Design and Experimental Validation of a Model-Free Controller for Beam Stabilization in Adaptive Optics Systems

Sicheng Guo, Tao Cheng, Kangjian Yang, Lingxi Kong, Chunxuan Su, Shuai Wang, and Ping Yang

***Abstract*—Stabilization of optical beams has always been a key factor affecting the performance of many optical systems. The Adaptive optics (AO) beam stabilization system requires further development to cope with increasingly complex application scenarios and challenges. Motivated by this, a new filter-based off-policy policy iteration (FB-OPPI) control scheme is proposed and experimentally verified in this paper to provide AO systems with a flexible beam stabilization method. The FB-OPPI is based on the policy iteration, a model-free controller design principle. To address the challenges such as convergence speed, data requirements and control stability that it faces in practice, we have proposed an implicit state reconstruction method based on the Kalman filter and introduced the adaptive transverse filter technology. Additionally, the off-policy learning mechanism is deployed to simplify the optimization process. An AO beam stabilization system was constructed to verify the effectiveness of the proposed method. Experimental results show that the FB-OPPI method features simple design and fast training, releases the requirement for additional sensors or model recognition. The FB-OPPI method is superior to traditional integral controllers, effectively handling high-frequency narrowband and complex beam jitters. Despite not requiring model identification, it is on par with the advanced Linear Quadratic Gaussian (LQG) control.**

***Index Terms*—Adaptive optics system, beam stabilization, model-free control, policy iteration.**

I. Introduction

BEAM stabilization is a ubiquitous problem that occurs in most of the existing optical systems. Pointing errors or beam fluctuations due to beam jitter adversely affect the system's ability to obtain the optical diffraction limit, improve the optical communication capability and focus the beam quality, making it a pressing concern in in astronomical telescopes, free-space optical communication and precise laser beam pointing and many other applications [1], [2], [3], [4]. The adaptive optics (AO) [5], [6], [7], [8], recognized as an effective technique for detecting and correcting wavefront aberrations, has been extensively utilized to mitigate beam jitter induced by atmospheric turbulence. However, beam jitter may also originate from within the system itself. Mechanical vibrations could be introduced by drive motors, pumps or moving components in the devices, etc. These vibrations are typically narrowband and have higher center frequencies [9], [10], presenting a more complex challenge than the broadband, low-frequency atmospheric turbulence. According to the statistics in [3], vibrations on NAOS have caused a decrease in Strehl ratio of up to 25%.

AO systems typically use photoelectric sensors (such as cameras) to detect changes in the position of the light spot as feedback, and calculate control actions based on the error feedback using classical integral control law. Although the integral controller benefits from simple structure and strong stability, the control bandwidth is limited by the system loop delay and limited sampling rate. This makes it difficult to achieve satisfactory control effects on high-frequency narrowband and complex beam jitter [11]. To meet the system's strict requirements for beam stabilization, many advanced control methods that follow model-based design approaches have been developed. One of the most successful implementations is the Linear Quadratic Gaussian (LQG) control, which designs a Kalman filter based on the system dynamics and noise characteristics and then solves the Algebraic Riccati Equation (ARE) to obtain the optimal control gain. It could provide optimal control in the sense of minimizing the residual jitter variance. The LQG control has been successfully applied and verified in various AO systems, demonstrating superior control performance over traditional integrators [12], [13], [14], [15]. Other methods, such as robust control, adaptive control, and model predictive control, have also been studied accordingly [16], [17], [18], [19]. However, these aforementioned model-based methods are mainly limited by their dependence on accurate a priori models in practical applications. LQG-type controllers cannot estimate or compensate for jitter signals not included in the model. In the worst-case scenario, the propagation and accumulation of unmodeled system dynamics and model errors can lead to

Received 3 September 2024; revised 27 September 2024; accepted 28 September 2024. Date of publication 30 September 2024; date of current version 28 November 2024. This work was supported by the National Natural Science Foundation of China under Grant 62005285 and Grant 62105336. *(Corresponding author: Ping Yang.)*

Sicheng Guo and Lingxi Kong are with the National Laboratory on Adaptive Optics, Chengdu 610209, China, also with the the Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu 610209, China, and also with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: guosicheng20@mails.ucas.ac.cn; konglingxi18@mails.ucas.ac.cn).

Tao Cheng, Kangjian Yang, Chunxuan Su, Shuai Wang, and Ping Yang are with the National Laboratory on Adaptive Optics, Chengdu 610209, China, and also with the Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu 610209, China (e-mail: chengtao@ioe.ac.cn; yangkangjian@ioe.ac.cn; suchunxuan16@mails.ucas.ac.cn; wangshuai@ioe.ac.cn; pingyang2516@163.com).

Digital Object Identifier 10.1109/JPHOT.2024.3471827

undesirable system behavior, causing a deterioration of system performance.

Many studies have attempted to alleviate this problem from the model identification side. Online estimation of jitter models by fitting the power spectrum density (PSD) was studied in [15] and [20]. However, recognizing the jitter model in a closed-loop system typically requires the establishment of Pseudo-Open Loop (POL) data, which still necessitates knowledge of the AO system's delay and TTM dynamics. Subspace identification methods have been widely used to identify AO system models and have been verified in experimental systems [21], [22], [23], [24]. With the development of deep learning technology, deep neural networks have also been used for reconstruction and prediction in AO systems [25], [26], [27]. However, it should be noted that accurate model identification is merely a prerequisite for model-based beam stabilization control, the model parameters will be used to complete the subsequent controller design process. Considering that the model identification process itself requires a large amount of data and iterative computations, and the design of the optimal controller is based on the solution of complex algebraic equations, the separation of model identification and controller design might increase the computational complexity of the system.

Another attempt is to design a model-free approach to achieve (near) optimal control of beam stabilization for AO systems. Reinforcement Learning (RL) [28] addresses optimization challenges by interacting with the real environment and learning from trial and error. It bypasses the system identification process and iteratively learns the optimal controllers directly from measured data. Many studies have already been conducted on the use of reinforcement learning techniques to compensate for wavefront or beam jitter in AO systems [29], [30], [31], [32]. Although better control performance has been achieved, the convergence of deep neural network controllers requires a large amount of time and data, which is contrary to the control demands of real systems. We therefore consider the development of Policy Iteration (PI) [33] technique, which interprets RL as a standard optimal control problem and solves it numerically.

Applying the PI method to practical AO beam stabilization faces three challenges. The first is the partially observable problem [32]. In addition to model parameters, the complete state information of the AO system and jitter is usually unknown, and the available information consists only of noisy jitter errors and control actions. Inferring the behavior patterns of jitter and the AO system from this information is crucial for precise control. The second challenge lies in the difficulty of solving the mathematical expectation in a stochastic environment [34], especially in the absence of accurate system models and state. This requires us to design a fast, causal solution to reduce the computational burden when the algorithm is actually deployed. The third challenge is how to reduce the data samples required by the algorithm while improving control stability at the same time.

Motivated by the above discussion, this paper proposes and experimentally verifies a new filter-based off-policy policy iteration (FB-OPPI) method. To the best of our knowledge, this is the first time PI has been used to solve beam stability problems in AO systems. To address the first challenge, we have developed an implicit state representation method based on the Kalman filter, which can successfully separate unknown model parameters. Unlike [35], [36], our algorithm does not rely on the assumption of complete system state information, nor does it require the construction of an actual state observer as in [37], [38]. Compared with the Output Feedback (OPFB) method developed in an ideal deterministic environment [39], [40], [41], [42], our method can work properly in more practical situations where stochastic noise is taken into consideration. At the same time, there is no need to estimate the system noise characteristics as in [38]. We have circumvented the computation of the mathematical expectation by designing an adaptive transversal filter to tackle the second challenge. Compared to the work in [43], we do not need to estimate the expectation with the numerical average of repeatedly generated trajectories. The third challenge is addressed by introducing an off-policy learning mechanism. The target controller can reuse the data sampled by an initialized controller for learning and converge after a few iterative steps. The target controller does not participate in the closed loop before convergence, thus ensuring the stability of the system during the learning process. The effectiveness of the FB-OPPI method is successfully verified in an actual AO beam stabilization system. The advantages of this FB-OPPI method for AO beam stabilization systems are:

1) The method releases the dependency on model identification, state observation, and additional sensors in systems, allowing for effective application across different systems and environments.
2) It provides an automatic and flexible controller design with satisfactory data sample requirements and convergence speed.
3) The performance of the beam stabilization control far exceeds that of conventional integrator and is comparable to the advanced model-based LQG control.

The structure of this paper is organized as follows: Section II outlines the principles of the classic AO beam stabilization system, representing it in state-space form and providing a general solution to the control problem. The filter-based off-policy policy iteration control scheme is proposed in Section III, and an experimental system for beam stabilization control is established in Section IV to verify the effectiveness of the algorithm. Finally, we conclude the paper in Section V.

