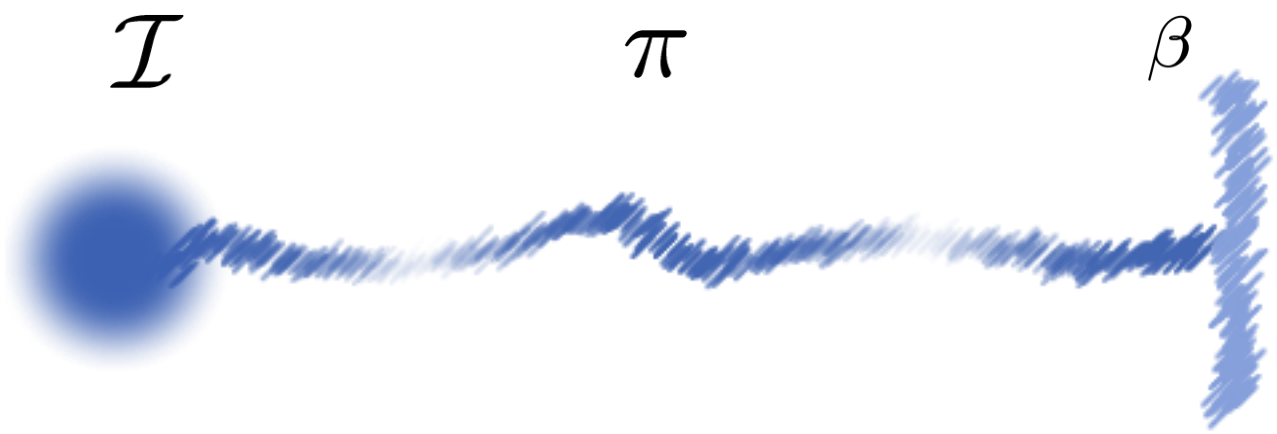
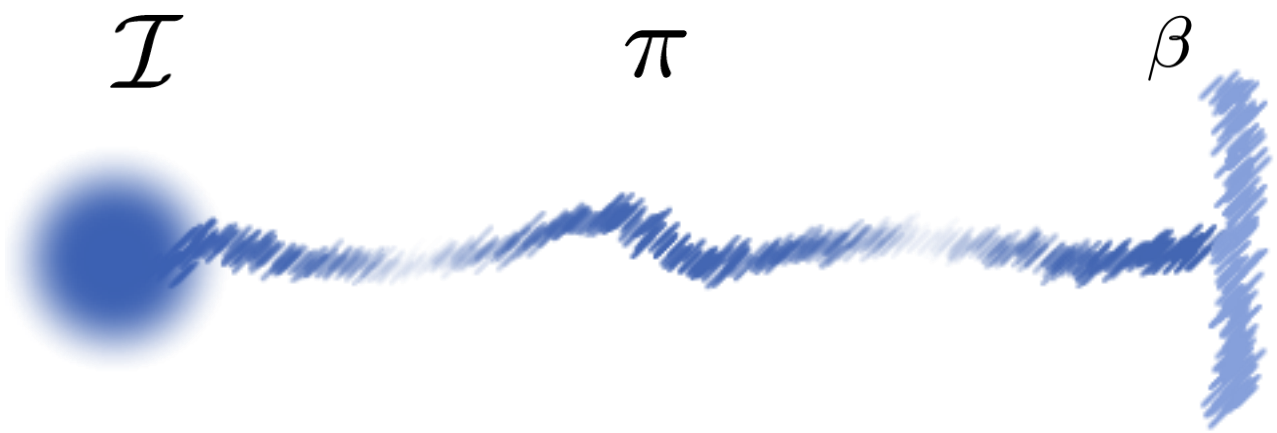
Policy Switching Towards Transferring Behavioural Knowledge

