Reinforcement learning for microgrid energy management

Elizaveta Kuznetsova, Yan-Fu Li, Carlos Ruiz , Enrico Zio,
Graham Ault, Keith Bell
Energy, Volume 59, 2013, Pages 133–146

Presenter: Elham Foruzan

Elham.foruzan@huskers.unl.edu

Overview Day One

- Goal & objectives
- System design
- Markov chain model for wind gen
- System Model
- Reinforcement Learning at Customers

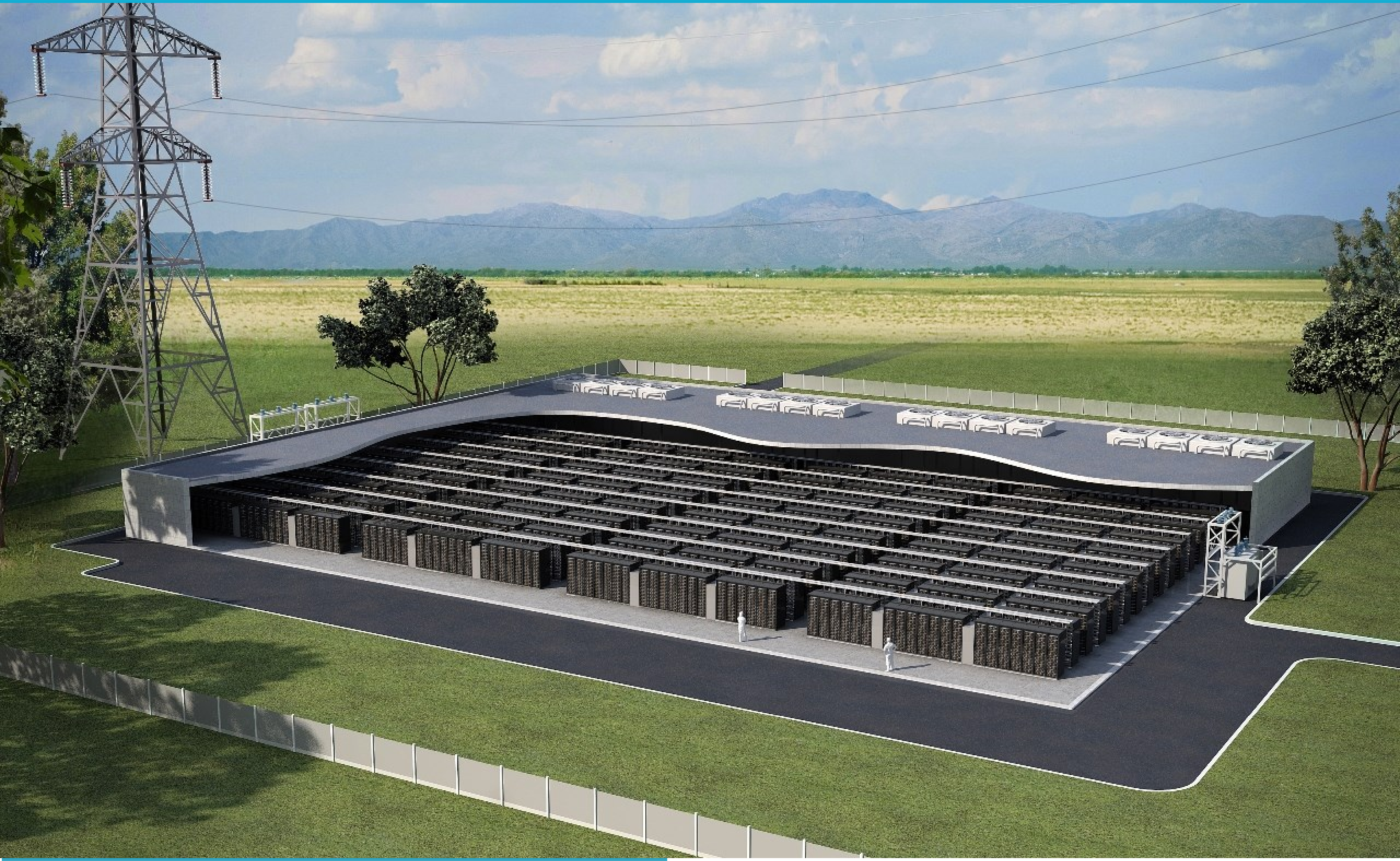
Overview Day Two

- Sensitivity analysis of learning parameters
- Simulation results and analysis
- Conclusion

Overview

- Goal & objectives
- System design
- Markov chain model for wind gen
- System Model
- Reinforcement Learning at Customers

Battery storage

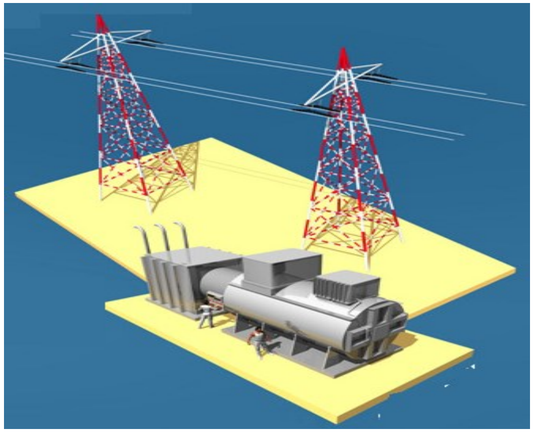


Source: http://engineeringinsightspodcast.com/episode002/

Goal & objectives

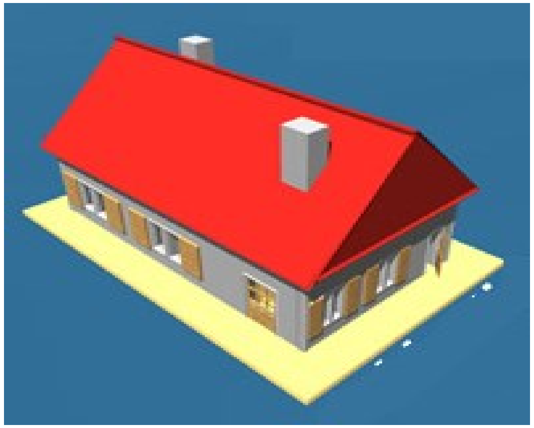
Paper goal: Increasing the utilization rate of the battery during high electricity demand, and increasing the utilization rate of the wind turbine for local use.
Customer: Consuming electricity.
Wind Turbine: Renewable generation.
Main Grid: External grid.
Battery Storage: Charge and discharge electricity.
Method: Two steps-ahead reinforcement learning algorithm to plan the battery scheduling.

Source: www.betterworldsolutions.eu/



Main Grid

Source: energy.gov



Residential

Microgrid and Battery

Source: http://www.energiestro.net/applications/

Overview

- Goal & objectives
- System design
- Markov chain model for wind gen
- System Model
- Reinforcement Learning at Customers

System Design

- The algorithm integrates two blocks:
  1. A forecasting step: design for stochasticity of the wind speed conditions.
  2. An optimization: design for adaptive task for finding the strategy of battery scheduling optimal

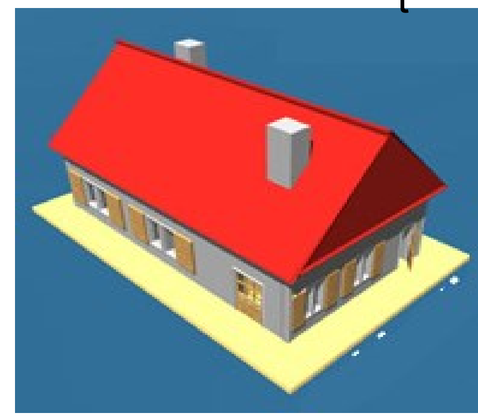
- The time step for the energy system optimization is set to be 1 h.

Microgrid design

The external grid imposes technical constraints and sets the market electricity price Pt.

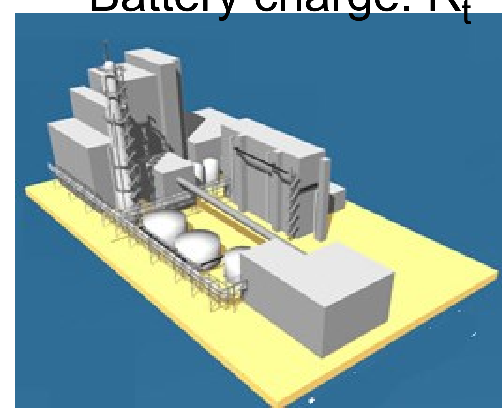
Source: ww.hydrogencarsnow.com

Inelastic load D_t



Power output P_{wt}

Battery charge: R_t



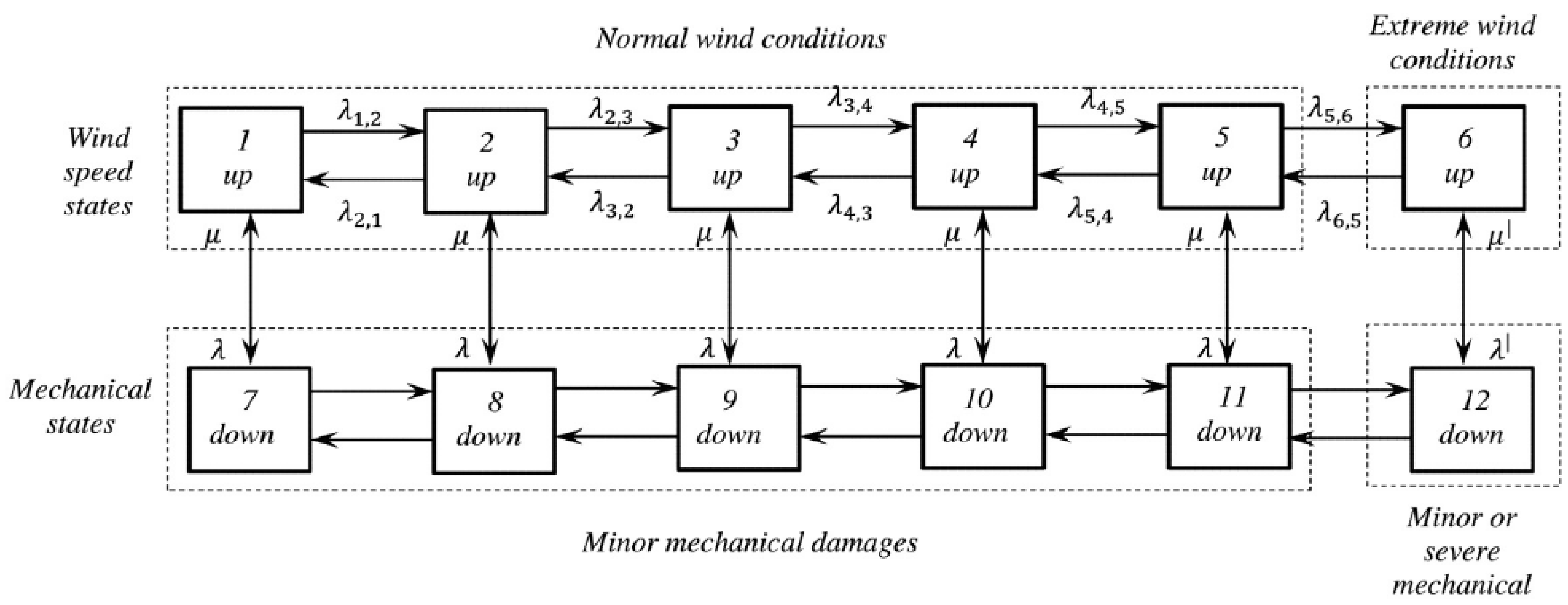
Overview

- Goal & objectives
- System design
- **Markov chain model for wind gen**
- System Model
- Reinforcement Learning at Customers

Model of the wind generator

- The amount of electricity output from the wind depends on :
  - Availability of the wind source.
  - Random mechanical failures of the wind generator components.

