Transmission Power Management of LEACH Wireless Sensor Network Cluster Using Multi-Agents Reinforcement Learning Power Control

Dwi Widodo Heru Kurniawan, Adit Kurniawan and M. Sigit Arifianto

School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
dwiwidodo29@gmail.com, adit@stei.itb.ac.id, msarif2a@gmail.com

Abstract: Energy consumption conservation is a hot topic in a wireless sensor network (WSN). The popular algorithm to solve the problem involves reducing communication range using the cluster method, and the low energy adaptive clustering hierarchy (LEACH) is one of the most famous techniques. The cluster based wireless sensor network experiences a heavy intra cluster interference caused of imperfect power control, which can deteriorate the network lifetime. The existing power control technique applies a distance-based power control to select transmission power, which cannot adapt to environmental changes; that is why the network failed to control the interference and calling for a new method to improve the existing power control. The research evaluates the reinforcement learning (RL) based power control that can bring the sensor to select the lowest transmission power while keeping the signal and interference to noise ratio (SINR) of sensor in cluster head (CH) above the threshold. To improve the RL-based power control, the author proposes a dual policy technique with SINR as a control metric. Finally, the simulation exhibits the proposed algorithm's effectiveness and performance improvement over the original one that is network lifetime, packet delivered, and average energy consumption per round.

Keywords: LEACH; reinforcement learning; power control; dual policy; SINR

1. Introduction

Clustering is a leading method to control interference and energy expenditure in a wireless sensor network. The method reduces the transmission distance between the sensors and the sink by inserting one node. The elected sensor behaves as an intermediate node that received data from the cluster's member and sends it into the sink. The most popular technique is the low energy adaptive clustering hierarchy (LEACH) [1][2]. The sensors run the LEACH algorithm to select CH and join with the closest CH. After cluster formation, the sensors begin to send data into the cluster head. However, the algorithm does not have a specific method to control the transmission power, and they use a distant estimation as a basis to select transmission power. The technique has a weakness in that it cannot adjust the power when the CH's signal quality is low and need to increase the sensor's power level [3][4]. Then the cluster-based sensor network experiences a heavy intra cluster interference coming from neighbor nodes[5]. The interference is caused by non-adaptive power control used by the nodes.

Zheng et al. [6] has proposed adaptive transmission power control (ATPC) which the objective is to select the minimum transmission power based on the power distance between sender and receiver. The idea is to find the shortest path between the sender and receiver and allocate the transmission power $P_{tx} \in \{P_{short\text{-}range}, P_{default\text{-}range}, P_{long\text{-}range}\}$. The sensor will create a multi-hop routing table following the minimum power distance, which can be solved by the Dijkstra algorithm. The ATPC benefit is achieved through two conditions:

a. The short-range communication selection

Figure 1 shows three areas of transmission that is short-range, default-range, and long-range. The method will guide the sensor to choose short-range transmission power when sending the data into Receiver1 rather than default-range. The selection will reduce the transmission power by 62.9% compare with the default setting.

Received: July 18th, 2020. Accepted: September 2nd, 2021
DOI: 10.15676/ijeei.2021.13.3.5

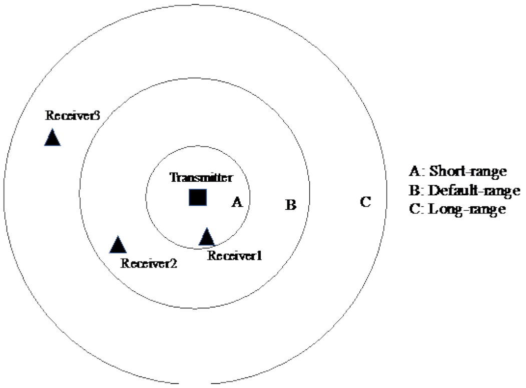

Figure 1. Transmission range for different power level[6]

b. The long-range communication selection

The second case if the sensor sends the data into Receiver3, which is longer than the default range. The sensor can choose another sensor as a relay sensor or increase the power to cover the long-range. Comparing the long-range power distance and the sum of the multi-hop power distance is the primary consideration. Indeed, the long-range selection can reduce the transmission power by 40% after the optimization and decrease one hop. Both selections in wireless sensor networks generate a power reduction of around 30-40% compared to the default setting. The adaptive transmission power control method has similarity with the proposed method which apply variable power control to meet the neighborhood link quality by selecting minimum transmission power. However, the method does not adapt to channel changes and interference from other nodes that can decrease link quality.

Masood et al. [7] have introduced an adaptive on-demand transmission power control (AODTPC) algorithm, the revised version of ODTPC (on demand transmission power control) [8]. The main idea is to apply the Kalman Filter to generate the predictive RSSI (receive signal strength indicator); then, the RSSI will guide the sensor to select appropriate transmission power to overcome the varying channel. The sensor uses the method to track the channel condition, estimate the next RSSI, and respond to it with increased or decreased transmission power based on comparing the next RSSI with the threshold. The method has been evaluated and exhibits superior performance compared with ATPC and ODTPC, reducing energy consumption by 53%.

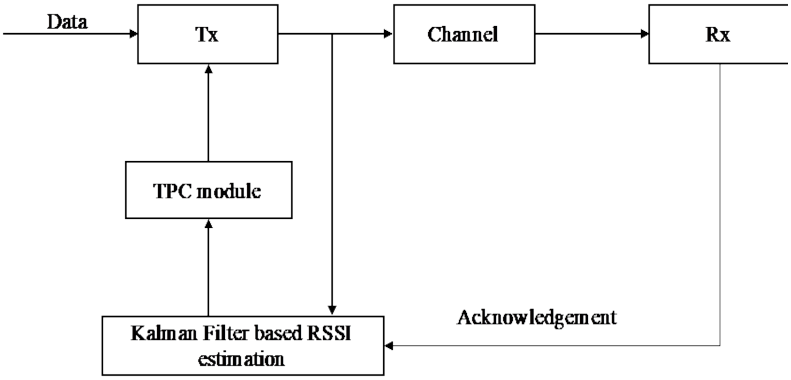

Figure 2. AODTPC system model[7]

Figure 2 shows the AODTPC system model that enhances the ODTPC method with the Kalman filter. The AODTPC is like the proposed method in that it tracks the control metric in the receiver, generates the prediction, and selects the power. However, the proposed method uses different control metrics, SINR rather than RSSI, which is powerful in evaluating the channel condition due to fading and interference. Another difference is the prediction method. The proposed method uses reinforcement learning to learn the environment changes and estimate the future based on the previous state. The system creates a Q table that contains previous experience.