II. State-Space Model and Problem Statement

A. Classic Beam Stabilization Control System

For an AO system, a classical beam stabilization control structure is shown in Fig. 1, which mainly consists of a wavefront sensor (WFS), a tip-tilt mirror (TTM), a high-voltage amplifier (HVA), a digital-to-analog converter (DAC) and the controller. The incident beam with jitter phase ϕ^{vib} is reflected by the TTM with phase ϕ^{cor}, and the residual phase is ϕ^{res}. The compensated beam is then focused and exposed on the Charge-coupled Device (CCD) camera of the WFS, so the jitter error e can be obtained by calculating the centroid of the far-field spot. The control voltage u calculated according to e is transformed into an analog signal through zero-order hold, and finally drives the TTM after

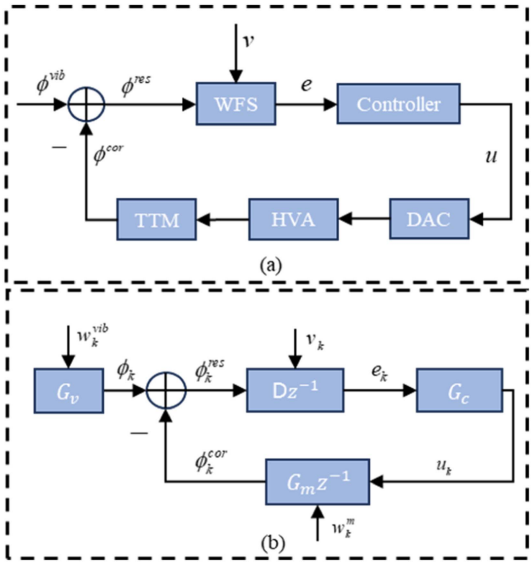

Fig. 1. The classic AO beam stabilization system: (a) Schematic diagram, (b) control block diagram.

high-voltage amplification, which makes the TTM compensate for the jitter error.

Assuming that The WFS is linear and the response relationship between phase and measurement can be described by a constant matrix D. The system experiences a two-frame delay due to CCD read-out and slope calculation and there is no coupling between the two axes of the TTM. The jitter control system could be simplified into the block diagram shown in Fig. 1(b), and the complete state-space model containing the jitter dynamics and the system model is of the form

$$\begin{cases} X_{k+1} = \begin{bmatrix} A_m & 0 \\ 0 & A_{vib} \end{bmatrix} X_k + \begin{bmatrix} B_m \\ 0 \end{bmatrix} u_{k-1} + \begin{bmatrix} w_k^m \\ w_k^{vib} \end{bmatrix}, \\ e_k = D \begin{bmatrix} -C_m & C_{vib} \end{bmatrix} X_{k-1} + v_k \end{cases} \quad (1)$$

where $X_k = [(x_k^m)^T \quad (x_k^{vib})^T]^T$, x_k^m and x_k^{vib} are the states of the AO system and unknown jitter, respectively. A_m, B_m and C_m are the unknown system matrices of the plant, A_{vib} and C_{vib} are the unknown matrices of the jitter dynamics. w_k^m, w_k^{vib} and v_k are independent Gaussian white noises. Generally, the following assumptions are easily satisfied: (a) A_{vib} is Hurwitz; (b) (A_m, B_m) is controllable; (c) (A_m, C_m) and (A_{vib}, C_{vib}) are observable. According to [44], the system model (1) could be brought into a standard delay-free form by using further state augmentation to have

$$\begin{cases} X_{k+1} = AX_k + Bu_k + w_k \\ e_k = CX_k + v_k \end{cases}, \quad (2)$$

and such a model guarantees stabilizability and observability. The system model (2) that contains both deterministic and stochastic components might be problematic for joint identification and the separate identification for subsystem is a cumbersome process [21].

B. Problem Statement

For the beam stabilization in system (2), the problem is to determine an optimal feedback controller that minimizes the following infinite time horizon performance index

$$J = E\left[\sum_{i=k}^{\infty} \gamma^{i-k} \left(X_i^T \bar{Q} X_i + u_i^T R u_i\right)\right], \quad (3)$$

where $\gamma \in (0, 1)$ is the discount factor that keeps the performance index bounded. $Q \geq 0$ and $R > 0$ are weighting matrices that regulate control performance and control energy consumption, and $\bar{Q} = C^T Q C$.

Notice that the jitter control problem now has the form of a stochastic linear quadratic tracking problem. The certainty equivalence principle [45] shows the optimal control policy has the same form as the solution of an equivalent deterministic system after replacing all random variables with their estimates, which is of the form $u_k = -K^* X_k$, where $K^* = (R + \gamma B^T P B)^{-1} \gamma B^T P A$ is the optimal feedback control gain derived from the following ARE [39]

$$P = \bar{Q} + \gamma A^T P A - \gamma^2 A^T P B \left(R + \gamma B^T P B\right)^{-1} B^T P A. \quad (4)$$

However, it is also difficult to get access to complete state information in practical systems. According to the separation principle [46], the optimal control policy under incomplete state information can be obtained by replacing $X_k$with its estimate $\hat{X}_k$ and the resulting solution is

$$u_k = \pi^*\left(\hat{X}_k\right) = -K^* \hat{X}_k. \quad (5)$$

It is clear that the solution to the control problem (3) relies on complete system model information and state estimation. The methods based on system identification require accurate fitting or solving of the model parameters before the feedback control gain is calculated. This not only increases the computational cost but also propagates the model error further into the control gain, thus affecting the stability of the entire control process. Therefore, in the following section, an attempt is made to use a model-free approach to directly optimize the controller to circumvent the errors brought about by model identification.

III. Filter-Based Off-Policy Policy Iteration Scheme

The designed FB-OPPI control scheme is shown in Fig. 2. According to the state space model (2) and relationship between error feedback and control action, the AO jitter control system can be divided into two parts: unknown system dynamics and the proposed model-free controller. The external jitter system and the components of the AO system are treated as an augmented unknown system. The controller is required to calculate the u_k based on the available historical jitter error and control action sequences, thereby driving the TTM for beam stabilization.

Therefore, a data buffer has been established to implement experience replay, which is used to break the correlation between data while facilitating the reuse of experience data [47], [28]. The concept of the Kalman filter is then utilized to implicitly reconstruct the system state from noisy jitter errors and control

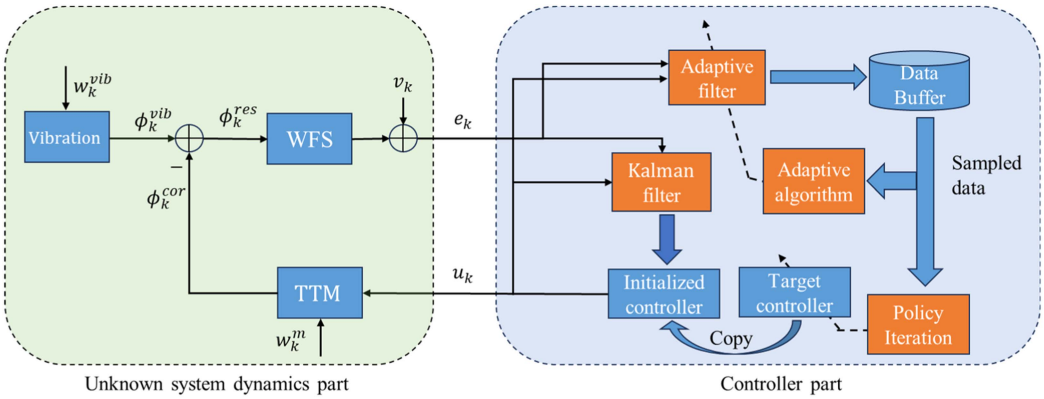

Fig. 2. Block diagram of the proposed FB-OPPI control scheme.

actions. Implicit reconstruction means that there is no need to actually solve for the Kalman filter parameters. An adaptive filter is employed to estimate future jitter errors from historical data and complete the dataset. Using such a composed dataset can avoid the calculation of mathematical expectations and repetitive trajectory sampling. There are two controllers under the off-policy learning mechanism. One is an initialized controller for ensuring system closed-loop stability and data sampling (e.g., an integrator), and the other is a target controller that updates online according to the PI algorithm. Once the target controller has iteratively converged, it will replace the initializing controller to achieve better beam stabilization control.

This FB-OPPI control scheme infers the dynamic evolution mechanism by interacting with the augmented unknown system (2). As a result, the controller design process does not necessitate concern for specific model parameters or state information of the unknown systems. Oriented towards the practical needs and challenges of beam stabilization, this scheme is characterized by its ease of deployment, rapid convergence, and less data requirements. In this section, we will introduce the working principle of the model-free control scheme.