Gheorghe Comanici

Options



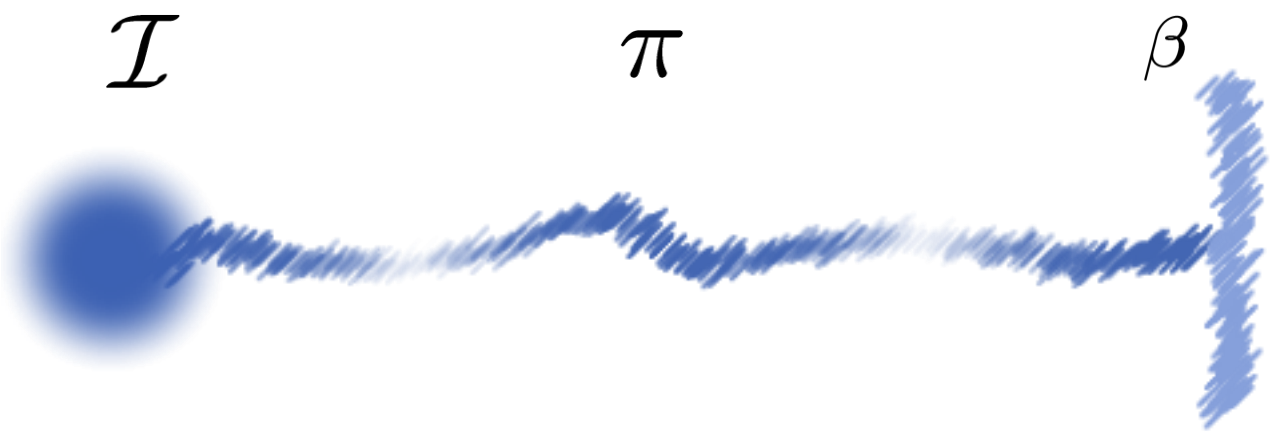
Options



Keywords : start, use, stop

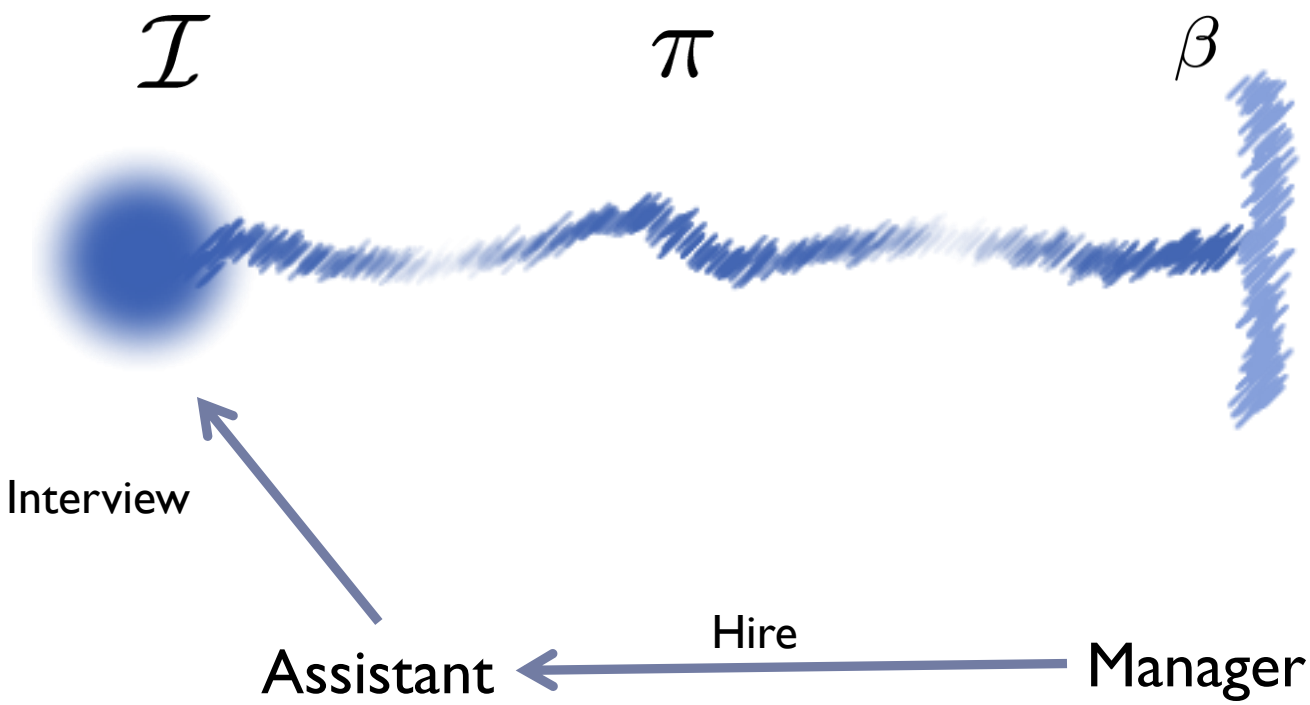
Options

- Manager hires an assistant



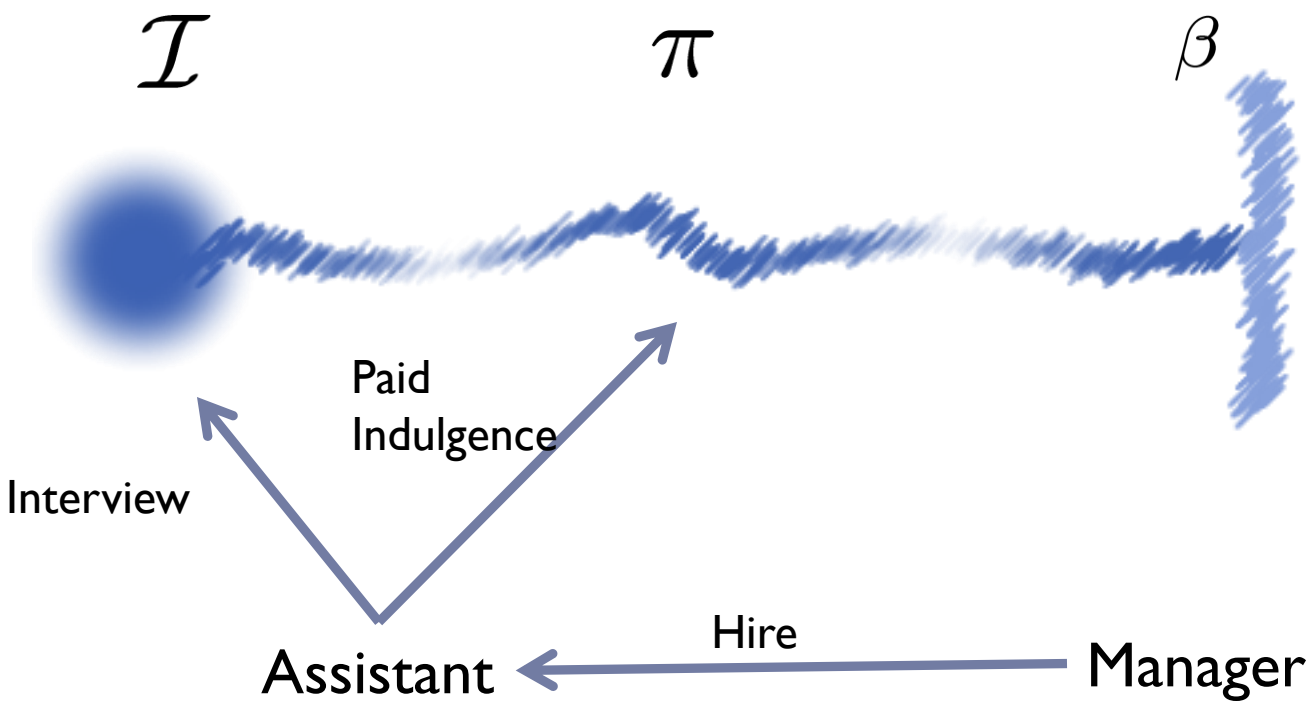
Options

- Manager hires an assistant



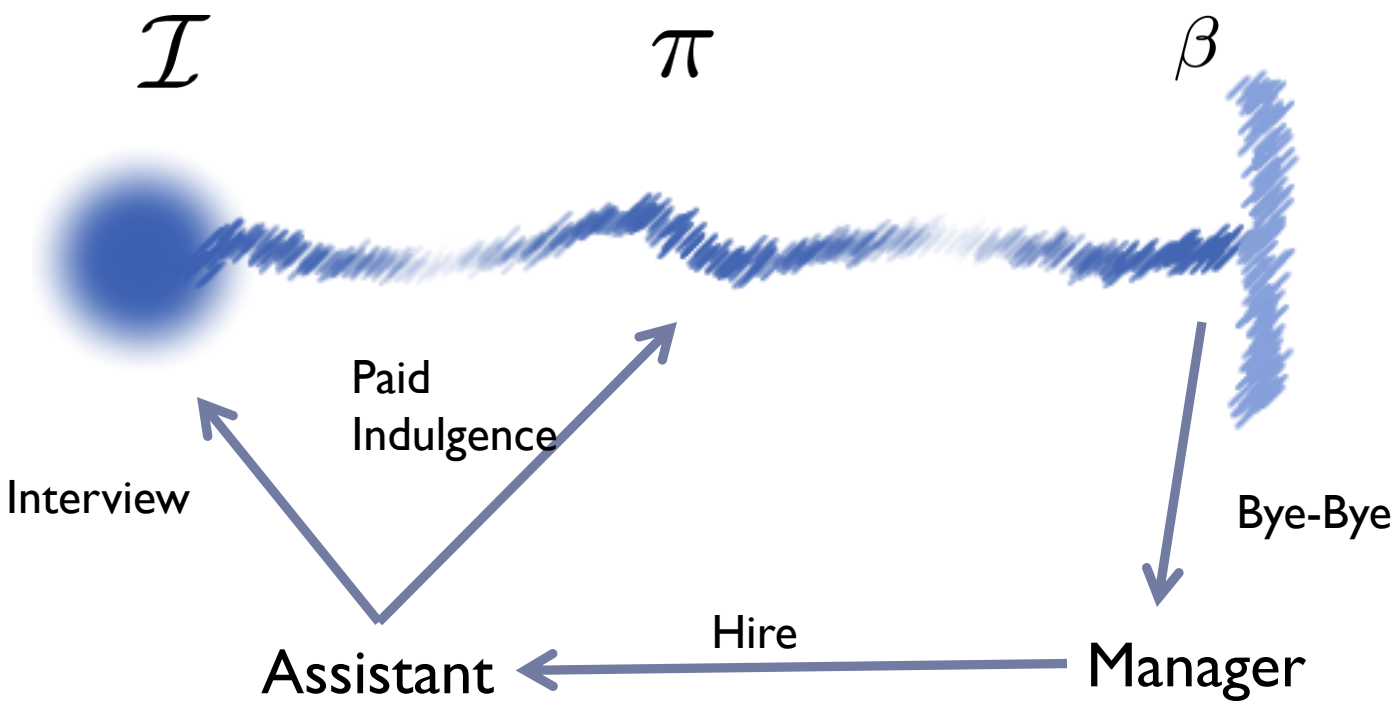
Options

- Manager hires an assistant



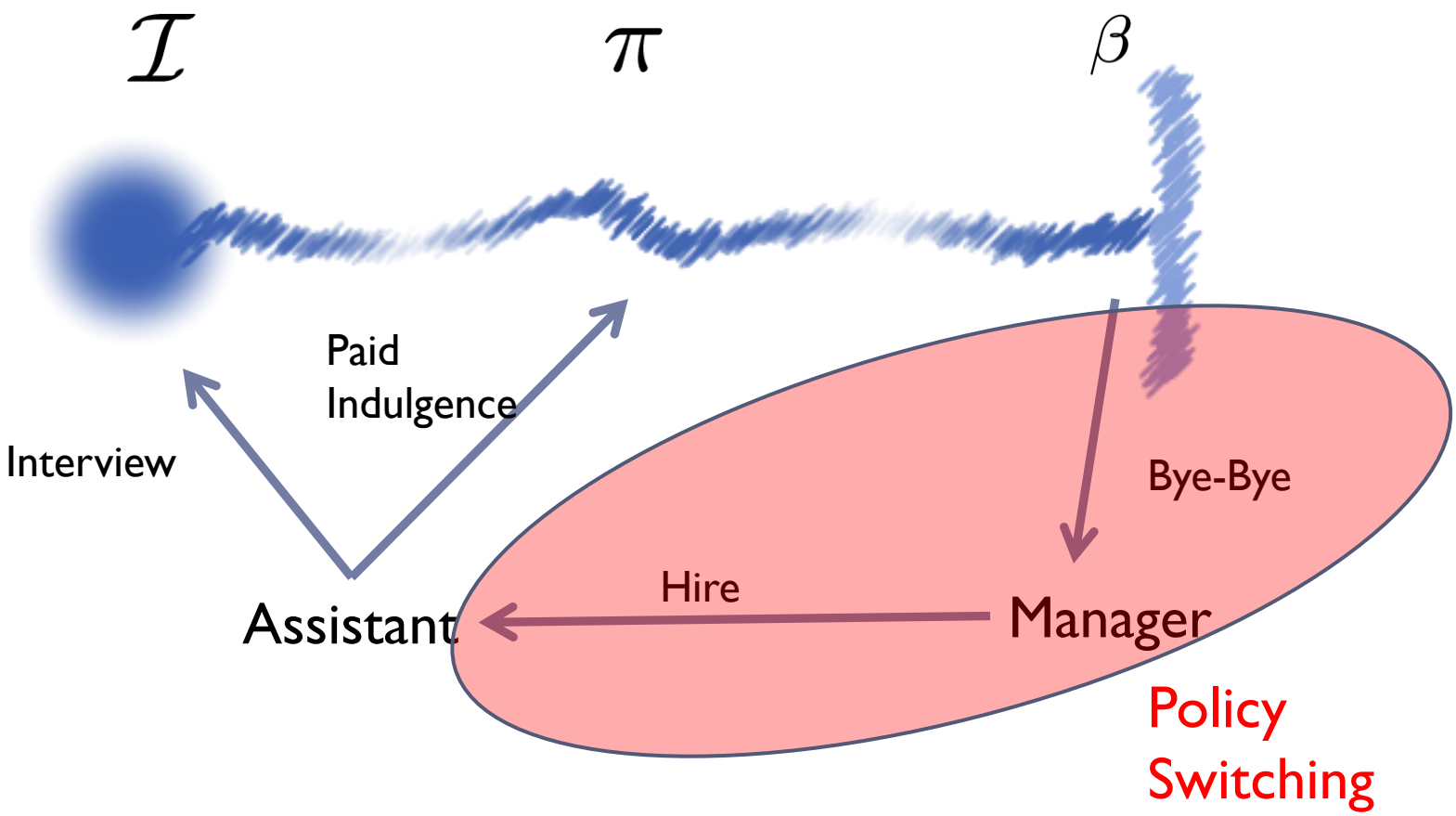
Options

- Manager hires an assistant



Options

- Manager hires an assistant



Options

- Why would you ever use options?

Options

- Why would you ever use options?
  - Why would you ever get an assistant?

Options

- Why would you ever use options?
  - Why would you ever get an assistant?

- Comes with education → Prior info

Options

- Why would you ever use options?
  - Why would you ever get an assistant?

- Comes with education → Prior info
- You're the boss! → Specific design and subgoals

Options

- Why would you ever use options?
  - Why would you ever get an assistant?

- Comes with education → Prior info
- You're the boss! → Specific design and subgoals
- You get more coffee breaks → Computational efficiency

Termination model

- Usual model : flip a biased coin after every step

