Sensorimotor Learning (Spring'23)

Pulkit Agrawal

Lecture 3: Bandits and Policy Gradients

Feb 14 2023

Course Logistics

Updated Assignment Release Schedule

Signup for Project Presentations

Weekly Status Report

Everyone getting Piazza E-mails?

From Course Status Reports

*“I'm a bit confused about the term "model-based RL",
I thought RL methods are by its nature model-free,
or maybe it's not?”*

“I am also unsure about the differences between RL, intuitive models and physics models.”

*“I thought the question about whether the objective function for RL vs SL was harder .. was kind of confusing,
because usually discussing "hardness" in algorithms / optimization is asking
whether one problem can reduce to the other problem.”*

Lecture Outline

Wrap up Bandits / Contextual Bandits

Understand Policy Gradients

Credit Assignment Problem

Variance Reduction Techniques

- Causality
- Discounting
- Baselines

CONTINUING

MULTI ARM BANDITS

Which arm to choose?

$$a_2, a_3, a_3, a_1, a_3, a_2, a_1, a_1, a_3 \ldots$$

Explore phase

Exploit phase

$$r(a_2), r(a_3), r(a_3), r(a_1), r(a_3), r(a_2), r(a_1), r(a_1), r(a_3) \ldots$$

time K

time T

$$\sum_{t=1}^{T} r(a_i^t) \quad i \in [1,N]$$

(total reward)

$$\mu_i = \frac{1}{k_i} \sum_{k_i} r(a_i)$$

(mean reward of action i)

One Strategy: Explore-First

Sample each arm equally $\approx \frac{K}{N}$

After K rounds, choose arm with highest average reward μ_i

Only take the highest rewarding action for remaining T-K rounds

Is this the best we can do?

(slide co-designed with Cathy Wu)

What do we mean by best?

$$a_2, a_3, a_3, a_1, a_3, a_2, a_1, a_1, a_3 \ldots$$

$$r(a_2), r(a_3), r(a_3), r(a_1), r(a_3), r(a_2), r(a_1), r(a_1), r(a_3) \ldots$$

time T

(total reward of selected actions)

$$R = \sum_{t=1}^{T} r(a_i^t)$$

(total reward of best actions)

$$R^* = \sum_{t=1}^{T} r(a^*)$$ oracle

Why is this called oracle?

Assume $r \in [0, 1]$

regret $||R^* - R||$

As in life, goal is to minimize regret

Poll time!

Not just the asymptotic performance, but how fast we reach!

(slide co-designed with Cathy Wu)

How bad can the agent possibly do?

(**i.e., what is the worst possible regret?**)

A. log(T)

B. T/2

C. T

D. None of the above

Poll time!

(slide co-designed with Cathy Wu)

What do we mean by best?

$$a_2, a_3, a_3, a_1, a_3, a_2, a_1, a_1, a_3 \ldots$$

$$r(a_2), r(a_3), r(a_3), r(a_1), r(a_3), r(a_2), r(a_1), r(a_1), r(a_3) \ldots$$

time T

(total reward of selected actions)

$$R = \sum_{t=1}^{T} r(a_i^t)$$

(total reward of best actions)

$$R^* = \sum_{t=1}^{T} r(a^*)$$

oracle

Why is this called oracle?

regret $||R^* - R||$

As in life, goal is to minimize regret

Assume $r \in [0, 1]$

Worst that we can do: T

Explore-First: $T^{2/3} \times O(N \log T)^{1/3}$

Not just the asymptotic performance, but how fast we reach!

(slide co-designed with Cathy Wu)

What do we mean by best?

$$a_2, a_3, a_3, a_1, a_3, a_2, a_1, a_1, a_3 \ldots$$

$$r(a_2), r(a_3), r(a_3), r(a_1), r(a_3), r(a_2), r(a_1), r(a_1), r(a_3) \ldots$$

Does there exist an optimal algorithm?

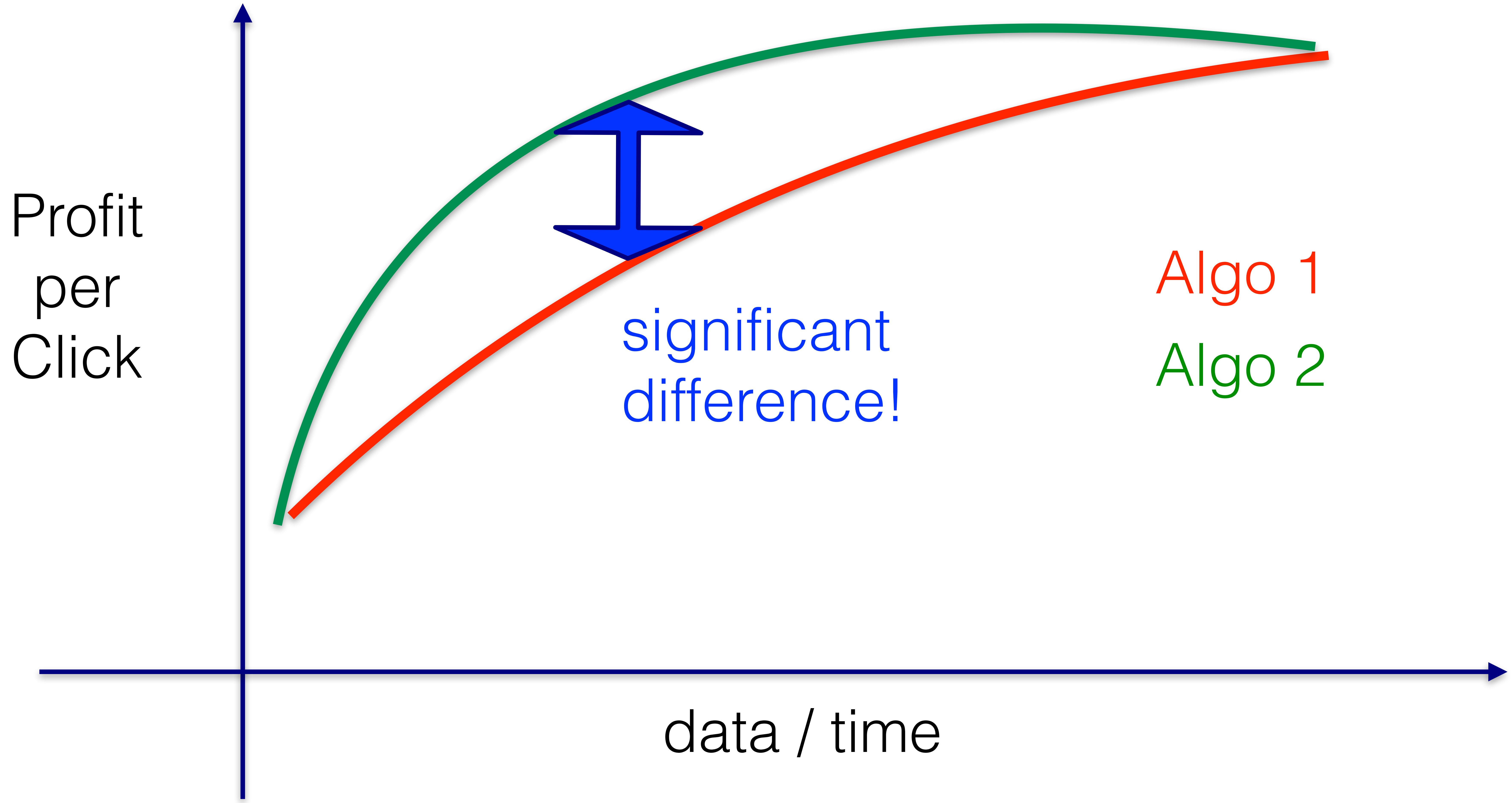
time T

$$R = \sum_{t=1}^{T} r(a_i^t) \qquad R^* = \sum_{t=1}^{T} r(a_i^{t*})$$ oracle

regret $||R^* - R||$

Not just the asymptotic performance, but how fast we reach!

Using a sub-optimal algorithm decreases profit



What do we mean by best?

$a_2, a_3, a_3, a_1, a_3, a_2, a_1, a_1, a_3 \ldots$

$r(a_2), r(a_3), r(a_3), r(a_1), r(a_3), r(a_2), r(a_1), r(a_1), r(a_3) \ldots$

Does there exist an optimal algorithm?

time T

Upper Confidence Bound (UCB) Algorithm

$R = \sum_{t=1}^{T} r(a_i^t)$ $\qquad$ $R^* = \sum_{t=1}^{T} r(a_i^*)$ oracle

regret $\quad ||R^* - R||$

Not just the asymptotic performance, but how fast we reach!

What do we mean by best?

Whats the notion of optimality?

Does there exist an optimal algorithm?

Upper Confidence Bound (UCB) Algorithm

regret $||R^* - R||$

Not just the asymptotic performance, but how fast we reach!

Explore-First

$$a_2, a_3, a_3, a_1, a_3, a_2, a_1, a_1, a_3 \ldots$$

Explore phase

Exploit phase

$$r(a_2), r(a_3), r(a_3), r(a_1), r(a_3), r(a_2), r(a_1), r(a_1), r(a_3) \ldots$$

time K

time T

Non-adaptive exploration

Explore + exploit <u>separately</u>

vs

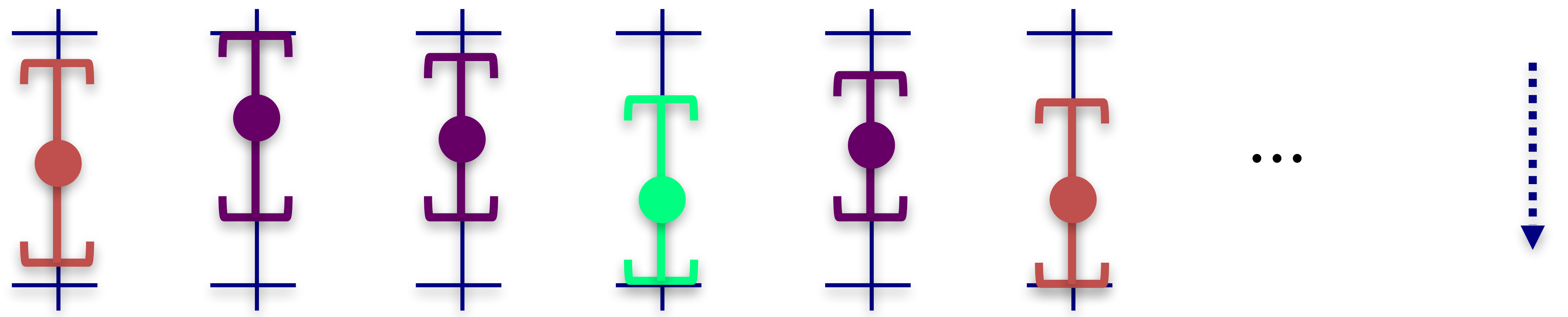
Adaptive exploration

Explore + Exploit <u>simultaneously</u>

(slide co-designed with Cathy Wu)

Upper Confidence Bound (UCB) Algorithm

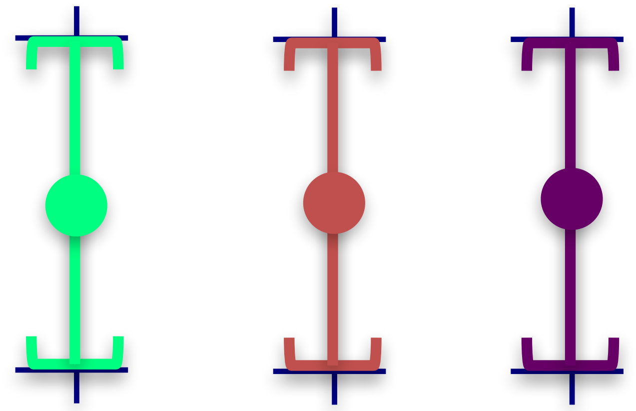
$a_2, a_3, a_3, a_1, a_3, a_2, a_1, a_1, a_3 \ldots$



$r(a_2), r(a_3), r(a_3), r(a_1), r(a_3), r(a_2), r(a_1), r(a_1), r(a_3) \ldots$

time T

Initial confidence intervals:



$a_1 \quad a_2 \quad a_3$

Exploitation

$$a_{t+1} = \arg\max_i \mu_i(t) + \sqrt{\frac{4 \log t}{k_i}}$$

$$\mu_i = \frac{1}{k_i} \sum_{k_i} r(a_i)$$

(mean reward of action i)

Optimism in face of uncertainty

Exploration bonus for rare actions (optimism)

(slide co-designed with Cathy Wu)

How good is UCB?

$$a_2, a_3, a_3, a_1, a_3, a_2, a_1, a_1, a_3 \ldots$$

$$r(a_2), r(a_3), r(a_3), r(a_1), r(a_3), r(a_2), r(a_1), r(a_1), r(a_3) \ldots$$

time T

(total reward of selected actions)

$$R = \sum_{t=1}^{T} r(a_i^t)$$

(total reward of best actions)

$$R^* = \sum_{t=1}^{T} r(a^*)$$

oracle

regret $||R^* - R||$

As in life, goal is to minimize regret

Optimal! (up to log factors)

Assume $r \in [0, 1]$

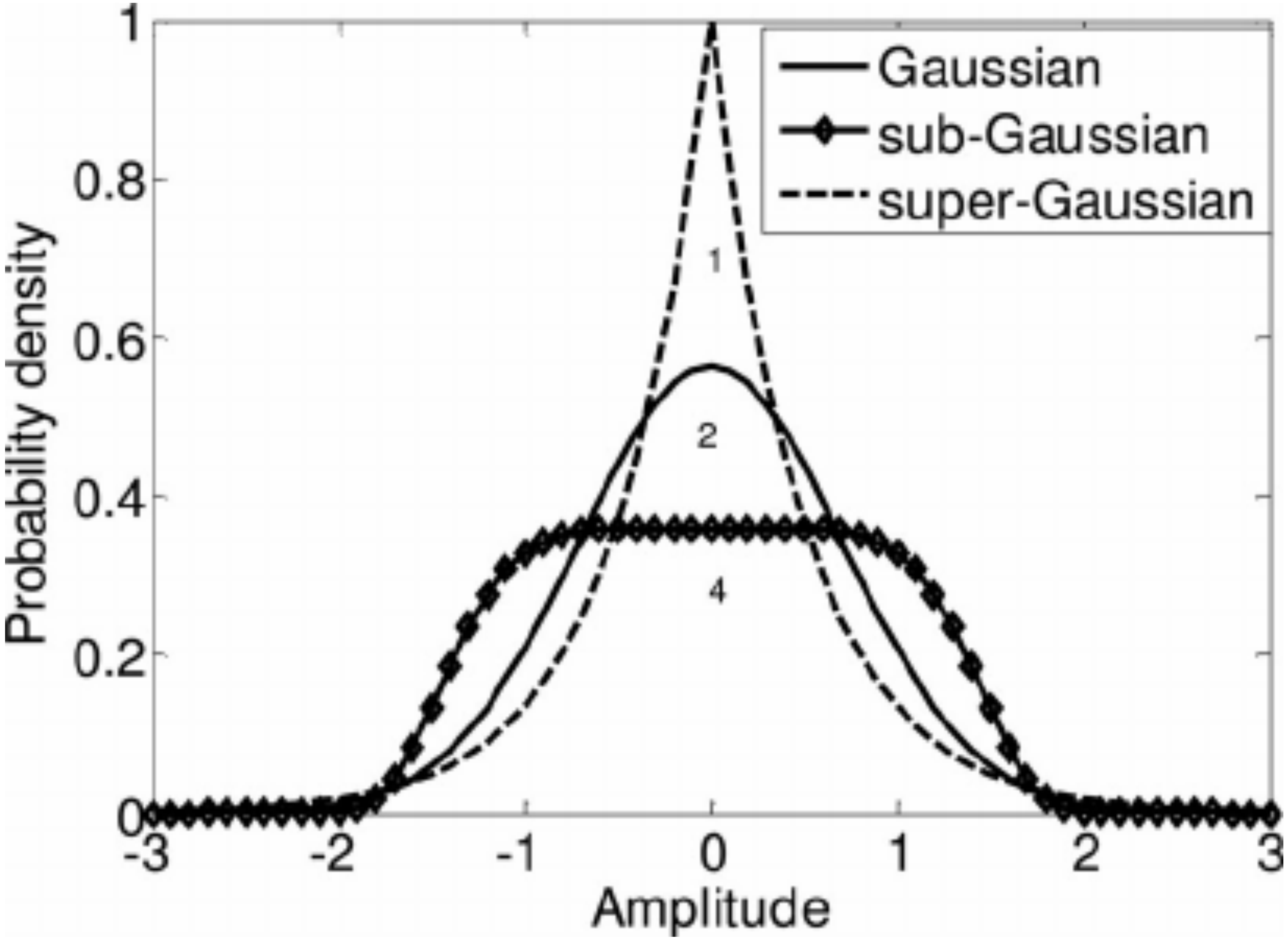
Worst that we can do: T

Explore-First: $T^{2/3}$ x $(N \log T)^{1/3}$

UCB: $(NT \log T)^{1/2}$

(slide co-designed with Cathy Wu)

Assume Payoffs are Sub-Gaussian



Upper Confidence Bound (UCB) Algorithm

$$\arg\max_i \hat{\mu}_i(t) + \sqrt{\frac{4\log t}{k_i}}$$

$$\hat{\mu}_i = \frac{1}{k_i}\sum_{k_i} r(a_i)$$

Where this come from?

Empirical estimate of μ_i (unknown)

We want

$$\arg\max_i \mu_i$$

We have

$$\arg\max_i \hat{\mu}_i$$

Initially (with few samples)
This estimate is going to be bad

Construct

$$\arg\max_i \hat{\mu}'_i$$

Principle of optimism: find $\hat{\mu}'_i \geq \mu_i$

$$p\left(\mu_i \geq \hat{\mu}'_i\right) \leq \delta$$

Is true!

If $r(a_i)$ are 1-subgaussian and if $\hat{\mu}'_i : \hat{\mu}_i + \sqrt{\frac{2\log\frac{1}{\delta}}{k_i}}$

Upper Confidence Bound (UCB) Algorithm

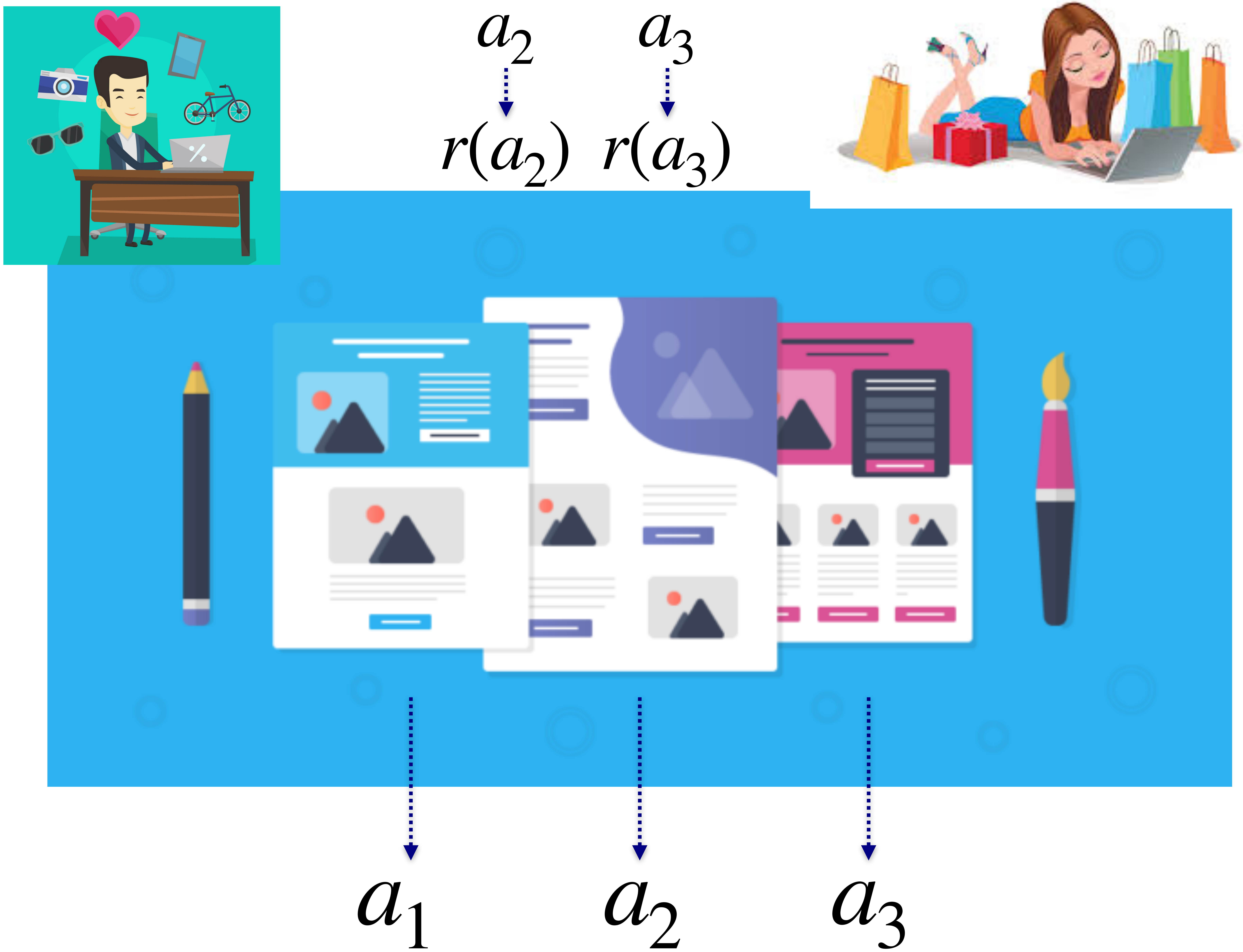
$$\arg\max_i \mu_i(t) + \sqrt{\frac{4\log t}{k_i}} \qquad \mu_i = \frac{1}{k_i}\sum_{k_i} r(a_i)$$

Upper Bound on average number of sub-optimal actions

$$\frac{16|A|\log T}{\Delta^2} + O(1)$$

$\Delta = \mu_{best} - \mu_{second\_best}$

$|A|$: number of actions



a_2
a_3
$r(a_2)$
$r(a_3)$
a_1
a_2
a_3

$a_2 \to r(a_2)$ $\quad$ $a_3 \to r(a_3)$

(female, 20s, computer-savvy)

(male, 30s, computer-savvy)

How to use these "features" in decision making?

