We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

24

Optimal Cardiac Pacing with Q Learning

Rami Rom[1] and Renzo DalMolin[2]
SORIN CRM
[1]*Israel*
[2]*France*

1. Introduction

With Reinforcement Learning (RL), an agent learns optimal behavior through trial-and-error interactions with a dynamic environment. On each step of interaction, the RL agent receives as input some indication of the current state of the environment. The agent then chooses an action to generate as output. The action changes the state of the environment, and the value of this state transition is communicated to the agent through a scalar reinforcement signal. The agent behavior should choose actions that tend to increase the long run sum of values of the reinforcement signal[1].

Cardiac Resynchronizayion Therapy (CRT) is an established therapy for patients with congestive heart failure (CHF) and intraventricular electrical or mechanical conduction delays. It is based on synchronized pacing of the two ventricles [5-7] according to the sensed natural atrium signal that determines the heart rhythm. The resynchronization task demands exact timing of the heart chambers so that the overall stroke volume for example is maximized for any given heart rate (HR). Optimal timing of activation of the two ventricles is one of the key factors in determining cardiac output. The two major timing parameters which are programmable in a CRT device and determine the pacing intervals are the atrioventricular (AV) delay and interventricular (VV) interval.

The adaptive Cardiac Resynchronization Therapy (CRT) cardiac pacemaker control system [2-4], solves a reinforcement learning problem. Accordingly, an implanted cardiac pacemaker is an agent connected to its environment, the patient heart and body, through an implanted electric leads and a hemodynamic sensor. The agent chooses the actions to be delivered, which are the stimulation AV delay and VV interval parameters that are used to resynchronize the right and left ventricles contractions in each heart beat. The agent task is to learn the optimal AV delay and VV interval that maximize the long run cardiac performance in all heart rates.

In order to simulate the resynchronization RL problem a responsive electro-mechanical heart model is needed for generating the expected environment responses to the agent CRT pacemaker stimulations with different AV delays and VV intervals. The responsive electro-mechanical heart model needs to simulate both the heart electrical activity and the correlated heart and body mechanical activities responsive to electrical stimulation delivered in the right and left ventricles with different AV delay and VV interval.

P. Glassel et al [8], provided a system for simulating the electrical activity of the heart that included a computer controlled heart model for generating and displaying the simulated electrogram signals. The simulation system included various hardware components and

software designed to realize the electrical activity of a responsive heart model. However, P. Glassel et al heart model did not simulate the mechanical activity of the heart such as the left ventricle stroke volume, the volumes and pressures of the heart chambers during the systole and diastole cardiac cycles and hence cannot be used for developing a CRT device agent.

A development system of adaptive CRT devices control systems that includes a simplified hemodynamic sensor model was presented by Rom [9]. The aim of the simplified hemodynamic sensor model was to allow a machine learning algorithm to be developed and tested in a simulation with no need to develop a full responsive electro-mechanical heart model. Developing a responsive electro-mechanical heart model is an immense task that needs a model of both the full cardiovascular system and the autonomous nerve system of a CRT patient.

The hemodynamic effects of changes in AV delays and VV intervals delivered to CRT patients were studied by Whinnett et al [10]. In this study, the authors applied non-invasive systolic blood pressure (SBP) monitoring, by continuous finger photoplethysmography (Finometer), to detect hemodynamic responses during adjustment of the AV delay of CRT, at different heart rates. The authors presented CRT response surfaces of systolic blood pressure measurement dependence on paced AV delay and VV intervals. The CRT response surface changed from patient to patient and depended also on the heart rate. The authors suggested that optimization of CRT devices is more important at higher heart rates where CRT patients are more symptomatic. The authors concluded that continuous non-invasive arterial pressure monitoring demonstrated that even small changes in AV delay from its hemodynamic peak value have a significant effect on blood pressure. This peak varied between individuals, was highly reproducible, and was more pronounced at higher heart rates than resting rates.

P. Bordachar et al [11] in a prospective echocardiographic study investigated the respective impacts of left ventricular (LV) pacing, simultaneous and sequential biventricular pacing on ventricular dyssynchrony during exercise in 23 patients with compensated heart failure and ventricular conduction delays. The authors concluded that the optimal interventricular delay was different in rest from exercise in 57% of the patients. In addition the authors showed that changes from rest to exercise in LV dyssynchrony were correlated with changes in stroke volume and changes in mitral regurgitation.

Odonnell et al [12] showed, in 43 CHF patients after CRT implantation in a follow-up study, that the optimal AV delay and VV interval found with echocardiography changed significantly over 9 months of follow-up period.

G. Rocchi et al [13] showed recently that exercise stress Echo is superior to rest echo in predicting LV reverse remodelling and functional improvement after CRT. The authors reported that exercise stress Echo enables identification of CRT responders with about 90% success rate comparing to the current methods that give only about 70% success rate which is still a major problem with CRT today.

According to the clinical studies recited above, the AV delay and VV interval need to be optimized for each CRT patient, may have different optimal values in exercise comparing to rest condition, and may change during 9 months follow up period.

Several optimization methods of control parameters of pacemaker devices in correlation with hemodynamic performance were published. D. Hettrick et al [14], proposed to use the real time left atrial pressure signal as a feedback control mechanism to adjust one or more device parameters. D. Hettrick et al proposed to identify specific characteristics and attributes of the left atrial pressure signal that correlate to hemodynamic performance and to adjust the AV delay parameter of implanted dual chamber pacemaker accordingly.

R. Turcott [15], provided a technique for rapid optimization of control parameters of pacemakers and implanted cardioverters and defibrillators. Turcott proposed to pace the heart with a sequence of consecutive short evaluation periods of equal duration. Turcott proposed to monitor the transient cardiac performance during each of the evaluation phases and to estimate the optimal parameter settings based on changes in the transient cardiac performance from one parameter settings to another.

Hettrick et al and Turcott proposed to use a gradient ascent scheme to adjust the AV delay and VV interval control parameters based on the left atrial pressure signal features (Hettrick et all), and changes in the transient cardiac performance from one parameter settings to another measured by any hemodynamic sensor (Turcott). Hettrick et al and Turcott did not propose to use advanced optimization algorithms for the adjustments of the AV delay and VV interval. Gradient ascent methods may converge slowly, especially in a biological noisy environment, such as the cardiac system. Furthermore, gradient ascent methods may converge to a sub optimal local maximum. Hence, a simple gradient ascent method may result in sub-optimal therapy delivered to CRT patients. The mentioned gradient ascent methods disadvantages together with the clear clinical need of CRT patients to receive optimal therapy may open the door to a more sophisticated machine learning methods that can guarantee convergence and delivery of tailored to the patient optimal therapy.

An adaptive CRT device control system based on reinforcement learning (RL) and using spiking neurons network architecture was presented in [2-4]. The adaptive CRT device control system architecture used a RL method combined with a Hebbian learning rules for the synaptic weights adjustments. The adaptive CRT device control system aim was to optimize online the AV delay and VV interval parameters according to the information provided by the implanted leads and a hemodynamic sensor.

The adaptive CRT device control system used several operational states with a built in priority to operate in an adaptive state aimed to achieve optimal hemodynamic performance. Other operational states were used to initialize the system and to operate as fallback states. The adaptive CRT device control system architecture and operation is described in section 2 herein below.

A Q Learning (QL) and a probabilistic replacement schemes were integrated with the adaptive CRT control system in [16] and are presented in section 3 herein below. QL guarantees convergence online to optimal policy [17], and implemented in a CRT device controller, QL achieves optimal performance by learning the optimal AV delay and VV interval in all heart rates.

With QL, an iterative equation that converges to the optimal policy is solved and a lookup table is calculated. A probabilistic replacement scheme is utilized that replaces an input from a hemodynamic sensor with an input from the lookup table when selecting the next applied AV delay and VV interval. The probability to replace the hemodynamic sensor input with the calculated lookup table value depends on the lookup table difference sign and magnitude that are used as confidence measure for the convergence of the QL scheme. QL combined with the probabilistic replacement scheme improve system performance over time that reach optimal performance even in the face of noisy hemodynamic sensor signal expected with the cardiac system, see Whinnett et al for example [10].

The major advantages of the adaptive CRT control system presented in this chapter are:

1. QL scheme guarantees convergence to optimal policy which in the adaptive CRT application translates to a guarantee to learn the optimal pacing timings (i.e. guarantee to learn online the optimal AV delays and VV intervals).

