Direct Heuristic Dynamic Programming Method for Power System Stability Enhancement*

Miao Yu, Chao Lu and Yongjun Liu

Abstract—In this paper a neural network-based approximate dynamic programming method, namely direct heuristic dynamic programming (direct HDP), is applied to power system stability control. Direct HDP is a learning and approximation based approach to address nonlinear system control under uncertainty. In the present paper, real-time system responses provided by wide area measurement system (WAMS) are used to construct such controllers which are uniquely tailored for the problems under consideration. In addition, the controller learning objective is formulated as a reward function that reflects global characteristics of the power system low frequency oscillation under the consideration of coupling effect among system components. The contribution of the paper includes a convergence proof of the direct HDP algorithm using an LQR framework, as well as case study to illustrate the proposed learning control algorithm. The case study aims at providing a new solution to a difficult large scale system coordination problem where the China Southern Power Grid is used for.

I. INTRODUCTION

Power system is a complex network composed of many different components such as generators, AC or DC transmission lines, and various types of load. The most important task of power system control is to maintain stable operation of the system which includes angle stability and voltage stability. Exciter and governor control of generators are traditional methods. High voltage direct current (HVDC) and flexible AC transmission system (FACTS) equipment are modern control devices that are frequently adopted in recent years. However, system nonlinearity and uncertainty as disturbances, plus coordination of multiple controllers are three main challenges for controller design in such a huge scale system.

Many nonlinear phenomena are apparent in a power system such as dead zone and control limits, to name a few. In addition, some special characteristics of nonlinear systems, bifurcation and chaos for example, have also been observed in power systems[1]. For a nonlinear control system design, two approaches are usually adopted: linearization and energy function. The former includes linearization around the operating point [2] and exact feedback linearization (EFL) [3], then linear design principles can be applied, however, EFL depends greatly on the accuracy of the system model. In the second approach, energy storage functions or Hamiltonian functions are constructed based on the dissipation or Lyapunov stability theory to obtain the nonlinear control law [4], but the energy function can not be easily constructed such that only simplified power system elements can be taken into consideration.

Owing to the development of power electronics technologies, power systems are becoming more controllable. For some global stability problems such as low-frequency oscillations in the entire network, it is essential that coordination among the controls of various devices must be considered. In addition to step-by-step field tuning [5], sensitivity information is also used to achieve this goal[6]. Moreover, multi-input multi-output (MIMO) system design is also an effective technique to achieve coordination[7]. Nevertheless, the manner in which these problems can be considered simultaneously, i.e., the manner in which coordinated stability controllers can be designed on a large scale for complicated nonlinear and uncertain systems has not been studied in depth. This constitutes the main objective of this paper.

Most classic control designs are based on mathematical models, which lend more accuracy to the controller analyses and designs; however, the limitations of these models seriously affect the practical control performance. Several approximate dynamic programming (ADP) methods [8] are based on real system responses, which reflect the most important system dynamics, and using these methods, the constraints of nonlinearities and uncertainties can be potentially avoided. For example, generator control using dual heuristic programming (DHP) was achieved via digital simulations and physical experiments[9]. In this approach, a system model implemented using a neural network was pretrained to predict the system responses in the next time step, and the training quality of the model greatly determines the performance of the controller. In this paper, we consider a model-independent ADP approach [10], which can be viewed as model-free action-dependent heuristic dynamic programming in adaptive critic designs (ACD); this will be referred to as direct HDP in the followings. In [11], direct HDP was employed to control the flight of an Apache helicopter—a complex, continuous state/control, MIMO nonlinear system with uncertainty. The learning performance of the direct HDP application in multimachine power systems is studied in [12]. These results suggest the potential of adaptive critic designs for scalable complex system control applications.

In this paper, the direct HDP method is employed to solve the low-frequency oscillation problem in multimachine power system. The swing period is usually greater than 1 s; therefore, the controller has sufficient time to adapt. The rest of this
paper is organized as follows: the general direct HDP framework is explained in section II. In section III, the relationships between optimal control and dynamic programming are described. Further, the convergence of the direct HDP in the linear quadratic regulator (LQR) problem is also deduced, which is the foundation for applications to power systems. The implementation details of the direct HDP controller in section IV. The case is a classic two-area four-generator system proposed by Kundur, in which the learning abilities and the generalization of a static var compensator (SVC) supplementary controller in a nonlinear and uncertain environment are validated. Section V summarizes this paper.

