Exact POMDP Solutions: α-vectors

Recap

Recap

- POMDP

Recap

- POMDP $(S, A, O, R, T, Z, \gamma)$

Recap

- POMDP
- Belief Updates

$$(S, A, O, R, T, Z, \gamma)$$

Recap

- POMDP $(S, A, O, R, T, Z, \gamma)$
- Belief Updates

$$b_t(s) = P(s_t = s \mid h_t)$$

Recap

- POMDP $(S, A, O, R, T, Z, \gamma)$
- Belief Updates

$$b_t(s) = P(s_t = s \mid h_t)$$

$$b' = \tau(b, a, o)$$

Recap

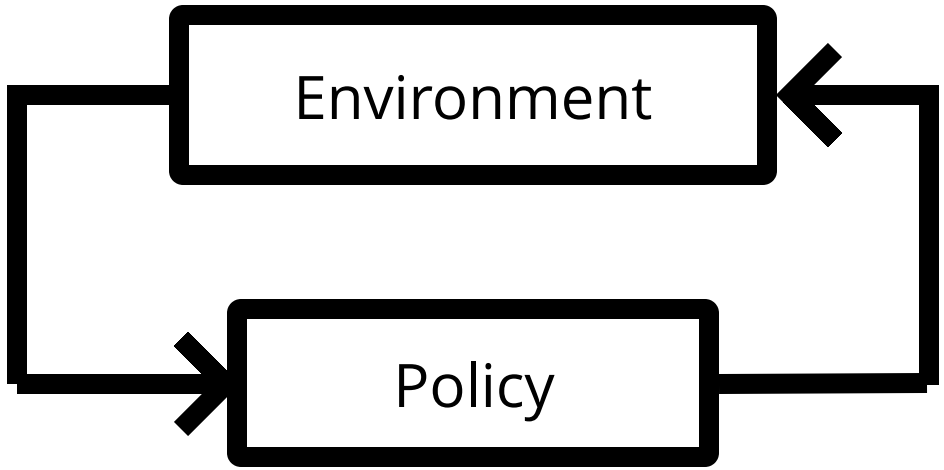
- POMDP $(S, A, O, R, T, Z, \gamma)$
- Belief Updates

$$b_t(s) = P(s_t = s \mid h_t)$$

$$b' = \tau(b, a, o)$$

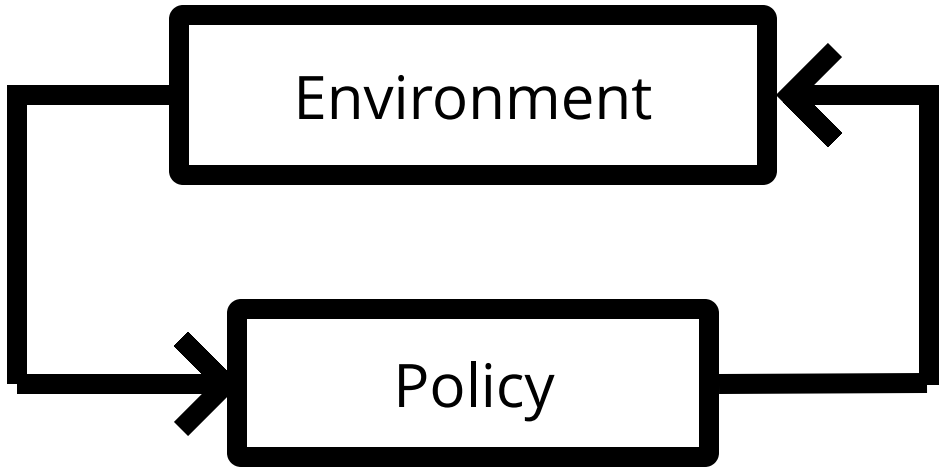
$$b'(s') \propto Z(o \mid a, s') \sum_s T(s' \mid s, a)\, b(s)$$