Contextual Bandits

a_1 $\quad$ a_2 $\quad$ a_3

a_2 a_3

$\cancel{r(a_2)}$ $\cancel{r(a_3)}$

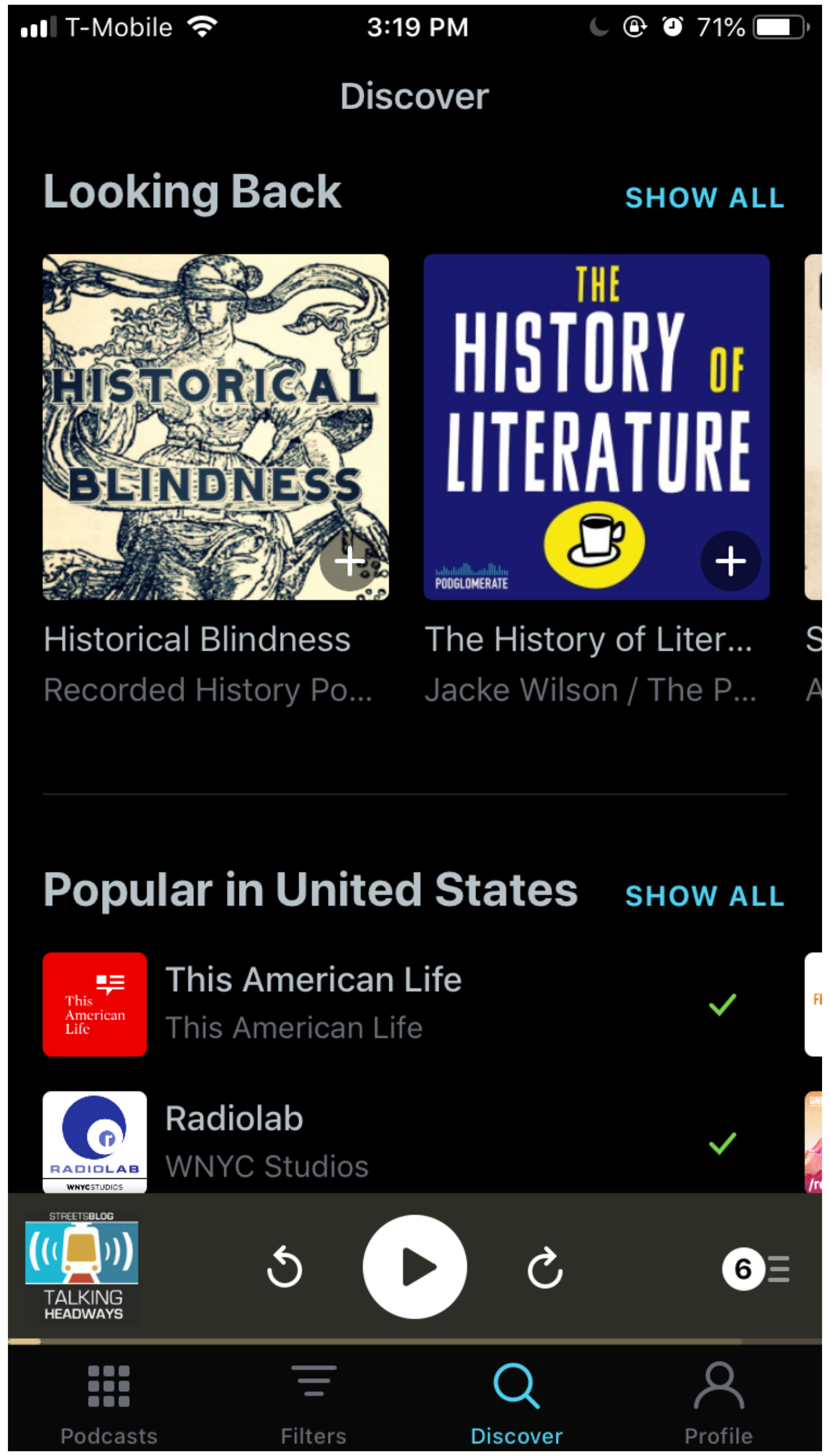
$r(a_2, x_2)$ $r(a_3, x_3)$

x_3

(female, 20s, computer-savvy)

x_2 **(male, 30s, computer-savvy)**

Example:
Podcast
recommendations



(slide co-designed with Cathy Wu)

a_2 a_3

$\cancel{r(a_2)}$ $\cancel{r(a_3)}$

$r(a_2, x_2)$ $r(a_3, x_3)$

x_3

(female, 20s, computer-savvy)

x_2

(male, 30s, computer-savvy)

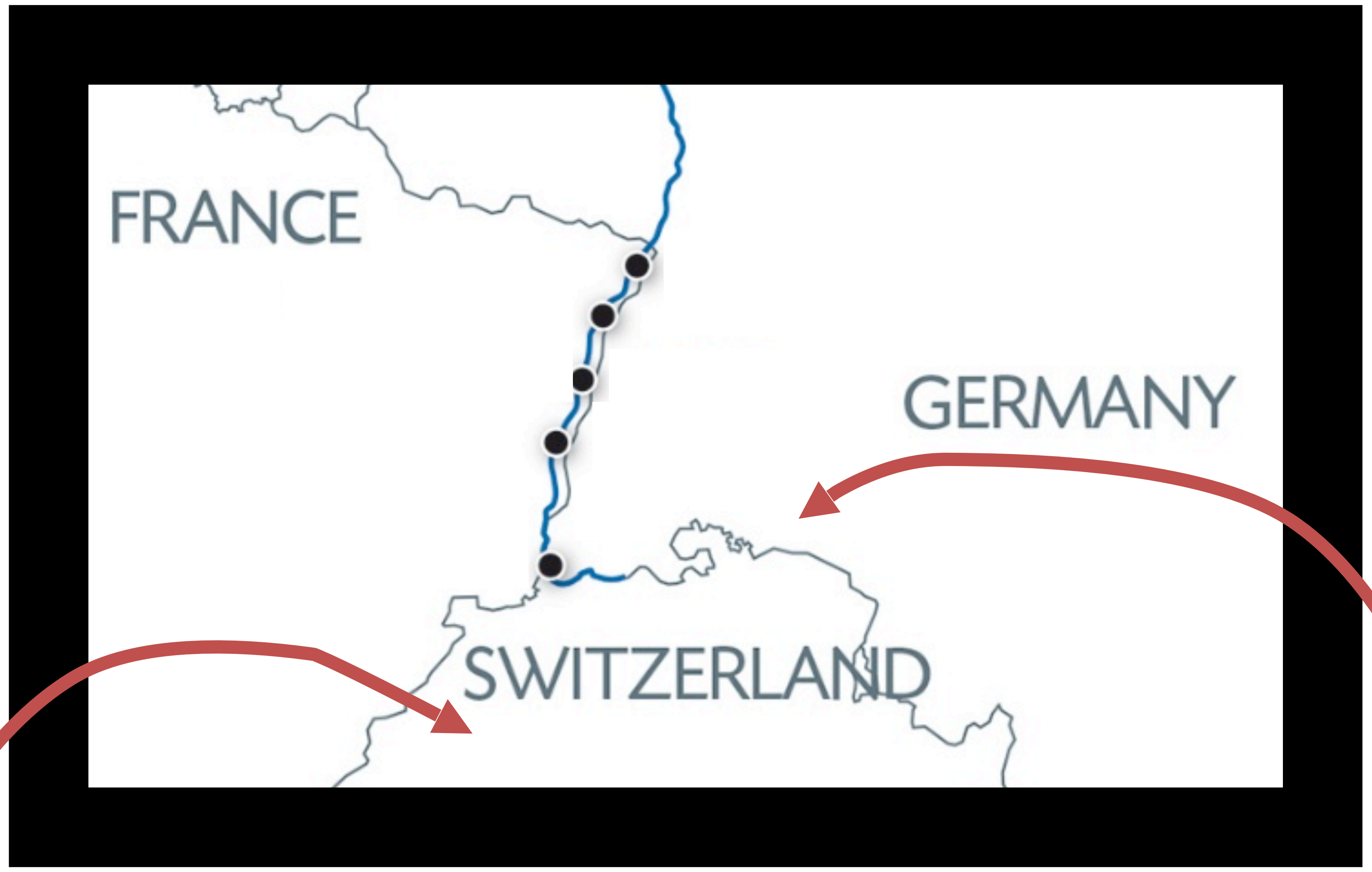
Example types of contexts

- User demographics
  - Discrete vs continuous

Switzerland

vs

(lat, long) coordinates



Wine or beer podcast?

vs

(slide co-designed with Cathy Wu)

$$a_2 \qquad a_3$$

$$\cancel{r(a_2)} \qquad \cancel{r(a_3)}$$

$$r(a_2, x_2) \qquad r(a_3, x_3)$$

x_3

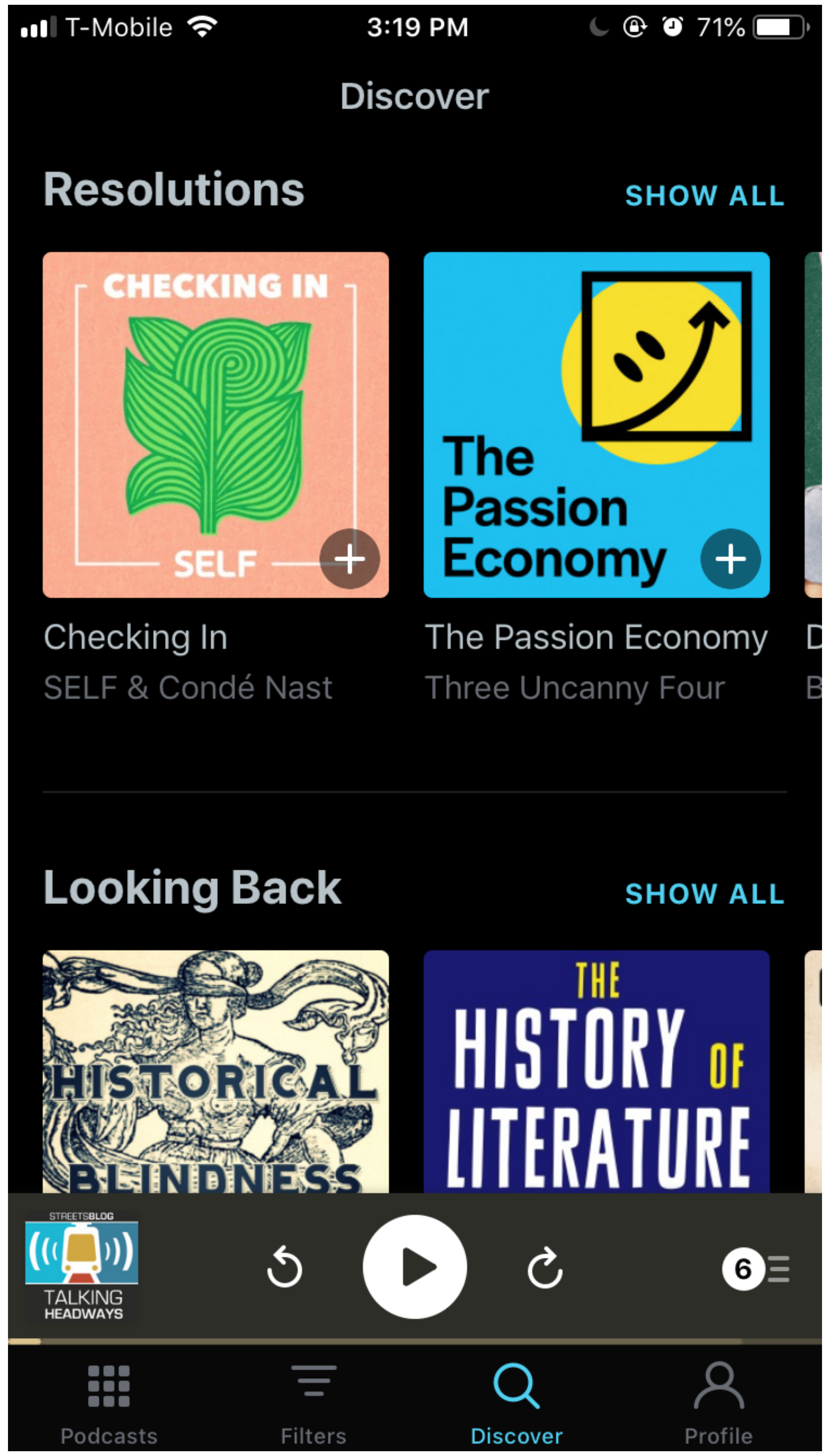
(female, 20s, computer-savvy)

x_2

(male, 30s, computer-savvy)

Example types of contexts

- User demographics
- Time of day/year
- User mental state



Happy 2023!
New years resolutions?

(slide co-designed with Cathy Wu)

a_2 a_3

$\cancel{r(a_2)}$ $\cancel{r(a_3)}$

$r(a_2, x_2)$ $r(a_3, x_3)$

x_2

(male, 30s, computer-savvy)

x_3

(female, 20s, computer-savvy)

Naïve approach:
independent bandit problems
(one for each context)

$\dashrightarrow$

for every user solve
a separate bandit problem

Challenge:
may not handle continuous
contexts well

(slide co-designed with Cathy Wu)

a_2 a_3

$\cancel{r(a_2)}$ $\cancel{r(a_3)}$

$r(a_2, x_2)$ $r(a_3, x_3)$

x_2

(male, 30s, computer-savvy)

x_3

(female, 20s, computer-savvy)

Better than context-free UCB:

LinUCB

Assume:

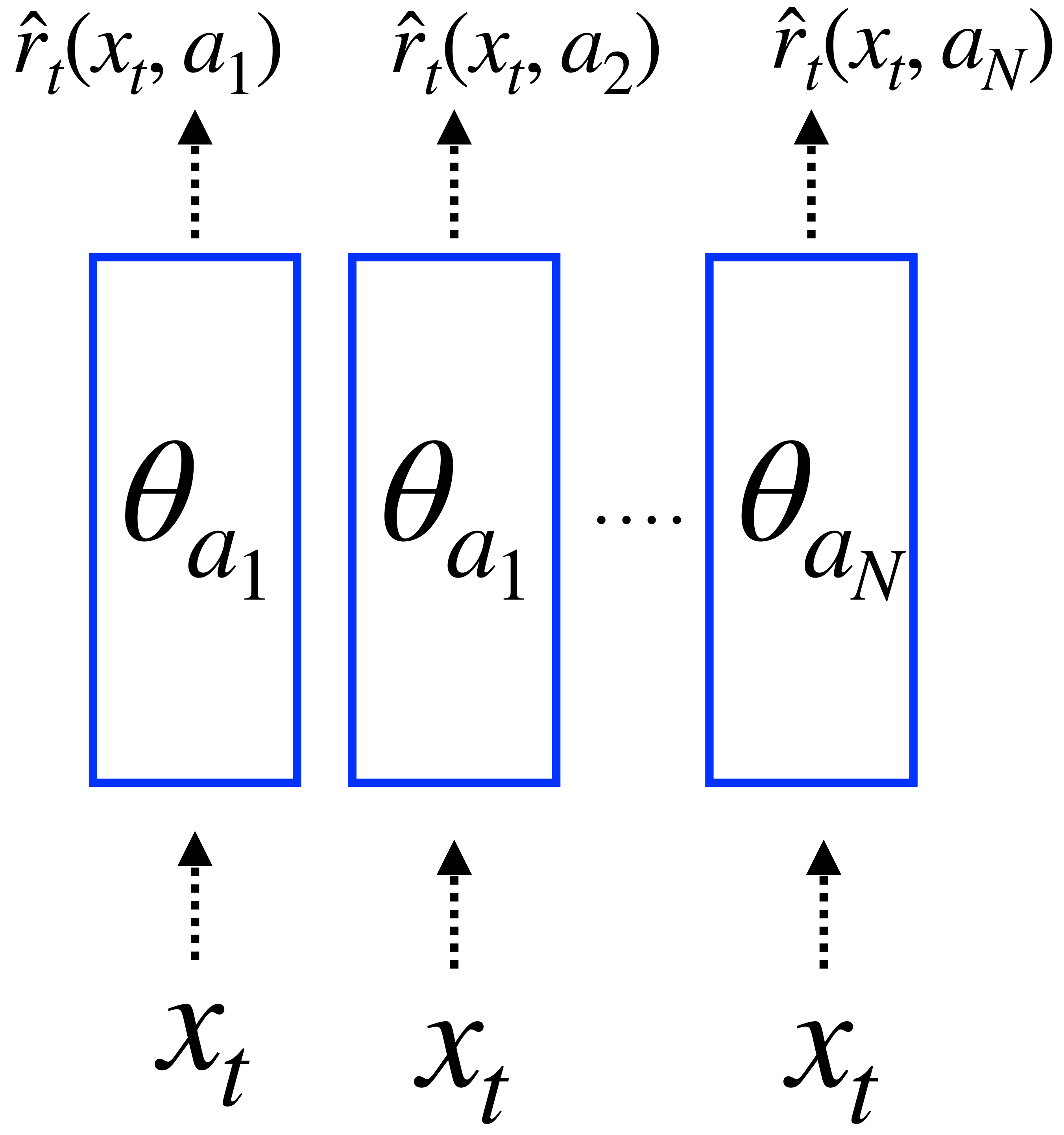
expected rewards are linear in context

$$\mu(a \,|\, s) = x_a^T \theta_a$$

Optimal! (up to log factors)

(slide co-designed with Cathy Wu)

Dis-Joint LinUCB



(female, 30s, computer-savvy)

Ridge Regression

$$\theta_t^a = \underbrace{\boxed{\left(X_{0:t}^T X_{0:t} + \lambda I\right)^{-1}}}_{A_a}\underbrace{\boxed{X_{0:t}^T R_{0:t}^a}}_{b_a}$$

Estimate reward for each action

$$\hat{r}_t^a = x_t \theta_t^a$$

Choose the best one (or sample)

$$a_t \leftarrow \max_a \hat{r}_t^a$$

All time steps until t

$$\hat{R}_{0:t}^a = X_{0:t} \theta^a$$

Solve for the parameters

$$\min_{\theta_a} ||R_{0:t}^a - \hat{R}_{0:t}^a||_2^2$$

Need to solve
Online!

$$\theta_t^a = \underbrace{\left(X_{0:t}^T X_{0:t} + \lambda I\right)^{-1}}_{A_a} \underbrace{X_{0:t}^T R_{0:t}^a}_{b_a}$$

Algorithm 1 LinUCB with disjoint linear models.

Inputs: $\alpha \in \mathbb{R}_+$
for $t = 1, 2, 3, \ldots, T$ **do**
  Observe features of all arms $a \in \mathcal{A}_t$: $\mathbf{x}_{t,a} \in \mathbb{R}^d$
  for all $a \in \mathcal{A}_t$ **do**
    $\boldsymbol{\theta}_a \leftarrow \mathbf{A}_a^{-1}\mathbf{b}_a$
  end for
end for

Exploration Bonus

Online Update

A Contextual-Bandit Approach to Personalized News Article Recommendation, Li et al. 2012

$$\theta_t^a = \underbrace{\boxed{\left(X_{0:t}^T X_{0:t} + \lambda I\right)^{-1}}}_{A_a}\underbrace{\boxed{X_{0:t}^T R_{0:t}^a}}_{b_a}$$

Algorithm 1 LinUCB with disjoint linear models.

Inputs: $\alpha \in \mathbb{R}_+$

for $t = 1, 2, 3, \ldots, T$ **do**

  Observe features of all arms $a \in \mathcal{A}_t$: $\mathbf{x}_{t,a} \in \mathbb{R}^d$

  for all $a \in \mathcal{A}_t$ **do**

    $\boldsymbol{\theta}_a \leftarrow \mathbf{A}_a^{-1}\mathbf{b}_a$

    $p_{t,a} \leftarrow \hat{\boldsymbol{\theta}}_a^\top \mathbf{x}_{t,a} + \alpha\boxed{\sqrt{\mathbf{x}_{t,a}^\top \mathbf{A}_a^{-1}\mathbf{x}_{t,a}}}$ Exploration Bonus

  end for

  Choose arm $a_t = \arg\max_{a \in \mathcal{A}_t} p_{t,a}$ with ties broken arbitrarily, and observe a real-valued payoff r_t

  $\mathbf{A}_{a_t} \leftarrow \mathbf{A}_{a_t} + \mathbf{x}_{t,a_t}\mathbf{x}_{t,a_t}^\top$

  $\mathbf{b}_{a_t} \leftarrow \mathbf{b}_{a_t} + r_t\mathbf{x}_{t,a_t}$ Online Update

end for

Pros and Cons of "disjoint models" (separate θ_a for each action) ?

A Contextual-Bandit Approach to Personalized News Article Recommendation, Li et al. 2012

$$\theta_t^a = \underbrace{\boxed{\left(X_{0:t}^T X_{0:t} + \lambda I\right)^{-1}}}_{A_a} \underbrace{\boxed{X_{0:t}^T R_{0:t}^a}}_{b_a}$$

Algorithm 1 LinUCB with disjoint linear models.