II. FRAMEWORK OF DIRECT HDP CONTROL

Direct HDP belongs to the ACD family. Its basic control framework is shown in Fig.1, where \( u \) is the control signal computed by an ADP controller and is outputted to the environment. \( X \) is the state vector, that is, the responses of the environment to \( u \). At the same time, the effect of \( u \) is evaluated by using a cost function and is used to update the control policy. Since a control environment/plant is usually represented by a set of differential equations, a time-domain computer simulation, or the real dynamic system itself, strong nonlinearities can be easily embedded in it. The iterations of the controller parameters are based on real-time system responses, which reflect the conditions of a practical system, thereby implying the uncertainties. The controller performance is determined by the cost function, which can be used to reflect the global system dynamics and coordination among designed controllers.

Direct HDP control comprises two main parts: an action network and a critic network. The former produces control signals according to the learned policy, while the latter approximates the function \( J \) of the Bellman equation in dynamic programming. The structure of these two parts can be look-up tables, neural networks, or decision trees. In most cases, neural networks are employed because of their universal approximation capability and the associated simple learning algorithm based on the gradient descent.

In power systems, the real-time system dynamics fed back to the direct HDP controller are provided by a WAMS, which is based on synchronized phasor measurement techniques and modern digital communication networks. Through this system, some key system variables (e.g., internal voltage angles of the generators in different areas) unavailable in the past can now be measured directly and used as remote feedback to improve system performance [13].

Fig.2 shows the schematic representation of direct HDP control [10]. The reinforcement signal \( r(t) \) is obtained from the external environment and its value is typically either “0” or “1” corresponding to “success” or “failure,” respectively. During online learning, the controller is “naive” when it starts to control, that is, both the action and critic neural networks are randomly initialized for their weights. Once a system state is observed, an action will be subsequently produced based on the parameters in the action network. A “better” control value under the specific system state will make the equations of the principle of optimality more balanced. This set of system operations will be reinforced through memory or other associations between the states and control outputs in the action network. Otherwise, the control value will be adjusted by tuning the weights in the action network.

The output of the critic network (the \( J \) function) approximates the discounted total reward-to-go. Specifically, it approximates \( R(t) \) given by

\[
R(t) = \sum_{k=1}^{\infty} \alpha^{k-1} r(t+k) \quad (1)
\]

where \( R(t) \) is the future accumulative reward-to-go value at time \( t \) and \( \alpha \) is a discount factor for the infinite-horizon problem (\( 0 < \alpha < 1 \)).

A. Critic Network

The critic network is trained to approximate the “value function” \( J(t) \) by minimizing the objective function, which represents the balance of the principle of optimality, as follows:

\[
E_c(t) = \frac{1}{2} e_c(t)^2 \quad (2)
\]

Where

\[
e_c(t) = \alpha J(t) - J(t-1) - r(t) \quad (3)
\]

B. Action Network

The principle behind adapting the action network is to back-propagate the error between the desired ultimate objective, denoted by \( U_* \), and the cost function \( R(t) \). Either the actual cost function \( R(t) \) or its approximation \( J(t) \) is used depending on whether an explicit cost function or a critic network is available. In the latter, back-propagation is achieved through the critic network, as shown in Fig.1. In this paper, for notational simplicity, \( J(t) \) represents either the actual or the approximate cost function, depending on which is being used.

The weight updated in the action network adjusts the action network weights to minimize the following objective function:

\[
E_a(t) = \frac{1}{2} e_a(t)^2 \quad (4)
\]
where

\[ e_i(t) = J_i(t) - U_i(t) \]  

**III. DIRECT HDP CONTROL FOR LQR PROBLEM**

**A. Optimal Control and Dynamic Programming**

Optimal control and dynamic programming can both be categorized as optimization theories. There exist tight relationships between optimal control based on calculus of variation or Pontryagin’s minimum principle and dynamic programming based on the principle of Bellman optimality. For example, the system model of a deterministic nonlinear continuous system is expressed as

\[ x = f(x, u, t), \ t \geq t_0, \ t_0 \text{ fixed} \]  

with the performance index given by

\[ J(x(t_0), t_0) = \phi(x(T), T) + \int_{t_0}^{T} L(x, u, t)dt \]  

and the final state constraint given by

\[ \psi(x(T), T) = 0 \]  

If we define the Hamiltonian function as

\[ H(x, u, t) = L(x, u, t) + \lambda^T f(x, u, t) \]  

the optimal control law obtained using the Euler-Lagrange equations in variant calculus is

\[
\begin{align*}
\dot{x} &= \frac{\partial H}{\partial x} = f, \ t \geq t_0 \\
\dot{\lambda} &= \frac{\partial H}{\partial u} + \frac{\partial f}{\partial x} \lambda, \ t \leq T \\
0 &= \frac{\partial H}{\partial u} + \frac{\partial f}{\partial x} \lambda \\
(\phi + \psi' \lambda - \lambda) \left[ dx(T) + (\phi + \psi' \lambda + H) \right] dt = 0
\end{align*}
\]

which correspond to the state and costate equations, stationary and boundary conditions, respectively.