MDP Sense-Plan-Act Loop

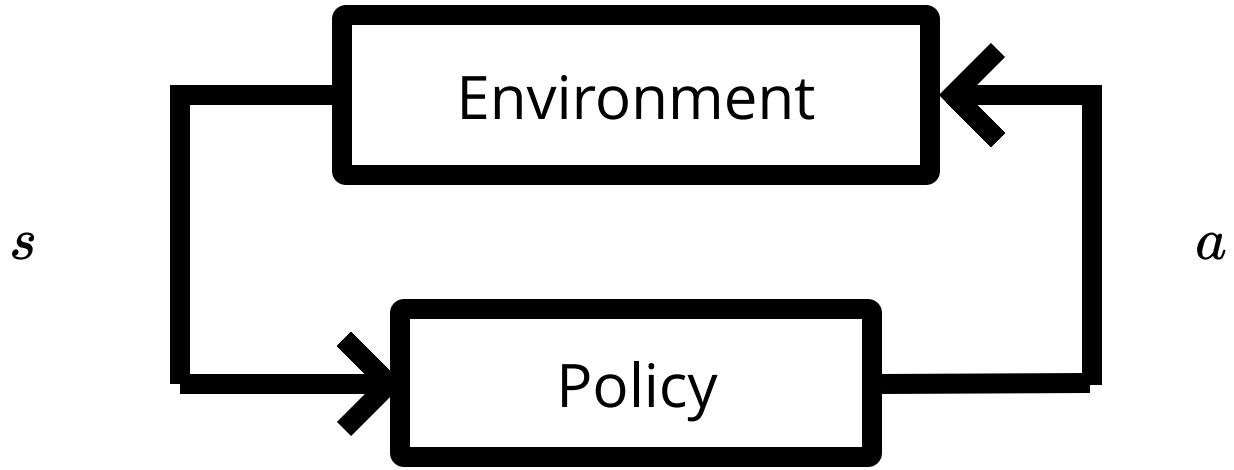


MDP Sense-Plan-Act Loop

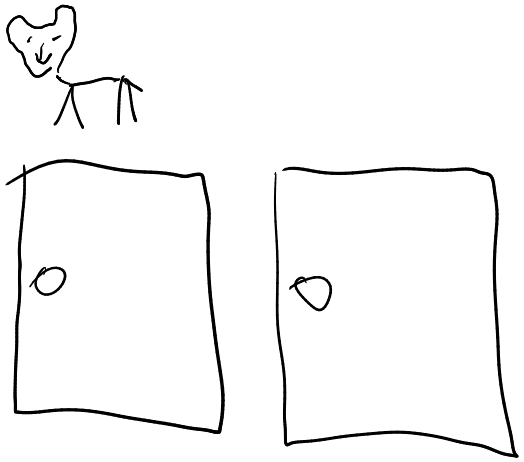
s

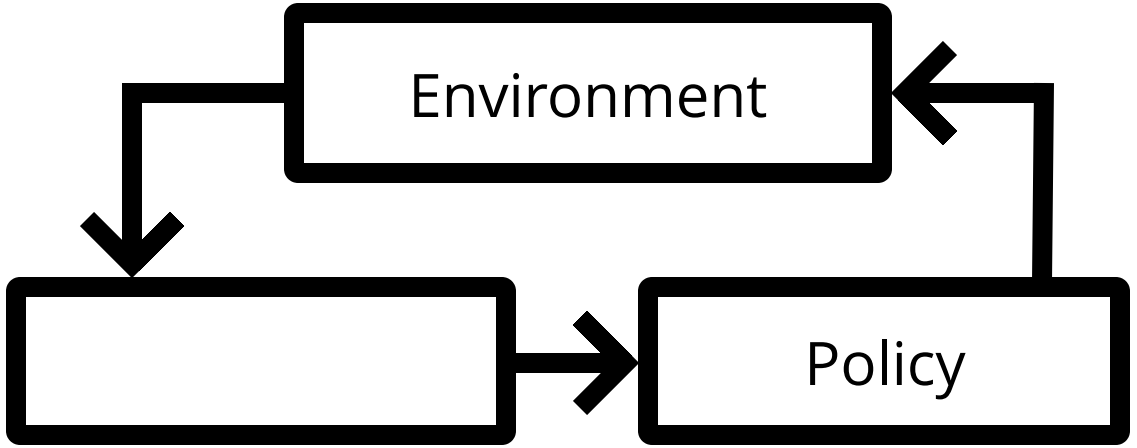


MDP Sense-Plan-Act Loop



POMDP Sense-Plan-Act Loop

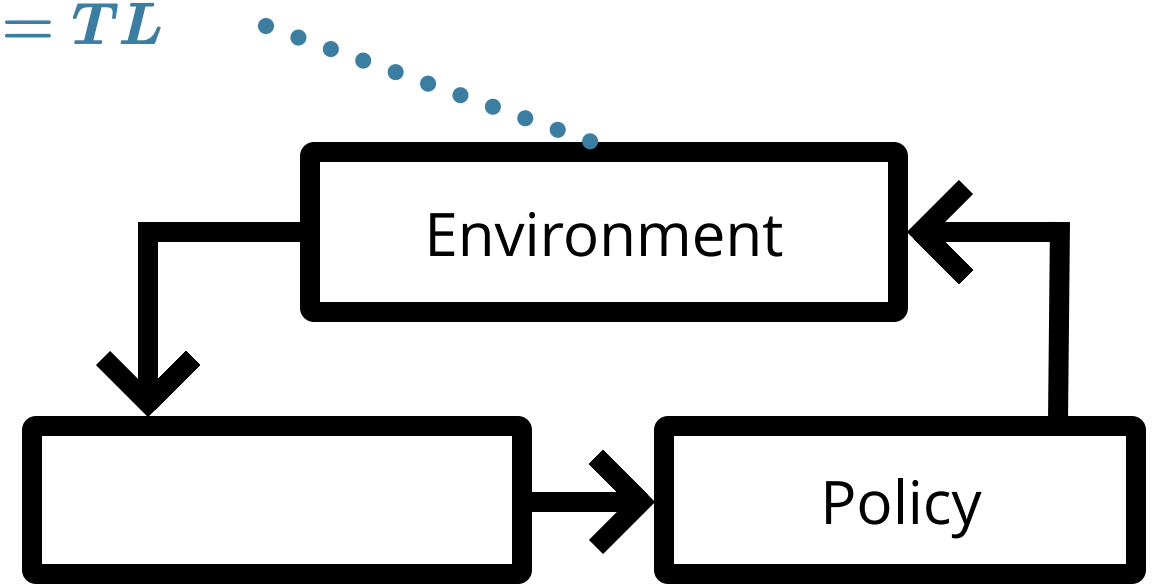




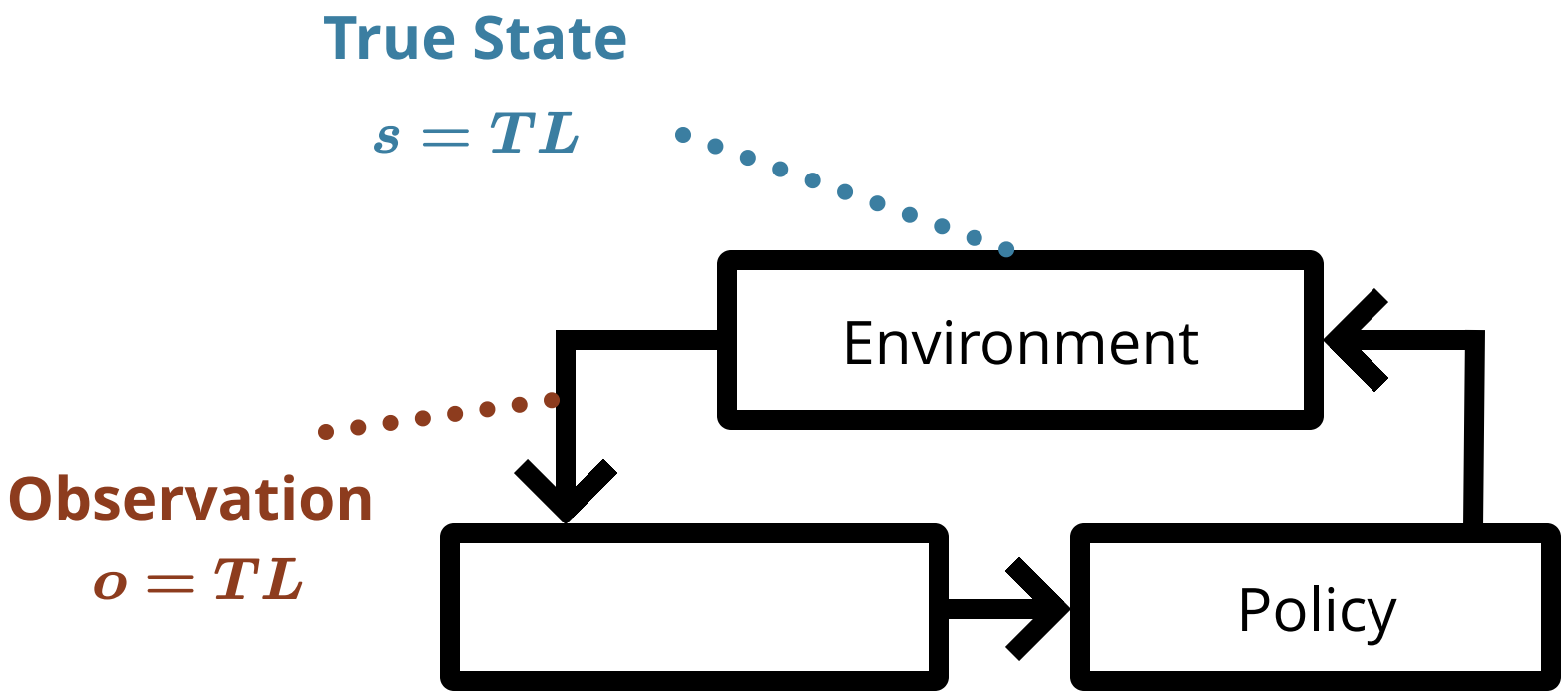
POMDP Sense-Plan-Act Loop

True State

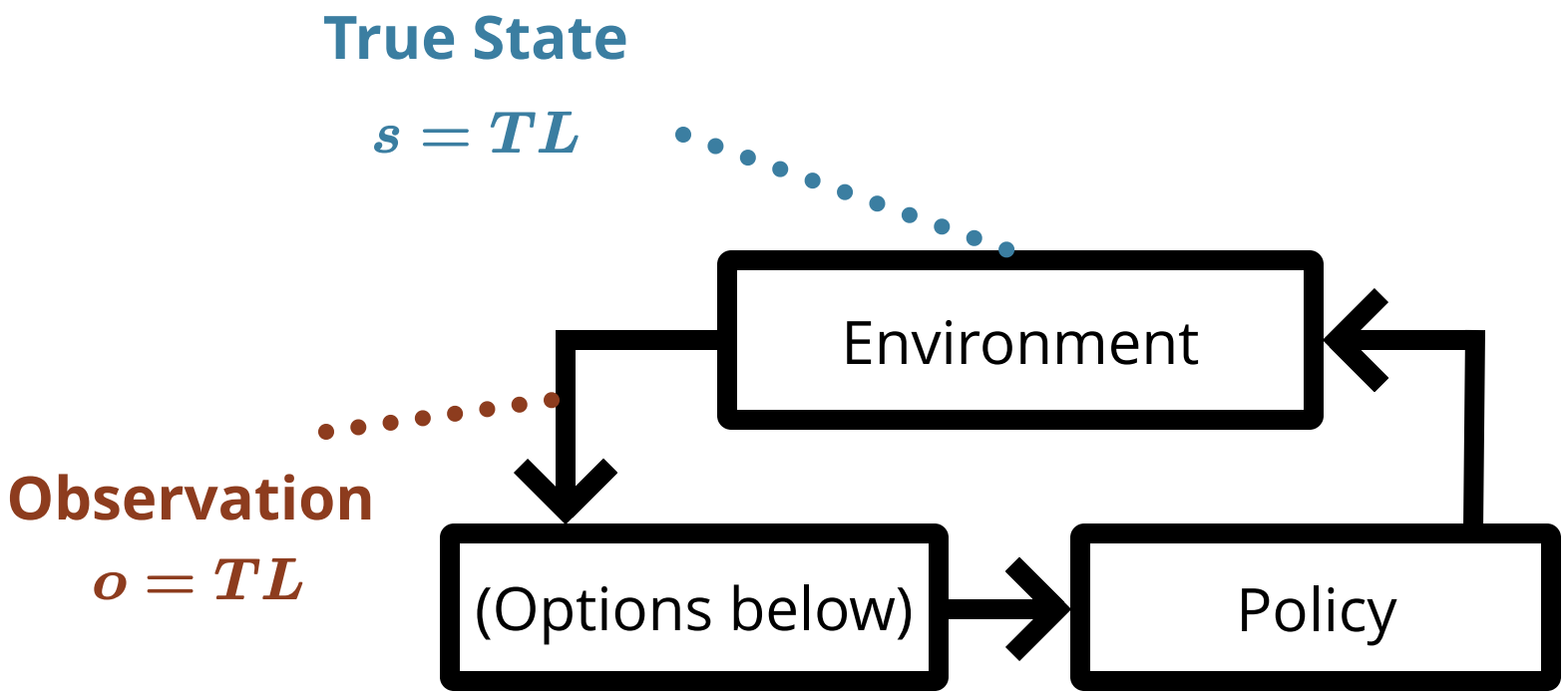
$s = TL$



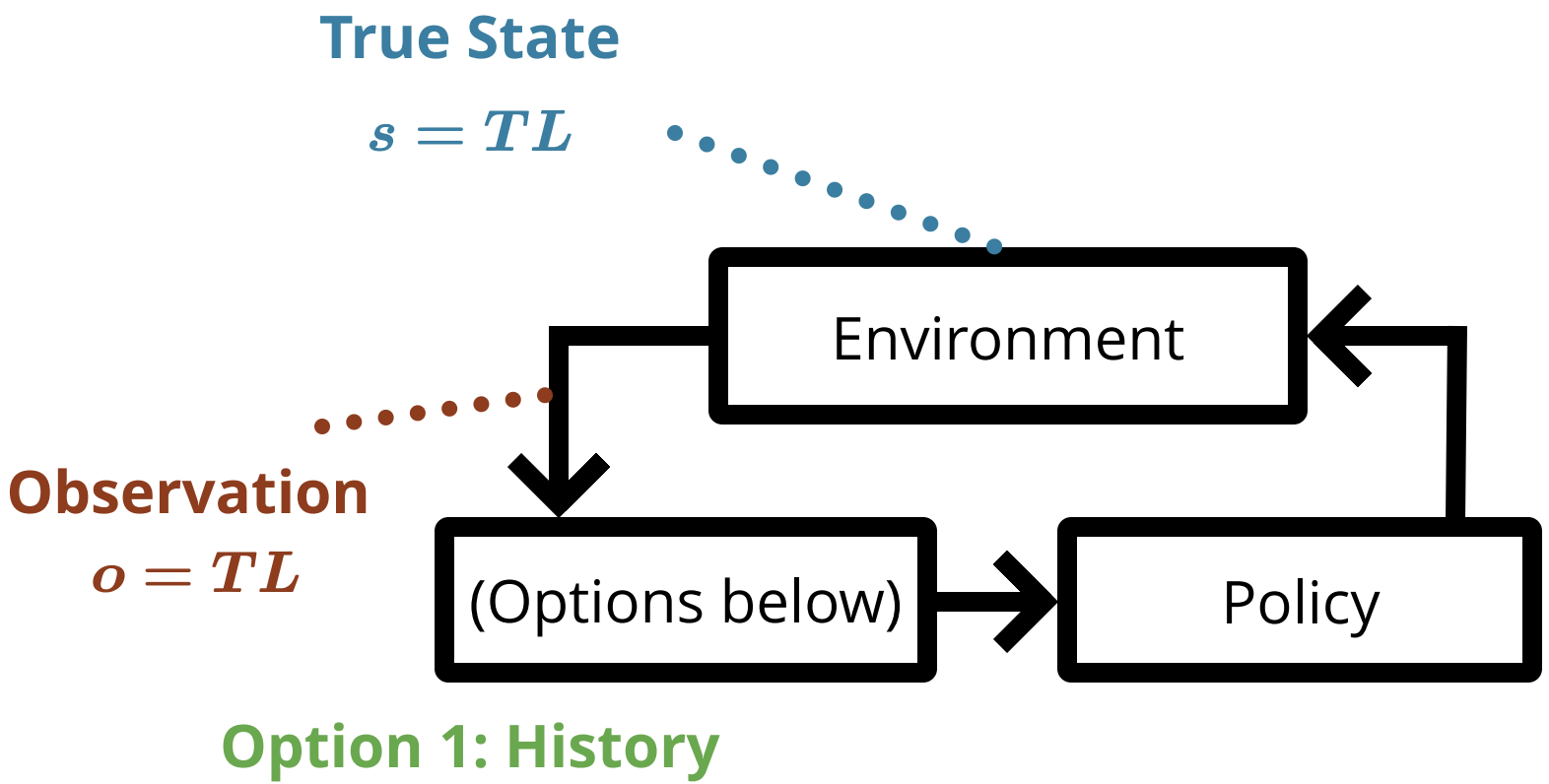
POMDP Sense-Plan-Act Loop



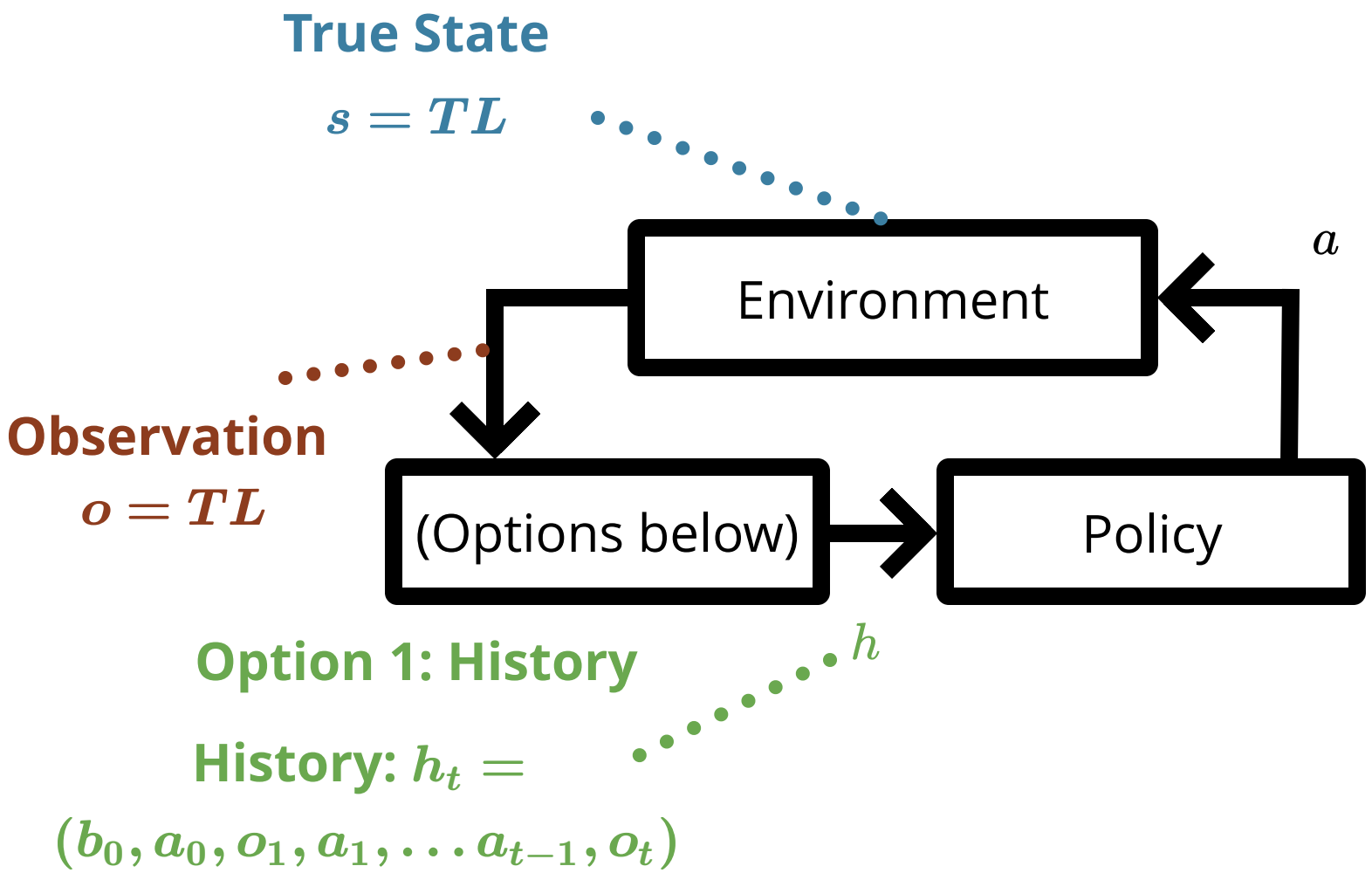
POMDP Sense-Plan-Act Loop



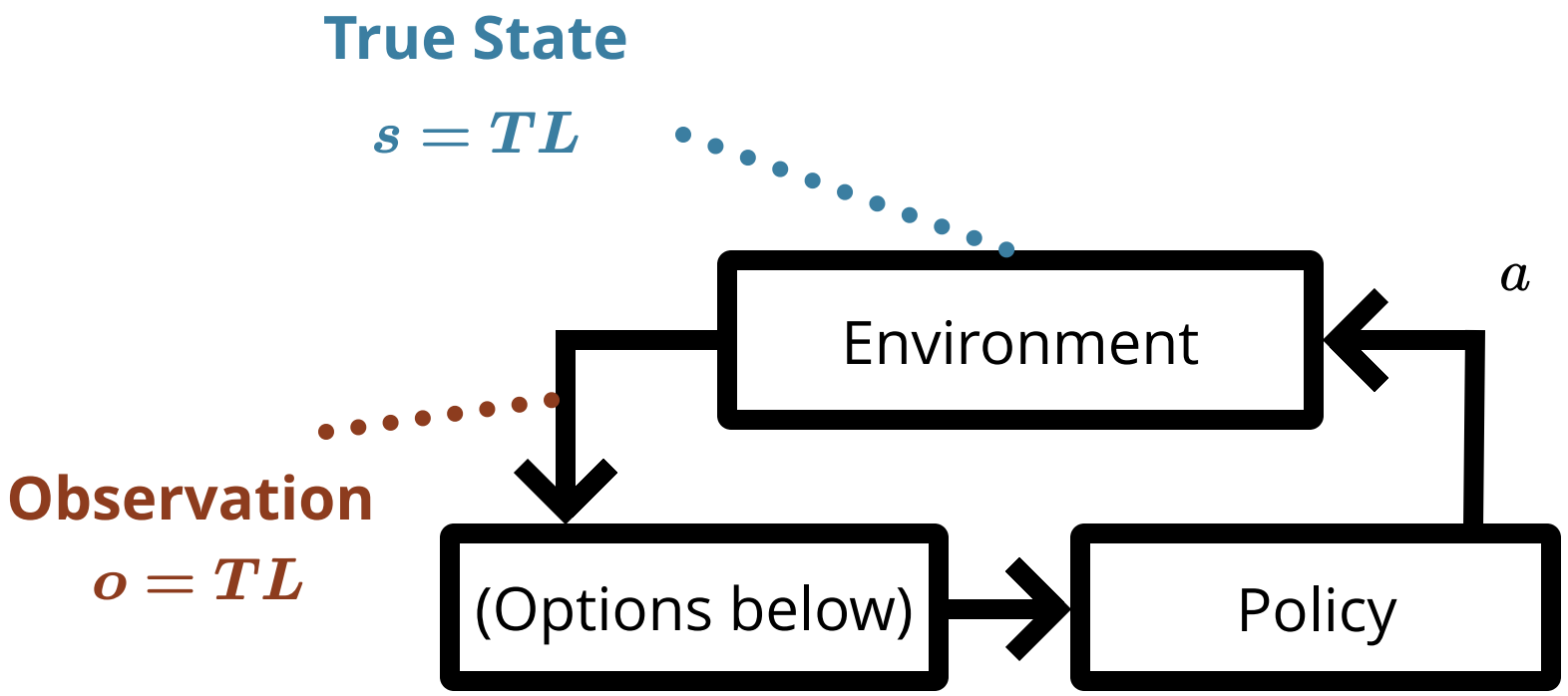
POMDP Sense-Plan-Act Loop



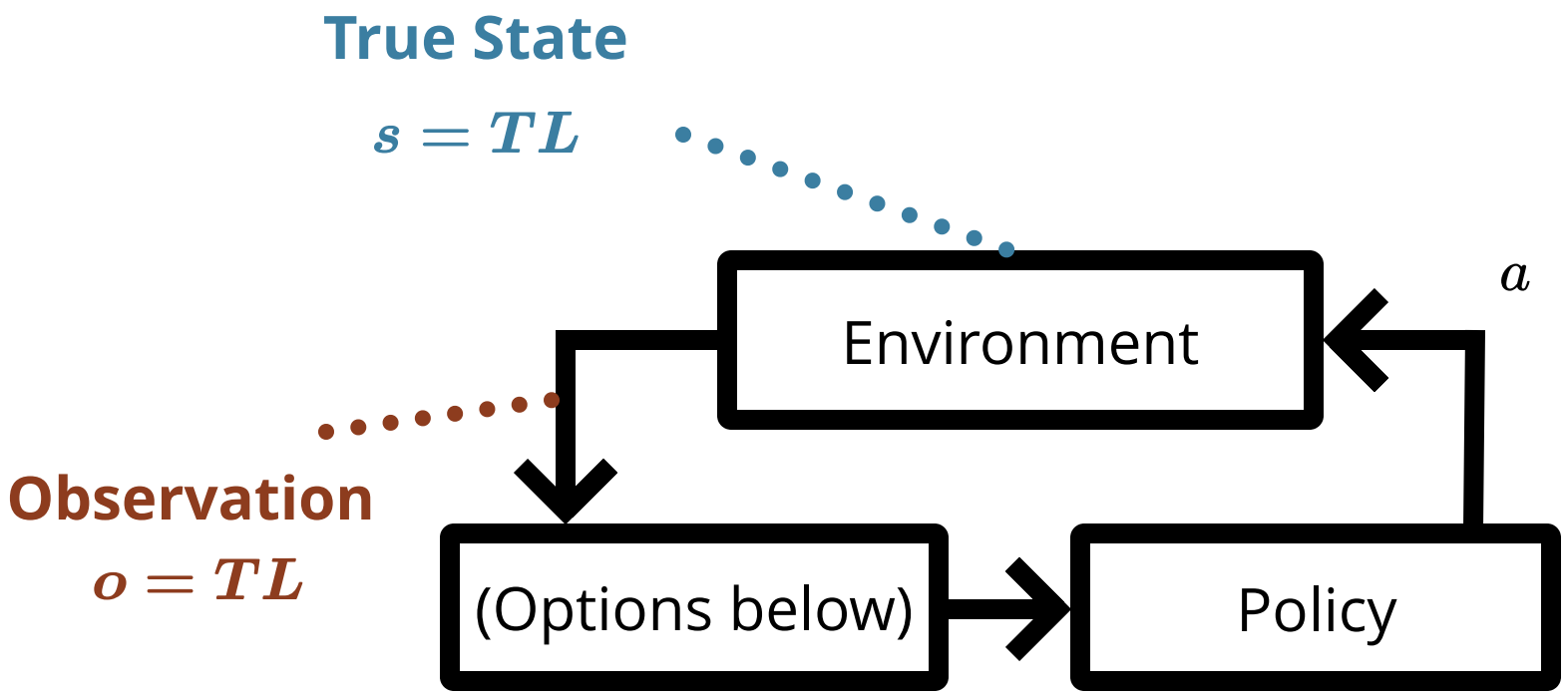
POMDP Sense-Plan-Act Loop



POMDP Sense-Plan-Act Loop



POMDP Sense-Plan-Act Loop



Exercise 1: Crying Baby Belief Update

Exercise 1: Crying Baby Belief Update

$S = \{h, \neg h\}$

$A = \{f, \neg f\}$

$O = \{c, \neg c\}$

Exercise 1: Crying Baby Belief Update

$S = \{h, \neg h\}$

$A = \{f, \neg f\}$

$O = \{c, \neg c\}$

$R(s, a) = R(s) + R(a)$

Exercise 1: Crying Baby Belief Update

$$S = \{h, \neg h\}$$

$$A = \{f, \neg f\}$$

$$O = \{c, \neg c\}$$

$$R(s, a) = R(s) + R(a)$$

$$R(s) = \begin{cases} -10 \text{ if } s = h \\ 0 \text{ otherwise} \end{cases}$$