In adaptive and robust topology control (ART), Chincoli et al. [9] have applied control metric PRR (packet received ratio) as guidance for selecting transmission power. The sensor calculates PRR as a ratio of the ACK message from receiver over L number of transmitted packets within a window W. Then, PRR will be compared to two thresholds Th_l and Th_h; if PRR is higher than Th_h, the sensor decreases the power by one level; otherwise, if PRR is below Th_l, the sensor increases the power level. In any other case, P_{tx} stays constant. After window W ended, then the calculation starts again. The author has verified the method using simulation and compares it with the MaxPow method. ART has shown better performance compared to MaxPow in latency and throughput indicators. The method has similarities with the proposed method, in which both methods use the variable transmission power with different control metrics. However, PRR selection has generated instability where PRR near the threshold and the simulation do not consider transmission interference directly.

Another transmission power research has been conducted by Chincoli. In self-learning power control, Chincoli et al. [10] evaluated reinforcement learning (RL) method in power control to overcome the weakness of the previous method. The method has been categorized as a learning technique using the past event to generate a learning table and use it to optimize the transmission power selection. The author has simulated the method using a small sensor network and considers interference from adjacent nodes. The method utilizes Clear Channel Assessment (CCA) and number of retransmissions as a control metric to identify if interference happens in the receiver. The states of the nodes are determined based on a combination of CCA and retransmission within windows W. The power selection effect will be valued using a reward system based on PRR and power level; the higher PRR with lower power transmission will be rewarded with a high reward score. This reward will be accounted for Q table calculation, a table that represents a good or bad power selection. After the system converges, the node can utilize the Q table as guidance to select the next transmission power. The RL-based power control is more advanced than other methods that provide a smart capability to adapt to channel fading, interference, and other environmental changes. The method applies RL-based power control, like the proposed method. However, it utilizes CCA and PRR as control metrics to measure the interference effect and use the information to create a Q table. The technique has been tested in a few numbers of sensors and shows outstanding performance.

In power control based on multi-agent deep Q network, Gengtian et al. [11] have proposed RL-based power control in the Device to Device (D2D) communication with Q value approximator using neural network. The system uses SINR as a control metric to maintain the signal quality above the threshold. When the system trains, it applies a deep Q network to minimize the loss function in every episode. The action, state, and reward will be stored in the memory as an input of the next process. The author has simulated the system and shows that the system has a superior performance compared with open-loop power control and MaxPower Control in terms of system throughput.

The previous work on power control can be categorized into four groups, namely proactive (ATPC), reactive (ODTPC and ART), predictive (AODTPC), and learning (RL) methods. The proactive method has estimated the transmission power based on the network model; however, this method cannot adapt to the environmental changes and lead to wrong action that degrades the performance. In the reactive method, the sensor monitors link quality (RSSI) continuously and compares it to the threshold. If the quality falls below the threshold, the sensor will adjust transmission power to overcome the situation. The method does not consider past events as part of the prediction and leads to oscillation as the environment varies every time. The predictive method has generated the next value and compared it with the threshold. If the next value is higher than the threshold, then the sensor will lower the transmission power and vice versa. However, the predictive system has considered the previous value only, and it has no memory of the best action taken in the past. The situation will lead to the wrong action, and the system will enter a competition which can bring the sensor applies high transmission power. The latest research on this area has led to a smart capability by utilizing past events to create a table representing good or bad selection history. After the table converges, the node uses it as guidance to select new power. However, the previous RL power control research applies in a small network or D2D communication, opening the research in a vastly complex network.

Our work studies the RL-based power control in a large-scale LEACH network to reduce interference and energy consumption with the main contribution are two folds:

a. System model and algorithm which control interference with collaboration between the nodes in one cluster to achieve an equilibrium, then every node sends the data.
b. Improving the RL algorithm with the introduction of the dual policy in the testing stage, which replace e-greedy policy

The following section discusses the system model for wireless sensor network power control, section 3 introduces the simulation design, section 4 discusses the results and performance, and the last chapter discusses the conclusion.

2. System Modelling

The research applies LEACH cluster based wireless sensor network as a model. LEACH is an eminent high-performance topology control algorithm that splits WSN into two-layer and transmitting the data in two hops. It is a distributed clustering scheme proposed for uniform distribution of energy consumption among all the nodes in WSN. The algorithm groups the sensor nodes in WSN into CH and the member sensors. The member sensors sense, collect data, and send it into CH. The CH performs processing operations (such as de-redundancy and data fusion) and delivers the data packets to the base station (BS) [12][1][13].

The LEACH method executes in rounds, and each round runs two phases: setup phase and stable state phase. In the setup phase, WSN specify pch% of n sensor nodes as CH. The sensor i generates a random number between 0 and 1 and compares it with the threshold value T(i) as follows[14]:

$$T(i) = \begin{cases} \frac{pch}{1-pch*(r \bmod \frac{1}{pch})}, if\ i\ \in\ G \\ 0, otherwise \end{cases} \qquad (1)$$

where pch is the desired number of CH, r is the current round, and G is the set of those nodes which are non-CHs in the last *1/pch* rounds. If the random value is below the threshold, then the sensor becomes CH. After the CH election, the non-CH sensors will join to the nearest power distance CH. In the stable state phase, the members transmit the sensed data using Time Division Multiple Access (TDMA) scheme to the CH. The CH performs the information fusion over all the nodes in the cluster and then transmits it to the BS.