- Describe the dynamics of stochastic transition among different levels of wind **speed conditions** and **mechanical states**.

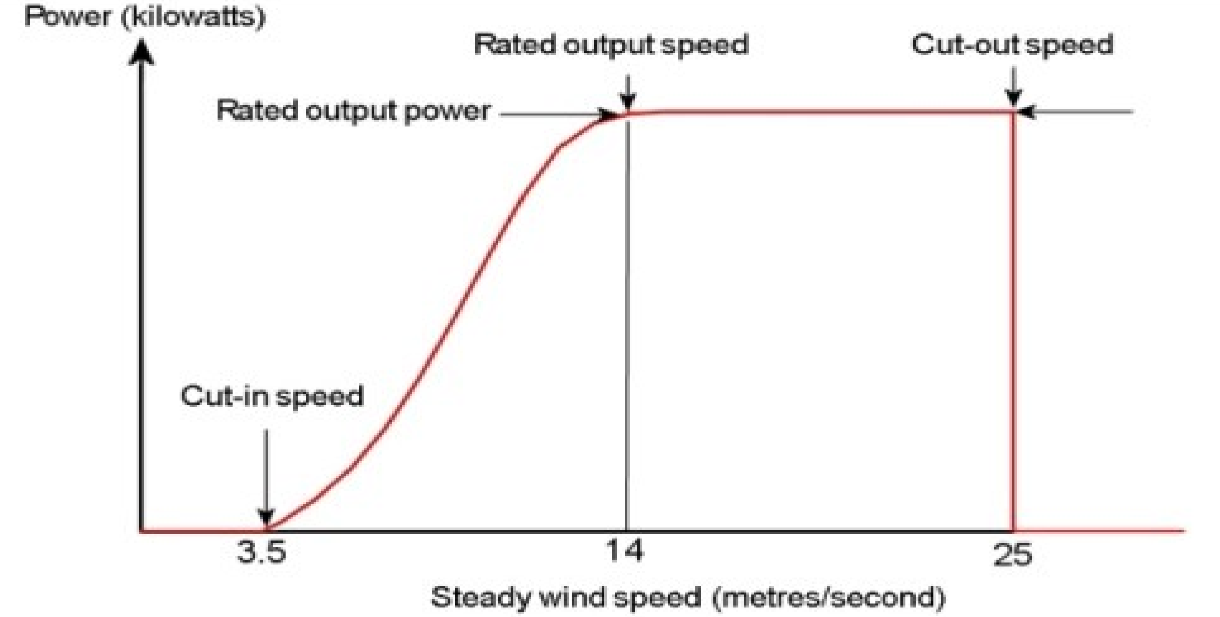


Wind power curve

- Wind power based on available wind

$$P(v) = \begin{cases} 0, & v \le v_{ci} \text{ or } v > v_{co} \\ P_N, & v_N \le v \le v_{co} \\ f(v), & v_{ci} < v < v_N \end{cases}$$

v_{ci} , cut-in
v_{co} , cut-out
v_N, nominal wind speed

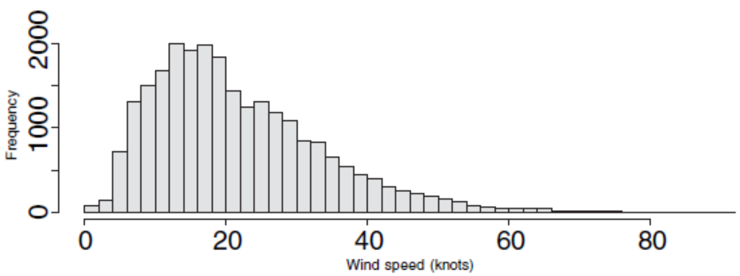


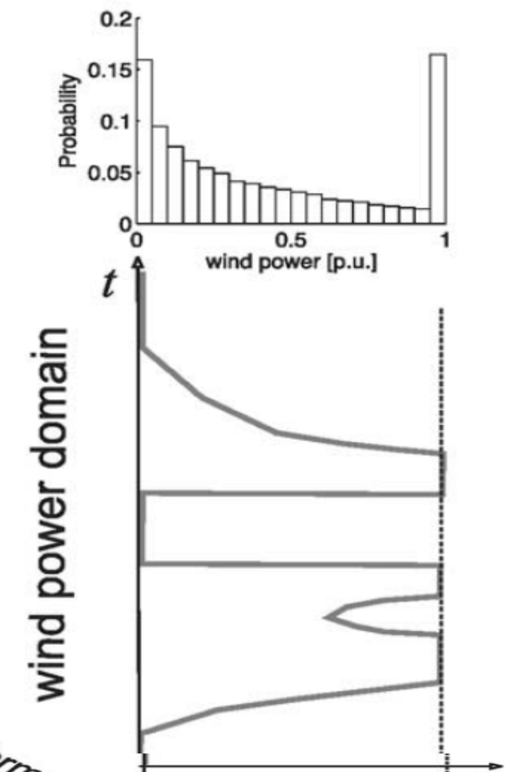
Typical wind turbine power output with steady wind speed.

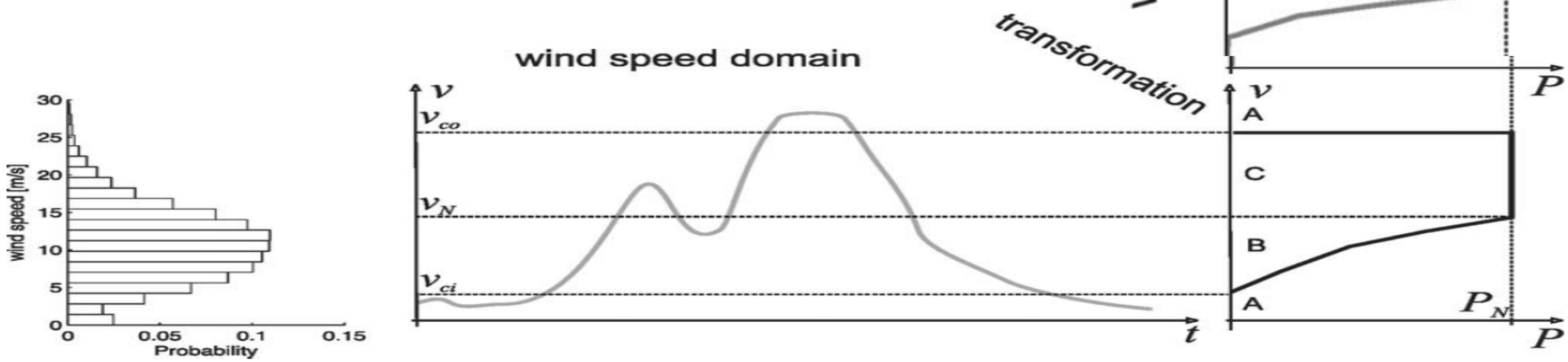
Obtaining wind power distribution function

❖ Obtaining the statistical property associated to the wind energy output over time.

✓ **Wind Speed frequency histogram**







MCMC for Wind Power Simulation

- **Markov chain** is define:
- A set of state S and transition probability from any two states.

$$\Pr(X_t = j | X_{t-1} = i) = p_{ij}$$

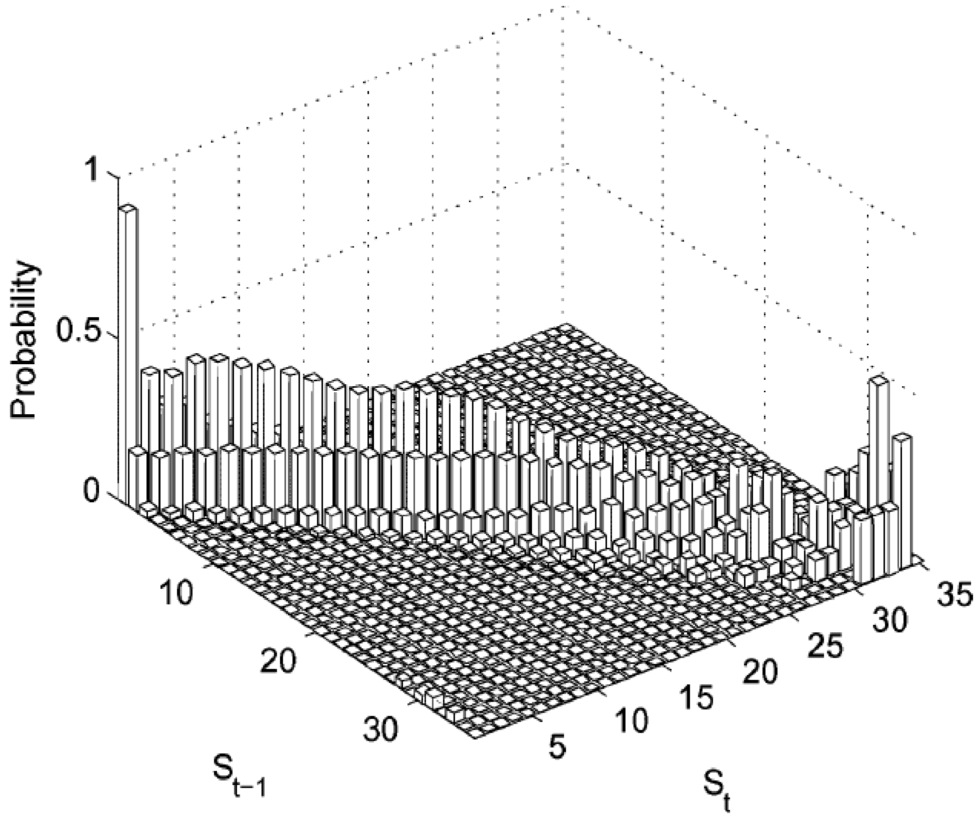
- Transition Probability

$$\mathbf{P} = \begin{array}{cc} & S_t \rightarrow \\ S_{t-1} \downarrow & \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{bmatrix} \end{array}$$