Exercise 1: Crying Baby Belief Update

$S = \{h, \neg h\}$

$A = \{f, \neg f\}$

$O = \{c, \neg c\}$

$$R(s, a) = R(s) + R(a)$$

$$R(s) = \begin{cases} -10 \text{ if } s = h \\ 0 \text{ otherwise} \end{cases}$$

$$R(a) = \begin{cases} -5 \text{ if } a = f \\ 0 \text{ otherwise} \end{cases}$$

Exercise 1: Crying Baby Belief Update

$S = \{h, \neg h\}$ $\quad$ $T(h \mid h, \neg f) = 1.0$

$A = \{f, \neg f\}$ $\quad$ $T(h \mid \neg h, \neg f) = 0.1$

$O = \{c, \neg c\}$ $\quad$ $T(\neg h \mid \cdot, f) = 1.0$

$$R(s, a) = R(s) + R(a)$$

$$R(s) = \begin{cases} -10 \text{ if } s = h \\ 0 \text{ otherwise} \end{cases}$$

$$R(a) = \begin{cases} -5 \text{ if } a = f \\ 0 \text{ otherwise} \end{cases}$$

Exercise 1: Crying Baby Belief Update

$S = \{h, \neg h\}$

$A = \{f, \neg f\}$

$O = \{c, \neg c\}$

$T(h \mid h, \neg f) = 1.0$

$T(h \mid \neg h, \neg f) = 0.1$

$T(\neg h \mid \cdot, f) = 1.0$

$$b'(s') \propto Z(o \mid a, s') \sum_s T(s' \mid s, a)\, b(s)$$

$$R(s, a) = R(s) + R(a)$$

$$R(s) = \begin{cases} -10 \text{ if } s = h \\ 0 \text{ otherwise} \end{cases}$$

$$R(a) = \begin{cases} -5 \text{ if } a = f \\ 0 \text{ otherwise} \end{cases}$$

$Z(c \mid \cdot, h) = 0.8)$

$Z(c \mid \cdot, \neg h) = 0.1$

Exercise 1: Crying Baby Belief Update

$S = \{h, \neg h\}$

$A = \{f, \neg f\}$

$O = \{c, \neg c\}$

$T(h \mid h, \neg f) = 1.0$

$T(h \mid \neg h, \neg f) = 0.1$

$T(\neg h \mid \cdot, f) = 1.0$

$$b'(s') \propto Z(o \mid a, s') \sum_s T(s' \mid s, a)\, b(s)$$

$$R(s, a) = R(s) + R(a)$$

$$R(s) = \begin{cases} -10 \text{ if } s = h \\ 0 \text{ otherwise} \end{cases}$$

$$R(a) = \begin{cases} -5 \text{ if } a = f \\ 0 \text{ otherwise} \end{cases}$$

$Z(c \mid \cdot, h) = 0.8)$

$Z(c \mid \cdot, \neg h) = 0.1$

$\gamma = 0.9$

Exercise 1: Crying Baby Belief Update

$S = \{h, \neg h\}$

$A = \{f, \neg f\}$

$O = \{c, \neg c\}$

$T(h \mid h, \neg f) = 1.0$

$T(h \mid \neg h, \neg f) = 0.1$

$T(\neg h \mid \cdot, f) = 1.0$

$$b'(s') \propto Z(o \mid a, s') \sum_s T(s' \mid s, a)\, b(s)$$

Starting at a $b(h) = 0$, calculate b' with $a = \neg f$ and $o = c$.

$$R(s, a) = R(s) + R(a)$$

$$R(s) = \begin{cases} -10 \text{ if } s = h \\ 0 \text{ otherwise} \end{cases}$$

$$R(a) = \begin{cases} -5 \text{ if } a = f \\ 0 \text{ otherwise} \end{cases}$$

$Z(c \mid \cdot, h) = 0.8)$

$Z(c \mid \cdot, \neg h) = 0.1$

$\gamma = 0.9$

$b'(h) \propto Z(c \mid \neg f, h)\,(T(h \mid h, \neg f)\, b(h) + T(h \mid \neg h, \neg f)\, b(\neg h))$

$0.8\,(1.0 \cdot 0 + 0.1 \cdot 1.0)$

$b'(h) \propto 0.08$

$b'(\neg h) \propto Z(c \mid \neg f, \neg h)\,(T(\neg h \mid h, \neg f)\, b(h) + T(\neg h \mid \neg h, \neg f)\, b(\neg h))$

$0.1\,(0.0 + 0.9 \cdot 1.0)$

$b'(\neg h) \propto 0.09$

$b'(h) = \frac{0.08}{0.08 + 0.09} = 47\%$

$b'(\neg h) = 53\%$

Belief Dynamics

Belief Dynamics

$$b'(s') \propto Z(o \mid a, s') \sum_{s} T(s' \mid s, a)\, b(s)$$

Belief Dynamics

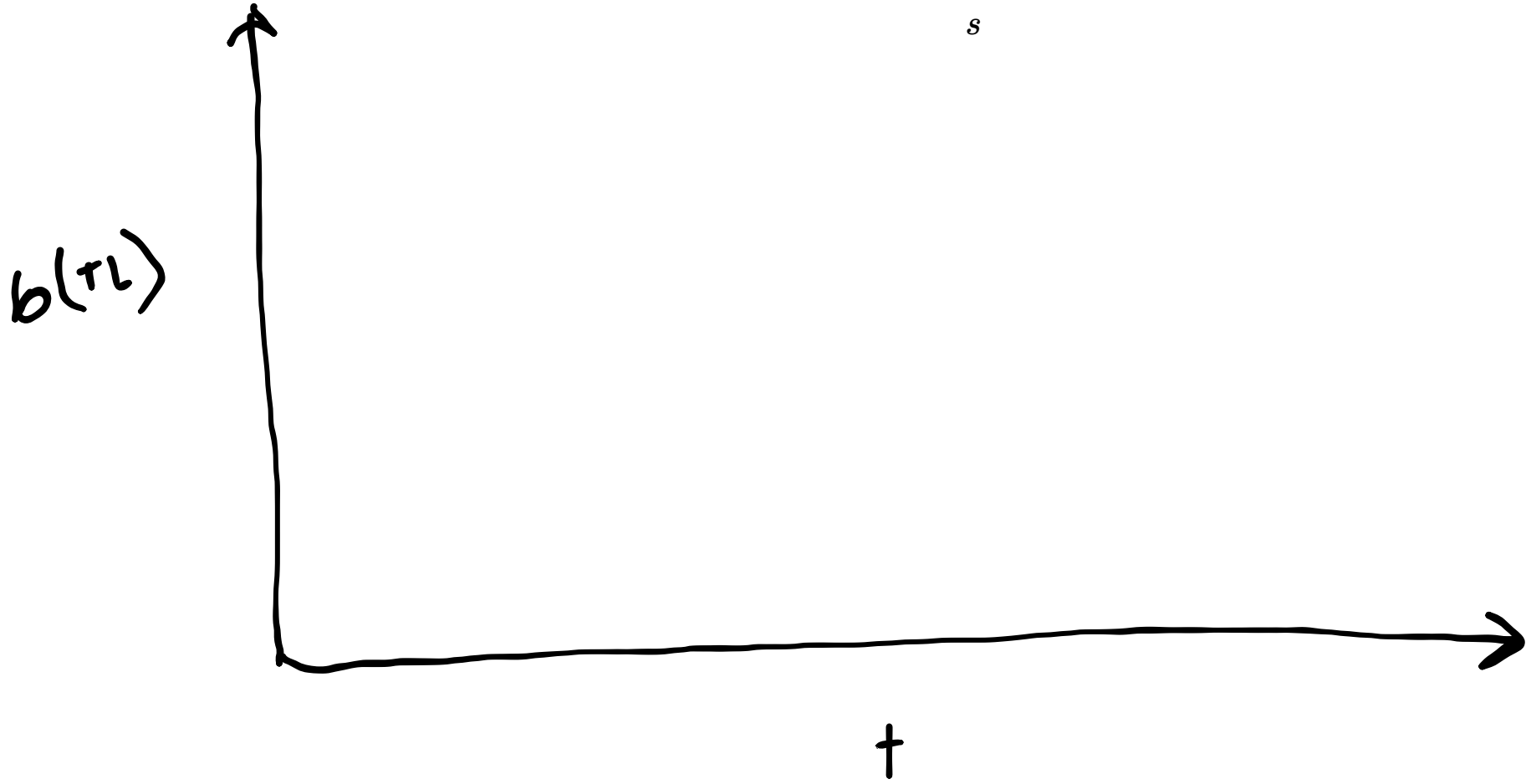
$$b'(s') \propto Z(o \mid a, s') \sum_{s} T(s' \mid s, a)\, b(s)$$

b(TL)