Inputs: $\alpha \in \mathbb{R}_+$

for $t = 1, 2, 3, \ldots, T$ **do**

  Observe features of all arms $a \in \mathcal{A}_t$: $\mathbf{x}_{t,a} \in \mathbb{R}^d$

  for all $a \in \mathcal{A}_t$ **do**

    $\boldsymbol{\theta}_a \leftarrow \mathbf{A}_a^{-1}\mathbf{b}_a$

    $p_{t,a} \leftarrow \hat{\boldsymbol{\theta}}_a^\top \mathbf{x}_{t,a} + \alpha \boxed{\sqrt{\mathbf{x}_{t,a}^\top \mathbf{A}_a^{-1} \mathbf{x}_{t,a}}}$ Exploration Bonus

  end for

  Choose arm $a_t = \arg\max_{a \in \mathcal{A}_t} p_{t,a}$ with ties broken arbitrarily, and observe a real-valued payoff r_t

  $\mathbf{A}_{a_t} \leftarrow \mathbf{A}_{a_t} + \mathbf{x}_{t,a_t}\mathbf{x}_{t,a_t}^\top$

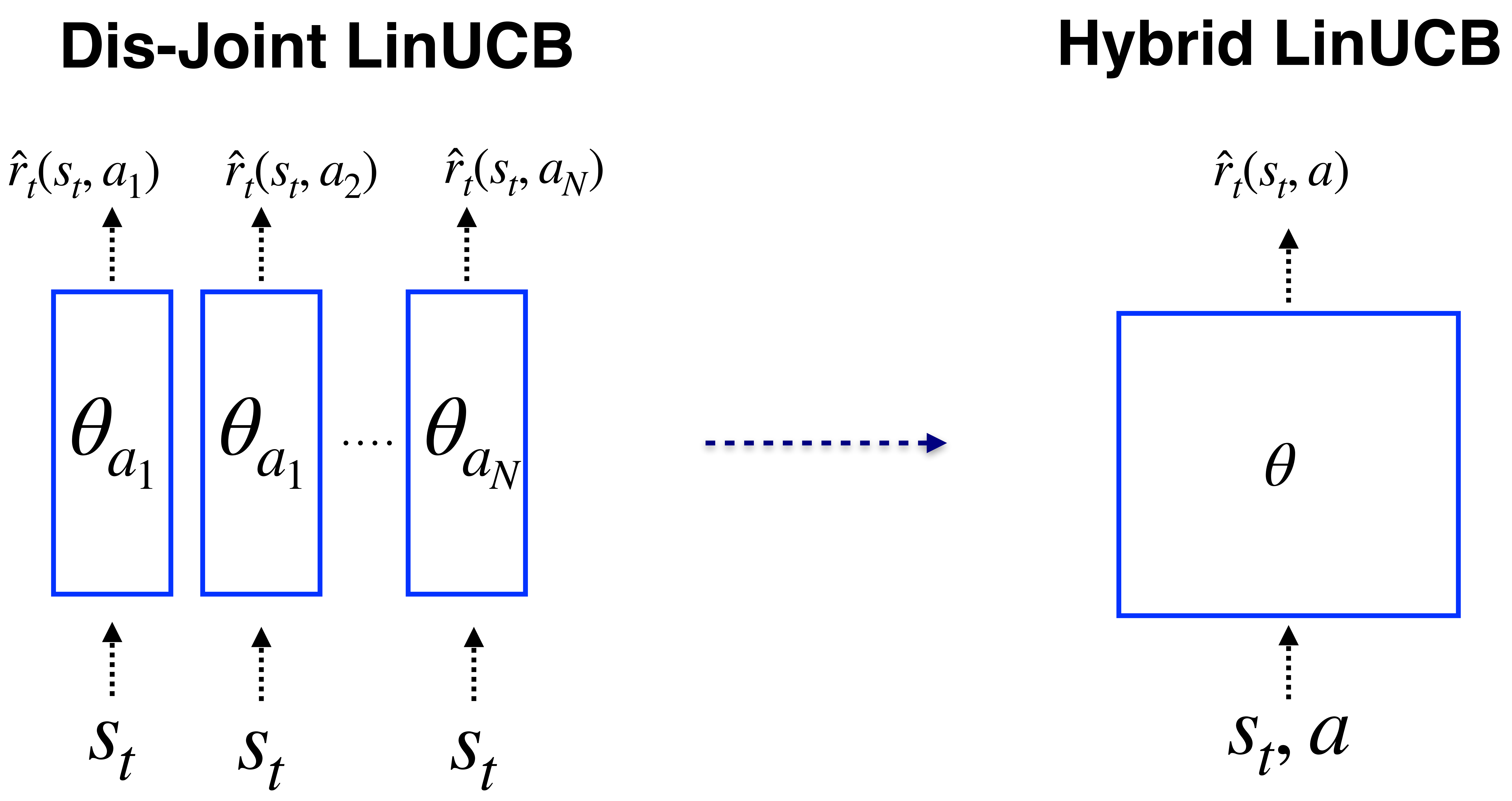
  $\mathbf{b}_{a_t} \leftarrow \mathbf{b}_{a_t} + r_t \mathbf{x}_{t,a_t}$

Online Update

end for

What if there are new news articles?

A Contextual-Bandit Approach to Personalized News Article Recommendation, Li et al. 2012



Going beyond the linear assumption

Beyond UCB: Optimal and Efficient Contextual Bandits with Regression Oracles, Foster & Rakhlin, 2020

$$a_2 \quad a_3$$

$$r(a_2) \quad r(a_3)$$

(female, 20s, computer-savvy)

(male, 30s, computer-savvy)

How to use these "features" in decision making?

Contextual Bandits

Optimal Exploration-Exploitation Tradeoff?

(Square CB Algorithm)

$$a_1 \quad a_2 \quad a_3$$

$$a_2 \qquad a_3$$

$$r(a_2) \quad r(a_3)$$

(female, 20s, computer-savvy)

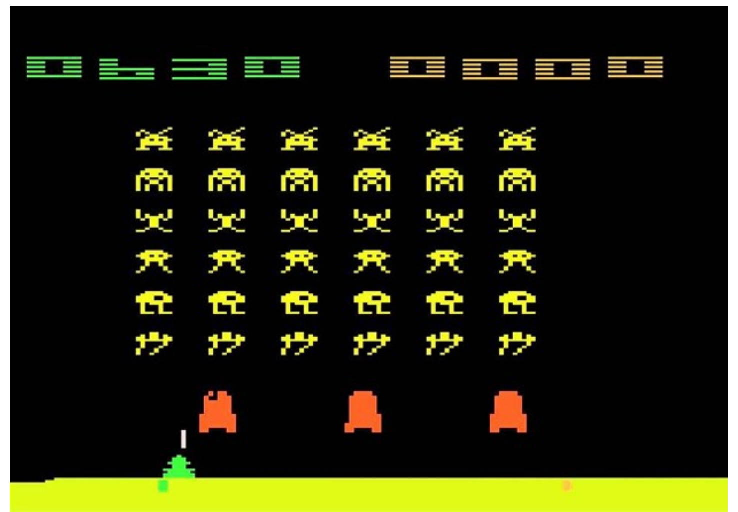
(male, 30s, computer-savvy)

How to use these “features” in decision making?