2. QL converges to the optimal AV and VV values in rest and in exercise where CRT patients are more symptomatic and the converged optimal values are stored in a lookup table that guides the controller operations.
3. Since the Adaptive CRT control system converges to the optimal AV and VV values online, a stress echo test proposed by P. Bordachar et al [11] and G. Rocchi et al [13] in a follow up procedure may not be needed.
4. AV delay and VV interval optimization methods that use a pre-defined sequence of control parameters in a scan test may fail to converge to the true optimal values. Different pre-defined sequences of varying control parameters may lead to different heart conditions and responses, resulting in different estimated values of AV delay and VV intervals since the cardiac system is regulated by the autonomous nerve system and has a delayed time response till it stabilize in a new heart condition.

In summary, an optimization method of the AV delay and VV interval that gradually converges to the optimal set of values is described in this chapter. The optimization method aim is to allow the cardiac system and the autonomous nerve system to stabilize gradually and reach optimum hemodynamic performance in correlation with a learned set of optimal control parameters delivered by the implanted pacemaker. Furthermore, the optimization method aim is to learn the optimal AV delay and VV interval in different heart conditions, and to identify and deliver the learned optimal values safely and efficiently.

This chapter is organized as follows: In section 2 the adaptive CRT device control system architecture and operation are presented and the integration of QL and a probabilistic replacement scheme with the adaptive CRT device control system is presented in section 3. In section 4 simulation results performed with CRT response surface models are presented and section 5 is a conclusion.

2. Adaptive CRT device control system architecture and operation

The adaptive CRT device control system learns to associate different optimal AV delay and VV interval in each heart condition for a CHF patient treated with an implanted CRT device. The adaptive CRT control system uses a deterministic master module to enforce safety limits and to switch between operational states online with a build-in priority to operate in an adaptive state (implemented with a build-in priority state machine). The adaptive CRT control system uses further a slave learning module that implements QL and a probabilistic replacement schemes. The learning module includes leaky I&F neural networks and sigmoid neural networks in order to learn to identify the heart state and to deliver optimal therapy. The adaptive control system uses both supervised learning and a model free reinforcement learning scheme. Supervised learning is used at initialization and fall back states of the priority state machine while QL is used in the higher priority adaptive atate. In the higher priority adaptive state, hemodynamic sensor signal and a QL lookup table calculted online are used. The control system architecture and operation is described herein below.

2.1 Adaptive CRT device control system architecture

The adaptive CRT device control system includes the following main modules:

5. Spiking neurons network.
6. Pattern recognition sigmoid neurons network.
7. Built-in priority state machine.
8. Configuration and register file.

2.1.1 Spiking neurons network architecture

Neural network architectures are inspired by the human brain [18]. Spiking neural networks [19] are closer to biological neural networks and have advantages over other neural networks architectures (such as sigmoid neurons based network) in real time control applications. The spiking neural network perform parallel computation locally and concurrently, in real time. A leaky integrate-and-fire neuron module and a dynamic synapse module are the building blocks of the spiking neurons architecture.

Leaky integrate-and-fire (I&F) neuron

The leaky I&F neuron module is a simplified model of a biological neuron and is naturally adapted for control tasks where the learning objective is a time interval as in the adaptive CRT device where the learned control parameters are the AV delay and VV interval. The leaky I&F neuron is implemented as a digital state machine and two leaky I&F neurons networks are used, one for learning the right AV delay and the second for learning the left AV delay. The interventricular (VV) interval is the time difference between the right and the left optimal AV delays learned by the two leaky I&F neurons. Each leaky I&F neuron is connected to a series of dynamic synapses, typically about 80 dynamic synapses are connected to each leaky I&F neuron. The dynamic synapses weights are adjusted online, in each synapse locally and concurrently (in a hardware version of the controller), according to a set of learning rules in the non-adaptive state and to a second set of learning rules in the adaptive state.

The leaky I&F neuron digital state machine is set initially to idle state waiting for an atrial sensed event. When an atrial sensed event occurs (sensed by an implanted lead in the right atria) the leaky I&F neuron state machine transits to a wait state where in each time step (typically a 1 milli second time step) the outputs of all dynamic synapses connected to the leaky I&F neuron are added to the value stored in a membrane potential register and compared with a threshold value. When the accumulated membrane potential value crosses the threshold value the state machine transits to a fire state, a spike is emitted through the leaky I&F neuron output, and the membrane potential register is reset to 0. The timing of the emitted spike measured relative to the sensed atrial event in milliseconds is the AV delay and the CRT device stimulates the right ventricles accordingly (the left ventricle is stimulated when the left leaky I&F neuron fires a spike similarly). Next the state machine transits to a refractory state for a predefined time period. The leaky I&F neuron state machine transits back to the initial idle state after the refractory period expired and it waits in the idle state to the next atrial sensed event.

A leakage function that reduces the membrane potential value gradually at a pre-defined rate is implemented as a constant value subtracted from the membrane potential register at a constant rate. The leakage function adds timing sensitivity to the leaky I&F neuron and is used to generate a coherent operation of the dynamic synapse. The I&F neuron membrane potential threshold is set to a value that can be crossed only when 3 to 5 dynamic synapses in a short time period emit a maximal post synaptic response (PSR). The dynamic synapses module is described below.

Dynamic synapse

Each dynamic synapse is implemented as a digital state machine. When an atrial event is sensed, a milli second timer starts to count and is used to trigger the dynamic synapses in a time sequence with a pre-defined time delay of 4 msec typically. After receiving the trigger from the timer, each synapse state machine is propagated using it's own local timer from state to state. The dynamic synapse states are: IDLE, WAIT, PRE-HEBB, HEBB, POST-HEBB,

REFRACTORY. Each dynamic synapse releases a post synaptic response in the HEBB state. The PSR magnitude is equal to the adjustable stored synaptic weight and is a time decaying function after the initial PSR is released. All the dynamic synapses PSR's are accumulated in the leaky I&F neuron membrane potential, and when the leaky I&F neuron emits a spike, the dynamic synapse state at the time of the spike in each synapse (may be WAIT, PRE-HEBB, HEBB, POST-HEBB or REFRACTORY state) is captured and stored. The adjustments of the synaptic weights occur at the next sensed atrial event according to the locally captured synapse state and to the learning scheme (supervised learning in the non adaptive state and reinforcement learning in the adaptive state). Typically the synaptic weight stored in each synapse has values between 0 and 31.

Dynamic synapse sub groups and time interleaving

The dynamic synapses are divided to 5 sub groups according to heart rate ranges, from low heart rate range, to high heart rate range and are interleaved according to their excitation time order and heart rate group. The excitation timer triggers the appropriate dynamic synapses sub group according to the time relative to the sensed atrial event in each heart beat and to the current heart rate. The division of the dynamic synapse to sub groups allows learning and adjusting the optimal AV delay and VV interval in each heart rate range in real time throughout the CRT device operation in a patient body which is typically 5 to 7 years. This architecture allows efficient delivery and adjustment of the learned optimal values with faster convergence to the current optimal values.

Supervised learning in the non adaptive state

In the initial and the fall-back non-adaptive CRT state, the adaptive CRT device stimulates the two ventricles using the AV delay and VV interval values programmed by a clinician. The supervised learning task is to train the leaky I&F neurons to fire (i.e. emit a spike) at the programmed values relative to the sensed atrial event in each heart beat. The learning task is to create significant and coherent post synaptic responses of 3 to 5 synapses at the proper times. The released PSR's are then accumulated in the leaky I&F neuron membrane potential that crosses the threshold and fire at the target time (the programmed AV delay and VV interval). Generally, the learning rule increases the synaptic weights in those dynamic synapses that release a PSR just before the target time and reduces synaptic weights values of other dynamic synapses.

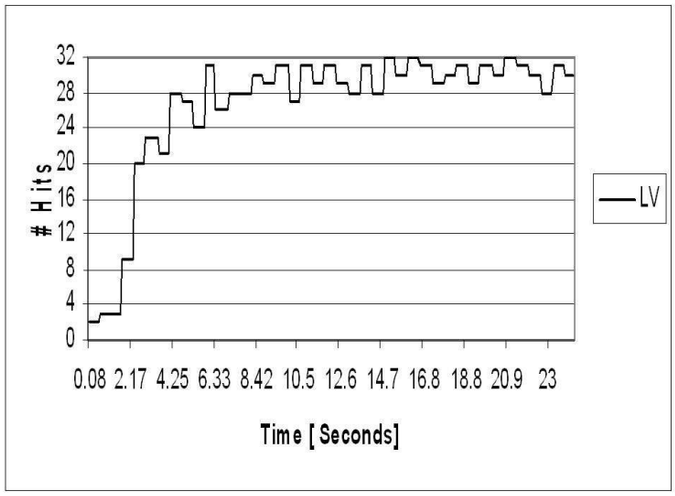


Fig. 1. Hit count rate convergence.