†

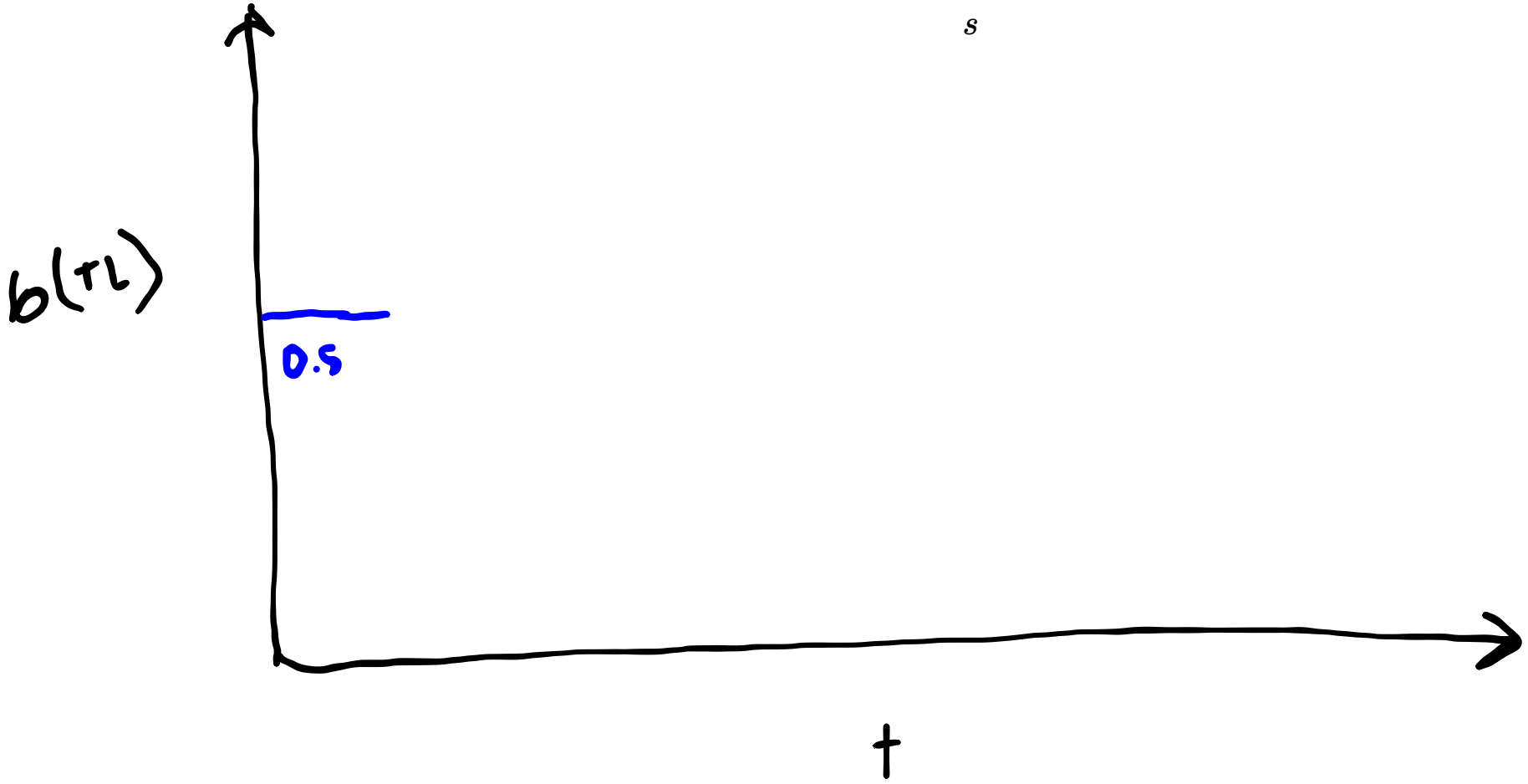
Belief Dynamics

$$b'(s') \propto Z(o \mid a, s') \sum_{s} T(s' \mid s, a)\, b(s)$$



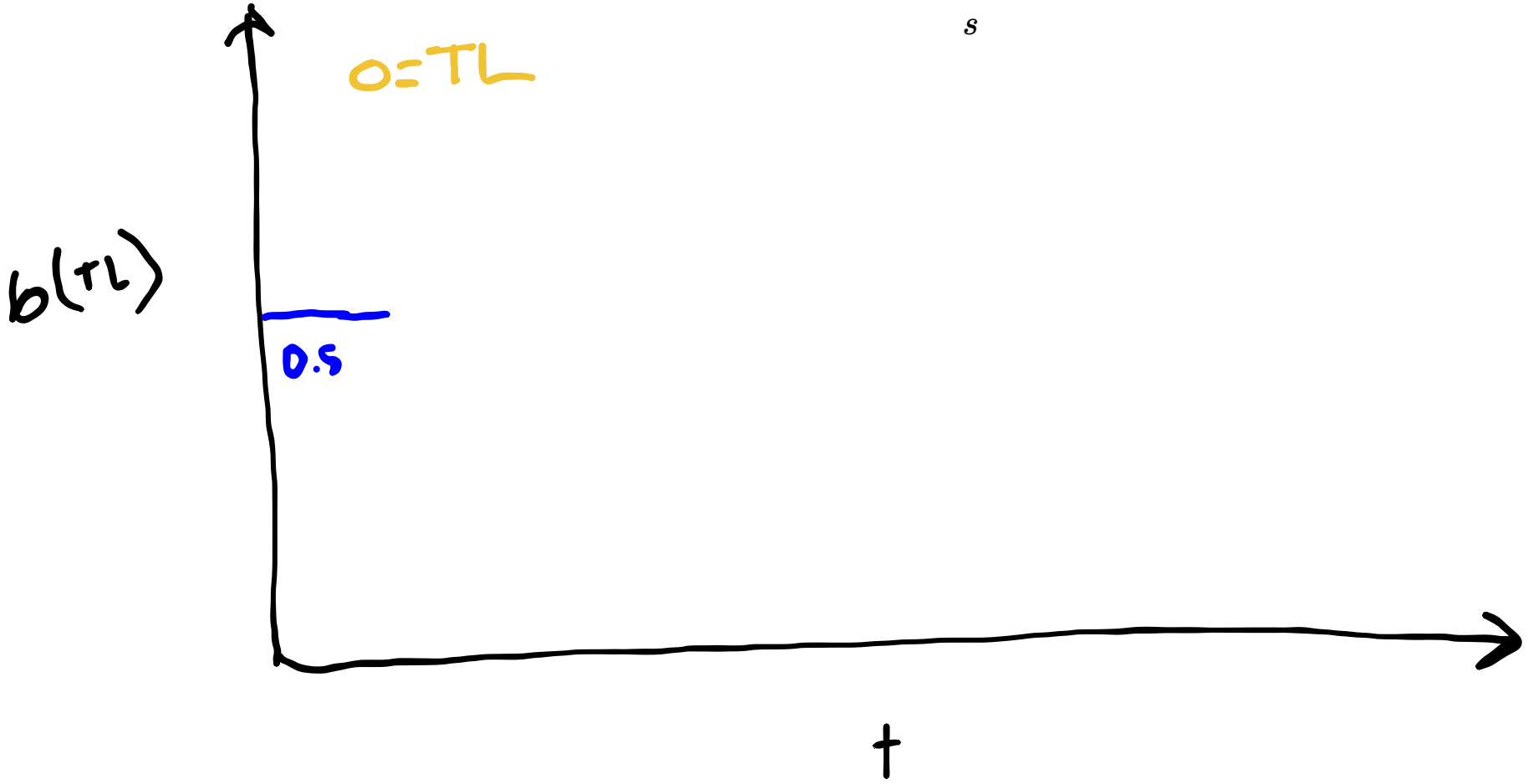
Belief Dynamics

$$b'(s') \propto Z(o \mid a, s') \sum_{s} T(s' \mid s, a)\, b(s)$$



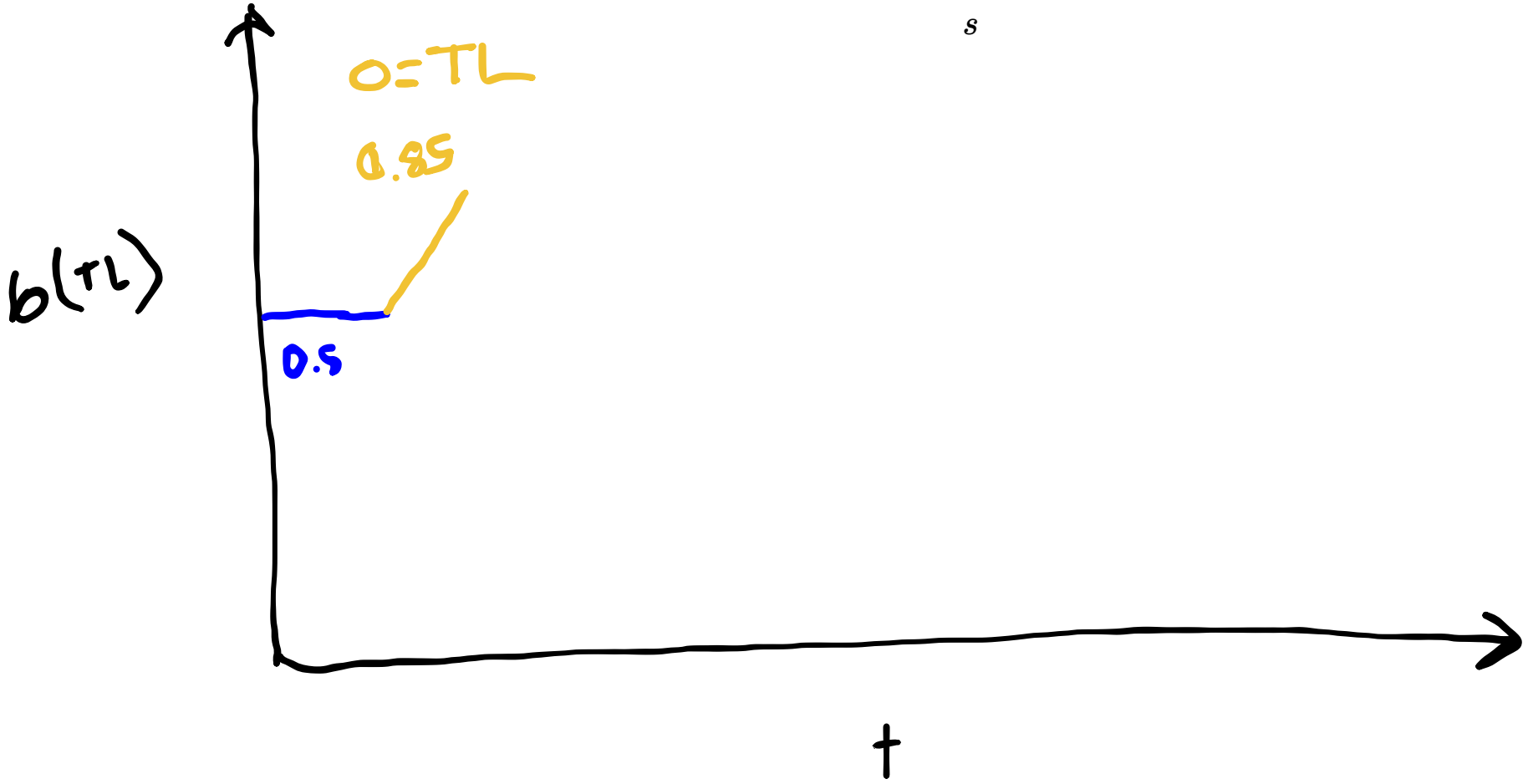
Belief Dynamics

$$b'(s') \propto Z(o \mid a, s') \sum_{s} T(s' \mid s, a)\, b(s)$$



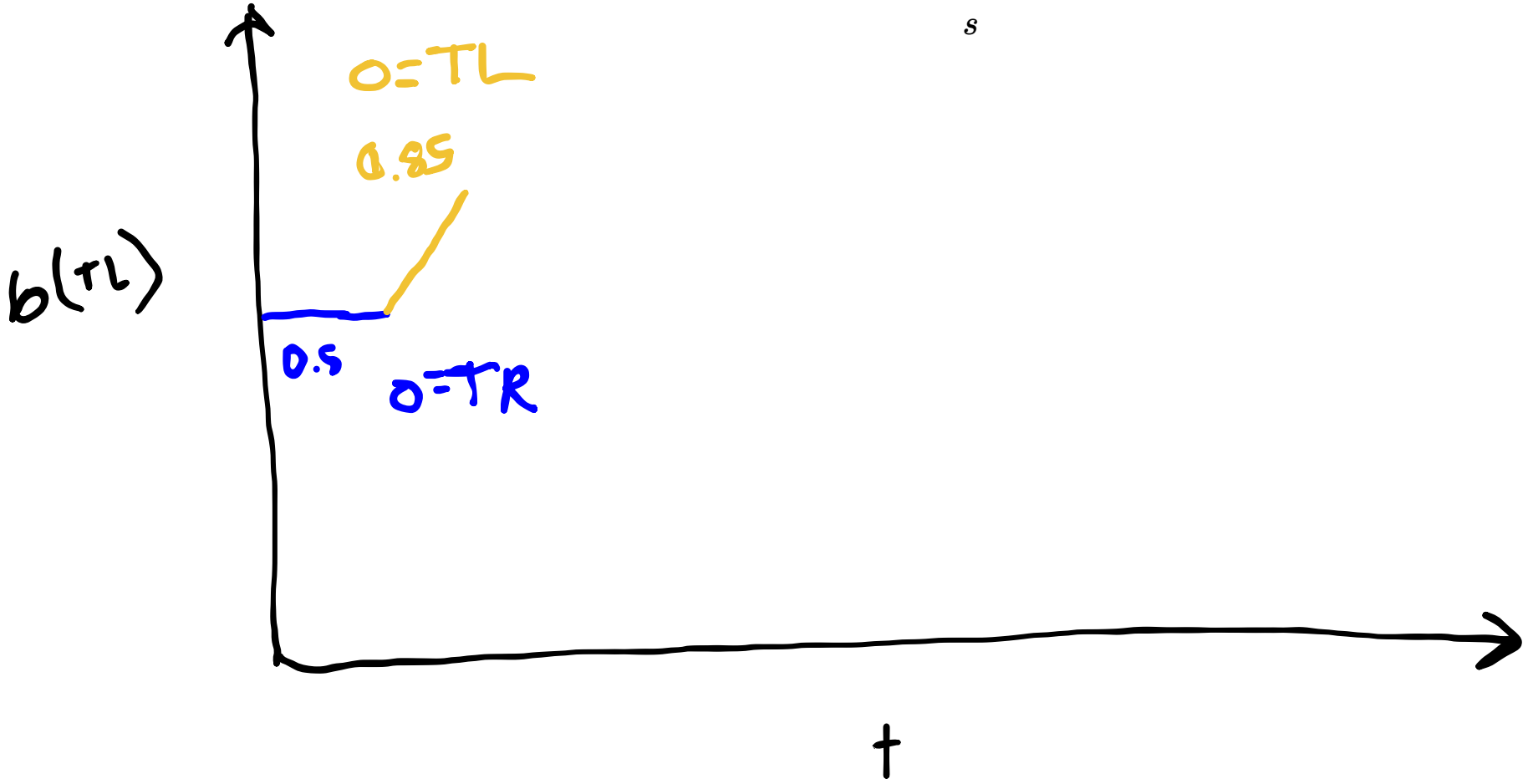
Belief Dynamics

$$b'(s') \propto Z(o \mid a, s') \sum_{s} T(s' \mid s, a)\, b(s)$$



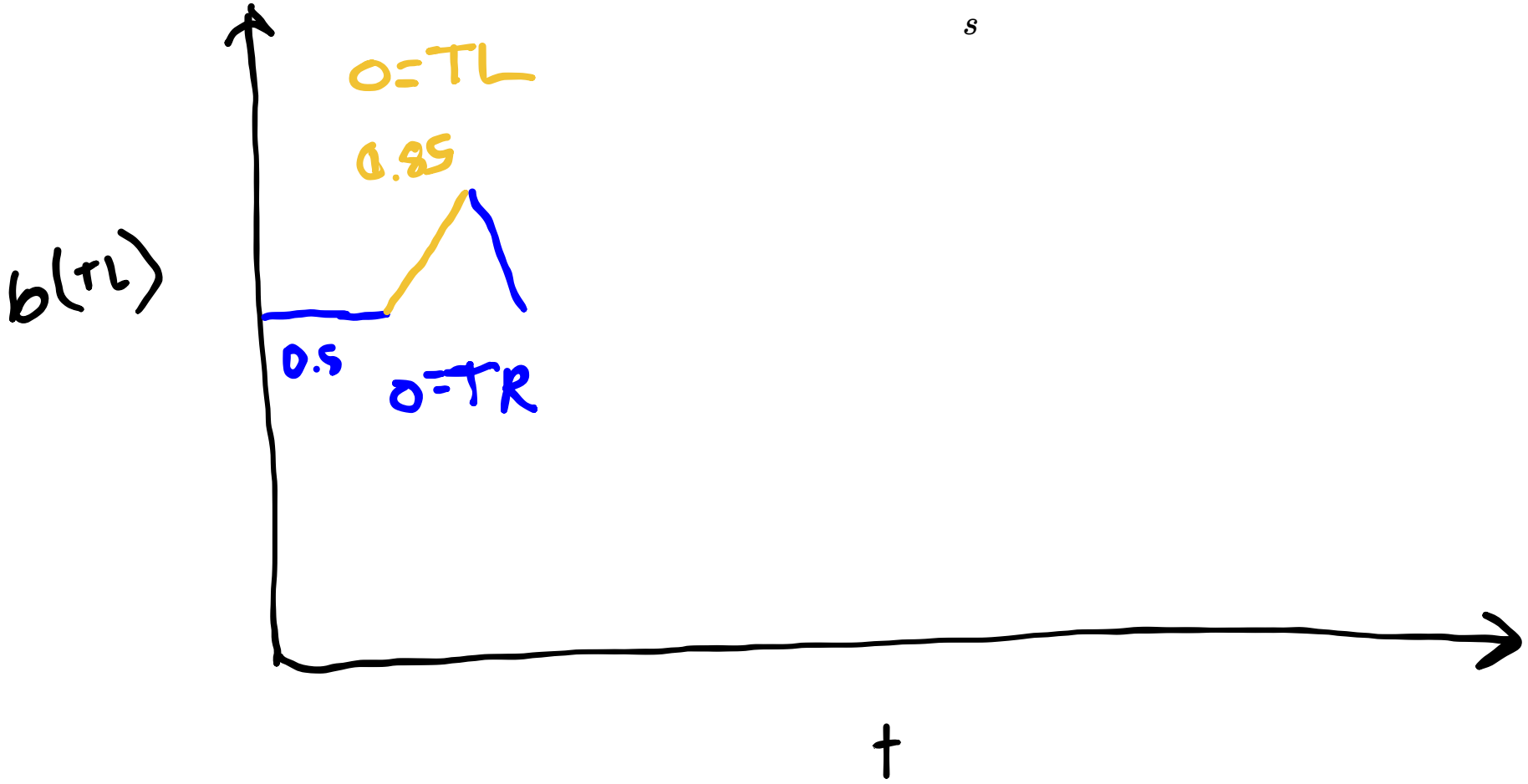
Belief Dynamics

$$b'(s') \propto Z(o \mid a, s') \sum_{s} T(s' \mid s, a)\, b(s)$$



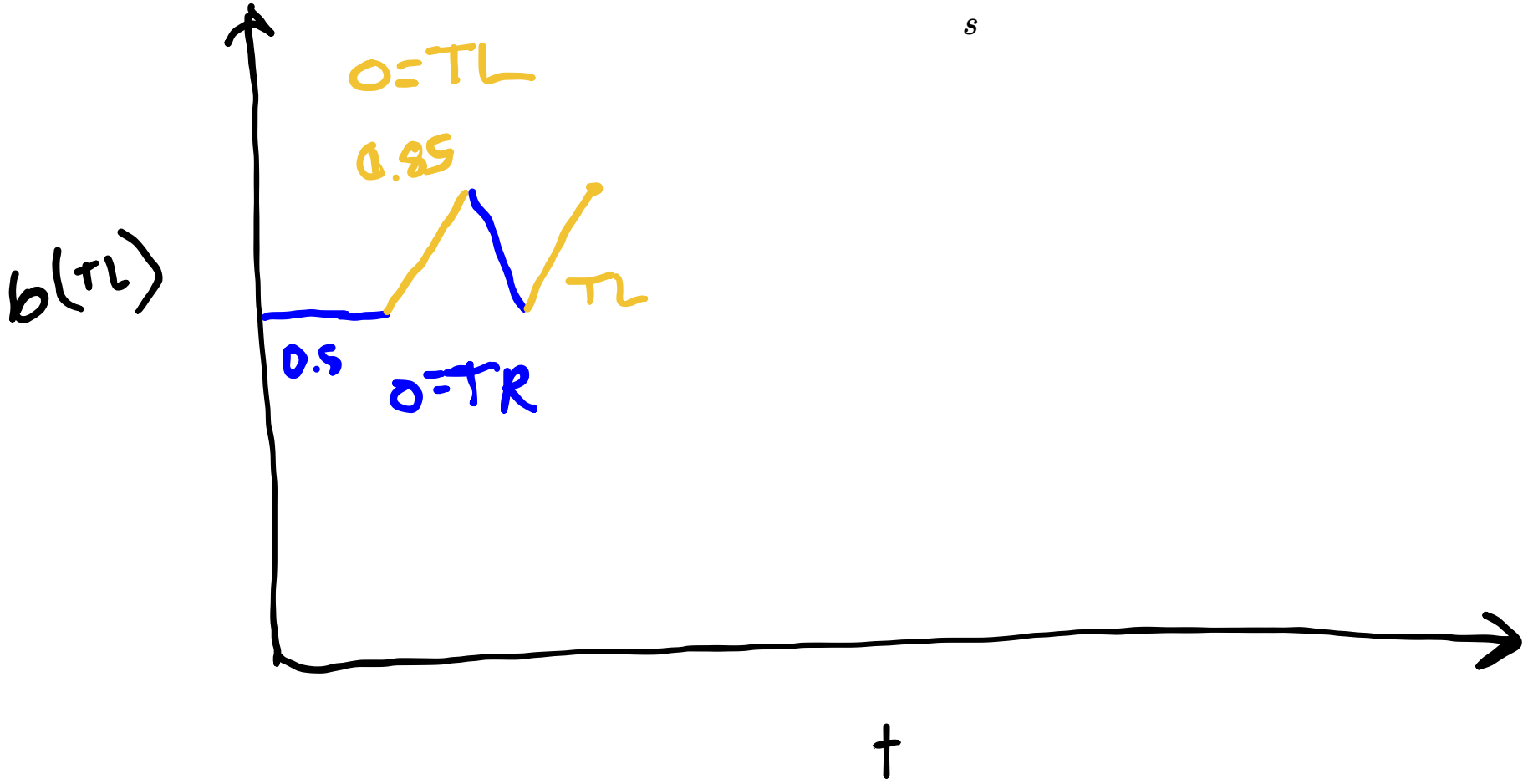
Belief Dynamics

$$b'(s') \propto Z(o \mid a, s') \sum_{s} T(s' \mid s, a)\, b(s)$$



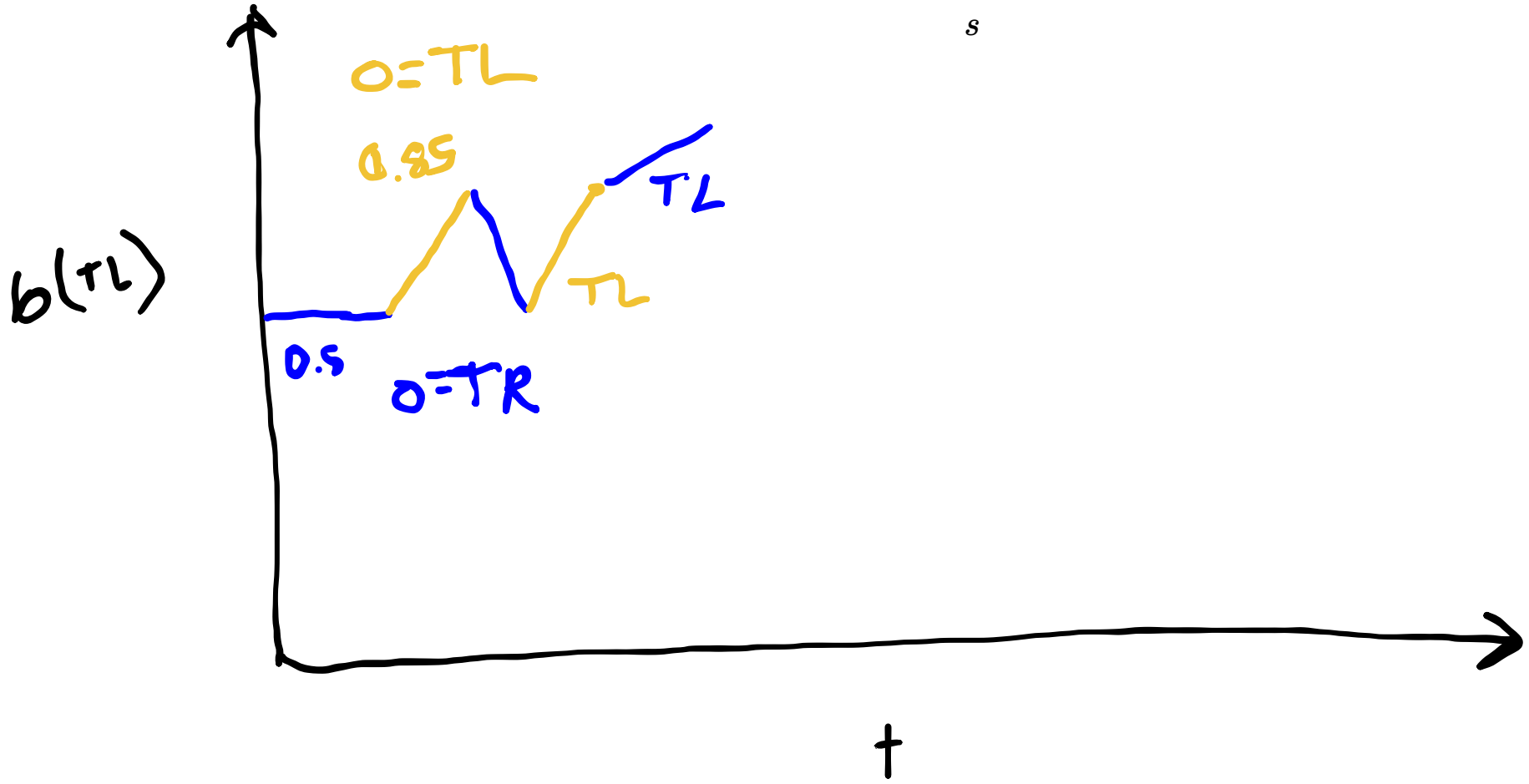
Belief Dynamics

$$b'(s') \propto Z(o \mid a, s') \sum_{s} T(s' \mid s, a)\, b(s)$$



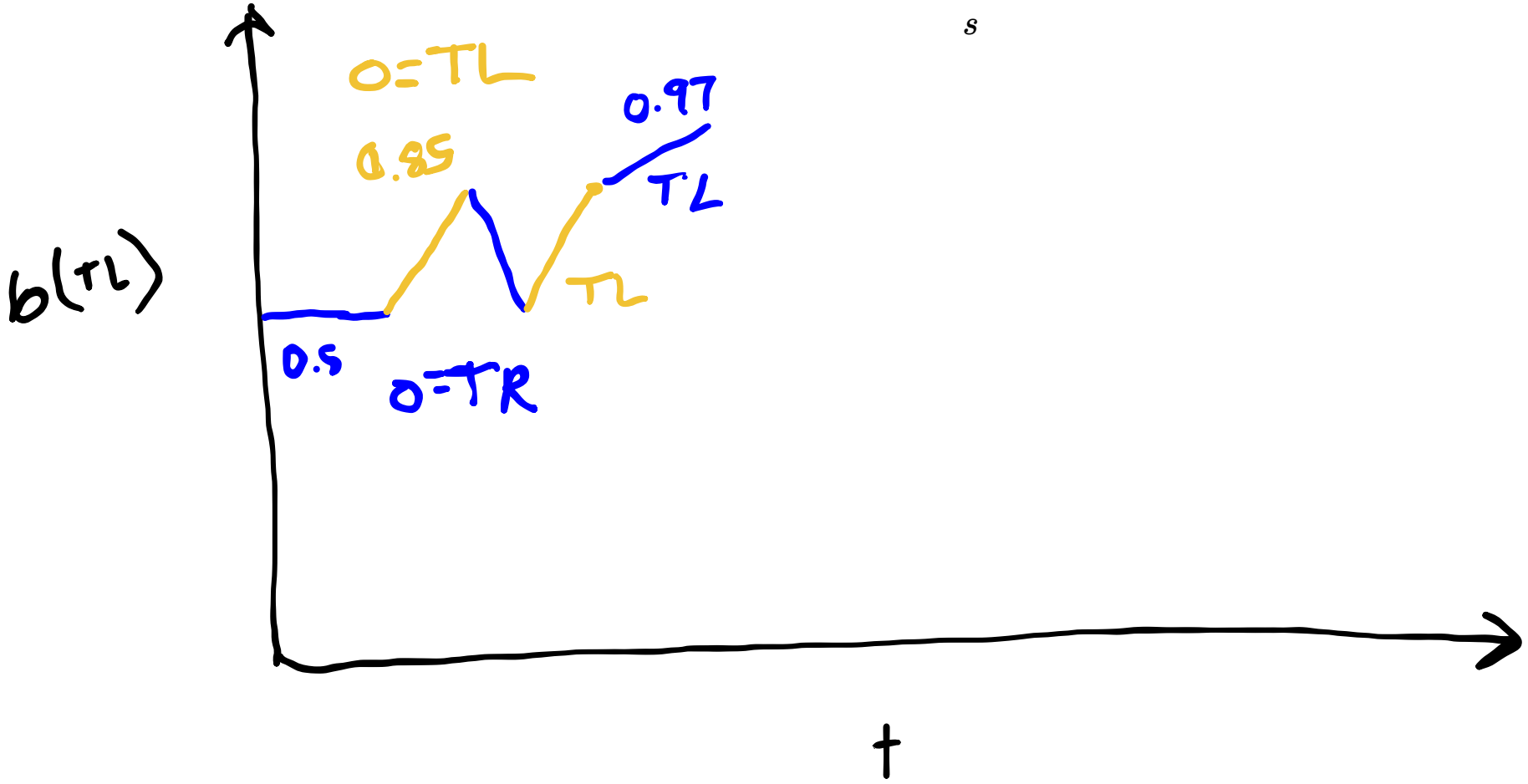
Belief Dynamics

