MDP

We are given a set of disconnected substations $\mathcal{C}$, with cardinality $|\mathcal{C}| = N$ (in practice in Trieste they are always < 20), between two remotely controlled substations, where these last will not be included in the problem, since they are already reconnected (in grey in Figure 1).

We define the cost of the process as the amount of 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. Our objective is to minimize this cost.

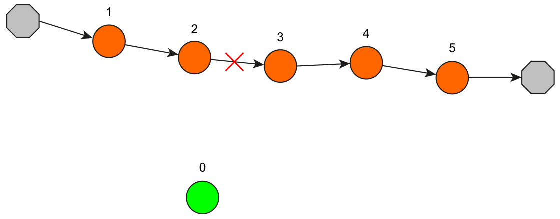

Figure 1. The fault has just occourred. We are in substation 0 and all the substations are disconnected (orange).

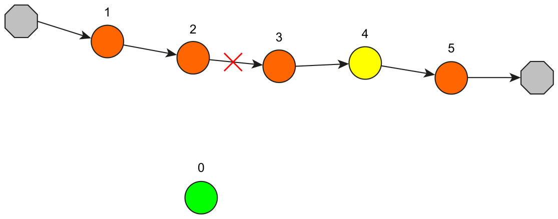

Figure 2. We visit substation 4 (yellow).

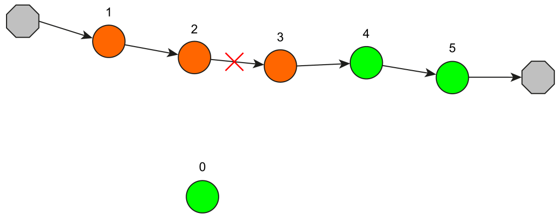

Figure 3. We have reconnected substations 4 and 5 (green).

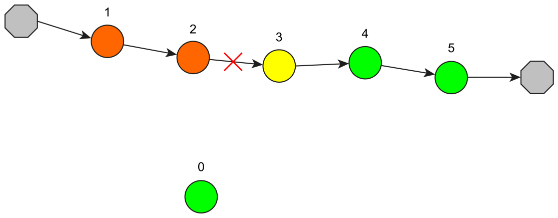

Figure 4. We go in substation 3 (yellow).

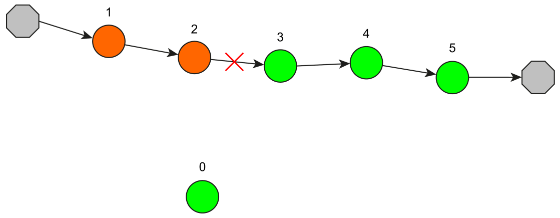

Figure 5. We have riconnected substation 3 (green).

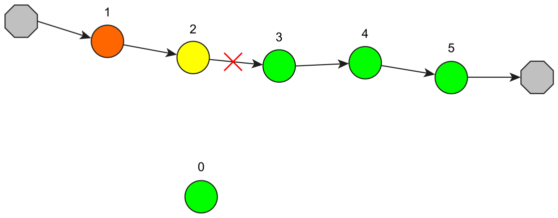

Figure 6. We go in substation 2 (yellow).

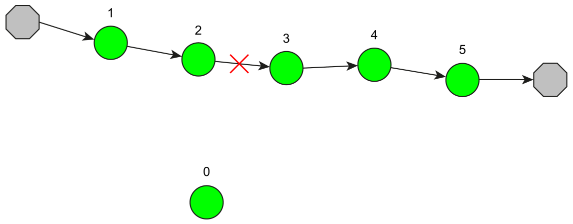

Figure 7. We reconnected substations 1 and 2 (green). All the subsations are reconnected.

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 the technician is, and $\{v\}$ is the set of the still disconnected substations after the technician operates in the current substation v_k. We could also use the substation already reconnected, since they are complementary: given one of the two sets we can always retrieve the other. We have that the variable x_g is **hidden**, while the variables v_k and $\{v\}$ are **observable**, and we will also write $s = (x_g, o)$ where $o = (v_k, \{v\})$ is the observation.
When the fault occurs the technician can be everywhere: at home if it is the middle of the night, at the company, be around, etc. So we introduce an extra "fake" substation, called substation 0, that is the position of the technician when the fault occurs. So the **initial state** is always $s_0 = (\, x_g, o_0 = (0, \mathcal{C})\,)$, thus we have $|x_g| = 2N + 1$ initial states, one for every possible position of the fault.
Instead, the **terminal state** is of the form $s_t = (x_g, v_k, \varnothing)$, where we have that, if the fault is on a cable, v_k will be one of the two substations at the ends of that faulty cable , so we would have two terminal states, while if the fault is in a substation, v_k would be that exact substation, so the terminal state would be only one.
So there is an initial cost which has a random component which depends on the position of the technician when the fault occurs. But what is important in our problem, based on how we are dealing with it, is the average cost, so we can think of doing an average with respect to all the possible distributions of the position of the technician, and this gives me a first average cost, which is the idea of the substation 0. The substation 0 represents the average position of the technician, so the associated cost to go to one random substation from this position. This is a rather brutal approximation of what happens in reality, but to make it more detailed we should introduce a spatial structure of the problem besides the graph representation..... Giving different costs to go from the substation 0 to ever other substation introduces a kind of metric
- the **observation** is $o = (v_k, \{v\})$. We define the observable o as a function of s:

