1 Bellman's learning equation for post-decision state

Final learning equations are, at nth iteration and in time t

$$\bar{V}^n(S^x_{t-1}) = (1-\alpha^n)\bar{V}^{n-1}(S^x_{t-1}) + \alpha^n(\min_{x_t}[C^n(S_t,x_t) + \beta\bar{V}^n(S^x_t)]) \tag{1}$$

$$x_t(S^t) = arg\min_{x_t}[C^n(S_t,x_t) + \beta\bar{V}^n(S^x_t)] \tag{2}$$

2 Value Function Approximation

In general, the regression VFA is written as

$$\bar{V}^n(S^x_t|\theta) = \sum_{f\in\mathcal{F}} \theta_f\phi_f(S^x_t), \tag{3}$$

where $\bar{V}^n(S^x_t|\theta)$ is the estimated value function in post-decision state S^x_t, x is the decision taken in pre-decision state S_t, θ_f is a parameter vector, ϕ_f is a set of basis functions, and $f \in \mathcal{F}$ is called feature, which is an integer that denotes the number of terms in the regression model. In ADP, the above regression value function approximation can reduce large state-spaces and their values into a small number of features f, which avoids the storage of millions of individual states and their values.

2.1 Classical Wavelets:

Wavelet is a topic in advanced digital signal processing. Unlike the Fourier functions that are long, periodic and sinusoidal, wavelets are very short waves. Their extensive usage is in signal processing for data compression, data decomposition in time at multiple frequency levels (multiresolution analysis), pattern recognition, denoising of data, and anomaly detection in data. There are several types of wavelets and the most popular are Haar (mother wavelet), Daubechies, coiflets and symlets. These are used to analyze data in 1-D and 2-D Euclidean space. Like the Taylor and Fourier series, wavelets are used for function representation: a property exploited in this research. They can be stretched (dilated to different frequency levels) and moved in time (translated) to fit any shape of an underlying function.

Wavelets have a natural tendency to represent a large function on a very compact set of scaling and wavelet coefficients through multiresolution analysis: a property that is unique to wavelets and very useful for VFA. Mathematically, the wavelet-based value function approximation is achieved by the successive decomposition of the value function into several frequency levels in the wavelet domain (multiresolution analysis), which is written as

$$\bar{V}(S^x) = \sum_{k=-\infty}^{\infty} c_{(j_0,k)_{f1}}\phi_{(j_0,k)_{f1}}(S^x) + \sum_{j=j_0}^{\infty}\sum_{k=-\infty}^{\infty} d_{(j,k)_{f2}}\psi_{(j,k)_{f2}}(S^x), \tag{4}$$

where $c_{(j_0,k)_{f1}}$ and $d_{(j,k)_{f2}}$ are scaling and wavelet coefficients respectively (they replace θ in equation (3)), $\phi_{(j_0,k)_{f1}}(S^x)$ and $\psi_{(j,k)_{f2}}(S^x)$ are scaling and wavelet basis functions respectively (they replace ϕ in equation (3)), j is the dilation index (j_0 is the coarsest frequency level of decomposition), k is the translation index (for time) used in classical wavelet theory, and f_1 and f_2 are the number of features for the scaling and wavelet functions respectively. The coefficients c and d are obtained using the convolution operation *i.e.* inner products, $c = \langle\bar{V}(S^x), \phi(S^x)\rangle$ and $d = \langle\bar{V}(S^x), \psi(S^x)\rangle$.

Unlike the MDP with known transition probabilities in which the value functions increase monotonically for a given state, learning-based ADP induces oscillations (jumps) in the value of the states. Being short waves on compact support wavelets can represent jumps, cusps, and discontinuities with relative ease and are best suited for VFA.

2.2 Diffusion Wavelets:

Diffusion Wavelet is a compact multi-level representation of a Markov diffusion process on graphs and is used in analyzing n-D data. Classical wavelet is a special case of diffusion wavelet in which the basis functions ϕ and ψ are derived from the mother wavelet, which must be specified. In contrast to classical wavelet, the diffusion wavelet method builds best-basis functions to approximate the state value functions by *exploiting the state-space structure* Σ, which is represented as a graph. In the classical wavelet analysis, which is performed on 1-D Euclidean spaces, dilations by powers of two and translations by integers are applied to a mother wavelet to obtain the wavelet bases. However, for diffusion wavelet, the diffusion operators T acting on functions of the state-space are used to build the wavelet basis functions.