- States are discretized wind speeds values.
- Two extra discrete states, namely $P \equiv 0$ and $P \equiv P_N$

Example

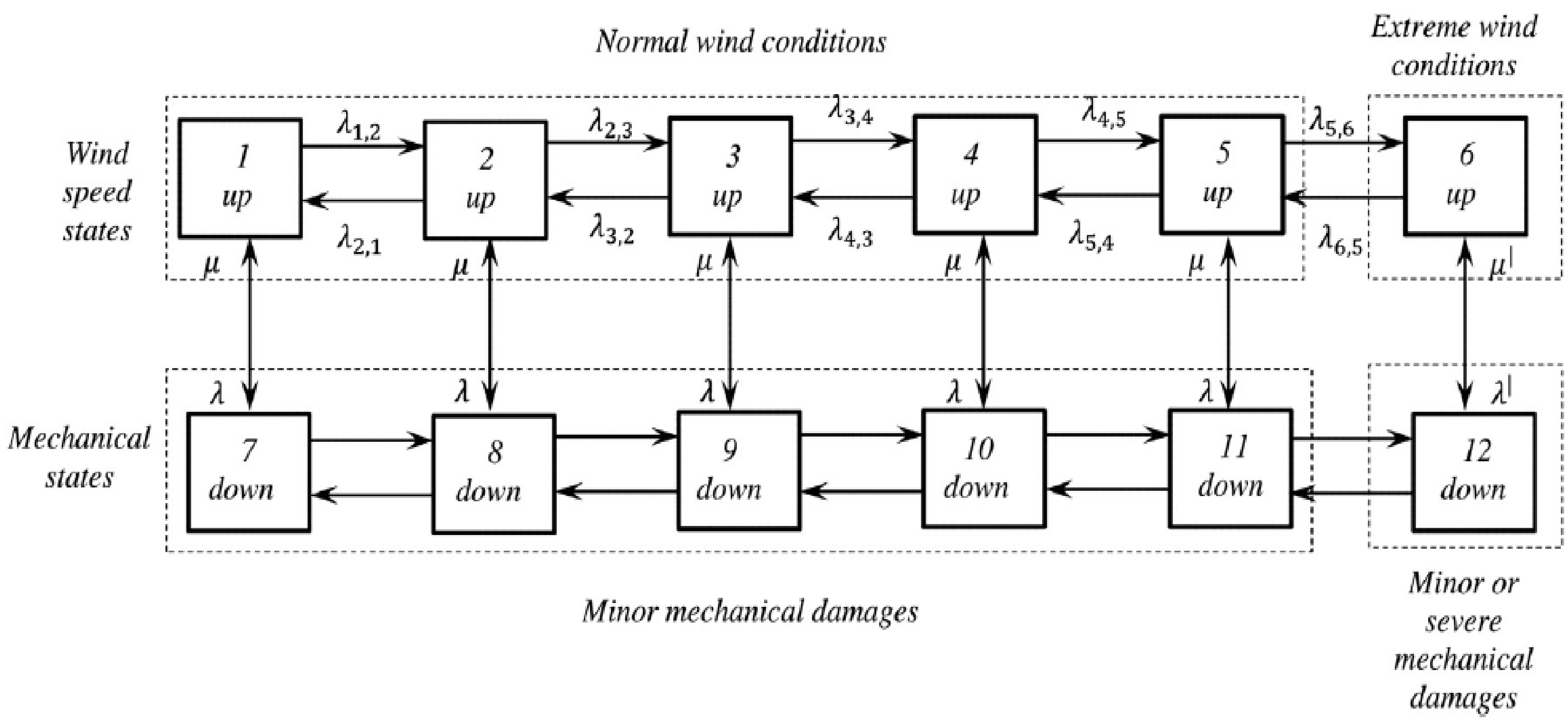
- The maximum recorded wind speed is 34.4 m/s
- Wind speed between 0-35 is divided to 35 states



Model of the wind generator

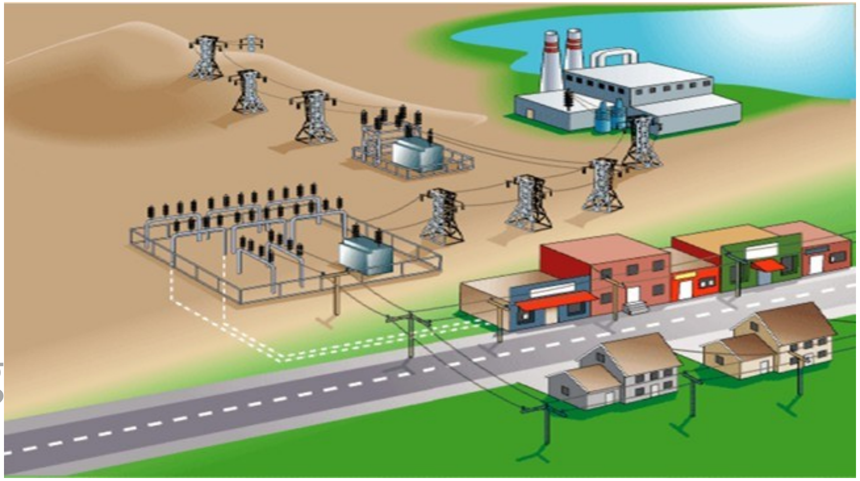
Markov chain for modelling:

- ❑ Describe the dynamics of stochastic transition among different levels of wind **speed conditions** and **mechanical states**.
- ❑ Discrete wind speed width is 3 m/s.
- ❑ States 7-12 represent the wind turbine failure states.



Overview

- Goal & objectives
- System design
- Markov chain model for wind g
- **System Model**
- Reinforcement Learning at Customers



Model of load

$$D_t = D^{\text{peak}} \cdot r_w^{\text{peak}} \cdot r_d^{\text{peak}} \cdot r_h^{\text{peak}}$$

D^{peak}: Is maximum hourly peak of power demand over a year

r_w^{peak} : Is the weekly peak of power demand

r_d^{peak} : Is the daily peak of power demand

r_h^{peak} : Is the hourly peak defined for working days and weekends

Model of battery storage

$$R_t = R_{t-1} + R_t^{\text{stor,charge}} - R_t^{\text{stor,discharge}}$$

R_t : Level of the energy stored in the battery at time t (Wh)

R_{t-1} : Level of the energy stored in the battery at time t_1 (Wh)

$R_t^{\text{stor,discharge}}$: The power flows over time step interval t between battery and consumer (Wh)

$R_t^{\text{stor,charge}}$: The power flows over time step interval t between wind generator and battery (Wh)

Overview

- Goal & objectives
- System design
- Markov chain model for wind gen
- System Model
- **Reinforcement Learning at Customers**

Definition of scenarios and actions

State And Scenarios

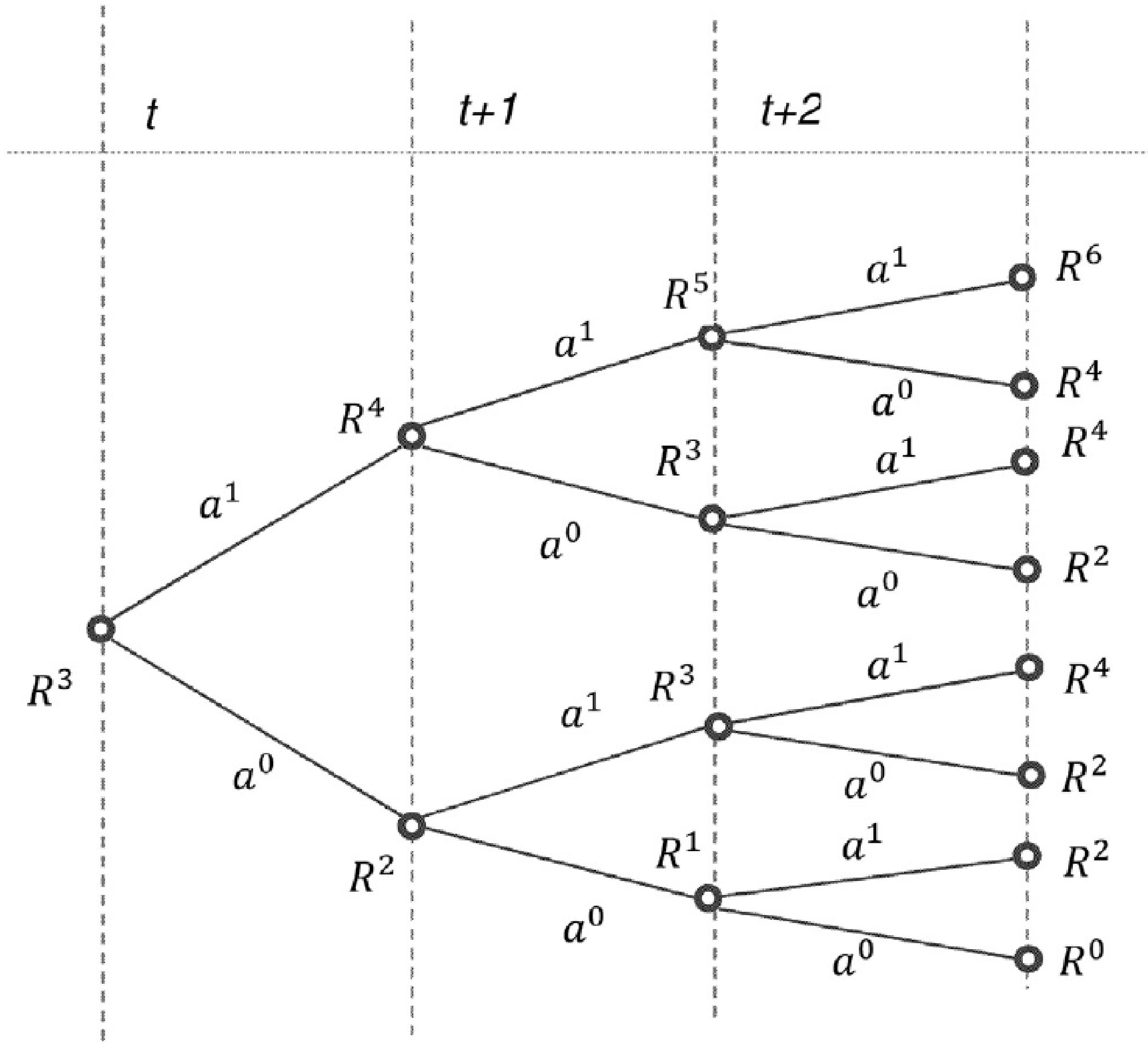
- S^i_t at time t is a set of $[D_t , P^{WT}_t]$.
- Scenario $S_l=[S^i_t , S^n_{t+1} , S^p_{t+2}]$
- At time t, battery decide for action at time interval [t, t+1 and t+2]
- Battery States $R_t = [R^0, R^1, R^2, R^3, R^4, R^5, R^6]$