A hit count rate is defined as the number of hits of the leaky I&F neuron spikes at a time window overlapping the target time (the programmed AV delay and VV interval). The I&F neuron learns to fire at the target time window and the number of hits in a time frame of 32 cardiac cycles is used as a performance measure (shown in Fig. 1 above). When the leaky I&F neurons hit count rate achieves a high value (~30) in a time frame, the learning task is converged and a transition to the adaptive state is allowed. When the leaky I&F neurons hit count rate falls below a predefined value (~10) in a time frame, the learning task failed to and a transition to a lower priority state is forced by the build-in priority state machine.

Reinforcement learning in the adaptive state

In the adaptive CRT state a hemodynamic sensor signal responsive to pacing with different AV delay and VV interval is used as the reinforcement immediate reward [1]. Whinnett et al showed in a clinical study [10] that a CRT response surface with a global maximum as a function of the stimulation intervals AV delay and VV interval exist. The adaptive CRT device control system reinforcement learning scheme [2-4], assumes that a CRT response surface exists, and accordingly the synaptic weights reach a steady state values that causes the leaky I&F neurons to fire at the correct timings correlated with the CRT response surface maximum.

The synaptic weights adjustments in the RL scheme are performed in two adjacent cardiac cycles as described in details below. In the first cardiac cycle, a pacing register is increased or decreased by a pre programmed step, Δ. In the next cardiac cycle, the adaptive CRT controller stimulate the heart with the new value and the hemodynamic response is received. Using the current and the previous hemodynamic response and the stored HEBB states of each dynamic synapse, the synaptic weights adjustments are made in each synapse locally and concurrently.

A random stepping mechanism is utilized as follows. In the first cardiac cycle a pacing register value is increased or decreased according to the I&F neuron spike timing, initialized by the sensed atrial event, and compared with the current pacing register value:

$$T_{Spike} > P \qquad P = P + \Delta \tag{1a}$$

$$T_{Spike} < P \qquad P = P - \Delta \tag{1b}$$

4 possible states are defined according to the flow diagram shown in Fig. 2 below

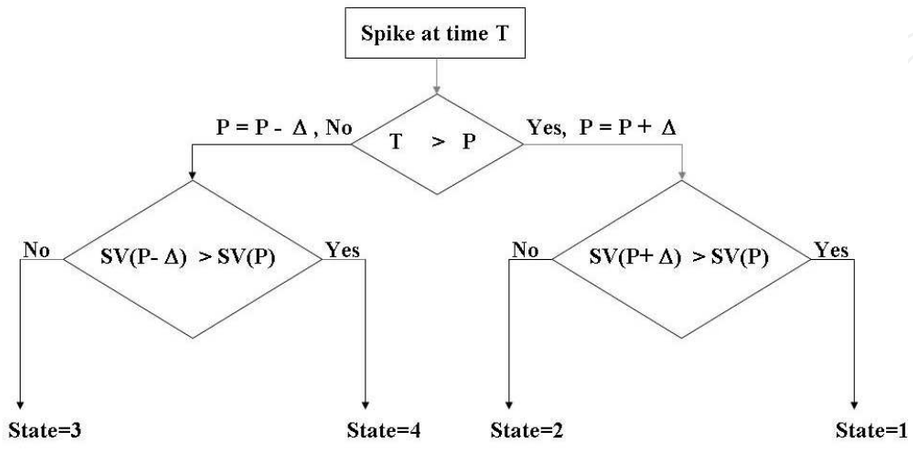


Fig. 2. Synaptic adjustments flow diagram.

Where SV(P) and SV(P+/- Δ) are the hemodynamic response immediate reward stored in the current and the previous cardiac cycles (SV is the stroke volume extracted from the hemodynamic sensor signal and is used as a CRT response surface). A hemodynamic sensor model [9] is used in the simulations presented in section 4 to extract the SV values with different AV delay, VV interval and heart rate.
Next, according to the 4 possible states shown in Fig. 2 and the stored HEBB state in each synapse (PRE_HEBB, HEBB and POST HEBB) , the synaptic adjustments are :

$$W_i=W_i+\lambda \text{ when } \{PRE\_HEBB, 3 \text{ or } 1\} \text{ or } \{HEBB, 4 \text{ or } 2\} \text{ or } \{POSTHEBB, 4 \text{ or } 2\} \quad (2a)$$

$$W_i=W_i-\lambda \text{ when } \{HEBB, 3 \text{ or } 1\} \text{ or } \{POST\ HEBB, 3 \text{ or } 1\} \text{ or } \{PRE\ HEBB, 4 \text{ or } 2\} \quad (2b)$$

The synaptic weights are typically limited to the values of 0 to 31, with a a step value λ, typically 0.125.

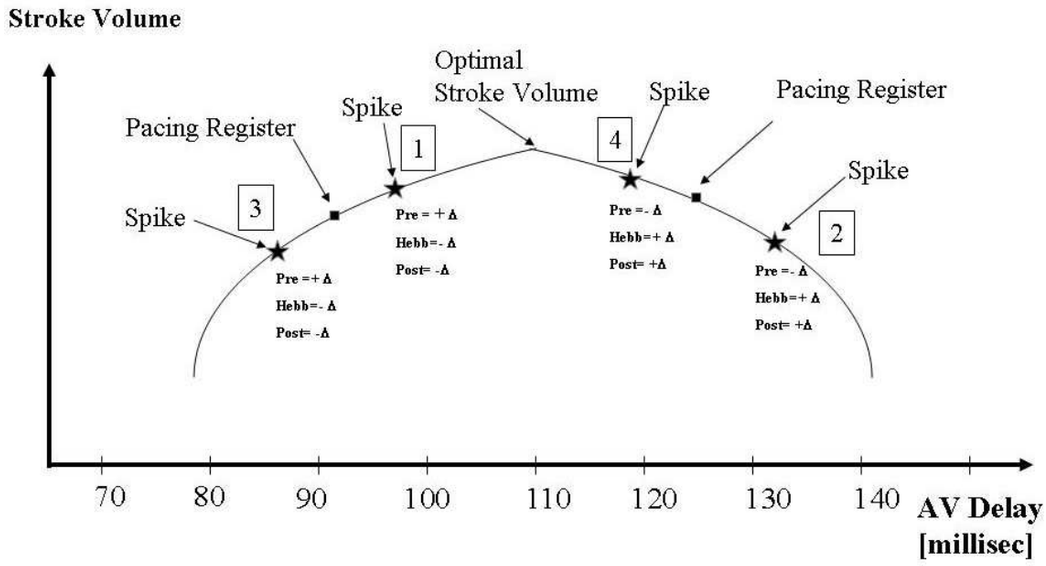


Fig. 3. Synaptic adjustments and the CRT response surface.

In summary, the synaptic adjustments learning rule uses the Hebbian states stored in each dynamic synapse and the hemodynamic responses in two adjacent cardiac cycles to train the leaky I&F neurons to fire at the optimal timing that correlates with the maximal value of the CRT response surface (as a result of coherent release of several dynamic synapse PSR's at the approperiate learned time). The adaptive CRT device control system learns to track the CRT response surface maximum online. When the heart rate changes, the CRT response surface shape changes too [10], and new optimal AV delay and VV interval values are learned and other steady state values of synaptic weights are obtained and stored at the dynamic synapses. Since these changes of the CRT response surface shape and the correlated optimal AV delay and VV interval are learned and stored at the dynamic synapses continuousuly in both the non adaptive and adaptive CRT states, the method maximizes the long term sum of the immediate rewards (i.e. the hemodynamic responses). In section 3 a QL scheme is presented that use Watkins and Dayan iterative equation and adds a probabilistic replacement scheme to QL [16].

2.1.2 Pattern recognition sigmoid network architecture

The pattern recognition sigmoid network includes two sigmoid neuron networks where each network has 16 sigmoid neurons in a layer, 16 synapses are connected to each sigmoid neuron, 3 hidden layers and one linear output neuron. The two sigmoid networks are trained by a standard supervised learning delta rule [18]. The inputs to the pattern recognition networks are temporal patterns of the last 16 hemodynamic sensor responses stored at the controller memory in each heart beat. The supervised training values that the sigmoid network recieves every heart beat is the firing time of the leaky I&F neurons relative to the sensed atrial signal, i.e. the right AV delay for one network and the left AV delay for the second network. The pattern recognition networks learns to associate the learned optimal AV delay and VV interval of the leaky I&F neurons network with a temporal patterns of hemodynamic responses, i.e. hemodynamic performance signals exteracted from the hemodynamic sensor. Hence, the pattern recognition network learns to associate optimal AV delay and VV interval with a heart condition characterized by the temporal patterns of hemodynamic sensor. The operation of the build-in priority state machine described below depends on the successes and failures of the pattern recognition network to output the correct AV delay and VV interval values comparing to the values obtained by the leaky I&F neutons.

