투 스위치 포워드 컨버터의 성능 및 안정성 향상을 위한 심층강화학습 제어

지상혁, 권상협, 배성우†
한양대학교

Deep Reinforcement Learning Control for Performance and Stability Enhancement of Two-Switch Forward Converter

Sanghyuk Ji, Sanghyeob Kwon, Sungwoo Bae†
Hanyang University

ABSTRACT

본 논문은 투 스위치 포워드 컨버터의 성능 및 안정성 향상을 위한 심층강화학습 제어기법을 제안한다. 컨버터의 폐루프 제어 시스템은 컨버터의 성능과 안정성을 보장하기 위해 필수적이다. 하지만, 전력변환장치는 비선형성을 가지고 있어 모델기반 제어 시스템 적용 시 특정 동작점을 벗어나면 컨버터의 성능과 안정성이 저하될 수 있다. 심층강화학습 방법 중 하나인 Twin Delayed Deep Deterministic policy gradient algorithm은 전력변환장치의 모델을 요구하지 않으며 전력변환장치의 넓은 동작 조건에서도 성능 및 안정성을 보장할 수 있다. 제안한 심층강화학습 제어는 Matlab/Simulink 환경에서 PI 제어 기법과 비교 검증된다.

1. 서 론

전력변환장치의 제어 방식은 모델기반 제어가 일반적으로 사용되고 있다. 모델기반 제어는 정밀한 모델링하더라도 오차가 발생하게 되므로 비선형과 모델링되지 않는 요소로 인해 컨버터의 안정성과 성능이 저하될 수 있다. 만약, 추가적으로 고려해야 될 요소가 있으면 처음부터 시스템을 다시 모델링해야 하는 단점이 있다. 또, 제어 파라미터 설정에 따라 느린 동적 응답을 가질 수 있고 동작 조건의 변화에 대해 안정성이 저하될 수 있다[1]. Proportional-Integral (PI)제어는 시행착오, 지글러-니콜스와 같은 설계 방법이 있으나 제어기가 복잡하여 제어 파라미터가 많아지면 설계하는데 많은 시간이 소요되고 최적화하기 어렵다[2].

심층강화학습 알고리즘 중 Twin Delayed Deep Deterministic policy gradient algorithm (TD3)는 Critic을 두 개 활용하여 과추정문제를 방지하므로 전력변환장치의 성능과 안정성을 향상시킬 수 있다. TD3를 통해 자동으로 최적의 제어 파라미터를 학습하여 시행착오나 복잡한 튜닝 과정이 필요없다. 또, Model-free 방식으로 모델링에 대한 전문지식이 요구되지 않으며 전력변환장치의 측정 값을 기반으로 최적 제어를 도출할 수 있다. TD3는 여러 동작 조건을 학습하여 PI 제어보다 넓은 동작 범위에서도 뛰어난 성능과 안정성을 보장한다. 본 논문에서는 넓은 동작 조건에서도 성능과 안정성을 보장할 수 있는 TD3를 투 스위치 포워드 컨버터에 적용한다.

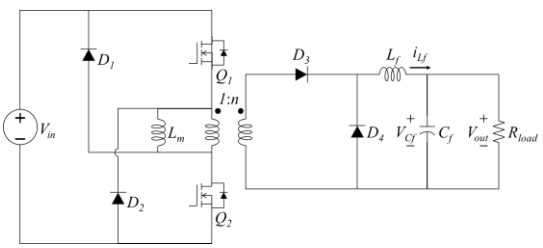

그림.1 투 스위치 포워드 컨버터 회로

2. 본론

2.1 투 스위치 포워드 컨버터

투 스위치 포워드 컨버터는 벅 컨버터에 변압기를 추가해 절연과 추가적인 전압 이득을 얻은 컨버터이며, 회로도는 그림1과 같다. 투 스위치 포워드 컨버터의 동작 모드는 다음과 같다. 스위치 Q1과 Q2 두 개의 스위치가 동시에 턴-온, 턴-오프가 된다. 턴-온 시 변압기 1차 측에서 2차 측으로 에너지가 전달되고 변압기는 자화된다. 턴-오프 시 변압기의 자화 에너지는 입력 측으로 회생되고 2차 측에 인덕터는 부하 측으로 에너지를 방출한다. 투 스위치 포워드 컨버터의 상태 공간 평균 모델은 아래와 수식과 같다.