Actions

- $A_j{}^t =[a_t, a_{t+1} , a_{t+2}]$

- a_t $\begin{cases} a0 & \text{Covering part of the consumer electricity demand by discharging the battery} \\ a1 & \text{Purchasing all the electricity demanded by the consumer from the external grid} \end{cases}$

Example



t
t+1
t+2
R^3
a^1
a^0
R^4
R^2
a^1
a^0
a^1
a^0
R^5
R^3
R^3
R^1
a^1
a^0
a^1
a^0
a^1
a^0
a^1
a^0
R^6
R^4
R^4
R^2
R^4
R^2
R^2
R^0

Problem Formulation

1-Initialize to 0 the Q-values of all possible actions sequences for each scenario and set time t =0.

2- Identify $s^i_t=[D_t , P^{WT}_t]$

Make the forecast of available wind power output P^{WT}_{t+1}, P^{WT}_{t+2} and load D_{t+1}, D_{t+2}

Identify Scenario $S_l=[S^i_t , S^n_{t+1} , S^p_{t+2}]$

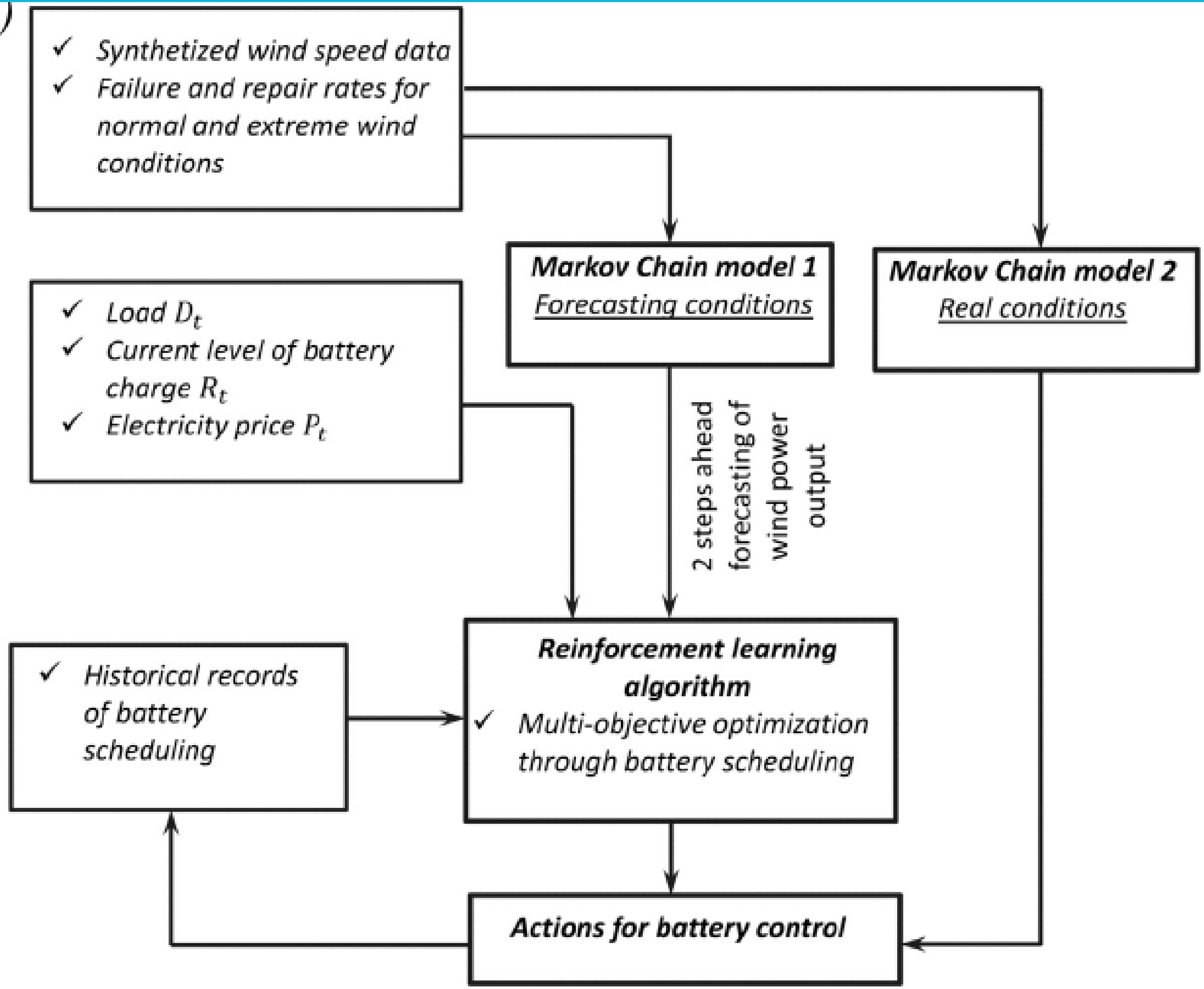
Based on identified S_l and battery charge R_t ,define all possible actions sequences of battery scheduling for 2 steps ahead.

Apply the policy for selection of sequence of actions.

Perform the selected sequence $A_j^{\,t}$ under real system conditions, simulated using the Markov chain model for real wind conditions. Update the value of the sequence performed.

3- Move to time step t+3; repeat step 2.

Algorithm


✓ Synthetized wind speed data
✓ Failure and repair rates for normal and extreme wind conditions
Markov Chain model 1
Forecasting conditions
Markov Chain model 2
Real conditions
✓ Load D_t
✓ Current level of battery charge R_t
✓ Electricity price P_t
2 steps ahead forecasting of wind power output
✓ Historical records of battery scheduling
Reinforcement learning algorithm
✓ Multi-objective optimization through battery scheduling
Actions for battery control

Reward functions

Final Goal: Increase the consumer independence from the external grid

i. Increasing the utilization rate of the battery during high electricity demand.

ii. Increasing the utilization rate of the wind turbine for local use.

$$f_t(a_t) = \begin{cases} \frac{P_t^{\text{wt}}}{D_t}\left(D_t - R_t^{\text{stor,discharge}}\right), & \text{if } a_t = a^0 \\ k \cdot \left(P_t^{\text{wt}} - R_t^{\text{stor,charge}}\right), & \text{if } a_t = a^1 \& P_t^{\text{wt}} > 0 \\ 0, & \text{if } a_t = a_1 \& P_t^{\text{wt}} = 0 \end{cases}$$

Link

Q learning and reward

- Maximizing the reward of current and future by performing sequence of actions $A_j^{\ t} = [a_t, a_{t+1}, a_{t+2}]$,

$$r\left(\text{Scenario}_t^l, A_t^j\right) = \gamma^0 \cdot f_t(a_t) + \gamma^1 \cdot f_{t+1}(a_{t+1}) + \gamma^2 \cdot f_{t+2}(a_{t+2})$$

Updating the Q value

$$Q\left(\text{Scenario}_t^l, A_t^j\right)_p = Q\left(\text{Scenario}_t^l, A_t^j\right)_{p-1} + \alpha\left[r\left(\text{Scenario}_t^l, A_t^j\right)_p - Q\left(\text{Scenario}_t^l, A_t^j\right)_{p-1}\right]$$

Recap

- Introduction & background to Batteries
- System design
- Markov chain model for wind gen
- Reinforcement Learning at Customers

Overview Day Two

- ❖ **Sensitivity analysis of learning parameters**
- ❖ **Simulation results and analysis**
- ❖ **Discussion &Conclusions**

Sensitivity analysis

- To understand the role of the learning parameters:
  - The weight coefficient k
  - The discounted rate γ
  - The learning rate α

Two scenarios

- Two scenarios
  - Scenario 1 : low wind power output, and medium and high values of load
  - Scenario 2 : high wind power output and low load.

- Initial battery size is 3000 W

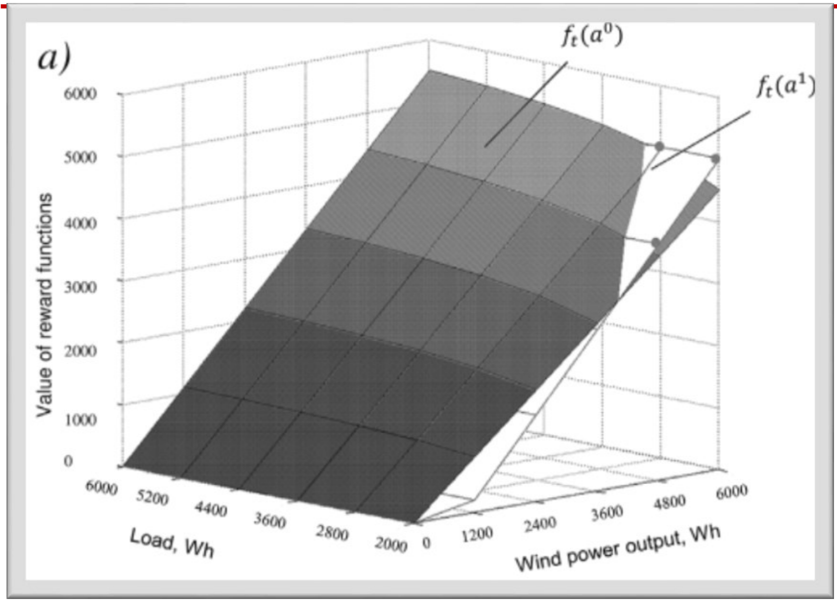
Parameters	Time steps		
	t	$t+1$	$t+2$
Scenario 1 with initial battery charge $R_t - 3000$ Wh			
Wind power output (P_t^{wt}), Wh	1200	1200	1200
Load (D_t), Wh	4400	5200	5200
Scenario 2 with initial battery charge $R_t - 3000$ Wh			
Wind power output (P_t^{wt}), Wh	6000	4800	4800
Load (D_t), Wh	2800	2800	2800
Possible sequences of actions $[a_t, a_{t+1}, a_{t+2}]$	a^0	a^0	a^0
	a^0	a^0	a^1
	a^0	a^1	a^0
	a^0	a^1	a^1
	a^1	a^0	a^0
	a^1	a^0	a^1
	a^1	a^1	a^0
	a^1	a^1	a^1