$$\begin{aligned} o(s):\quad & \mathcal{S} && \rightarrow && \mathcal{O} \\ & s = (x_g,\, o = (v_k, \{v\})\,) && \mapsto && o = (v_k, \{v\}), \end{aligned} \tag{1}$$

which for different states $s = (x_g,\, o = (v_k, \{v\})\,)$ that differ only on the position of the fault x_g associates the same observable $o = (v_k, \{v\})$. We have that o is an equivalence class for s. But for brevity we will keep implicit this dependency, and most of the times we will write o instead of $o(s)$, meaning that $o = o(s)$.
- the **action** is the intervention we do in the specific substation we decide to visit, so $a \in \mathcal{C}$. Actually, since we visit only disconnected substations, we have that $a \in \{v\}$.
- the **next state** is $s' = (x_g, v_{k+1} = a, \{v'\})$, where $\{v'\}$ is the set of disconnected substations after the technician operates in substation v_{k+1}. Since the technician can always at least reconnect the substation they visit, we have that $\{v'\} \subseteq \{v\} \backslash a$, so the set of disconnected substations decreases after each action. We are therefore positive that the process terminates.
- the **reward** is the **cost** of going in a certain substation (as the time *in seconds* that it takes to go there from where we are) multiplied for the number of disconnected users. Let's define as $d_{v_k, v_{k+1}}$ the time *in seconds* to go from the substation v_k to the next substation v_{k+1}, and n_k the number of users still disconnected <u>before</u> operating in substation v_{k+1}. So if we are in state $s = (x_g, v_k, \{v\})$, we make an action a and we end up in state $s' = (x_g, v_{k+1} = a, \{v'\})$, we have that the number of disconnected users is $n_k = \sum_{v \in \{v\}} u_v$, where u_v is the number of users underneath substation v. So the reward depends only on the previous state s and the action taken a, and has formula

$$r\Big(s = (x_g, v_k, \{v\}), a, s' = (x_g, v_{k+1} = a, \{v'\})\Big) = r(s, a) = d_{v_k, a} \cdot n_k = d_{v_k, a} \cdot \sum_{v \in \{v\}} u_v \,. \tag{2}$$

> For now, in the cost we will ignore the cost of discovering if the fault is left or right, which might rise the total cost significantly. This is due to a lack of data. To improve the computation of the cost, we need to take note carefully of the operations the technicians perform when a fault occurs. This will be done with a Telegram bot (see Section ?).

Idea for the state: We 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. We have to find a way of representing forks.

Example 1: Looking at the Figures 1-5, we have that the sequence of states and actions is:

- $s_0 = (x_g, 0, \mathcal{C} = \{1,2,3,4,5\}), a = 4$,
- $s_1 = (x_g, 4, \{1,2,3\}), a = 3$,
- $s_2 = (x_g, 3, \{1,2\}), a = 2$,
- $s_3 = (x_g, 2, \varnothing)$.

Since we know every aspect of the problem and we have a model of the environment, this is a **model-based** problem. Besides, this is a problem of **partial observability**, since part of the state s is hidden, which is the position of the fault x_g, that we don't know (in partially observable problems the full state is not available to the agent). So, we can not solve it using dynamic programming (cita Bellman, *Dynamic programming*), because in this way the policy would depend on the whole state, thus also on the hidden state, which can not be.

The partial observability of the system stems from the fact that multiple states give the same sensor reading, since the agent can only sense a limited part of the environment. The partial observability can lead to "perceptual aliasing": different parts of the environment appear similar to the agent's sensor system, but require different actions.

An example of obtaining Markov states through a state-update function is provided by the popular Bayesian approach known as *Partially Observable MDPs*, or *POMDPs*. In this approach the environment is assumed to have a well defined *latent state* X_t that underlies and produces the environment's observations, but is never available to the agent (and is not to be confused with the state S_t used by the agent to make predictions and decisions). The natural Markov state, S_t, for a POMDP is the *distribution* over the latent states given the history, called the *belief state*. [pag 467 (489) [1]]

Actually, if we know where the fault is, the solution is straightforward: if it is on an edge to solve the fault we need to visit the two substations at the ends of it (we might have to choose the right order), if it is in a substation we visit it directly.