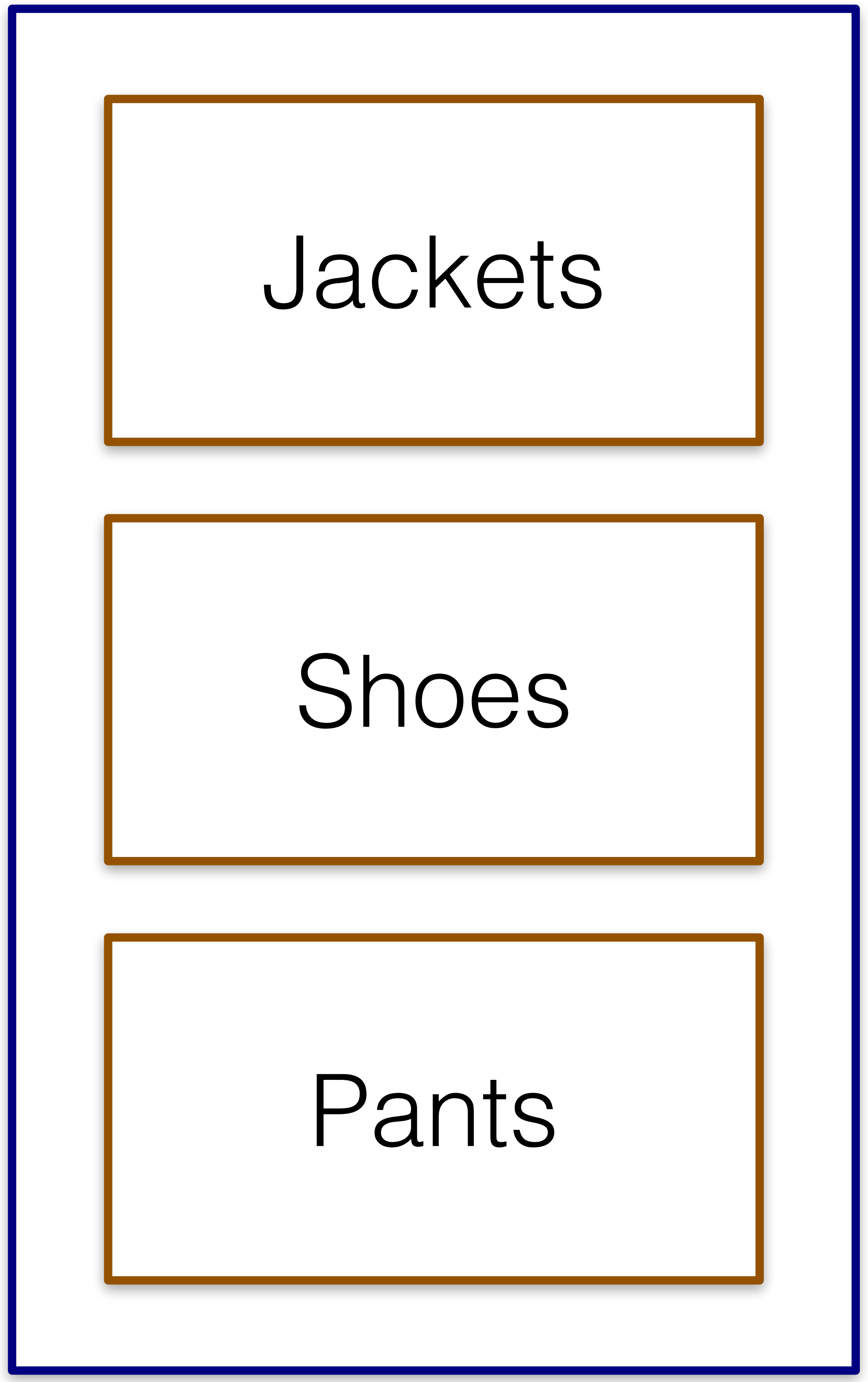
Contextual Bandits

BUT, Actions don’t change future state

Model Free Reinforcement Learning

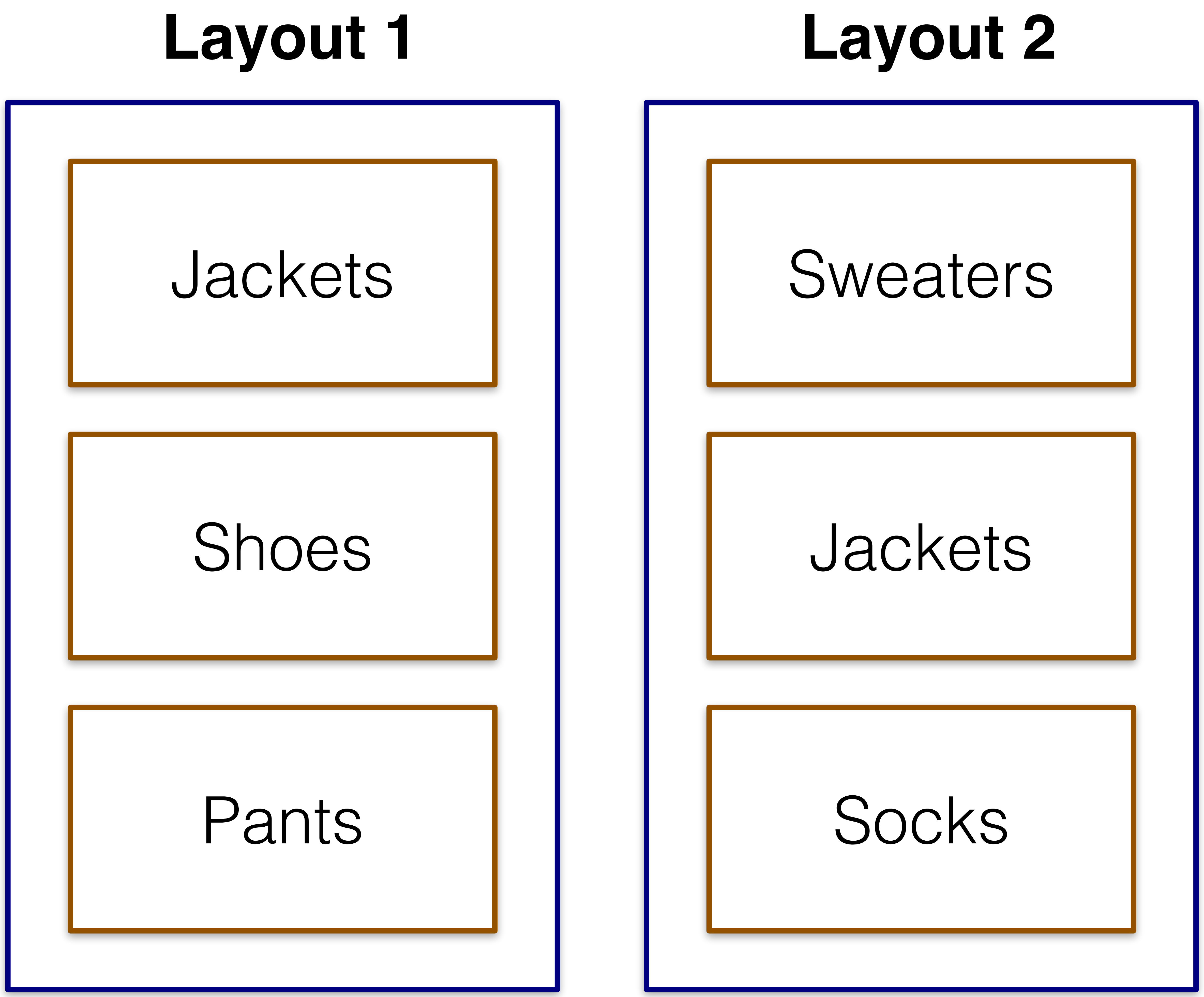


Layout 1



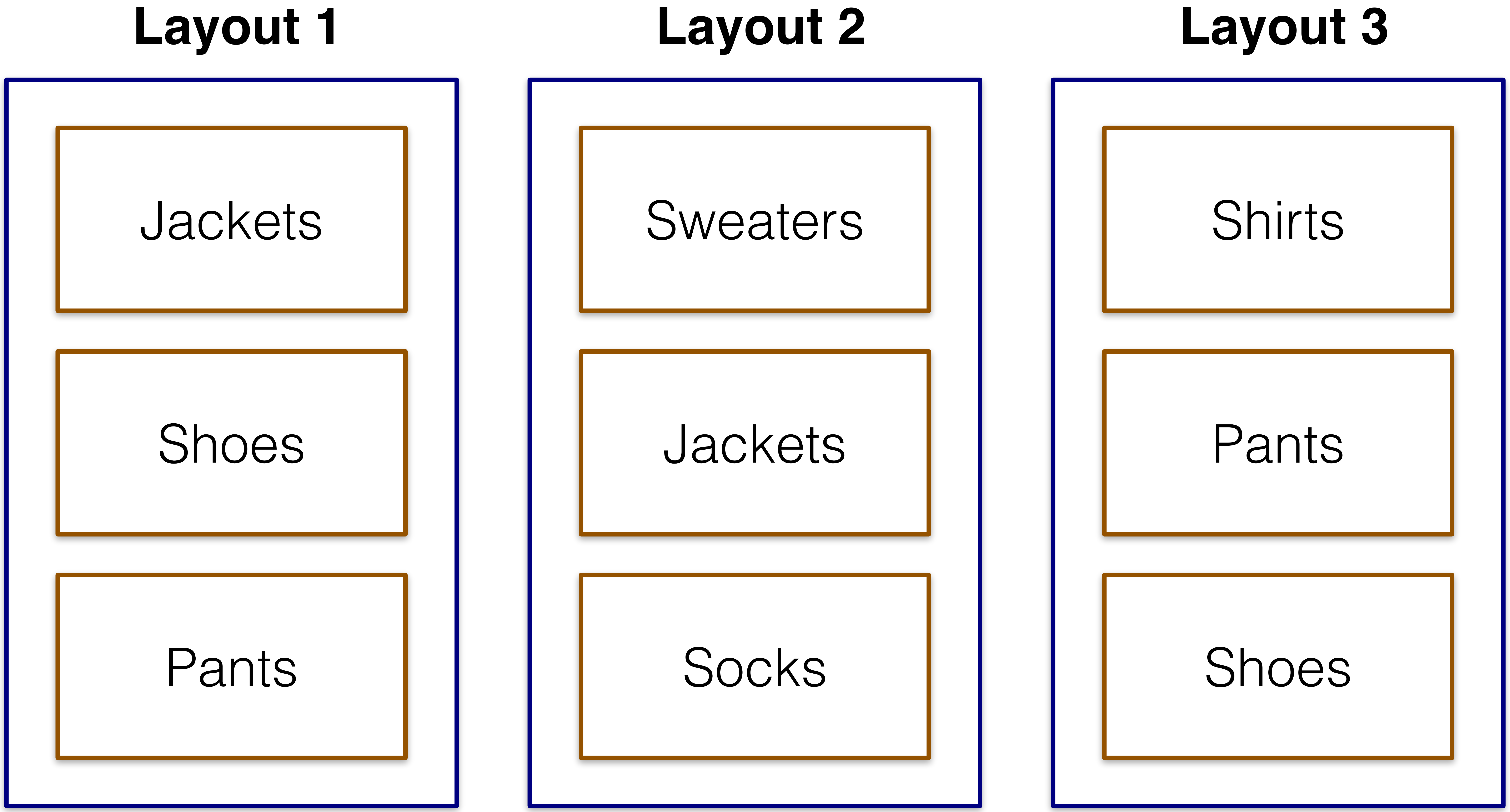
state: x_t x_t: user features

action: a_1



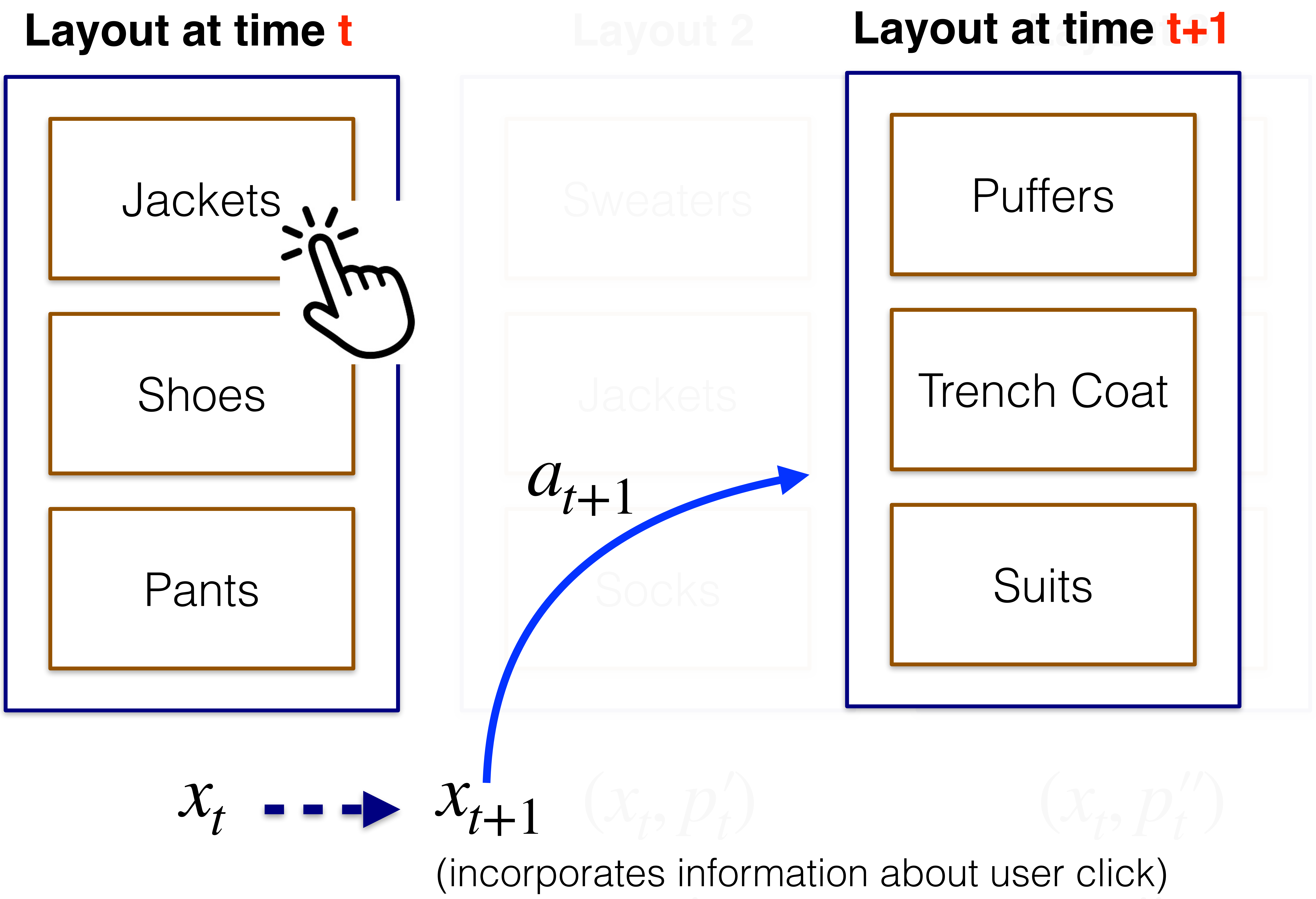
state: x_t

action: a_1 a_2



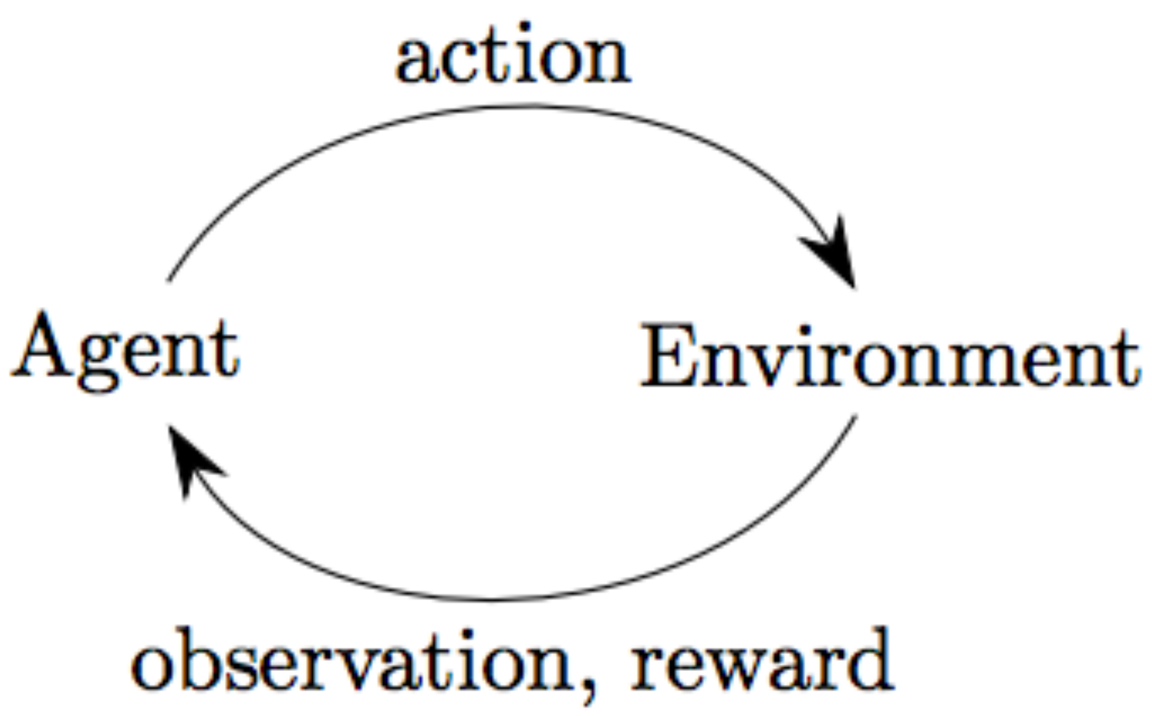
state: x_t

action: a_1 a_2 a_3



State of the system evolves with actions

The problem

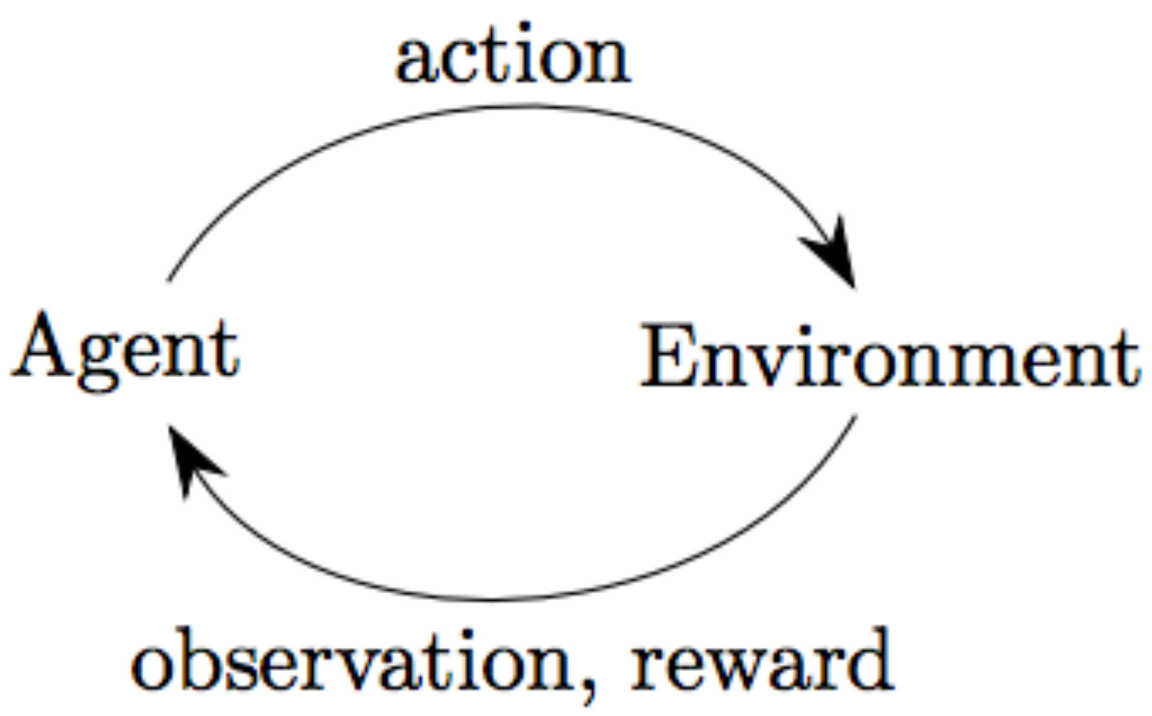


$s_1, a_1, r_1, s_2, a_2, r_2, s_3, a_3, \ldots.$

(State-action-reward trajectory)

(trajectory or rollout)

