**ISBN** 978-981-14-1684-2
**Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering**
(WCSE 2019 SUMMER)
*Hong Kong, 15-17 June, 2019, pp. 996-1001*
doi: 10.18178/wcse.2019.06.149

# A Deterministic Policy Gradient Based Load Control Policy in Direct Current Distribution Networks

Hong Duan [1], Xu Zhou [1], Xianhong Kang [2+] and Zhongjing Ma [1]

[1] School of Automation, Beijing Institute of Technology, Beijing, P.R.China
[2] Shanxi Information Industry Technology Reasearch Institute Co., Shanxi, P.R.China

**Abstract.** Developing algorithms for global optimum seeking of non-convex optimization problems has special potential in the real world. Previous researches in this field suffer from resulting a local optimum or losing some accuracy by convex relaxation. In this paper, we consider a demand side management (DSM) problem in direct current (DC) distribution networks as an application to study the global optimum seeking of non-convex optimization. Due to the voltage and network constraints, non-convexity appears in the objective function taking into account the tradeoff between the operation costs and users' preferences. By the freedom to express learning problem as a non-convex optimization, we explore a deterministic policy gradient (DPG) based algorithm to calculate the global optimum. A policy network and a polynomial regression critic are built to learn the optimal policy under an exploration noise. Numerical results are provided to demonstrate the DPG algorithm increasing the probability of convergence to the global optimum.

**Keywords:** Distribution networks, Demand-side management (DSM), Deterministic policy gradient, Reinforcement learning

## 1. Introduction

With the development of electronic technology and distributed energy sources, direct current distribution networks have become a popular research area. The power supply reliability can be increased and the transmission losses can be decreased, due to the ability of accommodating distributed energy generation, renewable resources and electric vehicles of distribution networks [1]. If loads are considered as controllable power electronic loads (PELs), they can be adjusted to reduce the peak power and alleviate the fluctuations from distributed generations.

In this paper, we develop the relationship between the PEL loads, bus voltage and the real power of each bus in distribution networks, based on the model in [2]. Then we formulate a demand-side management (DSM) problem as an optimization problem. It is easy to solve this DSM problem because the bus voltage is the main factor that matters and we don’t need to consider the power and the resistance.

However, this DSM problem is a non-convex optimization problem. Previous studies have been focused on using certain optimization methods to solve different types of non-convex problems, such as branch and bound methods [3], particle swarm optimization (PSO) [5] and trust region methods [6]. However, it takes a long time to calculate the results. They also fail to consider the online solution. Reinforcement learning (RL) is a kind of machine learning algorithm aiming at learning the optimal policy from dynamic interaction process with the environment. RL algorithms don't learn from stable databases. In the RL framework，a decision maker can choose action at possible states/scenarios and get reward from the environment. Then, based on the reward, the decision maker will learn which action/policy is better because it wants to maximize/ the cumulated discounted reward. With the development of RL, more and more researchers are

+ Corresponding author. Tel.: + 86-18811409037.
*E-mail address*: duanhong_bit@foxmail.com.

focusing on using RL to solve optimization problems, such as storage system management in smart-grids [7], on-line demand response of smart buildings [8] and non-convex economic dispatch problems [9].

In this paper, the DSM problem requires an optimal control policy of voltage for each PEL bus, which is a continuous variable. Thus, we focus on a RL algorithm called deterministic policy gradient (DPG) [10] and its latest development combined with deep neural networks (DDPG) [11]. The reason why DPG is chosen to solve our proposed DSM problem are as follows.

- The deterministic policy is compatible with the voltage control in distribution network.
- The state space and action space of DSM problem are continuous.

The rest of the paper is organized as follows. In section 2, we develop a voltage model of PEL units and formulate the DSM optimization problem. In section 3, we propose a DPG based optimal algorithm and use it to solve the DSM problem. Section 4 gives the numerical results. Conclusions and future works are given in Section 5.

## 2. Review of Formulation of Demand-side Optimization Problem

The model in this work comes from our previous work [12]. For purpose of demonstration, we give the model below.

We consider a distribution network that contains a group of DC buses denoted by $\mathcal{N}$. Each bus can connect a combination of fixed load or PELs. As for PELs, we regard PELs as adjustable resistances connected with the bus through a convertor. And they can be controlled to meet the objective. We denote the load buses as $\mathcal{L}$, non-load buses as $\mathcal{B}$, fixed load buses as $\mathcal{F}$, PEL buses as $\mathcal{P}$, buses with distributed generations as $\mathcal{G}$. For bus $i \in \mathcal{N}$, $\mathcal{N}_i$ denotes neighbouring buses of bus $i$.