We assume N sensor nodes are randomly distributed in $A \times A$ area to form the WSN. The analysis made several assumptions. Firstly, channel communication is modeled by free-space loss, and the two-ray ground propagation model depends on the node and CH distance [15]. The base station (or sink) is in the center of the area and has enough power to cover it. The nodes are stationary, homogeneous, and have a variable transmission power adjusted to reach the CH. The research uses the energy model shown in Figure 3 to analyze energy consumption. We adopt the first-order radio model to describe the energy consumption [16]. If the distance is less than the threshold d_0 [16], which is calculated using Equation (2), then the free space model is used; otherwise, the two-ray ground propagation is used [15][17].

$$d_0 = \sqrt{\epsilon_{fs}/\epsilon_{MP}} \qquad (2)$$

where ϵ_{fs} is free space propagation model (*pJ/bit/m*^{2}) and ϵ_{MP} is multipath propagation model (*pJ/bit/m*^{4}).

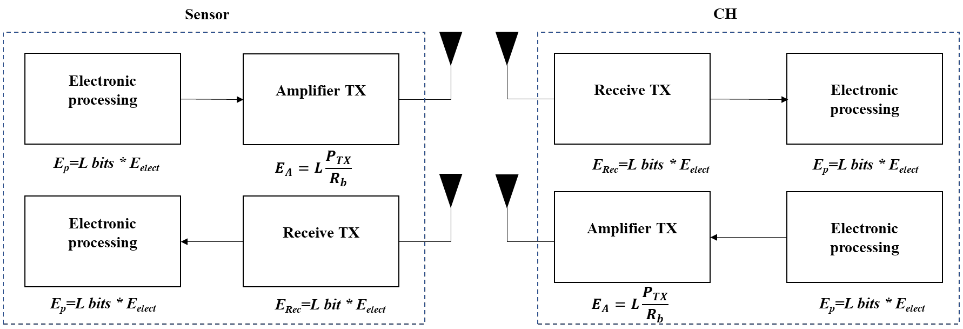
Sensor
CH
Electronic processing
Amplifier TX
Receive TX
Electronic processing
E_p=L bits * E_{elect}
$E_A = L\frac{P_{TX}}{R_b}$
E_{Rec}=L bits * E_{elect}
E_p=L bits * E_{elect}
Electronic processing
Receive TX
Amplifier TX
Electronic processing
E_p=L bits * E_{elect}
E_{Rec}=L bit * E_{elect}
$E_A = L\frac{P_{TX}}{R_b}$
E_p=L bits * E_{elect}

Figure 3. Energy Model[4]

We assume that E_{elect} is the dissipated energy per bit, and d is the distance between the transmitting node and the receiving node. Therefore, the energy consumption of transmitting L-bits (E_{TX}) is computed as [16][18]:

$$E_{TX}(L,d) = \begin{cases} LE_{elect} + L\epsilon_{fs}d^2, if\ d < d_0 \\ LE_{elect} + L\epsilon_{MP}d^4, if\ d \geq d_0 \end{cases} \quad (3)$$

where L is the length of the transmitting data package, d_0 is the threshold of the free space propagation model, generally about 80 m. The energy consumption of receiving L-bits (E_{RX}) is computed as [16][18]

$$E_{RX}(L) = 2LE_{elect} \quad (4)$$

LEACH algorithm applies a distant based transmission power control method. The sensors send advertisement message with maximum transmission power in the setup phase. The sensors estimate the distance between them and its adjacent node utilizing the received signal strength. Next, every sensor will decide itself as CH or sensor member with compare a random number and the threshold as Equation (1). If sensor i become cluster member, then it will join with nearest CH and calculates transmission power ($P_{TX,i}$) as follows [15]:

$$P_{TX,i} = \begin{cases} \frac{P_{rx,i}}{C_f} d_{i-CH}^2,\ d_{i-CH} \leq d_0 \\ \frac{P_{rx,i}}{C_f} d_{i-CH}^4,\ d_{i-CH} > d_0 \end{cases} \quad (5)$$

where C_f is a transceiver characteristic number, $P_{rx,i}$ is received power of sensor i in CH, $d_{i\text{-}CH}$ is the distance sensor i to CH. The transmission power $P_{TX,i}$ will be used to send the information from the sensor i to CH in a particular round. If a new round begins, then the sensor will re-evaluate the transmission power according to the new cluster formation.

The RL-based power control is one of the adaptive power controls that can track a specific control metric and set up an insight. The sensor will use the insight to decide a new transmission power that is suitable for the environment. Figure 4 shows the block diagram of the RL-based power control. There is two-component that interact to achieve a target, namely the sensor and CH. The RL-based power control embedded in the LEACH method considers the SINR as a metric as follows:

$$SINR_{i,CH} = \frac{P_{rx,i}}{\sum_{k=1,k\neq i}^{C} P_{rx,k\notin CH} + N_0} \quad (6)$$

where C is the number of cluster member, i is the cluster member, and N_o is gaussian noise.

The sensor starts to send the data into CH with a selected power level. Then, CH evaluates the SINR of the received signal interfered with by other members and noise, as in Equation (6), and sends SINR information back to the sensor. Based on it and transmission power level, the sensor creates a Q table to memorize the action and chooses the appropriate power level to send the data in the next round. However, the sensor should not choose a high-power level that can decrease the other sensor quality and generate action from other sensors. The sensors collaborate indirectly to maintain the SINR through adjusting the power level until their *SINR>threshold*. The collaboration utilizes the learning process guided by the reinforcement learning method.

