Q-Learning con grafos de coordinación

Requiere grafo de coordinación del sistema $G = (\mathcal{V}, \mathcal{E})$; orden de eliminación de los agentes $\mathcal{O} \subseteq \mathcal{V}$; conjunto de vecinos para cada agente $\Gamma(i)$

Para cada arco $(i, j) \in \mathcal{E}$ **hacer**
    Inicialice $Q_{ij}(s_{ij}, a_i, a_j)$ de manera optimista
Fin Para
Para cada episodio **hacer**
    Inicialice el estado observado s_i para todos los agentes $i \in \mathcal{V}$
    Para cada periodo de decisión k **hacer**
        Obtenga la acción conjunta $\mathbf{a}$ con $\epsilon-$greedy como método de selección
        Para cada agente $i \in \mathcal{V}$: aplique $a_i \in \mathbf{a}$, observe r_i y s_i^k
        Para todos los arcos $(i, j) \in \mathcal{E}$: forme $s_{ij}^k = s_i^k \cup s_j^k$
        Obtener $\mathbf{a}^* = \text{máx}_{\mathbf{a}}\, Q(\mathbf{s^k}, \mathbf{a}) \rightarrow \textsc{EliminaciónVariable}(s_{ij}^k \forall (i, j) \in \mathcal{E})$
        Para cada arco $(i, j) \in \mathcal{E}$ **hacer**
            $Q_{ij}(s_{ij}^{k-1}, a_i^{k-1}, a_j^{k-1}) := (1-\alpha) Q_{ij}(s_{ij}^{k-1}, a_i^{k-1}, a_j^{k-1}) + \alpha \left[\frac{r_i^k}{|\Gamma(i)|} + \frac{r_j^k}{|\Gamma(j)|} + \gamma Q_{ij}(s_{ij}^k, a_i^*, a_j^*) \right]$
        Fin Para
        Para cada agente $i \in \mathcal{V}$ actualice $s_i^{k-1} \leftarrow s_i^k$
    Fin Para
Fin Para

Algoritmo de eliminación de variable

Función ELIMINACIÓNVARIABLE(s_{ij}^k)
  Para cada agente $i \in \mathcal{O}$ **hacer**
    Calcular función de pago condicional ϕ
    si el agente i es el primero a eliminar **entonces**
      $k \equiv$siguiente agente a eliminar, $k \in \Gamma(i)$
      $\phi_{kj} = \text{máx}_{a_i} \left[\sum_{j \in \Gamma(i)} Q_{ij}(s_{ij}^k, a_i, a_j)\right]$
    sino si el agente i es el último a eliminar **entonces**
      $\phi_i = \text{máx}_{a_i}[\phi_i(a_i)]$
    sino
      $k \equiv$siguiente agente a eliminar, $k \in \Gamma(i)$
      $\phi_{kj} = \text{máx}_{a_i} \left[\sum_{j \in \Gamma(i)} Q_{ij}(s_{ij}^k, a_i, a_j) + \phi_{ik}(a_i, a_k)\right]$
    Fin si
    Transmitir ϕ al siguiente agente a eliminar
  Fin Para

  Para cada agente i en el orden inverso de $\mathcal{O}$ **hacer**
    si el agente i fue el último eliminado **entonces**
      Obtener $a_i^* = \text{argmax}_{a_i}[\phi_i(a_i)]$
      Transmitir al agente anterior a_i^*
      Transmitir $\text{máx}_{\mathbf{a}} Q(\mathbf{s}, \mathbf{a}) = \text{máx}_{a_i} \phi_i(a_i)$
    sino
      Esperar de los agentes vecinos a_k^*, a_j^* y $\text{máx}_{\mathbf{a}} Q(\mathbf{s}, \mathbf{a})$
      Obtener $a_i^* = \text{argmax}_{a_i}[\phi_{kj}(a_k^*, a_j^*)]$
    Fin si
  Fin Para
  Retornar $a_i^* \quad \forall \quad i \in \mathcal{V}$
Fin Función

Q-Learning con *best response*

Requiere conjunto de agentes N, conjunto de vecinos para cada agente NB_i

Para cada agente $i \in \{1, 2, \cdots, |N|\}$ **hacer**
  Para cada agente vecino $j \in \{1, 2, \cdots, |\mathrm{NB}_i|\}$ **hacer**
    Inicializar $Q_{i,\mathrm{NB}_i[j]}(s_{i,\mathrm{NB}_i[j]}, a_i, a_{\mathrm{NB}_i[j]})$ de manera optimista
    Inicializar modelo para la estimación de política $\theta_{i,\mathrm{NB}_i[j]}(s_{i,\mathrm{NB}_i[j]}, a_{\mathrm{NB}_i[j]}) = 1/|\mathcal{A}_{\mathrm{NB}_i[j]}|$
  Fin Para
Fin Para