A. Implicit State Representation Based on the Principle of Kalman Filter

The system state reflects the evolution characteristics of beam jitter and the response characteristics of the AO system, which plays an important role in state feedback control. Considering that the complete system state is typically unknown in practice, the first and foremost task is to reconstruct the expression for the state estimate based on the available information. The expression for the state estimation $\hat{X}_k$ under a steady-state Kalman filter [48] is first given as

$$\begin{cases} \hat{X}_k = \bar{X}_k + L\left(e_k - C\bar{X}_k\right) \\ \bar{X}_k = A\hat{X}_{k-1} + Bu_{k-1} \end{cases}, \tag{6}$$

where L is the steady-state Kalman gain. Although in most studies, directly solving (6) can directly obtain an estimate of the state. The requirement for obtaining the Kalman gain L from another ARE is difficult to be achieved when the system model parameters and noise covariance matrix are unknown. The core of the problem at hand is how to provide a new form of the controller using only the available sequences of jitter error e_k and control action u_k, instead of the unknown system state. The control performance should also be maintained.

Noticing that (6) includes the e_k and u_k we need, but the information contained in these two terms alone is far from sufficient to reconstruct the system state. Hence expanding it over the time interval [k-N, k] yields

$$\begin{aligned} \hat{X}_k = {} & A_L^N \hat{X}_{k-N} \\ & + \left[B_L A_L B_L \cdots A_L^{N-1} B_L\right] \begin{bmatrix} u_{k-1} \\ u_{k-2} \\ \vdots \\ u_{k-N} \end{bmatrix} \\ & + \left[L A_L L \cdots A_L^{N-1} L 0\right] \begin{bmatrix} e_k \\ \vdots \\ e_{k-N+1} \\ e_{k-N} \end{bmatrix}, \end{aligned} \tag{7}$$

where $A_L = A - LCA$, $B_L = B - LCB$.

Considering that for a stable Kalman filter, the state estimate will asymptotically converge to a steady-state solution, indicating the matrix A_L is Hurwitz. Accordingly, when the data length N is large enough, the first term in (7) will become small enough to be negligible, allowing the state estimate to be represented as a linear combination of input-output data that

$$\hat{X}_k = \Gamma \bar{z}_k \tag{8}$$

where $\Gamma = [L, \cdots, A_L^{N-1}L, 0, B_L, \cdots, A_L^{N-1}B_L]$ and $\bar{z}_k = [e_k^T, \ldots, e_{k-N}^T, u_{k-1}^T, \ldots, u_{k-N}^T]^T$. The optimal controller could also change from the state feedback form in (5) to the following output feedback form

$$u_k = -K^* \hat{X}_k = -K^* \Gamma \bar{z}_k = -\bar{K}^* \bar{z}_k. \tag{9}$$

Based on this linear representation (8), which separates the state estimate into the form of a product of the unknown coefficient matrix Γ and the available historical data vector $\bar{z}_k$, we can rewrite the performance index, which concerns the unknown

system state, as a function of the available data sequence $\bar{z}_k$ that:

$$\bar{J} = E\left[\sum_{i=k}^{\infty}\gamma^{i-k}\left(\bar{z}_i^T\tilde{Q}\bar{z}_i + u_i^T R u_i\right)\right], \tag{10}$$

where $\widetilde{Q} = \Gamma^T\bar{Q}\Gamma$.

This approach differs from the OPFB in that the introduced Kalman filter enables the system state to be reconstructed from noisy input/output measurements. Moreover, the state estimation is implicitly integrated into the controller optimization process, eliminating the need for separate iterative solving of the state estimation [37] or offline training of the neural network [38].

B. Bellman Equation Based on Adaptive Transversal Filter

The goal of policy iteration is to evaluate the performance of the current controller based on the performance index thereby optimizing it properly. In order to establish the relationship between the performance index and the control policy, the performance index can be defined as a value function with respect to the historical data $\bar{z}_k$ and the current control action u_k according to (10) as

$$\begin{aligned} Q^{\pi}(\bar{z}_k, u_k) &= E\left[\sum_{i=k}^{\infty}\gamma^{i-k}\left(\bar{z}_i^T\tilde{Q}\bar{z}_i + u_i^T R u_i\right)\right] \\ &= U_k + \gamma E\left[Q^{\pi}(\bar{z}_{k+1}, u_{k+1})\,|e_{0,\ldots,k}, u_{0,\ldots,k}\right], \end{aligned} \tag{11}$$

where the superscript π indicates the control policy followed by the current output control action u_k. The value function in (11) conforms to the form of the Bellman equation [28], which reflects the relationship between the cost functions of two adjacent moments. However, all the information that can be obtained by the controller at time k is the jitter errors $e_{0,\cdots,k}$ and control actions $u_{0,\cdots,k}$. The value function $Q^{\pi}(\bar{z}_{k+1}, u_{k+1})$ for the next moment cannot be determined due to the presence of stochastic noises, and can only be given in the form of mathematical expectations.

To avoid complex calculations or repeated sampling due to mathematical expectations, an adaptive transversal filter is adopted here, which uses the data available at a given time k to predict the jitter error at the next moment. The structure of the adaptive transversal filter is shown in Fig. 3.

According to Fig. 3, the prediction of e_{k+1} can be written as

$$\hat{e}_{k+1} = \sum_{i=1}^{N-1}(\alpha_i e_{k-i+1}, \beta_i u_{k-i+1}), \tag{12}$$

here N is the filter order. Considering the requirement for fast convergency speed, the coefficients α_i and β_i are updated using the recursive least squares (RLS) [49] algorithm. Then the prediction (12) is substituted into (11) to make all variables in the equation deterministic for a given moment k and the available data vector $\bar{z}_k$. The Bellman equation is then rewritten as

$$Q^{\pi}(\bar{z}_k, u_k) = U_k + \gamma Q^{\pi}(\hat{z}_{k+1}, \pi(\hat{z}_{k+1})), \tag{13}$$

where $\hat{z}_{k+1} = [\hat{e}_{k+1}^T, e_k^T, \ldots, e_{k+1-N}^T, u_k^T, \ldots, u_{k+1-N}^T]^T$.

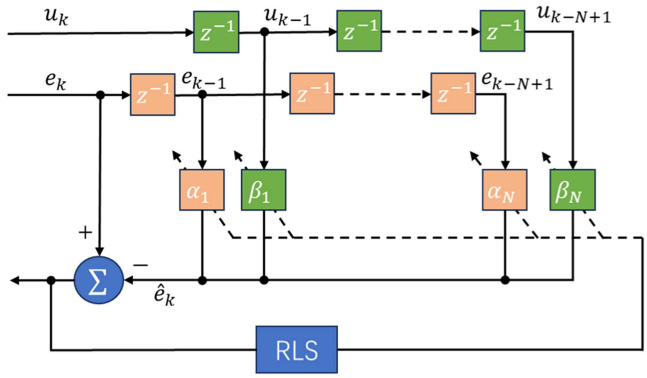

Fig. 3. Block diagram of the adaptive transversal filter.

The challenge of computing the mathematical expectation in the stochastic system optimization problem is converted into a parameter tuning problem for an adaptive filter, which greatly improves the efficiency of the algorithm execution.

C. Policy Iteration With Off-Policy Mechanism

Through the analyses in the previous two sections, we have addressed the problems of state reconstruction and expectation computation that controllers encounter when facing real stochastic systems. This section builds on this foundation by giving further information on how to use the policy iteration algorithm to circumvent the dependence on the precise model when designing the controller.

The policy iterative algorithm is built on the basis of Bellman's optimality equation, and by splitting the direct solution process of the optimal problem into the iterative solution process of the sub-problem, the requirement for accurate system model parameters can be eliminated. According to [50], the value function (11) can be expressed as a quadratic form

$$\begin{aligned} &Q^{\pi}(\bar{z}_k, u_k) \\ &= \begin{bmatrix}\bar{z}_k \\ u_k\end{bmatrix}^T \begin{bmatrix}\Gamma^T\left(\bar{Q} + \gamma A^T P A\right)\Gamma & \gamma\Gamma^T A^T P B \\ \gamma B P A\Gamma & R + B^T P B\end{bmatrix}\begin{bmatrix}\bar{z}_k \\ u_k\end{bmatrix} \\ &\triangleq \begin{bmatrix}\bar{z}_k \\ u_k\end{bmatrix}^T \begin{bmatrix}H_{zz} & H_{zu} \\ H_{uz} & H_{uz}\end{bmatrix}\begin{bmatrix}\bar{z}_k \\ u_k\end{bmatrix} \\ &\triangleq Z_k^T H Z_k. \end{aligned} \tag{14}$$

Equation (14) is of a very important form, completely separating the unknown parameters from the known data. Where Z_k contains the measurable historical jitter error and the control action sequence, the kernel matrix H contains all the unknown system model parameters such as the dynamics matrices in model (2) and the coefficient matrix of the Kalman filter Γ. Thus, the optimization problem of the controller at hand becomes an optimization problem with respect to the parameters of the kernel matrix H. This process is achieved by means of bootstrapping [28], which is completely independent of model parameters. Specifically, the policy iteration algorithm is calculated

Algorithm 1: The FB-OPPI Algorithm.