$$b'(s') \propto Z(o \mid a, s') \sum_{s} T(s' \mid s, a)\, b(s)$$



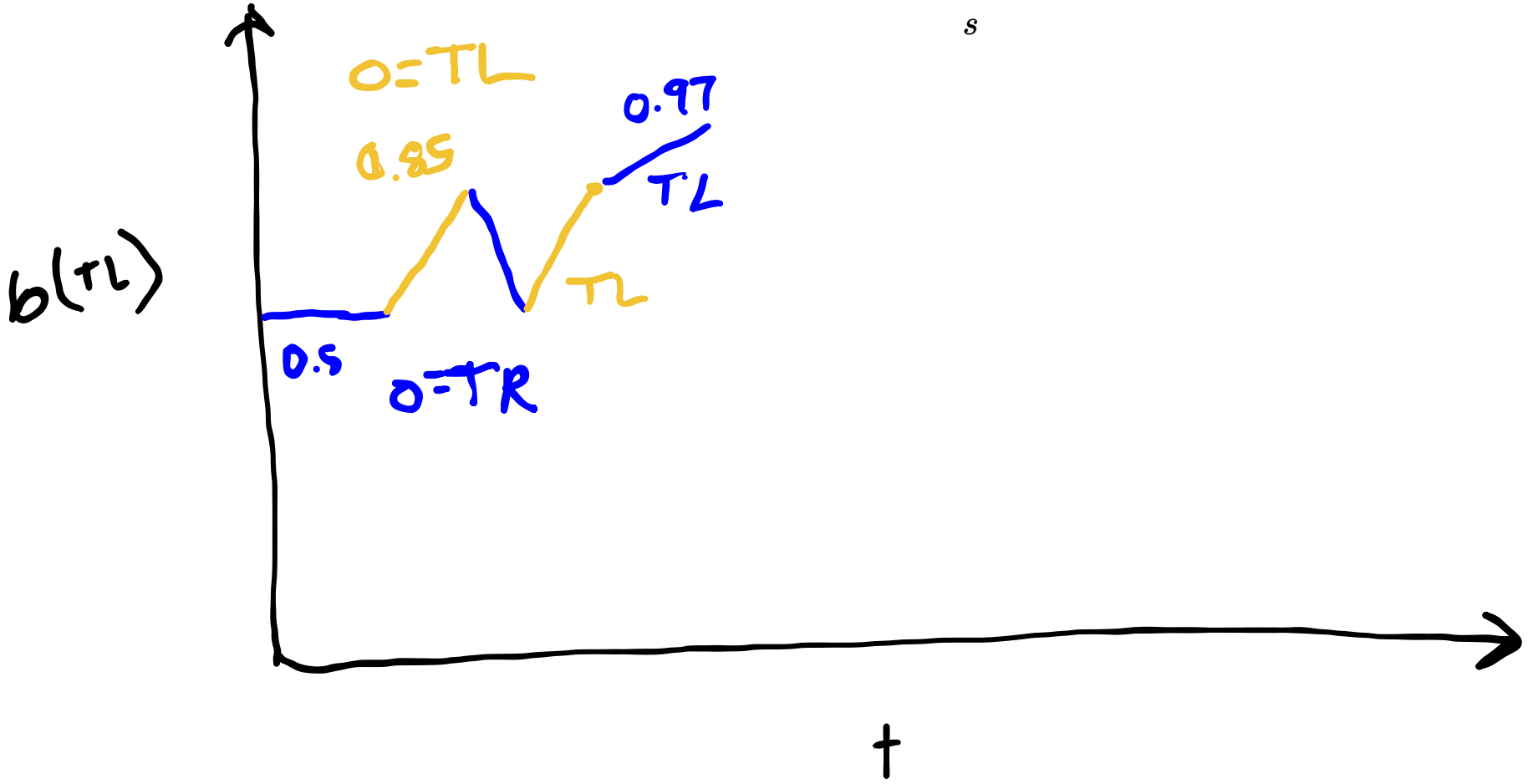
Belief Dynamics

$$b'(s') \propto Z(o \mid a, s') \sum_{s} T(s' \mid s, a)\, b(s)$$



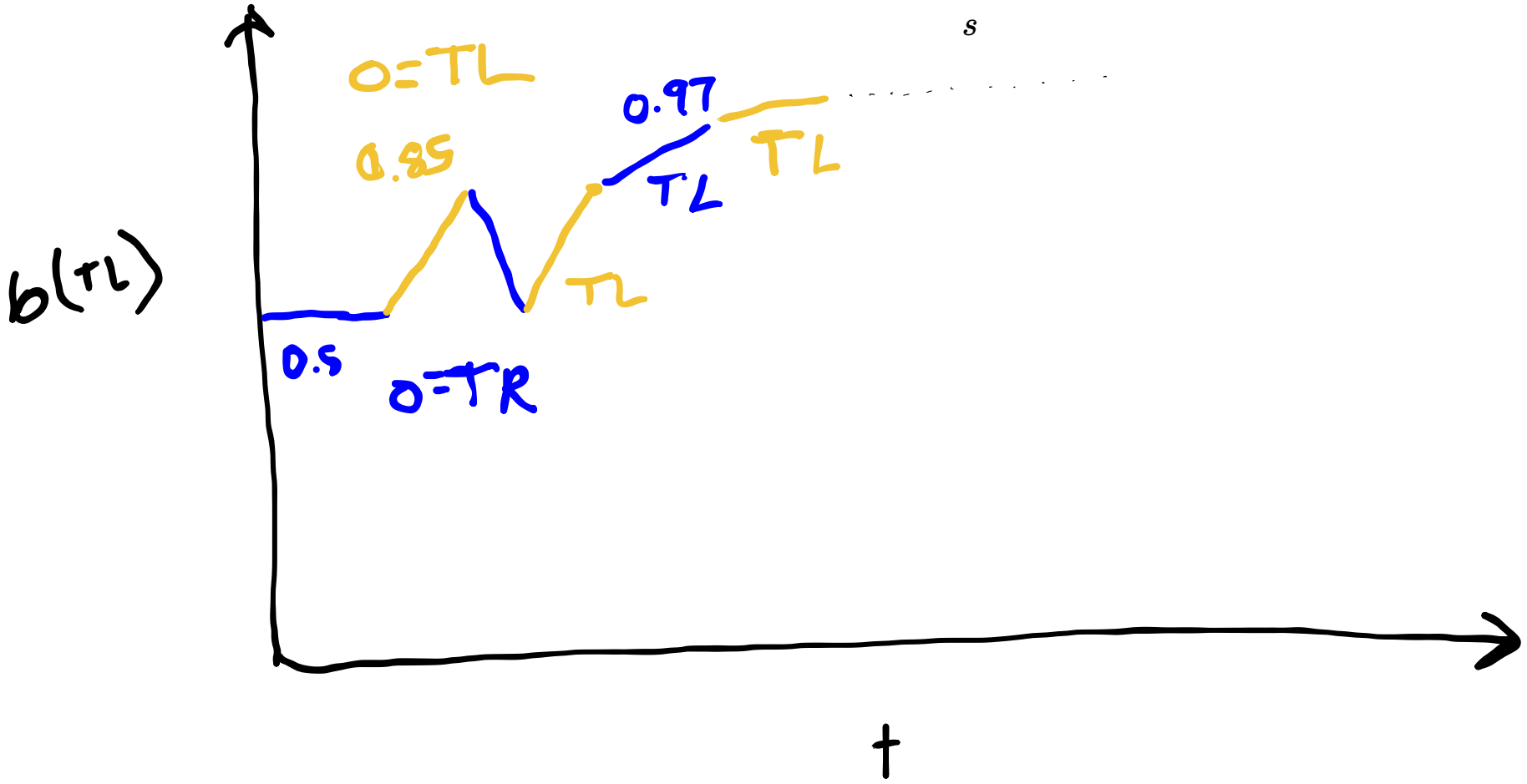
Belief Dynamics

$$b'(s') \propto Z(o \mid a, s') \sum_{s} T(s' \mid s, a)\, b(s)$$

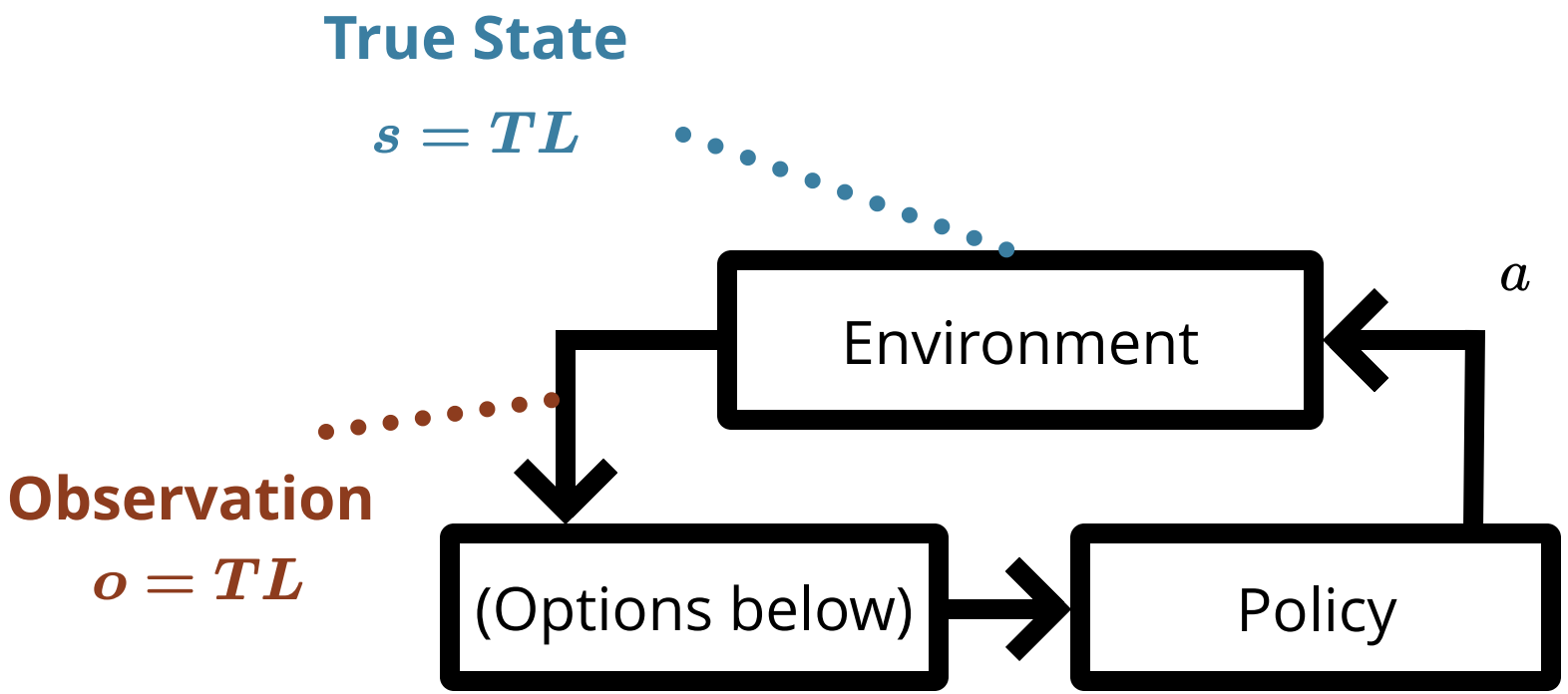


Belief Dynamics

$$b'(s') \propto Z(o \mid a, s') \sum_{s} T(s' \mid s, a)\, b(s)$$



POMDP Sense-Plan-Act Loop



Guiding Quesiton

How do we calculate the optimal action in a POMDP?

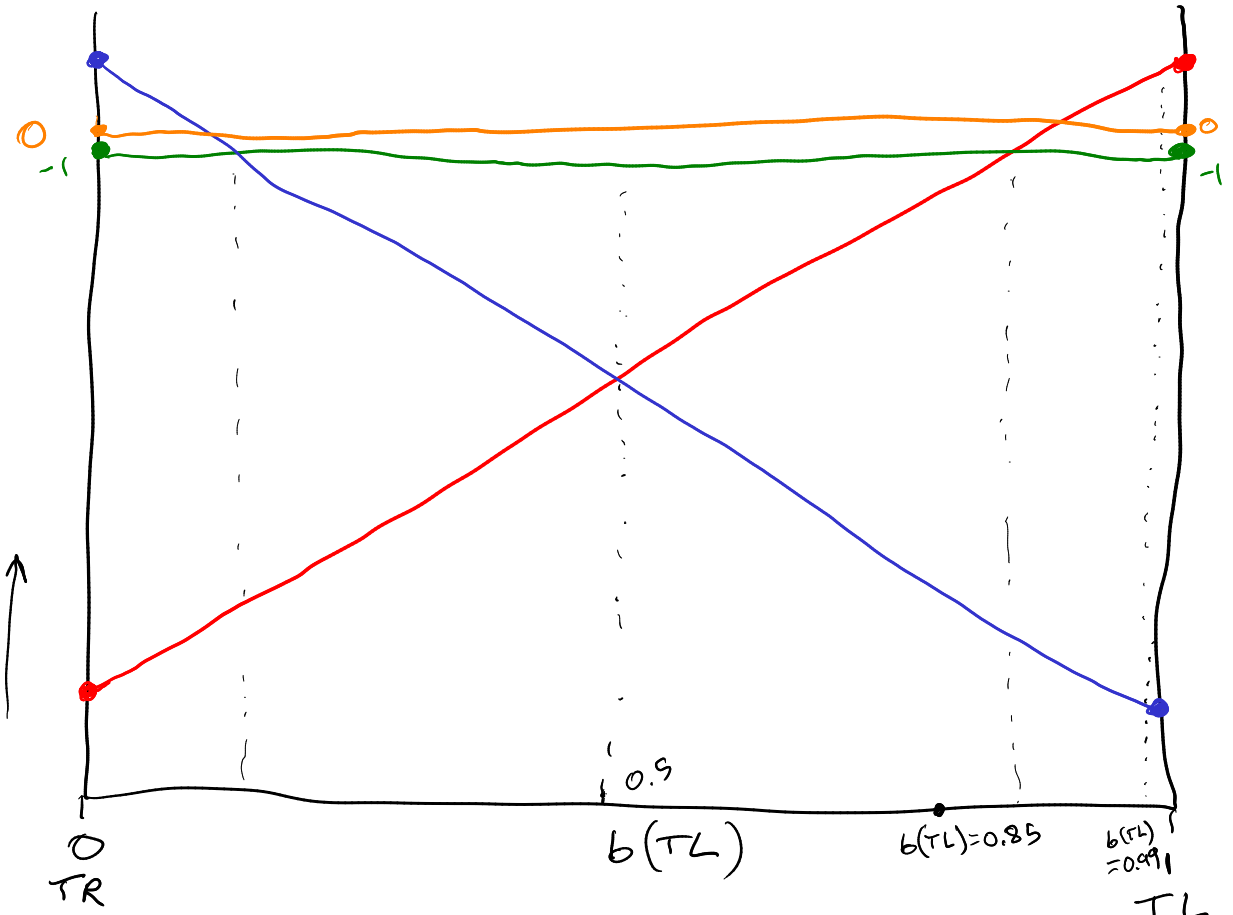
One-step utility

Reward
+10 open empty door
−1 Listen
−100 open tiger door
0 donothing

$a = L$
$R(TR, L) = -1$
$R(TL, L) = -1$

$a = OL$
$R(TR, OL) = +10$
$R(TL, OL) = -100$

$a = OR$
$R(TR, OR) = -100$
$R(TL, OR) = +10$



one-step alpha vector for action a

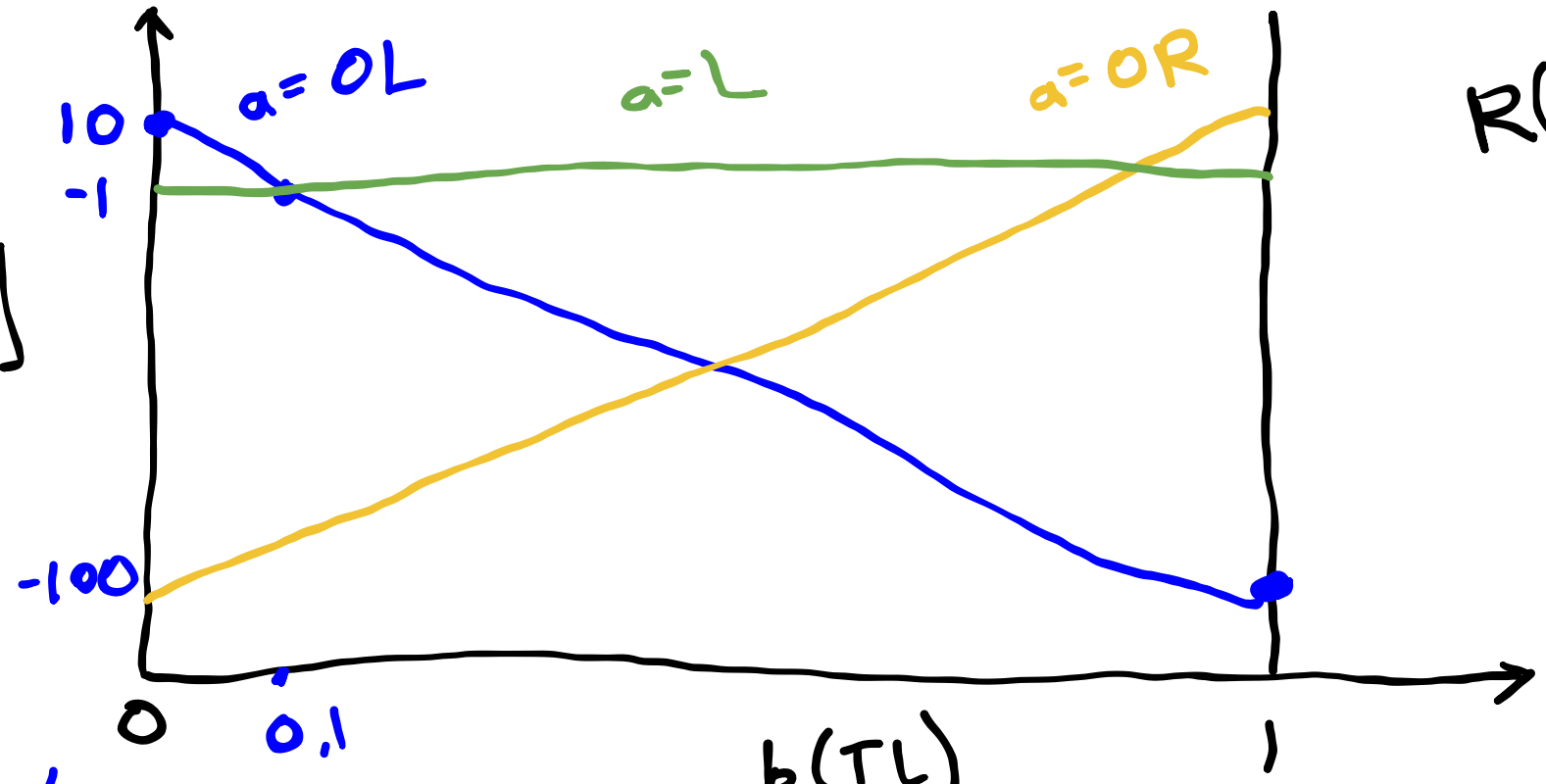
$$R(b,a) = \vec{R}^a \cdot \vec{b}$$