2.1.3 Configuration and registers file

The configuration and register file unit stores programmable parameters, such as the initial AV delay and VV interval, and other parameters needed for the initialization of the adaptive CRT control system. The programmable values of the AV delay and VV interval are used in the initialization and fall back non adaptive CRT state while in the non adaptive state the adaptive CRT device controller deliver stimulations with the learned optimal AV delay and VV intervals that correlates with the maximal hemodynamic responses values of the CRT response surface.

2.1.4 Build-in priority state machine

Fig. 4 shows the adaptive CRT priority state machine that has a build in logic that continuously directs the state machine to prefer and to transit to the highest priority adaptive state [21]. Switching to higher priority states require meeting convergence criteria and failing to meet convergence criteria results in transitions back to lower priority states. The lower priority initial and fallback state, the non adaptive CRT state, is the starting state. In the non adaptive CRT lower priority state, the leaky I&F neurons networks adjust their synaptic weights until convergence conditions are met (hit count rate is high) and the build-in priority state machine can switch to a higher priority state, delivering optimal therapy with best hemodynamic performance. The build-in priority state machine in the higher priority adaptive state is guaranteed to deliver the optimal AV and VV Intervals using QL and a probabilistic replacement scheme. In the non adaptive CRT lower priority state the AV delay and VV interval programmed by a clinician are delivered as initialization and safety fallback values. The adaptive CRT build in priority state machine operation and switching conditions are described below.

Non-adaptive CRT state

In the non adaptive CRT state, the CRT device uses a programmed AV delay and VV interval delivering biventricular pacing with fixed AV delay and VV interval. In the non-adaptive CRT state, a leaky integrate and fire (I&F) neurons synaptic weights are trained

using a supervised learning scheme and the synaptic weights reach a steady state values that bring the leaky I&F neurons to fire at the programmed AV delay and VV interval timings with high hit count rate as shown in Fig. 1 above, and after convergence is achieved switching to adaptive state is allowed.

Gradient ascent (GA) state

In the GA CRT state the AV delay and VV interval values are changed according to a random stepping mechanism (see equation 1a and 1b above), and the leaky I&F neurons synaptic weights are trained using a Hebbian and reinforcement learning scheme shown in Figs. 2 and 3 above. The leaky I&F neurons synaptic weights reach a steady state values that bring the leaky I&F neurons to fire at the optimal AV delay and VV interval correlated with the maximum of a CRT response surface extracted from a hemodynamic sensor that reflect for example the stroke volume dependence on the changing AV and VV delays. The GA scheme is designed to track continuously the maximum stroke volume on the CRT response surface as a function of pacing intervals in all heart condition. The leaky I&F neurons output the learned optimal pacing intervals with changing heart rates efficiently using a division of the dynamic synapses to sub groups according to the heart rate range.

QL state

In the QL state, the QL lookup table calculated according to Watkins and Dayan iterative equation [17], are used in addition to the hemodynamic sensor input according to a probabilistic replacements mechanism described in section 3. Q Learning combined with the probabilistic replacement mechanism enables the system to perform optimally also in a noisy biological environment and to improve the overall system performance online using its own predictions. The QL state brings the best hemodynamic performance, learned from the patient hemodynamic responses. The Adaptive CRT build-in priority state machine directs the control system to this highest priority QL state continuously [21].

Fail QL state

In the FAIL-QL state the pattern recognition sigmoid neurons networks re-adjust their synaptic weights in order to map the input temporal patterns of CRT response with the leaky I&F neurons networks outputs.

Switching criteria

Switching between the four states occurs automatically back and forth during operation according to the heart condition and system performance with a build in preference to operate in the QL state that brings the best hemodynamic performance.

Switching from the non-adaptive CRT state to the GA state occurs according to convergence of the leaky I&F neurons networks supervised learning scheme in the non-adaptive CRT state. The two leaky I&F neurons (one for the right AV delay and the second for the left AV delay) need to hit a target times with high rates in a time frame in order to enable a transition as shown in Fig. 1 above.

Switching from the GA state to the optimal QL state occur according to successful association of the temporal pattern recognition sigmoid neural networks predictions (predicted AV delay and VV interval) compared with the I&F neurons network predictions. A hit count rate is calculated for the pattern recognition sigmoid neural networks similar to the hit count rate calculated for the leaky I&F neurons and the hit count rate value of the sigmoid neural networks are used as a performance measure that allows transition to the QL state when it crosses a predefined threshold value.

Fallback from the GA state to the non-adaptive CRT state occurs according to pre-defined system failures that can be for example too low or too high AV delay and VV interval

(crossing pre-defined safety limits), too low or too high heart rate (or other arrhythmia detected) or a poor neural networks performance expressed as a too low hit count rate of the leaky I&F neuron due to sudden drifts of the networks outputs.
Fallback from the QL state to the FAIL QL state occurs if too low hit count rate of the temporal pattern recognition sigmoid neurons networks are obtained. Fallback from the QL state to the FAIL QL state occurs when a sudden change in the heart condition occurs resulting in unfamiliar to the pattern recognition neural networks temporal patterns of hemodynamic sensor values. In such case, the pattern recognition sigmoid neurons networks need to learn to associate the new temporal patterns with new learned optimal values achieved by the leaky I&F neurons network in the new heart condition in order to switch back to a higher priority state.

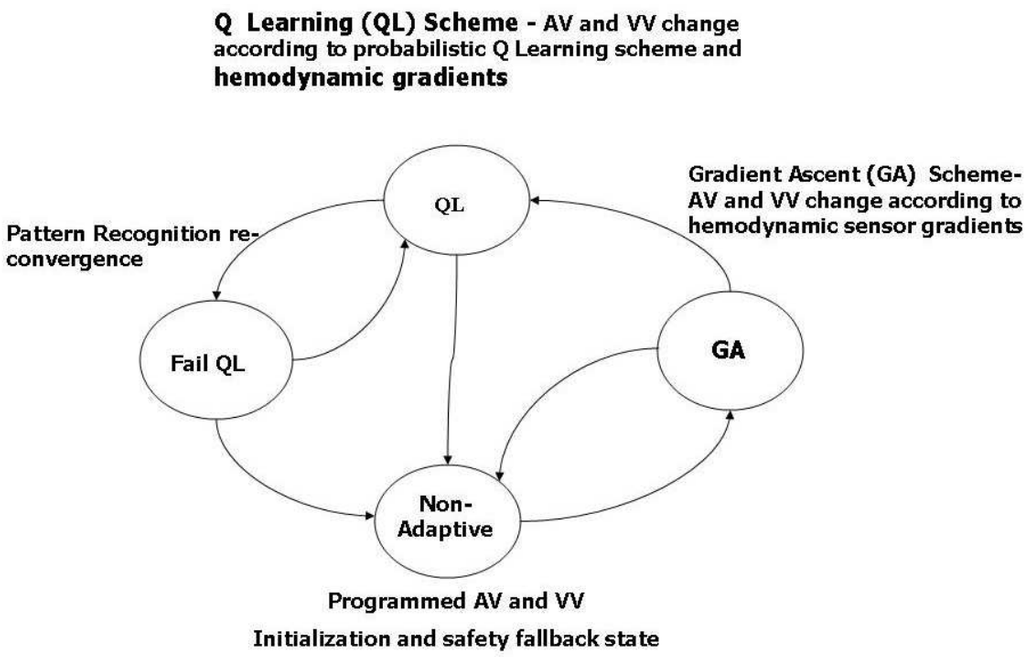


Fig. 4. Build-in priority state machine

2.2 Device optimization during Implantation

Due to the complexity and the cost of the follow up procedures using echocardiography, about 80% of CRT patients are not optimized in the US according to studies presented in Cardiostim conference, France 2006. It is known that more then 30% of CRT patients do not respond to CRT and that CRT non-responders are identified only after 3 to 6 months with quality of life (QOL) questioners or 6 minutes hall walk distance test.
The CLEAR study [22], with 156 pateints enrolled in 51 centers in 8 countries, demonstrated reduced mortality and heart failure related hospitalization in patients whose CRT device was optimized on a regular basis. Final results showed that regular optimization of CRT using Sorin Group's SonR sensor technology improved clinical response rate to 86% as compared to 62% in patients recieving standard medical treatment.
An adaptive CRT device, presented in this chapter, may be used to validate and identify responders to CRT in acute way[20]. The RL algorithm that changes automatically pacing

delays and converge gradually to maximal stroke volume of a CRT response surface will enable a clinician to identify a responder in 2-5 minutes during CRT implantation as simulated in Fig. 5 below. A clinician may monitor the device operation on a programmer screen and validate the hemodynamic improvement according to a CRT responder curve shown in Fig. 5. Optimal CRT and lead positioning during CRT device implantation may turn a non-responder to a responder and a responder to a better responder [6]. The adaptive CRT device implant may allow a clinician using a responder curve to change and validate lead position and achieve optimal lead positioning sites during the implantation procedure. Hence, in addition to the potential long term benefits of a machine learning based adaptive CRT device control system, aimed to manage an implanted CRT device continuously in a patient body for typically 5 to 7 years, the adaptive CRT device control system presented in this chapter may allow:

1. Acute Identification of CRT responders during implantation procedure.
2. Optimal lead positioning validation during implantation procedure.