Initialization: Set $j = 0$. Select proper data length N and ε. Select an initially admissible control policy $\pi_0 = -K_0\bar{z}_k + \mu_k$, where $\mu_k \sim \mathcal{N}(0, \sigma_\mu)$is the exploration noise. Initialize the data buffer $\mathcal{D}$.

Data sampling: Run the system under the initialized controller and put the sampled data tuple $\{Z_k, U_k, \hat{Z}_{k+1}\}$ into $\mathcal{D}$. Meanwhile, the adaptive transversal filter is trained according to (12) until convergency.

Repeat:

Policy evaluation: Update the value function according to (15).

Policy improvement: Get an improved target controller by (16).

$j \leftarrow j + 1$

Until $\|H^j - H^{j-1}\|_2 < \varepsilon$.

iteratively between the policy evaluation stage

$$Z_k^T H^{j+1} Z_k - \gamma \hat{Z}_{k+1}^T H^{j+1} \hat{Z}_{k+1} = U_k, \tag{15}$$

and the policy improvement stage

$$\pi^{j+1}(\bar{z}_k) = \arg\min_{u_k} Q^{\pi_j}(\bar{z}_k, u_k) = -\left(H_{uu}^{j+1}\right)^{-1} H_{uz}^{j+1} \bar{z}_k, \tag{16}$$

where the policy evaluation (15) can be solved as a least square (LS) problem. The updated value function after (15) can better approximate the infinite-horizon performance index (10), so as to guide the optimization of the control policy in the policy improvement (16).

The main feature of off-policy learning mechanism is the presence of two controllers, an initialized controller to interact with the system and a target controller to achieve (near) optimal beam stabilization. The initialized controller (e.g., an integrator) is primarily used to maintain closed-loop stability of the system, with no requirements for control performance. The target controller can utilize the sampling data generated by the initialized controller for training and it does not participate in closed-loop control before convergence. Therefore, by separating the controllers used for learning and sampling, existing sampling data can be reused for optimization calculations, significantly reducing the time cost associated with sampling. Based on the above discussion, Algorithm 1 provides a detailed procedure for the proposed FB-OPPI method.

Notice that in the data sampling stage, the admissible controller adds exploration noise to the control voltage, which is to meet the persistence of excitation (PE) condition in adaptive control, more detailed discussion can be found in [33].

IV. Experimental Results

A. Experimental Setup

A beam jitter control system was constructed to verify the effectiveness of the proposed method, as depicted in Fig. 4.

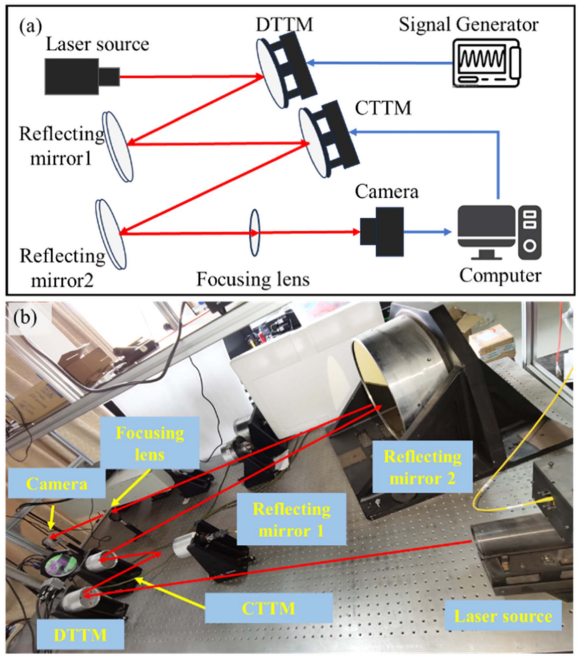

Fig. 4. The adaptive optics beam jitter control system: (a) Experimental setup, (b) schematic diagram.

A collimated laser beam at a wavelength of 635 nm simulates the observed target. Two TTMs were used, including the disturbance tilting mirror (DTTM) and the correction tilting mirror (CTTM). A sinusoidal analog signal generated by a signal generator is then amplified by a high-voltage amplifier to drive the DTTM to generate jitter in the optical path, while the CTTM is controlled by a computer workstation (Intel Core i7-8700K CPU) for jitter compensation. Both TTMs have the same specifications, with an aperture size of 70 mm, a stroke of ± 100 μrad, and a resonance frequency of 1200 Hz. For the experiments, we used the CL600x2/M industrial high-speed CCD camera from Optronis as the photoelectric detector to obtain beam jitter control error data. The effective working area of the camera in the experiment is 256 pixels x 256 pixels, with a single pixel size of 14 μm. The CCD camera operates at a sampling frequency of 1000 Hz and uses the centroiding algorithm to calculate the beam spot displacement as the beam jitter error. Therefore, we had neglected the resonant effect of the TTMs. In addition, only the jitter on the x-axis of the TTM was considered during the experiment, and the y-axis remained stationary.

For the FB-OPPI algorithm used in the experiment, the data length N is set to 25 after weighing the computational complexity and accuracy. Based on the number of unknown parameters in the kernel matrix H, it can be calculated that at least $(2N + 2)(2N + 3)/2$ sets of sampling data are required, hence the size of the data buffer is set to 1500. Therefore, it takes only 1.5 seconds to sample the enough data required for controller training. To better minimize jitter errors rather than energy consumption, the weighting matrices Q and R in the performance index were set to 100 and 0.01, respectively. The exploration noise variance in Algorithm 1 was 0.15.

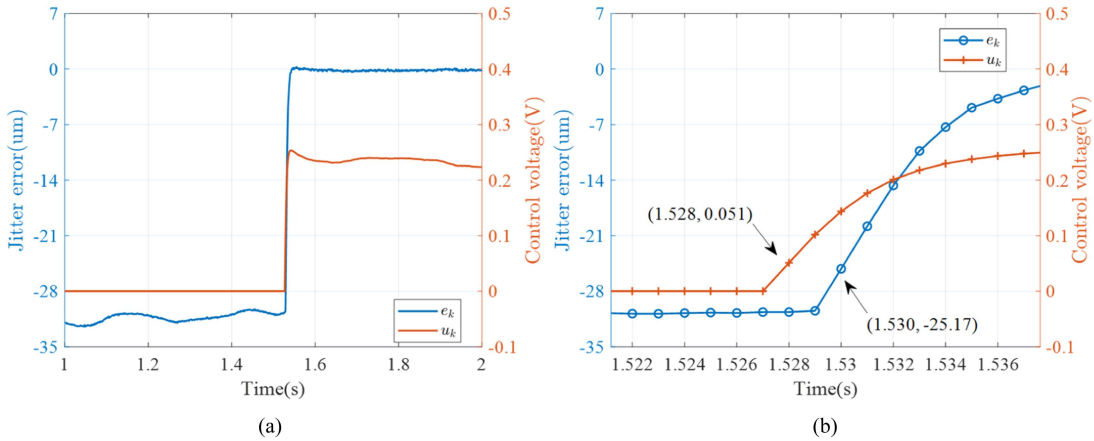

Fig. 5. Analysis of the system delay: (a) The response relationship between control action and jitter error, (b) local zoom.

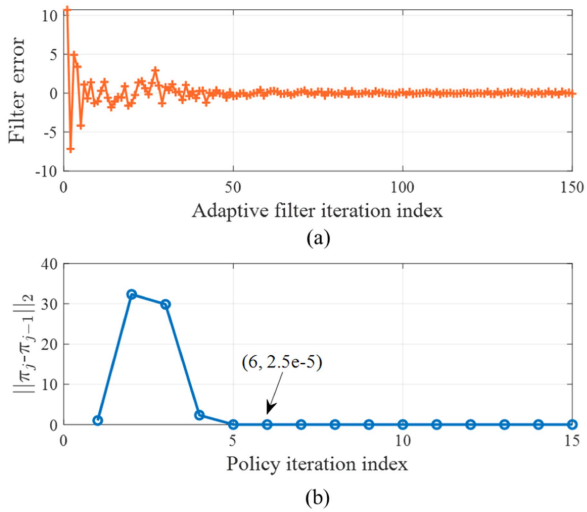

Fig. 6. Convergence of the FB-OPPI method: (a) Convergence curve of the adaptive transversal filter, (b) convergence curve of the target controller.