$$\beta : (S \times A \times R)^* \rightarrow [0, 1]$$

- Proposed model : increased dissatisfaction

$$\tau : (S \times A \times R)^* \rightarrow [0, 1]$$

  - with monotonicity on the set of histories
  - 0 for empty histories

Termination model

- The two formulations are equivalent
  - Given one ...

$$\beta(h) = \frac{\tau(h) - \tau(h_-)}{1 - \tau(h_-)}$$

  - Given the other ...

$$\tau(h) = \tau(h_-) + (1 - \tau(h_-))\beta(h)$$

Termination model

- Why use the new formulation?

Generalizing biased coins is ridiculous!

- We seek to generalize functions that preserve the structure of histories :

MONOTONICITY

Termination model

- Mathematical formulation proposed based on "bounded exponential growth"

$$\frac{dp}{dt} = mp - \frac{m}{P}p^2$$

- Termination is logistic :

$$\tau(h) = [1 + \exp(-(l(h) - \theta_0))]^{-1}$$

  - on "history length" : linear combination of positive features

$$l(h) = \sum_{s,a,r \in h} \phi_{s,a,r}\theta$$

Comanici, G. , Precup D., 2010

Policy Switching

- Combine termination and option choice
- Exploit differentiation in *policy gradient methods:*

Policy Switching

- Combine termination and option choice
- Exploit differentiation in *policy gradient methods:*