Para cada episodio **hacer**
  Inicializar el estado observado s_i para todos los agentes $i \in N$
  Para cada periodo de decisión k **hacer**
    Obtener la acción conjunta **a** con $\epsilon-$greedy como método de selección
    Aplicar $a_i \in \mathbf{a}$ para cada agente $i \in \mathcal{V}$
    Para cada agente $i \in \{1, 2, \cdots, |N|\}$ **hacer**
      Para cada agente vecino $j \in \{1, 2, \cdots, |\mathrm{NB}_i|\}$ **hacer**
        Observar s_i^k, r_i^k, $s_{\mathrm{NB}_i[j]}^k$ y $a_{\mathrm{NB}_i[j]}^{k-1}$
        Formar estado conjunto $s_{i,\mathrm{NB}_i[j]}^k = s_i^k \cup s_{\mathrm{NB}_i[j]}^k$ y acción conjunta $a_{i,\mathrm{NB}_i[j]}^{k-1} = a_i^{k-1} \cup a_{\mathrm{NB}_i[j]}^{k-1}$
        Actualizar modelo de estimación de la política para el vecino $\mathrm{NB}_i[j]$:

$$\theta_{i,\mathrm{NB}_i[j]}(s_{i,\mathrm{NB}_i[j]}^{k-1}, a_{\mathrm{NB}_i[j]}^{k-1}) = \frac{v(s_{i,\mathrm{NB}_i[j]}^{k-1}, a_{\mathrm{NB}_i[j]}^{k-1})}{\sum_{a_{\mathrm{NB}_i[j]} \in \mathcal{A}_{\mathrm{NB}_i[j]}} v(s_{i,\mathrm{NB}_i[j]}^{k-1}, a_{\mathrm{NB}_i[j]})}$$

        Encontrar la *best response* respecto al vecino $\mathrm{NB}_i[j]$:

$$br_i^k = \max_{a_i \in \mathcal{A}_i} \left[\sum_{a_{\mathrm{NB}_i[j]} \in \mathcal{A}_{\mathrm{NB}_i[j]}} Q_{i,\mathrm{NB}_i[j]}(s_{i,\mathrm{NB}_i[j]}^k, a_{i,\mathrm{NB}_i[j]}) \times \theta_{i,\mathrm{NB}_i[j]}(s_{i,\mathrm{NB}_i[j]}^k, a_{\mathrm{NB}_i[j]}) \right]$$

Actualizar factores Q:

$$Q^{k}_{i,\mathrm{NB}_i[j]}(s^{k-1}_{i,\mathrm{NB}_i[j]}, a^{k-1}_{i,\mathrm{NB}_i[j]}) = (1-\alpha) Q^{k-1}_{i,\mathrm{NB}_i[j]}(s^{k-1}_{i,\mathrm{NB}_i[j]}, a^{k-1}_{i,\mathrm{NB}_i[j]}) + \alpha \left[r^{k}_{i} + \gamma b r^{k}_{i} \right]$$

Actualizar $s^{k-1}_{i,\mathrm{NB}_i[j]} \leftarrow s^{k}_{i,\mathrm{NB}_i[j]}$

Fin Para

Acción que corresponde a *best response* con todos los vecinos:

$$a^{k}_{i} = \underset{a_i \in \mathcal{A}_i}{\text{argmax}} \left[\sum_{j \in \{1,2,\cdots,|\mathrm{NB}_i|\}} \sum_{a_{\mathrm{NB}_i[j]} \in \mathcal{A}_{\mathrm{NB}_i[j]}} Q_{i,\mathrm{NB}_i[j]}(s^{k}_{i,\mathrm{NB}_i[j]}, [a_i \cup a_{\mathrm{NB}_i[j]}]) \times \theta_{i,\mathrm{NB}_i[j]}(s^{k}_{i,\mathrm{NB}_i[j]}, a_{\mathrm{NB}_i[j]}) \right]$$

Fin Para

Fin Para

Fin Para

Discretización del espacio de estados

Procedimiento VECTOR QUANTIZATION

  Para cada hora $h \in 0, 1, \cdots, 23$ **hacer**

    Para cada fase k **hacer**

      Observar pasivamente el sistema durante n días

      Guardar todos los vectores estado $s \in \mathbb{R}^{2(2+2i)-1}$ y concatenar las n observaciones en una matriz $\mathcal{S}$, donde cada columna j represente una dimensión del estado

      Calcular la desviación estándar de cada columna $\textbf{Stdev}(\mathcal{S})_j$

      Calcular el número de centroides $C_{h,k} = \sum_j \textbf{Stdev}(\mathcal{S})_j$

      Aplicar K-Means sobre $\mathcal{S}$ usando el número de centroides calculado $\rightarrow \mathrm{VQ}_{h,k}(\mathcal{S})$