The system delay was first analyzed by examining the response between the control action and the jitter error in a closed loop. As illustrated in Fig. 5, the control action is initiated at 1.528 s, yet the compensation for the beam jitter is observed to commence only at 1.530 s. Consequently, it can be deduced that the system has a two-frame delay.

B. Convergence Analysis

First, it is necessary to verify the convergence of the FB-OPPI algorithm in a practical AO system, including the convergence of adaptive transversal filter parameters and the control gain of the target controller. The convergence curve is shown in Fig. 6.

We trained the parameters of the adaptive transversal filter using the RLS algorithm. The error of the RLS is shown in Fig. 6(a), indicating that the filter parameters can achieve convergence and stabilize within approximately 100 iterations. At the same time, the training of the target controller also maintained a fast and stable convergence. The difference between the controller gains of two consecutive iterations was measured using the L2 norm, as shown by the curve in Fig. 6(b). The results show that by the sixth iteration, the error can be reduced below 2.5e-5.

The above results demonstrate that the rapid and stable convergence of the FB-OPPI method under experimental conditions, despite measurement noise and computational errors in the data. To emphasize that the FB-OPPI method is driven by online data within a closed-loop framework without the need for model identification, we conducted another experiment to verify whether the FB-OPPI algorithm could effectively cope with varying external jitter, and Fig. 7 shows the entire control trajectory. The system initially suffered from a narrowband jitter of 59 Hz, which was then accurately suppressed by the proposed model-free controller after training. To simulate the effects of environmental changes, the center frequency of the jitter signal shifted to 179 Hz. The controller learned under 59 Hz jitter could only keep the control stability and the jitter suppression effect was reduced. However, as shown in stages 5 and 6, reapplying Algorithm 1 for controller quickly restored accurate jitter suppression performance.

The trajectory described in Fig. 7 represents a continuous closed-loop operation. Recalling that the jitter error (the blue curve) and control action (the orange curve) in Fig. 7 constitute the entirety of the data utilized throughout the learning process. This highlights the remarkable adaptability and flexibility of the proposed method. The FB-OPPI method does not rely on additional sensors to provide velocity and acceleration information but directly uses closed-loop data for training, avoiding extra errors introduced by sensor noise, POL data construction, and model identification processes. This method is an autonomous optimization process that reduces the frequent adjustment of controller parameters.

C. Beam Stabilization Control Performance Analysis

Beam stabilization control performance is a direct standard for verifying the performance of controllers. In our experiments, the proposed FB-OPPI method was compared with traditional integral controllers and the LQG controller proposed in [51].

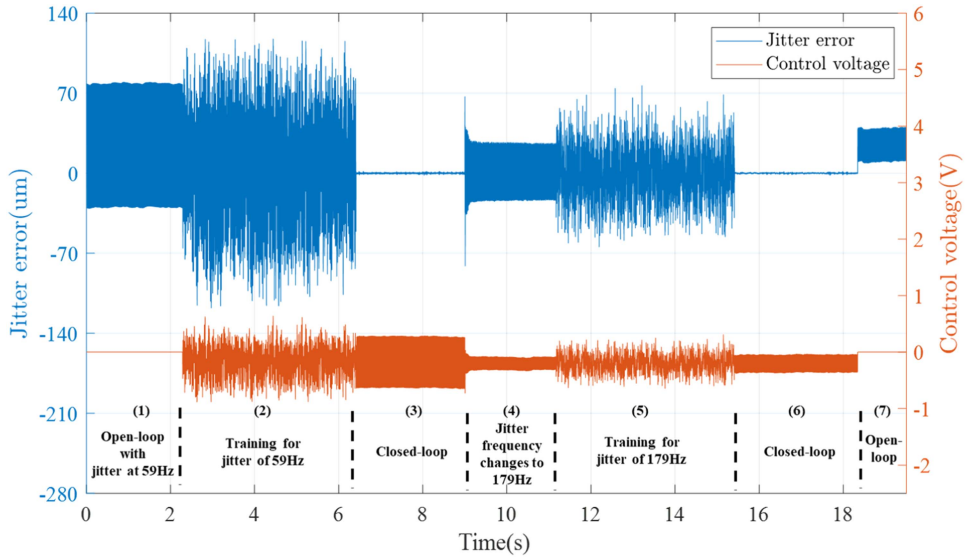

Fig. 7. Training and control trajectories for jitter signal with varying frequencies.

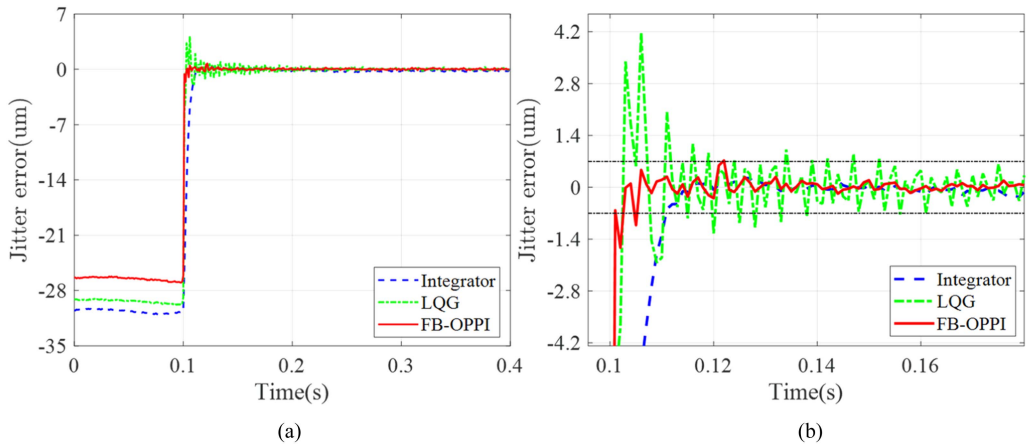

Fig. 8. Comparison results for jitter suppression in response to step disturbances: (a) Residual disturbance curves, (b) local zoom.

The time delay in the AO system loop has always been a critical factor affecting system performance, especially for beam stabilization tasks that require high control bandwidth. Traditional integral controllers are designed without considering the impact of system delay and typically require to be compensated by pre-calculating the corrections based on the dynamic response of an accurate AO system model to achieve better performance. However, in the FB-OPPI control scheme, the system delay is part of the unknown system model, as shown in model (2). This approach allows the dynamic characteristics of the AO system, including time delays, to be integrated into the controller design process without adversely affecting performance.

To verify this, a comparison was made between the dynamic performance of different control methods in terms of suppressing step disturbances. The results depicted in Fig. 8 indicate that the FB-OPPI method outperforms both the integrator and the LQG controller in terms of response speed and tracking accuracy. Notably, the rise time of the step response for the FB-OPPI method is only 0.007 s (7 frames), which is more than twice as fast as the 0.016 s (16 frames) of the integral control method. Additionally, it was observed that while the LQG controller can also compensate for system delay, it exhibits a more pronounced damping effect, prolonging the time the system takes to reach a stable state. The overshoot of the LQG controller was measured to be 4.172 μm, significantly higher than the 0.725 μm of the FB-OPPI method, which represents a reduction of 82.6%.

TABLE I
COMPARISON OF BEAM STABILIZATION CONTROL RESULTS FOR THE SINGLE NARROWBAND JITTER

Frequency (Hz)	RMSE (um)			
	Open-loop	Integrator	LQG	FB-OPPI
30	25.057	11.591	0.494	**0.415 (1.7%)**
50	20.705	16.288	0.577	**0.284 (1.4%)**
100	13.675	17.656	0.615	**0.365 (2.7%)**

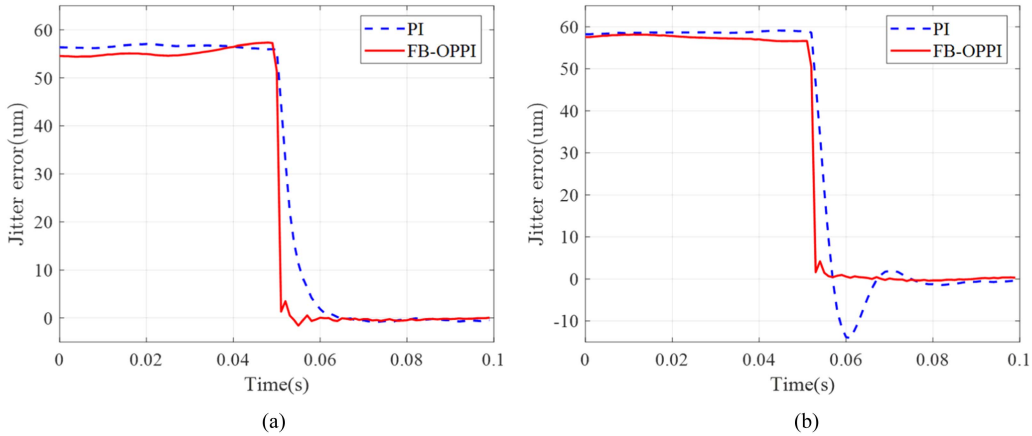

Fig. 9. Comparison results of different control methods for step disturbances with different system delay: (a) 2-frame delay, (b) 4-frame delay.

