Saturday, March 6, 2021 8:33 AM

Let $s_i^{(t)} = (\theta_b^{(t)}, \theta_h^{(t)})$ be the state s_i observed at time t.
i indexes unique states.

$\theta_b^{(t)}$: body angle of worm relative to target angle at time t. $\theta_b^{(t)} \in [-180°, 180°)$
$\theta_h^{(t)}$: head angle of worm relative to $\theta_b^{(t)}$ at time t. $\theta_h^{(t)} \in [-180°, 180°)$

In practice, we discretize w/ steps of 30°.

Let $r^{(t)}$ be the reward observed at time t, $r^{(t)} \in \mathbb{R}$.
Let $a^{(t)}$ be the action performed at time t, $a^{(t)} \in \{0, 1\}$.

Now we define a state-action value $q(s, a)$:

$$q(s_i^{(t)}, a^{(t)}) = \sum_{i=t}^{t+\tau} \gamma^{i-t} r^{(t)}$$

where τ is the number of timesteps to look ahead and γ is an optional discount factor, $\gamma \in (0, 1]$.
Thus far, we leave $\gamma = 1$.

Then we can say there is a distribution for each unique state s_i such that

$\{q(s_i^{(t)}, a^{(t)} = 1)\}$ are samples from $Q(s_i, a=1) \sim \mathcal{N}(\mu_{s_i, a=1}, \sigma^2_{s_i, a=1})$

$\{q(s_i^{(t)}, a^{(t)} = 0)\}$ are samples from $Q(s_i, a=0) \sim \mathcal{N}(\mu_{s_i, a=0}, \sigma^2_{s_i, a=0})$

where t takes on all values such that $s_j^{(t)} = s_i$.

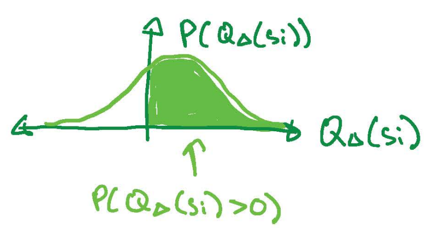
Next, let us define a random variable for each s_i

$$Q_\Delta(s_i) = Q(s_i, a=1) - Q(s_i, a=0).$$

We want to shine light (set $a=1$) on states s_i where $Q_\Delta(s_i) > 0$.
Thus, we would like to infer from samples the probability $P(Q_\Delta(s_i) > 0)$.
We can do this by assuming all the random variables Q & Q_Δ are normally distributed (or t-distributed if the number of samples for a state is <30).
So, we write

$$P(Q_\Delta(s_i) > 0) = 1 - \Phi\left(\frac{0 - \hat{\mu}_{s_i}}{\hat{\sigma}_{s_i}/\sqrt{n}}\right)$$

- where $\Phi(\cdot)$ is the cdf of the standard normal or t-dist for $n < 30$.
- n is the number of samples collected at state s_i
- $\hat{\mu}_{s_i}$ and $\hat{\sigma}_{s_i}$: are the mean and standard deviation of realizations of $Q_\Delta(s_i)$ as observed during sampling.



We therefore have a probability $P(Q_\Delta(s_i) > 0)$ for every state s_i.

We want an optimal policy $\pi(a|s)$ that maximizes reward for every state (assuming previous states do not affect the best policy — the Markov assumption).
$\pi(a|s_i)$ is a probability that action a will be taken in state s_i.
And from above, we can use $P(Q_\Delta(s_i)>0)$ as our optimal light activation rate

$$\pi(a=1|s_i) = P(Q_\Delta(s_i)>0).$$

π is the optimal policy, but it is not the policy we want to explore with. The policy to explore with will aim to visit the states with the most uncertainty in π.
We write it $\xi(a=1|s_i)$, the probability we will sample at a given state s_i by setting $a=1$. Note that $\sum_{s_i}\xi(a=1|s_i) \neq 1$ and $\sum_{A=\{0,1\}}\xi(a=A|s_i)=1$ for every s_i.

The states with the most uncertainty in $\pi(a|s_i)$ are the states with the highest entropy

$$H(\pi^*(a|s_i)) = -\pi(a=1|s_i)\log\pi(a=1|s_i) - \pi(a=0|s_i)\log\pi(a=0|s_i)$$
$$= h(s_i), \text{ for short.}$$

To maximize information/entropy while the worm is moving around, we keep in mind the following terms.