The problem



$s_1, a_1, r_1, s_2, a_2, r_2, s_3, a_3, \ldots.$

Goal

$$a_t = \pi_\theta(s_{0:t}) \qquad s.t. \quad \max \sum_t r_t$$

Goal

$$a_t = \pi_\theta(s_{0:t}) \qquad s.t. \boxed{\max \sum_t r_t}$$

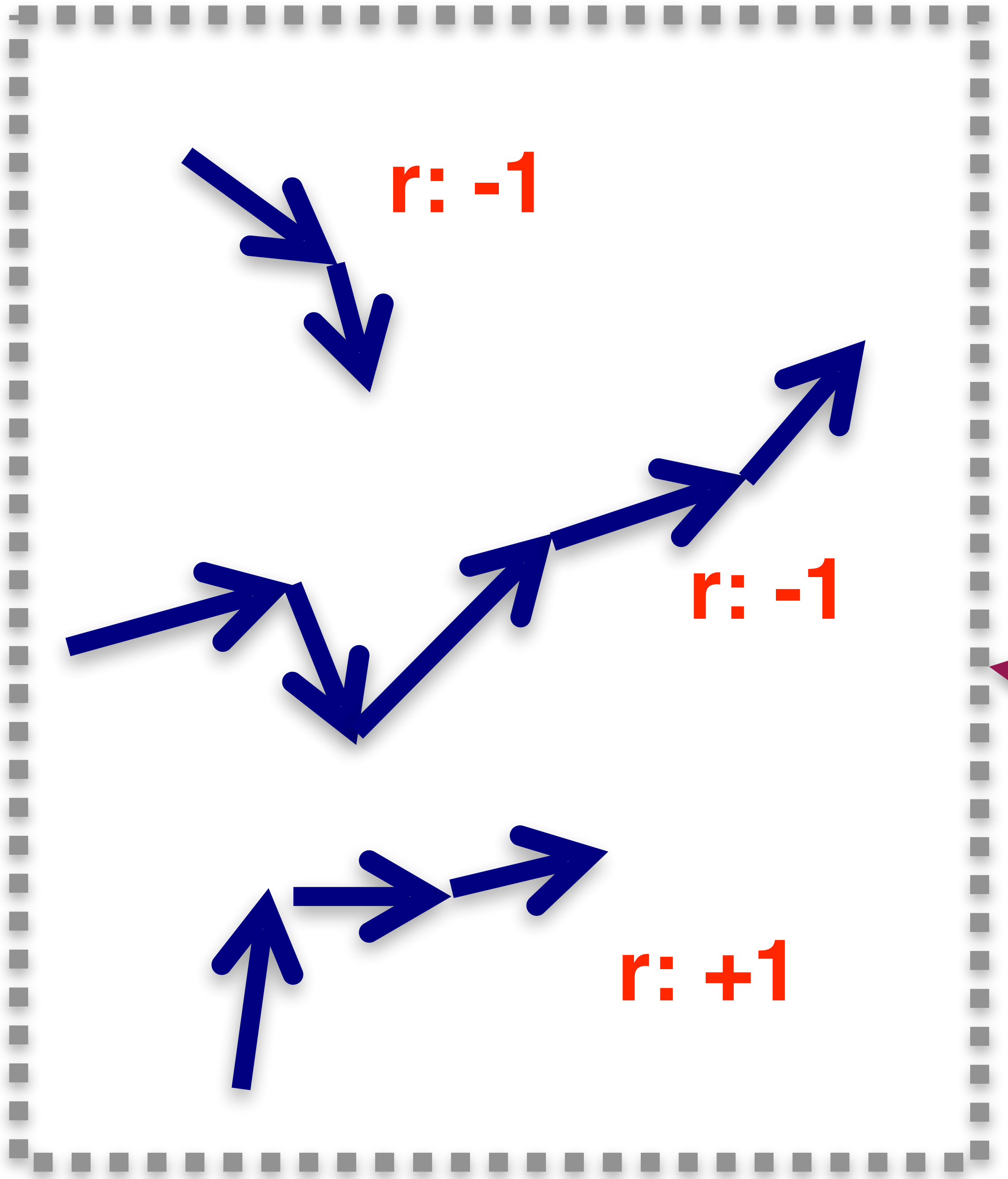
Infinite Time Horizon

$$\sum_t r_t$$

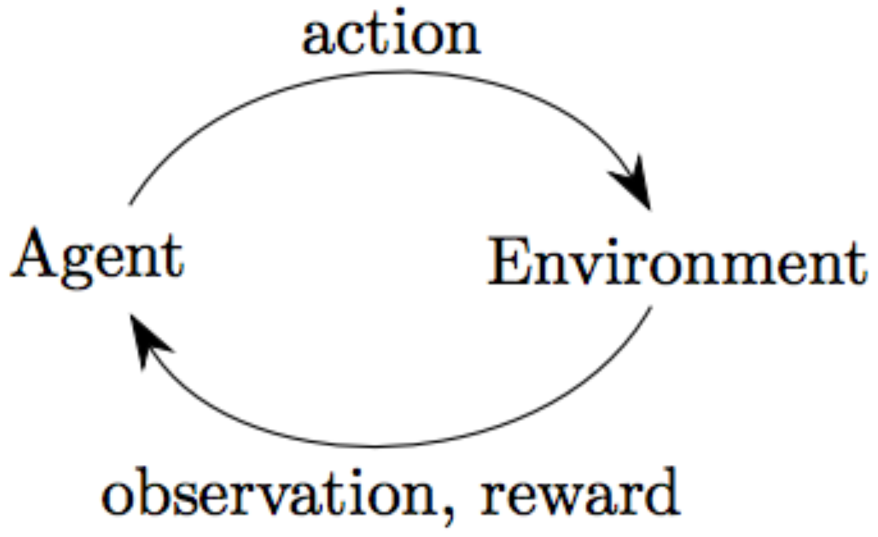
Finite Time Horizon

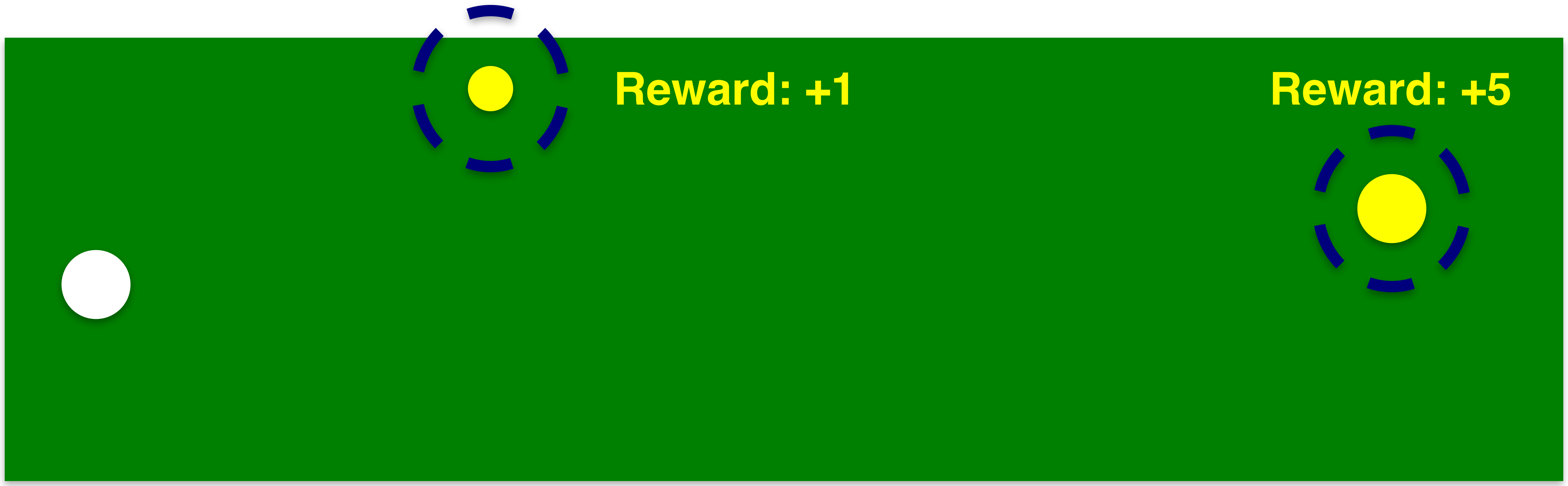
$$\sum_{t=1}^{T} r_t$$

Dataset

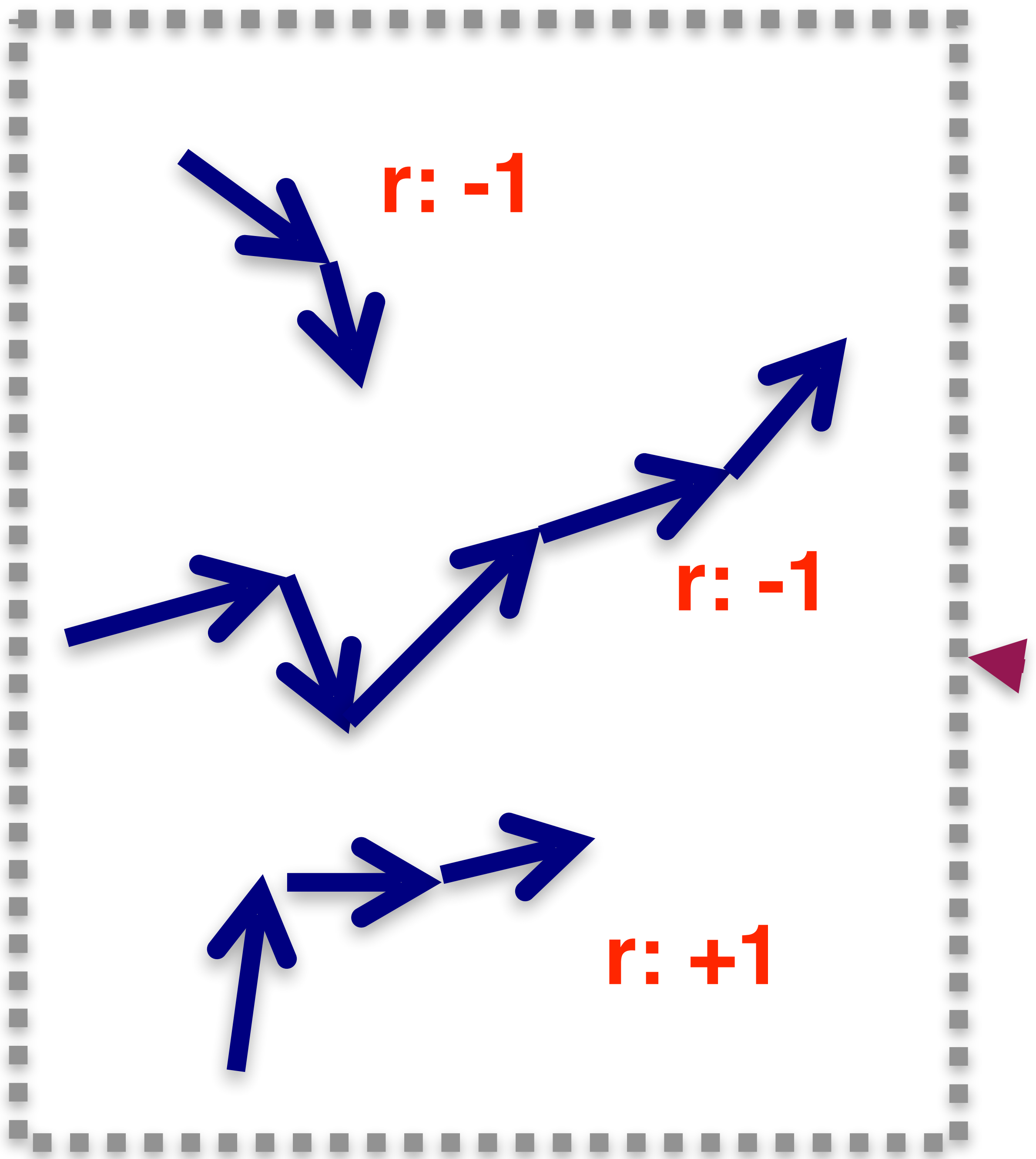


$$\sum_{t=1}^{T=100} r_t \qquad a_t = \pi_\theta(s_{0:t})$$

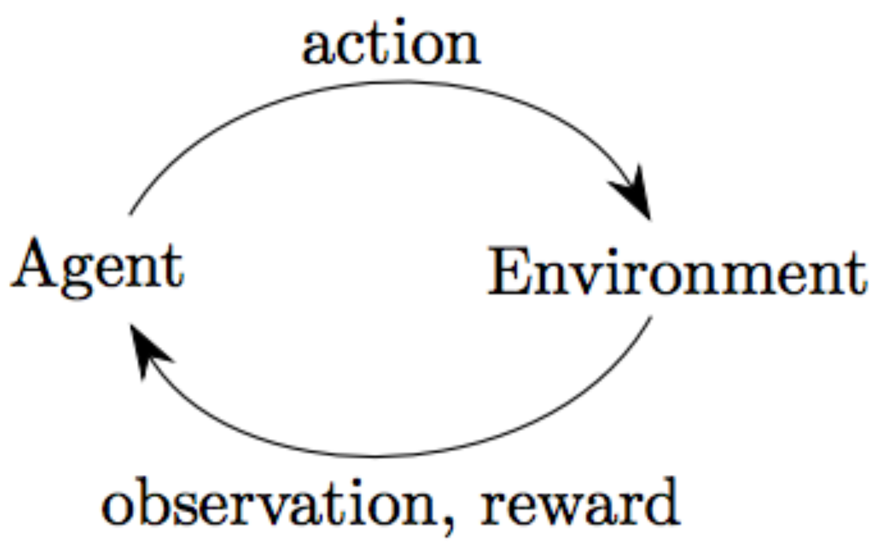


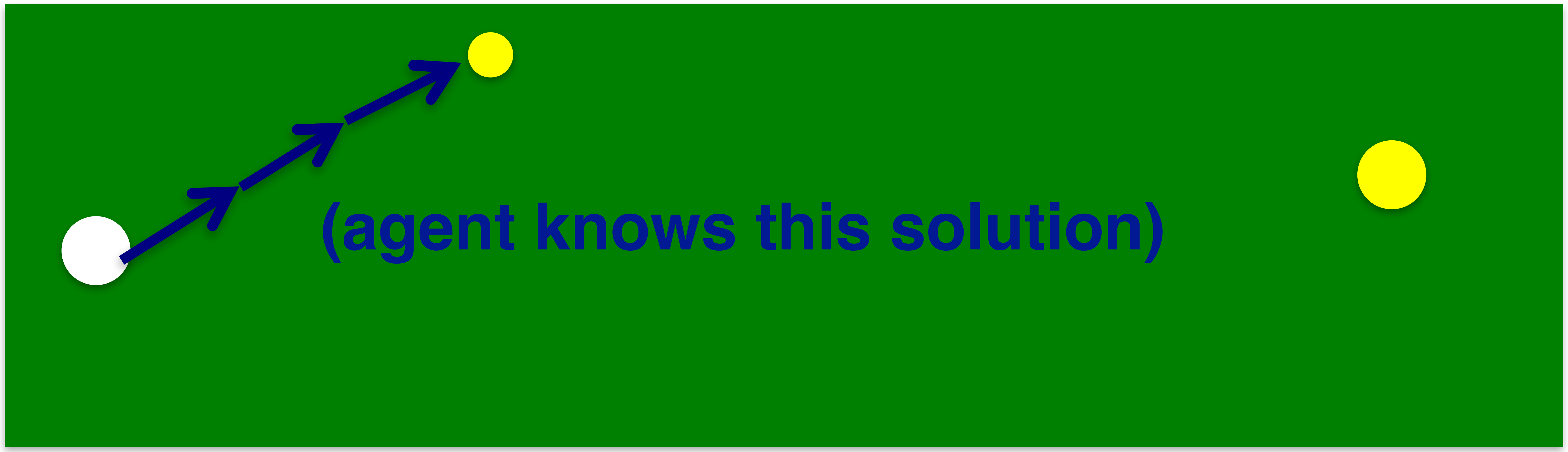


Dataset

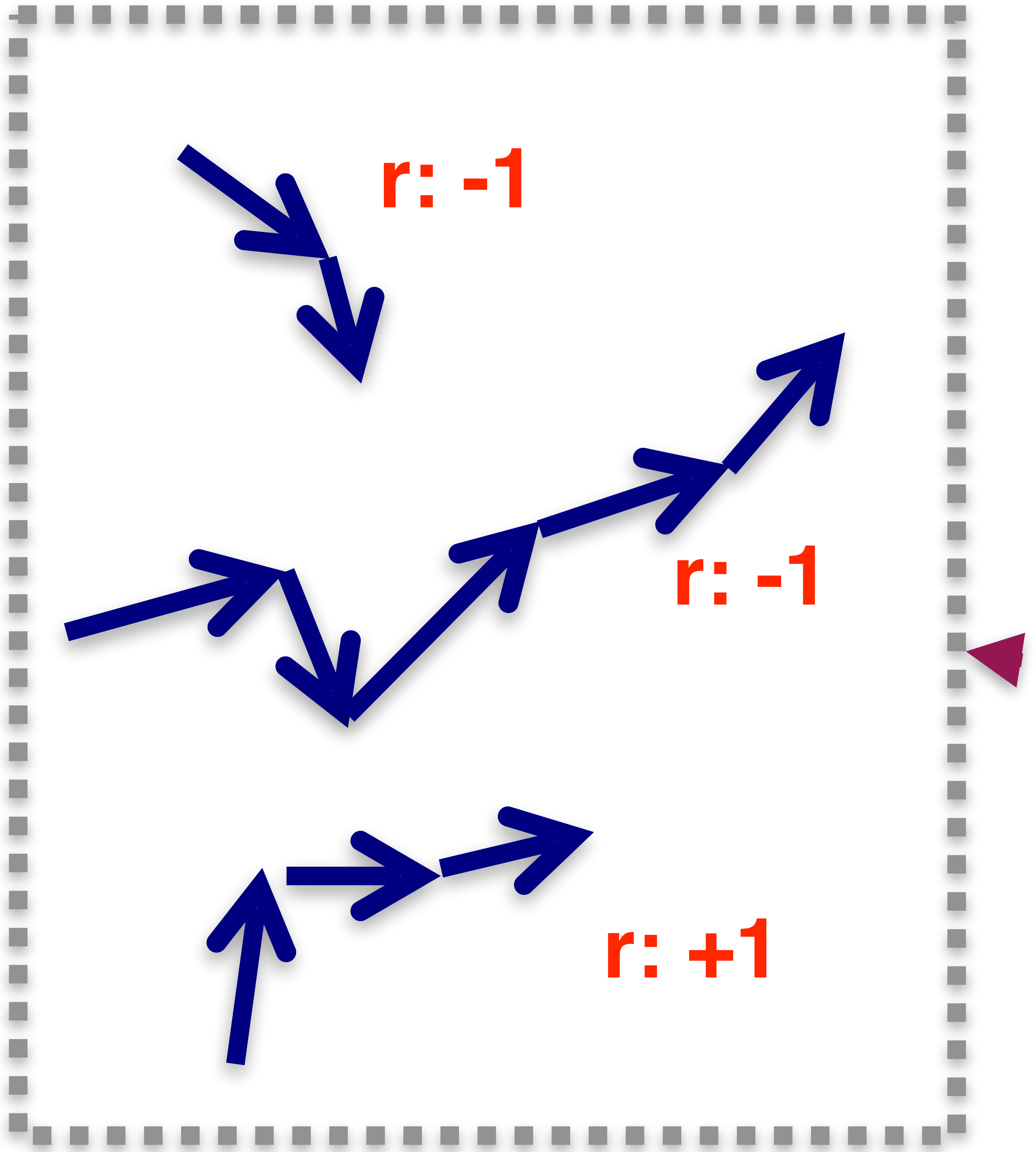


$$\sum_{t=1}^{T=100} r_t \qquad a_t = \pi_\theta(s_{0:t})$$

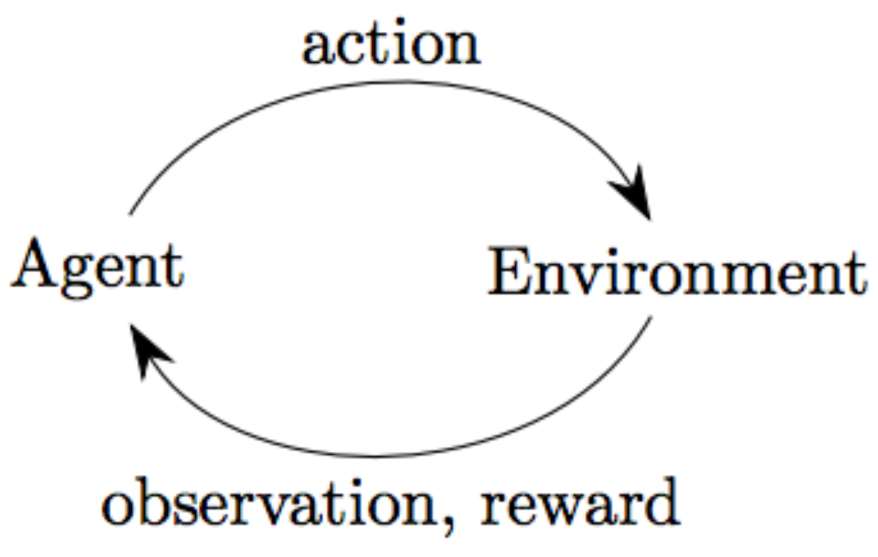




Dataset

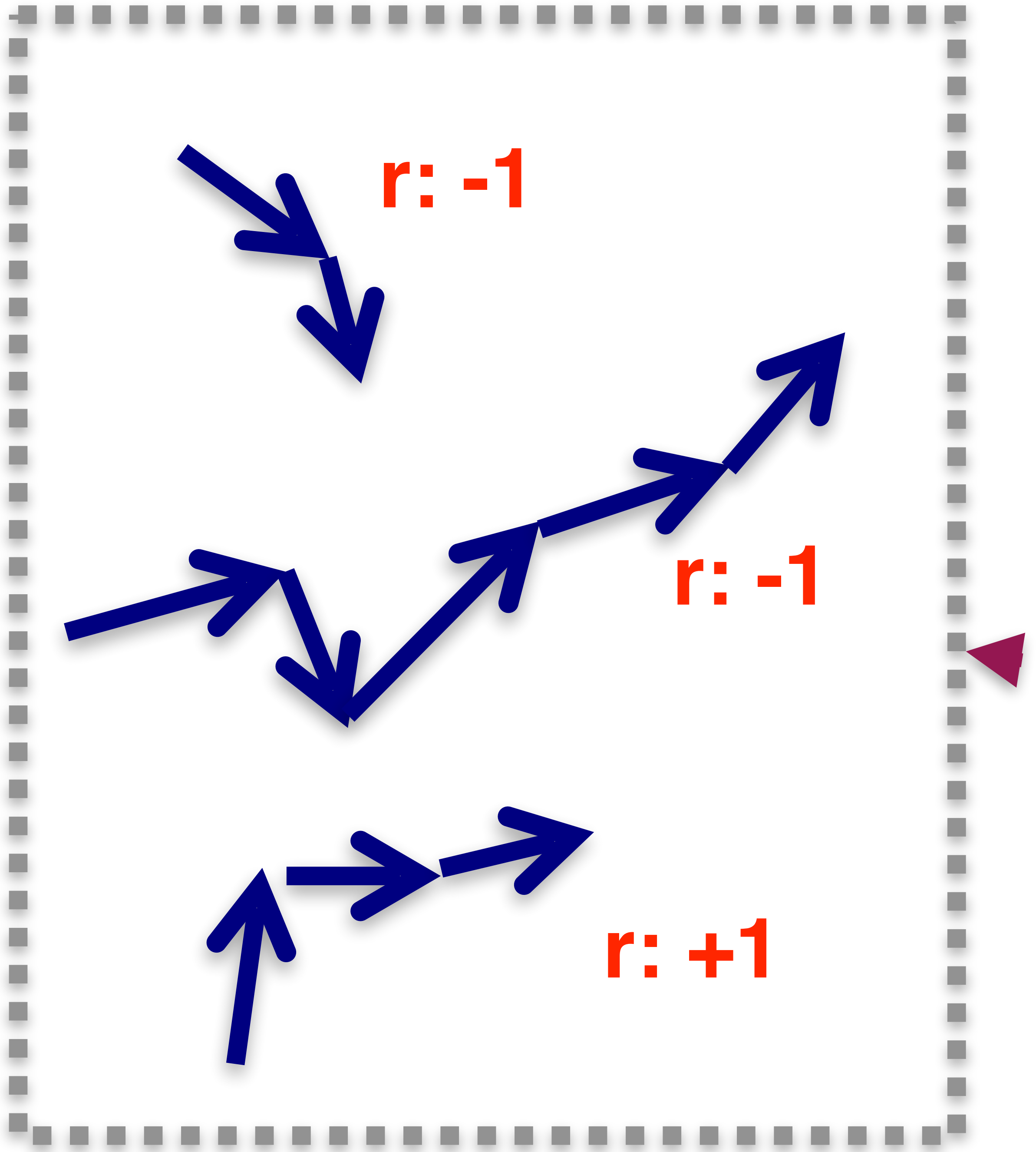


$$\sum_{t=1}^{T=100} r_t \qquad a_t = \pi_\theta(s_{0:t})$$

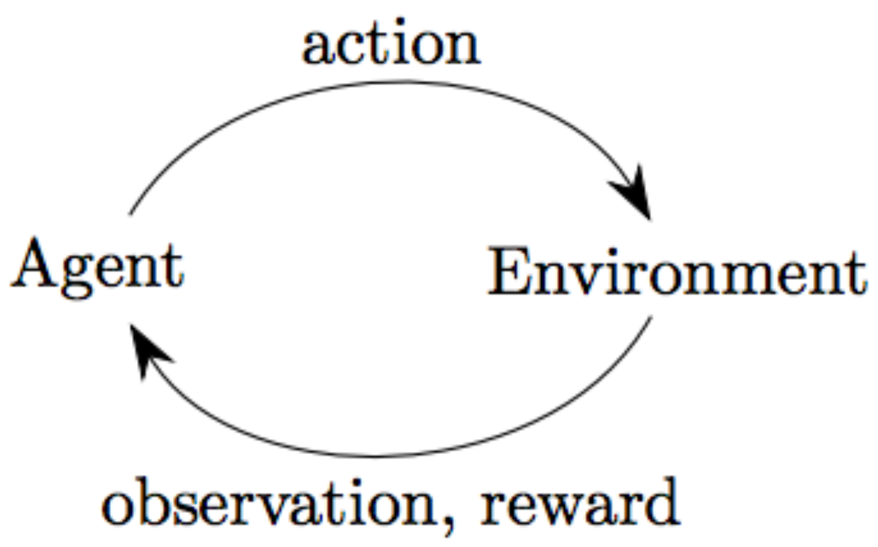


time t

Dataset

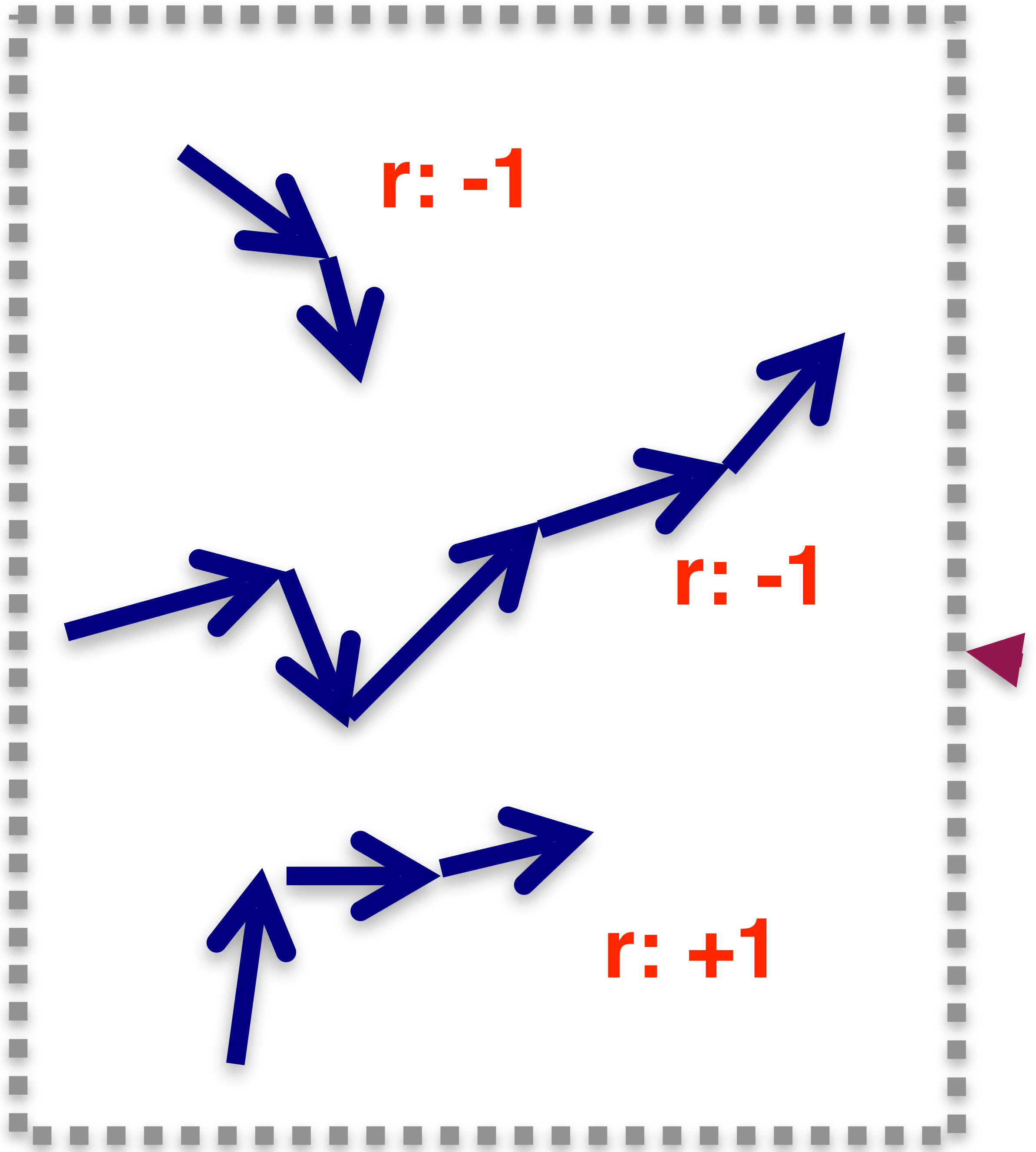


$$\sum_{t=1}^{T=100} r_t \qquad a_t = \pi_\theta(s_{0:t})$$

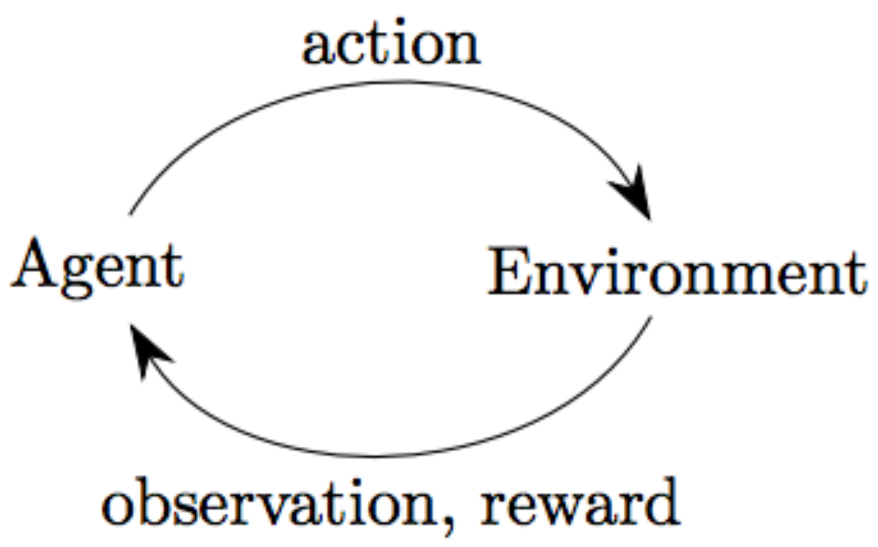


time t **t=97**

Dataset

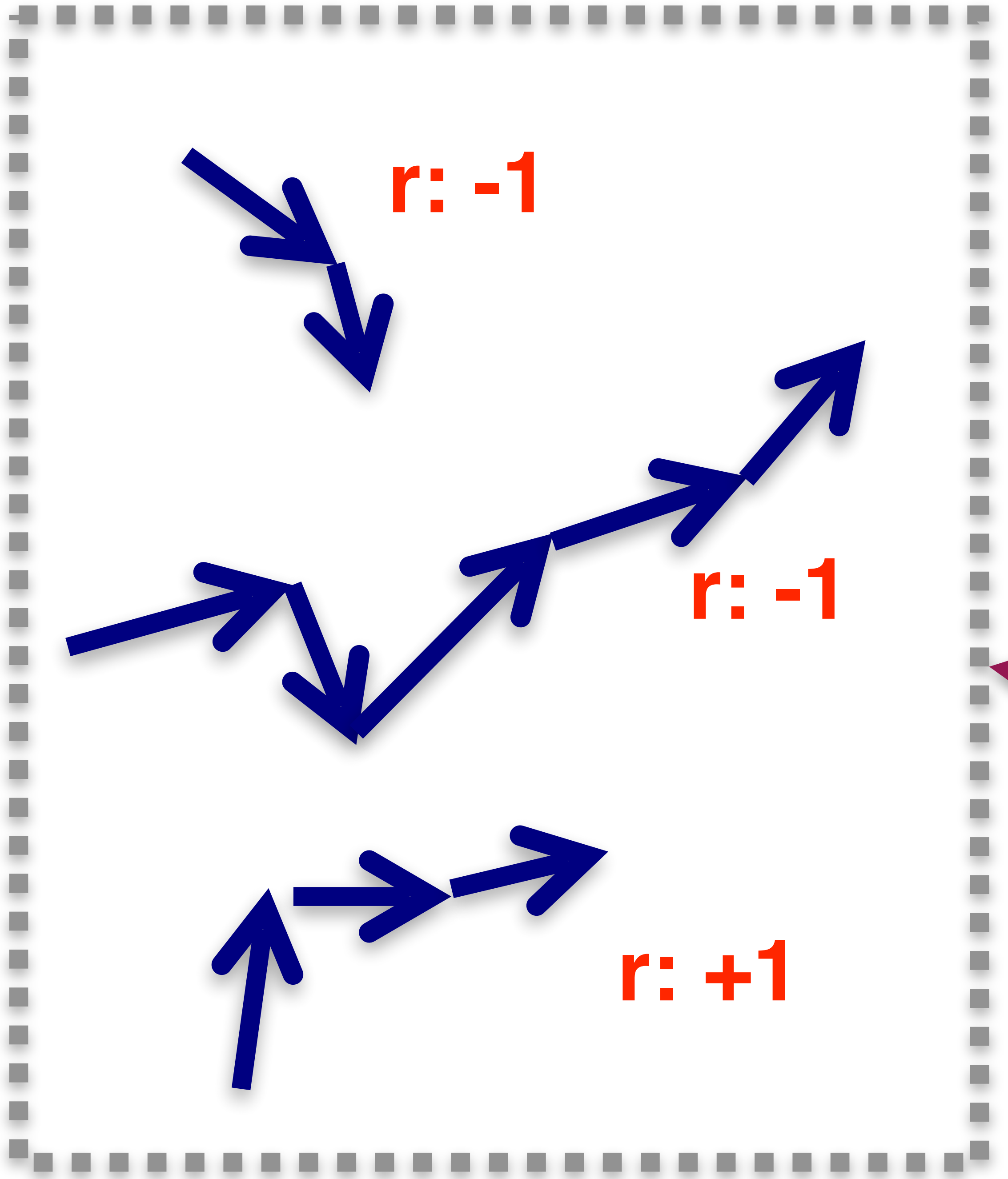


$$\sum_{t=1}^{T=100} r_t \qquad a_t = \pi_\theta(s_{0:t})$$



time t **t=97** **t=10**

Dataset

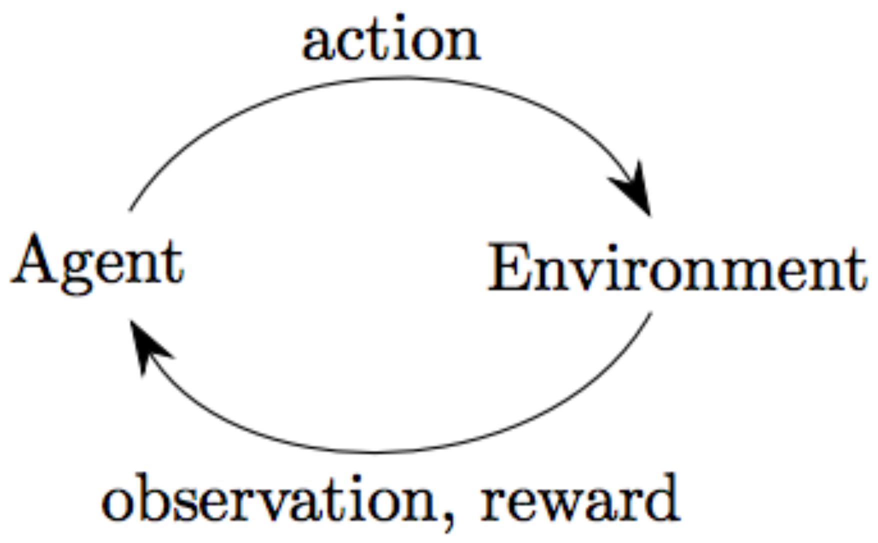


$$\sum_{t=1}^{T=100} r_t$$

$$a_t = \pi_\theta(s_{0:t})$$

$$a_t = \pi_\theta(s_{0:t}, T - t)$$

(is this a problem?)



