MDP

We are given a set of disconnected substations $\mathcal{C}$, with cardinality $|\mathcal{C}| = N$ (< 20), between two remotely controlled substations, where these last will not be included in the problem, since they are already reconnected.

We define the cost of the process as the time each underlying user of each substation remains disconnected. So we compute the cost multiplying the time of disconnection for the number of users of a substation, and we sum them. We want to minimize this cost.

In this MDP we have that

- the **state** is $s = (x_g, v_k, \{v\})$ where x_g is the position of the fault, $v_k \in \mathcal{C}$ is the substation in which I am, and $\{v\}$ is the set of the still disconnected substations (we could also use the substation already reconnected, since they are complementary: given one of the two sets I can always retrieve the other). We have that v_k and $\{v\}$ are *observable*, while x_g is *hidden*.
- the **action** is the intervention I do in the specific substation I decide to visit, so $a \in \mathcal{C}$. Actually, since I can visit only disconnected substations, we have that $a \in \{v\}$.
- the **reward** is the cost of going in a certain substation.
- the **next state** is $s' = (x_g, v_{k+1} = a, \{v'\})$ with $\{v'\} \subseteq \{v\} \backslash a$.

The system is **deterministic**, so given an admissible action a we will surely perform it and end up in the state in which that action leads. We can say that there are no execution errors, and I will always do what I want to do. So, mathematically we have that the **transition probability** is

$$p(s' \mid s, a) = \mathbb{I}(s' = s + a) = \delta_{s', s+a} = \begin{cases} 1 & \text{if } s' = s + a \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where δ is the Kronecker delta and $\mathbb{I}$ is the characteristic function of a set.

In our case, the transition probability is 1 only when, starting from state $s = (x_g, v_k, \{v\})$, the new substation v_{k+1} of $s' = (x_g, v_{k+1}, \{v'\})$ is equal to the action $a \in \mathcal{C} \backslash \{v\}$ that we took:

$$p(s'|s, a) = \begin{cases} 1 & \text{if } v_{k+1} = a \\ 0 & \text{if } v_{k+1} \neq a \end{cases} \tag{2}$$

The **policy** depends only on the observable states, so it doesn't know where the fault is. Let's define a parametrized policy:

$$\pi(a|s) = \frac{e^{\theta a v}}{\sum_b e^{\theta b v}} \tag{3}$$

where θ are the parameters.

This is a problem with terminal state, which occurs when I reconnect all the substations. It will always be reached, since with every action I visit a substation, and at the very least I remove it from the set of disconnected substations (if I'm lucky I can remove half of the substations from the set of disconnected substations every time). Let's call T our horizon. So since we are in a finite-horizon problem we have that $\gamma = 1$. So here we can think the time as the structure of visited / non-visited states.

Let's define J as the sum of all the costs you have if you are in state $s = (x_g, v_k, \{v\})$ summed in time until the process is concluded. The steps of the process are formal steps, since in a step we pass from one substation to another, so the physical time is in the costs (as the time / cost of going from one substation to another).

So we associate a *cost function* J_π to each state $s \in \mathcal{S}$ that is the *accumulated cost from that state to the end of the process, following policy* π.

Let's define as $d_{v_k, v_{k+1}}$ the time to go from the substation v_k to the next substation v_{k+1}, and n_k the number of users still disconnected at time k ($n_k = \sum_{v \in \{v\}} u_v$, where u_v is the number of users under substation v and $\{v\}$ is the set of disconnected users), we have that the cost is

$$J_\pi(s) = \mathbb{E}_\pi \left[\sum_{k=0}^{T} d_{v_k, v_{k+1}} \cdot n_k \;\middle|\; s_0 = s \right] \tag{4}$$

If I can find an optimal path in the states, I can find an optimal path in the substations, since that state contains the current substation and other information.

Can we find a recursive relation like in the traveling salesman?

$$J_\pi\Big(s = (x_g, v_k, \{v\})\Big) = \sum_{a \in \{v\}} \pi(a|s) \Big[d_{v_k, v_{k+1}} \cdot n_k + J_\pi\Big(s' = (x_g, a, \{v'\})\Big) \Big] \tag{5}$$

(from the recursive equation of V, since $V = J$).