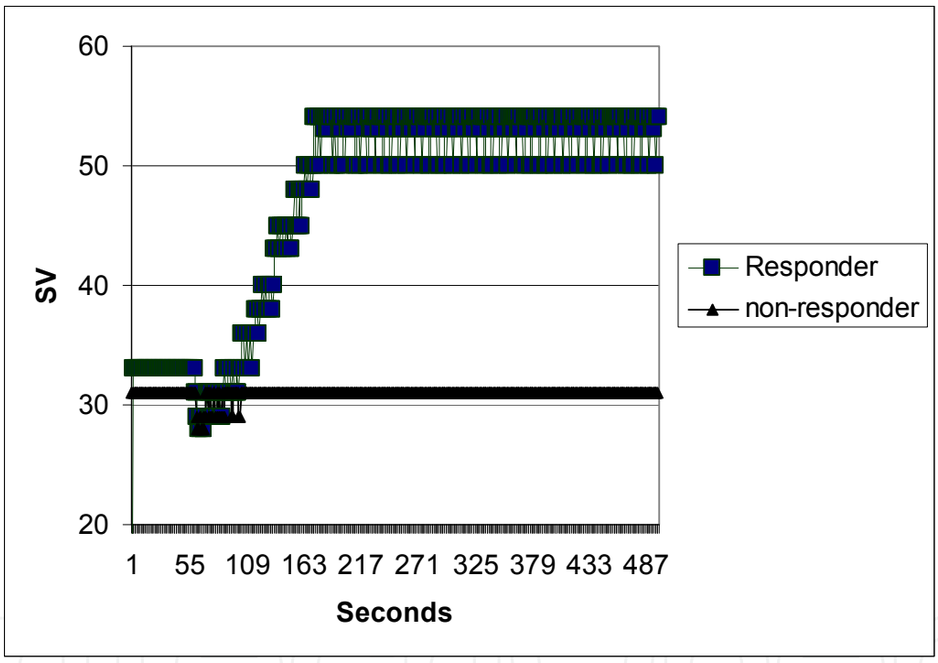


Fig. 5. CRT responder curve

3. Q Learning and cardiac resynchronization therapy

"Reinforcement learning differs from the more widely studied problem of supervised learning in several ways. The most important difference is that there is no presentation of input/output pairs. Instead, after choosing an action the agent is told the immediate reward and the subsequent state, but is not told which action would have been in its best long-term interests. It is necessary for the agent to gather useful experience about the possible system states, actions, transitions and reward actively to act optimally. Another difference from supervised learning is that on-line performance is important; the evaluation of the system is often concurrent with learning"[1].

Watkins and Dayan QL is a model free reinforcement learning scheme where the agent converge to the optimal policy online solving an iterative equation, shown below, and without apriori knowledge of the environment states transitions [17].

$$(S,A)= Q(S,A)+ \alpha(R(S,A)+ \gamma Q_{\max A} (S,A) -Q(S,A)) \quad (1)$$

A is the agent action, S is the environment state, Q(S,A) is the expected discounted reinforcement of taking action A in state S, R(S,A) is an immediate reward response of the environment, α is a small learning rate factor ($\alpha << 1$), γ is a discount factor (smaller then 1), $Q_{\max A}$ (S,A) is the learned optimal policy, i.e. the optimal action A that give maximum Q value at a given state, S, out of the possible set of actions A. The converged solution of Watkins and Dayan iterative equation is stored in a lookup table.

With a CRT device, the two parameters that need to be optimized are the AV delay and the VV interval. Watkins and Dayan QL lookup table is calculated for each configuration of AV and VV values and for each configuration the possible actions assumed are limited to an increase or a decrease by constant value ΔP at a time (typically 5 ms step size is used) applied in the next cardiac cycle.

$$(S,A) = Q(AV, VV, AV +/- \Delta P, VV +/- \Delta P) \quad (2)$$

A represents the pacemaker stimulation timings, AV delay and VV interval, S is the heart hemodynamic performance extracted from a hemodynamic sensor signal, Q(S,A) is the calculated lookup table using a specific AV delay and VV intervals parameters and action A, R(S,A) is the immediate reward (a stroke volume extracted from the hemodynamic sensor signal as an examples). $Q_{\max A}$ (S,A) is the converged Q value expected with the optimal AV delay and VV intervals and optimal action A.

Watkins and Dayan proved [17] that by solving the iterative equation, the agent learns the optimal policy in a model free reinforcement leaning problem with a probability of 1 when the action space is visited enough times such that exploration of the action space is sufficient. The importance of Watkins and Dayan proof, adopted here for CRT pacemakers, is that the stimulation timings obtained by solving the iterative equation are guaranteed to converge to the optimal AV delay and VV without making any assumptions regarding the CRT responses surface shape. The guarantee to converge to the optimal AV delay and VV interval is the valuable benefit of using a sophisticated machine learning method, such as QL, in a CRT pacemaker. Since the AV delay and VV interval parameters are crucial for the success of the therapy [10-13, 21], the guarantee to converge to the optimal AV delay and VV interval is an important advantage over other optimization methods that do not guarantee convergence to optimal values. Furthermore, this advantage should open the door to implementation of machine learning methods, such as QL, in implanted medical devices such as CRT pacemakers and defibrillators.

In the adaptive CRT control system presented here two control parameters, the AV delay and VV interval, were optimized at different heart rates. The QL scheme will be even more beneficial if more control parameters are needed to be optimized. In general, a QL scheme will be more beneficial when the action space is big and the agent needs to select its action from a bigger set of possible actions (i.e. control parameters).

3.1 Probabilistic replacement scheme

QL combined with a probabilistic replacement mechanism allows the adaptive CRT device to replace input gradients with its own predictions learned from the hemodynamic

responses to pacing with different AV delay and VV interval. Q Learning combined with probabilistic replacement mechanism enables the system to perform optimally also in a noisy biological environment, such as the cardiac system, and to improve the overall system performance using its own predictions. The probabilistic replacement scheme selects between input from a hemodynamic sensor and a calculated value obtained from the QL lookup table with a probability that depends on the calculated lookup table. The magnitude of the difference of the optimal action Q value and a sub-optimal action Q value is used as a confidence measure in the probabilistic replacement scheme [16].

Fig. 6 is a flow chart diagram, explaining the leaky I&F spiking neurons synaptic weight adjustmants combined with the probabilistic replacement scheme. The modification shown in Fig. 6 comparing to the flow diagram of Fig. 2 is that the selection conditions depends now on the stroke volumes difference (calculated in the current and previous cardiac cycle) or the QL lookup table difference magnitude which is used as a confidence measure of the probabilistic replacement scheme. After a selection of one out of four possible states is performed, the synaptic weight adjustment are performed in the same manner as described in [2-4 and 16]. The probabilistic replacement mechanism affects the synaptic weight adjustments directly since it determines the selection of one of the four possible states shown in Fig. 6 , and it affects the random stepping mechanism indirectly since the spike timing , T, depends on the values of the adjusted synaptic weights and T value compared to the pacing register value, P, determines the step selection.