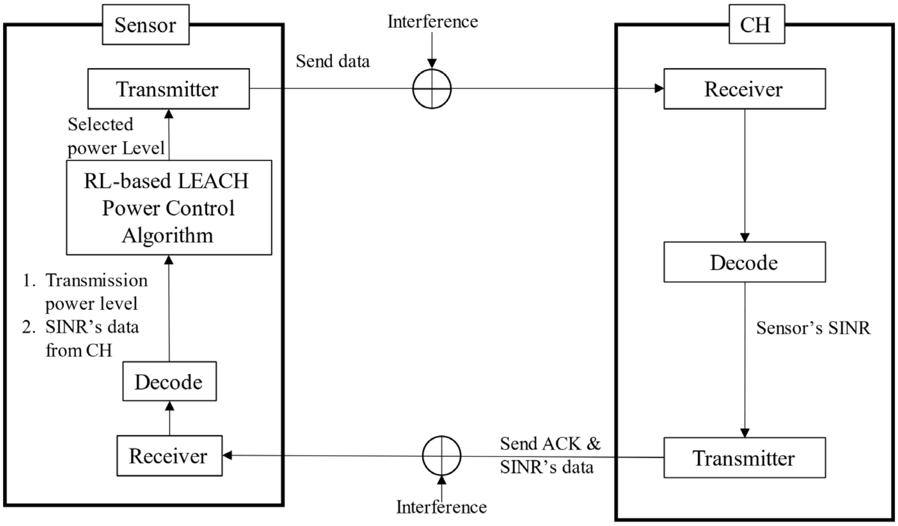

Figure 4. RL-based LEACH Power Control Block Diagram

The RL-based power control method consists of the reward system, Q value update, and action selection. The method applies three cycles of activities, namely select an action, reward the action based on the response of environment and value the new state. The cycles will continue until the system converges to the target.

A. Reinforcement Learning Method

Figure 5 shows the generic model for multi-agent reinforcement learning. The model consists of many agents that interact with the environment. The agent takes an action a_k that is select a power level and transmits data into the cluster head at a time steps k. The environment responds to the agent's action by giving a reward that represents how well the action is taken. When the agent does an action, it expects something good to happen in the environment. However, if the environment behaves otherwise, then the agent will get a low reward. After the environment gives a response, at time *k+1* the agent enters new state $s_{k+1} \in \{s_1, s_2, s_3\}$. In this research the states are s_1 ($SINR<Th_{low}$), s_2 (SINR between Th_{low} and Th_{high}) and s_3 ($SINR>Th_{high}$). The SINR transition is driven by the transmission power of the agent and interference. The objective is to accumulate the reward as much as possible until the state terminates (return of the system); the sensor updates Q value function based on current action and state $Q(s_k,a_k)$. There are three components in Q value that are current Q value, reward number, and the gap of current Q value and the best of Q value at the next state. The sensor will try to close the gap with select the right action. The next action selection will be guided by a policy π_k, where $\pi_k(s,a)$ is the probability that $a_k=a$ if $s_k=s$. The policy maps the states into the probability of selecting an action. In this research, the policy uses the Q value as guidance to select the next action selection that generates the maximum return. The system involves many agents that collaborate to achieve a common objective. In the research, the objective is to minimize energy consumption E_t in the network, which comprises energy sensor to transmit L bits (E_{TX}) and energy CH to receive and transmit data into the base station (E_{CH}).

$$E_t = min\left(\sum_{i=1,i\notin CH}^{N} E_{TX,i} + \sum_{k=1}^{M} E_{CH,k\in CH}\right) \quad (7)$$

with the constraint:

$$SINR_{i,CH} \geq Th_{low} \quad (8)$$

$$SINR_{i,CH} < Th_{high} \tag{9}$$

$$P_{tx,i} < P_{threshold} \tag{10}$$

where N is the number of nodes, M is the number of clusters, E_{CH} is receiving and transmission energy of CH, Th_{low} is the minimum SINR, Th_{high} is the maximum SINR, and $P_{threshold}$ is the maximum transmission power.

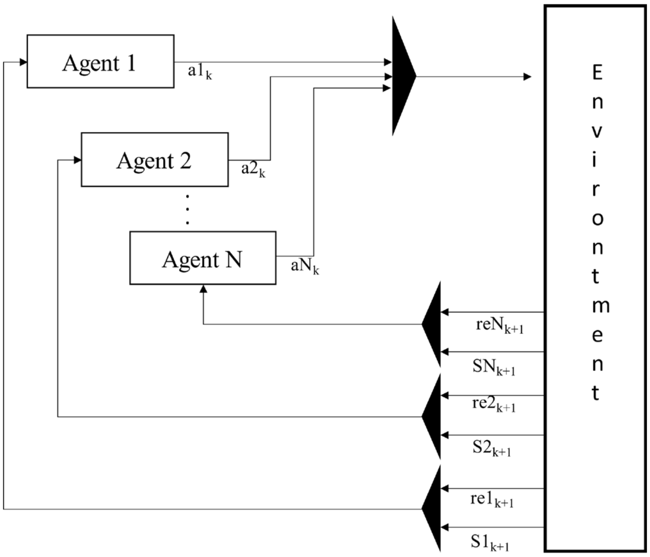

Figure 5. Multi-Agents Reinforcement Learning Method

B. Reward System

The reward system applies a positive reward that is the sensor will get a high reward if the environment's response close to the target and a low reward if the response is far from the target. The author uses the positive reward to drive the system to move to the target as fast as possible. The reward (re) is calculated as follows:

$$re_{SINR} = \begin{cases} 1 - \left[\frac{Th-SINR}{\Delta SINR}\right]^p, SINR \leq Th_{low} \\ 1 - \left[\frac{SINR-Th}{\Delta SINR}\right]^p, SINR > Th_{high} \end{cases} \tag{11}$$

$$re_a = 1 - \left(\frac{a_{max}-a_{min}}{\Delta a}\right)^p \tag{12}$$

$$re_{k+1} = re_{SINR} re_a \tag{13}$$

with Th is the SINR's threshold, p is a coefficient reward, a_{max} and a_{min} are maximum and minimum power level, ΔSINR is the range of SINR, and Δa is the range of power level. The component of the SINR reward value is determined by its distance from the threshold number (Th). The closer to threshold Th, the higher the reward, conversely the farther away from threshold Th, the smaller it is. Figure 6 displays the reward value curve that use value $0<p<1$ to implement a positive reward. The target is distance = zero, and we can see that the velocity is faster when the system approaching the target zero distance.