Gradient

We have that

$$\nabla_\theta J = \sum_{s,a} \eta_\pi(s) Q_\pi(s,a) \nabla_\theta \pi(a|s) \tag{6}$$

To optimize J we perform a gradient descend on θ. The quantities η and Q, given a policy π, are computed through *linear* operations (since they obey to linear equations), and are computed for all the states. They don't contain the position of the fault, then we sum over all possible positions of the fault.

Initial state: I didn't visit any substation and I have all the possible positions of the fault.

Idea for the state: I could use as state everything that is included in the ordered pair of two substations: $\{v\} = (v_l, v_r)$. And every intervention sends me from one state to another. Topological assumption: we have a tree structure. For this, check what happens with forks and the instrumental test!

We have that

$$\begin{aligned} Q_\pi(s,a) &= \sum_{s'} p(s'|s,a) \left(r(s,a,s') + \gamma \sum_{a'} \pi(a'|s') Q(s',a') \right) \\ &= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}) \;\middle|\; s_0 = s, a_0 = a \right] \end{aligned} \tag{7}$$

is the **state-action value function** or the **quality** of the state-action pair, while

$$\begin{aligned} \eta(s) &:= (I - \gamma P)^{-1} \rho_0 = \sum_{t=0}^{\infty} \gamma^t P^t \rho_0 \\ &= \sum_{t=0}^{\infty} \gamma^t \sum_{\bar{s}} \mathbb{P}\text{rob}(s_t = s \mid s_o = \bar{s}) \rho_0(\bar{s}) \end{aligned} \tag{8}$$

$$= \sum_{t=0}^{\infty} \gamma^t \sum_{\bar{s}} P^t(s|\bar{s})\rho_0(\bar{s})$$

is **the time spent in state s before the process dies**.

In (7) we have that for every value of the parameters θ we know $\pi(a'|s')$, so it is a linear equation in Q.

We have that that states are in the order of the number of substations cubed N^3, since we have the position of the fault and the two substations that identify the not yet visited substations, and the actions are in the order of N. So Q is an equation in N^4 variables. We solve it iteratively using value iteration (we find the fixed point).

The Q equation

Given that the state is $s = (x_g, v_k, \{v\})$, the action is $a \in \{v\}$ and the new state is $s' = (x_g, v_{k+1} = a, \{v'\})$, the equation of Q in our case, given (2), is:

$$\begin{aligned} Q_\pi(s,a) &= \sum_{s'} p(s'|s,a) \left(r(s,a,s') + \gamma \sum_{a'} \pi(a'|s') Q(s',a') \right) \\ &= \left(d_{v_k,a} \cdot n_k + \sum_{a' \in \{v'\}} \pi(a'|s') Q(s',a') \right) \end{aligned} \tag{9}$$

I can go from any substation to any other substation that has not already been connected. Let's suppose that my policy is simply to go in a random substation still disconnected, so π is a uniform distribution over the disconnected substations, so it is

$$\pi\Big(a \,|s = (x_g, v_k, \{v\})\Big) = \frac{1}{|\{v\}|} \tag{10}$$

So since $s' = (x_g, v_{k+1} = a, \{v'\})$ we have that Q becomes

$$Q_\pi\Big(s = (x_g, v_k, \{v\}), a\Big) = d_{v_k,a} \cdot n_k + \sum_{a' \in \{v'\}} \frac{1}{|\{v'\}|} Q\Big(s' = (x_g, a, \{v'\}), a'\Big) \tag{11}$$

Let's try to estimate how much does it cost computing Q. If we use the QR factorization for rectangular matrices on $Q \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$, is costs $O(|\mathcal{S}| \times |\mathcal{A}|^2)$. Since we have that the number of substations is $|\mathcal{C}| = N$ and $|\mathcal{S}| \sim O(N^3)$ and $|\mathcal{A}| \sim O(N)$, we have that solving this linear system requires $O(N^3 \times N^2) = O(N^5)$.