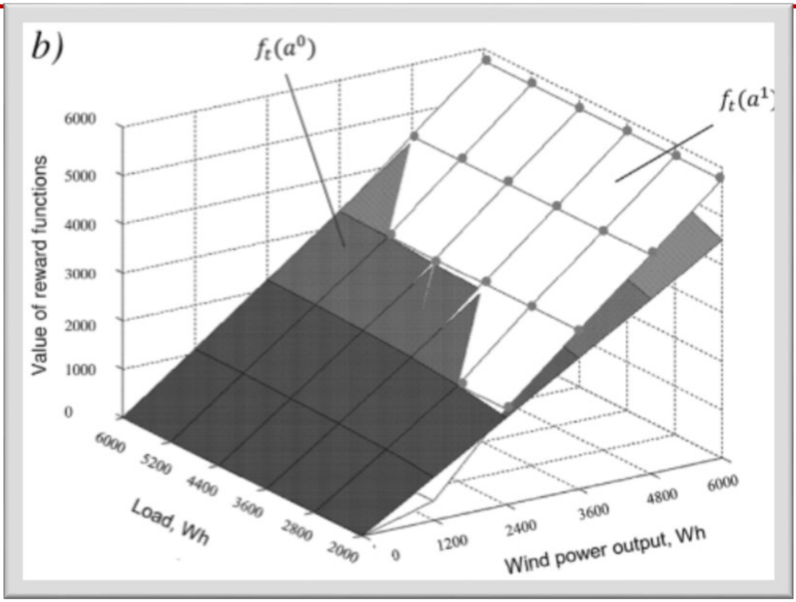
Possible values of the weight k

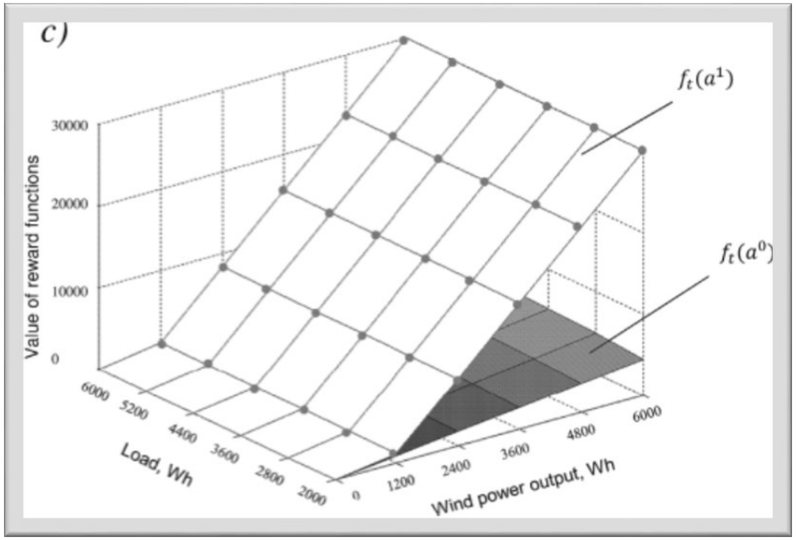
- The possible scenarios can be divided into three groups, depending on which of the following conditions is met.

- $f_t(a_0) > f_t(a_1)$
  - High loads and low wind power outputs.
- $f_t(a_0) < f_t(a_1)$
  - Low loads and high wind power outputs.
- $f_t(a_0) = f_t(a_1)$,
  - Where both actions a_0 and a_1 are equally valuable.

Possible values of the weight k







$f_t(a_0)$: grey-coloured surface
$f_t(a_1)$: white-coloured surface

a) $k = 1$

b) $k = 2^{\frac{1200}{P_t^W}}$

c) $k = 6$

Link

Influence of the weight k on the optimal sequence of actions

- Use sensitivity analysis to pick *k*:

  γ = .8, α=0.6

  - large *k* increase the selection of action a_1
  - Small k favors actions a_0

- For long term benefits, they consider the potential of absence of wind
  - k=6

Value of weight coefficient k	Scenario 1	Scenario 2
1	$[a^0, a^0, a^0]$	$[a^1, a^1, a^1]$
$2^{1200/P_t^{wt}}$	$[a^0, a^0, a^0]$	$[a^1, a^1, a^1]$
6	$[a^1, a^1, a^1]$	$[a^1, a^1, a^1]$

Discounted rate γ

k = 6, α=0.6

γ

If set to zero : values of actions undertaken at time steps t + 1 and t + 2 are neglected and only the first action at time step t is valuable

For the range 0.2 to 1: do not influence the sequence of actions with highest Q*-value.

Final Value set to 0.8

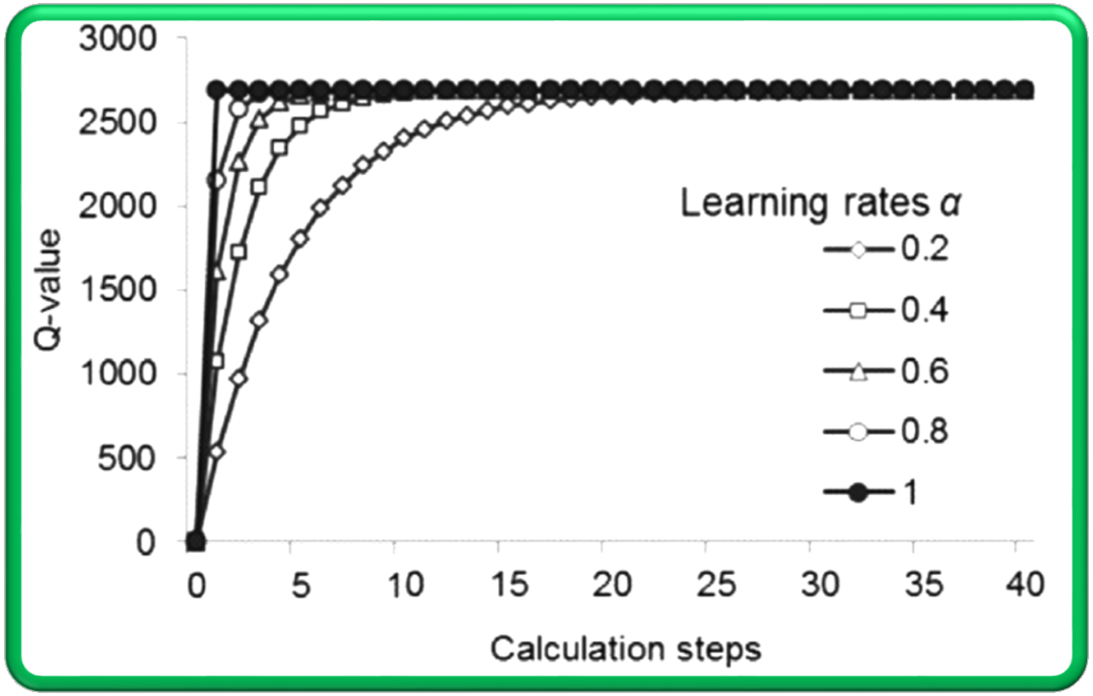
Value of discounted rate γ	Scenario 1/Scenario 2
0	$[a^1, a^1, a^1]$, $[a^1, a^1, a^0]$, $[a^1, a^0, a^1]$, $[a^1, a^0, a^0]$
0.2	$[a^1, a^1, a^1]$
0.4	$[a^1, a^1, a^1]$
0.6	$[a^1, a^1, a^1]$
0.8	$[a^1, a^1, a^1]$
1	$[a^1, a^1, a^1]$

Learning rate

- The value of the learning rate a influences the speed of convergence to Q*-values but not to the final highest Q*-values.
- The $\boldsymbol{\alpha}$ close to zero slowdown the convergence of Q values.

  $\boldsymbol{\gamma} = .8$, $\boldsymbol{k}=6$

- They select $\boldsymbol{\alpha}=1$



Overview Day Two

- Sensitivity analysis of learning parameters
- **Simulation results and analysis**
- Discussion &Conclusions

Simulation results and analysis

- The values of D_t , $P^{WT}{}_t$, and R_t are divided to six discrete values.

- Wind Power : [0, 1200, 2400, 3600, 4800, 6000] Wh
- Load: [2000, 2800, 3600, 4400, 5200, 6000] Wh
- Battery: [0, 1000, 2000, 3000, 4000, 5000, 6000] Wh
- Charging or discharging at each time step is : 1000 Wh

Wind turbine parameter

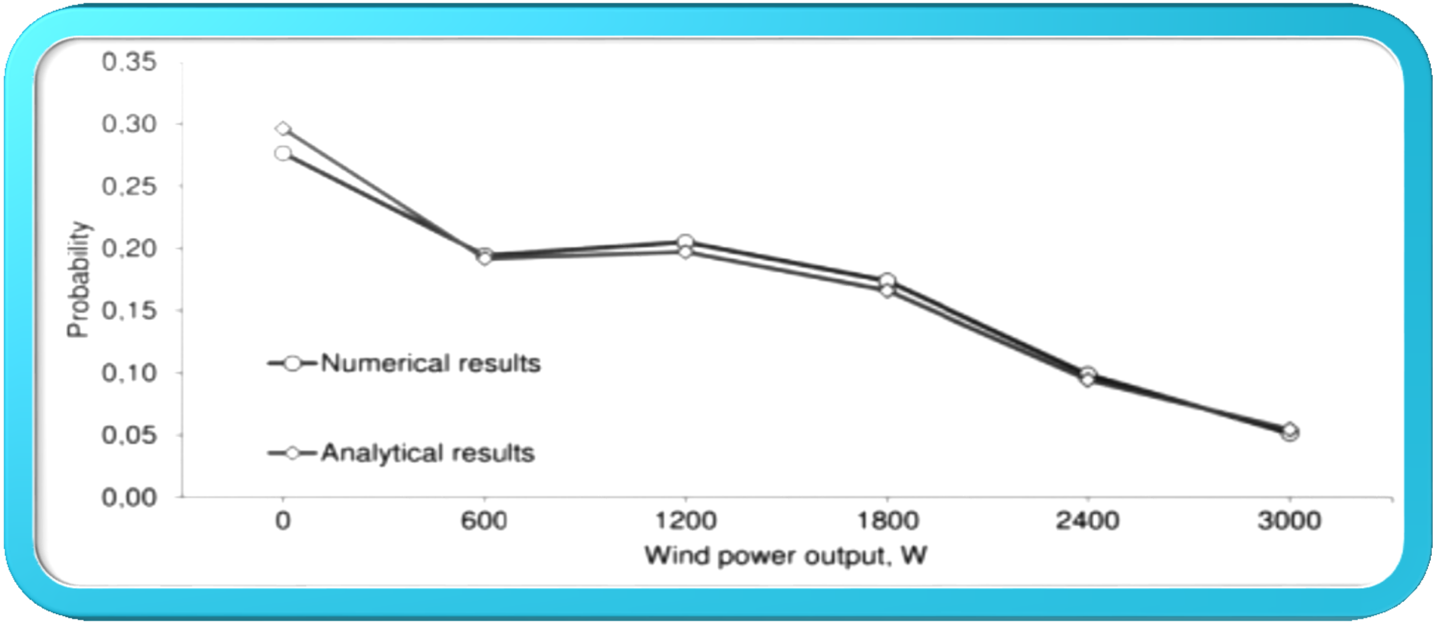
❖ Wind power output is proportional to the rated power of the wind generator

