Bellman公式的矩阵化

对于时刻t，$s_{t+1} \in \mathcal{S}$，其状态值函数可以写为如下形式：

$$
\begin{aligned}
v^\pi(s_t) &= E_{T\sim\pi}[\sum_{k=0}^{T-1}\gamma^k r_{t+k+1}] \\
&= E_{a_t\in\mathcal{A}}[E_{s_{t+1}\in\mathcal{S}}[r(s_t,a_t,s_{t+1})+\gamma v^\pi(s_{t+1})]] \\
&= \sum_{a_t\in\mathcal{A}}\pi(a_t|s_t)\sum_{s_{t+1}\in\mathcal{S}}p(s_{t+1}|s_t,a_t)[r(s_t,a_t,s_{t+1})+\gamma v^\pi(s_{t+1})] \\
&= \sum_{a_t\in\mathcal{A}}\pi(a_t|s_t)[\sum_{s_{t+1}\in\mathcal{S}}p(s_{t+1}|s_t,a_t)r(s_t,a_t,s_{t+1})+\sum_{s_{t+1}\in\mathcal{S}}p(s_{t+1}|s_t,a_t)\gamma v^\pi(s_{t+1})] \\
&= \sum_{a_t\in\mathcal{A}}\pi(a_t|s_t)[K(s_t,a_t)+\sum_{s_{t+1}\in\mathcal{S}}p(s_{t+1}|s_t,a_t)\gamma v^\pi(s_{t+1})] \\
&= \sum_{a_t\in\mathcal{A}}\pi(a_t|s_t)K(s_t,a_t)+\sum_{a_t\in\mathcal{A}}\pi(a_t|s_t)\sum_{s_{t+1}\in\mathcal{S}}p(s_{t+1}|s_t,a_t)\gamma v^\pi(s_{t+1}) \\
&= M(s_t)+\sum_{a_t\in\mathcal{A}}\pi(a_t|s_t)\sum_{s_{t+1}\in\mathcal{S}}p(s_{t+1}|s_t,a_t)\gamma v^\pi(s_{t+1})
\end{aligned}
$$

其中$M(s_t)$为状态s_t时做出一步action的期望收益，其中：

$$
\begin{gathered}
\sum_{s_{t+1}\in\mathcal{S}}p(s_{t+1}|s_t,a_t)\gamma v^\pi(s_{t+1}) \\
=\begin{bmatrix} p(s_{t+1}^1|s_t^1,a_t) & . & . & . & p(s_{t+1}^{|\mathcal{S}|}|s_t^{|\mathcal{S}|},a_t) \end{bmatrix}\gamma\begin{bmatrix} v^\pi(s_{t+1}^1) \\ . \\ . \\ . \\ v^\pi(s_{t+1}^{|\mathcal{S}|}) \end{bmatrix}
\end{gathered}
$$

则：

$$
\begin{gathered}
\sum_{a_t\in\mathcal{A}}\pi(a_t|s_t)\sum_{s_{t+1}\in\mathcal{S}}p(s_{t+1}|s_t,a_t)\gamma v^\pi(s_{t+1}) \\
=\begin{bmatrix} \pi(a_t^1|s_t) & . & . & . & \pi(a_t^{|\mathcal{A}|}|s_t) \end{bmatrix}\begin{bmatrix} p(s_{t+1}^1|s_t,a_t^1) & . & . & . & p(s_{t+1}^{|\mathcal{S}|}|s_t,a_t^1) \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ p(s_{t+1}^1|s_t,a_t^{|\mathcal{A}|}) & . & . & . & p(s_{t+1}^{|\mathcal{S}|}|s_t,a_t^{|\mathcal{A}|}) \end{bmatrix}\gamma\begin{bmatrix} v^\pi(s_{t+1}^1) \\ . \\ . \\ . \\ v^\pi(s_{t+1}^{|\mathcal{S}|}) \end{bmatrix} \\
=\begin{bmatrix} \sum_{k=1}^{|\mathcal{A}|}\pi(a_t^k|s_t)p(s_{t+1}^1|s_t,a_t^k) & . & . & . & \sum_{k=1}^{|\mathcal{A}|}\pi(a_t^k|s_t)p(s_{t+1}^{|\mathcal{S}|}|s_t,a_t^k) \end{bmatrix}\gamma\begin{bmatrix} v^\pi(s_{t+1}^1) \\ . \\ . \\ . \\ v^\pi(s_{t+1}^{|\mathcal{S}|}) \end{bmatrix} \\
=\begin{bmatrix} p(s_{t+1}^1|s_t) & . & . & . & p(s_{t+1}^{|\mathcal{S}|}|s_t) \end{bmatrix}\gamma\begin{bmatrix} v^\pi(s_{t+1}^1) \\ . \\ . \\ . \\ v^\pi(s_{t+1}^{|\mathcal{S}|}) \end{bmatrix}
\end{gathered}
$$

表述成矩阵形式，同时考虑所有$s_{t+1} \in \mathcal{S}$的可能取值则可以写为如下形式：

$$\begin{bmatrix} v^{\pi}(s_t^1) \\ v^{\pi}(s_t^2) \\ \cdot \\ \cdot \\ \cdot \\ v^{\pi}(s_t^{|\mathcal{S}|}) \end{bmatrix} = \begin{bmatrix} M(s_t^1) \\ M(s_t^2) \\ \cdot \\ \cdot \\ \cdot \\ M(s_t^{|\mathcal{S}|}) \end{bmatrix} + \begin{bmatrix} p(s_{t+1}^1|s_t^1) & \cdot & \cdot & \cdot & p(s_{t+1}^{|\mathcal{S}|}|s_t^1) \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ p(s_{t+1}^1|s_t^{|\mathcal{S}|}) & \cdot & \cdot & \cdot & p(s_{t+1}^{|\mathcal{S}|}|s_t^{|\mathcal{S}|}) \end{bmatrix} \gamma \begin{bmatrix} v^{\pi}(s_{t+1}^1) \\ \cdot \\ \cdot \\ \cdot \\ v^{\pi}(s_{t+1}^{|\mathcal{S}|}) \end{bmatrix}$$

又由于Markov性质：

$v^{\pi}(s_{t+1}^i) = v^{\pi}(s_t^i)$

上式可以写成：

$\mathbf{V} = \mathbf{M} + \gamma \mathbf{P} \mathbf{V}$

$\mathbf{V} = (\mathbf{I} - \gamma \mathbf{P})^{-1} \mathbf{M}$