A POMDP model consists of 6 components, i.e.:

1. $S$: 3 tiger in left room (TL), tiger in right room (TR)

2. $A$: 3 open left, open right, listen

3. $O$: 1 tiger sound from left room (OL), no obs.

4. $T$: The tiger can't change room, so:
   \[
   T(TL, *, TL) = 1.0, \quad T(TL, *, TR) = 0.0, \quad T(TR, *, TL) = 0.0
   \]

5. $Z$: Localization accuracy: 0.85
   \[
   Z(TL, listen; OL) = 0.85, \quad Z(TL, listen; OR) = 0.15, \quad Z(TR, listen; OR) = 0.85, \quad Z(TR, listen; OL) = 0.15
   \]
   \[
   Z(TL, listen, NO) = Z(TR, listen, NO) = 0
   \]

6. $R$: $R(\cdot, listen) = -1 \quad \rightarrow \quad \cdot$: any state $s \in S$
   \[
   R(TL, open left) = R(TR, open right) = 20, \quad R(TL, open right) = R(TR, open left) = 10
   \]
Optimal policy for 1 planning horizon (when the robot can only perform 1 action) depends only on the immediate reward (R).

First, let's compute the immediate reward of performing an action a ∈ A for various beliefs:

- a = open left; b is distribution over S

\[ R(b, a) = \sum_s b(s) R(s, \text{open left}) \]

\[ = b(\text{TL}) \cdot R(\text{TL, open left}) + b(\text{TR}) \cdot R(\text{TR, open left}) \]

\[ = -20b(\text{TL}) + 10b(\text{TR}) \]

- a = open right

\[ R(b, a) = \text{similar to above but swap TL and TR} \]

\[ = -20b(\text{TR}) + 10b(\text{TL}) \]

- a = listen

\[ R(b, a) = -1 \]

Second, we want to find what action gives highest value (in this case: R(b, a)) given different beliefs (b). We can do this using linear programming.

...
Optimal policy:
For $b(\text{TR}) \leq \frac{11}{30}$: open right
\(\frac{11}{30} < b(\text{TR}) \leq \frac{19}{30}\): listen
\(b(\text{TR}) > \frac{19}{30}\): open left.

Recall that $b(\text{TR}) + b(\text{TL}) = 1$ (probability)
Therefore stating what action to perform w.r.t. $b(\text{TR})$ is sufficient.