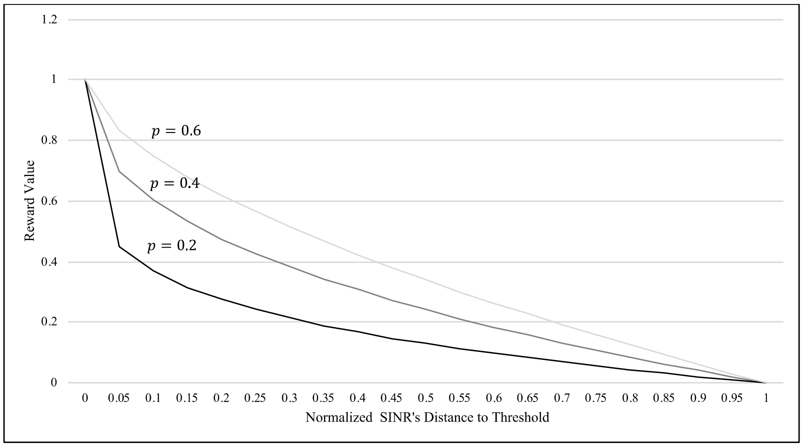

Figure 6. The Reward Values of Normalized the SINR's Distance

C. *Q Value*

The Q value represents how well the action selection through evaluation of the reward and the difference between Q value at k and discounted maximum Q value at *k+1* as follows [10]:

$$Q_{k+1} = Q_k(s_k, a_k) + b(r_{k+1} + y_{max} Q_a(s_{k+1}, a_k) - Q_k(s_k, a_k) \quad (14)$$

with b is learning factor and y is discounted factor.

The equation (14) is the Bellman equation, which is a simple value iteration update, using the weighted average of the previous value and the current data. The learning factor or step size determines how fast the system converges. If b is 0, then the agent ignores the future value (the agent exploits the prior knowledge), and if *b=1*, the agent considers the importance of the new value. The discounted factor determines the importance of the future value, if *y=0* means the agent only considers the current value and *y=1* means the agent strives for the long-term future value.

D. *Action Selection*

The agent will refer to the Q value in a specific state to guide an action selection. The action selection applies e-greedy method which use the exploration and exploitation phase. The agent generates random value z in each step and compare it with a predefined number e. If e is close to 1, then the agent enters exploration phase and if e is close to 0, the agent enters exploitation phase. The action selection policy is as follows [10]:

$$a_k = \begin{cases} U(a_{min}, a_{max}), z \leq e \\ Arg\ max\ Q_k(s_k, a_k), z > e \end{cases} \quad (15)$$

The agent applies the policy in Equation (15) at learning period until the system is converged, however after the converged states is achieved, the sensor will follow the new state policy as follows:

$$a_{t+1} = \begin{cases} arg\ max\ Q_k(s_{1,t+1}, a), if\ s_{t+1} = s_3 \\ arg\ max\ Q_k(s_{2,t+1}, a), if\ s_{t+1} = s_1 \end{cases} \quad (16)$$

with s_1 is *SINR<Th$_{low}$*, s_2 is *8<SINR<10*, and s_3 is *SINR>Th$_{high}$*.

When the state is converged, the system will follow the best Q value to select the power level. However, this raises the situation that the system cannot change to a lower power level when it is in s_3 or conversely the system cannot change to a higher power level when it is in s_1 because the system is already in a balanced environment. The author adopts the dual policy (Equation 15) to solve the situation.

E. Proposed Algorithm

Figure 7 exhibits the algorithm to run RL-based power control which is embedded in LEACH algorithm. At time the new round starts, the sensors execute LEACH method to select CH and form the clusters, then the RL-based power control algorithm is run.

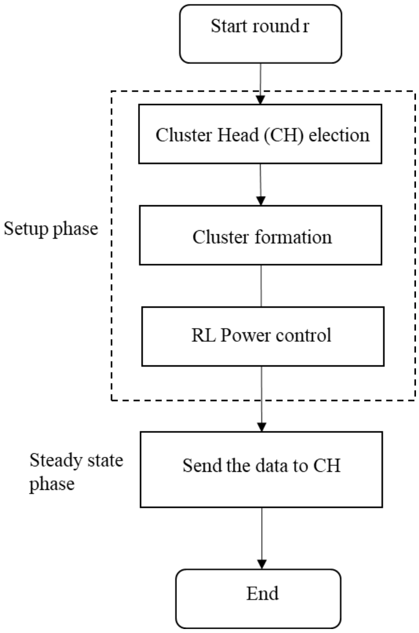

Figure 7. The System Flowchart

Algorithm 1 Reinforcement Learning Based Power Control

```
Initialize max_round, e, s , $a_{min}$, $a_{max}$;
Create LEACH cluster
for round=1:max_round --> round iteration
if all node dead
 break
end
for cl=1: number of cluster --> RL iteration
for i=1: number of iterations
Training session (learning and convergence)
for zz=2: cluster member
 set z=random
 if $z<e$
 select action random ($a_{min}$,$a_{max}$)
 else if i in testing stage
 if $s=s_3$
 Find a which max Q($s_1$:)
 else if $s=s_1$
 Find a which max Q($s_3$:)
 else
 Find a which max Q($s_2$)
 else
 Find a which max Q(s:)
 end
```

```
Convert a into power $P_{TX}=-24+(a-1)$
Set transmision power $P_{TX}= p=-24+(a-1)$
if $d>d_0$
  Calculate received power $P_{rx}=P_{TX}/d^4$
else
  Calculate received power $P_{rx}=P_{TX}/d^2$
end
Calculate SINR
Calculate reward re
Calculate Q(s,a,re)
Update Table Q(s,a)
end
end
Transmit data
end
```

F. System Parameter

The simulation is run using MATLAB with the parameters shown in Table 1. There are three group parameters, namely LEACH network parameters, RL parameters, and sensor parameters. The simulation covers sensing area *100x100* m^2 involving 100 sensors where 60 sensors have initial energy E_0, 20 sensors $1.5E_0$, and 20 sensors $1.75E_0$. The network applies the LEACH method to create clusters with LEACH predefined number d and round numbers 500. The RL-based power control uses the e-greedy factor (e), learning factor (b), and discounted factors (y). The rest is sensor parameters, namely data rate, packet length, and output power level.