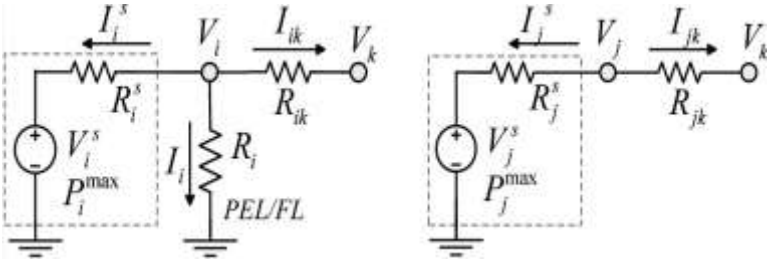



Fig. 1: An illustration of Kirchhoff's law, load bus and non-load bus.

As shown in Fig. 1, based on kirchhoff's current law, the relationship between bus voltage and load is:

$$P_p = \frac{V_p^2}{R_p} = V_p\Big(\sum_{k\in\mathcal{P}} c_{pk}V_k + d_p\Big), \quad \forall p \in \mathcal{P} \tag{1}$$

where $V_n$ represent the voltage of bus $n$, $P_p$ represent the power of bus $p$, $R_p$ is the load resistance, $c_{pk}$ and $d_p$ are coefficients. The object is to minimize the preference of each node and transmission loss:

$$J(V) \triangleq \sum_{p\in\mathcal{P}} \xi_p\Big(V_p\Big(\sum_{k\in\mathcal{P}} c_{pk}V_k + d_p\Big) - P_{p,ref}\Big)^2 + \rho \sum_{k\in\mathcal{N}}\sum_{l\in\mathcal{N}_k} (V_k - V_l)^2 / R_{kl} \tag{2}$$

where $P_{p,ref}$ is the expected power level of PEL $p$, $\xi_p$ is positive parameter represent the buses' tolerance due to load control, $\rho$ is a positive constant and $R_{kl}$ denotes resistance between bus $k$ and bus $l$.

As for constraints, we consider the power sources capacities, current constraints of feeder line and bounds of PEL by (3)-(5), respectively.

$$V_p^s - \frac{P_p^{\max}R_p^s}{V_p^s} \le V_p \le \max_{k\in\mathcal{G}}\{V_k^s\}, \quad \forall p \in \mathcal{P} \text{ and } \sum_{k\in\mathcal{P}} a_{jk}V_k \ge V_j^s - \frac{P_j^{\max}R_j^s}{V_j^s} - b_j, \quad \forall j \in \mathcal{B}\cup\mathcal{F} \tag{3}$$

$$V_j = \sum_{k\in\mathcal{P}} a_{jk}V_k + b_j \le I_j^{\max}R_j, \ \forall j \in \mathcal{F}, \text{ and } \frac{V_p}{R_p} = \sum_{k\in\mathcal{P}} c_{pk}V_k + d_p \le I_p^{\max}, \ \forall p \in \mathcal{P} \tag{4}$$

$$\sum_{k\in\mathcal{P}} \mathfrak{m}_{pk} V_k \geq R_p^{\min} d_p \text{ and } \sum_{k\in\mathcal{P}} \mathfrak{n}_{pk} V_k \leq R_p^{\max} d_p, \tag{5}$$

The DSM problem is:

$$V^* = \arg\min_V J(V) \tag{6}$$

Subject to: (3)-(5)

## 3. DPG based Algorithm for DSM Problem

Deterministic Policy Gradient (DPG) method is the limiting case of stochastic policy gradient method [13] as the policy variance of the stochastic policy tends to zero. Thus, it is capable to solve real physical control problems. The DPG algorithm usually has an Actor-Critic (AC) [14].

In this paper, we formulate the DSM problem as an one-step decision making problem, which is similar to the continuous bandit problem [15]. We consider state $s \in \mathcal{S}$, and $s = [P_{1,ref}, P_{2,ref}, \cdots, P_{p,ref}]$ consists of the power reference $P_{ref}$ of each PEL bus. We consider action $a \in \mathcal{A}$, and $a = \mu_\theta(s) = [V_1, V_1, \cdots, V_p]$ consists of the continuous voltage of each PEL bus. As for reward $r(s,a)$, which is equal to the value function $Q(s,a)$ in one step problem, We formulate it as the system cost (objective function) by (2):

$$J(V) = J(\mu_\theta) = Q^{\mu_\theta}(s,a). \tag{7}$$

In our study, the main objective is to gain the optimal voltage control policy for minimizing the system cost. The system constraints are used to check whether the action we take is rational. .