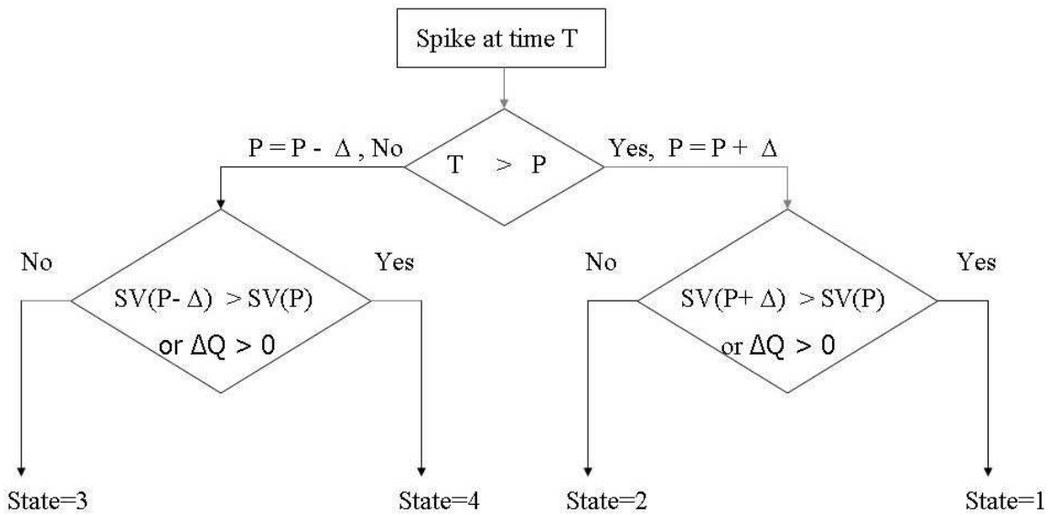


Fig. 6. Synaptic adjustments with the probabilistic replacements scheme

Regulation of α and threshold parameters

Watkins and Dayan iterative equation (Eq. 1) learning rate parameter α, determines the convergence rate of the solution of the iterative Q Learning lookup table. With a high value of α the QL iterative equation will converge faster. However, in noisy environment a too big α parameter may cause instability in the Q Learning scheme. Hence an automatic regulation of α is used to ensure proper performance. The regulation scheme is based on a replacements rate counter calculated online and a maximal steady replacement rate value is typically 80%. This high limit value determines the effectiveness of the probabilistic replacement scheme.

When a QL replacement is performed as shown in Fig. 6, a replacement counter is incremented and the replacement counter is reset to 0 every 2000 cardiac cycles typically.

The number of replacements performed in a time period depends on the α value and on 3 threshold values that are used in the replacement probabilistic mechanism. The α parameter is set initially to a low value (typically 0.02) and is incremented slowly if the replacement rate is below the programmed high limit (typically 80%) until it reach the maximal value allowed for α (typically 0.05). When the replacement rate is higher then the maximal value allowed α is decreased (lower limit for α is typically 0.002).
The 3 thresholds values regulation scheme depends on the value of α and on the calculated replacement rate. The initial thresholds values are set to low values (typically 10, 20 and 30). The values of the 3 thresholds determine three ranges for selecting the lookup table prediction replaceing the hemodynamic sensor input and determining 3 confidence ranges. The magnitude of the difference of the optimal action Q value and a sub-optimal action Q value is compared with the 3 threshold values and accordingly a replacement of the hemodynamic sensor input with the lookup table prediction is selected with a probability that depends on the 3 ranges. When the diffefence magnitude is high, a replacement is performed with a high probability and vice versa.
When α is maximal (0.05) and the replacement rate is still too low the thresholds will be lowered. When α is minimal (typically 0.002) and the replacement rate is still too high the 3 thresholds will be incremented gradually till the replacement rate will be lowered. The aim of both α and the 3 thresholds values regulation is to maintain a steady replacement rate close to the maximal value required (typically 80%). The replacement rate value defines the efficiency of the QL scheme to correct errors of noisy biological inputs using the learned environment responses acquired in the QL lookup table (the probabilistic replacement mechanism use the magnitudes of the diffefnce in addition to it's sign).

4. Simulation results

In the simulation results section we first show that the adaptive CRT control system learns to deliver the optimal pacing timings, i.e. the optimal AV delay and VV interval that maximize the CRT response with varying heart rate (Fig. 7). Next we show that with the combined QL and probabilistic replacement mechanism, the adaptive CRT control system reach the optimal performance in a noisy environment almost independent to the noise level (Fig. 8). We compare simulation results with and without the combined QL and probabilistic replacement mechanism to show that it out perform a simple gradient ascent scheme with varying noise levels (Fig. 9), and finally we show that QL scheme enables the system to escape from local maximum and to converge to the global maximum of a CRT response surface (Fig. 10).
Simulations of the adaptive CRT device control system and the CRT response surface were performed using Matlab-Simulink version 7.6. The adaptive CRT control system application was coded in C and compiled as a Simulink S-function. Simulink S-functions were also used for implementing timers (Atrial-Atrial timer that defines the simulated heart rate and for the AV and VV delays for example) in order to simulate a real time, interrpts based application.
Fig. 7 shows the AV delay and VV interval obtained in a simulation that starts at a heart rate of 60 beats per minute (BPM), then the heart rate changes to 100 BPM and to 120 BPM and relax back to 100 BPM and to 60 BPM periodically. Pre programmed optimal AV delay of the simulated CRT response surface were 160 ms at 60 BPM, 130 ms at 100 BPM and 90 ms at 120 BPM. Optimal pre programmed VV intervals were 0 at 60 BPM, -10 at 100 BPM and -30 at 120 BPM. The simulation results shown in Fig. 7 follow accurately the optimal values.