$$P_t^{\mathrm{wt}} = \begin{cases} 0, & \text{if } \; v < v_{\mathrm{ci}} \\ P^r \cdot \dfrac{(v_t - v^{\mathrm{ci}})}{(v^r - v^{\mathrm{ci}})} \cdot \Delta t, & \text{if } \; v_{\mathrm{ci}} \leq v < v_r \\ P^r \cdot \Delta t, & \text{if } \; v_r \leq v < v_{\mathrm{co}} \\ 0, & \text{if } \; v > v_{\mathrm{co}} \end{cases}$$

Parameters	P^r	v_{ci}	v_r	v_{co}
Values	6000 W	3 m/s	12 m/s	20 m/s

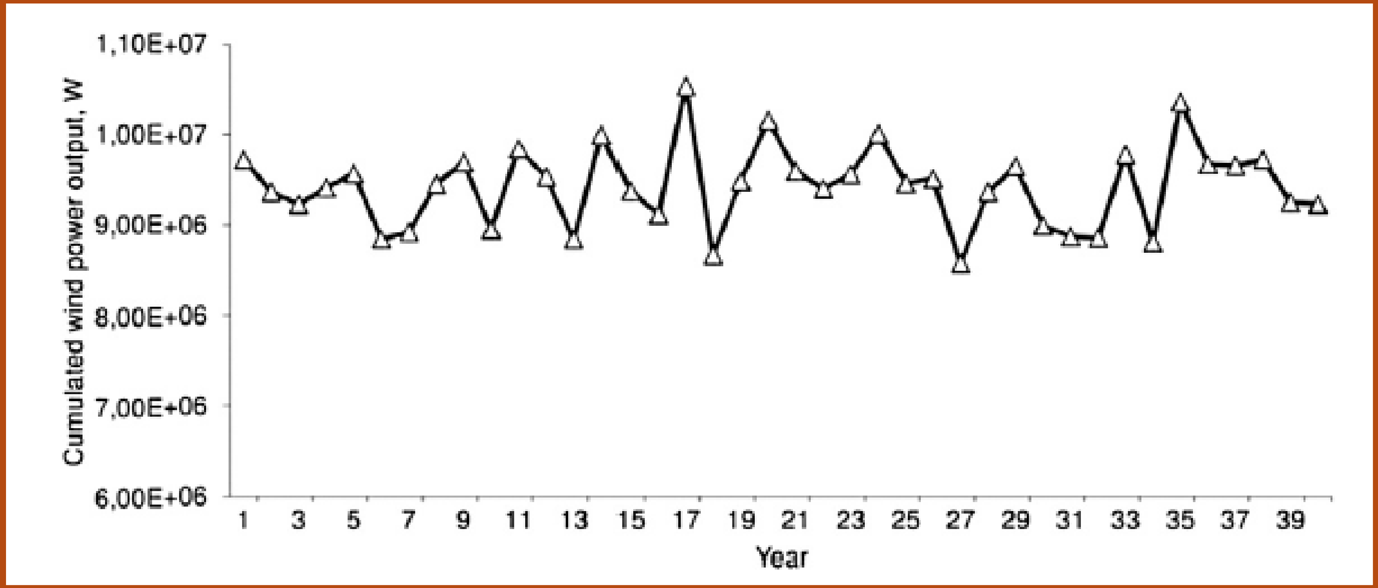
Available wind power output

- Method of eigen-vectors : Numerical Value
- Markov chain model : Analytical Value



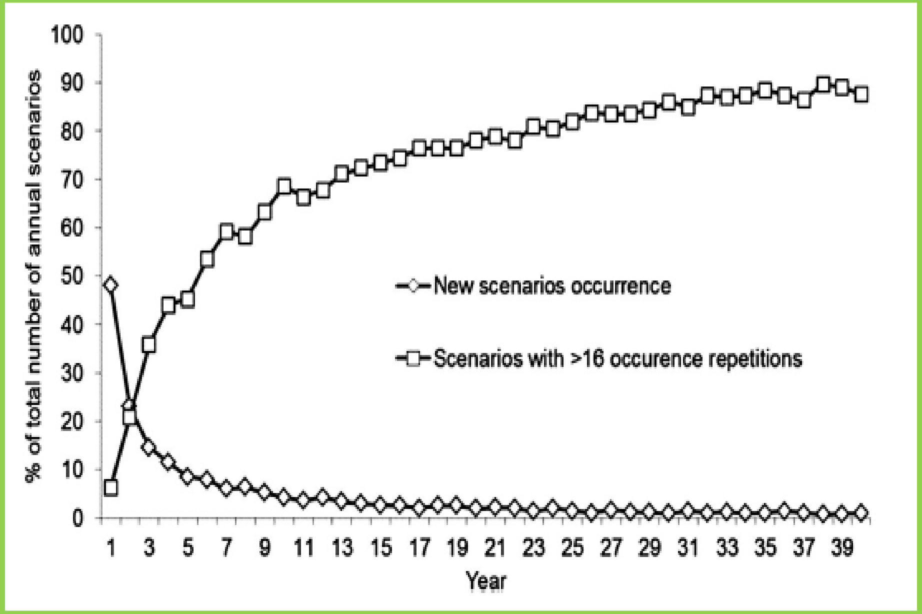
Wind Output for 40 years

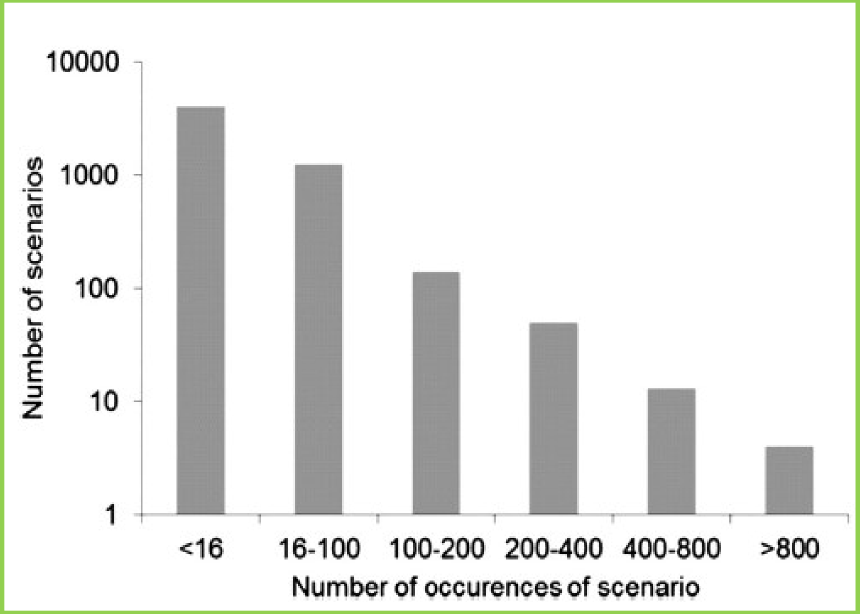
❖ Wind power output was calculated with the Markov chain



Evaluation of the microgrid performance

- The threshold for learning each scenario is set to 16.
- After 10 years of learning, number of new scenarios is less than 1.5% of available scenario at that year.
- Number of learned scenarios are 87% in the year 40
- Still large number of unlearned scenarios





Evaluation of the microgrid performance

- Three indexes for analyzing the performance of reinforcement learning:

$$V_0 = \frac{\sum R_t^{\text{stor,discharge}}}{\sum D_t},$$

$$V_1 = \frac{\sum R_t^{\text{stor,charge}}}{\sum P_t^{\text{wt}}},$$

$$E = \left(\sum D_t - \sum R_t^{\text{stor,discharge}}\right) \cdot P_t$$

- Where P_t is assumed to be constant
- The values all are calculated as a cumulative values in a year

Evaluation of the microgrid performance

- Ns =50 independent simulation runs are executed.
- For each run, wind profile for a year was generated.