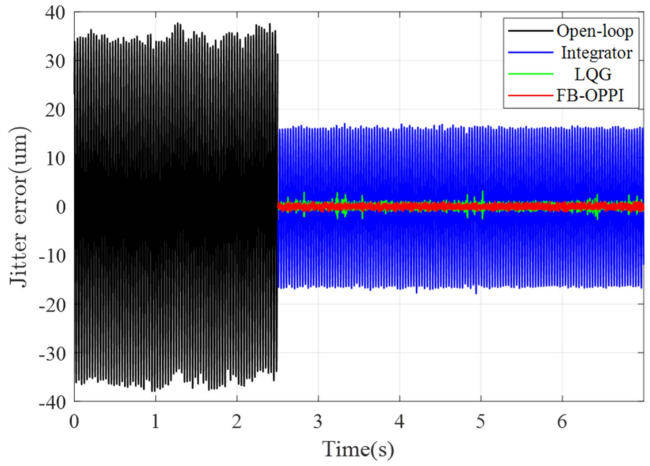

Fig. 10. Experimental comparison of beam stabilization for 30 Hz jitter.

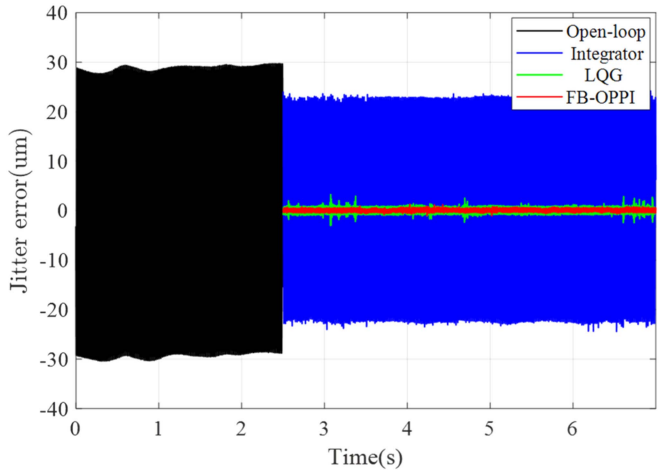

Fig. 11. Experimental comparison of beam stabilization for 50 Hz jitter.

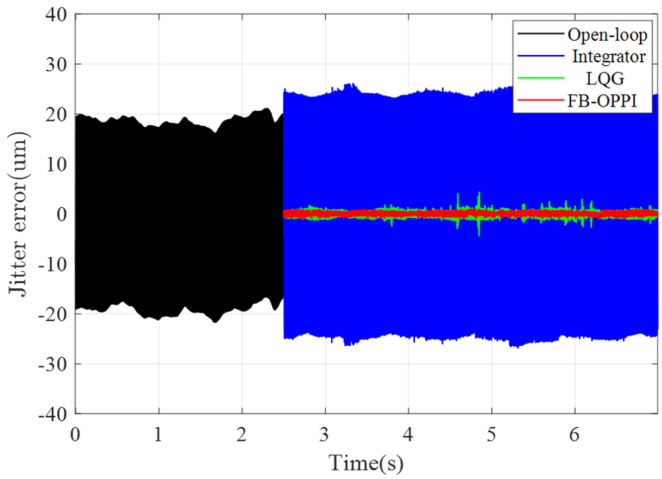

Fig. 12. Experimental comparison of beam stabilization for 100 Hz jitter.

To further illustrate the impact of system delay on the control performance of different methods, we compared the suppression capabilities against step disturbances under varying degrees of delay. The results in Fig. 9 demonstrate that as the system's delay increases, the stability margin of the system significantly decreases. The integral controller shows significant overshoot due to its inability to actively compensate for the time lag and causes the system to oscillate and destabilize. To maintain system stability, it is necessary to optimize the control parameters of the integrator, which will inevitably result in a reduction of its response speed and control bandwidth. In contrast, the proposed FB-OPPI control method can autonomously identify the delay characteristics within the system. Consequently, the controller derived from the FB-OPPI method ensures control stability while maintaining a high response speed. In summary, the FB-OPPI controller can effectively address the common time delay issues in AO systems, exhibiting superior dynamic performance.

Subsequently, we compared the stabilization effects of these methods on narrowband beam jitter at 30 Hz, 50 Hz, and 100 Hz, with the results displayed in Figs. 10 –12. It can be observed that traditional integral control exhibits the poorest beam stabilization performance, with jitter error increasing as frequency rises. When facing 100 Hz beam jitter, the uncompensated phase lag caused by the controller itself and the system delay leads to an increase in jitter error, severely affecting the

TABLE II
COMPARISON OF BEAM STABILIZATION CONTROL RESULTS FOR THE MIXED-FREQUENCY JITTER

Frequency (Hz)	RMSE (um)				PSD (dB)			
	Open-loop	Integrator	LQG	FB-OPPI	Open-loop	Integrator	LQG	FB-OPPI
30+106+179	13.35	9.67	1.21	0.80	23.74 (30Hz)	20.35	-6.16	-9.16
					21.69 (106Hz)	16.09	-10.06	-9.69
					10.70 (179Hz)	18.44	-3.74	-5.80

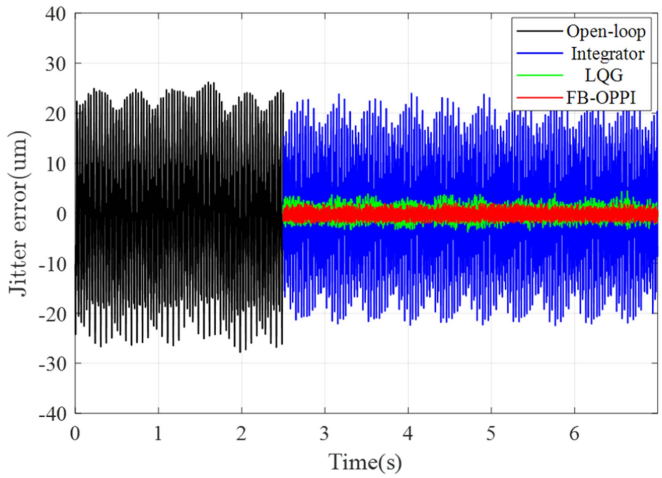

Fig. 13. Experimental comparison of mixed-frequency beam stabilization.

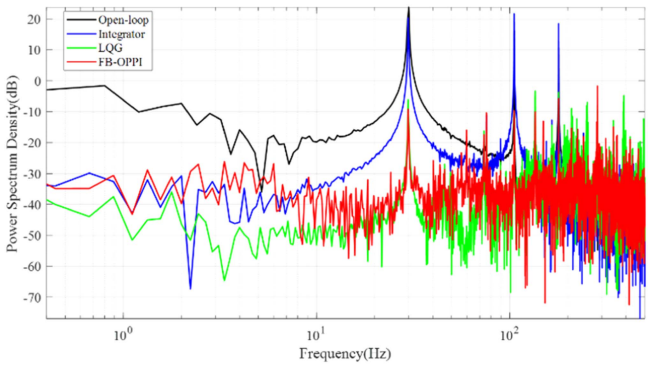

Fig. 14. PSDs comparison of mixed-frequency beam stabilization.

stability of the AO system. In contrast, both LQG and FB-OPPI can realize effective control of these narrowband beam jitters.

However, it is noticed that there are significant fluctuations and abrupt changes in the jitter error under LQG control, while the beam jitter error controlled by FB-OPPI is smoother and more stable. The reason is that the LQG controller only considers the linear response of the TTM and a single beam jitter model, without adequately accounting for unmodeled high-frequency dynamics and parasitic jitters generated by the TTM during operation. On the contrary, the FB-OPPI controller circumvents the model errors. The use of closed-loop data for controller design enabled it to fully account for the dynamics of parasitic jitter and yielded better control performance.

For a more intuitive comparison, Table I presents the root-mean-square error of the three methods. It is evident that the traditional integral control has limited performance in controlling the high-frequency narrowband beam jitter. Calculations show that the root-mean-square error (RMSE) of the FB-OPPI method has been reduced by more than 97.3% compared to the open loop, demonstrating consistent control performance across various frequencies of beam jitter. The data in Table I further illustrate that the FB-OPPI method is far superior to the conventional integral control, and it also maintains an advantage compared to the advanced LQG control.