If the dynamic programming approach is employed, according to the principle of optimality, when the optimal policy is reached, the cost can be expressed as

\[ J^*(x(t), t) = \min_{u(t)} \left[ \int_{t_0}^{t_0} L(x, u, t) dt + \int_{t_0}^{T} J^*(x + \Delta x, t + \Delta t) dt \right] \]  

From the above equation, the Hamilton-Jacobi-Bellman (HJB) equation can be derived as follows:

\[ -\frac{\partial J^*}{\partial t} = \min_{u(t)} \left( L + \frac{\partial J^*}{\partial x} f \right) \]  

This is a partial differential equation for the optimal cost \( J^*(x, t) \), which is usually solved backward in time. Since \( u(t) \) is unconstrained, the minimization in the HJB equation can be carried out by setting

\[ \frac{\partial L}{\partial u} + \frac{\partial f}{\partial x} \lambda = 0 \]  

This is the stationary condition in equation (10). To examine the dynamics of \( \lambda(t) \), we write

\[ -\frac{d\lambda}{dt} = \frac{\partial J^*}{\partial \lambda} \frac{\partial J^*}{\partial x} \dot{x} \]  

Next, we take the partial derivative of the HJB equation such that

\[ -\frac{\partial J^*}{\partial \lambda} = \min_{u(t)} \left( \frac{\partial L}{\partial x} + \frac{\partial f}{\partial x} \lambda + f \frac{\partial \lambda}{\partial x} + \frac{\partial f}{\partial \lambda} \lambda^T \dot{x} \right) \]  

At the minimum, equation (13) holds so that

\[ -\frac{\partial J^*}{\partial \lambda} = \frac{\partial L}{\partial x} + \frac{\partial f}{\partial \lambda} \lambda + \frac{\partial f}{\partial \lambda} \dot{x} \]  

On using the state equation and equation (16), we obtain

\[ \frac{d\lambda}{dt} = \frac{\partial f}{\partial \lambda} \lambda + \frac{\partial f}{\partial \lambda} \dot{x} \]  

which is exactly the costate equation.[14]

If \( u(t) \) is constrained, the same result can be deduced from the HJB equation as easily as from Pontryagin’s minimum principle.

Within the framework of optimal control, solving the state and costate equations is difficult for a general nonlinear system, especially when \( u(t) \) is constrained. On the other hand, the solution of the HJB equation is more difficult; however, by using a dynamic programming algorithm, this problem can be numerically and exactly computed, and the more the number of constraints on the control and state variables, the easier it is to obtain the solutions.[14] Nevertheless, if the number of possible states and admissible controls in the discretization is very large, the “curse of dimensionality” is inevitable and therefore various approximation methods including the direct HDP method are presented.

**B. Direct HDP Control for LQR Problem**

The convergence and accuracy of approximation are important issues and have been widely investigated.[15]-[19] The scope of these researches encompass the reinforcement learning (especially TD learning) and adaptive critic design methods. The LQR problem was focused on in many papers because of its classic formulation and clear results. In this paper, the convergence of direct HDP control for the LQR problem will also be proved below.

In the direct HDP control approach, multi-layer perceptrons (MLP) neural networks with the back-propagation (BP) algorithm are used as the critic and action networks. According to the universal approximation theorem [20], a single hidden layer is sufficient for an MLP to compute a uniform \( \epsilon \) approximation to any continuous nonlinear function. In addition, BP (a type of least mean square algorithm) is assured of mean square convergence provided the learning rate is small enough[21].

Consider the discrete-time multi-variable linear time-invariant system

\[ x_{i+1} = f(x_i, u_i) = Ax_i + Bu_i \]  

with feedback control

\[ u_i = g(x_i) = Lx_i \]  

where the matrices \( A \in \mathbb{R}^{n \times n} \) and \( B \in \mathbb{R}^{n \times \omega} \) constitute a controllable pair \((A, B)\), which means that a stabilizing feedback matrix \( L \in \mathbb{R}^{n \times \omega} \) can be obtained. The instantaneous cost associated with this system is a quadratic form given by[22][24]

\[ x_i^T P x_i + u_i^T R u_i \triangleq r(x_i, u_i) \]  