$h(s_i)$, the information of each state.
This was based on our inferred optimal policy π.
$h(s_i) = H(\pi(a|s_i))$

$\xi(a|s_i)$, our yet-unknown exploration policy.
This gives the probability of action a being performed at state s_i.

$p(s_i)$, the worm's probability of being in state s_i.
(first time this is mentioned.)

The term we want to maximize is

$$I = \sum_{s_i} \underbrace{p(s_i)\xi(a=1|s_i)}_{\text{together, the proportion of time spent sampling from } s_i} h(s_i)$$

in other words, the expected information gained by following our exploration policy $\xi(a|s_i)$.

Our constraints are — knowing that $p(s_i)$ and $h(s_i)$ are fixed given a fixed number of samples —

$$0 \le \xi(a=1|s_i) \le 1$$
That is, $\xi(a=1|s_i)$ is a probability

$$\sum_{s_i} p(s_i)\xi(a=1|s_i) \le \ell, \quad \ell\in[0,1]$$
where ℓ is a parameter indicating the maximum proportion of time we allow the light to be on.

Which gives us a Lagrangian

$$\mathcal{L} = \sum_{s_i} p(s_i)\xi(a=1|s_i)h(s_i) + \sum_{s_i}\lambda_{s_i}\underbrace{\left(\xi(a=1|s_i)-1+\epsilon_{s_i}^2\right)}_{\le 1 \text{ constraint}}$$
$$+ \sum_{s_i}\alpha_{s_i}\underbrace{\left(\xi(a=1|s_i)-\delta_{s_i}^2\right)}_{\ge 0 \text{ constraint}} + \beta\left(\sum_{s_i}p(s_i)\xi(a=1|s_i)-\ell\right)$$

We can solve for $\xi(a=1|s_i)$, λ_{s_i}, $\epsilon_{s_i}^2$, $\delta_{s_i}^2$, α_{s_i}, β. [Did this and got soln described below; need to check over & write it up]

However, we can also use an example to see what the $\xi(a=1|s_i)$'s must be.

Say there are 3 states s_1, s_2, s_3. They have entropies $h(\pi(a|s_1)) > h(\pi(a|s_2)) > h(\pi(a|s_3))$.
The probability of occupying each state s_i is given by $p(s_i)$.
Say the states are being sampled with probability $\xi(a=1|s_i)$ where $a=1$ indicates sampling.
We want to find $\xi(a=1|s_i)$ for each s_i.

Given constraint $\sum_{s_i} \xi(a=1|s_i)\,p(s_i) < l$, where $l \in [0,1]$ and is the total proportion of time spent sampling, we can see that if $l = p(s_1)$, then

$$\sum_{i=1}^{3} \xi(a=1|s_i)\,p(s_i)\,h(\pi(a|s_i))$$

is maximized when $\xi(a=1|s_1)=1$,
$\xi(a=1|s_2)=0$,
$\xi(a=1|s_3)=0$.

The policy that maximizes sampling in information-rich states is the one that samples the highest-entropy states as much as it can, where the entropy cutoff is $\underset{h(\pi(a|u_k))}{\arg\max} \sum_{i=1}^{k} p(u_i)\,\xi(a=1|u_i) \le l$, for u_i, the states s_i but sorted in order of decreasing entropy.

A problem with this policy is that lower-information states may never be sampled again if the first few observations are randomly low in $h(\pi(a|s_i))$.
We need a convergence guarantee, that $\hat{\mu}_{s_i}$ approaches the true value μ_{s_i} as sampling time $t \to \infty$.
This will not happen if any state's sampling rate $\xi(a=1|s_i)=0$.

If we simply approximate $\hat{\mu}_{s_i}$ as the $\frac{1}{N}\sum_{j=1}^{N} q_{\Delta j}(s_i) = \overline{q_\Delta}(s_i)$, where $q_{\Delta j}(s_i)$ is the j^{th} observed sample at s_i, then

$$\overline{q_\Delta}(s_i) \sim \mathcal{N}(\mu_{s_i}, \sigma_{s_i}^2/n)$$

and as long as $\xi(a=1|s_i) \neq 0 \;\forall s_i$ as $t \to \infty$, we will approach μ_{s_i}.