time t **t=97** **t=10**

Goal

$$a_t = \pi_\theta(s_{0:t}) \qquad s.t. \boxed{\max \sum_t r_t}$$

Finite Time Horizon

$$\sum_{t=1}^{T} r_t \dashrightarrow a_t = \pi_\theta(s_{0:t}, T - t)$$

Infinite Time Horizon

$$\sum_t r_t \dashrightarrow \sum_t \gamma^t r_t \dashrightarrow a_t = \pi_\theta(s_{0:t})$$

$$0 < \gamma < 1$$

discount factor

Commonly Used

Maximizing Rewards

$$a_t = \pi_\theta(s_{1:t})$$

$$\tau = (s_1, a_1, r_1, s_2, a_2, r_2, \ldots) \dashrightarrow p_\theta(\tau)$$

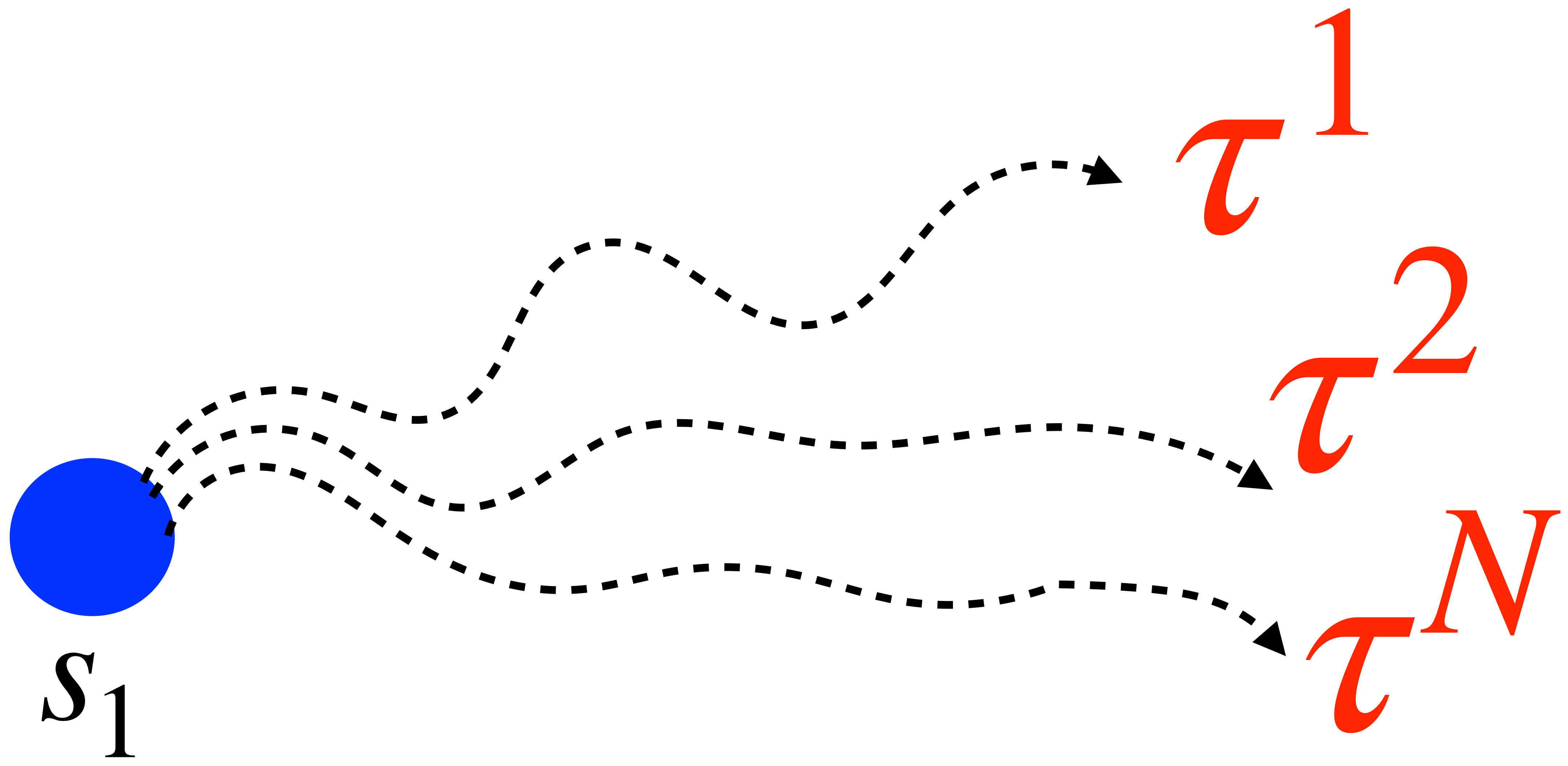
Why do we need probability of a rollout?

Rollouts from the same state can be different

Stochastic Environment

Stochastic Rewards

Stochastic Policy



Need for stochastic policy

- Two–player game of rock–paper–scissors
  - Scissors beats paper
  - Rock beats scissors
  - Paper beats rock
- Consider policies for **iterated** rock–paper–scissors
  - A deterministic policy is **easily exploited**
  - A **uniform random policy** is optimal (i.e., Nash equilibrium)

Slide adapted from Marcello Restelli

Policy Optimization

$$a_t = \pi_\theta(s_{1:t})$$

$$\tau = (s_1, a_1, r_1, s_2, a_2, r_2, \ldots) \dashrightarrow p_\theta(\tau)$$

$$R(\tau) = \sum_t r_t$$

Average reward

$$\sum_\tau p_\theta(\tau) R(\tau) = E_\tau\left[R(\tau)\right]$$

Maximize Reward

Policy Gradients!

$$\max_\theta E_\tau\left[R(\tau)\right] \dashrightarrow E_\tau[\nabla_\theta\left(\log p_\theta(\tau)\right) R(\tau)]$$

POLICY GRADIENTS

Deriving Policy Gradient

$$\max_{\theta} E\tau[R(\tau)]$$

Deriving Policy Gradient

$$\max_{\theta} E\tau[R(\tau)]$$

$$\nabla_{\theta} E\tau[R(\tau)]$$

Deriving Policy Gradient

$$\max_\theta E\tau[R(\tau)]$$

$$\nabla_\theta E\tau[R(\tau)]$$

$$\nabla_\theta \int p_\theta(\tau)R(\tau)d\tau$$

Deriving Policy Gradient

$$\max_{\theta} E\tau[R(\tau)]$$

$$\nabla_\theta E\tau[R(\tau)]$$

$$\nabla_\theta \int p_\theta(\tau) R(\tau) d\tau$$

$$\int \nabla_\theta (p_\theta(\tau)) R(\tau) d\tau$$

(Leibniz Integral Rule)

Deriving Policy Gradient

$$\max_{\theta} E\tau[R(\tau)]$$

$$\nabla_{\theta} E\tau[R(\tau)]$$

$$\nabla_{\theta} \int p_{\theta}(\tau) R(\tau) d\tau$$

$$\int \nabla_{\theta}(p_{\theta}(\tau)) R(\tau) d\tau$$

$$\int \boxed{p_{\theta}(\tau)} \frac{\nabla_{\theta}(p_{\theta}(\tau))}{\boxed{p_{\theta}(\tau)}} R(\tau) d\tau$$

Deriving Policy Gradient

$$\max_{\theta} E\tau[R(\tau)]$$

$$\nabla_{\theta} E\tau[R(\tau)]$$

$$\nabla_{\theta} \int p_{\theta}(\tau) R(\tau) d\tau$$

$$\int \nabla_{\theta}(p_{\theta}(\tau)) R(\tau) d\tau$$

$$\int p_{\theta}(\tau) \boxed{\frac{\nabla_{\theta}(p_{\theta}(\tau))}{p_{\theta}(\tau)}} R(\tau) d\tau$$

Deriving Policy Gradient

$$\max_{\theta} E\tau[R(\tau)]$$

$$\nabla_\theta E\tau[R(\tau)]$$

$$\nabla_\theta \int p_\theta(\tau) R(\tau) d\tau$$

$$\int \nabla_\theta (p_\theta(\tau)) R(\tau) d\tau$$

$$\int p_\theta(\tau) \boxed{\frac{\nabla_\theta (p_\theta(\tau))}{p_\theta(\tau)}} R(\tau) d\tau$$

$$\int p_\theta(\tau) \boxed{\nabla_\theta (\log p_\theta(\tau))} R(\tau) d\tau$$

Deriving Policy Gradient

$$\max_{\theta} E\tau[R(\tau)]$$

$$\nabla_{\theta} E\tau[R(\tau)]$$

$$\nabla_{\theta} \int p_{\theta}(\tau) R(\tau) d\tau$$

$$\int \nabla_{\theta}(p_{\theta}(\tau)) R(\tau) d\tau$$

$$\int p_{\theta}(\tau) \frac{\nabla_{\theta}(p_{\theta}(\tau))}{p_{\theta}(\tau)} R(\tau) d\tau$$

$$\int p_{\theta}(\tau) \nabla_{\theta}(\log p_{\theta}(\tau)) R(\tau) d\tau$$

$$E\tau[\nabla_{\theta}(\log p_{\theta}(\tau)) R(\tau)]$$

Policy Gradients

$$E\tau[\nabla_\theta(\log p_\theta(\tau))R(\tau)]$$

Intuitive Interpretation

Roll out multiple trajectories

Policy Gradients

$$E\tau[\nabla_\theta(\log p_\theta(\tau))R(\tau)]$$

Intuitive Interpretation

Roll out multiple trajectories

Increase the log-prob of trajectories that result in high rewards!

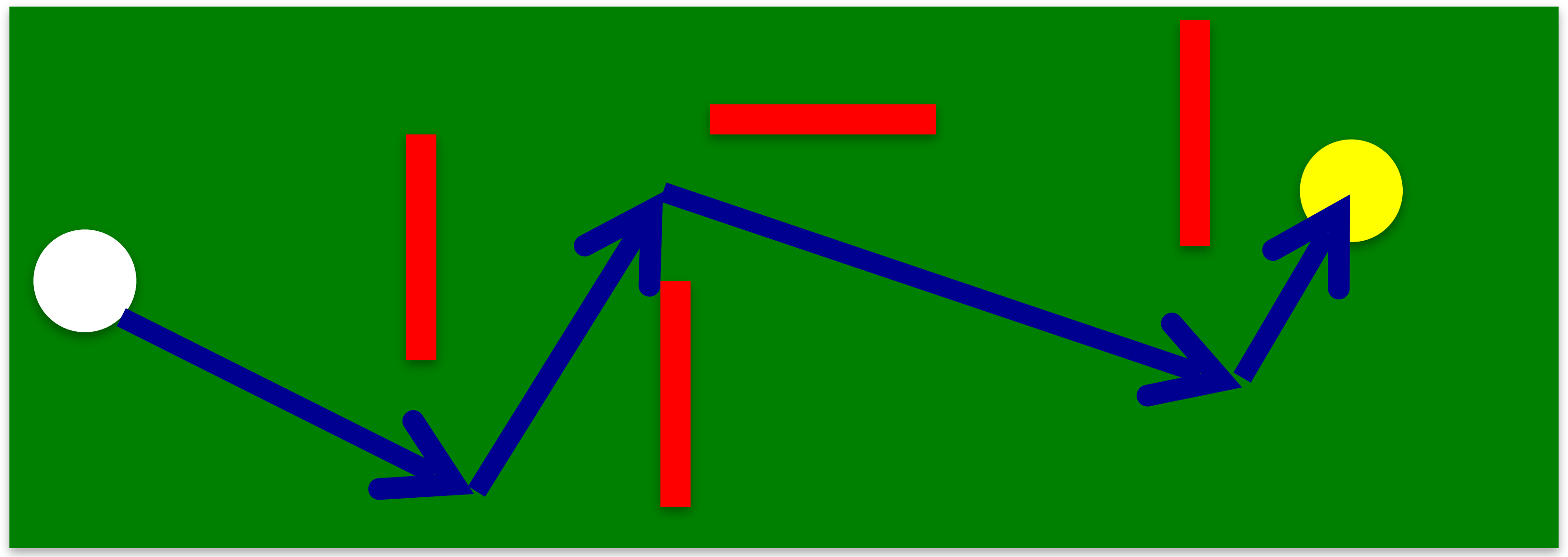
Policy Gradients

$$E\tau[\nabla_\theta(\log p_\theta(\tau))R(\tau)]$$

Intuitive Interpretation

Roll out multiple trajectories

Increase the log-prob of trajectories that result in high rewards!



Increase log-prob

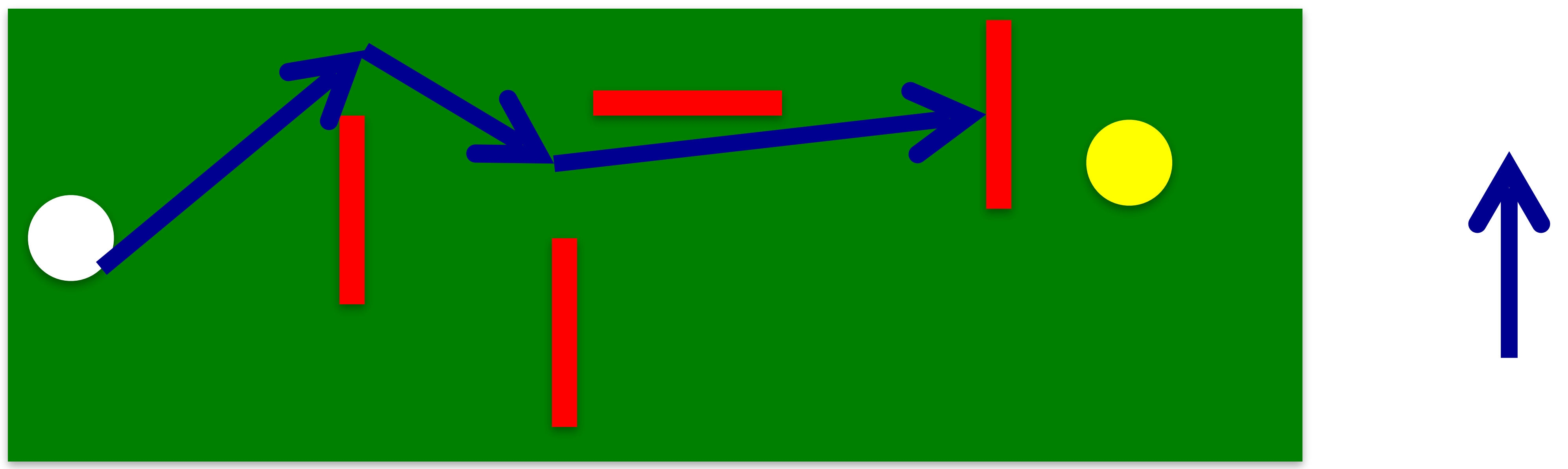
Policy Gradients

$$E\tau[\nabla_\theta(\log p_\theta(\tau))R(\tau)]$$

Intuitive Interpretation

Roll out multiple trajectories

Increase the log-prob of trajectories that result in high rewards!



Increase
log-prob by
small
amount

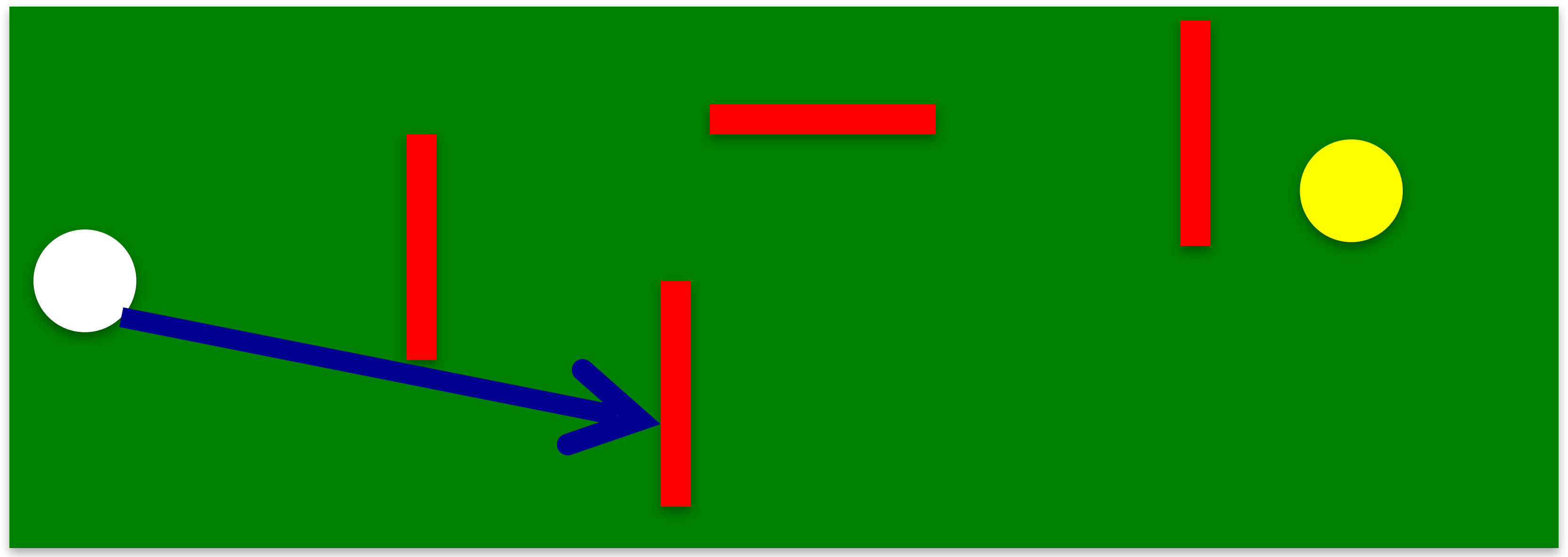
Policy Gradients

$$E\tau[\nabla_\theta(\log p_\theta(\tau))R(\tau)]$$

Intuitive Interpretation

Roll out multiple trajectories

Increase the log-prob of trajectories that result in high rewards!



Increase log-prob by smaller amount

Expanding on Policy Gradients

$$E_\tau[\nabla_\theta\left(\log p_\theta(\tau)\right)R(\tau)]$$

$$E_\tau\left[\sum_t\left(\nabla_\theta \log p_\theta(a_t \mid s_{1:t}, a_{1:t-1})\right)R(\tau)\right]$$

$$E_\tau\left[\sum_t\left(\nabla_\theta \log \pi_\theta(a_t \mid s_{1:t}, a_{1:t-1})\right)R(\tau)\right]$$

Does something feel off?

NO dependence on $p\left(s_t \mid s_{1:t-1}, a_{1:t-1}\right)$

Expanding on Policy Gradients

$$E_\tau[\nabla_\theta\left(\log p_\theta(\tau)\right)R(\tau)]$$

$$E_\tau\left[\sum_{t=1}^{T}\left(\nabla_\theta \log p_\theta(a_t \mid s_{1:t}, a_{1:t-1})\right)R(\tau)\right]$$

Model Free!

$$E_\tau\left[\sum_{t=1}^{T}\left(\nabla_\theta \log \pi_\theta(a_t \mid s_{1:t}, a_{1:t-1}; \theta)\right)R(\tau)\right]$$

Does something feel off?

NO dependence on $p\left(s_t \mid s_{1:t-1}, a_{1:t-1}\right)$

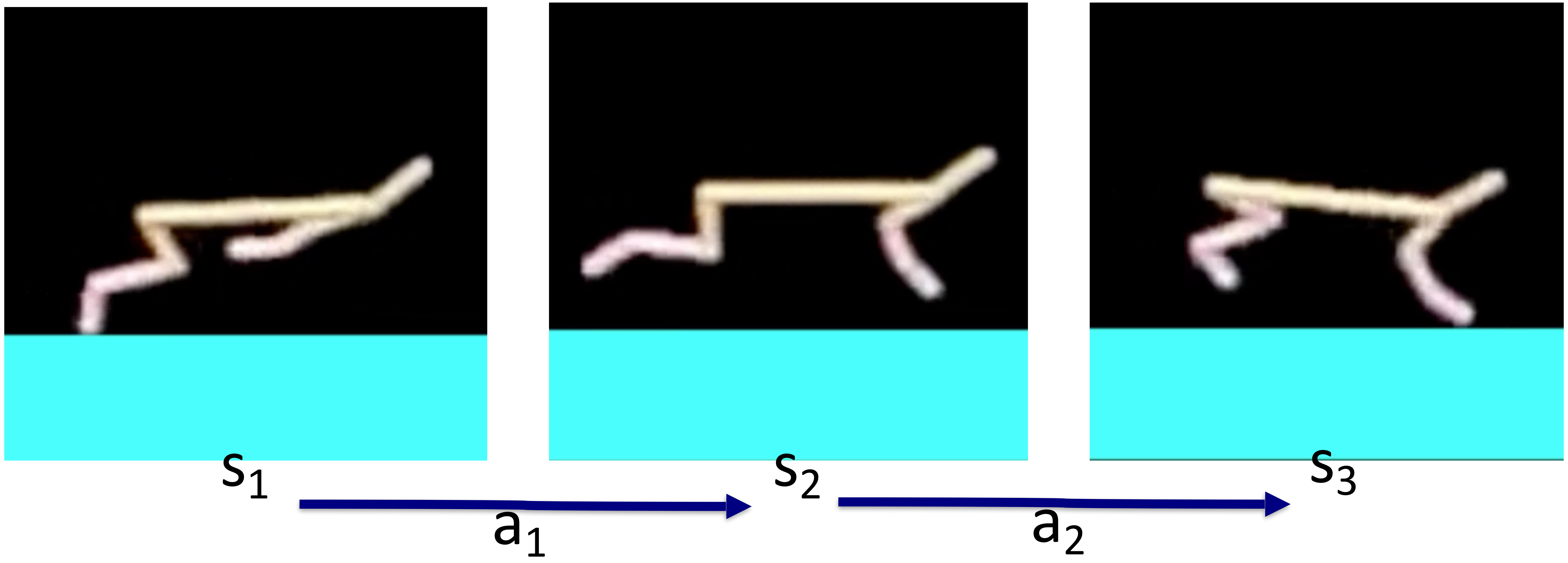
Expanding on Policy Gradients

$$E_\tau\left[\sum_{t=1}\left(\nabla_\theta \log \pi_\theta(a_t \mid s_{1:t}, a_{1:t-1})\right)R(\tau)\right]$$

Markov assumption not necessary!