749
where \( P \) is positive semidefinite and \( R \) is positive definite and both are symmetric. Using equation (10) for this LQR problem, the optimal control law can be deduced as follows:

\[
L' = -\left( B^T KB + R \right)^{-1} B^T KA \tag{21}
\]

where \( K \) is the unique positive semidefinite solution to the following discrete algebraic Riccati equation (DARE)\(^{[25]-[27]}\):

\[
K = A^T \left[ K - KB (B^T KB + R)^{-1} B^T K \right] A + P \tag{22}
\]

The objective is to derive the iteration expressions of direct HDP control in the LQR problem and to prove that it converges to equation (21). For this deterministic system, the discounted sum of the cost (the state-value function) is

\[
V(x) = \sum_{k=0}^{\infty} \gamma^k r(x, u)
\]

\[
= \sum_{x_i} \gamma^k (P + L RL)x_{i,1}
\]

\[
= x^T \left[ \sum_{x_i} \gamma^k \left( (A + BL)^T (P + L RL) (A + BL) \right) \right] x.
\]

\[
\triangleq x^T F x
\]

This sum is convergent and, the matrix \( F \) in the quadratic form is positive semi-definite and symmetric.

Direct HDP is based on the estimation of the action-value function; therefore,

\[
Q(x_i, u_i) = r(x_i, u_i) + V(x_{i,1})
\]

\[
= x^T \left( G + A^T \right) F \left( A \right) x_i
\]

\[
\triangleq x^T \left( H_{u} \right) x_i
\]

\[
\triangleq x_i^T \left( H_{u} x_i \right)
\]

where \( H \) is positive semidefinite and \( Q(x_i, u_i) \) is in the quadratic form. The approximation of the value function using the critic network is just the calculation of \( H \) through the MLP neural network; therefore, the above action-value function can be expressed as \( Q(x_i, u_i, H_k) \).

If the system model is known, the optimal cost (reward-to-go) and the optimal policy can be obtained by using equation (24). However, the available information is only the input and output of the plant. The role of the action network is to obtain \( u' = \arg \min_u Q(x, u) \) in this condition, and the method employed is the BP algorithm, which is based on the gradient \( \partial Q / \partial u \) backpropagated through the critic network. Since

\[
\frac{\partial Q(x, u)}{\partial u} = \frac{\partial}{\partial u} \left( x^T \left( H_{u} x_i \right) \right) x_i
\]

\[
= 2H_{u} x + 2H_{u} u
\]

at the minimum point, \( \partial Q / \partial u = 0 \) and the optimal control should be

\[
u = -H_{u}^T H_{u}^{-1} L \hat{x}
\]

therefore, a control law like equation (26) is obtained by using only \( Q \).

Considering equation (3), the desired cost function at time \( k \) is

\[
d(x,u,r,f,g,H_k) = Q(k-1) - r(k)
\]

\[
= Q(x_i, u, H_k) - r(f(x_i), g(f(x_i)))
\]

The update of the critic network is implemented through the minimization of the square of the error between equation (27) and the real cost function at time \( k \):

\[
H_k = \arg \min_h E \left\{ \delta(x_i, u, r, f, g, H_k) \right\}
\]

According to the definitions in the Appendix, let \( z^T = (x_i^T, u^T), h = \nu(H) \); then, equation (28) can be transformed to

\[
h_k = \arg \min_h \left\{ \left\| \delta(x_i, u, r, f, g, H_k) \right\|^2 \right\}
\]

The only minimum exists in this equation

\[
h_k = \left( E_k \right)^{-1} \left\{ E_k \delta(x_i, u, r, f, g, H_k) \right\}
\]

In addition, equation (27) can be expressed as

\[
d = z^T H_k z - z^T (A + Bu) (LA + Bu) z
\]

By substituting equation (26) and (31) for equation (30), we get

\[
h_k = \nu \left( H_{k-1} - (A + Bu) (LA + Bu) (A + Bu)^T \right)
\]

On writing the equation in the matrix form

\[
H_k = H_{k-1} - (A + Bu) b_k (A + Bu)^T
\]

Equation (33) is the iteration formula of matrix \( H \), and it also is the update of the critic network. An error term is decreased at each time step, and it is determined by the plant model, the control in the previous time step, and the the weight matrices in the quadratic performance index.

This iteration has been proven to converge to the optimal parameter matrix \( L^* \), as shown in equation (21). For more information, see the example in \([28]\).