$$\begin{bmatrix} R(s^1,a) \\ R(s^2,a) \\ \vdots \end{bmatrix} \qquad \begin{bmatrix} b(s^1) \\ b(s^2) \\ \vdots \end{bmatrix}$$

$$R(b,a) = \underset{s\sim b}{\mathbb{E}}[R(s,a)] = \sum_s b(s) R(s,a) = b(TL)\, R(TL,a) + (1 - b(TL))\, R(TR,a)$$

One-step utility

Reward: +10 empty door
-1 Listen
-100 Tiger



$R(b,a) = E_{s\sim b}[R(s,a)]$

$R(b,a) = \vec{r}_a \cdot b$

α-vector

$R(b,a) = 10 \cdot (1 - b(TL)) - 100\, b(TL) = -110\, b + 10$

Exercise 2: Crying Baby 1-Step Utility

$S = \{h, \neg h\}$

$A = \{f, \neg f\}$

$O = \{c, \neg c\}$

$T(h \mid h, \neg f) = 1.0$

$T(h \mid \neg h, \neg f) = 0.1$

$T(\neg h \mid \cdot, f) = 1.0$

$$R(s, a) = R(s) + R(a)$$

$$R(s) = \begin{cases} -10 \text{ if } s = h \\ 0 \text{ otherwise} \end{cases}$$

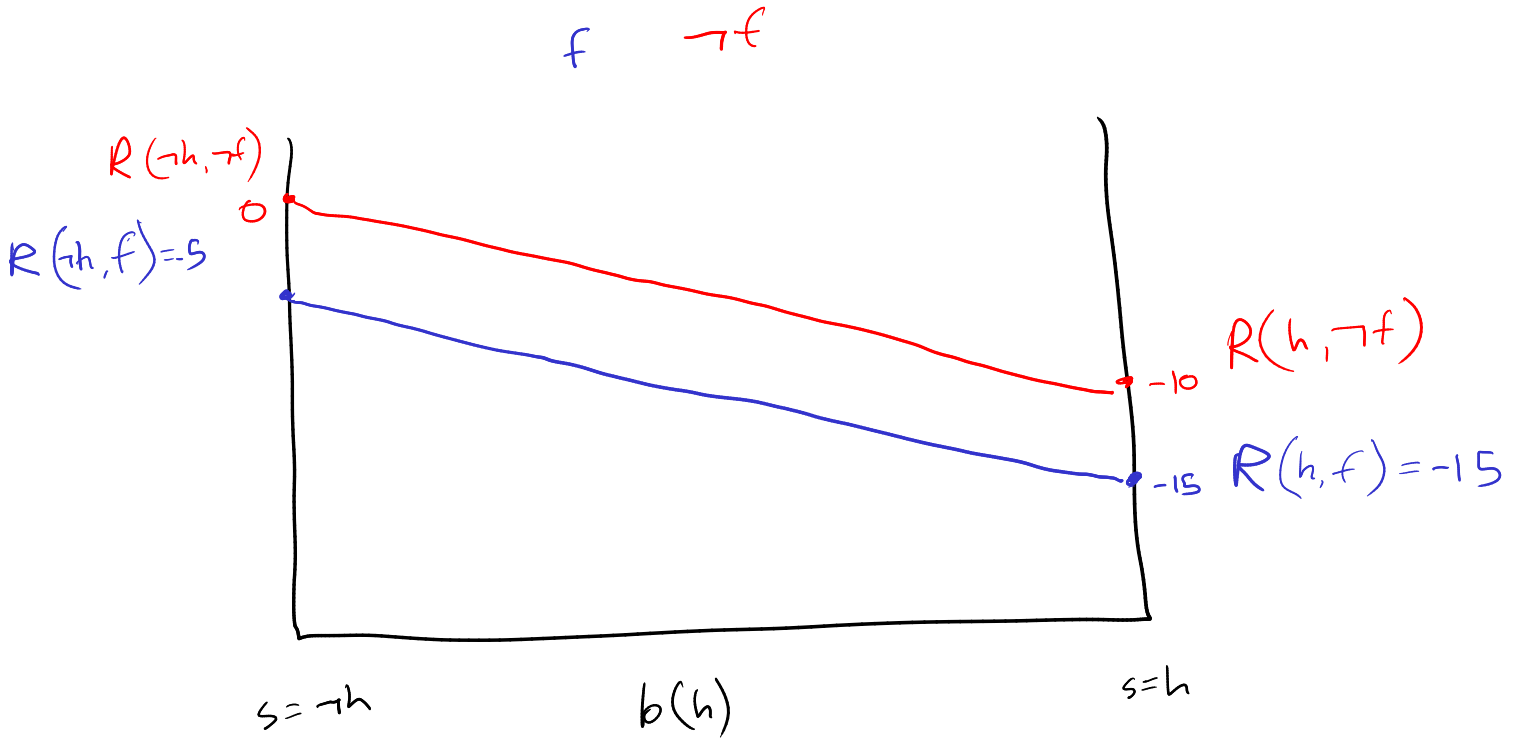
$$R(a) = \begin{cases} -5 \text{ if } a = f \\ 0 \text{ otherwise} \end{cases}$$

$$Z(c \mid \cdot, h) = 0.8)$$

$$Z(c \mid \cdot, \neg h) = 0.1$$

$$\gamma = 0.9$$

Draw the 1-step utility α-vectors for the Crying Baby problem.



