A closed-loop quasi-optimal dynamic braking resistor and shunt reactor control strategy for transient stability," *IEEE Trans. Power Syst.*, vol. 3, no. 3, pp. 879–886, Aug. 1988.

[11] A. H. M. A. Rahim, A. M. Al-Shehri, and A. I. J. Al-Sammak, "Optimum control strategies for transient as well as oscillatory instability of power systems," *IEEE Trans. Power Syst.*, vol. 8, no. 2, pp. 491–496, May 1993.

[12] B. Das, A. Ghosh, and P. Sachchidanand, "A novel control strategy for a braking resistor," *Int. J. Elect. Power Energy Syst.*, vol. 20, no. 6, pp. 391–403, 1998.

[13] T. Hiyama, M. Mishiro, and H. Kihara, "Fuzzy logic switching of thyristor controlled braking resistor considering coordination with SVC," *IEEE Trans. Power Del.*, vol. 10, no. 4, pp. 2020–2026, Oct. 1995.

[14] A. H. M. A. Rahim and S. A. Al-Baiyat, "Dynamic brake switching strategies for stabilization of power systems using artificial neural networks," *Expert Syst. Appl.*, vol. 18, pp. 101–109, 2000.

[15] J. Machowski, A. Smolarczyk, and J. W. Bialek, "Damping of power swings by control of braking resistors," *Int. J. Elect. Power Energy Syst.*, vol. 23, pp. 539–549, 2001.

[16] Russian Far East Interconnected Power System Emergency Stability Control, A. A. Grobovoy.*http://www.transmission.bpa.gov/orgs/opi/Power_Stability/* [Online]

[17] R. Bellman, *Dynamic Programming*. Princeton, NJ: Princeton Univ. Press, 1957.

[18] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1998.

[19] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Belmont, MA: Athena Scientific, 1996.

[20] A. W. Moore and C. G. Atkeson, "Prioritized sweeping: reinforcement learning with less data and less real time," *Mach. Learn.*, vol. 13, pp. 103–130, 1993.

[21] M. Glavic, D. Ernst, and L. Wehenkel, "A reinforcement learning based discrete supplementary control for power system transient stability enhancement," in *Proc. Intelligent Systems Application in Power Systems (ISAP'03)*, Lemnos, Greece, 2003, Paper ISAP03/012.

[22] C. C. Liu, W. Jung, G. T. Heydt, V. Vittal, and A. G. Phadge, "The Strategic Power Infrastructure Defense (SPID) system," *IEEE Control Syst. Mag.*, vol. 20, pp. 40–52, 2000.

[23] J. Jung, C. C. Liu, S. L. Tanimoto, and V. Vittal, "Adaptation in load shedding under vulnerable operating conditions," *IEEE Trans. Power Syst.*, vol. 17, no. 4, pp. 1199–1205, Nov. 2002.

[24] D. Ernst, M. Glavic, and L. Wehenkel, "Power system stability control: reinforcement learning framework," *IEEE Trans. Power Syst.*, vol. 19, no. 1, pp. 427–436, Feb. 2004.

[25] M. Pavella, D. Ernst, and D. Ruiz-Vega, *Transient Stability of Power System—A Unified Approach to Assessment and Control*. Norwell, MA: Kluwer, 2000.

[26] K. V. Katsikopoulos and S. E. Engelbrecht, "Markov decision process with delays and asynchronous cost collection," *IEEE Trans. Autom. Control*, vol. 48, no. 4, pp. 568–574, Apr. 2003.