Table 1. Simulation Parameters

Parameter	Value	Unit
Sensing area	100 x 100	m^2
Number of sensor (N)	100	sensor
Initial energy (E_0)	0.1	Joule
Threshold (Th)	8-10	dB
ϵ_{fs}	10^{-12}	$pJ/bit/m^2$
ϵ_{MP}	$0.0013x10^{-13}$	$pJ/bit/m^4$
E_{elect}	$50x10^{-9}$	Joule
E_{DA}	$50x10^{-9}$	Joule
Round (r)	500	
Distance (d)	0.1	
E-greedy factor (e)	0.7, 0.5, 0.1, 0.01	
Learning factor (b)	0.9, 0.5, 0.01	
Discounted factor (y)	0.8	
Data rate (R_b)	250	Kbps
Packet length (L)	4,000	Bits
Output power (programmable 8 step)	-24 sd 0	dBm

3. Results and Performance Evaluation

The proposed method is simulated to verify the performance. First, the convergence process will be discussed, then the system performance comparison with original LEACH will be presented.

A. *Convergence Process and Dual Policy Impact*

The first simulation is to evaluate the epsilon factor (e) variation in convergence process. The Figure 8 shows the simulation results which uses factor *e = 0.7, 0.5, 0.1* and *0.01*. The factor *e = 0.7* generates 70% of transmission power selection randomly and only 30% selection is guided by Q table, the system oscillates around 1.2. The factor *e=0.5* generates similar graphic with *e=0.7* but smaller deviation, and finally the simulation converges to 1.4 in *e=0.1* and *0.01*. Why this happen can be explained like this: the system uses factor *e=0.7* to learn the response of the environment and record it as Q value in the Q table, and then 30% selection will use the Q table to choose the power. This leads to a random graphic in Learning 1 with high deviation. However, the graphic is more stable in *e=0.5* as more power selections use the Q table. The broad variation is impacted by the variation of $Q(s_{k+1}, a_k)$-$Q(s_k,a_k)$; it is called the temporal difference learning. The system tries to achieve convergent value by minimizing the difference.

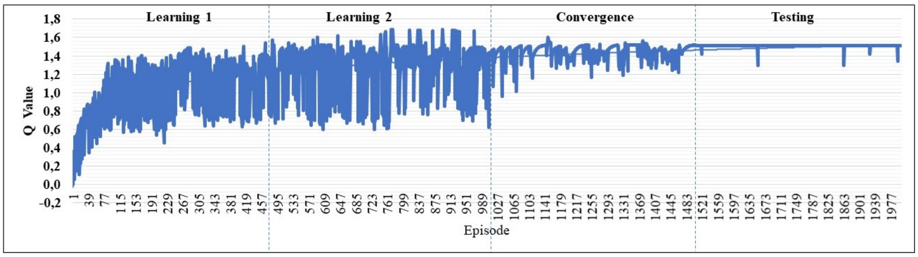

Figure 8. The Convergence Process

The second simulation is to evaluate the learning factor (b) variation in convergence. Figure 9 shows effect of learning factor (b) variation from 0.2, 0.5 dan 0.9. The curve is converging faster when b value is greater because it reduces the gap between maximum Q value and current Q value. Otherwise, if the b value is small then the system needs more action to achieve maximum Q value. However, the high b value will generate overshoot situation and take a time to converge. That is why for the simulation, the author selects *b=0.5* as learning factor for stability reason.

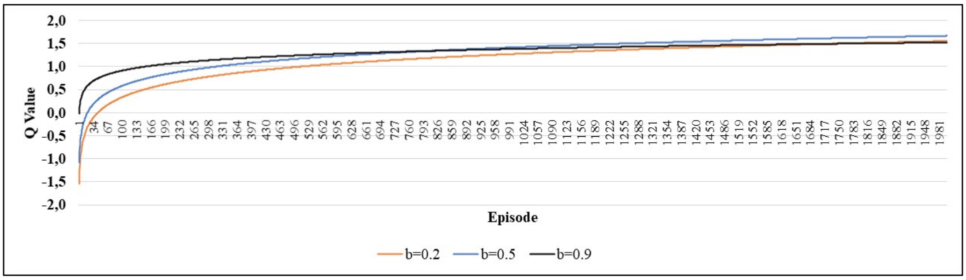

Figure 9. The Learning Factor Effect

Next, the system simulates the dual policy that is the policy when the system is converged and starts to send the data. The simulation compares if the system utilizes the e-greedy method only and if the system applies the dual policy (e-greedy method and new state policy method). Table 2 displays the result comparison where 71% of SINR of the single policy > Th, and 82% of SINR of the dual policy >Th; the dual policy has increased 11% of the signal quality (SINR). The dual policy has improved the system performance by choosing a power level from the lower state (s_1) if the *SINR>Th* (s_3), and vice versa.

Table 2. Comparison of single and dual policy

	Single Policy	Dual Policy
SINR>Th	71%	82%
SINR<Th	29%	18%

B. Performance Comparison

The author compares the proposed algorithm with the original LEACH to prove the improvement. The performance indicators are network lifetime (first node death/FND, half node death/HND, and all nodes death/AND), packet delivered, and average energy consumption per round. Figure 10 shows the network lifetime of both methods, where the original LEACH terminates in round 35 and the RL-based power control terminates in round 85. The graphic of both methods shows different patterns where the original one follows a staging pattern, and the proposed method displays a smoother one. The staging pattern comes from an equal energy consumption rate in CH that making a group of sensors death together. It makes the sensor is getting rarer and lengthen the distance between sensors. This drives sensor transmits with higher power and accelerates the whole sensor die. Otherwise, the proposed method will increase the LEACH capability to track SINR changes in CH and drive the sensors to lower the transmission power. It makes the energy consumption rate is not the same and generates a smoother graphic.