Diff. switching

Diff. termination

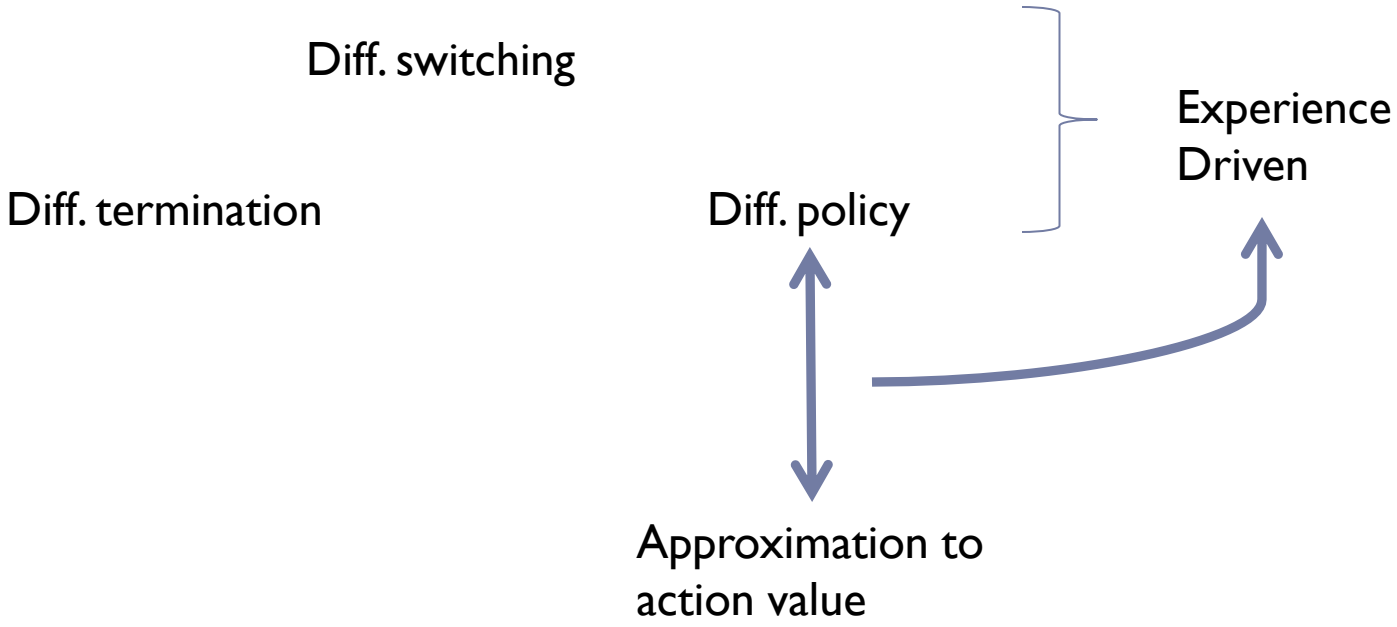
Diff. policy

Policy Switching

- Combine termination and option choice
- Exploit differentiation in *policy gradient methods:*

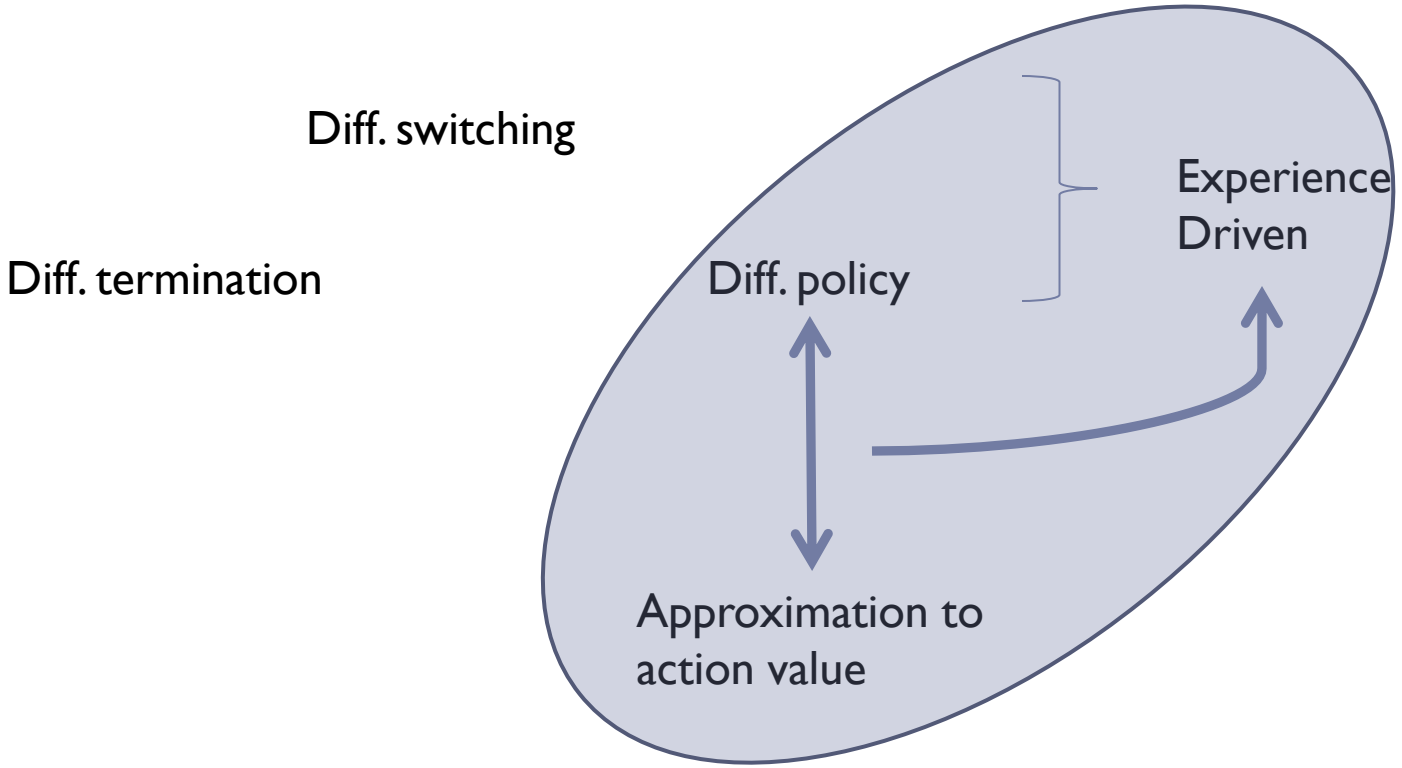
Policy Switching

- Combine termination and option choice
- Exploit differentiation in *policy gradient methods:*



Policy Switching

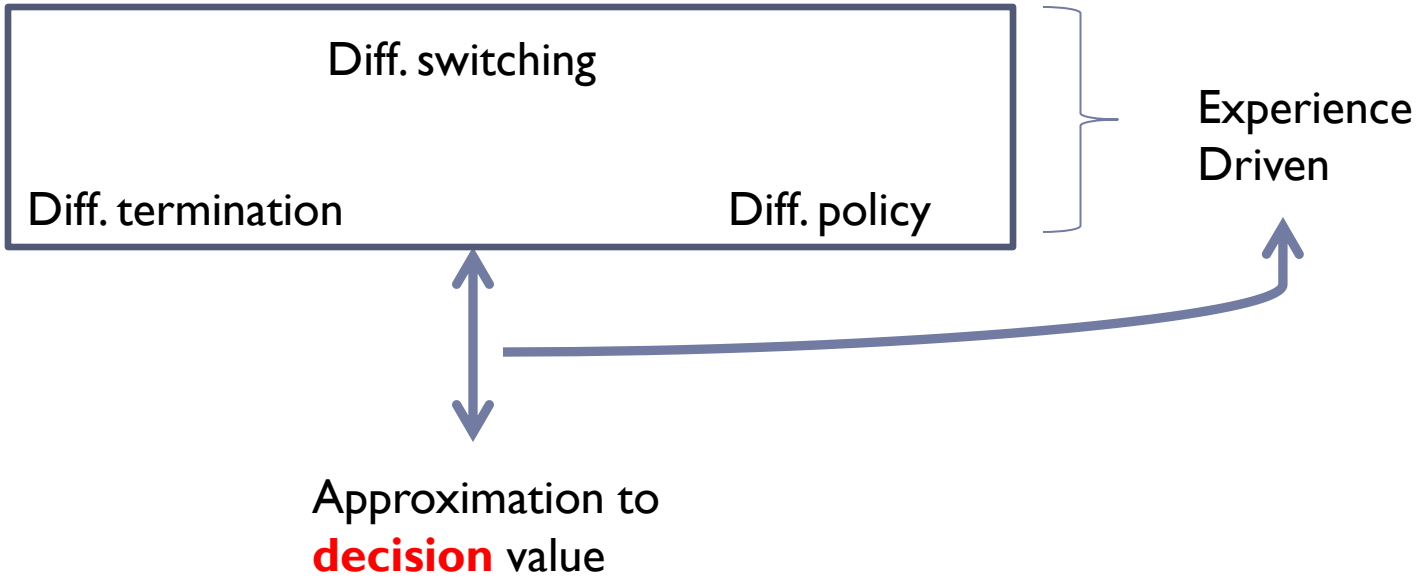
- Combine termination and option choice
- Exploit differentiation in *policy gradient methods:*