IV. LEARNING ABILITY OF DIRECT HDP CONTROLLER

A. Two-area System and SVC Supplementary Control

The learning ability of the direct HDP controller is demonstrated in the two-area system shown in Fig. 3\(^{[29]}\); four generators are located in two areas, between which two parallel long tie lines run from bus 7 to bus 9.
An SVC composed of an FC and a TCR is placed at the center of the tie lines to support the voltage and suppress swings. The traditional proportional-integral (PI) method is used for regulation, but this method by itself cannot guarantee security when a three-phase short circuit fault occurs near bus 9 on the tie line; therefore supplementary control is needed. In [29], Kundur’s design is based on the conventional pole-placement method (C1), and the corresponding block diagram is shown in Fig.4. The structure of this controller is identical to that of power system stabilizer (PSS), which includes a washout block and two lead-lag blocks that adjust the phase relationship between the input and output. Both these controllers provide additional damping signals to the main voltage regulators—one for the SVC and the other for the generator.

Here, $V_{\text{supmax}} = -V_{\text{supmin}} = 0.1$, $V_{\text{sup}}$ is the output, the input signal $I_{\text{line}}$ is the magnitude of the current in the line between buses 9 and 10 and is chosen according to the computations of system observability. The frequency response characteristics of the transfer function between the SVC input and this current also reveal a high gain at the frequency of the dominant oscillation mode [13].

**B. Two Direct HDP Control Approaches**

In order to compare the control performance of the direct HDP controller and that of C1, the same input signal $I_{\text{line}}$ is used. The phase adjustment of the supplementary control input signal is critical with regard to the oscillation damping problem, and it is implemented through the two phase-shift blocks in C1. However, if only one variable is inputted, the ordinary MLP neural network is not capable of changing the phase to the desired. For example, only one input signal $\sin(t)$ is not sufficient to train a two-layer neural network to learn $\sin(t - \pi/2)$. To achieve the control goal, an additional input—the differential of $I_{\text{line}}$—is necessary. An approximate differentiator as that given in formulation (34) is employed to reduce the noise amplification in the differential calculation.

$$\frac{1}{\tau_2 - \tau_1} * \left( \frac{1}{\tau_1 \tau_2 + 1} - \frac{1}{\tau_1 \tau_3 + 1} \right) 0 < \tau_1 < \tau_2 \quad (34)$$

The inter-area power oscillation is an interaction process among generator groups in different regions, and WAMS can provide a system-level view of the disturbance dynamics. The average rotor speed deviation of the generators ($\Delta \omega_{\text{inter-area}}$) is a good choice to reflect the swing nature directly. For the two-area system,

$$\Delta \omega_{\text{inter-area}} = \frac{(\omega_1 + \omega_2) - (\omega_3 + \omega_4)}{2} \quad (35)$$

where $\omega_i$ ($i = 1, 2, 3, 4$) is the rotor speed of generators $i$.

On the basis of the two abovementioned approaches, the structures of the supplementary damping controller using the direct HDP method are shown in Fig.5.

**C. Simulation Results**

The basic scheme is the same as that given in [23], i.e., the same load flow calculation results and the same disturbance in which a three-phase short circuit occurs near bus 9 and is cleared by tripping the line between buses 8 and 9 74 ms later. This case is used to train the direct HDP controller from random initialization, and a typical learning process of the C3 controller including two trials is shown in Fig.6.
V. CONCLUSIONS AND FUTURE WORK

The nonlinearities, uncertainties, and coordinate design are the three main problems in the stability control of a large-scale power system. In this paper, the direct HDP method that can converge to the optimal solution in the LQR problem is employed to damp low-frequency oscillations. Based on the real system responses instead of the exact system model, a direct HDP controller avoids the influences of modeling for nonlinearities and uncertainties. Different controllers share one common cost function so that the design and adaptations can be easily coordinated. The performance of the direct HDP controller is validated through two DC power modulation controllers in the CSG, and the advanced performances compared with the traditional control approach are presented.

APPENDIX

A. Definition 1

For a vector $x \in \mathbb{R}^n$, there is $\tau \in \mathbb{R}^{(n+1)/2}$, and

$$\tau = (x_1^2, \ldots , x_n^2, x_1, \ldots , x_n)^T,$$

i.e. all the quadratic functions over the elements of $x$.

B. Definition 2

For a square matrix $M \in \mathbb{R}^{n \times n}$, there is a vector $v(M) \in \mathbb{R}^{(n+1)/2}$, whose element is $M_{ii} + M_{ii} \ldots v(M)$ is ordered such that $x^T \tau x = \tau \cdot v(M)$ is satisfied. Note that if $M$ is symmetric, it can be retrieved from the components in $v(M)$.

REFERENCES