With Markov Assumption (discuss this later in detail)

$$E_\tau\left[\sum_{t=1}\left(\nabla_\theta \log \pi_\theta(a_t \mid s_t)\right)R(\tau)\right]$$



State (s_1, s_2 …)

- Location/rotation of joints
- Or, the image
- Or, both

Action (a_1, a_2 …)

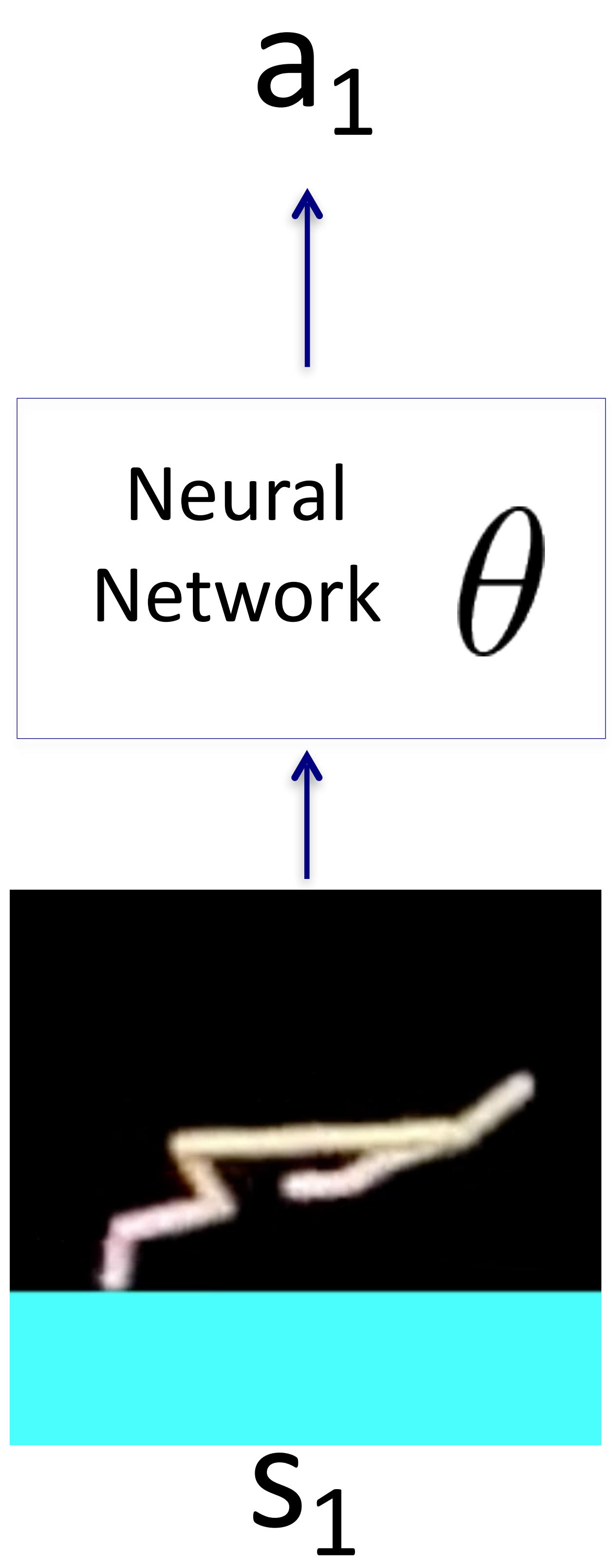
desired joint position

Rewards (r_1, r_2 …)

Speed of the Cheetah

Illustration of Policy Gradients

$$\boxed{E_\tau \Big[} \sum_{t=1} \left(\nabla_\theta \log \pi_\theta(a_t \mid s_t) \right) R(\tau) \Big]$$



$$\boxed{\frac{1}{N} \sum_{i=1}^{N}} \left(\sum_{t=1}^{T} \left(\nabla_\theta \log \pi_\theta\left(a_t^i \mid s_t^i\right) \right) R(\tau) \right)$$

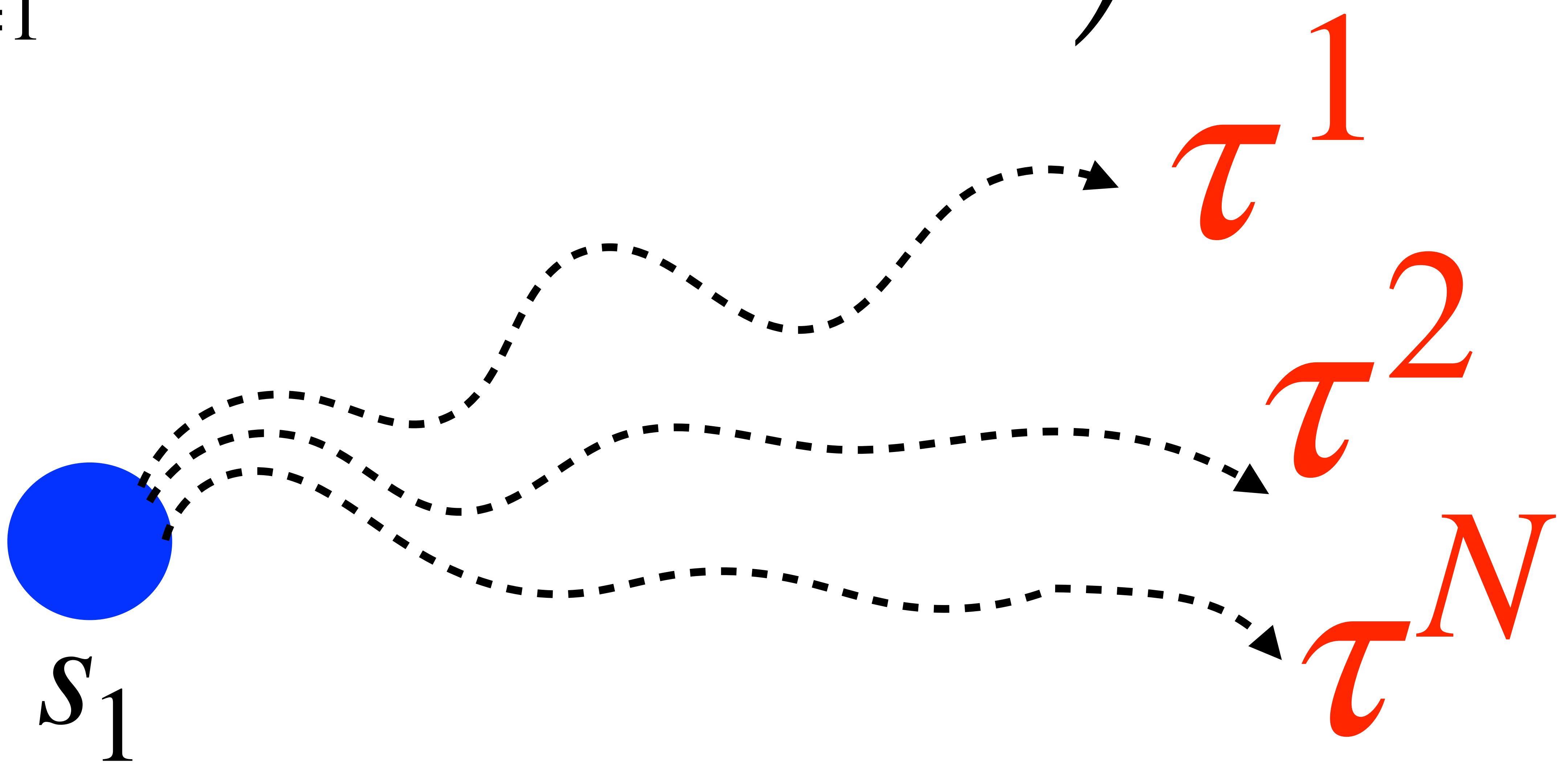
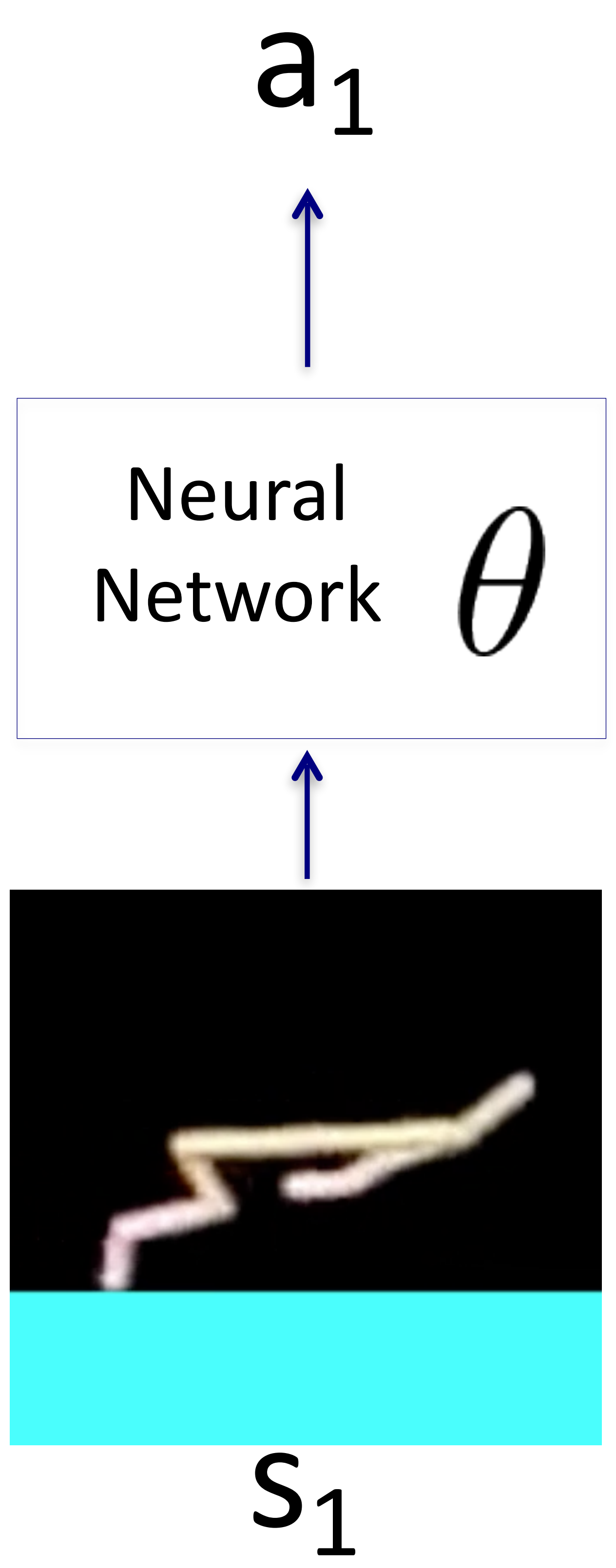


Illustration of Policy Gradients

$$E_\tau\left[\sum_{t=1} \left(\nabla_\theta \log \pi_\theta(a_t \mid s_t)\right) R(\tau)\right]$$

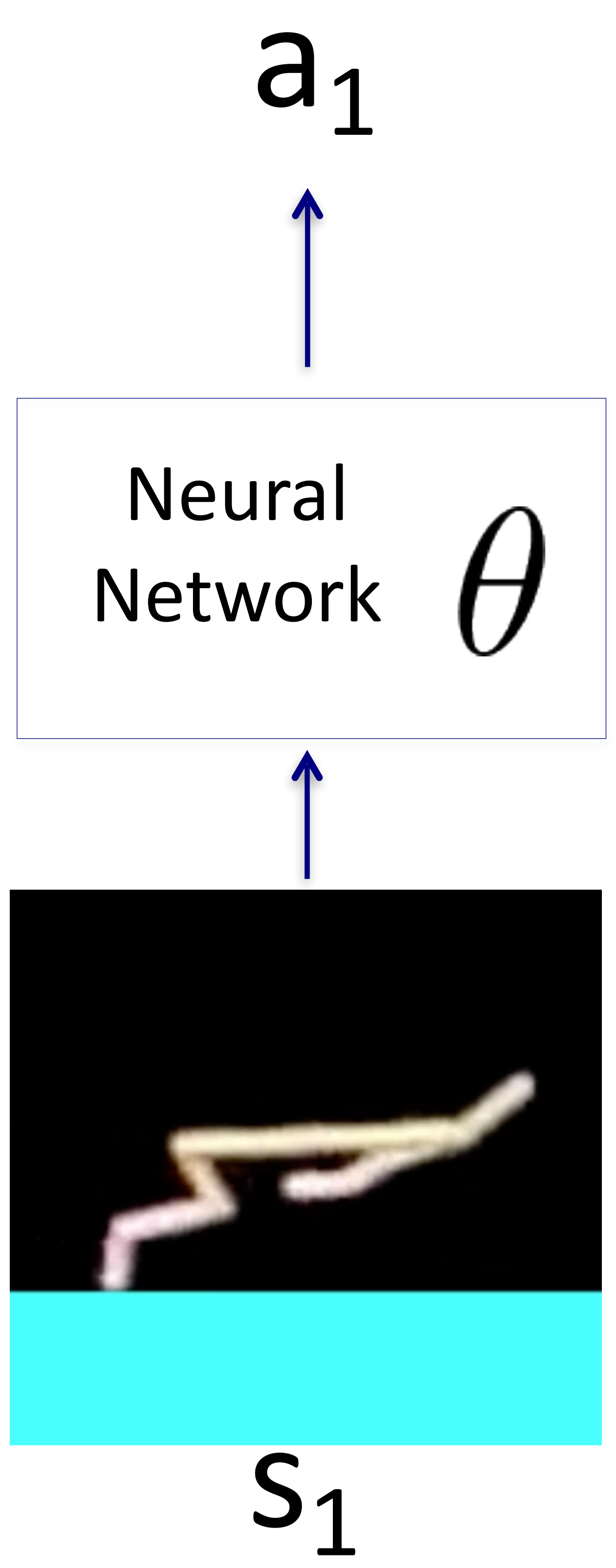


$$\frac{1}{N}\sum_{i=1}^{N}\left(\boxed{\sum_{t=1}^{T}}\left(\nabla_\theta \log \pi_\theta\left(a_t^i \mid s_t^i\right)\right)R(\tau)\right)$$

in practice can't roll out until infinity

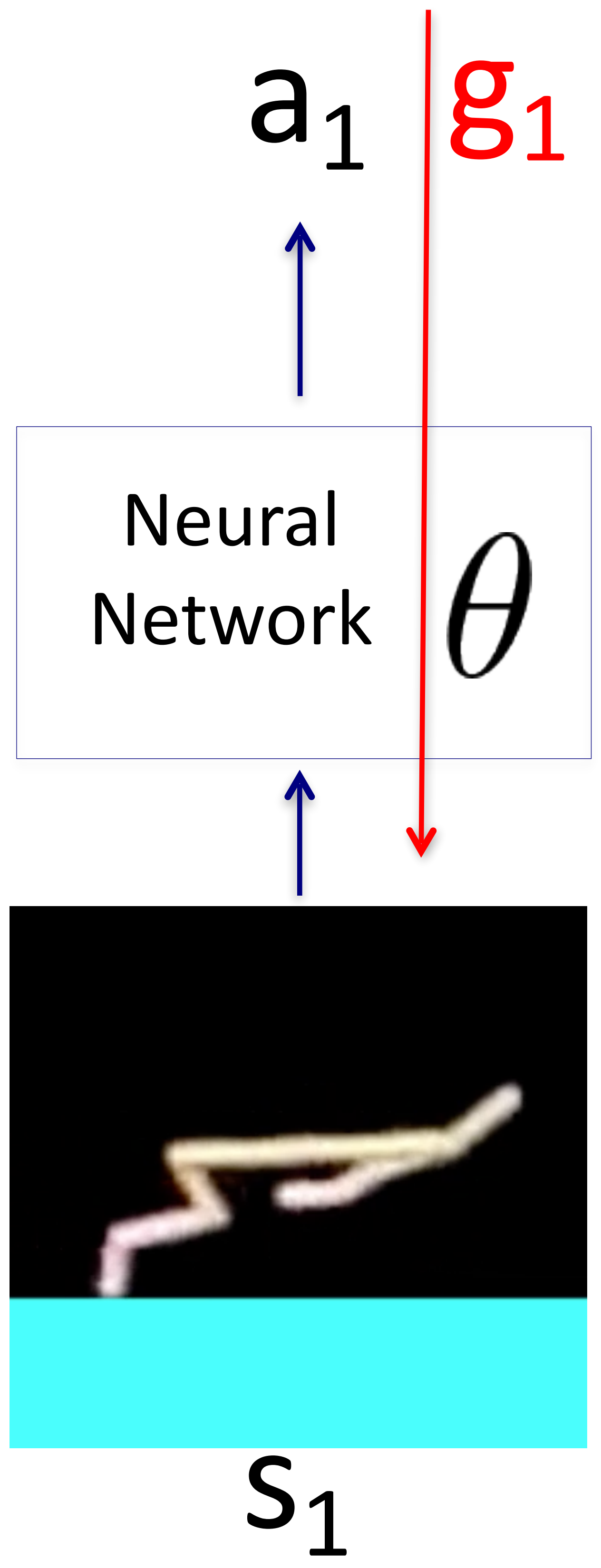
Treat **finite** horizon as **infinite** horizon with discount

Illustration of Policy Gradients



$$\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\left(\nabla_{\theta}\log\pi_{\theta}\left(a_t^i \mid s_t^i\right)\right)R^{\gamma}(\tau)\right)$$

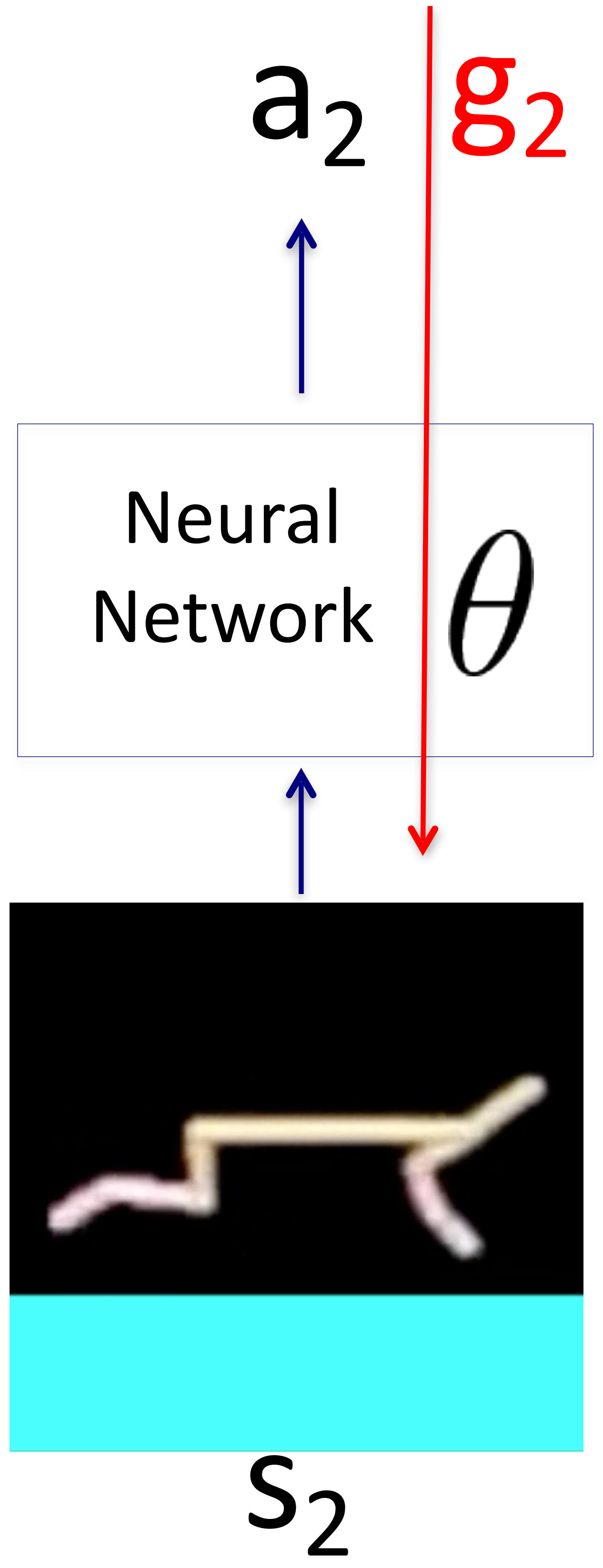
Illustration of Policy Gradients



$$\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\left(\nabla_\theta \log \pi_\theta(a_t^i \mid s_t^i)\right)R^\gamma(\tau)\right)$$

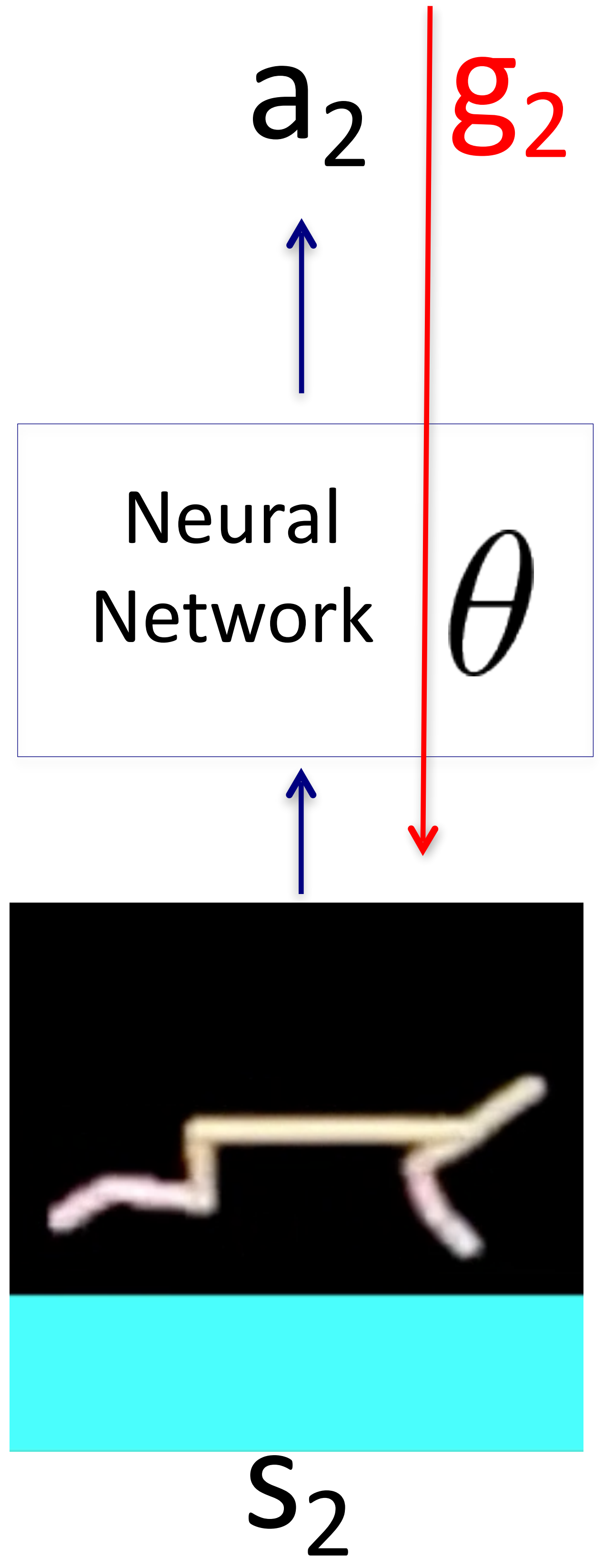
$$g_1 = \nabla_\theta \log \pi_\theta(a_1 | s_1)$$