Not knowing the position of the fault x_g, we can solve this problem with two different approaches. One is to use **approximate value iteration**: we work in a subspace of the states space corresponding to the observable states. Another possibility is to perform **policy iteration**, in particular gradient descent in the policy space. So we impose that the policy depends only on the observable variables, and we try to find the best policy in this subspace in order to optimize the MDP.

The approach with *beliefs* of the classical MDPs is a Bayesian approach, in which we start from the idea that our state has a certain distribution, the prior distribution, and then you make a probabilistic planning of where the state will go. The policy gradient method instead is a frequentist, and knowing the model we can do planning, and all the aspect of partial observability is contained in the parametrization of the policy. We are doing something that is a manner of doing dynamic programming, a form of policy gradient / policy iteration, but approximated, in the sense that we are searching in the space of all the policies the one variety that is parametrized with Boltzmann. This is a more direct approach that doesn't require beliefs, which are very heavy since require to write a probability distribution over the states and work with them. Algorithms that are guaranteed and scale well to solve a POMDP does not exist, there are point-based approaches like Perseus, but are very difficult to organize. This one instead is much more transparent and direct. A policy that is based on beliefs can express more, and is richer than the one that we are trying to express, since it contains all the history of the precedent observations, while we are using a reactive policy, that takes the observations done and parametrizes the policy in function of these observations. A reactive policy reacts on instantaneous observations, and there is no history of the preceding observations. In both methods we need all the model, but while the other model is more difficult, and potentially more powerful, and has a higher computational cost, and we evaluated that it is not worth it.

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. This means that $s' = \sigma(s, a)$. So, mathematically we have that the **transition probability** is

$$p(s' \mid s, a) = \mathbb{I}\big(s' = \sigma(s, a)\big) = \delta_{s', \sigma(s,a)} = \begin{cases} 1 & \text{if } s' = \sigma(s, a) \\ 0 & \text{otherwise} \end{cases}, \tag{3}$$

where δ is the Kronecker delta and $\mathbb{I}$ is the characteristic function of a set.
In our specific case, the transition probability is equal to 1 only when, starting from state $s = (x_g, v_k, \{v\})$, the new substation v_{k+1} of the next state $s' = (x_g, v_{k+1}, \{v'\})$ is equal to the action $a \in \{v\}$ that we took:

$$p(s'|s, a) = \begin{cases} 1 & \text{if } v_{k+1} = a \\ 0 & \text{if } v_{k+1} \neq a \end{cases}. \tag{4}$$

It is somewhat surprising and not widely recognized that function approximation includes important aspects of partial observability. For example, if there is a state variable that is not observable, then the parameterization can be chosen such that the approximate value does not depend on that state variable. The effect is just as if the state variable were not observable. Because of this, all the results obtained for the parameterized case apply to partial observability without change. In this sense, the case of parameterized function approximation includes the case of partial observability. [1, pag 464 (486)]

Since we don't know where the failure is, the **policy** depends only on the observable states, so it doesn't know where the fault is. Let's define a parameterized policy using Boltzmann parameterization:

$$\pi\Big(a \mid o = (v_k, \{v\}), \theta\Big) = \frac{e^{\theta_{o,a}}}{\sum_{b \in \{v\}} e^{\theta_{o,b}}}. \tag{5}$$

where θ are the parameters for each state — actually, the observable part o — and action, so they depend on the action a and on the observable variable o of the state:

$$\theta = (\theta_{o,a})_{o \in O, a \in A} = \begin{pmatrix} \theta_{o_1,a_1} & \theta_{o_1,a_2} & \cdots & \theta_{o_1,a_N} \\ \theta_{o_2,a_1} & \theta_{o_2,a_2} & \cdots & \theta_{o_2,a_N} \\ \vdots & & & \vdots \\ \theta_{o_{|O|},a_1} & \theta_{o_{|O|},a_2} & \cdots & \theta_{o_{|O|},a_N} \end{pmatrix} \tag{6}$$

(where N is the number of substations, so the number of possible actions). So we also have that

$$\pi\Big(a \mid o = (v_k, \{v\}), \theta\Big) = \begin{pmatrix} \pi(a_1|o_1) & \pi(a_2|o_1) & \cdots & \pi(a_N|o_1) \\ \pi(a_1|o_2) & \pi(a_2|o_2) & \cdots & \pi(a_N|o_2) \\ \vdots & & & \vdots \\ \pi(a_1|o_{|O|}) & \pi(a_2|o_{|O|}) & \cdots & \pi(a_N|o_{|O|}) \end{pmatrix}. \tag{7}$$

The policy can not depend on the position of the failure, otherwise we would have automatically have solve the problem: the solution would be to go in the substation in which the failure is or at the ends of the cable (edge) where the fault is.