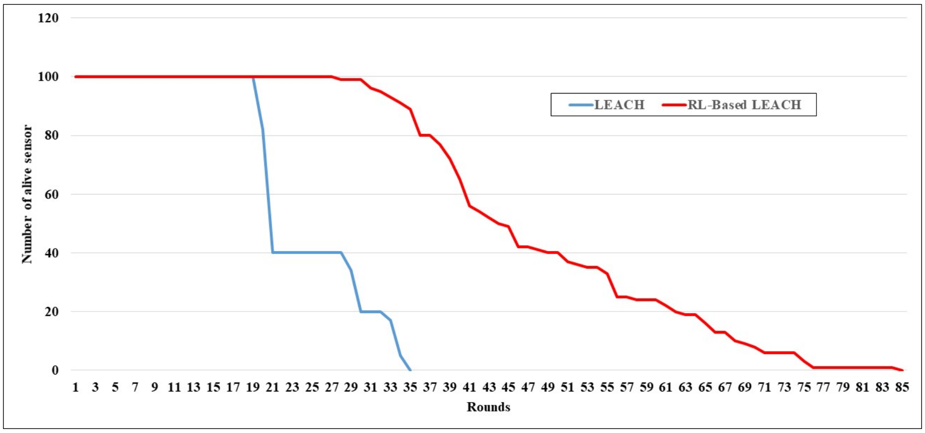

Figure 10. Network Lifetime Comparison

Figure 11 displays FND, HND, and AND indicators comparison with the proposed method is 35%, 52%, and 72% higher than the original method. The gap is increasing, which shows the energy saving of the proposed method better than the original method. The RL-based LEACH can track SINR changes in CH and adjust the next transmission power to keep SINR close to the threshold, leading to better energy saving and an impact on the lifetime of the sensor.

Figure 12a displays the delivered packet indicator, where the proposed method has increased by 162% compared to the original one. From the analysis in Figure 11, energy-saving resulting from minimum transmission power selection has lengthened the network lifetime by 72% compared with the original LEACH and reduces many sensors die every round. More sensors can transmit data into CH will increase the number of packets delivered in the proposed method.

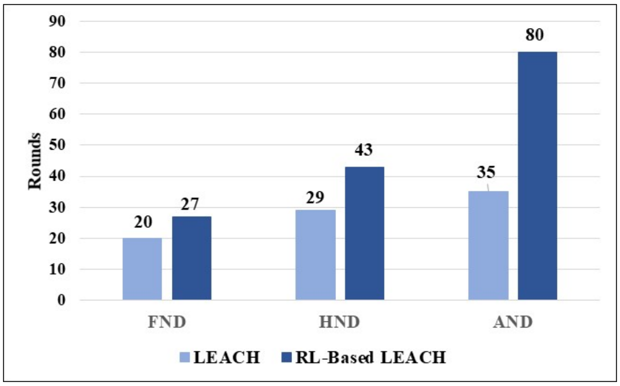

Figure 11. Sensor Lifetime Comparison

Figure 12b. compares the average energy consumption per round in both methods that show significant savings. The RL-based LEACH can reduce 55% energy spending compared to LEACH. The saving is caused by lower transmission power selection from a clever decision made by the sensor based on SINR observation. When the sensor goes to a new round, it creates a Q table which is updated in the learning and testing phase and guides on selecting a transmission power.

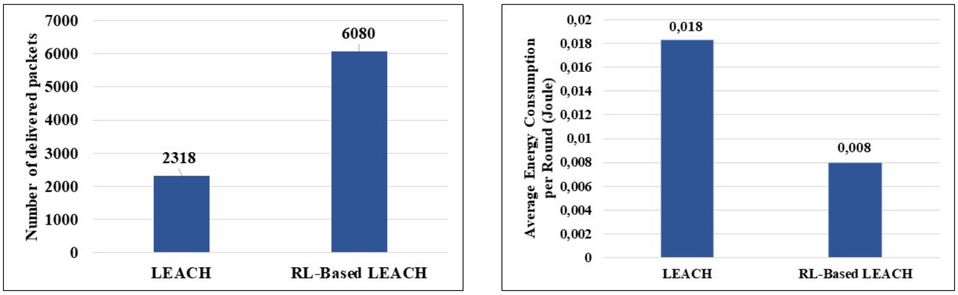

Figure 12a. Delivered Packets, b. Average Energy Consumption per Round

4. Conclusion

The RL-based LEACH has been evaluated to mitigate the interference and the sensor power consumption in a wireless sensor network under free space and ray-ground propagation. The RL-based power control simulation shows performance better than static power control in the LEACH cluster. The system has increased the low transmission power in the sensors by 11% and longer the network lifetime. We have shown that to reduce power consumption while maintaining the signal quality in CH, RL-based power control using SINR as a control metric is effective to lower transmission power and lengthen the network lifetime. The proposed method can track the interference changes and adjust the transmission power to anticipate the SINR degradation. The dual policy power control shows a better performance than that of the single policy in the steady-state situation. The system will lock to a certain transmission power even if the interference is degrading due to the small Q value variation. The dual policy enforces the power control to move to lower state and drives the system to change the transmission power. In this paper, we have shown the effectiveness of LEACH with double policy multi-agent reinforcement power control to mitigate the interference. However, we use a fixed SINR threshold to maintain the signal quality. In the future, the system can apply a dynamic threshold to fulfill the different types of data transfer in the network. The RL-based power control has increased the processing complexity and will burden the sensor as the number of sensors is increasing. This situation opens a new research area that can decrease the complexity while maintaining the SINR.

5. References

[1] A. Yousaf, F. Ahmad, S. Hamid, and F. Khan, "Performance Comparison of Various LEACH Protocols in Wireless Sensor Networks," in *2019 IEEE 15th International Colloquium on Signal Processing & Its Applications (CSPA)*, 2019, pp. 108–113, doi: 10.1109/CSPA.2019.8695973.

[2] D. W. H. Kurniawan, A. Kurniawan, and M. S. Arifianto, "Layering of CDMA Wireless Sensor Network Cluster to Improve Network Capacity," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 9, no. 4, p. 1129, Aug. 2019, doi: 10.18517/ijaseit.9.4.6502.

[3] S. K. Singh, P. Kumar, and J. P. Singh, "A survey on successors of LEACH protocol," *IEEE Access*, vol. 5, pp. 4298–4328, 2017.