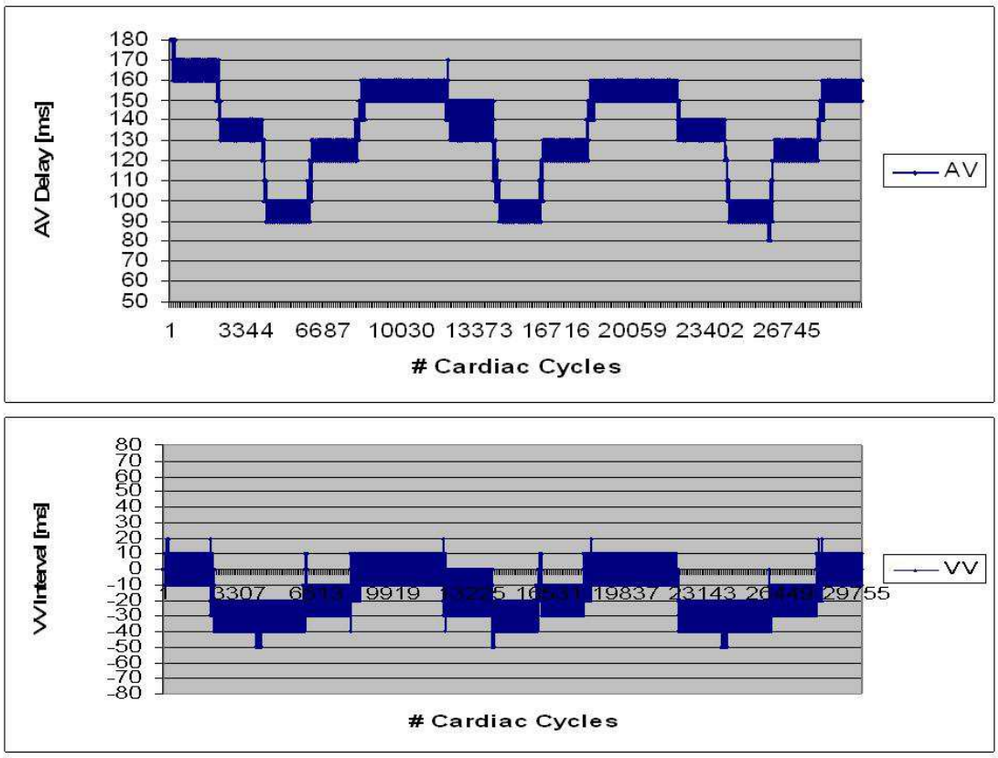


Fig. 7. Dynamic optimization of AV and VV intervals

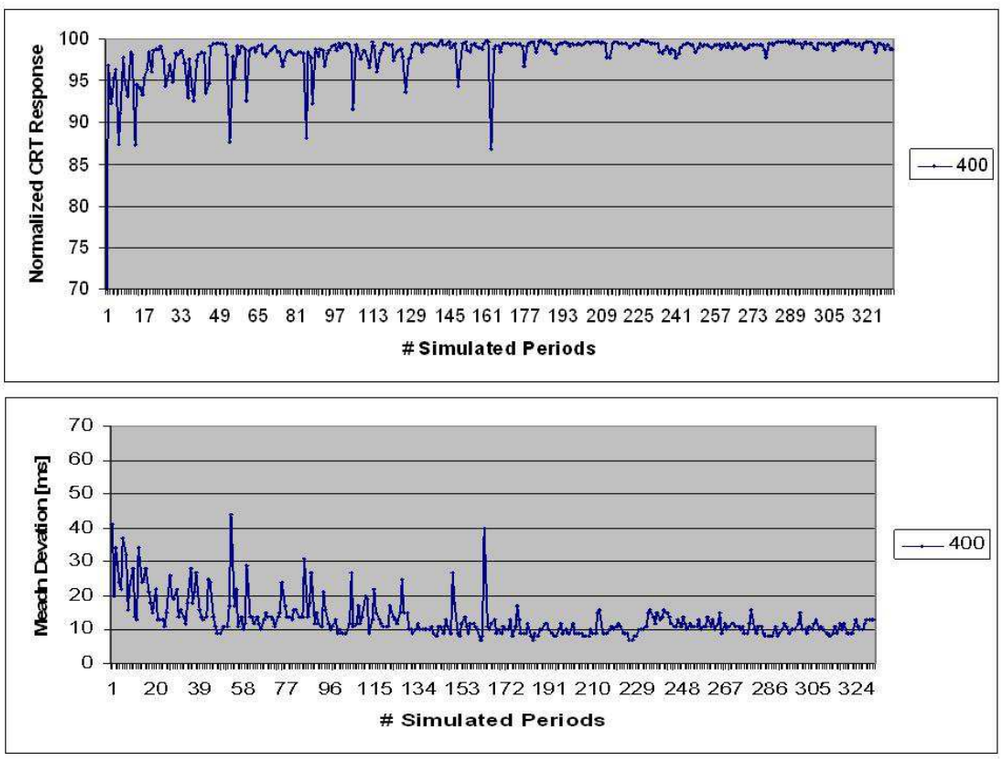


Fig. 8. RT responses convergence with QL in a noisy environment

Fig. 8 shows a normalized CRT response (defined further below) in its upper part and an average deviation from optimal values in its bottom part, calculated during a long simulations with varying heart rate. The normalized CRT response grows during the simulation and shows deeps at the first part of the simulation. After convergence of the combined QL and probabilistic replacement mechanism, the normalized CRT response reach the maximal value with almost no deeps and remain steady at the maximal value while the noise level added to the hemodynamic sensor signal is effetive during all the simulation. The average deviation from optimal AV delay and VV values shown in Fig 8 bottom part is high initially, shows some peaks during a convergence period with generally lower values and then reach a minimal steady value of 10 msec. Fig. 8 proves in a simulation that the adaptive control system learns to deliver pacing with optimal AV and VV intervals in rest and exercise conditions in a noisy environment and the overall system performance improves during the simulation and reach the optimal performance, i.e. the agent learns and acts according to the optimal policy in a noisy environment.

Since the CRT surface responses are proportional to the patient cardiac output, Fig. 8 shows that the combined QL and probabilistic replacement mechanism has the potential to increase the cardiac output of CRT patients which is a major goal of CRT. Hence machine learning methods may be clinically beneficial to CHF patients and this advantage may open the door for machine learning methods implemented in implanted medical devices [16, 23, 24].

Fig. 9 shows the adaptive CRT device system performance with and without learning with varying random noise levels added to the hemodynamic sensor input, i.e. to the CRT response surface. The system performance with a simple gradient ascent scheme (without learning) falls linearly with the growing random noise level to below 70% of the optimal performance while QL combined with the probabilistic replacement scheme is able to improve the system performance and keep it almost at the optimal system performance with no noise at all noise levels shown.

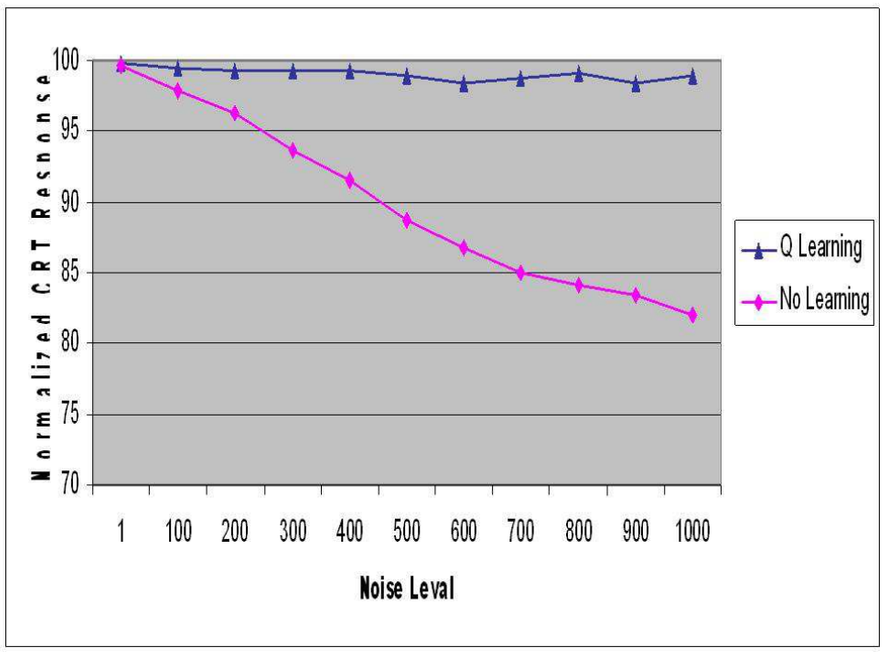


Fig. 9. System performance with QL in noisy environment

The normalized CRT response shown in Figs. 8 and 9 is defined as a normalizaed average of CRT responses calculated during each simulation period of 2000 cardiac cycles.

$$\text{Averaged CRT Response} = 1/2000 * \Sigma_1^{2000} \text{ CRT Response(i)} \quad (3)$$

A normalize CRT Response, that takes into account the surface global maximal value and the minimal value at a given heart rate is calculated according to Eq. 4 below:

$$\text{Normalized CRT Response} = [\text{Averaged CRT Response} - \text{CRT Response Min}]/[\text{CRT Response Max} - \text{CRT Response Min}] * 100 \quad (4)$$

Where at heat rate of 60 BPM :
CRT Response Min = CRT Response (worst values AV=60,VV=0)
CRT Response Max = CRT Response (optimal values AV=160,VV=0) .
An important aspect of the QL based adaptive CRT control system is its ability to converge to the global maximum of the CRT response surface when it includes also a local maximum. Fig. 10 shows the pacing histogram obtained with a long simulations of 1 million cardiac cycles with random noise and compares the results obtained with and without Q Learning. The simulation starts at low AV delay of 90 ms in vicinity of a local maximum in the simulated CRT response surface. The simulated pacing histograms shows that without QL the histogram has a stronger peak at the local maximum of 90 ms and a weaker peak at the global maxima of 160 ms. With Q Learning the histogram is peaked at the global maximum of 160 ms and only a small peak is seen at the local sub optimal maximum of 90 ms.

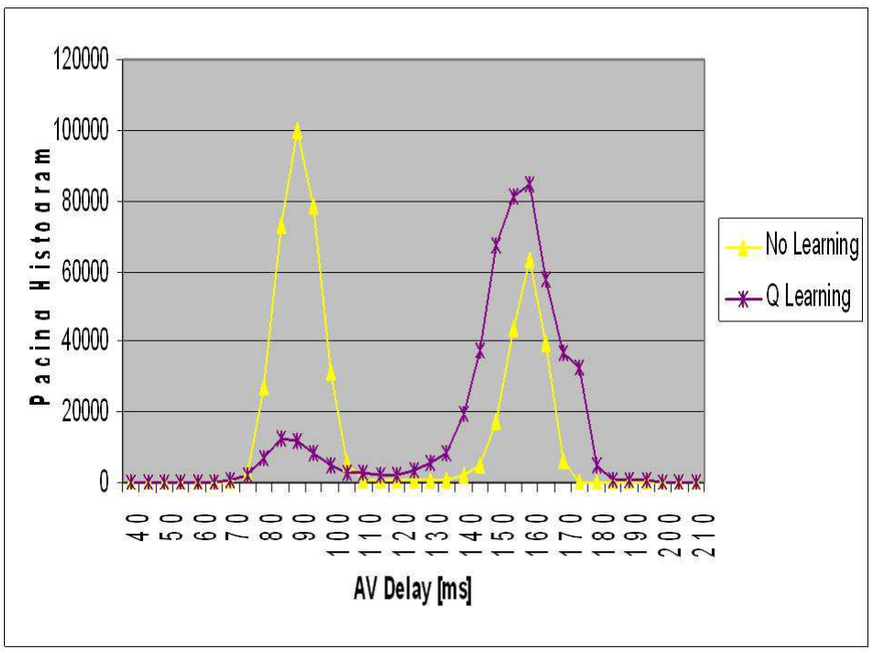


Fig. 10. Convergence to the global maxima with QL

7. Conclusions

In this chapter we present an adaptive control system for a CRT device based on QL and a probabilistic replacement mechanism aimed to achieve patient specific optimal pacing

therapy for CHF patients implanted with a CRT device. The learning target was to learn to deliver optimal AV delay and VV interval in all heart conditions and in a biological noisy environment. The adaptive control system uses a deterministic master module that enforce safety limits and switch between operational states online with a build-in priority to operate in the adaptive state, and a learning slave module that use QL and probabilistic replacement mechanism and includes leaky I&F neural networks and sigmoid neural networks in order to identify heart conditions and to learn to deliver optimal therapy.