If we search in this space of policies, this will give us a policy which doesn't depend on the time, since with any algorithm we try to find the optimal parameters to solve this problem. This gives us a **stationary policy**. The structure of the states is already a measure of time, since we have the number of steps already done: the substations we already visited. So the important is not to establish a policy at the different steps, but a policy with respect to the states.

This is a problem with terminal state, which occurs when we reconnect all the substations. It will always be reached, since with every action we visit a substation, and at the very least we remove it from the set of disconnected substations (instead, if we are lucky, every time we can remove half of the substations from the set of disconnected substations). So, for this specific problem it doesn't make sense to introduce a discount factor γ.

Let's define J as the sum of all the costs we incur ~~if we 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 define

$$\begin{aligned} J_\pi &= \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} r(s_t, a_t, s_{t+1}) \right] \\ &= \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} d_{v_k,a} \cdot n_k \right] . \end{aligned} \tag{8}$$

Besides, we associate a *cost function* V_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. In practice, it is the value of the state s:

$$\begin{aligned} V_\pi(s) &= \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} r(s_t, a_t, s_{t+1}) \;\middle|\; s_0 = s \right] \\ &= \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} d_{v_k,a} \cdot n_k \;\middle|\; s_0 = s \right] . \end{aligned} \tag{9}$$

If we can find an optimal path in the states, we 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? Using the recursive equation of the value function V we have that

$$V_\pi\Big(s = (\, x_g, o = (v_k, \{v\})\,)\Big) = \sum_{a \in \{v\}} \pi(a|o) \Big[d_{v_k,a} \cdot n_k + V_\pi\Big(s' = (x_g, a, \{v'\})\Big) \Big] . \tag{10}$$

Since the transition probabilities are deterministic, from each one of the two (or one) terminal states we can go back. But unlike in the Traveling Salesman, we won't solve the Bellman's equation at each step performing a $\max$ on the actions, since the policy can not depends on the whole state, but only on the observable part.

This is why we have to perform a continuous projection in the case of approximate value iteration or the gradient search. The approximate value iteration is a delicate operation and you have to pay a lot of attention, instead using the gradient is more direct. So we will use the latter.

Gradient descent in the policy space

We have that

$$J_\pi = \sum_s \rho_0(s) V_\pi(s) = \sum_s \rho_0(s) \sum_a \pi(a|o) Q_\pi(s,a) \,, \tag{11}$$

given (15), so we are averaging over the distribution of possible initial states. J is a property of the entire episode, and it is like an average value.

If we have a unique initial state $\bar{s}$, we have that $\rho_0(s) = \delta_{s,\bar{s}}$ and so we have that

$$J_\pi = \sum_s \delta_{s,\bar{s}} V_\pi(s) = V_\pi(\bar{s}) \tag{12}$$

which is what happens in our problem, since we have an unique initial state $s_0 = (x_g, 0, \mathcal{C})$. NO we have $|x_g| = 2N + 1$ possible initial states: one for every position of the fault!

From the theory we have that the gradient formula is

$$\nabla_{\theta_{o',a'}} J_\pi = \sum_{s,a} \eta_\pi(s) Q_\pi(s,a) \nabla_{\theta_{o',a'}} \pi(a|o)\,. \tag{13}$$

where we have that $s = (\, x_g, o = (v_k, \{v\})\,)$ and the policy depends only on the observable part o of the entire state s.

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: We didn't visit any substation and we have all the possible positions of the fault.

From the theory we have that

$$\begin{aligned} Q_\pi(s,a) &:= \sum_{s'} p(s'|s,a)\,\big(r(s,a,s') + V_\pi(s')\big) \\ &= \mathbb{E}\left[\sum_{t=0}^{\infty} r(s_t, a_t, s_{t+1}) \,\middle|\, s_0 = s, a_0 = a\right]. \end{aligned} \tag{14}$$

is the **state-action value function** or the **quality** of the state-action pair. In particular we have that

$$V_\pi(s = (\, x_g, o = (v_k, \{v\})\,)) = \sum_a \pi(a|o) Q_\pi(s,a)\,, \tag{15}$$

so we have that

$$Q_\pi(s,a) = \sum_{s'} p(s'|s,a) \left(r(s,a,s') + \sum_{a'} \pi(a'|o') Q_\pi(s',a') \right). \tag{16}$$

Given that the state is $s = (\, x_g, o = (v_k, \{v\})\,)$, the action is $a \in \{v\}$ and the new state is $s' = \sigma(s,a) = (\, x_g, o' = (v_{k+1} = a, \{v'\})\,)$, the equation of Q in our case, given (4), is:

$$Q_\pi(s,a) = \left(d_{v_k,a} \cdot n_k + \sum_{a' \in \{v'\}} \pi\big(a'|\sigma(o,a)\big) Q\big((\sigma(s,a), a')\right). \tag{17}$$

While for the other variable we have that

$$\eta_\pi(s') := \rho_0(s') + \sum_{s,a} \pi(a|o) p(s'|s,a) \eta_\pi(s) \tag{18}$$

is **the time spent in state s' before the process dies**, where $\rho_0(s')$ is the probability of starting in the initial state s'.