[4] T. A. Chit and K. T. Zar, "Lifetime Improvement of Wireless Sensor Network using Residual Energy and Distance Parameters on LEACH Protocol," in *2018 18th International Symposium on Communications and Information Technologies (ISCIT)*, 2018, pp. 1–5, doi: 10.1109/ISCIT.2018.8587930.

[5] D. W. H. Kurniawan, A. Kurniawan, and M. S. Arifianto, "An analysis of optimal capacity in cluster of CDMA wireless sensor network," in *Proceedings - 2017 International Conference on Applied Computer and Communication Technologies, ComCom 2017*, 2017, vol. 2017-Janua, pp. 1–6, doi: 10.1109/COMCOM.2017.8167080.

[6] L. Zheng, W. Wang, A. Mathewson, B. O'Flynn, and M. Hayes, "An adaptive transmission power control method for wireless sensor networks," in *IET Irish Signals and Systems Conference (ISSC 2010)*, 2010, pp. 261–265, doi: 10.1049/cp.2010.0523.

[7] M. M. Y. Masood, G. Ahmed, and N. M. Khan, "A Kalman filter based adaptive on demand transmission power control (AODTPC) algorithm for wireless sensor networks," in *2012 International Conference on Emerging Technologies*, 2012, pp. 1–6, doi: 10.1109/ICET.2012.6375499.

[8] M. M. Y. Masood, G. Ahmed, and N. M. Khan, "Modified on demand transmission power control for wireless sensor networks," in *2011 International Conference on Information and Communication Technologies*, 2011, pp. 1–6, doi: 10.1109/ICICT.2011.5983544.

[9] M. Chincoli, C. Bacchiani, A. A. Syed, G. Exarchakos, and A. Liotta, "Interference Mitigation through Adaptive Power Control in Wireless Sensor Networks," in *2015 IEEE International Conference on Systems, Man, and Cybernetics*, 2015, pp. 1303–1308, doi: 10.1109/SMC.2015.232.

[10] M. Chincoli and A. Liotta, "Self-learning power control in wireless sensor networks," *Sensors*, vol. 18, no. 2, p. 375, 2018.

[11] S. Gengtian, T. Koshimizu, M. Saito, P. Zhenni, L. Jiang, and S. Shimamoto, "Power Control Based on Multi-Agent Deep Q Network for D2D Communication," in *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 2020, pp. 257–261, doi: 10.1109/ICAIIC48513.2020.9065192.

[12] A. Almasri, A. Khalifeh, and K. A. Darabkh, "A Comparative Analysis for WSNs Clustering Algorithms," in *2020 Fifth International Conference on Fog and Mobile Edge Computing (FMEC)*, 2020, pp. 263–269, doi: 10.1109/FMEC49853.2020.9144765.

[13] R. S. M. Alyousuf, "Analysis and Comparison on Algorithmic Functions of Leach Protocol in Wireless Sensor Networks [WSN]," in *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 2020, pp. 1349–1355, doi: 10.1109/ICSSIT48917.2020.9214149.

[14] D. Chauhan, S. Iyer, and B. Jajal, "Enhanced Prototype Model for Energy Efficient Algorithm of LEACH over Wireless Sensor Network," in *2019 6th International Conference on Computing for Sustainable Global Development (INDIACom)*, 2019, pp. 569–573.

[15] P. Santi, *Topology Control in Wireless Ad Hoc and Sensor Networks*. 2005.

[16] S. V Purkar and R. S. Deshpande, “Energy Efficient Clustering Protocol to Enhance Performance of Heterogeneous Wireless Sensor Network: EECPEP-HWSN,” *J. Comput. Networks Commun.*, vol. 2018, p. 2078627, 2018, doi: 10.1155/2018/2078627.

[17] Istikmal, A. Kurniawan, and Hendrawan, “Throughput performance of transmission control protocols on multipath fading environment in mobile ad-hoc network,” in *2017 11th International Conference on Telecommunication Systems Services and Applications (TSSA)*, 2017, pp. 1–5, doi: 10.1109/TSSA.2017.8272939.

[18] S. Gupta and N. Marriwala, “Improved distance energy based LEACH protocol for cluster head election in wireless sensor networks,” in *2017 4th International Conference on Signal Processing, Computing and Control (ISPCC)*, 2017, pp. 91–96, doi: 10.1109/ISPCC.2017.8269656.

Dwi Widodo Heru Kurniawan graduated from the Surabaya Institute of Technology (ITS), Indonesia, in 1993, received the Master of Technology degree from STT Telkom, in 2009 and the Doctor degree from the Bandung Institute of Technology, in 2021, in telecommunication engineering. He joins PT Telkom Indonesia since 1995 until now with digital business expertise. His research interests are IoT system, cellular communication system, and radio communication.

Mohammad Sigit Arifianto received the B.S. degree in Electrical Engineering from the Institut Teknologi Bandung, Indonesia, in 1998, the M.S. degree in Electrical Engineering from the University at Buffalo, NY, USA, in 2003, and the Ph.D. degree in Telecommunication from the Universiti Malaysia Sabah, Malaysia, in 2010. From 2008 to 2010, he was a lecturer in the Computer Engineering Program of the School of Engineering and Information Technology, the Universiti Malaysia Sabah. In 2010, he joined the School of Electrical Engineering and Informatics, the Institut Teknologi Bandung, in the Telecommunication Engineering Program, where he is currently an Assistant Professor (appointed in 2016). His research interests include the development of new techniques for future wireless communications in the areas of multiple access, multiple-input–multiple-output systems, channel coding, cognitive radio, wireless optical communications, and wireless sensor networks.

Adit Kurniawan graduated from the Department of Elecrical Engieneeeing, Bandung Institute of Technology (ITB), Indonesia, in 1986. He then received M.Eng. degree from RMIT, Australia, in 1996 and Ph.D. degree from the University of South Australia, in 2003, both in Telecommunication Engineering. He is currently Professor and serves as the Chair of Telecommunication Engineering Research Group at the School of Electrical Engineering and Informatics, Bandung Insitute of Technology, Indonesia. His research interest covers the antenna and wave propagation, cellular communication system, and spread spectrum communications.