The best basis functions are obtained as follows. Using a sample set (of size M) of system states $S^x \in \Sigma$, a Gaussian kernel is used to build a graph (G, E, W^g) using spectral graph theory. The graph can be imagined as a cloud of interconnected system states in n-D space. The graph (G, E, W^g) represents the structure of state space Σ, where E is the edge and W^g is the weight on the edge. It should be noted that the random walk on graph (G, E, W^g) is a very special case of a Markov chain. The Laplacian operator L of the graph is obtained using spectral graph theory, and $I - L = T^1$ is obtained, where I is an identity matrix and T^1 is the diffusion operator at level $j = 1$. Sparse factorization (QR factorization) of T^1 yields the scaling basis functions ϕ, which are obtained from Q. The wavelet basis functions ψ are then obtained from the scaling basis functions ϕ by sparse factorization of $(I - \phi * \phi)$, where I is an identity matrix. The concept is that the basis functions in T^1 intrinsically represent the structure of the state space Σ and its geometric constraints. To obtain basis functions at the next level, the diffusion operator T^2 is first obtained from R as follows: $T^2 = R \times R^*$ where R^* is the complex conjugate of R. Sparse factorization of T^2 yields the scaling basis functions at the next level, and the procedure is continued until no further decomposition is possible. The number of features f_1 and f_2 are automatically selected by the size of the basis functions and the diffusion wavelet theory guarantees compact representation.

3 IMPLEMENTING WAVELET-BASED VFA for ADP:

1. Define state, action, contribution function $C^n(S_t, x_t)$, and uncertainty model (simulator) to generate the pre-decision state.

2. Define β - fixed discount parameter, α^n - learning parameter, its start value, and decay scheme, and γ^n - exploration parameter, its start value, and decay scheme. Set number of iterations for exploration and learning. γ^n goes to zero at the end of exploration iterations. α goes to about 0.3 by the end of learning. In general, the learning phase has 10 times more iterations than the exploration phase.

3. EXPLORATION PHASE: Run exploration using parameter $0 < \gamma^n < 1$. Take a random number $r = U[0, 1]$. Take random actions, that is, do not use the max or min operator in equation (1) if $r < \gamma^n$, otherwise if $r > \gamma^n$. Decay γ^n and α^n.

4. A sample set of most frequently visited post-decision states S^x of size M (say $M = 1000$ states over 50000 iterations within exploration) and their values $V(S^x)$ is first collected. Arrange M in the descending order of frequency of visits.

5. Using S^x of size M the best basis functions $\phi(S^x)$ and $\psi(S^x)$ for the sample states are obtained using the diffusion wavelet procedure given above.

6. The initial estimates of coefficients c and d are obtained using the convolution operation *i.e.* inner products, $c = \langle \bar{V}(S^x), \phi(S^x) \rangle$ and $d = \langle \bar{V}(S^x), \psi(S^x) \rangle$.

7. LEARNING PHASE: When exploration phase is done, the algorithm moves to learning phase.

8. For the next 1000 iterations, values of future post-decision states $V^n(S_t^x)$ are obtained either from M if the state is in M or from Equation 4.

9. If using equation 4 then the basis functions ϕ and ψ are obtained (updated) using the diffusion wavelet procedure. Replace the last row of M with the new post-decision states (S_t^x).

10. These updated basis functions along with the current values of c, and d are used in equation (4) to obtain the estimate of the last term $\bar{V}^n(S_t^x)$ in equation (1) for the next set of all possible post-decision states (S_t^x).

11. Repeat the above last 3 steps till all actions are evaluated and equation (1) is updated. Continue this for 1000 iterations while keeping a frequency counter for the new states visited within the 1000 runs. If equation (1) updates the value of a post-decision state in M then increment its frequency counter also.

12. At the end of next 1000 runs of the ADP algorithm using equation (1), the next set of most frequently visited post-decision states (S_t^x) is used to update set M by replacing the states with lowest frequency of visits in set M with the new post-decision states from the 1000 runs whose frequencies are higher.

13. Then the basis functions ϕ and ψ are updated using the diffusion wavelet procedure given above. Also update the values of the parameters c and d.

14. Repeat another 1000 iterations and the steps above.

15. Learning continues through several iteration steps until the value functions have converged in a band and MSE has stabilized. When learning is complete, the most recent M sample states, and the values of ϕ, ψ, c, and d are taken as inputs to the learnt stage.

16. LEARNT PHASE OF ADP ALGORITHM: In this stage the ADP algorithm will be tested. The dynamic process will be simulated and the pre-decision state will be obtained.

17. The feasible decisions x in this pre-decision state will generate the possible post-decision states. The most recent M sample states are updated with the post-decision states. Next, ϕ and ψ are updated using the diffusion wavelet procedure. The VFA procedure will provide the estimated value functions of the post-decision states. Equation (2) will then provide the best resource allocation decision for the current pre-decision state. It is optional to update c, and d in the learnt phase.