Sutton, R. , McAllester, D., Singh, S., Mansour, Y. 1999

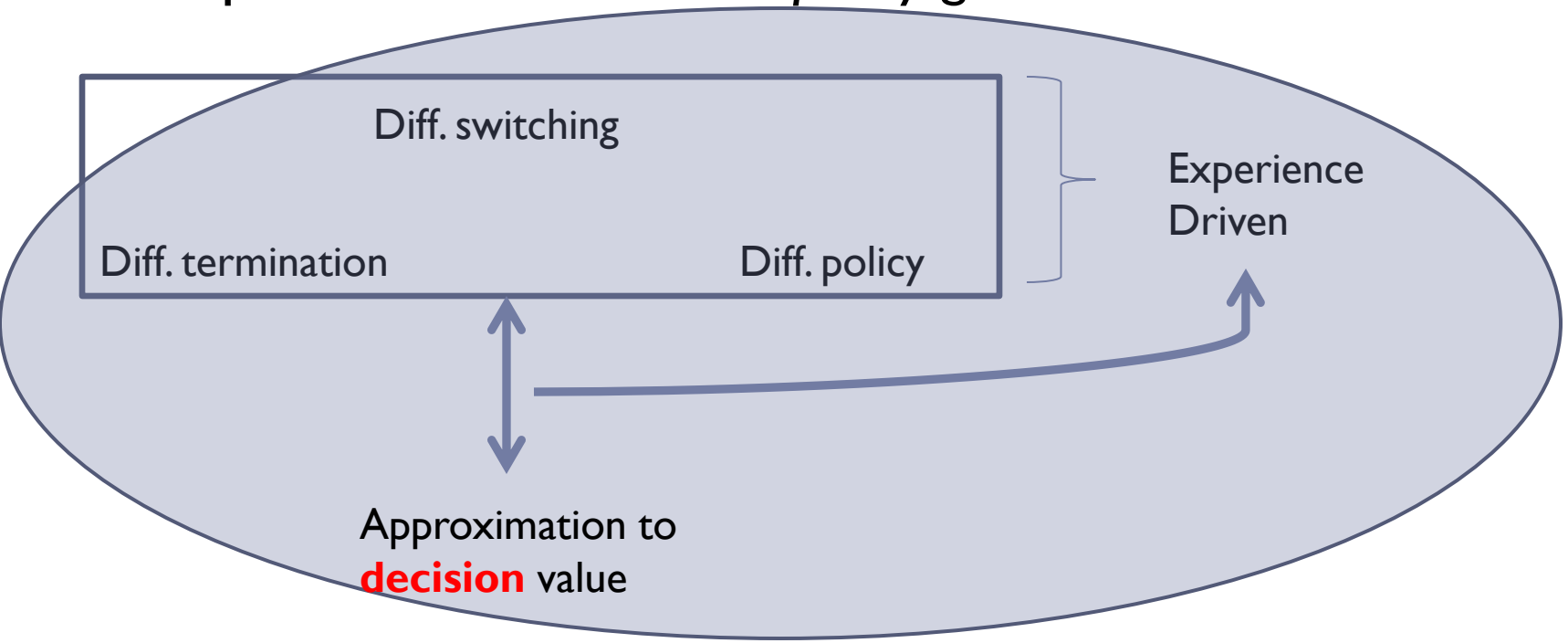
Policy Switching

- Combine termination and option choice
- Exploit differentiation in *policy gradient methods:*



Policy Switching

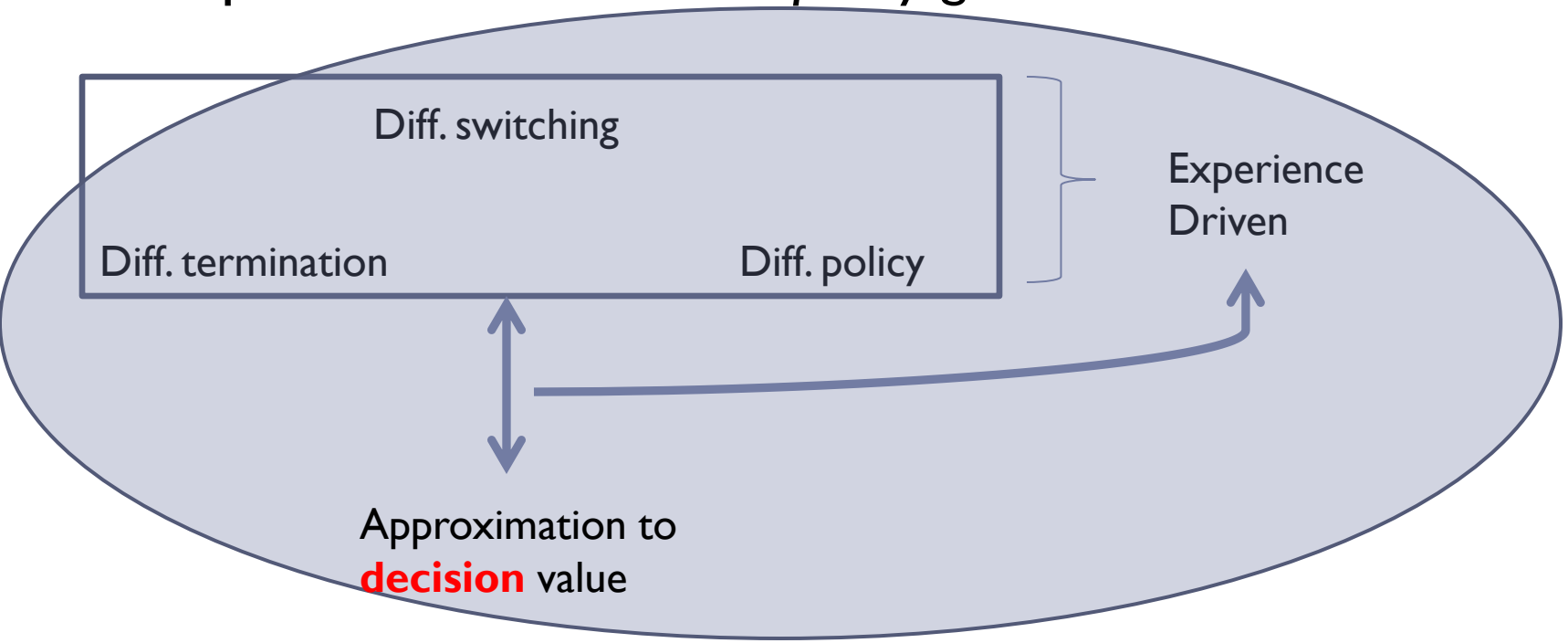
- Combine termination and option choice
- Exploit differentiation in *policy gradient methods:*