As for the actor (policy network), the target policy is a deterministic policy $\mu_\theta(s)$ with the parameter vector $\theta$ .However, the DPG method may lead to a suboptimum easily due to the lack of efficient exploration. To arise exploration, we propose a novel exploration noise based on the Ornstein-Uhlenbeck noise [11]:

$$\Omega_{t+1} = (|\Omega_t| + |\eta(\phi - \Omega_t)| + \sigma W_t) * \gamma \tag{8}$$

where $t$ denotes time, and $\Omega$ is the random noise which is related to the time $t$. $\phi$ represents the expectation of stochastic process, and $W_t$ is a random variable. $\eta$ and $\sigma$ are parameters related to the shape of noise. $\gamma$ is a random variable selected from $\{-1,1\}$ .As shown in Fig. 2, the novel noise changes with time $t$ and finally converges to the expectation value $\phi$ .

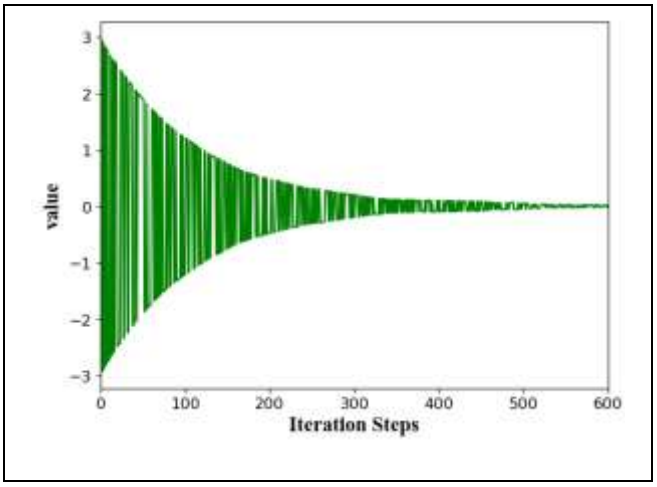



Fig. 2. Novel O-U noise with one-dimension.

As for the critic, considering the inaccuracy of model coefficient and computational complexity with high dimensions, we would like to build a real simulation system and sample the real voltage and power data. Then we can approximate the true $Q^\mu(s,a)$ such that $Q^\omega(s,a) = Q^\mu(s,a)$ by using polynomial regression or deep neural network. If the dimension is not high, we can obtain the action-value function by using (2) directly as mentioned. The gradient of objective function is as below:

$$\nabla_\theta J(V) = \nabla_\theta \mu_\theta(s) \nabla_a Q^\mu(s,a)|_{a=\mu_\theta(s)} = \nabla_\theta \mu_\theta(s) \nabla_a Q^\omega(s,a)|_{a=\mu_\theta(s)} \tag{9}$$

Based on the gradient of objective function, the parameters of policy can be updated by:

$$\theta_{t+1} = \theta_t - \alpha_\theta \nabla_\theta \mu_\theta(s_t) \nabla_a Q^\omega(s_t, a_t)|_{a=\mu_\theta(s)} \tag{10}$$

Where $\alpha_\theta$ is the learning rate and $\alpha_\theta > 0$.

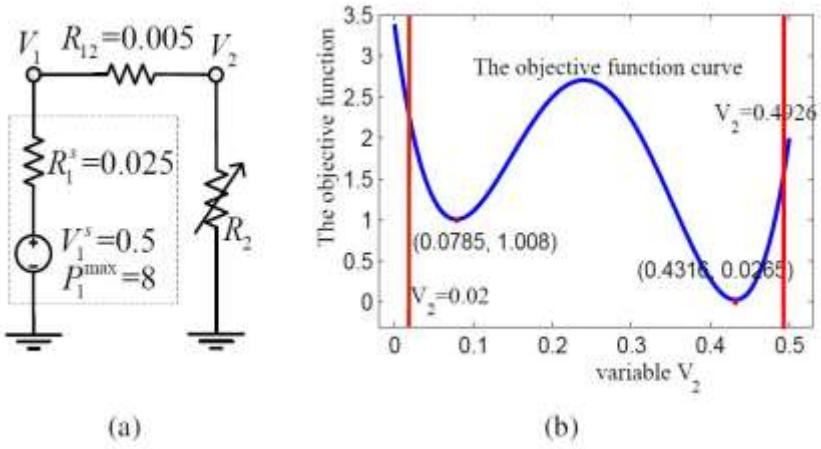



Fig. 3. (a) Distribution network with 2 buses. (b) The non-convex objective function curve.