    Fin Para

  Fin Para

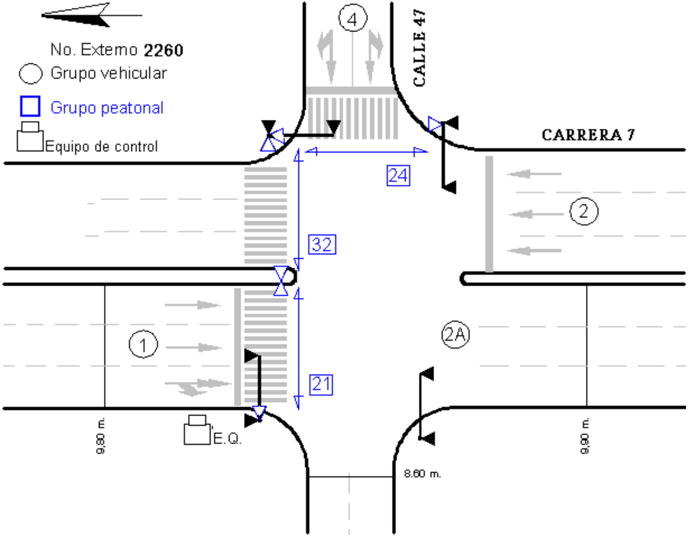
Fin Procedimiento

Procedimiento DISCRETIZACIÓN(estado continuo s_{ij})

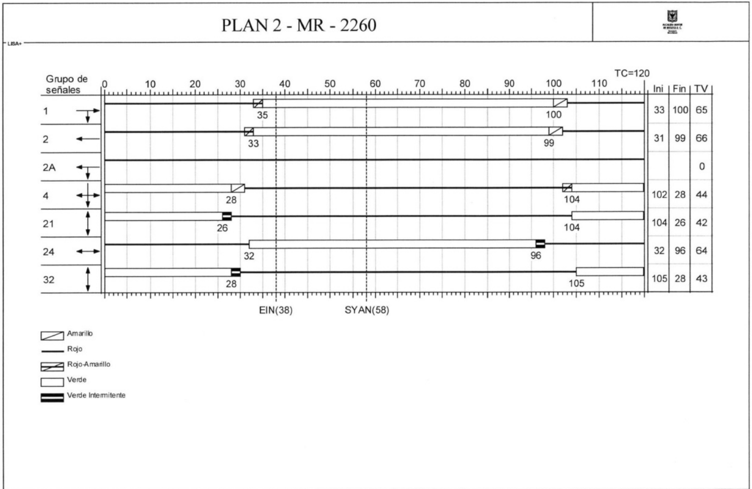
  Retornar $\mathrm{argmin}_{y \in \mathrm{VQ}_{h,k}(s_{ij})} \{d_E(s_{ij}, y)\}$ ▷ euclidiana

Fin Procedimiento

Ejemplo de un plan de fases



Ejemplo de un plan de fases



ZONA	PALOQUEMAO I				
INTERSECCIÓN	AK 7 X CL 47				
ELABORÓ.	ING. OTONIEL CRUZ GARCIA	VERSIÓN	1	FECHA	10/09/2014
CONTRATO No	2014447	Vo.Bo.		GRUPO	2010

De control óptimo a aprendizaje por refuerzo

La política π corresponde a un conjunto de funciones $\pi = \{\mu_0, \mu_1, \cdots\}$ en donde cada μ_k mapea el estado s_k en una acción de control $a_k = \mu_k(s_k)$, tal que $\mu_k \in A(s_k)$ para todo s_k.

Costo de seguir la política π:

$$J_\pi(s) = \lim_{N\to\infty} \mathbf{E}\left[\sum_{k=0}^{N} \gamma^k r(s_k, \mu_k(s_k), s_{k+1}) | s = s_0\right]$$

La política óptima $\pi^* \in \Pi$ corresponde a aquella que maximiza el funcional:

$$J_\pi^*(s_0) = \max_{\pi\in\Pi} J_\pi(s_0)$$

De control óptimo a aprendizaje por refuerzo

Solucionar la ecuación de Bellman para encontrar el valor óptimo del funcional para una iteración del problema:

$$J_\pi^*(s_i) = \max_{a \in A(s_i)} \sum_{j=1}^{n} p_{s_i,s_j}(a) \left(r(s_i, a, s_j) + \gamma J(s_j) \right), \quad i = 1 \cdots n$$

D. Bertsekas muestra que se puede obtener el funcional a partir de factores Q óptimos:

$$Q^*(s_i, a) = \sum_{j=1}^{n} p_{s_i,s_j}(a) \left(r(s_i, a, s_j) + \gamma J^*(s_j) \right), \quad \forall\, (s_i, a)$$

De control óptimo a aprendizaje por refuerzo

Reemplazando $J^*(s_i) = \text{máx}_{a \in A(s_i)} Q^*(s_i, a)$, se obtiene una ecuación de Bellman para sistemas de estados aumentados:

$$Q^*(s_i, a) = \sum_{j=1}^{n} p_{s_i, s_j}(a) \left(r(s_i, a, s_j) + \gamma \max_{a' \in A(s_j)} Q^*(s_j, a') \right), \quad \forall\, (s_i, a)$$

Una vez calculados los $Q^*(s_i, a) \,\forall\, (s_i, a)$, es directo obtener la política óptima π^*:

$$\mu^*(s_i) = \operatorname*{argmax}_{a \in A(s_i)} Q^*(s_i, a) \quad \forall\, s_i$$