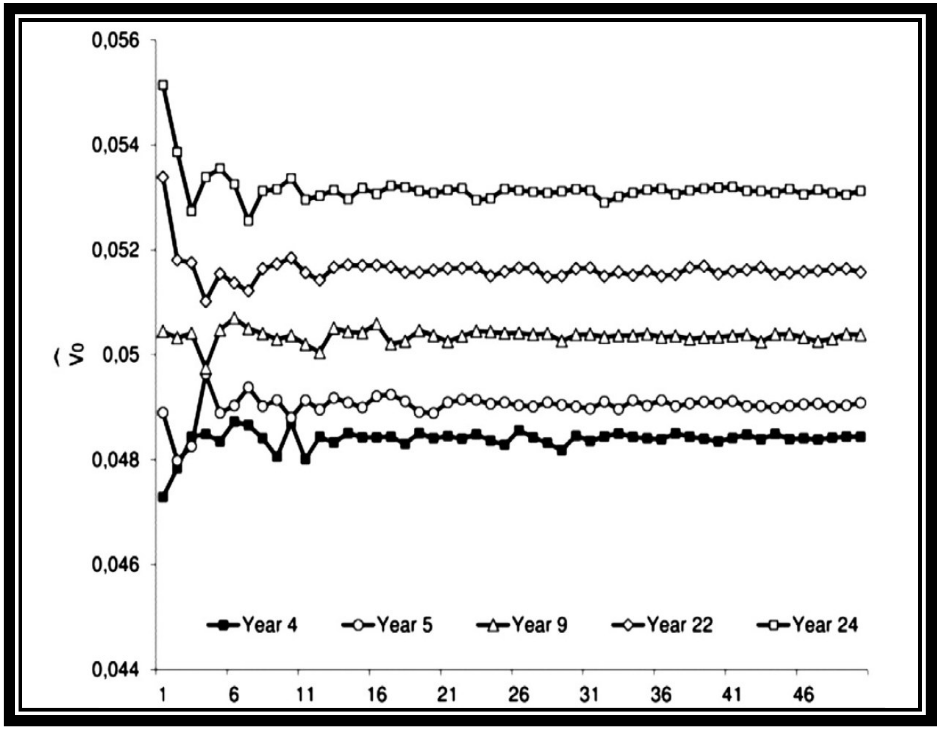
- Through 50 independent simulation runs, they evaluate the estimated V_1 and V_2:

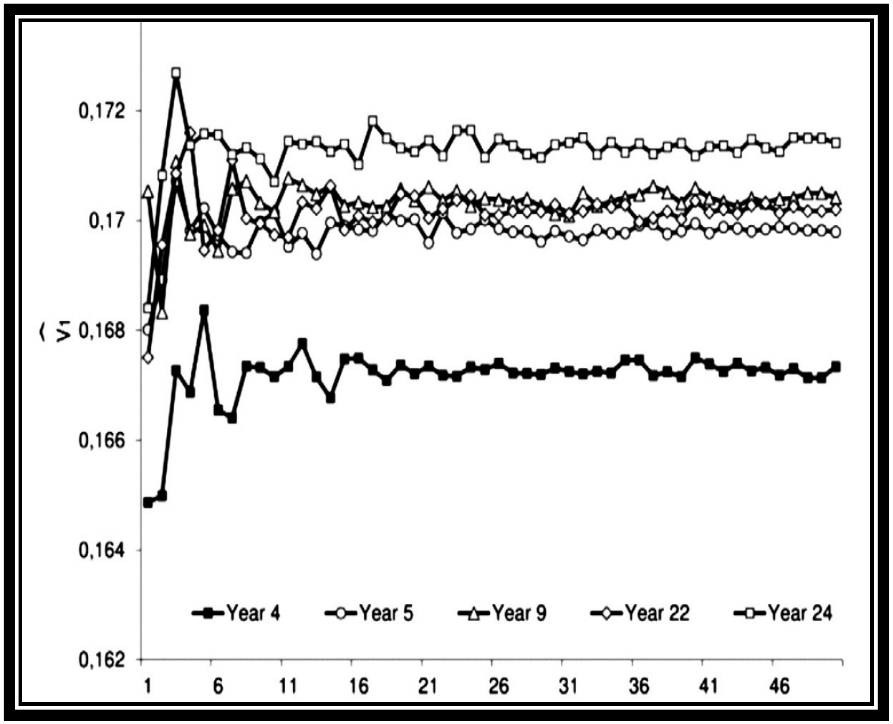
$$\widehat{V}_0 = \frac{\sum_{j=1}^{N_s} V_0^j}{N_s}$$

$$\widehat{V}_1 = \frac{\sum_{j=1}^{N_s} V_1^j}{N_s}$$

Evaluation of the microgrid performance

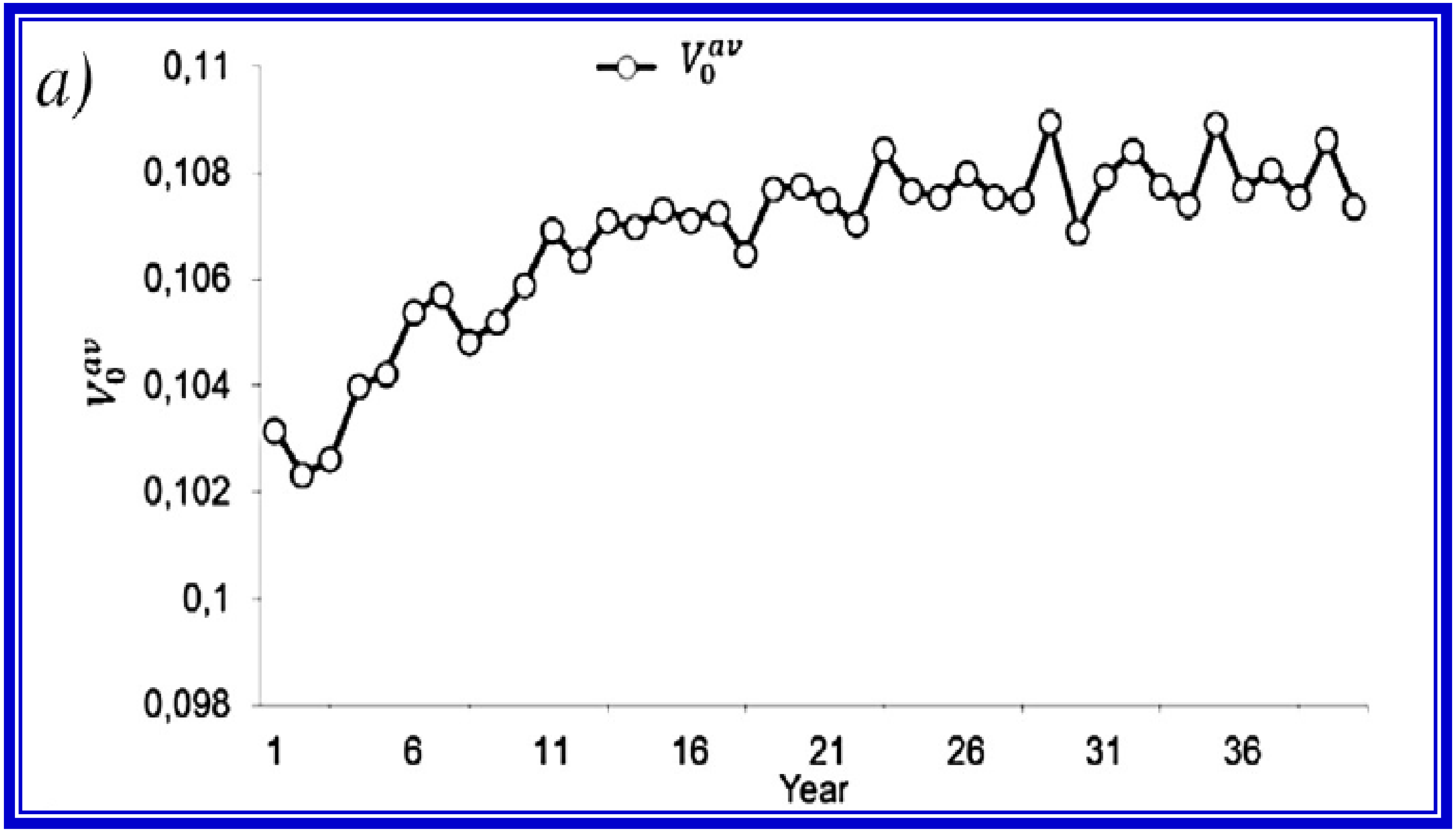
❖ The convergence of V_0 and V_1 for five randomly selected years.





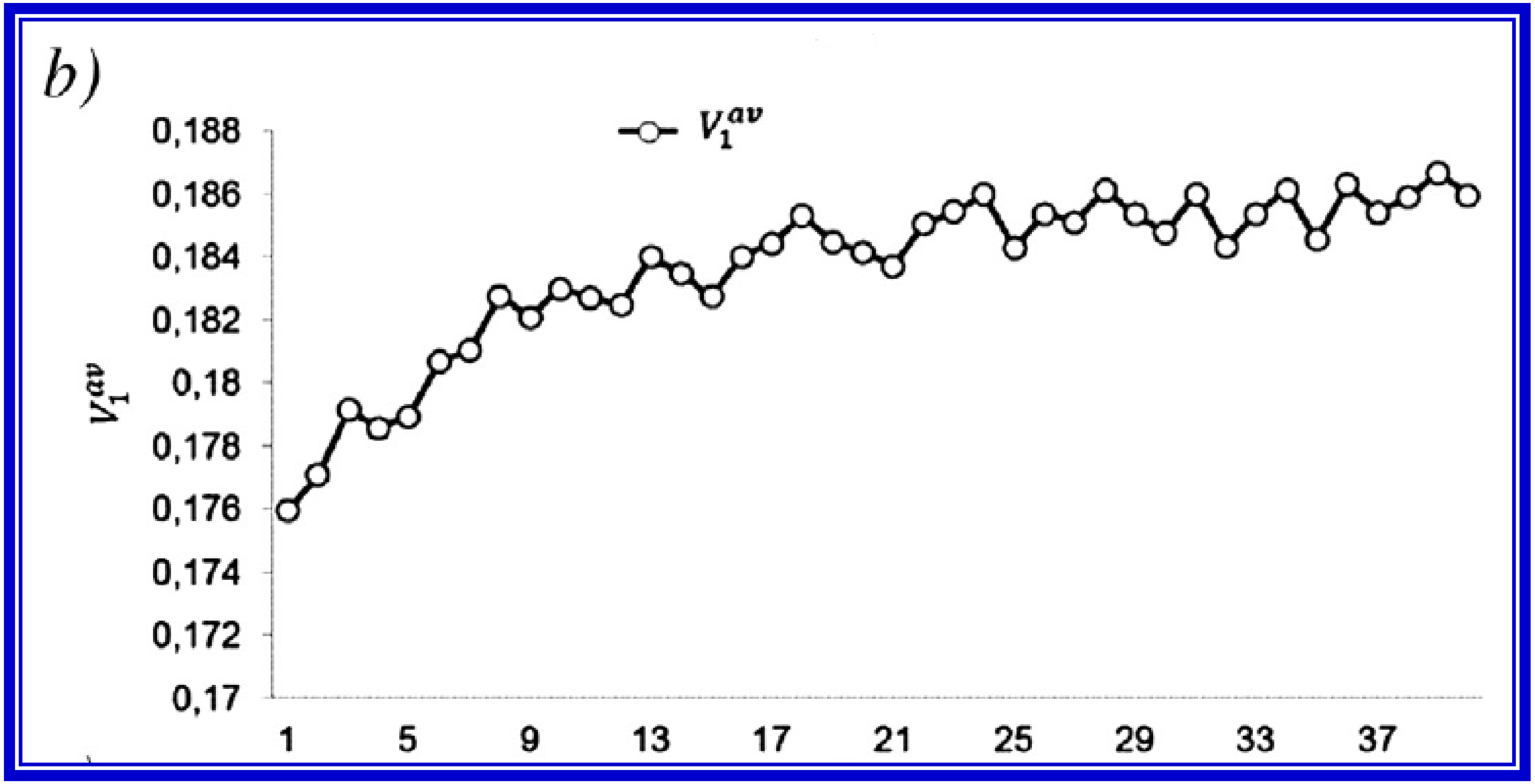
v_0^{av}

- The average values of the performance indicators for each year :
- v_0^{av}, v_1^{av}, and E^{av}
- Performance indicator v_0^{av} increases