Comanici, G. , Precup D., 2010

Policy Switching

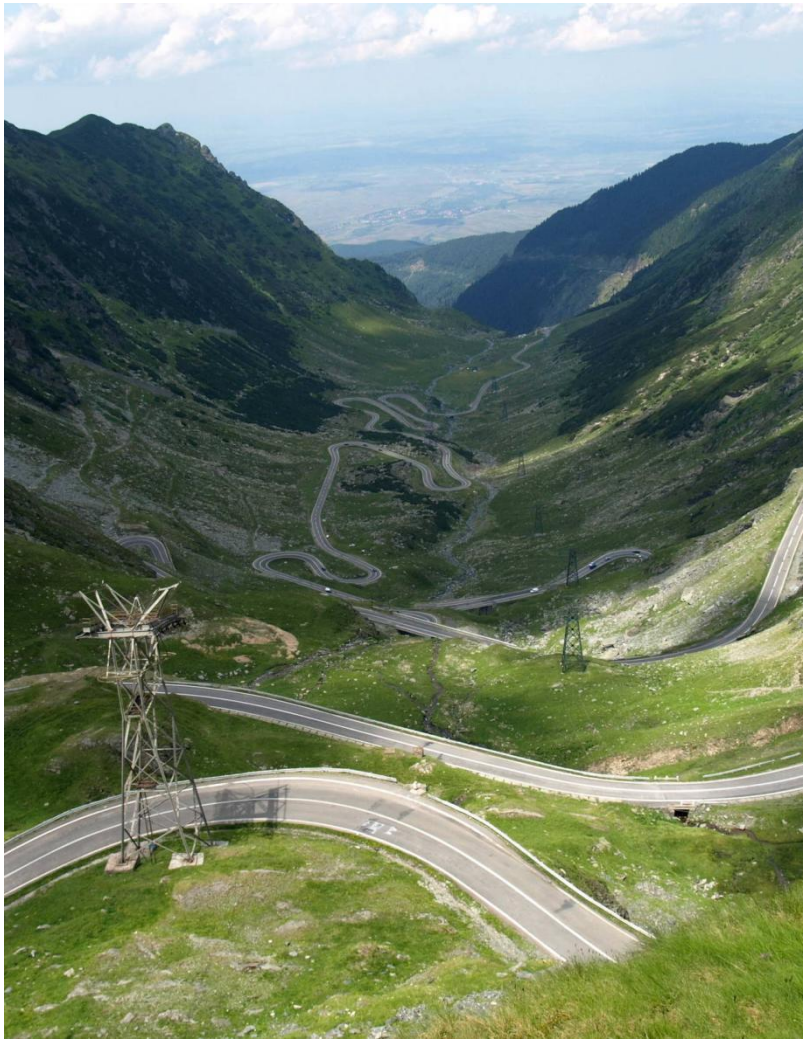
- Combine termination and option choice
- Exploit differentiation in *policy gradient methods:*



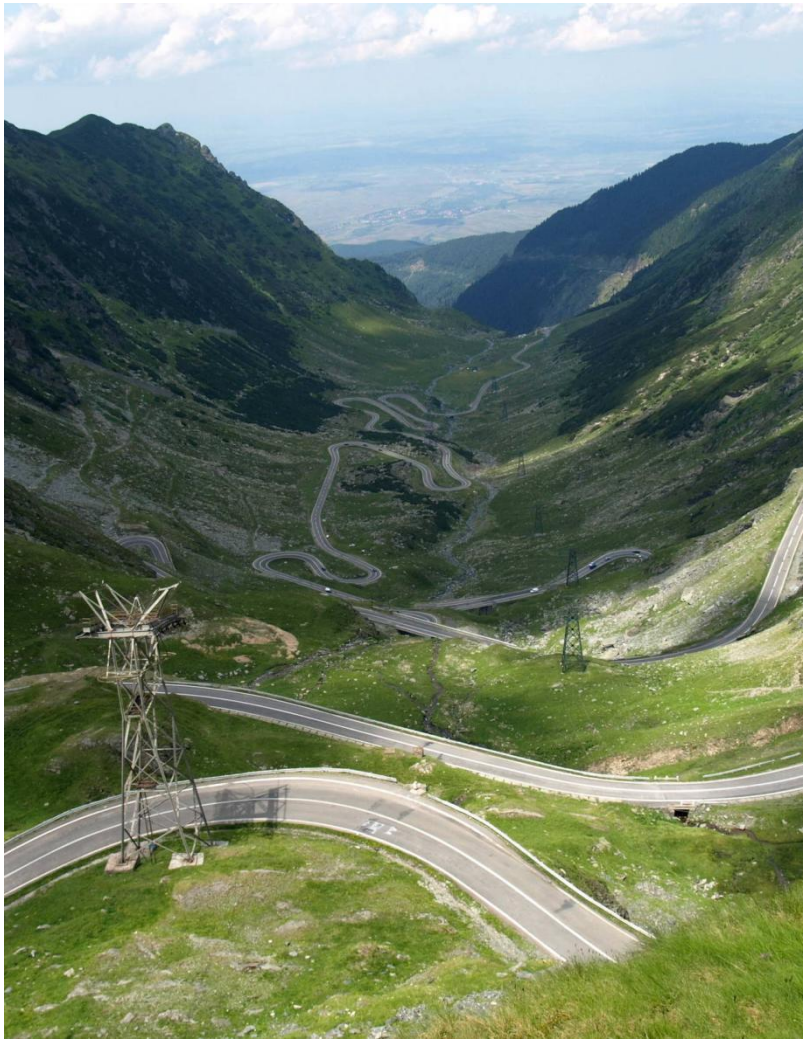
- Sufficient decisions : "use π_1" and "switch **from** π_1 **to** π_2 "