$$\begin{bmatrix} \frac{di_{L_f}}{dt} \\ \frac{dv_{C_f}}{dt} \end{bmatrix} = \begin{bmatrix} 0 & -\frac{1}{L_f} \\ \frac{1}{C_f} & -\frac{1}{RC_f} \end{bmatrix} \begin{bmatrix} i_{L_f} \\ v_{C_f} \end{bmatrix} + \begin{bmatrix} \frac{nd}{L_f} \\ 0 \end{bmatrix} V_{in} \quad (1)$$

$$V_{out} = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} i_{L_f} \\ V_{C_f} \end{bmatrix} \quad (2)$$

여기서, i_{Lf}는 인덕터에 흐르는 전류, L_f는 필터 인덕턴스, V_{cf}는 필터 커패시터 전압, C_f는 필터 커패시턴스, n은 권선 비, d는 듀티비, V_{in}은 입력전압, V_{out}은 출력 전압이다.

2.2 강화학습

2.2.1 Markov Decision Process

강화학습은 Sequential decision making에 대한 프레임워크를 갖는 Markov Decision Process(MDP)의

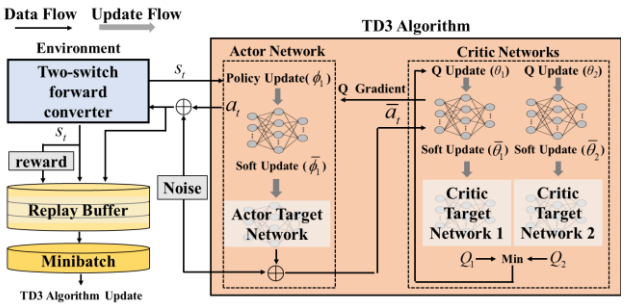

그림.2 Twin Delayed Deep Deterministic policy gradient algorithm의 구조

배경을 가진다.

MDP설정에서, 에이전트는 환경과 상호작용을 하면서 누적기대 보상이 최대가 되는 최적 정책(ϕ*)을 도출할 수 있다. MDP는 총 4개의 튜플이며 S, A, P, R로 표현된다. S는 상태, A는 행동, P는 상태 전이 확률, R은 보상을 의미한다. 예를 들어, 시점 t에서 환경으로부터 에이전트가 필요한 정보인 상태 (st)를 전달받으며, 에이전트는 정책에 따라 행동(at)을 도출한다. at가 환경에 적용될 때, P에 의해 다음 상태(st+1)가 결정되며, 에이전트는 다음 상태에 따라 보상을 받게 된다. MDP 조건에서 최적 정책은 에이전트가 받는 기대 누적 보상(Rt)이 최대가 되는 정책을 의미하며, 다음과 같이 표현된다.

$$\phi^* = \arg\max V^{\phi}(s_t) \tag{3}$$

$$V^{\phi}(s_t) = E[R_t \mid s_t] \tag{4}$$

$$R_t = \sum_{k=0}^{\infty} \gamma^k r(s_{t+k}, a_{t+k}) \tag{5}$$

여기서 V 는 상태 가치 함수로 s_t 부터 기대되는 누적 보상을 의미하며, 누적 보상은 시점 t 부터 매시간 간격마다 할인율 (γ)을 곱한 값의 총 합을 의미한다. r은 보상이다.

Q-함수는 State-Action Value Function이라 불리며, Model-free 강화학습의 근본이다. Q-함수는 s_t , a_t 상황에서 할인된 누적 보상의 기대 값을 의미하며 다음과 같이 정의된다.

$$Q^{\phi}(s_t, a_t) = E[R_t \mid s_t, a_t] \tag{6}$$

최적 $V(s_t)$는 s_t에서 총 예상 할인 보상의 최대치이므로 모든 가능한 행동에서 Q(s_t, a_t)의 최대치가 될 것이며, 다음과 같이 정의된다.

$$V^*(s_t) = \max_{a_t \in A} Q^*(s_t, a_t) \tag{7}$$

이때, 에이전트가 가능한 행동 중에서 최대 Q-값인 최적의 행동을 도출할 것이며, 수식으로 나타내면 다음과 같이 표현된다.

$$a_t^* = \arg\max_{a_t \in A} Q^*(s_t, a_t) \tag{8}$$

2.2.2 Twin Delayed Deep Deterministic policy gradient algorithm

그림 2는 TD3 알고리즘의 구조이다. TD3 알고리즘은 Actor-Critic으로 구성되며 Approximate dynamic programming 방법이다. 인공지능 기반 Actor-Critic은 최적정책을 도출하는 정책 그래디언트 접근 방식이며, Function approximator를 인공신경망으로 설계하여 연속 상태 및 행동 공간 다룰 수 있다. Deep Deterministic Policy Gradient와 같은 알고리즘은 Actor를 평가하는 Crtic이 오직 하나만 존재하여 과추정상태의 문제가 발생할 수 있다. TD3 알고리즘은 Critic을 두 개 활용하므로 과추정문제를 방지하면서 Actor의 최적정책 도출 성능을 높일 수 있다[3].

Actor-critic 접근방식에서 Critic은 Actor의 현재 정책을 평가한다. 이 때, Actor는 Critic에서 추정한 가치함수를 기반으로 정책 업데이트한다. Actor의 최적 정책은 환경의 s_t로부터 기대 보상을 최대화하는 행동을 도출한다.

TD3 에이전트의 행동은 Actor 네트워크의 정책과 노이즈의 함수로 다음과 같이 표현될 수 있다.

$$a_t = clip\left(\phi(s_t \mid \theta^{\phi}) + \varepsilon_t, a_{min}, a_{max}\right) \tag{9}$$

여기서, θ는 Gaussian 노이즈, a_{max}와 a_{min}은 각각 에이전트의 제어 범위의 최대와 최소값이다. 행동은 환경에 적용되며, s_{t+1}이 도출되며 다음 보상(r_{t+1})을 계산할 수 있다. 각 시간 간격마다 버퍼에 튜플 (s_t, a_t, r_{t+1}, s_{t+1})이 저장된다. 충분한 튜플이 버퍼에 쌓였다면, 버퍼에서 미니 배치 사이즈만큼 랜덤으로 추출하여 정책을 업데이트한다. TD3 정책은 두 개의 Critic 신경망 활용하여 이 중 Q-값의 최솟값을 선택하여 Bellman 방정식으로 다음과 같이 계산할 수 있다.

$$y(r_{k+1}, s_{k+1}) = r_{k+1} + \gamma \cdot \min_{i=1,2} \bar{Q}_i\left(s_{k+1}, \bar{\phi}\left(s_{k+1} \mid \theta^{\bar{\phi}}\right)\right) \tag{10}$$

여기서, $\bar{Q}$ 는 Critic의 타겟 네트워크이다. Q-정책을 계산이 완료된 후에 평균 제곱 오차 손실 함수를 활용하여 Critic 네트워크를 업데이트한다.

$$L(\theta_i) = \frac{1}{|B|} \sum_{B_k \in B} \left(Q_{\theta_i}(s_k, a_k) - y(r_{k+1}, s_{k+1})\right)^2, i = 1,2 \tag{11}$$

여기서, $L(\theta_i)$ 는 Critic 네트워크의 손실 함수이며 Critic 네트워크가 추정한 Q-값과 목표 Q-값 간의 차이를 최소화하도록 한다.

$$\theta_{k+1}^{Q} = \theta_k^{Q} - \eta_Q \nabla_{\theta^Q} \mathcal{L}(\theta^Q) \tag{12}$$

Actor network는 누적보상을 최대화하기 위해 최적정책을 도출할 수 있도록 정책 그래디언트를 기반으로 파라미터를 업데이트하며 다음과 같이 표현된다.

$$\theta_{k+1}^{\phi} = \theta_k^{\phi} + \eta_{\phi} \nabla_{\theta^{\phi}} \mathcal{J}(\theta^{\phi}) \tag{13}$$

여기서, 정책 그래디언트는 다음과 같이 표현된다.

$$\nabla_{\theta^{\phi}} J\left(\theta^{\phi}\right) \approx \frac{1}{N} \sum \left[\nabla_a Q\left(s_i, a_i \mid \theta^{Q}\right) \nabla_{\theta^{\phi}} \phi(s_i \mid \theta^{\phi})\right] \tag{14}$$

Critic와 Actor 네트워크의 파라미터가 업데이트된 후, 각 타겟 네트워크는 추정과 필터링를 위해 Polyak Averaging이라고 불리는 soft update 방법으로 업데이트된다. 이는 다음과 같이 표현된다.

$$\bar{\theta}_i \leftarrow (1-\tau)\theta_i + \tau\bar{\theta}_i, \quad i = 1,2 \tag{15}$$

$$\bar{\phi}_i \leftarrow (1-\tau)\phi_i + \tau\bar{\phi}_i \tag{16}$$

여기서, τ 는 soft update 비율이다.

2.3 컨버터 제어를 위한 심층강화학습 구성

심층강화학습 제어의 행동은 듀티비이다. 투 스위치 포워드 컨버터는 스위치 턴-오프 시 변압기 포화를 방지하기 위해 자화 에너지를 모두 방출해야 하므로 듀티비는 0에서 0.45로 제한된다.

심층강화학습을 통해 컨버터를 제어하기 위해서는 보상 설계가 필요하다. 보상은 원하는 목표를 달성하기 위한 성능 지표의 기준이 된다. 본 논문에서 설계한 보상 함수는 다음과

표 1 투 스위치 포워드 컨버터의 파라미터

파라미터	값	단위
입력 전압(V_{in})	311	V
출력 전압(V_{out})	80	V
스위칭 주파수	10	kHz
권선 비(n)	1.2	–
자화 인덕턴스(L_m)	300	μH
필터 인덕턴스(L_f)	8.8	mH
필터 커패시턴스(C_f)	22	μF

표 2 Twin Delayed Deep Deterministic policy gradient algorithm의 하이퍼 파라미터

하이퍼 파라미터	값
Actor 학습률	0.002
Critic 학습률	0.05
버퍼 사이즈	10000
미니 배치 사이즈	128
할인율	0.95
탐색 표준편차	0.015
탐색 표준편차 감쇠율	0.0001
정책 표준편차	0.015
정책 표준편차 감쇠율	0.002
최대 에피소드	1000

같다.

$$\begin{aligned} r &= -error && (error > 2.4) \\ r &= error && (1 < error \le 2.4) \\ r &= 20 && (0.5 < error \le 1) \\ r &= 40 && (error \le 0.5) \end{aligned} \quad (17)$$

여기서, *error*는 출력 전압과 레퍼런스 전압의 차이이다.

에이전트의 입력은 error 값, error 이전 값, error의 변화량, 이전 에이전트의 행동이다.

2.4 시뮬레이션

제안하는 제어기법의 성능 검증을 위해 MATLAB/Simulink 환경에서 PI 제어 기법과 비교 검증된다. 제안하는 제어기법은 출력 전압을 제어하도록 설계되어 PI 제어 기법도 출력 전압 제어를 수행하도록 설계하였다.

2.4.1 시뮬레이션 조건

투 스위치 포워드 컨버터의 파라미터는 표1과 같다. 해당 필터의 파라미터를 통해 전압 리플은 1% 이내, 전류 리플은 10% 이내로 설계되었다. 심층강화학습 제어를 위한 TD3의 하이퍼파라미터는 표2와 같다.

시뮬레이션은 Matlab/Simulink 환경에서 진행하였다. 부하 변동 조건을 주었으며 출력 레퍼런스 전압은 변동되지 않는다. 총 시뮬레이션 시간은 0.49초이며 0.07초마다 부하가 변동되도록 설계하였다. 부하율 100%에서 500W로 설계되었으며 0.07초마다 부하율이 100%, 75%, 50%, 25%, 50%, 75%, 100%로 변경된다. 이에 따라 부하 변동에 대한 출력 전압 및 출력 전류의 동적 응답결과로부터 TD3와 PI 제어의 성능과 안정성을 비교한다.

2.4.2 시뮬레이션 결과

강화학습의 각 에피소드에 대한 보상 결과는 그림 3과 같다. 125번 에피소드부터 수렴하기 시작하였으며 가장 보상이 좋은 에피소드를 기반으로 에이전트를 도출하였다. 부하 변동에 대한 출력 전압 및 출력 전류의 동적 응답 시뮬레이션 결과는 그림 4와 같다. 초기 부하율 100%에서 TD3의 경우 1ms에 정상상태로 도달하였으며 PI 제어의 경우 4ms에 정상상태로 도달하여 TD3가 PI 제어에 비해 4배 빠르게 정상상태에 수렴하였다. 부하 변동 시 TD3와 PI 제어 모두 오버 슈트가 발생하였으나 TD3의 경우 최대 오버 슈트는 114%, PI 제어의 경우 128%이다. 부하율 50%와 25%에서 PI제어의 경우 정상상태에서 비감쇠 진동이 발생하여 불안정하나 TD3는 모든 부하율 변동에 대해 정상상태에서 진동이 발생하지 않는다.

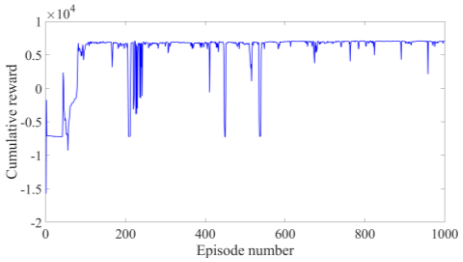

그림.3 각 에피소드에 대한 보상 결과

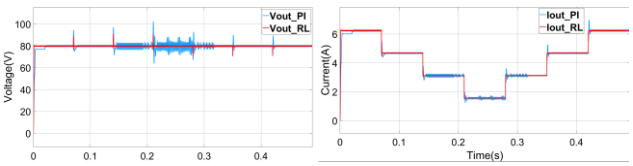

그림.4 부하 변동에 대한 출력 전압 및 전류 동적 응답 곡선

4. 결 론

본 논문에서는 투 스위치 포워드 컨버터의 성능 및 안정성 향상을 위한 강화학습기반 제어를 제안하였다. 강화학습기반 제어는 PI 제어에 비해 넓은 동작 조건 범위에서도 더 높은 성능과 안정성을 보장하였다. 추후 DSP에 강화학습제어를 다운로드하여 실제 환경에서의 강화학습 기반 컨버터 제어를 구현할 예정이다.

이 논문은 2023년도 정부(산업통상자원부)의 재원으로 한국에너지기술평가원의 지원을 받아 수행된 연구임(RS-2023-00234563, 분산에너지 기반 그리드포밍 적용 전력망 모델링 및 분석, 상호운용성 평가 기술 개발)

참 고 문 헌

[1] C. Cui, T. Yang, Y. Dai, C. Zhang and Q. Xu, "Implementation of Transferring Reinforcement Learning for DC–DC Buck Converter Control via Duty Ratio Mapping," in IEEE Transactions on Industrial Electronics, vol. 70, no. 6, pp. 6141–6150, June 2023

[2] Ho-Jun Lee, Yoon-Gun Jung, Don Hur, Heung-jae Lee, and Min-Han Yoon, "A3C 강화학습을 활용한 VSC 컨버터 PI 제어기 튜닝." 대한전기학회 학술대회 논문집 2021.10 (2021): 230-231.

[3] Fujimoto, Scott, Herke Hoof, and David Meger. "Addressing function approximation error in actor-critic methods." International conference on machine learning. PMLR, 2018.