To solve the DSM optimization problem, we design an algorithm expressed as Algorithm 1. In each episode, we firstly initialize the noise $\Omega$ and parameter vector $\theta$ randomly. And in each iteration step $t$ of episode $m$, we calculate the policy based on the parameter vector $\theta$ and add the noise to the policy. Then we obtain the action-value function $Q^\omega(s,a)$. Based on the gradient of objective function in (9), we update the $\theta$ by (10). The process is repeated until the change in policy $\mu_\theta^m$ is negligible. Then, we store this policy $\mu_\theta^m$. Finally, we compare different $\mu_\theta^m$ and find out the best policy from them.

**Algorithm 1** DPG Based Algorithm for DSM Problem

**Require:**

Initialize the distribution networks parameters $V_i^s$, $P_{i,ref}$, $\xi_p$, $\rho$, $R_0$ ;

Initialize the maximum iteration episode $M$, and the maximum iteration step $T$ ;

Initialize the state s=$[P_{1,ref}, P_{2,ref}, \cdots, P_{p,ref}]\top$ ;

**Ensure:**

PEL bus voltage $V^*$ ;

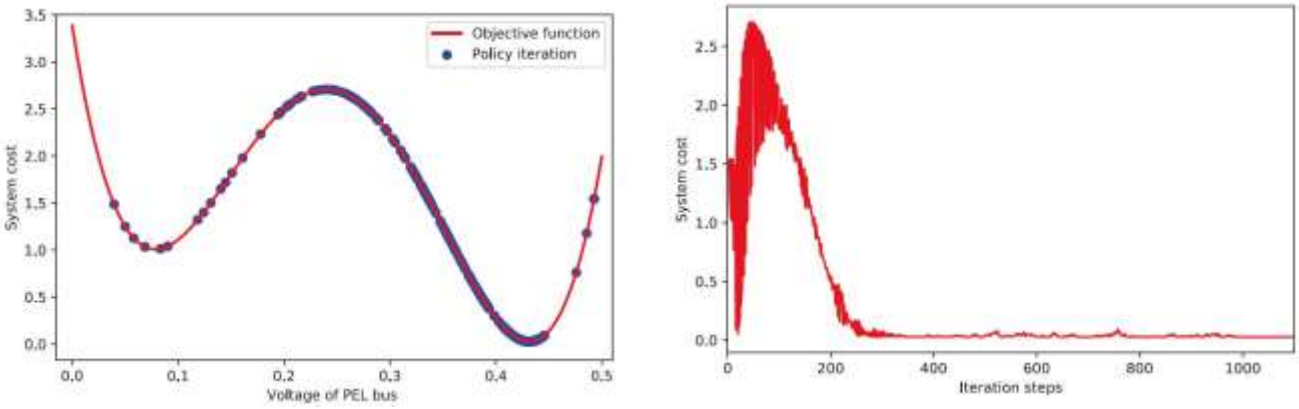
1: Calculate the feasible set of action $a$ by constraints (3)-(5);
2: **for** $m=1$ **to** $M$ do
3: Initialize a random process $\Omega$ for action exploration;
4: Randomly initialize the policy $\mu_\theta(s)=[V_1, V_2, \cdots, V_p]\top$ with parameter $\theta$ ;
5: **for** $t=1$ **to** $T$ do
6: Take action at state $s$ using the behavior policy with noise updated by (8)
7: Obtain the action-value function by equation (2);
8: Calculate the gradient of cost by (9);
9: Update parameter $\theta$ by (10);
10: **end for**
11: Store the policy $\mu_\theta^m$ of episode $m$;
12: **end for**
13: Select the best policy from all the $\mu_\theta^m$

## 4. Numerical Result

We demonstrate the performance of Algorithm 1 in a distribution network, as shown in Fig. 3(a). The distribution network has two buses.

Bus 1 is power source bus, while bus 2 is a PEL bus. The voltage of bus 2 is denoted by $V_2$. We set the initial resistance $R_2 = 2$ , $P_{2,ref} = 1$, $\rho = 1$ and $\xi_2 = 2$. Due to the constraint of power source capacities and the bounds of PEL resistance. The feasible set of bus voltage is: $0.02 \le V_2 \le 0.4926$. As shown in Fig. 3(b), the objective function has a four-order non-convex form:

$$J(V) = 2(-\frac{100}{3}V_2^2 + \frac{50}{3}V_2 - 1)^2 + 200(-\frac{1}{6}V_2 + \frac{1}{12})^2. \tag{11}$$




(a) (b)

Fig. 4: (a) The iteration of optimal policy. (b) The iteration of system cost.