This suggests that although the maximum information policy says we should only sample at the highest information states and otherwise have $\xi(a=1|s_i)=0$, we need a nonzero sampling rate at all states for estimate $\hat{\mu}_{s_i}$ to converge to μ_{s_i}.

One way to do this is to set a baseline sampling rate β such that $\min\{\xi(a=1|s_i)\}_{i=1}^{|S|} \ge \beta$ where $|S|$ is the number of unique states.
Combining that with the maximum information policy means $\xi(a=1|s_i)$ can only be one of two values: β or 1. Then the policy becomes:

The proportion of time light is on

For a given baseline sampling rate β and overall maximum sampling rate $l \in [0,1]$, let u_i be the states ordered in decreasing entropy $h(\pi(a|u_i))$. Find the maximum k s.t.

$$\beta\left(\sum_{i=k+1}^{|S|} p(u_i)\right) + \left(\sum_{i=1}^{k} p(u_i)\right) \le l$$

$$\beta\left(1-\sum_{i=1}^{k} p(u_i)\right) + \left(\sum_{i=1}^{k} p(u_i)\right) \le l$$

$$\beta + \sum_{i=1}^{k} p(u_i)\,[1-\beta] \le l$$

$$\sum_{i=1}^{k} p(u_i) \le \frac{l-\beta}{1-\beta}$$

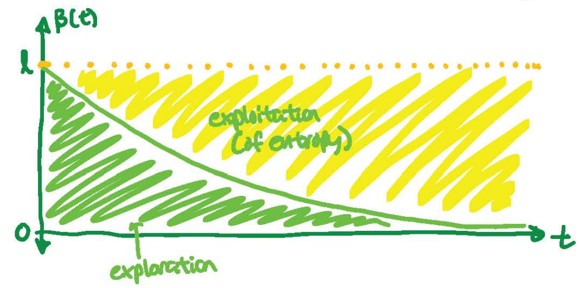
Then, $$\xi(a=1|u_i) = \begin{cases} 1 & \text{for } i=1,2,\dots,k \\ \beta & \text{for } i=k+1,k+2,\dots,|S|. \end{cases}$$

What is left now is a choice for β. Say β is a function of sampling time t. Some properties we want are that $\beta(t)$ satisfies (for $t = 0, 1, \ldots, \infty$)

- $\rightarrow \beta(t) \in (0, \ell]$ for $t \in [0, \infty)$ the baseline rate must be nonzero and $\leq \ell$, since ℓ is the maximum overall sampling rate
- $\rightarrow \beta(0) = \ell$ at first, every state is sampled as often as ℓ will allow.
- $\rightarrow \sum_{t=0}^{\infty} \beta(t) = \infty$ in the limit of $t \to \infty$, every state will be sampled ∞ times.

Not strictly necessary but preferable:

- $\rightarrow \lim_{t \to \infty} \beta(t) = 0$ in the limit, when estimates $\hat{\mu}_{s_i} \approx \mu_{s_i}$, we approach the max information policy.



Divergent series whose terms converge to 0 are some candidates. One simple option would be

$\beta(t) = \frac{c}{1+t}$ where c can be tuned based on experiment

but need to read more to figure this part out.

Plan from here:

- consider how to balance convergence & information maximization.
  - seems like an exploration/exploitation problem. Here, entropy is the "reward" from the usual RL framework and exploring to get better estimates of entropy takes the place of the usual RL approach of exploring to find reward.
  - conditions on $\beta(t)$ are essentially conditions on exploration/exploitation dilemma.
  - Thompson sampling may be applicable here: lots of convergence guarantees. Some optimal behavior guarantees as well.
- run experiments to get a sense of annealing rate (rate at which "exploration factor" $\beta(t)$ decreases) for $\beta(t)$ and expected timescales
- going from $g_\Delta(s_i)$ samples to $\hat{\mu}_{s_i}, \hat{\sigma}_{s_i}$ requires an uncertainty-aware model that can take advantage of what we know about our continuous state space.
  - We're currently using BART but the boundaries are choppy because of how trees split up an input space. Other options?