Alpha Vectors for Conditional Plans

Alpha Vectors for Conditional Plans

Conditional Plans: fixed-depth history-based policies

Alpha Vectors for Conditional Plans

Conditional Plans: fixed-depth history-based policies

1 Step:

Alpha Vectors for Conditional Plans

Conditional Plans: fixed-depth history-based policies

1 Step: 

Alpha Vectors for Conditional Plans

Conditional Plans: fixed-depth history-based policies

1 Step: 

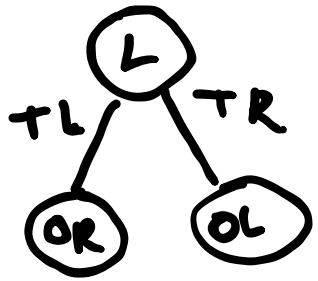
2 Step:

Alpha Vectors for Conditional Plans

Conditional Plans: fixed-depth history-based policies

1 Step:

2 Step:

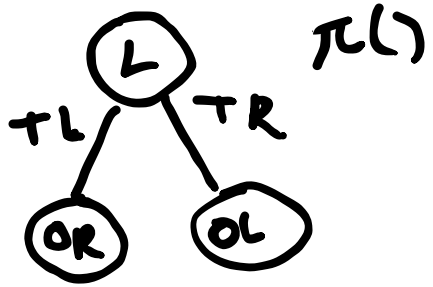


Alpha Vectors for Conditional Plans

Conditional Plans: fixed-depth history-based policies

1 Step:

2 Step:

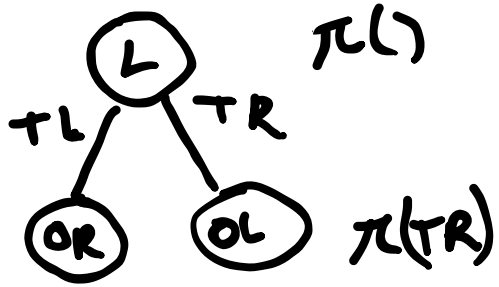


Alpha Vectors for Conditional Plans

Conditional Plans: fixed-depth history-based policies

1 Step:

2 Step:

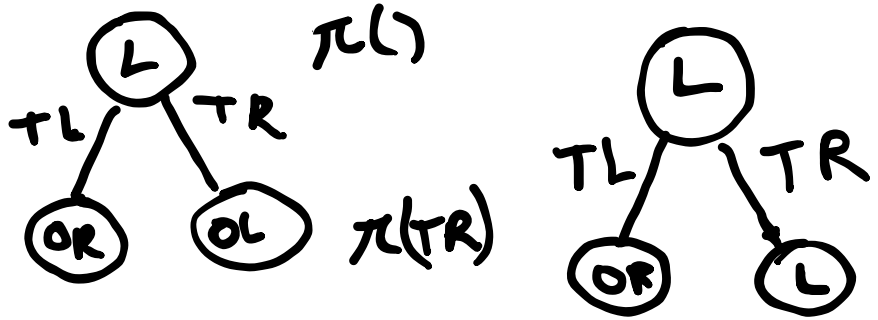


Alpha Vectors for Conditional Plans

Conditional Plans: fixed-depth history-based policies

1 Step:

2 Step:

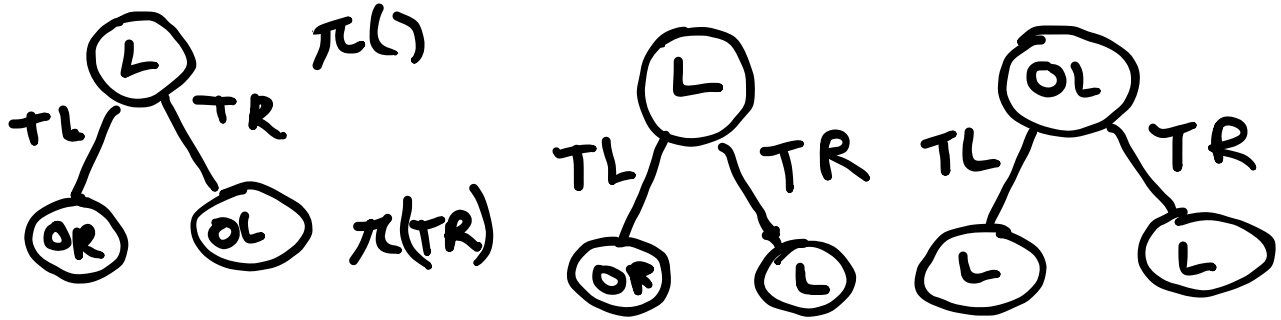


Alpha Vectors for Conditional Plans

Conditional Plans: fixed-depth history-based policies

1 Step:

2 Step:

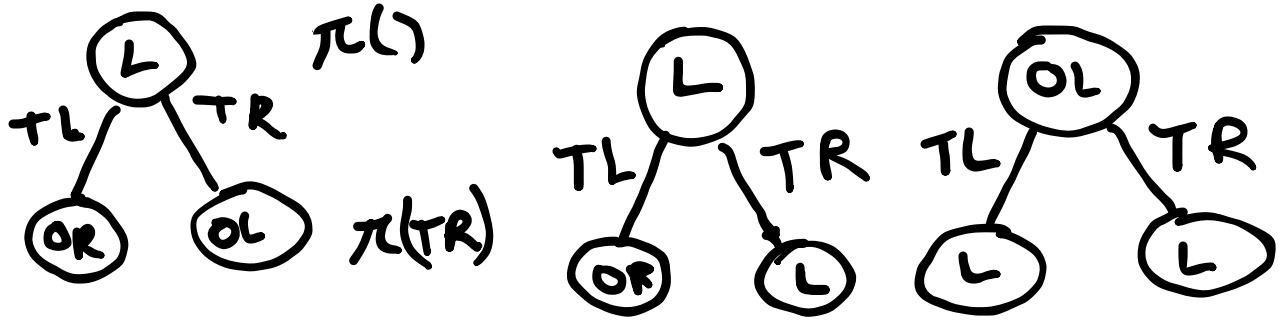


Alpha Vectors for Conditional Plans

Conditional Plans: fixed-depth history-based policies

1 Step:

2 Step:



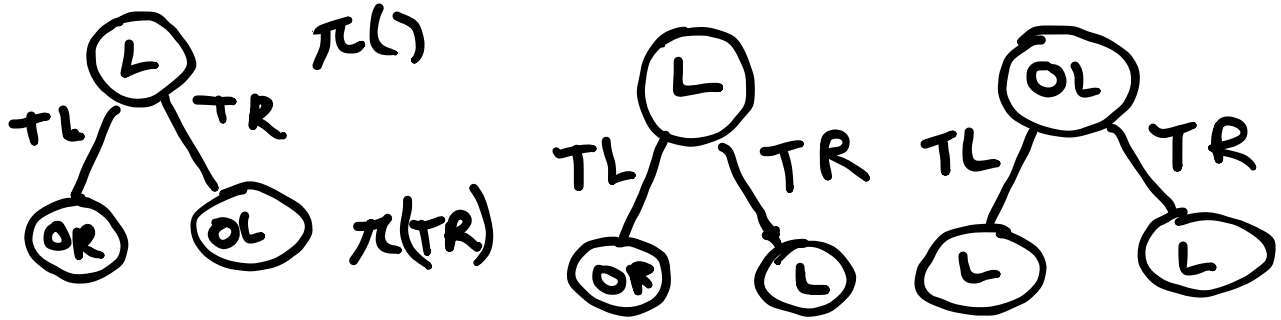
$$|A|^{\frac{(|O|^h-1)}{(|O|-1)}}$$

Alpha Vectors for Conditional Plans

Conditional Plans: fixed-depth history-based policies

1 Step:

2 Step:



$|A|^{\frac{(|O|^h-1)}{(|O|-1)}}$

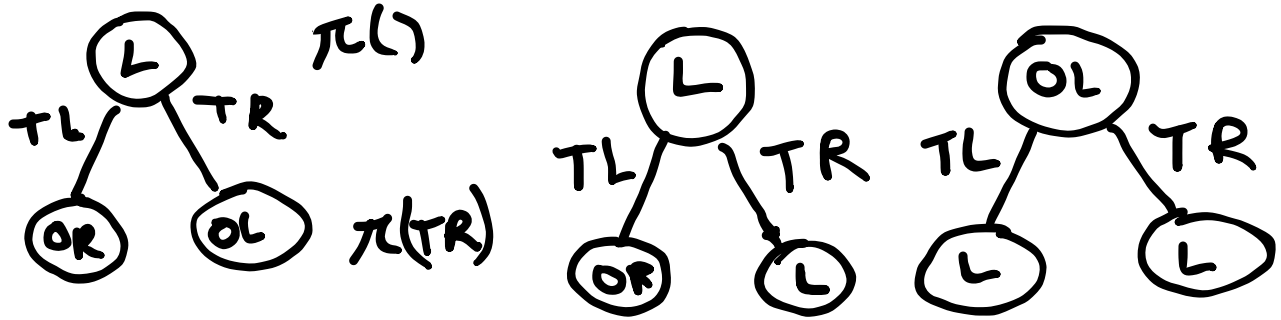
27 two step plans!

Alpha Vectors for Conditional Plans

Conditional Plans: fixed-depth history-based policies

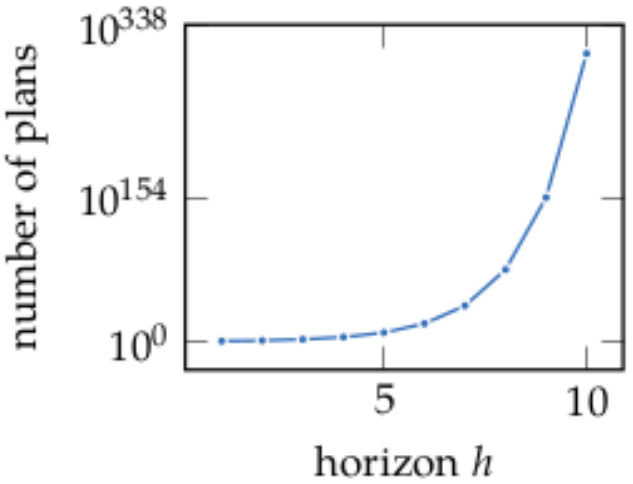
1 Step:

2 Step:



$$|A|^{\frac{(|O|^h-1)}{(|O|-1)}}$$

27 two step plans!



Alpha Vectors for Conditional Plans

Alpha Vectors for Conditional Plans

For 1-step: $U^{\pi}(s) = R(s, \pi())$

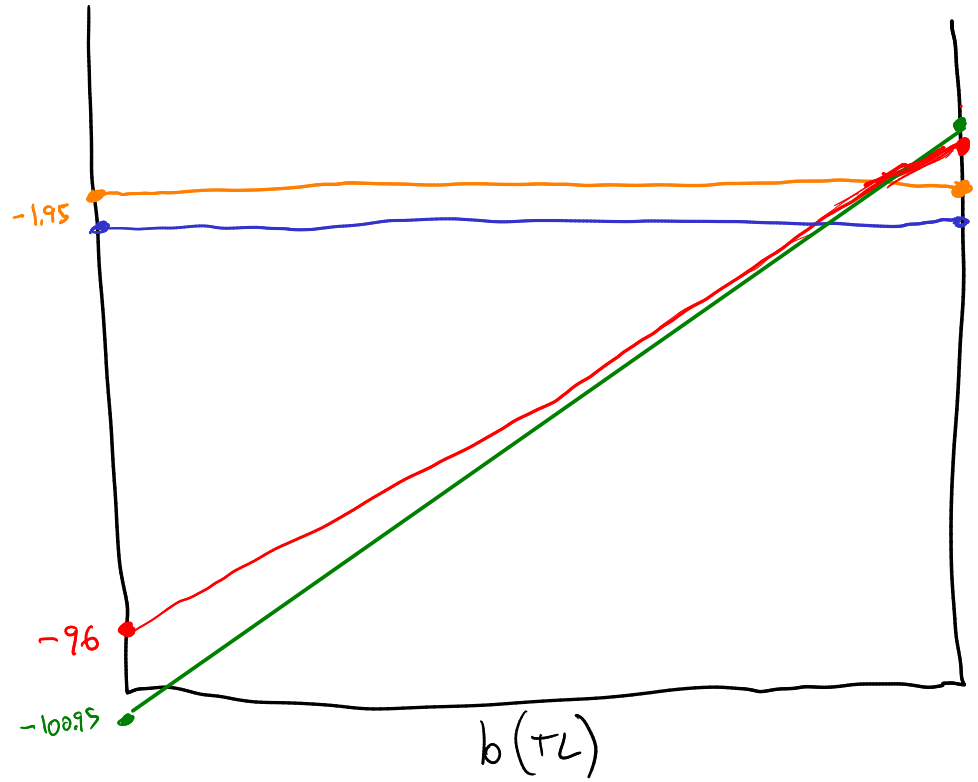
Alpha Vectors for Conditional Plans

For 1-step: $U^{\pi()}(s) = R(s, \pi())$

any-length plans

$$U^{\pi()}(s) = R(s, \pi()) + \gamma \left[\sum_{s'} T(s' \mid s, \pi()) \sum_{o} O(o \mid \pi(), s') U^{\pi(o)}(s') \right]$$

