Solving large structured Markov Decision Problems for perishable inventory management and traffic control

Haijema, R.

Publication date
2008

Appendix B

Relative values in periodic MCs

The content of this section is a reformulation of a technical note by Van der Wal [147].

When a policy π is periodic, a state $\mathbf{x}$ of the underlying Markov chain belongs to one of the D periodic classes. The expected direct costs to incur in a state are stored in the vector $\mathbf{c}$ and the transition probabilities under policy π are stored in matrix $\mathbf{P}$.

The system visits the classes in the order $1, 2, \ldots, D, 1, 2$, etc. and the long run average costs to incur in a slot is class dependent: $g_1, g_2, \ldots, g_D$. Therefore the value vector in a SA algorithm can not be used directly as a relative value vector, as was argued already in Section 1.3.2. In this section, we show how the bias terms are computed for periodic policies. These bias terms are relative values that can be used in a PI algorithm or in a one-step policy improvement algorithm.

The bias terms $\widetilde{\mathbf{v}}^\pi = \mathbf{v}$ and the gain $g^\pi = g$ are the unique solution to the following set of equations:

$$\begin{aligned} \mathbf{v} &= \mathbf{c} + \mathbf{P}\mathbf{v} - \mathbf{g}, && \text{(B.1)} \\ \mathbf{P}\mathbf{g} &= \mathbf{g}, && \text{(B.2)} \\ \mathbf{P}^*\mathbf{v} &= \mathbf{0}. && \text{(B.3)} \end{aligned}$$

where $\mathbf{g} = g \cdot \mathbf{1}$ and

$$\mathbf{P}^* \equiv \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \mathbf{P}^i = \lim_{N\to\infty} \mathbf{P}^N \frac{1}{D} \sum_{d=0}^{D-1} \mathbf{P}^d. \qquad \text{(B.4)}$$

By iteratively substituting $\mathbf{v} = \mathbf{c}^\pi + \mathbf{P}^\pi\mathbf{v} - g\mathbf{1}$ in the righthand-side of Equation (B.1), one gets after n iterations:

$$\mathbf{v} \;=\; \sum_{i=0}^{n-1}\mathbf{P}^i(\mathbf{c}-\mathbf{g}) + \mathbf{P}^n\mathbf{v} = \sum_{i=0}^{n-1}\mathbf{P}^i\mathbf{c} - n\cdot\mathbf{g} + \mathbf{P}^n\mathbf{v}.$$

$\sum_{i=0}^{n-1}\mathbf{P}^i\mathbf{c}$ is the total expected costs over an horizon of n slots when strategy π is applied. In the evaluation of the MC by the SA algorithm, these costs are denoted by $\mathbf{V}_n^\pi$. Hence

$$\mathbf{v} \;=\; \mathbf{V}_n^\pi - n\mathbf{g} + \mathbf{P}^n\mathbf{v}. \tag{B.5}$$

Summing Eq. (B.5) over $n = N$ to $N + D - 1$, and next diving both sides by D yields:

$$\mathbf{v} = \frac{1}{D}\sum_{n=N}^{N+D-1}\mathbf{V}_n^\pi - \frac{N+N+D-1}{2}\cdot\mathbf{g} + \frac{1}{D}\sum_{n=N}^{N+D-1}\mathbf{P}^n\mathbf{v}.$$

Taking the limit for N to ∞ results in:

$$\mathbf{v} = \lim_{N\to\infty}\left(\frac{1}{D}\sum_{n=N}^{N+D-1}\mathbf{V}_n^\pi - N\mathbf{g}\right) - \frac{D-1}{2}\mathbf{g}, \tag{B.6}$$

as $\lim_{N\to\infty}\sum_{n=N}^{N+D-1}\mathbf{P}^n\mathbf{v} = \mathbf{0}$, given Equations (B.3) and (B.4).

Since relative values are unique up to an additive constant, one may subtract any constant vector from it. For example, one may value all states against a fixed reference state, say state $\mathbf{z}$:

$$\lim_{N\to\infty}\frac{1}{D}\sum_{d=0}^{D-1}\left(\mathbf{V}_{N+d}^\pi - V_{N+d}^\pi(\mathbf{z})\cdot\mathbf{1}\right). \tag{B.7}$$

Alternatively, the value vector can be set to the bias terms: the expected difference in receiving the total expected costs over an infinitely long horizon compared to receiving every slot the class dependent average costs:

$$\lim_{N\to\infty}\frac{1}{D}\sum_{d=0}^{D-1}\left(\mathbf{V}_{N+d}^\pi - (N+d)\cdot\cdot\mathbf{g}\right). \tag{B.8}$$