Considering that beam jitter in real systems is affected by various factors, it usually has a more complex frequency spectrum. We next verify the control effect on the mixed-frequency beam jitter, as shown in Fig. 13. The mixed-frequency beam jitter signal mainly consists of narrowband jitter at 30 Hz, 106 Hz, and 179 Hz, as well as low-frequency broadband disturbance. The results indicate that the FB-OPPI method can also correctly understand the dynamic characteristics of mixed-frequency beam jitter and optimize an appropriate controller, without relying on the identification of the jitter model and parameter adjustment. The PSDs of the jitter error presented in Fig. 14 also shows that the FB-OPPI has the capability to suppress both low-frequency broadband and high-frequency narrowband jitter. Specific data are provided in Table II. These results demonstrate that the FB-OPPI method has a wideband beam stability performance, capable of correctly and efficiently handling complex beam stabilization control issues in practical AO system applications.

V. CONCLUSION

In order to avoid the influence of model errors on the control performance of practical AO beam stabilization systems, this paper develops a new model-free control scheme, the filter-based off-policy policy iteration (FB-OPPI) algorithm. The FB-OPPI algorithm is crafted to address the practical needs of applications. The challenges of traditional strategy iterative algorithms in the face of unknown system states and measurement noise are addressed through the design of Kalman and adaptive filters, which allows the algorithms to be directly applied using jitter errors and control actions in real operating trajectories. The off-policy learning mechanism for controller optimization features high data efficiency while ensuring system stability. The experimental results indicate that our FB-OPPI controller, operating in a closed-loop configuration, demonstrates swift and stable convergence. Compared to the traditional integral controller, the FB-OPPI scheme is more effective in handling high-frequency narrowband and complex frequency beam jitter

introduced by devices in AO systems. And in contrast to the advanced LQG control, the FB-OPPI approach does not require model identification, thus demonstrating higher control stability and dynamic performance in experiments.

Building on these results, further experimental validations will be conducted to assess the effectiveness of FB-OPPI on TTMs with larger apertures, higher time delays and more complex dynamics. Based on our preliminary validation, measurement noise is indeed one of the significant factors affecting the control performance of the proposed algorithm. Currently, many improved centroid algorithms can assist us in reducing the introduction of measurement noise. However, we will conduct a more in-depth quantitative analysis of the specific impact of measurement noise on control performance in our subsequent work and propose corresponding solutions. In addition, accelerated learning algorithms will be investigated to alleviate computational demands. This may require us to further consider high-performance computing and fuzzy control techniques [52], [53] to alleviate computational demands and better cope with uncertainties in the system. We are confident that this method shows significant potential for practical application in the near future.

REFERENCES

[1] M. Arikawa and T. Ito, "Performance of mode diversity reception of a polarization-division-multiplexed signal for free-space optical communication under atmospheric turbulence," *Opt. Exp.*, vol. 26, no. 22, Oct. 2018, Art. no. 28263, doi: 10.1364/OE.26.028263.

[2] X. Wang, X. Su, G. Liu, J. Han, R. Wang, and K. Wang, "Laser beam jitter control of the link in free space optical communication systems," *Opt. Exp.*, vol. 29, no. 25, Dec. 2021, Art. no. 41582, doi: 10.1364/OE.443411.

[3] Y. Clénet et al., "NAOS performances: Impact of the telescope vibrations and possible origins," in *Proc. SF2A: Semaine Astrophysique Francaise*, 2004, p. 179.

[4] J. R. Maly, D. Erickson, and T. J. Pargett, "Vibration suppression for the Gemini planet imager," *Proc. SPIE*, vol. 7733, pp. 506–514, 2010, doi: 10.1117/12.857699.

[5] A. V. Goncharov, N. Devaney, and C. Dainty, "Atmospheric dispersion compensation for extremely large telescopes," *Opt. Exp.*, vol. 15, no. 4, Feb. 2007, Art. no. 1534, doi: 10.1364/OE.15.001534.

[6] C. Correia and J.-P. Véran, "Woofer–tweeter temporal correction split in atmospheric adaptive optics," *Opt. Lett.*, vol. 37, no. 15, Aug. 2012, Art. no. 3132, doi: 10.1364/OL.37.003132.

[7] Y. Guo et al., "Adaptive optics based on machine learning: A review," *Opto-Electron. Adv.*, vol. 5, no. 7, 2022, Art. no. 200082, doi: 10.29026/oea.2022.200082.

[8] Y. Guo et al., "High-resolution visible imaging with piezoelectric deformable secondary mirror: Experimental results at the 1.8-m adaptive telescope," *Opto-Electron. Adv.*, vol. 6, no. 12, 2023, Art. no. 230039, doi: 10.29026/oea.2023.230039.

[9] J. Garcés et al., "Vibrations in magao: Resonance sources identification through frequency-based analysis," presented at the Adaptive Opt.: Anal., Methods Syst., Arlington, VA, USA, Jun. 7–11, 2015, Paper AOTh1D, doi: 10.1364/AOMS.2015.

[10] J. A. Stoesz et al., "Evaluation of the on-sky performance of Altair," *Proc. SPIE*, vol. 5490, pp. 67–78, Oct. 2004, doi: 10.1117/12.552493.

[11] Y. Su, J. Han, X. Wang, C. Ma, and J. Wu, "Adaptive compound control of laser beam jitter in deep-space optical communication systems," *Opt. Exp.*, vol. 32, no. 13, Jun. 2024, Art. no. 23228, doi: 10.1364/OE.521520.

[12] C. Petit, J.-M. Conan, C. Kulcsár, H.-F. Raynaud, and T. Fusco, "First laboratory validation of vibration filtering with LQG control law for Adaptive Optics," *Opt. Exp.*, vol. 16, no. 1, pp. 87–97, 2008, doi: 10.1364/OE.16.000087.

[13] G. Sivo et al., "First on-sky SCAO validation of full LQG control with vibration mitigation on the CANARY pathfinder," *Opt. Exp.*, vol. 22, no. 19, Sep. 2014, Art. no. 23565, doi: 10.1364/OE.22.023565.

[14] C. Petit et al., "SPHERE eXtreme AO control scheme: Final performance assessment and on sky validation of the first auto-tuned LQG based operational system," *Proc. SPIE*, vol. 9148, pp. 214–230, 2014, doi: 10.1117/12.2052847.

[15] S. Meimon, C. Petit, T. Fusco, and C. Kulcsar, "Tip–tilt disturbance model identification for Kalman-based control scheme: Application to XAO and ELT systems," *J. Opt. Soc. Amer. A*, vol. 27, no. 11, Nov. 2010, Art. no. A122, doi: 10.1364/JOSAA.27.00A122.

[16] K. Hinnen, M. Verhaegen, and N. Doelman, "A data-driven H2-optimal control approach for adaptive optics," *IEEE Trans. Control Syst. Technol.*, vol. 16, no. 3, pp. 381–395, May 2008, doi: 10.1109/TCST.2007.903374.

[17] M. Gluck, J.-U. Pott, and O. Sawodny, "Model predictive control of multi-mirror adaptive optics systems," in *Proc. IEEE Conf. Control Technol. Appl.*, 2018, pp. 909–914, doi: 10.1109/CCTA.2018.8511551.

[18] R. J. Watkins and B. N. Agrawal, "Use of least means squares filter in control of optical beam jitter," *J. Guid., Control, Dyn.*, vol. 30, no. 4, pp. 1116–1122, Jul. 2007, doi: 10.2514/1.26778.

[19] P. K. Orzechowski, N. Y. Chen, J. S. Gibson, and T.-C. Tsao, "Optimal suppression of laser beam jitter by high-order RLS adaptive control," *IEEE Trans. Control Syst. Technol.*, vol. 16, no. 2, pp. 255–267, Mar. 2008, doi: 10.1109/TCST.2007.903377.

[20] K. Yang, P. Yang, S. Wang, L. Dong, and B. Xu, "Tip-tilt disturbance model identification based on non-linear least squares fitting for Linear Quadratic Gaussian control," *Opt. Commun.*, vol. 415, pp. 31–38, May 2018, doi: 10.1016/j.optcom.2018.01.022.

[21] J. Mocci, M. Quintavalla, A. Chiuso, S. Bonora, and R. Muradore, "PI-shaped LQG control design for adaptive optics systems," *Control Eng. Pract.*, vol. 102, Sep. 2020, Art. no. 104528, doi: 10.1016/j.conengprac.2020.104528.