$s \rightarrow s'$



L

$U^{L}(\cdot) = -1$

OL

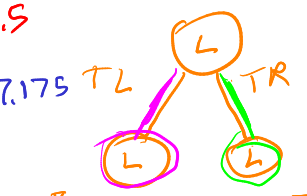
$U^{OL}(TR) = 10$

$U^{OL}(TL) = -100$

OR

$U^{OR}(TR) = -100$

$U^{OR}(TL) = 10$

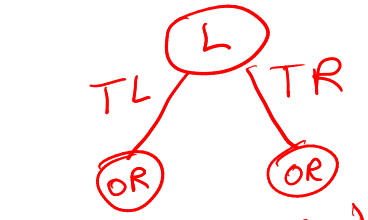


$U^{\pi}(TL) = -1 + \gamma \left[1 \left(0.85(-1) + 0.15(-1) \right) \right]$

T

$= -1 + \gamma(-1)$

$\gamma = 0.95$ $= -1.95$

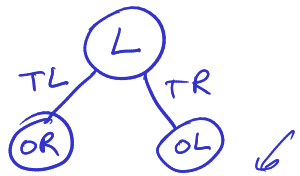


$U^{\pi}(TL) = -1 + \gamma(10)$

$= 8.5$

$U^{\pi}(TR) = -1 + \gamma(-100)$

$= -96$

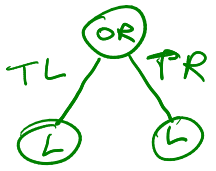


$Z(TL|L, TL)$

$U^{\pi}(TL) = -1 + \gamma(1 \cdot (0.85 \cdot 10 + 0.15(-100)))$

$= -7.175$

$U^{\pi}(TR) = -7.175$

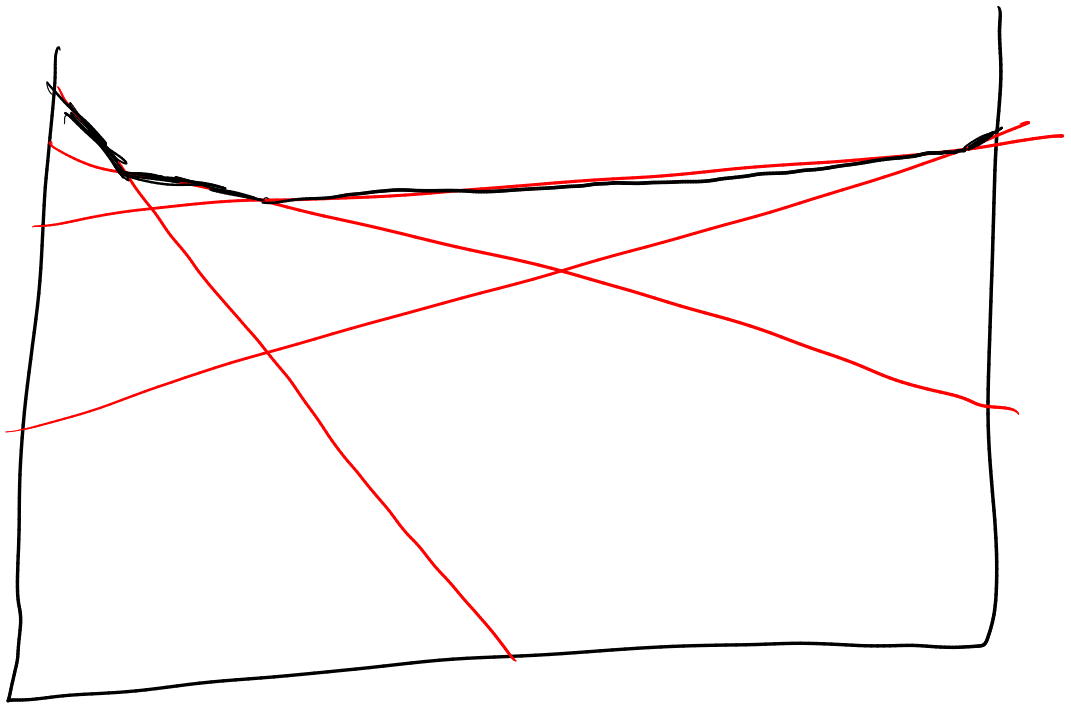


$U^{\pi}(TL) = 9.5$

$U^{\pi}(TR) = -100.95$

POMDP Value Functions

POMDP Value Functions

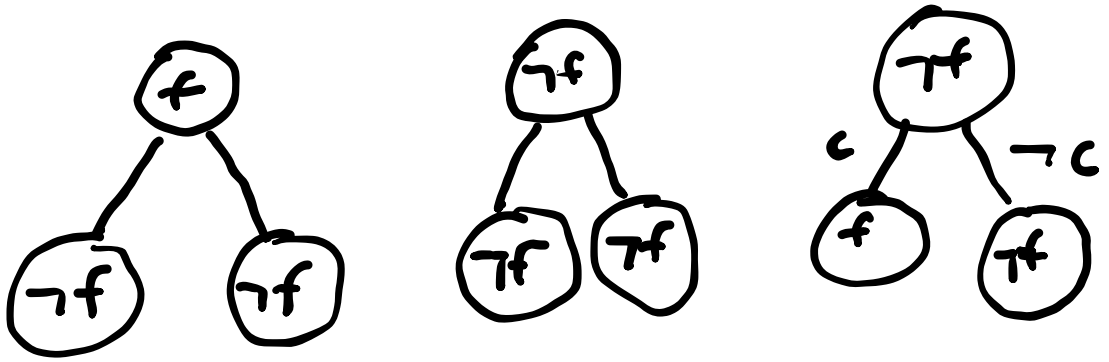


$$V^*(b) = \max_{\alpha \in \Gamma} \alpha^\top b$$

Exercise: 2-Step Crying Baby α Vectors

$$S = \{h, \neg h\}$$
$$A = \{f, \neg f\}$$
$$O = \{c, \neg c\}$$

$$T(h \mid h, \neg f) = 1.0$$
$$T(h \mid \neg h, \neg f) = 0.1$$
$$T(\neg h \mid \cdot, f) = 1.0$$



$$R(s, a) = R(s) + R(a)$$

$$R(s) = \begin{cases} -10 \text{ if } s = h \\ 0 \text{ otherwise} \end{cases}$$

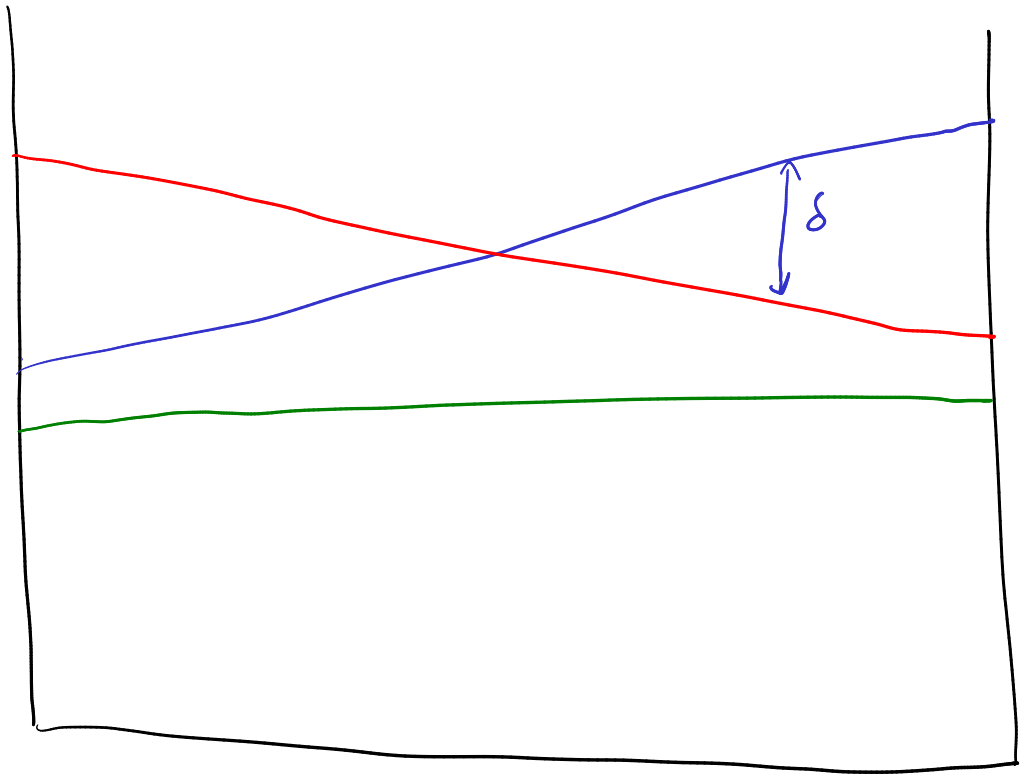
$$R(a) = \begin{cases} -5 \text{ if } a = f \\ 0 \text{ otherwise} \end{cases}$$

$$Z(c \mid \cdot, h) = 0.8)$$
$$Z(c \mid \cdot, \neg h) = 0.1$$

$$\gamma = 0.9$$

$$U^{\pi}(s) = R(s, \pi()) + \gamma \left[\sum_{s'} T(s' \mid s, \pi()) \sum_{o} O(o \mid \pi(), s') U^{\pi(o)}(s') \right]$$

α-Vector Pruning



$$\underset{\delta, b}{\text{maximize}} \quad \delta$$

$$\text{subject to} \quad \left.\begin{array}{l} b \geq 0 \\ \mathbf{1}^T b = 1 \end{array}\right\} \text{enforce that } b \text{ is a probability distribution}$$

$$\alpha^T b > \alpha'^T b + \delta \quad \forall \alpha' \in \Gamma$$

- If there is a positive δ^* solution the α is not dominated
- b^* is sometimes called "witness"

Alpha Vector Expansion

POMDP Value Iteration (horizon d)

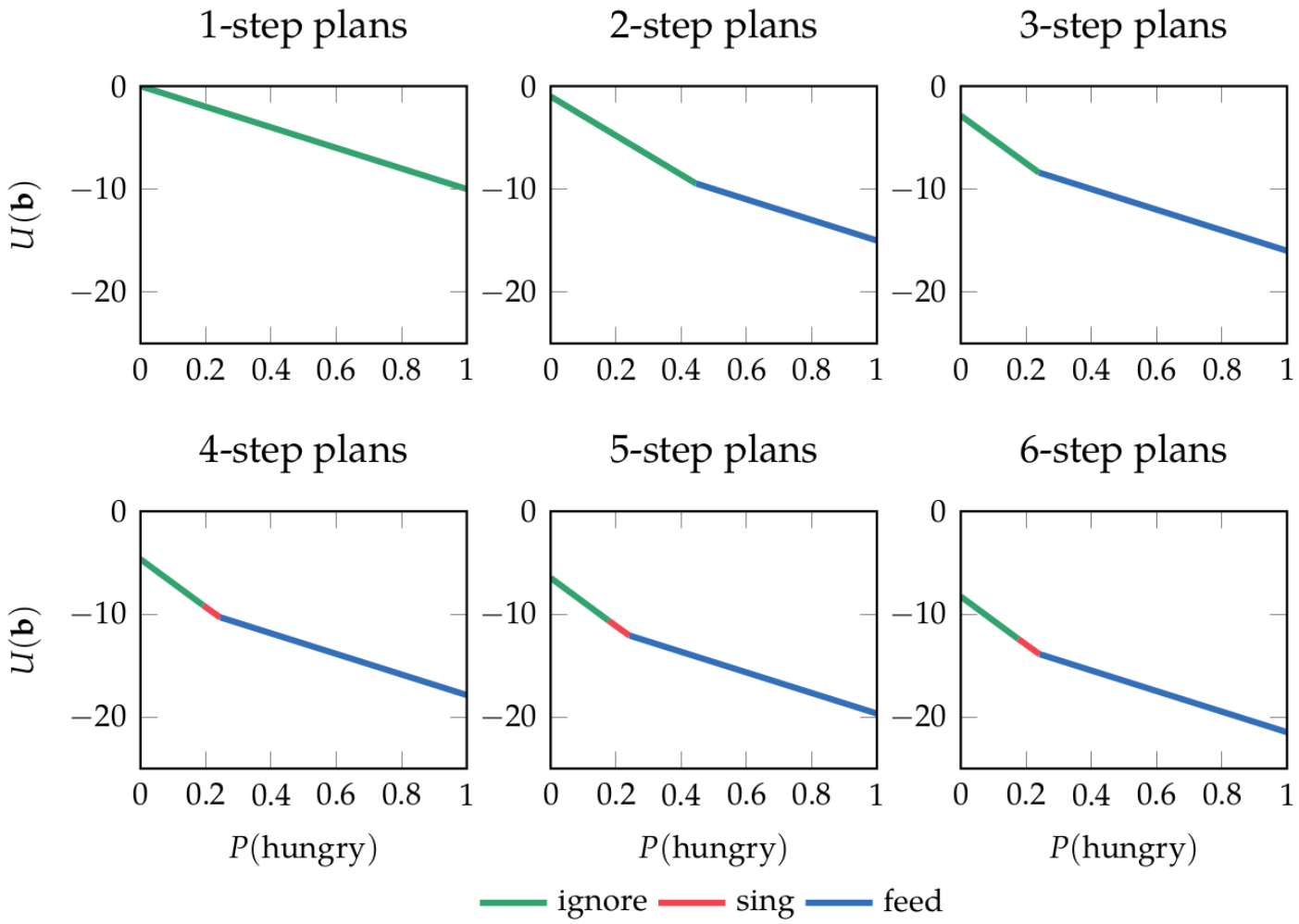
$\Gamma^0 \leftarrow \emptyset$

for $n \in 1 \ldots d$

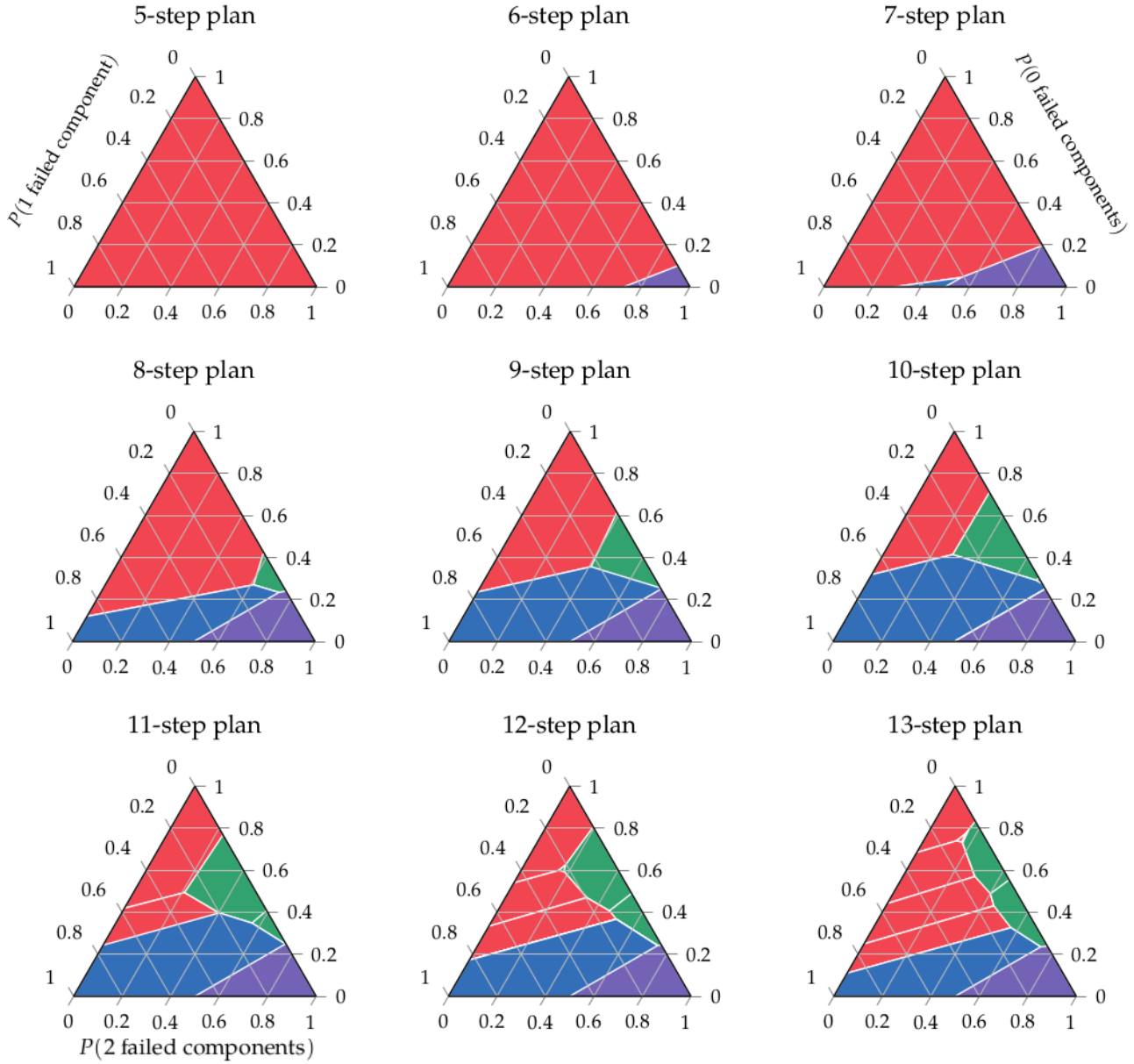
  Construct Γ^n by expanding with Γ^{n-1}

  Prune Γ^n

Finite Horizon POMDP Value Iteration



Finite Horizon POMDP Value Iteration



Recap

Recap

- A POMDP is an MDP on the _____

Recap

- A POMDP is an MDP on the belief space

Recap

- A POMDP is an MDP on the belief space
- The value function of a discrete POMDP can be represented by a set of _____

Recap

- A POMDP is an MDP on the belief space
- The value function of a discrete POMDP can be represented by a set of α-vectors

Recap

- A POMDP is an MDP on the belief space
- The value function of a discrete POMDP can be represented by a set of α-vectors
- Each α vector corresponds to a _____

Recap

- A POMDP is an MDP on the belief space
- The value function of a discrete POMDP can be represented by a set of α-vectors
- Each α vector corresponds to a conditional plan