A combined QL and probabilistic replacement mechanism may allow the adaptive CRT device to replace hemodynamic sensor inputs with its own calculated predictions learned from the environment responses to pacing. The combined QL and probabilistic replacement mechanism may enable the adaptive CRT control system to perform optimally in a biological noisy environment, such as the cardiac system.

The adaptive CRT device aim is to increase the patient hemodynamic performance (cardiac output for example) and to be clinically beneficial to CRT patients especially in high heart rates where CRT patients are more symptomatic [10]. Pre-clinical and clinical studies are needed to prove the clinical benefits of an adaptive CRT device bases on machine learning methods.

Adaptive control systems that learn to deliver optimal therapy were proposed for two other implanted medical devices: Vagal stimulation device that learns to regulate the patient heart rate combined in one can with a CRT device that improve cardiac efficiency by learning to optimize at the same time also the AV delay and VV interval [23], and a deep brain stimulation (DBS) device adaptive control system that learns the optimal control parameters that reduce uncontrolled movements of Parkinson's disease patients [24].

8. References

[1] Leslie P. Kaebling, Michael L. Littman and Anderw W. Moore, "Reinforcement Learning: A Survey", Journal of Artificial Intelligence Research 4 , 237-285, 1996.K.

[2] R. Rom, "Adaptive Resynchronization Therapy System", US Patent 7,657,313, July 2004.

[3] R. Rom, J. Erel, M. Glikson, K. Rosenblum, R. Ginosar, D. L. Hayes, "Adaptive Cardiac Resynchronization Therapy Device: A Simulation Report", Pacing and Clinical Electrophysiology. Vol. 28, pp. 1168-1173, November 2005.

[4] R. Rom, J. Erel, M. Glikson, K. Rosenblum, O. Binah, R. Ginosar, D. L. Hayes, "Adaptive Cardiac Resynchronization Therapy Device Based On Spiking Neurons Architecture", IEEE-TNN, Vol 18, Number 2, 542-550, March 2007.

[5] A. Ellenbogen,B. L. Wilkoff, and G. N. Kay , "Device Therapy for Congestive Heart Failure", Philadelphia, Pennsylvania, Elsevier Inc., 2004, pp 47-91.

[6] D. L. Hayes, P. J. Wang, K. Sackner-Bernstein, S. J. Avirsatham,"Resynchronization and Defibrillation for Heart Failure, A Practical Approach", *Oxford, UK,* Blackwell Publishing, 2004, pp 39-72.

[7] D. L. Hayes and S. Forman, "Cardiac Pacing. How it started, Where we are, where we are going", Pacing and Clinical Electrophysiology, vol. 27, pp. 693-704, May 2004.

[8] P. Glassel et al, in "Interactive Cardiac Rhythm Simulator", US Patent 5,692,907, August 1995.

[9] R. Rom, "Heart Simulator", US Patent application 2008/0103744, Oct 2006.

[10] Z. I. Whinnett, J. E.R. Davis, K. Wilson, C. H. Manisty, A. W Cohw, R. A Foale, D W. Davies, A. D Hughes, J. Mayet and D. P Francis, "Haemodynamic effects of

changes in AV and VV delay in cardiac Resynchronization Therapy show a consistent pattern: analysis of shape, magnitude and relative importance of AV and VV delay", Heart published online, 18 May 2006, doi:10.1136/hrt.2005.080721.

[11] P. Bordachar, S. Lafitte, S. Reuter, K. Serri, S. Garrigue, J. Laborderie, P. Reant, P. Jais, M. Haissaguerre, R. Roudaut, J. Clementy, "Echocardiography Assessment During Exercise of Heart Failure Patients with Cardiac Resynchronization Therapy", American Journal of Cardiology, Vol. 97 , June 2006 , pp. 1622-5.

[12] D. Odonnell, V. Nadurata, A. Hamer, P. Kertes and W. Mohammed, "Long Term Variations in Optimal Programming of Cardiac Resynchronization Therapy Devices", Pacing and Clinical Electrophysiology, vol. 28, January 2005, suppl. 1, pp. S24-S26.

[13] G. Rocchi et al, in "Exercise stress Echo is superior to rest echo in predicting LV reverse remodelling and functional improvement after CRT", European Heart Journal (2009), 30, 89-97.

[14] D. Hettrick et al, in "System and a Method for Controlling Implantable Medical Devices Parameters in Response to Atrial Pressure Attributes", US Patent Application 11,097,408, March 2005.

[15] R. Turcott, in "System and a Method for Rapid Optimization of Control Parameters of an Implantable Cardiac Stimulation Device", US Patent 7,558,627, Sep 2003.

[16] R. Rom, "Optimal Cardiac Pacing with Q Learning", WO 2010/049331, Oct 2008.

[17] Christopher Watkins and Peter Dayan, "Q Learning", Machine Learning 8, 279-292, 1992.

[18] Stamatios Kararalopoulos, "Understanding Neural Networks and Fuzzy Logic", IEEE Press, 1996.

[19] Wolfgang Maas and Christopher Bishop, "Pulsed Neural Networks", MIT Press, 2001.

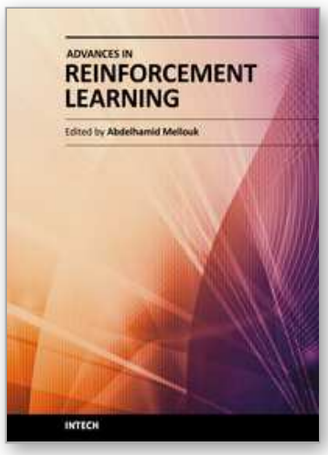
[20] R. Rom, "Optimizing and Monitoring Adaptive Cardiac Resynchronization Therapy Devices", PCT WO 2006/061822, June 2006.

[21] R. Rom, "Intelligent Controll System for Adaptive Cardiac Resynachronization Therapy Device", US Patent application 2010/145402, July 17 2006.

[22] Sorin Group Press Release, "CLEAR Study Demonstrates Automatic Optimization of CRT with SONR Improves Heart Failure Patients Response Rates ",June 17th, 2010 , http://www.sorin-crm.com/uploads/Media/clear_study_pressrelease.pdf.

[23] R. Rom, "Adaptive Resynchronization Therapy and Vagal Stimulation System ", US Patent application 2008/0147130, Aug 2007.

[24] R. Rom, "Optimal Deep Brain Stmilation with Q Learning", WO 2010/049331, Oct 2008.



Advances in Reinforcement Learning
Edited by Prof. Abdelhamid Mellouk

ISBN 978-953-307-369-9
Hard cover, 470 pages
Publisher InTech
Published online 14, January, 2011
Published in print edition January, 2011

Reinforcement Learning (RL) is a very dynamic area in terms of theory and application. This book brings together many different aspects of the current research on several fields associated to RL which has been growing rapidly, producing a wide variety of learning algorithms for different applications. Based on 24 Chapters, it covers a very broad variety of topics in RL and their application in autonomous systems. A set of chapters in this book provide a general overview of RL while other chapters focus mostly on the applications of RL paradigms: Game Theory, Multi-Agent Theory, Robotic, Networking Technologies, Vehicular Navigation, Medicine and Industrial Logistic.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Rami Rom and Renzo DalMolin (2011). Optimal Cardiac Pacing with Q Learning, Advances in Reinforcement Learning, Prof. Abdelhamid Mellouk (Ed.), ISBN: 978-953-307-369-9, InTech, Available from: http://www.intechopen.com/books/advances-in-reinforcement-learning/optimal-cardiac-pacing-with-q-learning

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen