第 50 卷 第 2 期：714-725　　高电压技术　　Vol.50, No.2: 714-725
2024 年 2 月 29 日　　High Voltage Engineering　　February 29, 2024
DOI: 10.13336/j.1003-6520.hve.20230104

基于卡尔曼滤波器及深度强化学习的双有源全桥变换器控制策略

武　涵，贾燕冰，韩肖清，石俊逸，孟祥齐
（电力系统运行与控制山西省重点实验室(太原理工大学)，太原 030024）

摘　要：高比例的新能源以及随机性负载大量接入微电网，其不确定性带来的大扰动对直流母线电压的稳定性造成不良影响。为了实现大扰动下快速、自适应的电压调节，针对双有源桥式 DC-DC 变换器(dual active bridge, DAB)提出了一种基于卡尔曼滤波器(Kalman filter, KF)及深度强化学习的新型复合控制策略。设计了基于 Actor-Critic 架构的深度确定性策略梯度强化学习智能体，采用 KF 的最佳观测结果作为前馈补偿提高输出电压调节的准确性，通过在线学习自动调整 DAB 变换器的控制参数，保证直流变换器在面临系统各种扰动问题时均保持稳定，最后通过仿真和实验验证了该控制策略的有效性。

关键词：双有源桥式 DC-DC 变换器；深度强化学习；DDPG 智能体；卡尔曼滤波器；大扰动

Research on Control Strategy of Dual-active Full-bridge Converter Based on Deep Reinforcement Learning and Kalman Filter

WU Han, JIA Yanbing, HAN Xiaoqing, SHI Junyi, MENG Xiangqi
(Shanxi Key Laboratory of Power System Operation and Control, Taiyuan University of Technology, Taiyuan 030024, China)

Abstract: A large proportion of new energy and random loads are connected to microgrids, and the large disturbance caused by its uncertainty has a negative impact on the stability of DC bus voltage. In order to realize fast adaptive voltage regulation under large disturbance, a new compound control strategy based on Kalman filter (KF) and deep reinforcement learning is proposed for dual-active bridge (DAB)DC-DC converter. A deep deterministic strategy gradient reinforcement learning agent based on the Actor-Critic architecture is designed. The best observation results of KF are used as feed-forward compensation to improve the accuracy of output voltage regulation. The control parameters of DAB converter are automatically adjusted through online learning to ensure that the DC converter is stable in the face of various system disturbances. Finally, the effectiveness of the control strategy is verified by simulation and experiments.

Key words: dual active bridge DC-DC converter; deep reinforcement learning; DDPG agent; Kalman filter; large disturbance

0　引言

直流微电网因能充分发挥分布式电源的价值和效益、协调大电网和分布式电源之间的矛盾[1-2]，得到了广泛应用。分布式电源、蓄电池、电动汽车等接入直流微电网，使得直流微电网成为典型非线性系统[3-4]。由于光伏和风电发电的间歇性以及负载侧扰动等因素，这导致直流微电网中母线电压波动较大，从而增加其稳定性风险[5-7]。双有源桥式 DC-DC 变换器(dual active bridge, DAB)拥有高功率密度、高效率、能量双向传输和便于级联和并联等优点[8-11]，是直流微电网中源荷并网的关键设备。对于直流微电网的正常运行，DAB 变换器的控制策略至关重要，它能够确保直流母线电压稳定[12-14]。

目前 DAB 广泛采用的 PI 控制器大多基于特定工作点线性化小信号模型设计，其控制参数相互耦合，对结构和参数不确定的强非线性系统控制效果不佳[15]；研究人员采用自适应 PI 增益控制器[16]、鲁棒 PI 控制[17]，或采用拉格朗日乘数法[18]、遗传算法[19]和牛顿法[20]等与 PI 控制相结合，根据不同

基金资助项目：国家自然科学基金联合基金重点项目(U1910216)；山西省重点研发计划项目(国际合作)(201803D421010)。
Project supported by Key Project of National Natural Science Foundation of China (U1910216), Key R&D Program of Shanxi Province (Internalional Cooperation)(201803D421010).

的工作点迭代修正 PI 增益参数，一定程度上提升了 PI 控制器的瞬态响应能力性能，但其本质仍然基于小信号分析模型，对小扰动稳定控制能力较强，当系统遭遇大扰动时，基于 PI 控制的线性控制策略易发生饱和、动态性能差，增大了直流微电网失稳风险。

为应对大扰动情况，文献[21]提出一种直接功率控制策略，提升了输入电压突变过程的动态响应能力；文献[22]采用大信号分析建模，采用逆系统解耦控制解决 Buck-Boost 变换器出现较大电压扰动的情况；文献[23]提出高阶滑模控制保持大扰动下 Buck-Boost 变换器的输出电压稳定性。直流微电网为典型的复杂非线性系统，高度依赖模型知识的 DAB 控制策略，适应场景有限，具备可适应性、自完善性的新型控制策略成为时之所趋。

深度强化学习凭借自学习、自完善、应用场景广的特点，已经广泛应用于控制问题的优化[24]，其中深度确定性策略梯度算法(deep deterministic policy gradient，DDPG) 通过决策过程中产生的大量数据训练智能体，是一种不受模型限制的数据驱动型控制方法。将 DDPG 控制应用到 DAB 变换器系统中，将可以实现 DAB 系统在运行过程中不断自我修正和进化，提升系统动态性能和稳定性。由于直流微电网中源荷扰动不确定性较高，为有效降低 DDPG 跟踪误差，提升系统控制效果，需提升对系统未知扰动的估计准确性。研究表明观测器技术是一种估计不确定扰动的有效方法，并且不需要额外的传感器，卡尔曼滤波器(Kalman filter, KF)[25-26]可估计三相无刷双馈感应电机的转子速度和功率，为抗干扰和不确定性补偿提供了一种很有前景的解决方案。

本文针对 DAB 系统在新能源并网和不确定性负载并网等各种扰动下的稳定问题，提出了一种基于 KF 的 DDPG 复合非线性控制策略。根据 DAB 系统输入输出状态设计了基于 Actor-Critic 神经网络架构的 DDPG 智能体控制器，并采用卡尔曼滤波器在每个控制周期内估计和补偿由源荷波动引起的未知扰动，根据决策过程中生成的系统数据在线学习不断自我修正和进化，并对其决策行为进行评价，通过最大化奖励提升输出动作的决策精度，增强 DAB 系统应对新问题的适应能力，保障其输出电压的稳定性，实现 DAB 系统的自适应和自稳定性。仿真和实验结果表明，本文所提新型控制策略，可以在不同程度扰动下均保证 DAB 稳定运行。

1　DAB 控制模型

1.1　DAB 的结构

隔离型双有源全桥 DC-DC 变换器的主电路如图 1 所示，由两个完全对称的全桥(full bridge，FB)和一个高频变压器 T(变比为 1:n)组成。其中，每个全桥包含 4 个功率开关管(S_1—S_4, S_5—S_8)，两个直流电容 C_i 和 C_o，L 中包含高频变压器的漏感和辅助电感。U_a 和 U_b 分别是高频变压器的初级电压和次级电压，U_i 和 U_o 表示 DAB 的输入、输出电压，U_L 和 i_L 分别是电感 L 的电压和电流，T 提供电气隔离和电压变换的功能，电感 L 提供能量暂存功能，反并联二极管提供功率双向流动通路。

脉宽调制控制和相移控制是 DAB 变换器的两种基本控制方法[27-28]。相比于脉宽调制控制，相移控制具有易于控制、高动态性能和软开关等优点得以广泛应用。

DAB 平均值等效电路如图 2，图中 I_i 和 I_o 分别表示输入侧和输出侧电流，输入电压源为 U_s，输入、输出电阻分别为 R_s 和 R，输入输出电容为 C_i 和 C_o。

在整个工作周期内，电感电流 i_L 是交流分量，平均值为零。因此，不考虑 i_L 状态变量，对 N_1、N_2 节点根据基尔霍夫定律可得 DAB 变换器状态方程如(1)、(2)所示，其中 f 是开关频率，D 为移相比。

$$\frac{\mathrm{d}U_i}{\mathrm{d}t}=-\frac{U_i}{R_sC_i}+\frac{nU_o}{2fLC_i}D(|D|-1)+\frac{U_s}{R_sC_i} \tag{1}$$

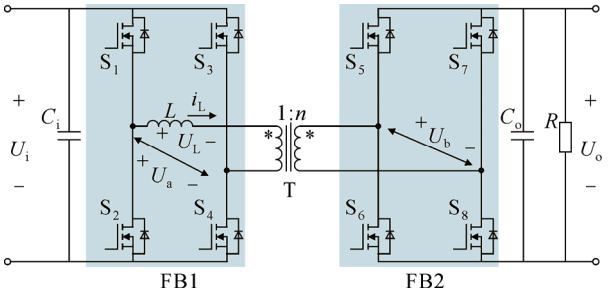


图 1　双有源全桥变换器结构

Fig.1　Dual active full-bridge converter structure

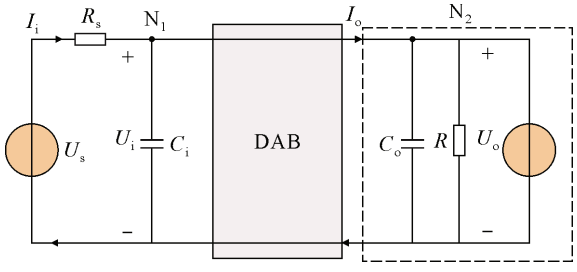


图 2　DAB 平均值等效电路

Fig.2　Average value equivalent circuit of DAB

$$\frac{\mathrm{d}U_{\mathrm{o}}}{\mathrm{d}t}=-\frac{U_{\mathrm{o}}}{RC_{\mathrm{o}}}+\frac{nU_{\mathrm{i}}}{2fLC_{\mathrm{o}}}D(1-|D|) \tag{2}$$

由式(2)可知移相控制方法能够通过调节 D 来控制输出电压。D>0 时，功率正向传输，D<0 时，功率负向传输。

1.2 基于 KF 和 DDPG 的 DAB 控制模型

在大规模新能源并网的复杂场景下，系统中将出现大量不同幅度的非线性扰动，传统 PI 控制方法超调量高、调节时间长、自适应能力差，随着扰动范围增大，不能确保 DAB 输出电压的稳定性。DDPG 控制可实现 DAB 系统在运行过程中不断自我修正和进化，提升系统动态性能和稳定性。但工程应用中，测量动态系统中的所有变量是不切实际的，而 KF 能通过动态的状态空间模型来估计动态过程的状态[29-30]，是存在噪声和干扰的情况下用于状态的最优估计器[31]。

本文将 KF 和 DDPG 用于 DAB 控制。通过 KF 在每个控制周期内对噪声和源荷等随机性非线性扰动进行估计后，再提供给 DDPG，提升 DDPG 控制器从环境摄取有效信息和训练的效率，从而快速确定 DAB 系统运行状态，同时由于 DDPG 具备自学习的能力，在运行过程中会根据 DAB 所遇到的运行场景不断修正其控制过程，从而提高 DAB 系统的抗扰动能力和稳定性。其控制器的结构如图 3 所示，其中 U_{ref} 表示输出电压参考值；a_t 代表 DDPG 控制器输出动作值；Z 表示在观测误差 N 存在下的输入、输出电压观测结果；M 表示系统中存在的过程误差；U 表示经 KF 后的输出电压。

2 Kalman 滤波器设计

选择 U_{o}、U_{i} 作为系统状态变量，$\boldsymbol{x}=[U_{\mathrm{o}}\ U_{\mathrm{i}}]^{\mathrm{T}}$，由式(1)、(2)得 DAB 系统状态方程：

$$\dot{\boldsymbol{x}}=\boldsymbol{A}\boldsymbol{x}+\boldsymbol{B}U_{\mathrm{s}} \tag{3}$$

其中 $\boldsymbol{A}=\begin{bmatrix}-\dfrac{1}{RC_{\mathrm{o}}} & \dfrac{nD(1-D)}{2fLC_{\mathrm{o}}}\\ \dfrac{nD(1-D)}{2fLC_{\mathrm{i}}} & -\dfrac{1}{R_{\mathrm{s}}C_{\mathrm{i}}}\end{bmatrix}$，$\boldsymbol{B}==\begin{bmatrix}0\\ \dfrac{1}{R_{\mathrm{s}}C_{\mathrm{i}}}\end{bmatrix}$

将式(3)离散化得

$$\boldsymbol{X}_{n+1}=\boldsymbol{A}_{\mathrm{e}}\boldsymbol{X}_n+\boldsymbol{B}_{\mathrm{e}}U_{\mathrm{s}} \tag{4}$$

式中：$\boldsymbol{A}_{\mathrm{e}}=\mathrm{e}^{\boldsymbol{A}T_{\mathrm{s}}}$，$\boldsymbol{B}_{\mathrm{e}}=\int_0^{T_{\mathrm{s}}}\mathrm{e}^{\boldsymbol{A}\tau}\boldsymbol{B}\mathrm{d}\tau$，$T_{\mathrm{s}}$ 为离散化步长(与 DDPG 控制采样时间相同)，考虑到实际系统中的过程噪声和测量噪声，则 DAB 状态方程和 KF 观测器的模型如下式：

$$\begin{cases}\boldsymbol{X}_{n+1}=\boldsymbol{A}_{\mathrm{e}}\boldsymbol{X}_n+\boldsymbol{B}_{\mathrm{e}}U_{\mathrm{s}}+\boldsymbol{W}_n\\ \boldsymbol{Z}_n=\boldsymbol{H}\boldsymbol{X}_n+\boldsymbol{V}_n\end{cases} \tag{5}$$

式中：$\boldsymbol{X}_n$ 为实际值；$\boldsymbol{Z}_n$ 为测量值；$\boldsymbol{H}$ 为二阶量测矩阵 $\boldsymbol{H}=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$；$\boldsymbol{W}_n$ 和 $\boldsymbol{V}_n$ 分别对应过程噪声和测量噪声，其特性如下：

$$\begin{cases}\boldsymbol{E}(\boldsymbol{W}_n)=0,\boldsymbol{C}_{\mathbf{ov}}(\boldsymbol{W}_k,\boldsymbol{W}_j)=\boldsymbol{M}\\ \boldsymbol{E}(\boldsymbol{V}_n)=0,\boldsymbol{C}_{\mathbf{ov}}(\boldsymbol{V}_k,\boldsymbol{V}_j)=\boldsymbol{N}\\ \boldsymbol{C}_{\mathbf{ov}}(\boldsymbol{W}_k,\boldsymbol{V}_k)=0\end{cases} \tag{6}$$

式中：$\boldsymbol{M}$ 为过程噪声 $\boldsymbol{W}_n$ 的方差阵；$\boldsymbol{N}$ 代表测量噪声 $\boldsymbol{V}_n$ 的方差阵。

基于(5)构建如式(7)的观测器方程

$$\begin{aligned}\hat{\boldsymbol{X}}_n&=\boldsymbol{A}_{\mathrm{e}}\hat{\boldsymbol{X}}_{\mathrm{n\text{-}1}}+\boldsymbol{B}_{\mathrm{e}}U_{\mathrm{s}}+\boldsymbol{K}_n(\boldsymbol{Z}_{n-1}-\hat{\boldsymbol{Z}}_{n-1})=\\ &\boldsymbol{A}_{\mathrm{e}}\hat{\boldsymbol{X}}_{n-1}+\boldsymbol{B}_{\mathrm{e}}U_{\mathrm{s}}+\boldsymbol{K}_n\boldsymbol{H}(\boldsymbol{X}_{n-1}-\hat{\boldsymbol{X}}_{n-1})\end{aligned} \tag{7}$$

式中：带有^表示估计值，通过求得观测器增益矩阵 $\boldsymbol{K}_n$ 可实现最优估计。KF 的计算过程如图 4 所示，包含滤波、增益计算两回路。

先验状态估计 $\hat{\boldsymbol{X}}_{n/n-1}$

$$\hat{\boldsymbol{X}}_{n/n-1}=\boldsymbol{A}_{\mathrm{e}(n/n-1)}\hat{\boldsymbol{X}}_{n-1}+\boldsymbol{B}_nU_{\mathrm{s}} \tag{8}$$

预测误差协方差矩阵 $\boldsymbol{P}_{n/n-1}$

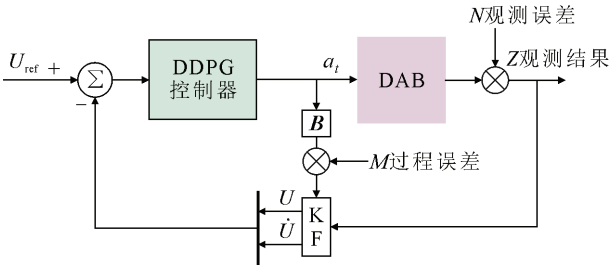


图 3 基于 KF 和 DDPG 的 DAB 复合控制框图

Fig.3 The composite control block diagram of DAB by KF and DDPG

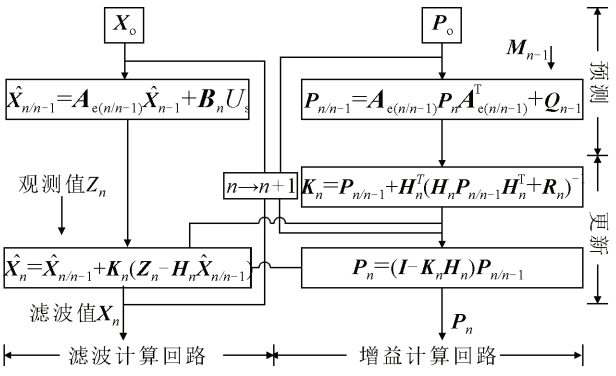


图 4 KF 计算过程

Fig.4 Calculation process of Kalman filter

$$\boldsymbol{P}_{n/n-1}=\boldsymbol{A}_{\mathrm{e}(n/n-1)}\boldsymbol{P}_n\boldsymbol{A}^{\mathrm{T}}_{\mathrm{e}(n/n-1)}+\boldsymbol{M}_{n-1} \tag{9}$$

增益矩阵 $\boldsymbol{K}_n$

$$\boldsymbol{K}_n=\boldsymbol{P}_{n/n-1}\boldsymbol{H}_n^{\mathrm{T}}\left(\boldsymbol{H}_n\boldsymbol{P}_{n/n-1}\boldsymbol{H}_n^{\mathrm{T}}+\boldsymbol{N}_n\right)^{-1} \tag{10}$$

计算状态变量的最优估计值 $\hat{\boldsymbol{X}}_n$

$$\hat{\boldsymbol{X}}_n=\hat{\boldsymbol{X}}_{n/n-1}+\boldsymbol{K}_n\left(\boldsymbol{Z}_n-\boldsymbol{H}_n\hat{\boldsymbol{X}}_{n/n-1}\right) \tag{11}$$

计算后验误差协方差矩阵 $\boldsymbol{P}_n$

$$\boldsymbol{P}_n=\left(\boldsymbol{I}-\boldsymbol{K}_n\boldsymbol{H}_n\right)\boldsymbol{P}_{n/n-1} \tag{12}$$

递推计算出来的最佳观测状态 $\hat{\boldsymbol{X}}_n$，作为 DDPG 控制器中神经网络的输入和环境变量，提升控制器的鲁棒性，在下个控制周期内继续循环计算过程。

3　基于 Kalman 滤波器的深度强化学习控制

3.1　强化学习 DDPG 设计

强化学习框架一般描述为马尔可夫决策(Markov decision process, MDP)过程[32]，是对强化学习问题的数学描述。DDPG 基于 DQN[33]和 Q-learning[34]，采用 value-based 的方法训练 Critic-Network，利用时间差分(temporal difference, TD)方法计算价值函数 $V^{\pi}(s)$[35]，算法的代码详见附录表 B1。

训练过程中，最小损失损失函数表示为 L_0：

$$L_0=\frac{1}{N}\sum_i\left(y_i-Q\left(s_i,a_i|\ \theta^Q\right)\right)^2 \tag{13}$$

式中：s_i, a_i 分别表示当前的状态值和基于该状态的动作值，$y_i=r_i(s_i,a_i)+\gamma Q'\left(s_{i+1},\mu'\left(s_{i+1}|\ \theta^{\mu'}\right)|\ \theta^{Q'}\right)$，利用(14)的策略梯度，更新 Actor 网络中的权重系数 θ^{μ}。

$$\nabla_{\theta^{\mu}}J^{\theta^{\mu}}\approx$$
$$E_{s_t\sim\rho^{\beta}}\left(\nabla_{\theta^{\mu}}Q\left(s,a|\ \theta^Q\right)\Big|_{a=\mu\left(s|\ \theta^{\mu}\right)}\nabla_{\theta}{}^{\mu}\mu\left(s|\ \theta^{\mu}\right)\right)= \tag{14}$$
$$E_{s_t\sim\rho^{\beta}}\left(\nabla_aQ\left(s,a|\ \theta^Q\right)\Big|_{a=\mu^{\theta}(s)}\nabla_{\theta^{\mu}}\mu\left(s|\ \theta^{\mu}\right)\right)$$

式中：ρ 是折扣系数；β是当前策略π的探索因子。在时间步长 t 内，DDPG 代理与环境交互，将经验元组 $e_t=\left(s_t,a_t,r_t,s_{t+1}\right)$ 存储到经验池 T 中。

为了保证 DDPG 学习的稳定性，过程中采用目标网络逼近。主网络、目标网络的权重系数 $\theta^{Q'}$、$\theta^{\mu'}$ 和通过(15)进行参数对更新[24]。

$$\begin{aligned}&\theta^{Q'}\leftarrow\tau\theta^{Q}+(1-\tau)\theta^{Q'}\\&\theta^{\mu'}\leftarrow\tau\theta^{\mu}+(1-\tau)\theta^{\mu'},\\&\tau<<1\end{aligned} \tag{15}$$

DDPG 控制器能够根据实时反馈信息自动地调整控制策略，能够处理非线性和时变系统，实现更好的输出电压控制。训练过程中过渡数据保存到经验缓存器中，随机抽取批量的经验值样本，用于训练 DDPG 神经网络相应参数。构建容量为 C_1 的内存库 M_1 存储当前状态 s_t，动作μ 完成之后，获得奖励 r_t 和下一时刻状态 s_{t+1}。初始化状态是 $s_1=[0,0,0]$，动作 $a_t=\mu(s_t|\theta^{\mu})+Nosie$ 是基于动作评估网络选择的，Critic-Network 通过最小化损失函数(13)来更新临界值，随后导入到方程(14)中更新 Actor-Network。

图 5 中详细描述了 DAB 电路与复合控制交互的过程，包括提取环境，存储数据，训练和输出动作值。以 DAB 变换器系统为环境对象，通过 KF 从中采集输出电压信号、输入电压信号。将理想输出电压作为参考量，环境信息中的输出电压信号 v_o 与 v_{ref} 之间的偏值 e 及其积分值、微分值作为深度神经网络 Actor-Critic 网络的输入，以减小 e 作为控制目标，经过 Actor 网络输出动作值 D_1 实现控制过程。

3.2　强化学习 Agent 的设计方案

DDPG 中 Actor-Critic 的结构关系如图 6 所示。图 6 左边是 Actor-Network 架构，网络的输入层包含 3 个神经元，对应于 3 个输入节点，即输出电压 U_o、输出电压偏差值 $e=\left(U_o-U_{ref}\right)/U_{ref}\times100\%$ 以及 $\mathrm{d}e/\mathrm{d}t$，输出层为 $a_t=\mu(s_{t+1}|\theta^{\mu-})$。

图 6 右半部分是 Critic-Network 架构，输入层中包含有 4 个神经元，对应于 4 个输入节点，即 U_o、e、$\mathrm{d}e/\mathrm{d}t$ 与 Actor-Network 的输出 a_t，输出层为状态动作值函数 $Q(s_{t+1},\mu(s_{t+1}|\theta^{\mu-})|\theta^{Q-})$。

Actor-Network 和 Critic-Network 都包含 3 个隐藏层，第 1 层和第 2 层中均包括 200 个神经元，激活函数类型均为ReLu6，第3层中包含10个神经元，激活函数类型为 ReLu；均采用 L2 正则化来防止网络过度拟合。

Critic-Network 输出层 Q 函数表示在 s 状态下采取策略μ输出动作所获得的奖励期望值，表示为：

$$Q^{\mu}\left(s_t,a_t\right)=E\left(r\left(s_t,a_t\right)+\gamma Q^{\mu}\left(s_{t+1},\mu\left(s_{t+1}\right)\right)\right) \tag{16}$$

其中 γ 表示折扣系数，用 $J_{\beta}(\mu)$表达策略的优劣：

$$\begin{aligned}J_{\beta}(\mu)&=\int_s\rho^{\beta}(s)Q^{\mu}(s,\mu(s))\mathrm{d}s=\\&E_{s\sim\rho^{\beta}}(Q^{\mu}(s,\mu(s))\end{aligned} \tag{17}$$

为保证输出动作能够找到 DAB 系统最佳的工作状态，加入探索因子提高动作的探索区域。β是通过在动作策略过程中加入从 Ornstein-Uhlenbeck(OU)噪声用于探索，OU 是一种随机过程，表述如下：

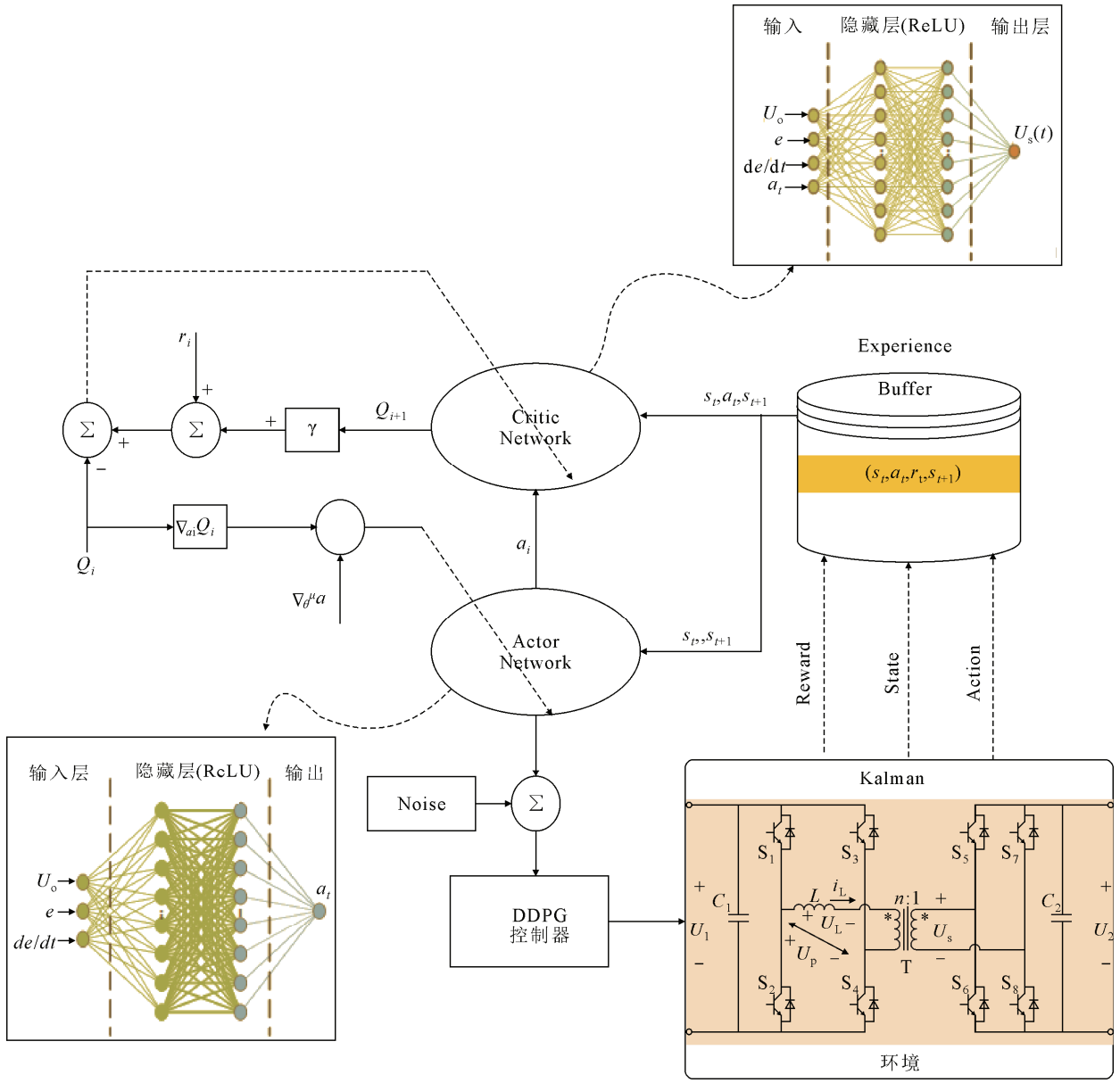


图 5 带有 Kalman 滤波器的 DDPG 控制器方案

Fig.5 DDPG controller scheme with kalman filter

$$dx_t = -\theta(x_t - \mu)dt + \sigma dW_t \tag{18}$$

式中：μ 为均值；θ、σ>0,W_t 是布朗运动过程；x_t 是状态变量。当 x_t>μ，下一状态 x_t 就会减小；反之，x_t 就会增加。总的来说，状态发生偏离就会拉回到均值。

对(18)进行求解，解得：

$$x_t = \mu + (x_0 - \mu)e^{-\theta t} + \sigma\int_0^t e^{-\theta(t-s)}dW_s \tag{19}$$

式中：θ 越大，干扰越小、系统向均值靠近越快。扰动项在每段时间间隔增量符合高斯分布：

$$W(t) - W(s) \sim N(0, \sigma^2(t-s)) \tag{20}$$

σ 决定噪音的变化尺度，强化学习中加入 OU 的动态探索可以提高 DAB 系统控制任务的探索效率。根据(18)噪声会随着学习频率的增加进行动态

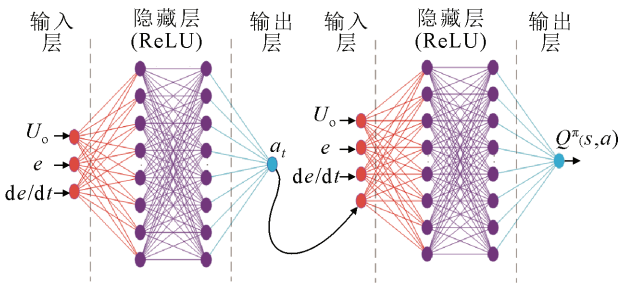


图 6 Actor-Critic 的神经网络架构

Fig.6 Actor-Critic neural network architecture

调整，此过程提高了神经网络的收敛速度。

本文针对 DAB 系统的控制目标是提高输出电压的稳定性，设定奖励函数 r_t 随着电压偏差$|e(t)|$增大而减小。利用(16)计算 Q 进而动态调整 Actor-Network 的输出 a_t，以寻求 $Q_{\max}$。为了加快收敛速度，根据电压偏差范围，设置(21)中 4 分段奖励函数 r_t。由于指数函数爆炸性增长的特点，r_t 随着 $\mathrm{e}^{f(|e(t)|)}$ 增减变化明显，在输出电压发生偏差时奖惩分明，能够提高训练的效率；四分段函数满足所有动作空间能实现对 DAB 输出电压稳定的连续性控制。

$$r_t=\begin{cases}-1000, & |e(t)|\ge 5\%\\ -\mathrm{e}^{50|e^2(t)|-0.5}+1, & 1\%\le|e(t)|<5\%\\ \mathrm{e}^{-500|e(t)|+6}, & 0.5\%\le|e(t)|<1\%\\ \mathrm{e}^{\frac{1}{50|e(t)|+0.1}+2}, & 0\le|e(t)|<0.5\%\end{cases}\tag{21}$$

1）系统允许最大电压偏差为 5%，故设定输出电压 U_o 偏差的边界值为$|e(t)|$=5%，当$|e(t)|\ge$5%，设定 $r_t=-1000$，使得 Actor 网络停止动作值输出，结束该周期的训练过程。

2）当$1\%\le|e(t)|<5\%$时，$|e^2(t)|$变化幅度明显，设 $f(|e(t)|)=50|e^2(t)|-0.5$，附加常数 1 奖励值来避免训练过程终止、实现奖励由负到正阶跃。

3）当$0.5\%\le|e(t)|<1\%$时，$|e(t)|$较 $e^2(t)$变换幅度大，设 $f(|e(t)|)=-500|e(t)|+6$。

4）当$0\le|e(t)|<0.5$时，DAB 的输出电压保持在相对稳定，设 $f(|e(t)|)=1/(50|e(t)|+0.1)+2$，随着$|e(t)|$的变 r_t 增减更加明显。

训练曲线如附录 A1 所示。可以看出，当训练次数达到 600 次时，奖励趋于 2400，能够使输出电压维持在 1%以内，维持电压稳定性。当 DAB 工作环境发生变换时，DDPG 智能体还能通过实时在线学习，在满足 r_t 最大化的目标下维持输出电压稳定，具有良好自适应性。

4　仿真结果

为验证所提控制策略的有效性，在 Matlab/Simulink 平台，搭建了图 1 所示的双有源桥 DC-DC 变换器模型，控制方案如图 5 所示，主电路及 PI 控制参数如表 1 所示。DDPG 控制中相关参数如表 2 所示，OU 噪声探索过程中使用 Actor-Network 的输出值作为分布均值 μ，θ=0.5，σ=1；取

表 1　元件参数

Table 1　Component parameters

参数	数值
输入电压 U_s/V	100
输出电压 U_o/V	100
变压器变比 1:n	1:1
变压器额定工作频率 f/kHz	50
辅助电感 L_r/μH	20
电容 C_i/μF	1000
电容 C_o/μF	1000
K_P	0.5
K_i	10

表 2　DDPG 算法的关键参数

Table 2　Key Parameters of DDPG Algorithm

参数	数值
学习率(α)	0.005
折扣系数	0.9
噪声方差	0.4
探索模型方差	0.1
梯度阈值	1
单次样本数	256
Actor 学习率	0.000 01
Critic 学习率	0.0001
存储器 M_1 容量 C_1	50 000

T_s=0.000 01 s、T_f=0.1 s。输入侧的额定电压为 100 V，输出侧的额定电压为 100 V。

1）不同扰动下控制性能

分别对比采用 PI 控制策略和本文所提制策略时 DAB 输出电压的稳定性。由于 PI 控制性能与控制参数选取密切相关，在负载 600 W 扰动下，设定 PI 控制参数以保持输出系统的稳定性。其中，U_i 和 U_o 分别表示输入电压和输出电压，I_i 和 I_o 分别表示输入电流和输出电流。不同扰动和控制策略下系统的性能特征参数如表 3 所示。

工况 1：t_1 时刻，负载端由 200 W 阶跃到 800 W，t_2 时刻降低到 200 W，输出波形如图 7 所示。

工况 2：t_1 时刻，负载端由 200 W 阶跃到 1400 W，t_2 时刻降低到 200 W，输出波形如图 8 所示。

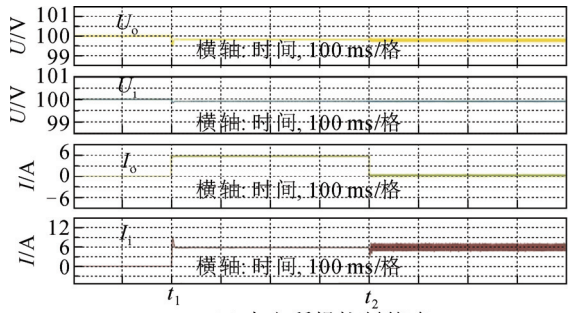
由图 7 和表 3 可知：当负载变化±600 W 时，本文所提控制策略超调量及电压偏差相较于 PI 控制均降低近 87.6%，调节时间缩短 73.6%；随着扰动增大，当负载变化±1200 W 时，如图 8 所示，本文所提策略的超调量及电压偏差相较于 PI 控制均降低近 98%，调节时间缩短 76%，PI 控制已无法保障系统稳定状态。

由此可见，随着扰动大，PI 控制调节能力逐步降低，甚至使得 DAB 系统失稳，而本文所提控制策略可保持系统的稳定性，受扰动大小的影响较小，具有较好的控制精度。

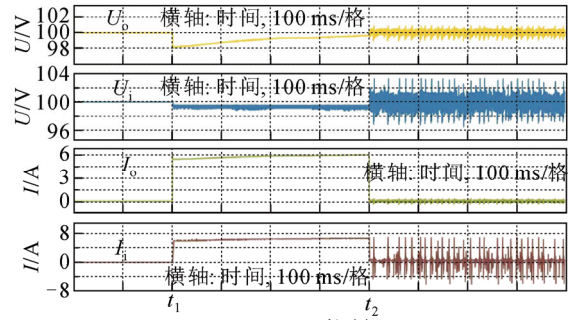
2）对输入电压变化的鲁棒性

为验证所提出的控制器在输入电压及负载同时出现扰动下的鲁棒性，用两种输入电压变化的情况来比较仿真结果。图 9(a)—(b)分别显示了使用阶跃变化的输入电压和三角波变化的输入电压得在图 9 中，负载在 t_1 时从 200 W 变为 800 W，在 t_4 时变为 200 W。图 9(a)中，输入电压加入阶跃变化，在 t_2 时从 0 V 变化到 35 V，在 t_3 时变成−15 V；在图 9(b)中，输入电压加入频率为 25 Hz 的三角波变化，其振幅在 t_2 时从 0 V 变化到 28 V，在 t_3 时变为 10 V。

仿真结果可以看出，在输入电压发生阶跃变化和三角变化以及输出负载侧出现扰动的情况下，系统的输出电压偏差和超调量能保持在±0.5%以内，平均调节时间为 18 ms。因此，本文所提出的控制



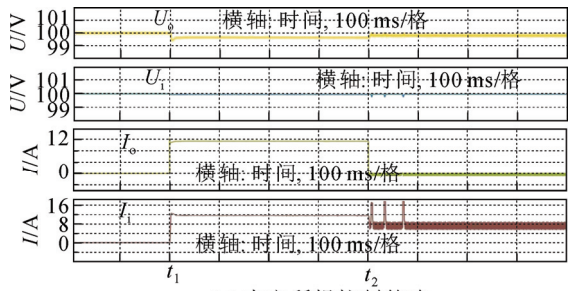
(a) 本文所提控制策略



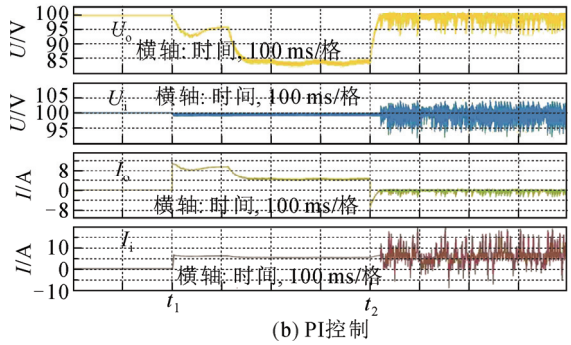
(b) PI控制

图 7 负载变化±600 W 控制效果对比图

Fig.7 DAB output voltage with ±600 W disturbance



(a) 本文所提控制策略



(b) PI控制

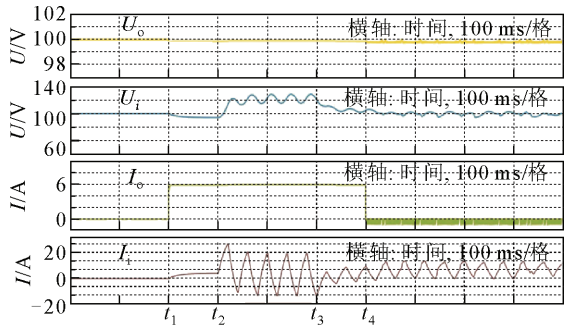
图 8 负载变化±1200 W 控制效果对比图

Fig.8 DAB output voltage with ±1200 W disturbance

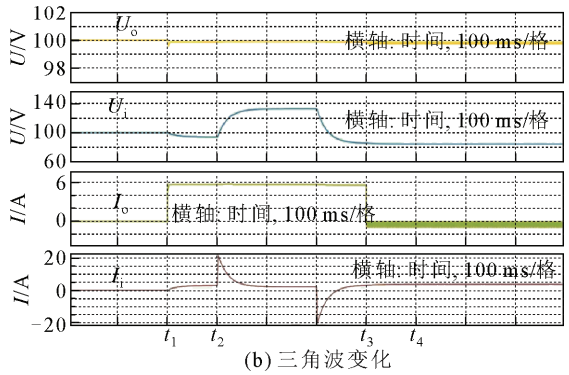
表 3 不同控制策略下系统的性能特征

Table 3 Characters of system performance with different control strategy

负载扰动		控制方法	偏差值/V	调节时间/ms	超调量/V
600 W	增大	PI	1.50	95	2.02
		本文	0.17	15	0.32
	减小	PI	1.23	45	1.52
		本文	0.10	10	0.25
1200 W	增大	PI	失稳	失稳	16.2
		本文	0.40	25	1.09
	减小	PI	5.53	105	15.5
		本文	0.23	25	0.45



(a) 阶跃式变化



(b) 三角波变化

图 9 输入电压变化的仿真波形。

Fig.9 Waveforms of input voltage variation

策略可以确保对输入电压和负载同时变化下系统的稳定性。

5　实验结果

为验证所提控制策略的有效性，搭建了如图 10 所示实验线路。开关管选用 CREE 公司的 CAS120M12BM2 型 SiC MOSFET，dSPACE 用于产生 DAB 中开关管的 PWM 信号，硬件参数见表 1。

为了证明本文所提控制策略的有效性，在两种情况下进行实验验证。图 11—14 中，I_o 为 DAB 的输出电流，I_i 为 DAB 的输入电流，U_o 为 DAB 的输出电压，U_i 为 DAB 的输入电压。

1）情况 1：为了验证在负载侧出现不同程度扰动下，本文所提控制策略对 DAB 输出电压的稳定性和系统动态性能，在 t_1 和 t_2 时刻，负载阶跃变化。

初始时刻输出侧功率为 200 W，输出电压稳定在 100 V。在 t_1 时刻，负载从 200 W 跃变为 700 W，在 t_2 时刻，负载从 700 W 跃变至 200 W。分别采用本文控制、模型预测控制[36](model prediction control，MPC)和 PI 控制策略，实验波形如图 11 所示。

初始时刻输出侧功率为 1200 W，输出电压稳定在 100 V。在 t_1 时刻，负载从 1200 W 跃变为 200 W，在 t_2 时刻，负载从 200 W 跃变至 1200 W。本文所提控制、MPC 和 PI 控制策略，实验结果见图 12。

由图 11—12 实验结果可知，在负载由 200 W—700 W—200 W 时：与 MPC 控制相比，本文所提控制超调量降低近 86%，调节时间缩短近 75%，平均误差降低近 85%；与 PI 控制相比，本文所提控制超调量降低近 90%，调节时间缩短近 65%，平均误差降低近 93%。在负载由 1200 W—200 W—1200 W 时：与 MPC 控制相比，本文所提控制超调量降低近 89%，调节时间缩短近 40%，平均误差降低近 93%；而 PI 控制不能维持系统稳定。表 4 为不同控制策略下的系统性能具体参数对比。

从表 4 可以看出，相比与模型预测和 PI 控制器，本文所提控制器的调节时间、超调量、稳态误差同比降低近 35%~90%。尤其在大扰动下，本文所提控制仍能保持良好的性能。因此本文所提控制方法改善了系统的动态性能和跟踪精度，并且保证了系统在负荷扰动下的稳定性。

2）情况 2：为了验证在复合扰动下，所提控制策略对 DAB 的稳定性和动态性能的影响，在 t_1 和 t_2 时刻，输入电压和负载阶跃变化。

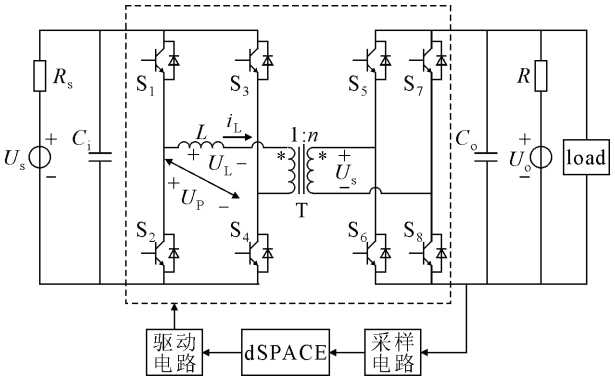
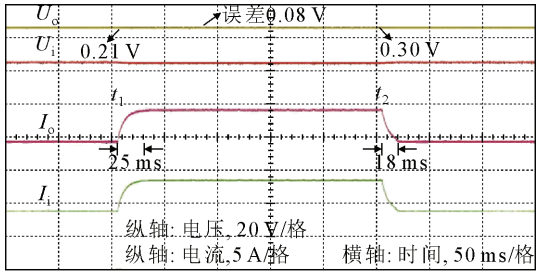
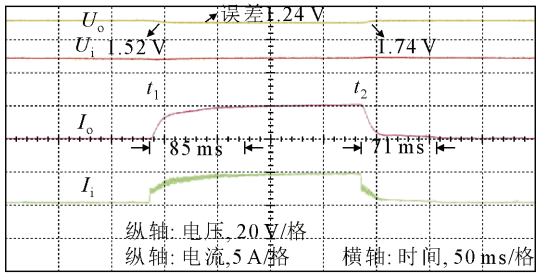


图 10　实验示意图

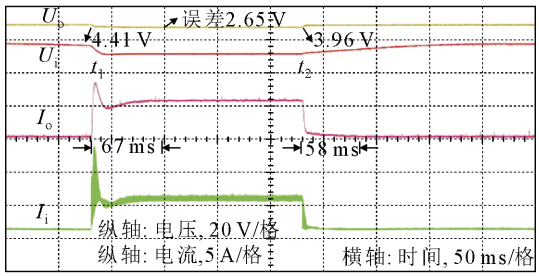
Fig.10　Schematic of experiment



(a) 本文所提控制策略



(b) MPC策略

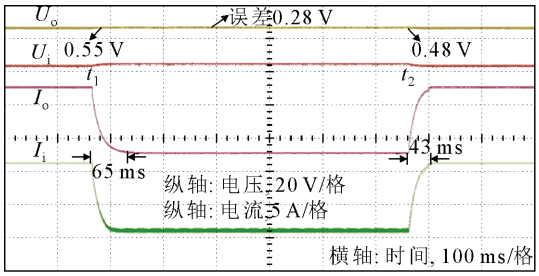


(c) PI策略

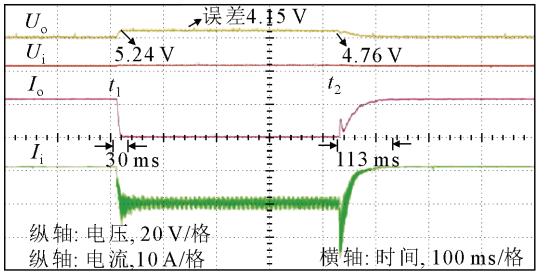
图 11　负载在 200 W—700 W—200 W 变化下实验波形

Fig.11　Experimental waveform with load varying from 200 W—700 W—200 W

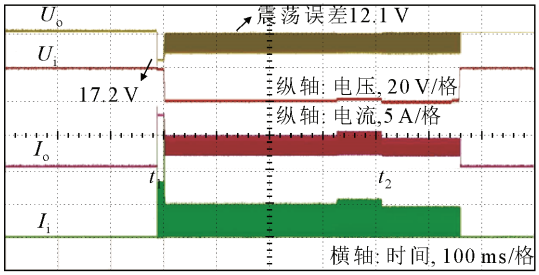
图 13 为 DAB 在输入电压与负载同趋势变化下的实验波形。输出侧额定功率为 200 W，此时输出电压稳定在 100 V。在 t_1 时刻，负载从 200 W 到 700 W 变化，输入电压从 100 V 变为 120 V。在 t_2 时刻，负载从 700 W 到 200 W 变化，输入电压从 120 V 变为 100 V。



(a) 本文所提控制策略



(b) MPC策略



(c) PI策略

图 12 负载在 1200 W—200 W—1200 W 变化下实验波形

Fig.12 Experimental waveform with load varying from 1200 W—200 W—1200 W

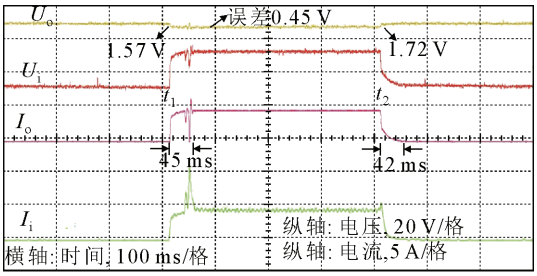
表 4 不同控制策略对比

Table 4 Comparison of different control strategies

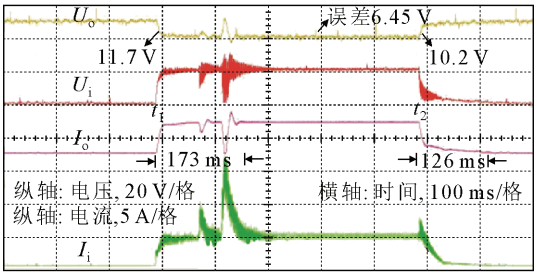
负载扰动/W	控制方式	调节时间 t_1/ms	调节时间 t_2/ms	超调量(t_1)/V	超调量(t_2)/V	平均误差/V
200—700—200	本文	25	18	0.21	0.30	0.18
	MPC	85	71	1.52	1.74	1.24
	PI	67	58	4.41	3.96	2.65
1200—200—1200	本文	65	43	0.55	0.48	0.28
	MPC	30	113	5.24	4.76	4.15
	PI			失稳		

图 14 为 DAB 在输入电压降低与负载不同趋势变化下的实验波形。输出侧额定功率为 200 W，此时输出电压稳定在 100 V。在 t_1 时刻，负载从 200 W 到 700 W 变化，输入电压从 100 V 变为 80 V。在 t_2 时刻，负载从 700 W 到 200 W 变化，输入电压从 80 V 变为 100 V。

由图 13、14 可以看出：本文所提控制策略在输入电压(±20 V)和负载(±500 W)同时出现扰动的情况下，平均调整时间在 35 ms 以内，输出电压偏差保持±0.7%以内，超调量小于±5%，能保证 DAB 的稳定运行，输出电压都能准确跟踪参考值，动态性能很快；而 PI 控制电压偏差超过 7%，同时输出电流波形紊乱，不能保证供电可靠性。



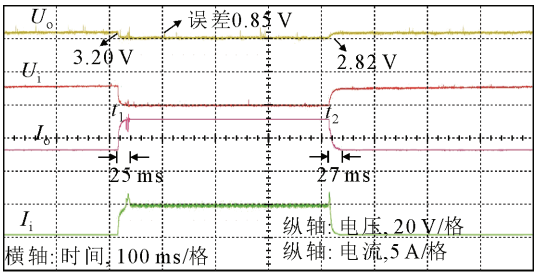
(a) 本文所提控制策略



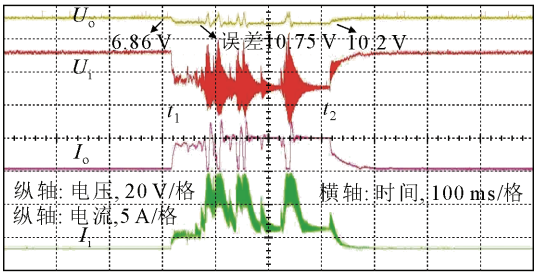
(b) PI控制策略

图 13 输入电压与负载同趋势变化下实验波形

Fig.13 Experimental waveforms at the same trend of input voltage and load



(a) 本文所提控制策略



(b) PI控制策略

图 14 输入电压与负载反趋势变化下实验波形

Fig.14 Experimental waveforms under counter-trend variation of input voltage and load

根据上述实验结果，本文提出的控制器有效地保证了准确的电压调节，具有快速的动态响应。在负载变化和不同工况下可以有效地保证 DAB 输出

电压精确快速地调节，并且保持运行稳定。动态性能和输出电压跟踪精度得到了有效改善，实验结果达到了理论预测的效果。

6　结论

本文提出了基于 KF 的 DDPG 控制策略，用于提高 DAB 变换器的自适应性和自稳定性得出以下结论：

1）本文所提控制策略具有智能地改善系统的动态响应和跟踪精度的特点，对运行环境的变化具有鲁棒性，可以保证系统在运行环境变化后的稳定运行，提高系统鲁棒性。

2）本文所提控制策略可以在不同工况下准确快速地调节输出电压，并且能够保持系统稳定运行。随着未来新能源占比逐渐提升，直流微电网面临的大扰动运行情况将更加频繁，本文提出的控制策略能够很好地应对电网系统中出现的不同种类的波动，提高系统的自稳定性和自适应性，保障电网稳定运行，具有广泛的应用前景。

附录见本刊网络版(http://hve.epri.sgcc.com.cn)。

参考文献　References

[1] 韩肖清，李廷钧，张东霞，等. 双碳目标下的新型电力系统规划新问题及关键技术[J]. 高电压技术，2021，47(9)：3036-3046.
HAN Xiaoqing, LI Tingjun, ZHANG Dongxia, et al. New issues and key technologies of new power system planning under double carbon goals[J]. High Voltage Engineering, 2021, 47(9): 3036-3046.

[2] 任海泉，贾燕冰，田　丰，等. 含多微能网的城市能源互联网优化调度[J]. 高电压技术，2022，48(2)：554-564.
REN Haiquan, JIA Yanbing, TIAN Feng, et al. Urban energy internet optimal dispatch with multi-micro energy network[J]. High Voltage Engineering, 2022, 48(2): 554-564.

[3] 肖　峻，莫少雄. 柔性直流配电网的 *N*–1 安全域[J]. 高电压技术，2021，47(10)：3626-3634.
XIAO Jun, MO Shaoxiong. *N*–1 security region of flexible DC distribution network[J]. High Voltage Engineering, 2021, 47(10): 3626-3634.

[4] 韦延方，王志杰，王　鹏，等. 基于限流电抗电压的柔性直流配电网故障保护方法[J]. 高电压技术，2022，48(12)：5068-5079.
WEI Yanfang, WANG Zhijie, WANG Peng, et al. Protection method for flexible DC distribution network faults based on current limiting reactance voltage[J]. High Voltage Engineering, 2022, 48(12): 5068-5079.

[5] 魏　斌，韩肖清，李　雯，等. 融合多场景分析的交直流混合微　电网多时间尺度随机优化调度策略[J]. 高电压技术，2020，46(7)：2359 -2369.
WEI Bin, HAN Xiaoqing, LI Wen, et al. Multi-time scale stochastic optimal dispatch for AC/DC hybrid microgrid incorporating multi-scenario analysis[J]. High Voltage Engineering, 2020, 46(7): 2359-2369.

[6] 盛万兴，段　青，王　良，等. 基于多代理协调机制的能量路由器群组与配电网综合规划[J]. 高电压技术，2021，47(1)：1-11.
SHENG Wanxing, DUAN Qing, WANG Liang, et al. Comprehensive planning for energy routers and distribution network based on multi-agent system coordination mechanism[J]. High Voltage Engineering, 2021, 47(1): 1-11.

[7] 曾　嵘，赵宇明，赵　彪，等. 直流配用电关键技术研究与应用展望[J]. 中国电机工程学报，2018，38(23)：6790-6802.
ZENG Rong, ZHAO Yuming, ZHAO Biao, et al. A prospective look on research and application of DC power distribution technology[J]. Proceedings of the CSEE, 2018, 38(23): 6791-6801.

[8] 郭东鑫，贾燕冰，任春光，等. 基于虚拟频率的隔离型双向 AC/DC 变换器最小电流应力控制策略[J]. 高电压技术，2021，47(8)：2914-2922.
GUO Dongxin, JIA Yanbing, REN Chunguang, et al. Minimum current stress control strategy of isolated bidirectional AC/DC converter based on virtual frequency[J]. High Voltage Engineering, 2021, 47(8): 2914-2922.

[9] 李　凯，赵争鸣，袁立强，等. 面向交直流混合配电系统的多端口电力电子变压器研究综述[J]. 高电压技术，2021，47(4)：1233-1250.
LI Kai, ZHAO Zhengming, YUAN Liqiang, et al. Overview on research of multi-port power electronic transformer oriented for AC/DC hybrid distribution grid[J]. High Voltage Engineering, 2021, 47(4): 1233-1250.

[10] 魏　星，朱信舜，葛　健，等. 级联型电力电子变压器并联运行的改进下垂控制策略[J]. 高电压技术，2021，47(4)：1274-1282.
WEI Xing, ZHU Xinshun, GE Jian, et al. Improved droop control strategy for parallel operation of cascaded power electronic transformers[J]. High Voltage Engineering, 2021, 47(4): 1274-1282.

[11] 赵　峰，甘延奇，陈小强，等. 基于频域分析的双有源桥串联谐振变换器的设计与闭环控制[J]. 高电压技术，2022，48(11)：4557-4567.
ZHAO Feng, GAN Yanqi, CHEN Xiaoqiang, et al. Design of dual active bridge series resonant converter based on frequency domain analysis and closed-loop control[J]. High Voltage Engineering, 2022, 48(11): 4557-4567.

[12] 杨向真，孔令浩，杜　燕，等. T 型三电平双有源桥变换器 ZVS 优化策略[J]. 高电压技术，2022，48(5)：1913-1923.
YANG Xiangzhen, KONG Linghao, DU Yan, et al. ZVS optimization strategy of T-type three-level dual active bridge converter[J]. High Voltage Engineering, 2022, 48(5): 1913-1923.

[13] KUNG S, LIU D L, PERERA C, et al. Exploiting buck–boost duality in dual active bridge modular multilevel converters to achieve high DC step ratios[J]. High Voltage, 2021, 6(3): 495-513.

[14] KAKIGANO H, MIURA Y, ISE T. Low-voltage bipolar-type DC microgrid for super high quality distribution[J]. IEEE Transactions on Power Electronics, 2010, 25(12): 3066-3075.

[15] MAHMOUD M S, UR RAHMAN M S, SUNNI F M A L. Review of microgrid architectures–a system of systems perspective[J]. IET Renewable Power Generation, 2015, 9(8): 1064-1078.

[16] BRANDO G, DEL PIZZO A, MEO S. Model-reference adaptive control of a dual active bridge dc-dc converter for aircraft applications[C]//2018 International Symposium on Power Electronics, Electrical Drives, Automation and Motion (SPEEDAM). Amalfi, Italy: IEEE, 2018: 502-506.

[17] KOCH G G, QUEIROZ S S, RECH C, et al. Design of a robust PI controller for a dual active bridge converter[C]//2016 12th IEEE International Conference on Industry Applications (INDUSCON). Curitiba, Brazil: IEEE, 2016: 1-6.

[18] HOU N, SONG W S, WU M Y. Minimum-current-stress scheme of dual active bridge DC–DC converter with unified phase-shift con-

trol[J]. IEEE Transactions on Power Electronics, 2016, 31(12): 8552-8561.

[19] MENG L X, DRAGICEVIC T, VASQUEZ J C, et al. Tertiary and secondary control levels for efficiency optimization and system damping in droop controlled DC–DC converters[J]. IEEE Transactions on Smart Grid, 2015, 6(6): 2615-2626.

[20] DU Z, TOLBERT L M, CHIASSON J N, et al. Reduced switching-frequency active harmonic elimination for multilevel converters[J]. IEEE Transactions on Industrial Electronics, 2008, 55(4): 1761-1770.

[21] 侯 聂，宋文胜，武明义. 全桥隔离 DC/DC 变换器的直接功率控制方法[J]. 电力系统自动化，2016，40(17)：204-209.
HOU Nie, SONG Wensheng, WU Mingyi. Direct power control scheme of full-bridge isolated DC/DC converters[J]. Automation of Electric Power Systems, 2016, 40(17): 204-209.

[22] 朱海猛. 带恒功率负载的 Buck-Boost 变换器数字解耦控制器设计[D]. 南宁：广西大学，2017.
ZHU Haimeng. Zhe design of digital decoupling controller for buck-boost converter with constant power load[D]. Nanning, China: Guangxi University, 2017.

[23] 吴 宇，皇甫宣耿，张 琳，等. 大扰动 Buck-Boost 变换器的鲁棒高阶滑模控制[J]. 中国电机工程学报，2015，35(7)：1740-1748.
WU Yu, HUANGFU Yigeng, ZHANG Lin, et al. A robust high order sliding mode for buck-boost converters with large disturbances[J]. Proceedings of the CSEE, 2015, 35(7): 1740-1748.

[24] KRISHNAMOORTHI. S, SHYAM G K. Review of deep reinforcement learning-based recommender systems[C]//. 2022 Third International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), Bengaluru, India: ICSTCEE 2022: 1-12.

[25] MADHUKAR P S, PRASAD L B. State estimation using extended Kalman filter and unscented Kalman filter[C]//2020 International Conference on Emerging Trends in Communication, Control and Computing (ICONC3). Lakshmangarh, India: IEEE, 2020: 1-4.

[26] LI S Z, LI Y, SUN Y, et al. State of charge estimation for lithium-ion batteries based on adaptive fractional extended Kalman filter[C]// 2020 IEEE 9th Data Driven Control and Learning Systems Conference (DDCLS). Liuzhou, China: IEEE, 2020: 266-271.

[27] COSTINETT D, MAKSIMOVIC D, ZANE R. Design and control for high efficiency in high step-down dual active bridge converters operating at high switching frequency[J]. IEEE Transactions on Power Electronics, 2013, 28(8): 3931-3940.

[28] LIU X, ZHU Z Q, STONE D A, et al. Novel dual-phase-shift control with bidirectional inner phase shifts for a dual-active-bridge converter having low surge current and stable power control[J]. IEEE Transactions on Power Electronics, 2017, 32(5): 4095-4106.

[29] ZHANG Y B, XIAO M L, WANG Z H, et al. Robust three-stage unscented Kalman filter for Mars entry phase navigation[J]. Information Fusion, 2019, 51: 67-75.

[30] LIU F, MA J, SU W X, et al. SOC estimation based on data driven exteaded Kalman filter algorithm for power battery of electric vehicle and plug-in electric vehicle[J]. Journal of Central South University, 2019, 26(6): 1402-1415.

[31] ZHANG X W, ZHAO T Y, GUI X G. Force ripple suppression of permanent magnet linear synchronous motor based on fuzzy adaptive Kalman filter[C]//2021 13th International Symposium on Linear Drives for Industry Applications (LDIA). Wuhan, China: IEEE, 2021: 1-5.

[32] HUANG L, FENG X, ZHANG C, et al. Deep reinforcement learning-based joint task offloading and bandwidth allocation for multi-user mobile edge computing[J]. Digital Communications and Networks, 2019, 5(1): 10-17.

[33] 张 博，刘 绚，于宗超，等. 基于人工智能的电力系统网络攻击检测研究综述[J]. 高电压技术，2022，48(11)：4413-4426.
ZHANG Bo, LIU Xuan, YU Zongchao, et al. Review on artificial intelligence-based network attack detection in power systems[J]. High Voltage Engineering, 2022, 48(11): 4413-4426.

[34] WANG Y D, SUN J, HE H B, et al. Deterministic policy gradient with integral compensator for robust quadrotor control[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2020, 50(10): 3713-3725.

[35] ZUO G Y, DU T T, LU J H. Double DQN method for object detection[C]// 2017 Chinese Automation Congress (CAC). Jinan, China: IEEE, 2017: 6727-6732.

[36] VAFAMAND N, YOUSEFIZADEH S, KHOOBAN M H, et al. Adaptive TS fuzzy-based MPC for DC microgrids with dynamic CPLS: Nonlinear power observer approach[J]. IEEE Systems Journal, 2019, 13(3): 3203-3210.



武 涵
2000—，男，硕士生
主要从事微电网运行与控制、电力电子接口变换器的控制研究工作
E-mail: wuhanemails@163.com

WU Han



贾燕冰(通信作者)
1980—，女，博士，教授
主要从事微电网运行控制、电力系统优化运行、电力系统稳定分析、综合能源系统优化运行和风险评估的研究工作
E-mail: jybtyut@163.com

JIA Yanbing
Ph.D., Professor
Corresponding author

收稿日期 2023-01-28 修回日期 2023-05-24 编辑 陈 蔓

附录 A

表 B1　DDPG 算法代码

Table B1　DDPG algorithm code

DDPG 算法
随机初始化在线 Critic-Network $Q(s,a\mid\theta^{Q})$ 和在线 Actor-Network $\mu(s\mid\theta^{\mu})$ 的网络参数 θ^{Q} 和 θ^{Q}
初始化目标网络 Q'及其权重 $\theta^{Q'}\leftarrow\theta^{Q},\theta^{\mu'}\leftarrow\theta^{\mu}$
初始化经验池 R
for episode =1, M do
初始化随机探索噪声 N
得到初始观测状态 s_1
for t=1, T do
根据当前的策略和探索噪音选择动作 $a_t=\mu\left(s_t\mid\theta^{\mu}\right)+\mathcal{N}_t$
根据动作 a_t 与环境进行交互产生奖励 r_t ，新的状态 s_{t+1}
将元组数据 (s_t,a_t,r_t,s_{t+1})存至经验池 R
从 R 中随机采样 N 个元组数据(s_i, a_i, r_i, s_{i+1})
令 $y_i=r_i+\gamma Q'\left(s_{i+1},\mu'\left(s_{i+1}\mid\theta^{\mu'}\right)\mid\theta^{Q'}\right)$
最小化损失来更新当前 Critic 网络: $L=\frac{1}{N}\sum_i\left(y_i-Q\left(s_t,a_i\mid\theta^{Q}\right)\right)^2$
通过样本的策略梯度来更新当前 Actor 网络: $\nabla_{\theta^{\mu}}J\approx\frac{1}{N}\sum_t\nabla_a Q\left(s,a\mid\theta^{Q}\right)\Big\vert_{s=s_t,a=\mu(s_t)}\nabla_{\theta^{\mu}}\mu\left(s\mid\theta^{\mu}\right)\Big\vert_{s_t}$
更新目标网络: $\theta^{Q'}\leftarrow\tau\theta^{Q}+(1-\tau)\theta^{Q'}$，$\theta^{\mu'}\leftarrow\tau\theta^{\mu}+(1-\tau)\theta^{\mu'}$
end for
end for

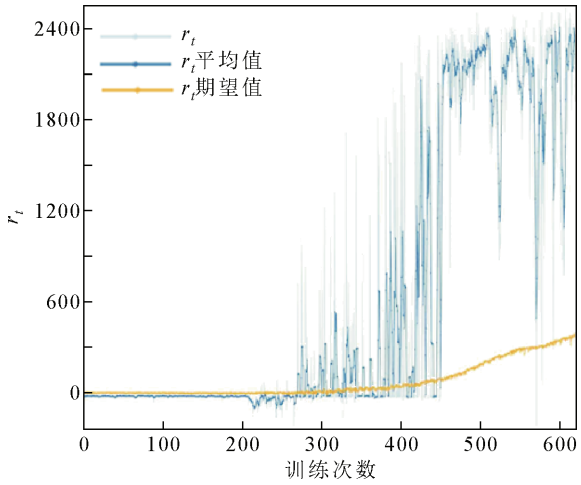


图 A1　控制器训练曲线

Fig.A1　The training curve of the controller.