We can notice that the value of Q depends on the values of Q of future states, while the value of η depends on the values of η for previous states.

In (16) we have that for every value of the parameters θ we know $\pi(a'|o')$, 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). Actually we can compute it in one sweep over all the possible states!

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

$$\pi\Big(a\,|o=(v_k,\{v\})\Big) = \frac{1}{|\{v\}|} \tag{19}$$

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

$$Q_\pi\Big(s=(x_g,v_k,\{v\}),a\Big) = d_{v_k,a}\cdot n_{k+1} + \sum_{a'\in\{v'\}} \frac{1}{|\{v'\}|} Q\Big(\sigma(s,a),a'\Big) \tag{20}$$

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 $|\mathcal{C}|=N$, given that a state can be thought as the current substation v_k and the pair of substations that delimit the segment of disconnected substation, we have that $|\mathcal{S}|\sim O(N\cdot N^2)=O(N^3)$, while $|\mathcal{A}|\sim O(N)$. So we have that solving this linear system requires $O(|\mathcal{S}|\times|\mathcal{A}|^2)=O(N^3\times N^2)=O(N^5)$. CHECK WHAT PROF SAID ABOUT SPEED

We have that in (22) the function $\rho_0(s')$ is the probability of starting in the initial state s'. Since we don't have any prior information of where the fault might be, ρ_0 doesn't depend on it, so ρ_0 will be uniform in x_g. This means that we divide by the number of possible positions of x_g, which is either in a substation or in a cable, so it is $N+(N+1)=2N+1$. Instead, the first substation in which to go is always the fake substation 0, which represents the position of the technician when the fault occurs, and the set of the disconnected substations must be equal to the set of all the substations $\mathcal{C}$. So ρ_0 must be 1 when the current substation is 0 and the set of the disconnected substations is equal to $\mathcal{C}$, and must be 0 for every other situation. So we have that

$$\begin{aligned}\rho_0\Big(s=(\,x_g, o=(v_k,\{v\})\,)\Big) &= \Pr(x_g)\Pr(o=o_0=(0,\mathcal{C}))\\ &= \frac{1}{2|\mathcal{C}|+1}\mathbb{I}\big(v_k=0,\{v\}=\mathcal{C}\big)\\ &= \frac{1}{2N+1}\mathbb{I}\big(v_k=0,\{v\}=\mathcal{C}\big)\,.\end{aligned} \tag{21}$$

Actually, according to the technicians of the company, the majority of the faults happen in the cables, usually when they are not anymore perfectly isolated (e.g. the wire's insulation breaks down) and they cause a short circuit, being connected to the ground and allowing charge to flow through it. We will try to take advantage of this information in a later moment, but for now we will suppose that every component has the same probability of being damaged.

So in our case, thanks to (21), (19), and (4), we have that

$$\begin{aligned}\eta_\pi\Big(s'=(x_g,v_{k+1},\{v'\})\Big) &:= \rho_0(s') + \sum_{s,a}\pi(a|o)p(s'|s,a)\eta_\pi(s)\\ &= \frac{1}{2N+1}\mathbb{I}\big(v_k=0,\{v'\}=\mathcal{C}\big) + \sum_{s\in\mathrm{pa}(s')}\pi(a=v_{k+1}|o)\eta_\pi(s)\\ &= \frac{1}{2N+1}\mathbb{I}\big(v_k=0,\{v'\}=\mathcal{C}\big) + \sum_{s\in\mathrm{pa}(s')}\frac{1}{|\{v\}|}\eta_\pi\Big(s=(x_g,v_k,\{v\})\Big)\end{aligned} \tag{22}$$

where $\mathrm{pa}(s')$ indicates the parents of the node s' in the dependency graph (probabilistic graphical model - Bayesian Network).

In an episodic task, the on-policy distribution is a little different in that it depends on how the initial states of episodes are chosen. Let $\rho_0(s)$ denote the **probability that an episode begins in each state s**, and let $\eta(s)$ denote **the number of time steps spent, on average, in state s in a single episode**. Time is spent in a state s if episodes start in s, or if transitions are made into s from a preceding state $\bar{s}$

in which time is spent:

$$\eta(s) = \rho_0(s) + \sum_{\bar{s}} \eta(\bar{s}) \sum_a \pi(a|\bar{s}) p(s|\bar{s}, a), \text{ for all } s \in \mathcal{S}\,. \tag{23}$$

This system of equations can be solved for the expected number of visits $\eta(s)$.
The **on-policy distribution** is then the fraction of time spent in each state normalized to sum to one:

$$\mu(s) = \frac{\eta(s)}{\sum_{s'} \eta(s')}, \text{ for all } s \in \mathcal{S}\,. \tag{24}$$

This is the natural choice without discounting. If there is discounting ($\gamma < 1$) it should be treated as a form of termination, which can be done simply by including a factor of γ in the second term of (23). ([1] pag. 199 (221))

Idea of the algorithm: We start from a certain policy, for example the random policy of equation (19), in which we choose randomly the substation to be visited. This means that in the parameterized policy (5) all the parameters θ are equal to 0: $\theta = 0 = (0, 0, \dots, 0)$. This is because in the random policy the parameters θ don't depend on the action a, so we have that

$$\pi\Big(a \mid o = (v_k, \{v\})\Big) = \frac{e^{\theta_o}}{\sum_{b\in\{v\}} e^{\theta_o}} = \frac{e^{\theta_o}}{e^{\theta_o} \sum_{b\in\{v\}} 1} = \frac{1}{|\{v\}|}\,. \tag{25}$$

This is true for every choice of θ which doesn't depend on the action, but in practice to construct a uniform policy we take $\theta = 0 = (0, 0, \dots, 0)$. But notice that this is not the only way.

In general, there is no guarantee that this is a convex problem in the parameters θ, we could have several different minima. So one should do random restarts with different policies than the random one to see if you can reach a different minimum. However, this doesn't guarantee to find the global minimum, but it is the best we can do.

Then we compute the gradient using (13). The gradients of the policy are simple computations (since we decided the parametrization of the policy, we can derive it), while the objects Q and η have to be done indirectly: you have to solve the linear equations (20) and (22) for that given policy. This can be more or less computationally heavy, but we have to solving two linear equations, with any chosen method. Given the form of our transition probabilities $p(s'|s, a)$, we have that the equations of Q don't depend on s', but only on a. So fixed the first state s, we have to solve two linear equations only on the actions (we can see the $Q(s, a)$ matrix as a series of vectors). **This operations can be parallelized!**

So having done these steps we have an expression for the gradient for every parameter. So we do a sweep over all the parameters and we find the value of the gradient for each one of them. Then we take a step in the parameters space and we descent the gradient. We stop when the value of J doesn't change "so much" in percentage.

If we have the intuition that there is a deterministic sequence of action to be performed, we have that the parameters θ tend to infinity. This is because the deterministic policies correspond to a one-hot vector in which we have 1 for only one action and 0 for all the other actions, and this happens when the parameter associated to this action becomes much bigger than the others, so that when we normalize it with the parametrization of the policy (which is like a softmax) it becomes 1 while the others are so small that becomes 0. If it is like this, the gradient descent never stops and there could be problems of overflow on θ since it goes to infinity.

A smart thing to do when this happens is to write the parametrization in the following way (the state s is fixed):

$$\pi(a) = \frac{e^{\theta_a}}{\sum_b e^{\theta_b}} = \frac{e^{-(\max_{a'} \theta_{a'} - \theta_a)}}{\sum_b e^{-(\max_{a'} \theta_{a'} - \theta_b)}}\,. \tag{26}$$

The benefit of writing it in this way is that, since the exponents of the two exponentials are negative (the parts in the parenthesis $(\max_{a'} \theta_{a'} - \theta_a)$) are always positive), this never explodes. This is a simple numerical trick that allows to improve the stability of the algorithm. (link)

Automatic computation of Q and η

The computation of Q and η is done for every possible state simultaneously, included all the possible positions of the fault. So they are not quantities that are observable, because they depend on the position of the fault.

The gradient is computed... It is a problem of planning in which the decisions are made based only over a subset of the quantities that define the states. So we have to solve the equations for Q and η for every position of the fault x_g.

We want to find an automatic way to compute the matrix Q and the vector η. We have that

$$Q[s,a] = P[s,a|s',a']Q[s',a'] + R[s,a] \tag{27}$$

where Q, P and R are matrices.

The matrix $R[s,a]$ represents the immediate cost of being in state s and doing action a, and it is computed using:

$$R\Big[s = (x_g, v_k, \{v\}), a\Big] = d_{v_k,a} \cdot n_k = d_{v_k,a} \cdot \sum_{v \in \{v\}} u_v \,. \tag{28}$$

For every position of the fault, we compute the matrix Q bottom-up and the vector η top-down. We can overwrite these matrices every time since we don't need them for long.

Compute the gradient

We have seen in (13) that, for $s = (\, x_g, o = (v_k, \{v\})\,)$ and $s' = (\, x_g, o' = (v_{k+1}, \{v'\})\,)$, we have:

$$\nabla_{\theta_{o',a'}} J = \sum_{s,a} \eta_\pi(s) Q_\pi(s,a) \nabla_{\theta_{o',a'}} \pi(a|o)\,.$$

Given that the vector θ is $\theta = (\theta_{o,a})_{o,a}$, where N is the number of substations, we have that

$$\nabla_{\theta_{o',a'}} J = \begin{pmatrix} \frac{\partial}{\partial \theta_{o_1,a_1}} J & \cdots & \frac{\partial}{\partial \theta_{o_1,a_N}} J \\ \frac{\partial}{\partial \theta_{o_2,a_1}} J & \cdots & \frac{\partial}{\partial \theta_{o_2,a_N}} J \\ \vdots & & \\ \frac{\partial}{\partial \theta_{o_{|O|},a_1}} J & \cdots & \frac{\partial}{\partial \theta_{o_{|O|},a_N}} J \end{pmatrix} . \tag{29}$$

Given the equation for the policy in (5), and given (13), we have that its derivative is

$$\begin{aligned} \frac{\partial}{\partial \theta_{o',a'}} \pi\Big(a \;\Big|\; o = (v_k, \{v\})\Big) &= \frac{\partial}{\partial \theta_{o',a'}} \left(\frac{e^{\theta_{o,a}}}{\sum_{b \in \{v\}} e^{\theta_{o,b}}} \right) \\ &= \delta_{o',o} \cdot \frac{\delta_{a',a} \cdot e^{\theta_{o,a}} \cdot \sum_{b \in \{v\}} e^{\theta_{o,b}} - e^{\theta_{o,a}} \cdot e^{\theta_{o,a'}}}{(\sum_{b \in \{v\}} e^{\theta_{o,b}})^2} \\ &= \delta_{o',o} \cdot \left(\frac{\delta_{a',a} \cdot e^{\theta_{o,a}} \cdot \sum_{b \in \{v\}} e^{\theta_{o,b}}}{(\sum_{b \in \{v\}} e^{\theta_{o,b}})^2} - \frac{e^{\theta_{o,a}}}{\sum_{b \in \{v\}} e^{\theta_{o,b}}} \cdot \frac{e^{\theta_{o,a'}}}{\sum_{b \in \{v\}} e^{\theta_{o,b}}} \right) \\ &= \delta_{o',o} \cdot \big(\delta_{a',a} \pi(a|o) - \pi(a|o)\pi(a'|o)\big) \\ &= \delta_{o',o} \left(\delta_{a',a} - \pi(a'|o)\right) \pi(a|o) \end{aligned} \tag{30}$$

since if $o \neq o'$ we have that $\theta_{o',a'}$ and $\theta_{o,a}$ are definitely different parameters. When $\theta = 0$ we have that

$$\frac{\partial}{\partial \theta_{o',a'}} \pi\Big(a \mid o = (v_k, \{v\})\Big)\Big|_{\theta=0} = \delta_{o',o} \left(\frac{1}{|\{v\}|} \delta_{a',a} - \frac{1}{|\{v\}|^2} \right) \tag{31}$$

In particular

$$\nabla_{\theta_{o',a'}} \pi(a|o) = \begin{pmatrix} \frac{\partial}{\partial \theta_{o_1,a_1}} \pi(a|o) & \frac{\partial}{\partial \theta_{o_1,a_2}} \pi(a|o) & \cdots & \frac{\partial}{\partial \theta_{o_1,a_N}} \pi(a|o) \\ \vdots & & & \vdots \\ \frac{\partial}{\partial \theta_{o_{|O|},a_1}} \pi(a|o) & \frac{\partial}{\partial \theta_{o_{|O|},a_2}} \pi(a|o) & \cdots & \frac{\partial}{\partial \theta_{o_{|O|},a_N}} \pi(a|o) \end{pmatrix} \tag{32}$$

So we have that

$$\begin{aligned} \nabla_{\theta_{o',a'}} J &= \sum_{s,a} \eta_\pi(s) Q_\pi(s,a) \nabla_{\theta_{o',a'}} \pi(a|s) \\ &= \sum_{s,a} \eta_\pi(s) Q_\pi(s,a) \delta_{s,s'} \Big(\left(\delta_{a,a'} - \pi(a'|o)\right) \pi(a|o) \Big) \\ &= \sum_{a} \eta_\pi(s') Q_\pi(s',a) \left(\delta_{a,a'} - \pi(a'|o')\right) \pi(a|o') \\ &= \eta_\pi(s') \sum_{a} Q_\pi(s',a) \left(\delta_{a,a'} - \pi(a'|o')\right) \pi(a|o') \,. \end{aligned} \tag{33}$$

So we have that

$$\theta = \theta - \alpha \nabla_\theta J \,, \tag{34}$$

where α is a step-size parameter, or learning rate.

Averaging for the position of the failure

THE PROBLEM IS THAT I DON'T HAVE ACCESS TO THE POSITION OF THE FAILURE! SO I can not use (13), because I cannot compute $\eta(s)$ and $Q(s,a)$ for the real position of the failure.

So I have to consider every possible position of the failure and find a way to determine a unique policy (instead of a policy for every position of the failure) that doesn't depend on the position of the fault.

As we already said, we have that the policy does not depend on the position of the failure, so $\pi = \pi(a|o)$, while the value Q_π depends on it: $Q_\pi = Q_\pi(s,a)$.

We can compute ∇J in two different ways:

1. First method: compute

$$\nabla_{\theta_{s',a'}} J = \eta_\pi(s') \sum_a Q_\pi(s',a) \nabla_{\theta_{s',a'}} \Big(\left(\delta_{a,a'} - \pi(a'|s')\right) \pi(a|s') \Big) \tag{35}$$

for every possible position of the failure (where $\pi(a|s')$ with $s' = (\,x_g, o'\,)$ is equal to the value of $\pi(a|o')$, so it is full of duplicates) and then compute

$$\Delta\theta_{o',a'} = \sum_{x_g : s' = (x_g, o')} \nabla_{\theta_{s',a'}} J, \tag{36}$$

so I sum all the states that have equal observation o' (raggruppo le variazioni delle θ/π per differenti x_g ma uguali $o \rightarrow$ raggruppo sulla direzione che non so: x_g!). Forse devo moltiplicare anche per $\Pr(x_g)$?

2. Second method: compute

$$\nabla_{\theta_{o',a'}} J = \bar{\eta}_\pi(o') \sum_a \bar{Q}_\pi(o',a) \nabla_{\theta_{o',a'}} \Big(\left(\delta_{a,a'} - \pi(a'|o')\right) \pi(a|o') \Big) \tag{37}$$

using the previous defined $\bar{\eta}$ and $\bar{Q}$.

Quindi o faccio la media delle variazioni dei θ oppure faccio la media direttamente sulle variabili di η e Q.

$$\frac{\partial}{\partial\theta_{o',a'}}\pi\Big(a \mid s=(x_g,\, o=(v_k,\{v\})\,)\Big) = \delta_{o',o(s)}\left(\delta_{a',a}-\pi(a'|o)\right)\pi(a|o) \tag{38}$$

$$\frac{\partial}{\partial\theta_{o',a'}}\pi\Big(a \mid o(s)\Big) = \delta_{o',o}\left(\delta_{a',a}-\pi(a'|o)\right)\pi(a|o) \tag{39}$$

$$\begin{aligned}\nabla_{\theta_{o',a'}}J &= \sum_{s,a}\eta_\pi(s)Q_\pi(s,a)\nabla_{\theta_{o',a'}}\pi(a|o(s))\\ &= \sum_{s}\eta_\pi(s)\sum_{a}Q_\pi(s,a)\nabla_{\theta_{o',a'}}\pi(a|o(s))\\ &= \sum_{s}\eta_\pi(s)\sum_{a}Q_\pi(s,a)\delta_{o',o(s)}\Big(\left(\delta_{a',a}-\pi(a'|o)\right)\pi(a|o)\Big)\\ &= \sum_{s}\delta_{o',o(s)}\eta_\pi(s)\sum_{a}Q_\pi(s,a)\Big(\left(\delta_{a',a}-\pi(a'|o)\right)\pi(a|o)\Big)\\ &= \sum_{x_g}\eta_\pi((x_g,o'))\sum_{a}Q_\pi((x_g,o'),a)\left(\delta_{a,a'}-\pi(a'|o')\right)\pi(a|o')\,.\end{aligned} \tag{40}$$

since $\sum_s \delta_{o',o(s)}$ is equivalent to do an average of every possible position of the fault.

Model free

If this is not possible, one option is to do it model free, so instead of doing a complete gradient we do a stochastic gradient, using the **policy gradient algorithms**. In this case we remove the cost of solving the equations for Q and η, since we proceed by trial and error, you try an action, compute a policy, upgrade the gradient and go on, and this operations are not computationally heavy: you just need to experience the costs. The downside is that it is stochastic, so it has a great variance.

Natural policy gradient

Another option that can be done both at the level of the deterministic gradient and at the level of the stochastic gradients is to use **natural policy gradient**. It's a minimal change: we only change the gradient of π, everything else remains the same.

The improvement is that instead of having a gradient that reaches a horizontal asymptote, the gradient keeps on moving and doesn't slow down. This allows to reach the convergence faster.

---

1. Sutton and Barto, Reinforcement Learning: An Introduction, 2018, MIT Press. ↩ ↩ ↩