v_1^{av}

❖ Performance indicator v_1^{av} increases



E^{av}

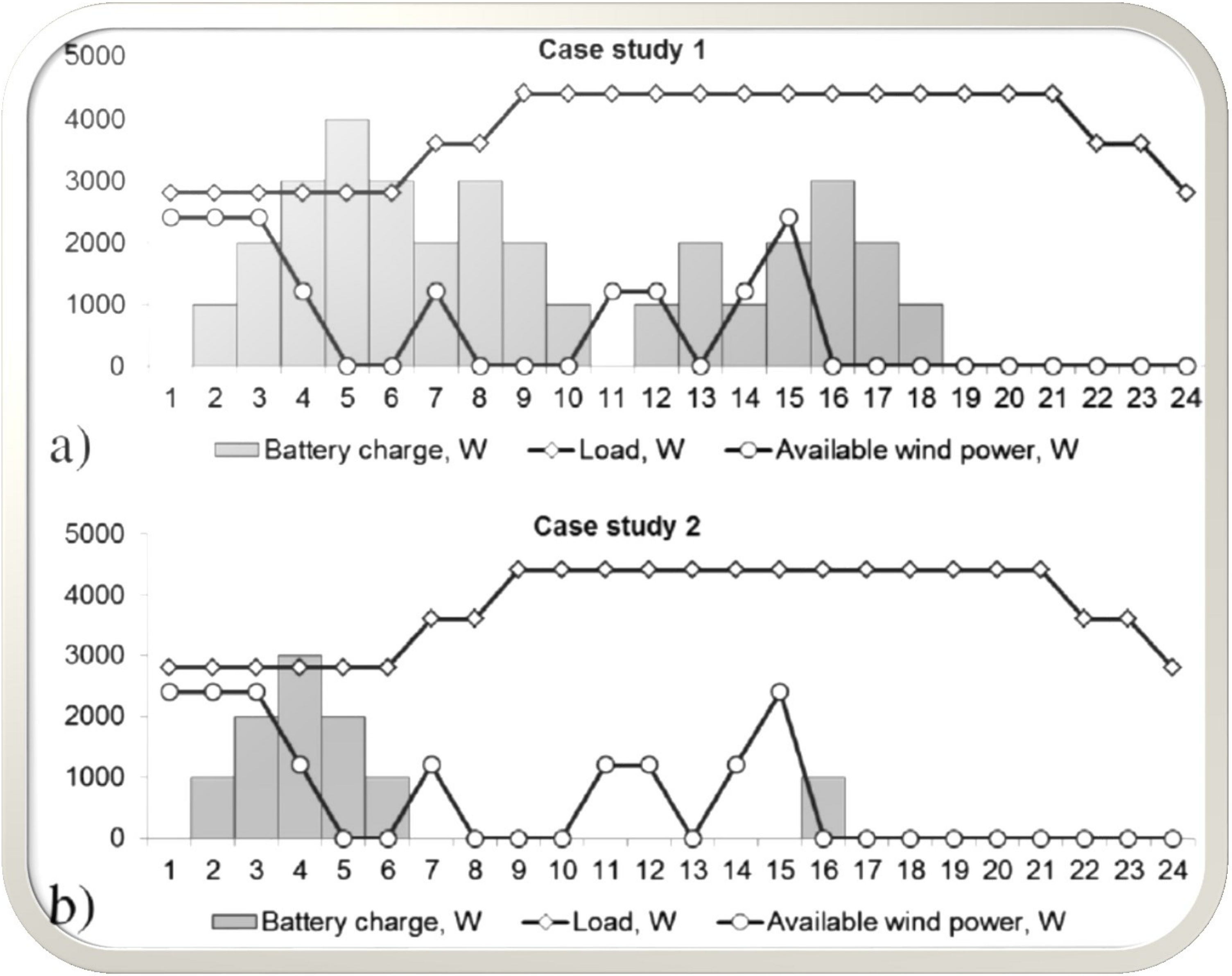
- Progressive decrease of the E^{av}

Case study for k

❖ It is more valuable for the consumer to adopt the strategy illustrated by the case study 1 with weight coefficient k = 6

		Case study 1. $k = 6$	Case study 2. $k = 2^{1200/P_t^{wt}}$
Average improvement	V_0	3.93%	2.72%
of performance indicators	V_1	5.37%	0.96%
after convergence	E	−0.47%[a]	−0.26%[a]

Battery scheduling process for a day of operation



Overview Day Two

- Sensitivity analysis of learning parameters
- Simulation results and analysis
- **Discussion &Conclusions**

Discussion &Conclusions

- The microgrid energy management is done for the benefit of the consumer, i.e. to maximize her or his personal objectives.

i. The paper used a two step ahead approach for learning and decision making using Q-learning for customers. Therefore, based on the current time and knowledge of system about current scenario, it will get a decision. Therefore system states needed to be learned increase significantly.

ii. It would be nice if we have a comparison between this framework and the regular q-learning.

Discussion &Conclusions

i. I believe, analyzing the sensitivity of α after determining the actions is not consider as a sensitivity analysis.

ii. I believe in this framework the battery charges always. The only case that it discharges is when wind output is zero. (they mentioned, they will choose maximum Q after training.)

- The proposed modelling framework is capable of accounting for generation uncertainty.

i. They needed to talk more about method of eigen-vectors or they could not mention it at all.

Discussion &Conclusions

- The optimization framework of reinforcement learning is analyzed through a sensitivity analysis aimed at understanding the role of the learning parameters.

i. They final chosen value of k is in conflict with the paper sensitivity analysis, which is not proper.

ii. One solution for k is to define a variable k based on their prediction for future wind power.

- For measuring the performance of the learning algorithm, three indicators have been introduced.

i. There is a conflict in the results in Figs 11 and 13.

ii. They need to define v_0^{av}, v_1^{av}, and E^{av} more carefully.

Future Work

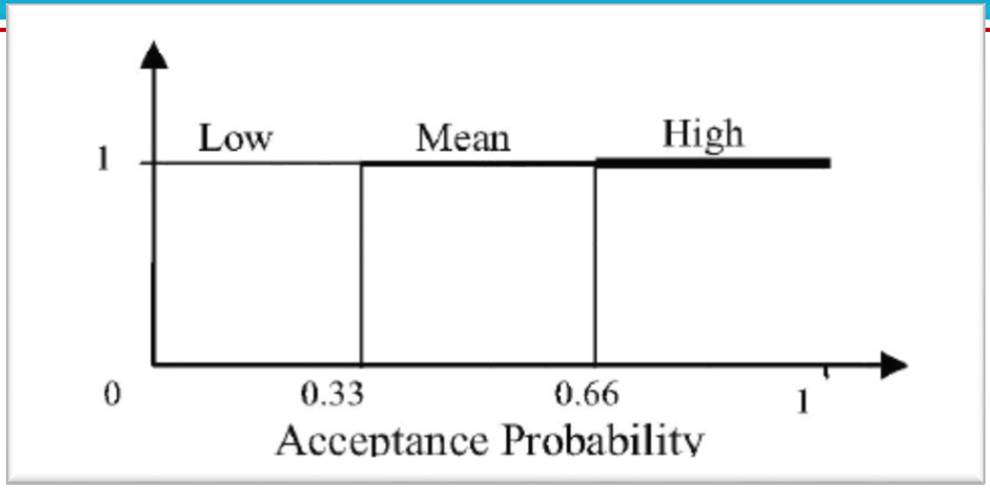
- The improvement for the forecasting and learning capabilities
- The extension to multiple agents integrating, diverse renewable generators, and several intelligent consumers with limited access to information about the power available and limited communication capabilities within the microgrid.

References

1) Kuznetsova, Elizaveta, et al. "Reinforcement learning for microgrid energy management." *Energy* 59 (2013): 133-146.

2) Papaefthymiou, George, and Bernd Klockl. "MCMC for wind power simulation." *IEEE Transactions on Energy Conversion* 23.1 (2008): 234-240.

Risk strategy based on parameters



1. Risk-Averse: Bid low to have HIGH acceptance
2. Risk-Indifferent : They are at MEAN
3. Risk-Taker : Bid high, they have LOW acceptance

- To trade off between exploitation and exploration, the ε - greedy chooses the action with maximum Q-value by the 1- ε probability and selects all possible actions with small probability ε.

Risk strategy based on parameters

- Risk-Averse (RA):
- The agent prefers to be greedy about new data and experience and pick the maximum immediate reward right away without exploring
  - The discounted rate γ : Low value of discounted value since they don't care about future.
  - The learning rate α : High value of learning factor to indicate a greedy feature
  - ε : Low value

Risk strategy based on parameters

- Risk-Taker (RT)
- in a risky situation, it likes to explore more (the high value of ε) to get new opportunities and is not greedy about new data.
  - The discounted rate γ : High value since The expected future reward is valuable for this type of agent.
  - The learning rate α : low value of learning factor to indicate a non-greedy feature
  - ε: High value

Risk strategy based on parameters

- Risk-Indifferent (RI):
- The normal values for the α, γ, ε parameters are suited for this strategy.

Tabular form of parameters and risk

AP: Acceptance probability

Agent	AP	α	γ	ε
RA	High	High	Low	Low
RI	Mean	Mean	Mean	Mean
RT	Low	Low	High	High

References

1) Kuznetsova, Elizaveta, et al. "Reinforcement learning for microgrid energy management." *Energy* 59 (2013): 133-146.
2) Papaefthymiou, George, and Bernd Klockl. "MCMC for wind power simulation." *IEEE Transactions on Energy Conversion* 23.1 (2008): 234-240.
3) Rahimiyan, Morteza, and Habib Rajabi Mashhadi. "Modeling the supplier agent's risk strategy based on fuzzy logic combined with the Q-learning algorithm." 2006 International Conference on Computational Intelligence and Security. Vol. 1. IEEE, 2006.
4) Rahimiyan, Morteza, and Habib Rajabi Mashhadi. "An Adaptive-Learning Algorithm Developed for Agent-Based Computational Modeling of Electricity Market." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 40.5 (2010): 547-556.

Thank you very much