Comanici, G. , Precup D., 2010

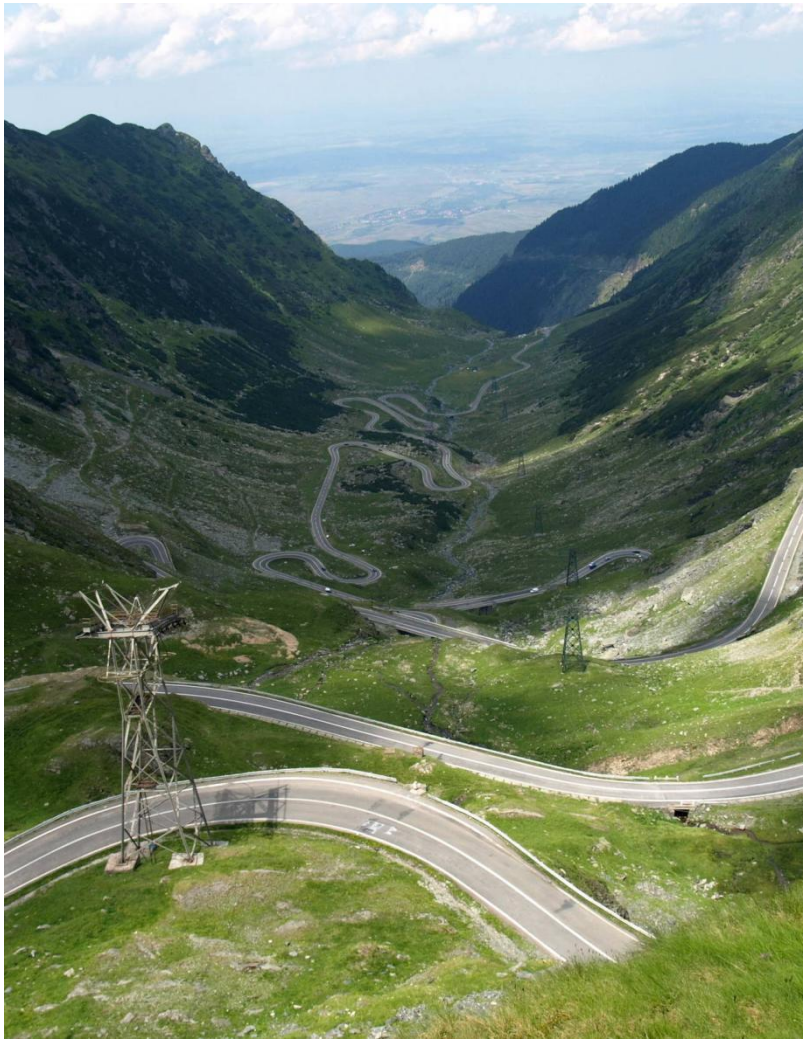
Behavioural Knowledge



Behavioural Knowledge

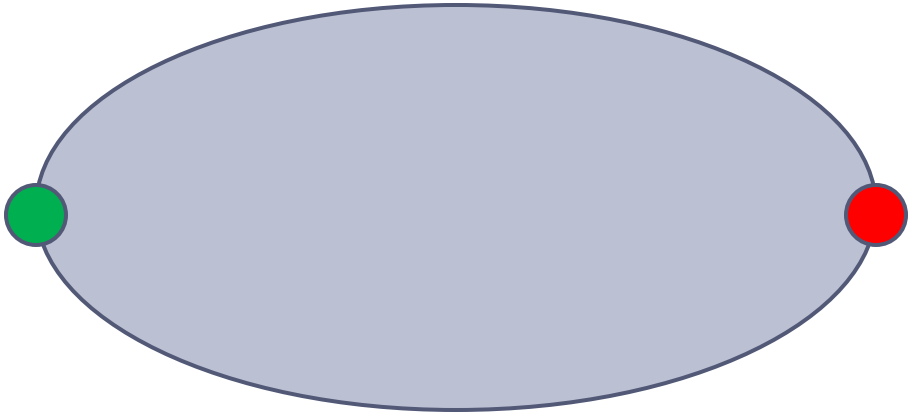


Behavioural Knowledge = Policies



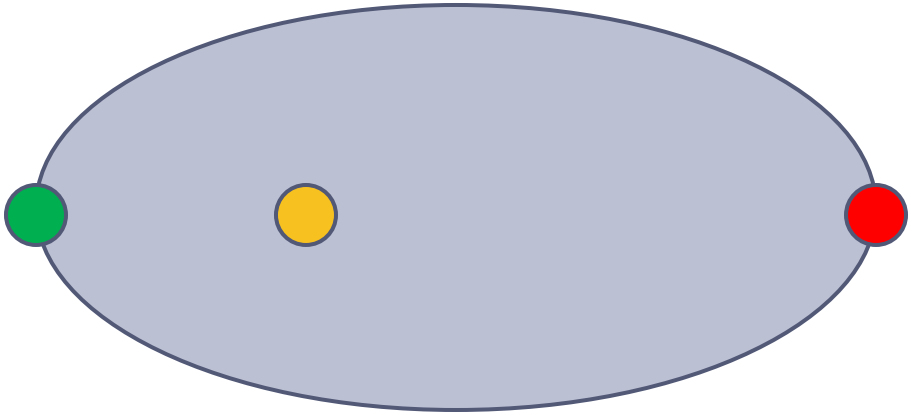
Policy transfer

- Example : subgoals



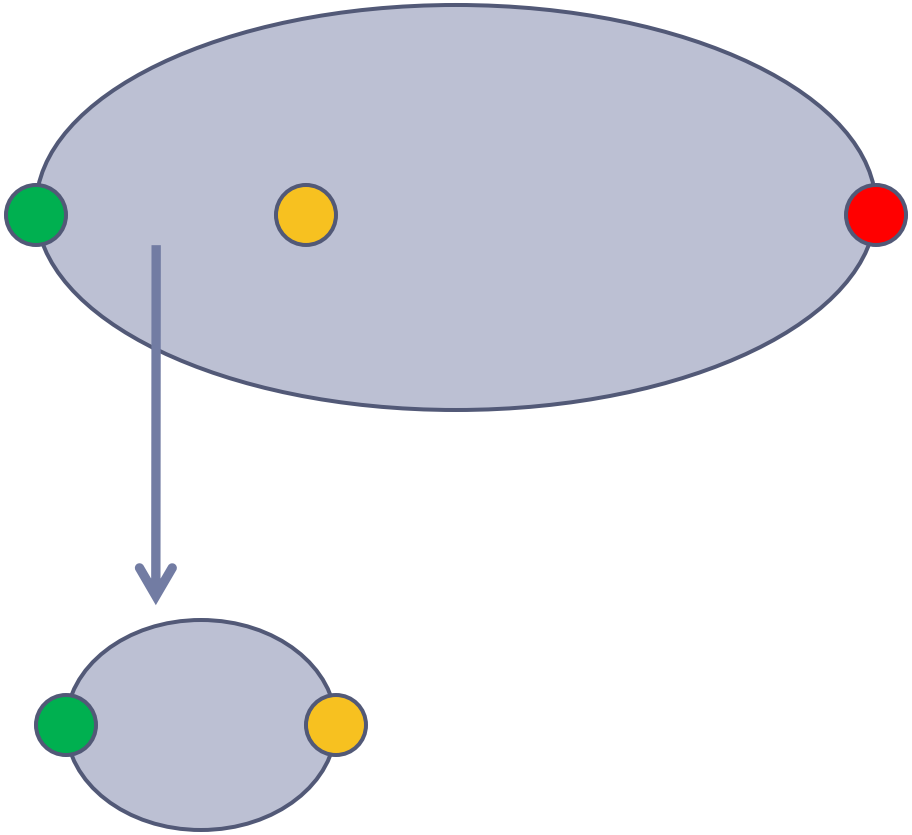
Policy transfer

- Example : subgoals



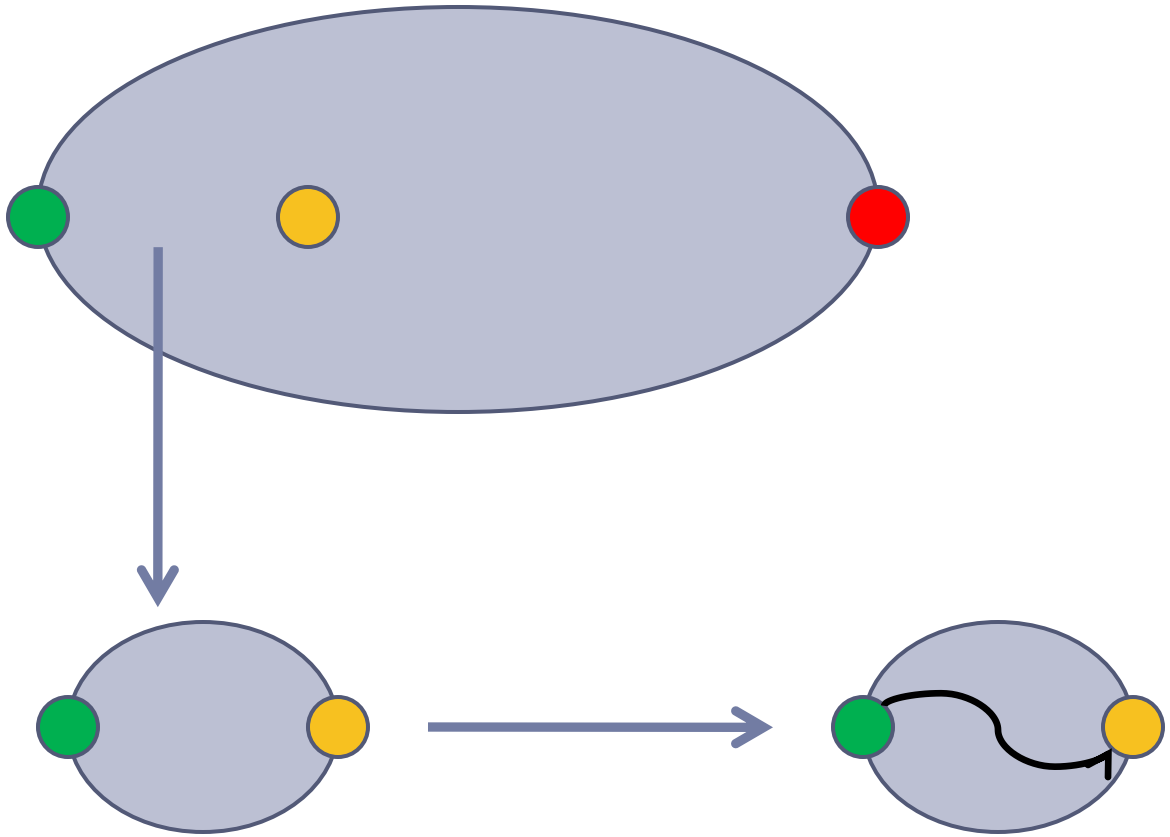