As shown in Fig. 4(a), discrete blue points represent the policy iteration in the learning process and the red curve represents the objective function of the network. As we can see, the policy reaches the suboptimal point (0.0785, 1.008) at first and jumps out of the suboptimal solution. Finally, the policy converges to the global optimum (0.4316, 0.0265). Fig. 4(b) shows the relationship between the objective function (system cost) and iteration steps. The objective function converges to the optimal value 0.0265 finally.

## 5. Conclusion

We model the relationship between the PEL resistance and the bus voltage in distribution networks. Then we formulate the DSM problem as a four-order non-convex optimization problem, which determines an optimal bus voltage policy to regulate the PEL and minimizes the losses and deviation cost of system. However, this optimization problem is non-convex and hard to find global optimum. Hence, we propose a DPG based algorithm to solve this problem. In the DPG framework, the actor is a policy neural network with parameter $\theta$ . The critic is estimated by polynomial regression from sampled action-value points. To increase the probability of convergence to the global optimum, we add noise to the policy when it interacts with the environment. The simulation results show that our algorithm can converge to the global optimum.

As future works, we would like to use this algorithm to solve high-dimensional non-convex optimization problem in distribution networks, and discuss the computational complexity and scalability of this algorithm.

## 6. Acknowledgements

This work is supported by National Natural Science Foundation (NNSF) of China under Grant 61873303.

## 7. References

[1] Justo J J, Mwasilu F, Lee J, et al. AC-microgrids versus DC-microgrids with distributed energy resources: A review. Renewable and sustainable energy reviews, 2013, 24: 387-405.

[2] Weaver Wayne W. Dynamic energy resource control of power electronics in local area power networks. IEEE Transactions on Power Electronics, 2011, 26(3): 852-859.

[3] Rider M J, Garcia A V, Romero R. Transmission system expansion planning by a branch-and-bound algorithm. IET generation, transmission & distribution, 2008, 2(1): 90-99.

[4] Wang M, Xu F, Xu C. A branch-and-bound algorithm embedded with DCA for DC programming. Mathematical Problems in Engineering, 2012, 2012.

[5] Sortomme E, El-Sharkawi M A. Optimal power flow for a system of micro-grids with controllable loads and battery storage. 2009 IEEE/PES Power Systems Conference and Exposition. IEEE, 2009: 1-5.

[6] Erway J B, Plemmons R J, Adhikari L, et al. Trust-region methods for nonconvex sparse recovery optimization. 2016 International Symposium on Information Theory and Its Applications (ISITA). IEEE, 2016: 275-279.

[7] Kuznetsova E, Li Y F, Ruiz C, et al. Reinforcement learning for micro-grid energy management. Energy, 2013, 59: 133-146.

[8] Mocanu E, Mocanu D C, Nguyen P H, et al. On-line building energy optimization using deep reinforcement learning. IEEE Transactions on Smart Grid, 2018.

[9] Abouheaf M I, Haesaert S, Lee W J, et al. Approximate and reinforcement learning techniques to solve non-convex economic dispatch problems. 2014 IEEE 11th International Multi-Conference on Systems, Signals & Devices (SSD14). IEEE, 2014: 1-8.

[10] Silver D, Lever G, Heess N, et al. Deterministic policy gradient algorithms. ICML. 2014.

[11] Lillicrap T P , Hunt J J , Pritzel A , et al. Continuous control with deep reinforcement learning. Computer Science, 2015, 8(6):A187.

[12] Zou S, Ma Z, Liu S. Load Control Problems in Direct Current Distribution Networks: Optimality, Equilibrium of Games. IEEE Transactions on Control Systems Technology, 2018, PP(99): 1-14.

[13] Sutton R S, McAllester D A, Singh S P, et al. Policy gradient methods for reinforcement learning with function approximation. Advances in neural information processing systems. 2000: 1057-1063.

[14] Degris T, Pilarski P M, Sutton R S. Model-free reinforcement learning with continuous action in practice. 2012 American Control Conference (ACC). IEEE, 2012: 2177-2182.

[15] Trovo F, Paladino S, Restelli M, et al. Budgeted multi–armed bandit in continuous action space. Proceedings of the Twenty-second European Conference on Artificial Intelligence. IOS Press, 2016: 560-568.