[22] A. Chiuso, R. Muradore, and E. Marchetti, "Dynamic calibration of adaptive optics systems: A system identification approach," *IEEE Trans. Control Syst. Technol.*, vol. 18, no. 3, pp. 705–713, May 2010, doi: 10.1109/TCST.2009.2023914.

[23] A. Haber and M. Verhaegen, "Modeling and state-space identification of deformable mirrors," *Opt. Exp.*, vol. 28, no. 4, pp. 4726–4740, 2020, doi: 10.1364/OE.382880.

[24] A. Haber and M. Krainak, "Data-driven estimation, tracking, and system identification of deterministic and stochastic optical spot dynamics," *Opt. Exp.*, vol. 31, no. 11, May 2023, Art. no. 17494, doi: 10.1364/OE.486642.

[25] R. Swanson, M. Lamb, C. M. Correia, S. Sivanandam, and K. Kutulakos, "Closed loop predictive control of adaptive optics systems with convolutional neural networks," *Monthly Notices Roy. Astron. Soc.*, vol. 503, no. 2, pp. 2944–2954, 2021, doi: 10.1093/mnras/stab632.

[26] R. Landman and S. Haffert, "Nonlinear wavefront reconstruction with convolutional neural networks for Fourier-based wavefront sensors," *Opt. Exp.*, vol. 28, no. 11, pp. 16644–16657, 2020, doi: 10.1364/OE.389465.

[27] X. Liu, T. Morris, C. Saunter, F. J. de Cos Juez, C. González-Gutiérrez, and L. Bardou, "Wavefront prediction using artificial neural networks for open-loop adaptive optics," *Monthly Notices Roy. Astron. Soc.*, vol. 496, no. 1, pp. 456–464, Jul. 2020, doi: 10.1093/mnras/staa1558.

[28] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An introduction*. Cambridge, MA, USA: MIT Press, 2018.

[29] J. Nousiainen et al., "Toward on-sky adaptive optics control using reinforcement learning," *Astron. Astrophys.*, vol. 664, p. A71, 2022, doi: 10.1051/0004-6361/202243311.

[30] H. Ke et al., "Self-learning control for wavefront sensorless adaptive optics system through deep reinforcement learning," *Optik*, vol. 178, pp. 785–793, Feb. 2019, doi: 10.1016/j.ijleo.2018.09.160.

[31] B. Pou, F. Ferreira, E. Quinones, D. Gratadour, and M. Martin, "Adaptive optics control with multi-agent model-free reinforcement learning," *Opt. Exp.*, vol. 30, no. 2, Jan. 2022, Art. no. 2991, doi: 10.1364/OE.444099.

[32] R. Landman, S. Y. Haffert, V. M. Radhakrishnan, and C. U. Keller, "Self-optimizing adaptive optics control with reinforcement learning for high-contrast imaging," *J. Astron. Telescopes, Instrum., Syst.*, vol. 7, no. 3, 2021, Art. no. 039002, doi: 10.1117/1.JATIS.7.3.039002.

[33] F. L. Lewis, D. Vrabie, and K. G. Vamvoudakis, "Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers," *IEEE Control Syst.*, vol. 32, no. 6, pp. 76–105, Dec. 2012, doi: 10.1109/MCS.2012.2214134.

[34] J. Lai, J. Xiong, and Z. Shu, "Model-free optimal control of discrete-time systems with additive and multiplicative noises," *Automatica*, vol. 147, Jan. 2023, Art. no. 110685, doi: 10.1016/j.automatica.2022.110685.

[35] J. Zhao, Y. Lv, Z. Wang, and Z. Zhao, "Adaptive Q-learning based model-free H∞ control of continuous-time nonlinear systems: Theory and application," *IEEE Trans. Emerg. Topics Comput. Intell.*, early access, Sep. 02, 2024, doi: 10.1109/TETCI.2024.3449870.

[36] J. Zhao, Y. Lv, Z. Zhao, and Z. Wang, "Adaptive optimal tracking control of servo mechanisms via generalised policy learning," *IEEE Trans. Instrum. Meas.*, vol. 73, 2024, Art. no. 3002311, doi: 10.1109/TIM.2024.3457963.

[37] M. Zhang, M.-G. Gan, and J. Chen, "Data-driven adaptive optimal control for stochastic systems with unmeasurable state," *Neurocomputing*, vol. 397, pp. 1–10, Jul. 2020, doi: 10.1016/j.neucom.2019.12.001.

[38] A. N. Putri, C. Machbub, D. Mahayana, and E. Hidayat, "Data driven linear quadratic Gaussian control design," *IEEE Access*, vol. 11, pp. 24227–24237, 2023, doi: 10.1109/ACCESS.2023.3254879.

[39] B. Kiumarsi, F. L. Lewis, M.-B. Naghibi-Sistani, and A. Karimpour, "Optimal tracking control of unknown discrete-time linear systems using input-output measured data," *IEEE Trans. Cybern.*, vol. 45, no. 12, pp. 2770–2779, Dec. 2015, doi: 10.1109/TCYB.2014.2384016.

[40] S. A. A. Rizvi, A. J. Pertzborn, and Z. Lin, "Reinforcement learning based optimal tracking control under unmeasurable disturbances with application to HVAC systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 7523–7533, Dec. 2022, doi: 10.1109/TNNLS.2021.3085358.

[41] F. L. Lewis and K. G. Vamvoudakis, "Reinforcement learning for partially observable dynamic processes: Adaptive dynamic programming using measured output data," *IEEE Trans. Syst., Man, Cybern. B*, vol. 41, no. 1, pp. 14–25, Feb. 2011, doi: 10.1109/TSMCB.2010.2043839.

[42] J. Zhao, Y. Lv, Q. Zeng, and L. Wan, "Online policy learning-based output-feedback optimal control of continuous-time systems," *IEEE Trans. Circuits Syst. II: Exp. Briefs*, vol. 71, no. 2, pp. 652–656, Feb. 2024, doi: 10.1109/TCSII.2022.3211832.

[43] M. Li, J. Qin, W. X. Zheng, Y. Wang, and Y. Kang, "Model-free design of stochastic LQR controller from reinforcement learning and primal-dual optimization perspective," 2021, *arXiv:2103.09407*.

[44] S. A. A. Rizvi, Y. Wei, and Z. Lin, "Model-free optimal stabilization of unknown time delay systems using adaptive dynamic programming," in *Proc. IEEE 58th Conf. Decis. Control*, 2019, pp. 6536–6541, doi: 10.1109/CDC40024.2019.9029600.

[45] T. Söderström, *Discrete-Time Stochastic Systems: Estimation and Control*. Berlin, Germany: Springer, 2002.

[46] B. D. Anderson and J. B. Moore, *Optimal Control: Linear Quadratic Methods*. North Chelmsford, MA, USA: Courier Corporation, 2007.

[47] N. Malla and Z. Ni, "A new history experience replay design for model-free adaptive dynamic programming," *Neurocomputing*, vol. 266, pp. 141–149, Nov. 2017, doi: 10.1016/j.neucom.2017.04.069.

[48] C. Petit, J.-M. Conan, C. Kulcsár, and H.-F. Raynaud, "Linear quadratic Gaussian control for adaptive optics and multiconjugate adaptive optics: Experimental and numerical analysis," *J. Opt. Soc. Amer. A*, vol. 26, no. 6, Jun. 2009, Art. no. 1307, doi: 10.1364/JOSAA.26.001307.

[49] S. A. U. Islam and D. S. Bernstein, "Recursive least squares for real-time implementation [lecture notes]," *IEEE Control Syst.*, vol. 39, no. 3, pp. 82–85, Jun. 2019, doi: 10.1109/MCS.2019.2900788.

[50] W. C. Wong and J. H. Lee, "A reinforcement learning-based scheme for direct adaptive optimal control of linear stochastic systems: Reinforcement learning-based scheme," *Optimal Control Appl. Methods*, vol. 31, no. 4, pp. 365–374, Jul. 2010, doi: 10.1002/oca.915.

[51] K. Yang et al., "Vibration identification based on Levenberg–Marquardt optimization for mitigation in adaptive optics systems," *Appl. Opt.*, vol. 57, no. 11, Apr. 2018, Art. no. 2820, doi: 10.1364/AO.57.002820.

[52] C. Chen, W. Zou, and Z. Xiang, "Event-triggered connectivity-preserving formation control of heterogeneous multiple USVs," *IEEE Trans. Syst., Man, Cybern.: Syst.*, early access, Sep. 17, 2024, doi: 10.1109/TSMC.2024.3458465.

[53] Y. Zhang, M. Chadli, and Z. Xiang, "Prescribed-time formation control for a class of multiagent systems via fuzzy reinforcement learning," *IEEE Trans. Fuzzy Syst.*, vol. 31, no. 12, pp. 4195–4204, Dec. 2023, doi: 10.1109/TFUZZ.2023.3277480.