Policy transfer

- Example : subgoals



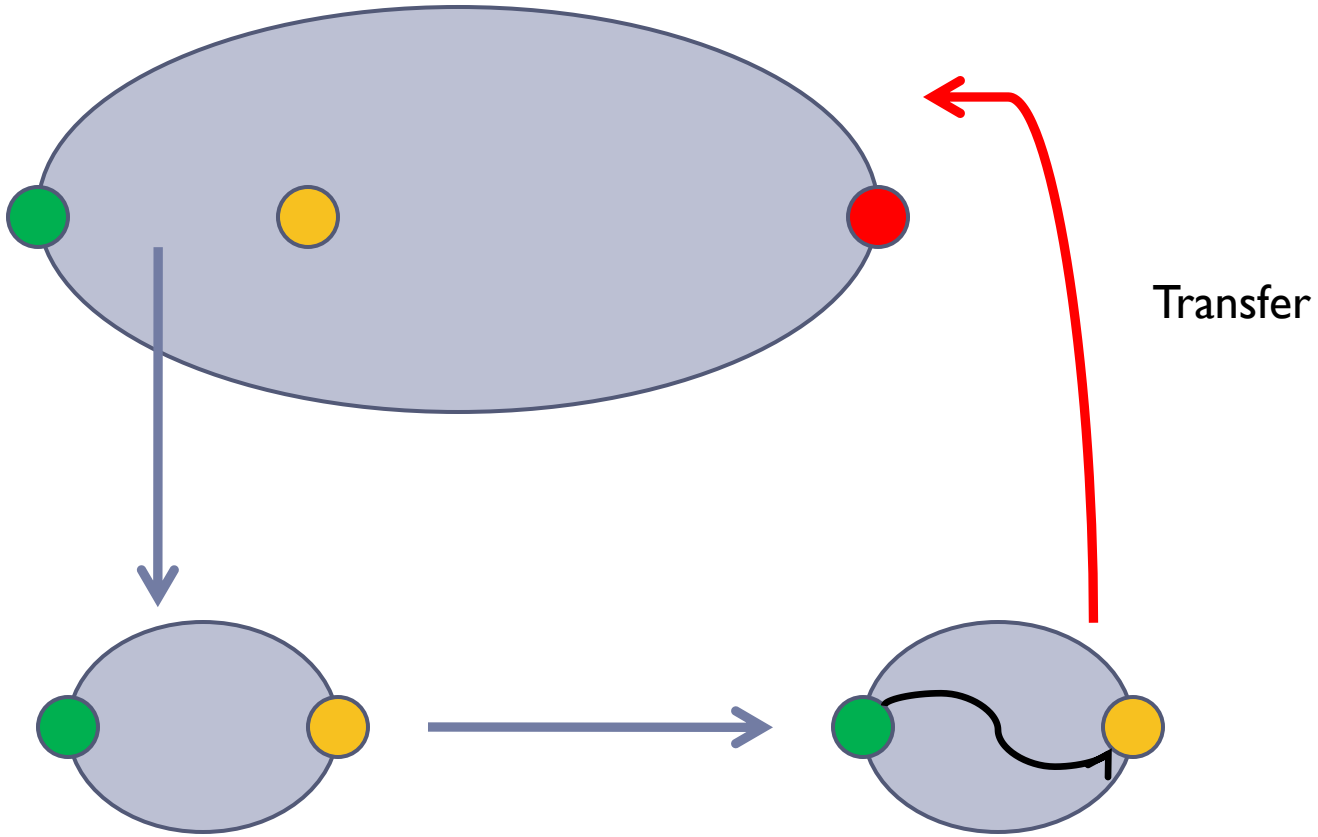
Policy transfer

- Example : subgoals



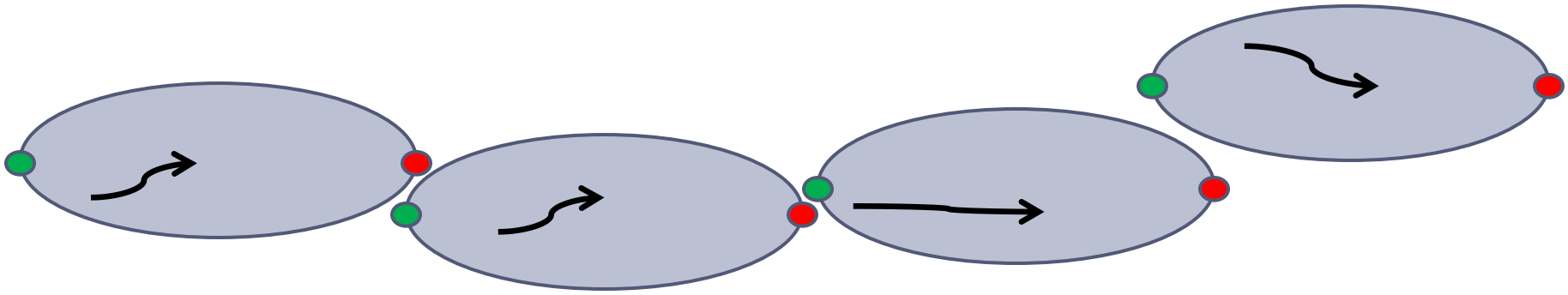
Policy transfer

- Example : subgoals



Policy transfer

- Using policy switching, we can transfer any policy, and learn how to use



Policy transfer

- Using policy switching, we can transfer any policy, and learn how to use