WHAT IS CLAIMED IS:

1. A dynamical symmetry modeling based offline reinforcement learning method, applied to a policy optimization, comprising:

obtaining an offline data set;

processing an original state-action pair in the offline data set by an encoder and a bidirectional dynamical model to obtain a low-dimensional state-action pair and a current time inversion state variation; and

inputting the low-dimensional state-action pair to a value network, and obtaining a value of a state-action value function corresponding to the original state-action pair,

wherein the current time inversion state variation comprises a forward time inversion state variation and a reverse time inversion state variation, and the forward time inversion state variation and the reverse time inversion state variation are used for setting a corresponding policy constraint item to obtain an optimized behavior policy.

2. The method of claim 1, wherein the processing the original state-action pair in the offline data set by the encoder and the bidirectional dynamical model to obtain the low-dimensional state-action pair and the current time inversion state variation comprises:

processing the original state-action pair by the encoder to obtain the low dimensional state-action pair; and

processing a low-dimensional state-action pair at a current moment and a low-dimensional state-action pair at a next moment respectively by the bidirectional dynamical model, and obtaining the forward time inversion state variation and the reverse time inversion state variation through a preset loss function.

3. The method of claim 2, wherein the preset loss function comprises a dynamical loss function, a low-dimensional dynamical loss function and a reconstruction loss function,

wherein formulas of the dynamical loss function, the low-dimensional dynamical loss function and the reconstruction loss function are as follows:

$$L_{dl} = \left\| \Delta X(s, s') - D_{\varphi}(\dot{Z}_{\theta}(X)) \right\|_2^2;$$

$$L_{ldl} = \left\| E_{\varphi}(\Delta X(s, s')) - \dot{Z}_{\theta}(X) \right\|_2^2;$$

$$L_{rl} = \left\| X(s, a) - D_{\varphi}(Z(X)) \right\|_2^2,$$

wherein, s is a state in the offline data set, a is an action in the state of s in the offline data set, $\Delta X(s, s')$ is a state variation at the next moment, s' is a state at the next moment in the offline data set, θ and φ are hyper-parameters, $\dot{Z}_{\theta}(X)$ is the forward time inversion state variation obtained by processing of the bidirectional dynamic model, $D_{\varphi}(\dot{Z}_{\theta}(X))$ is a state-action variation at the next moment obtained by decoding the forward time inversion state variation, $E_{\varphi}(\Delta X(s, s')$ is the forward time inversion state variation obtained by directly encoding of the encoder, $X(s, a)$ is the

original state-action pair in the offline data set and $D_{\varphi}(Z(X))$ is the original state-action pair obtained by decoding the low-dimensional state-action pair.

4. The method of claim 1, wherein the inputting the low-dimensional state-action pair to the value network, and obtaining the value of the state-action value function corresponding to the original state-action pair comprises:

calculating a low-dimensional policy state variation at the next moment based on a low-dimensional state and a low-dimensional policy output action in the low-dimensional state-action pair;

calculating the forward time inversion state variation based on the low-dimensional state and a low-dimensional representation of an original state at the next moment in the low-dimensional state-action pair;

obtaining the value of the state-action value function by taking a difference value between the low-dimensional policy state variation at the next moment and the forward time inversion state variation as a preset constraint item.

5. The method of claim 4, further comprising:

obtaining a behavior policy corresponding to an original state in the original state-action pair by taking the value of the state-action value function and the difference value between a low-dimensional policy action and the low-dimensional action in the low-dimensional state-action pair as a constraint item; or,

obtaining a behavior policy corresponding to an original state in the original state-action pair by taking the value of the state-action value function, and the difference value between the forward time inversion state variation obtained based on the low-dimensional state-action pair and the forward time inversion state variation obtained based on the low-dimensional state and the low-dimensional policy action in the low-dimensional state-action pair as a constraint item; or,

obtaining a behavior policy corresponding to an original state in the original state-action pair by taking the value of the state-action value function, the difference value between the forward time inversion state variation obtained based on the low-dimensional state-action pair and the forward time inversion state variation obtained based on the low-dimensional state and the low-dimensional policy action in the low-dimensional state-action pair, and the difference value between the low-dimensional policy action and the low-dimensional action in the low-dimensional state-action pair as a constraint item; or,

obtaining a behavior policy corresponding to an original state in the original state-action pair by taking the value of the state-action value function, the difference value between the forward time inversion state variation obtained based on the low-dimensional state-action pair and the forward time inversion state variation obtained based on the low-dimensional state and the low-dimensional policy action in the low-dimensional state-action pair, the difference value between the low-dimensional policy action and the low-dimensional action in the low-dimensional state-action pair, and the difference value between the low-dimensional policy action and an original action in the original state-action pair as a constraint item.

6. The method of claim 1, wherein the bidirectional dynamical model is obtained

by training a time inversion loss function.

7. A dynamical symmetry modeling based offline reinforcement learning apparatus, applied to a policy optimization, comprising:

a data obtaining module, used for obtaining an offline data set;

a data processing module, used for processing an original state-action pair in the offline data set by an encoder and a bidirectional dynamical model to obtain a low-dimensional state-action pair and a current time inversion state variation; and

a value obtaining module, used for inputting the low-dimensional state-action pair to a value network, and obtaining a value of a state-action value function corresponding to the original state-action pair,

wherein the current time inversion state variation comprises a forward time inversion state variation and a reverse time inversion state variation, and the forward time inversion state variation and the reverse time inversion state variation are used for setting a corresponding policy constraint item to obtain an optimized behavior policy.

8. An electronic device, comprising:

a processor;

a memory; and

a computer program stored on the memory and executable on the processor,

wherein the computer program, when executed by the processor, causes the electronic device to perform steps of the method of any one of claims 1 to 6.

9. A non-transitory computer-readable storage medium storing computer programs, when executed by a processor, cause the processor to perform steps of the method of any one of claims 1 to 6.

10. A computer program product comprising computer programs, when executed by a processor, cause the processor to perform steps of the method of any one of claims 1 to 6.