Illustration of Policy Gradients



$$\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\left(\nabla_\theta \log \pi_\theta\left(a_t^i \mid s_t^i\right)\right)R^\gamma(\tau)\right)$$

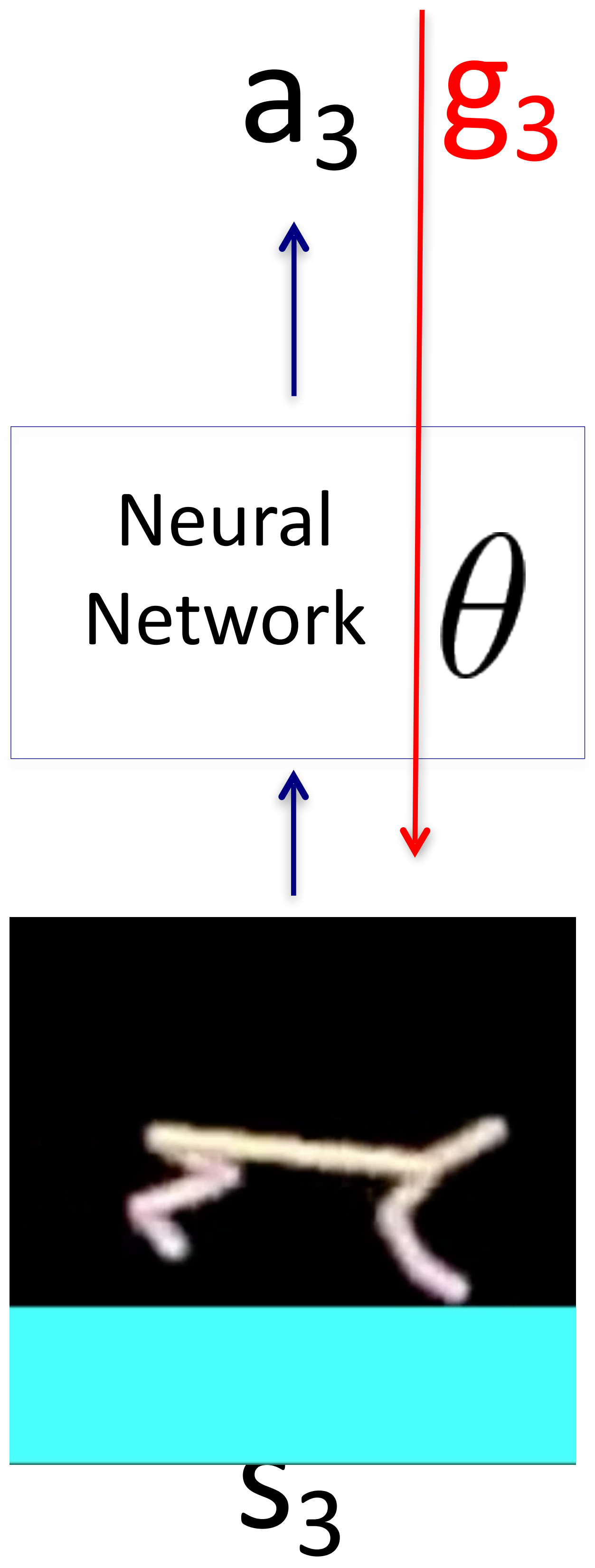
Illustration of Policy Gradients



$$\frac{1}{N}\sum_{i=1}^{N}\left(\underbrace{\boxed{\sum_{t=1}^{T}\left(\nabla_\theta \log \pi_\theta\left(a_t^i \mid s_t^i\right)\right)}}_{G} R^\gamma(\tau)\right)$$

$G = g_1 + g_2$

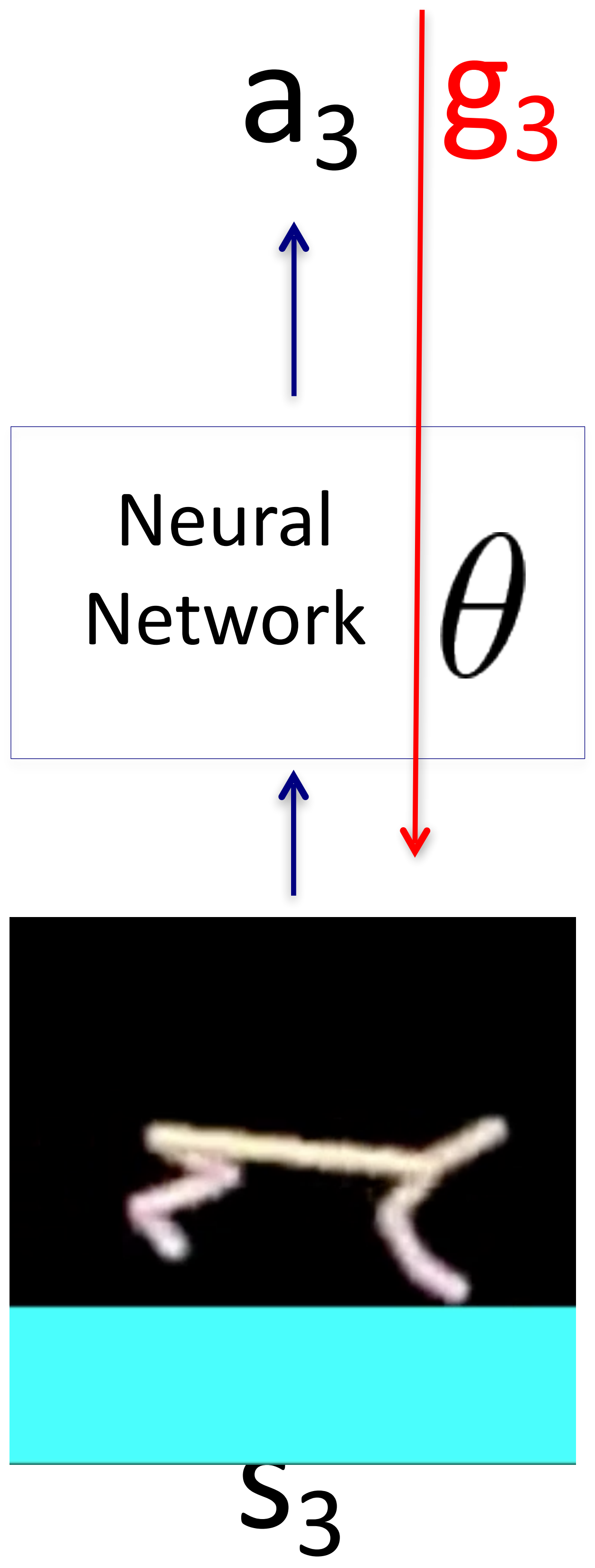
Illustration of Policy Gradients



$$\frac{1}{N}\sum_{i=1}^{N}\left(\underbrace{\boxed{\sum_{t=1}^{T}\left(\nabla_\theta \log \pi_\theta(a_t^i \mid s_t^i)\right)}}_{G} R^\gamma(\tau)\right)$$

$G = g_1 + g_2 + g_3$

Illustration of Policy Gradients



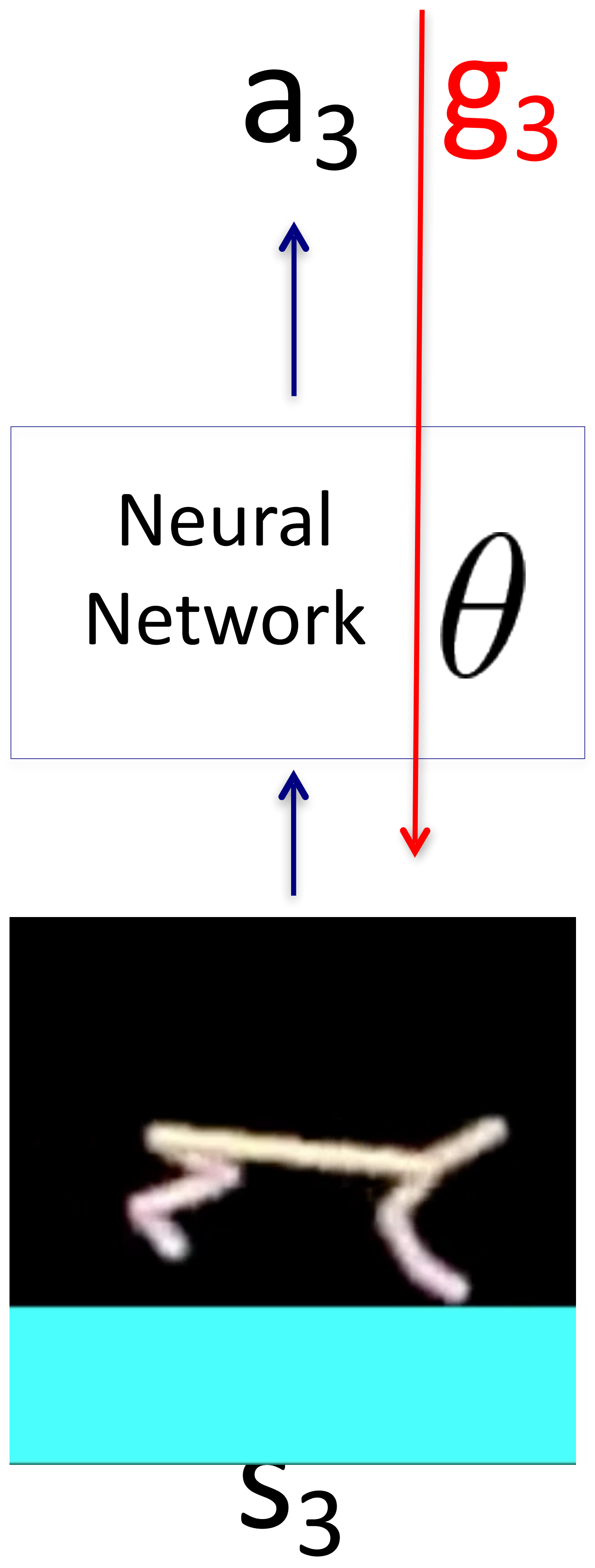
$$\frac{1}{N}\sum_{i=1}^{N}\left(\underbrace{\boxed{\sum_{t=1}^{T}\left(\nabla_\theta \log \pi_\theta\left(a_t^i \mid s_t^i\right)\right)}}_{G}\underbrace{\boxed{R^\gamma(\tau)}}_{V}\right)$$

Sum of velocities across time

$G = g_1 + g_2 + g_3$

Illustration of Policy Gradients

This is also called the REINFORCE Algorithm



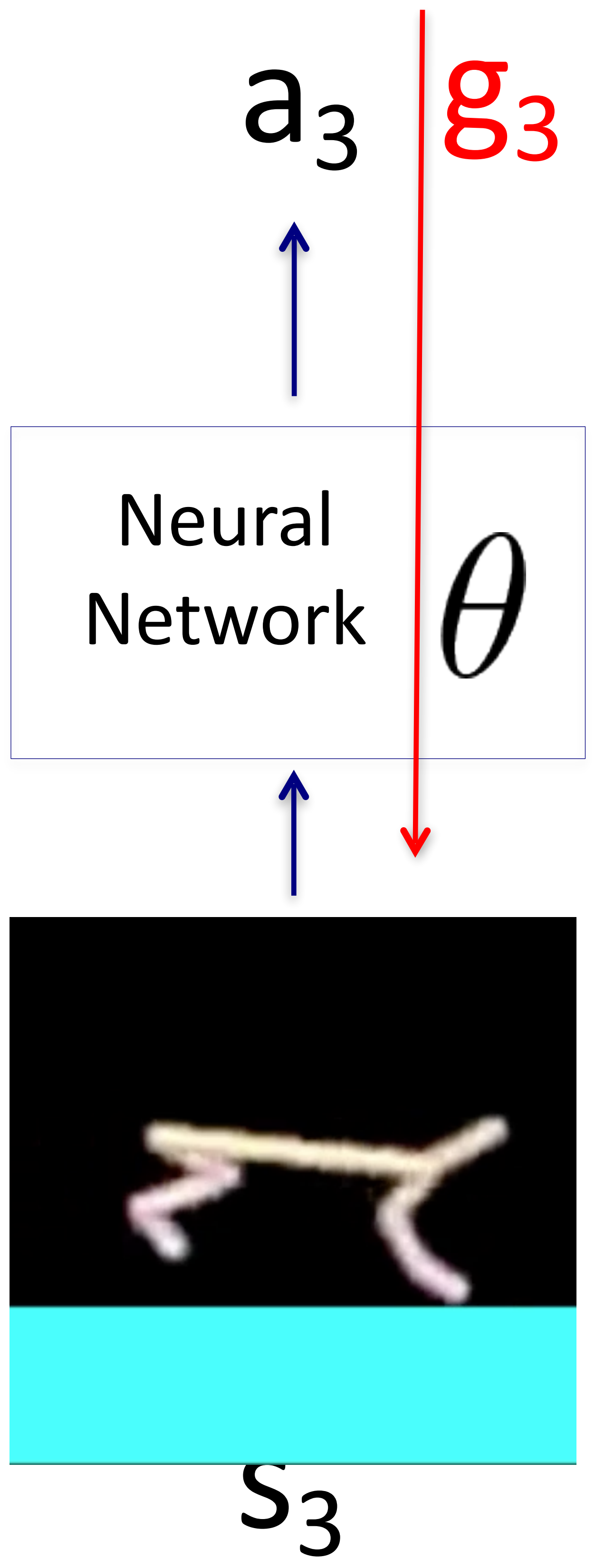
$$\frac{1}{N}\sum_{i=1}^{N}\left(\underbrace{\boxed{\sum_{t=1}^{T}\left(\nabla_\theta \log \pi_\theta\left(a_t^i \mid s_t^i\right)\right)}}_{G}\underbrace{\boxed{R^\gamma(\tau)}}_{v}\right)$$

Gradient Ascent

$$\theta(t+1) = \theta(t) + \alpha(vG)$$

Illustration of Policy Gradients

Discrete Action Space
Multinomial Policy

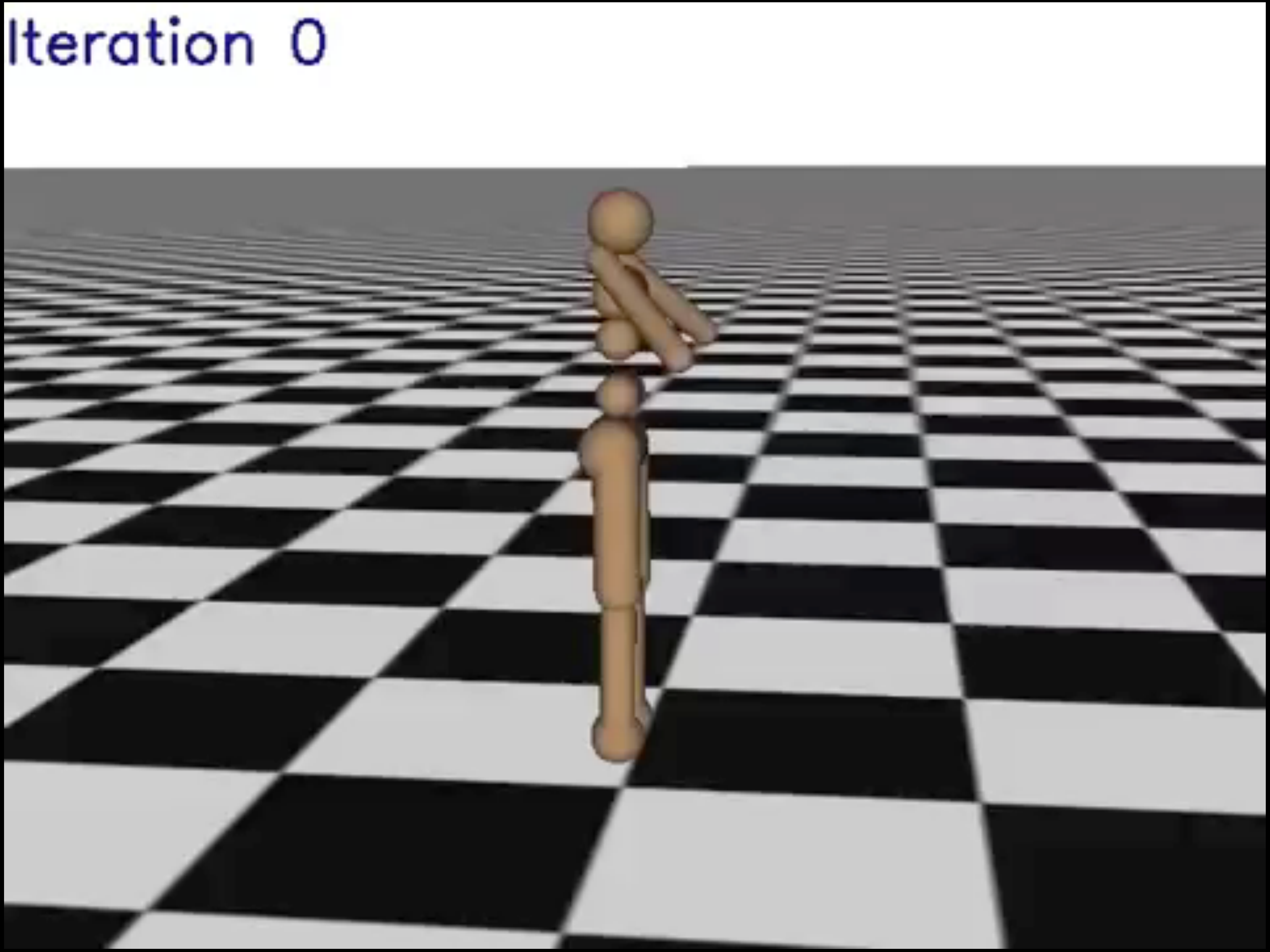


$$\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\left(\nabla_\theta \log \pi_\theta(a_t^i \mid s_t^i)\right)R^\gamma(\tau)\right)$$

Continuous Action Space
Gaussian Policy

Comparing with Supervised Learning

RL	Supervised Learning
$\sum_t r_t$	$\tau^{gt} = (s_1, a_1^{gt}, s_2, a_2^{gt}, \ldots)$
$E_\tau[\nabla_\theta \left(\log p_\theta(\tau)\right) R(\tau)]$	$E_{\tau^{gt}}[\nabla_\theta \left(\log p_\theta(\tau^{gt})\right)]$
Policy Gradients	Maximum Likelihood



High-Dimensional Continuous Control Using Generalized Advantage Estimation, Schulman et al., 2015

The Idea of Episode

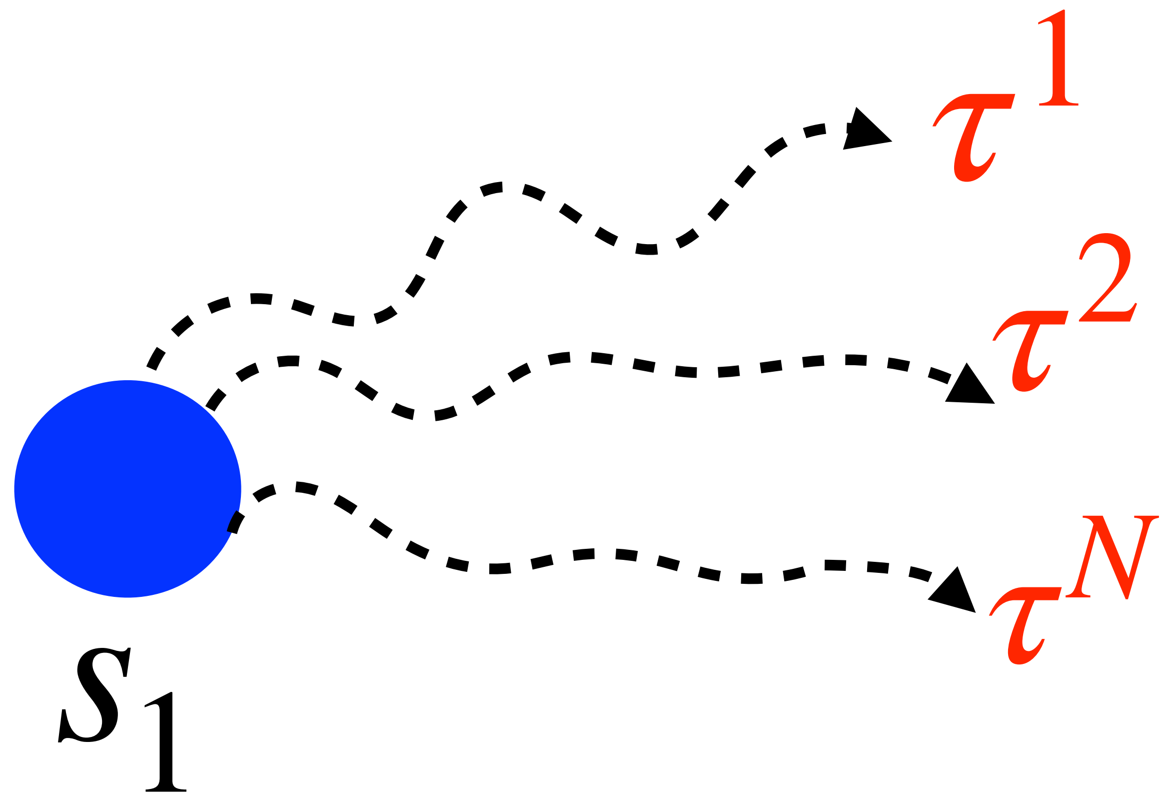
$$E_\tau[\nabla_\theta(\log p_\theta(\tau))R(\tau)]$$

$$\frac{1}{N}\sum_{i=1}^{N}\left(\boxed{\sum_{t=1}^{T}\left(\nabla_\theta \log \pi_\theta(a_t^i \mid s_t^i)\right)R(\tau)}\right)$$

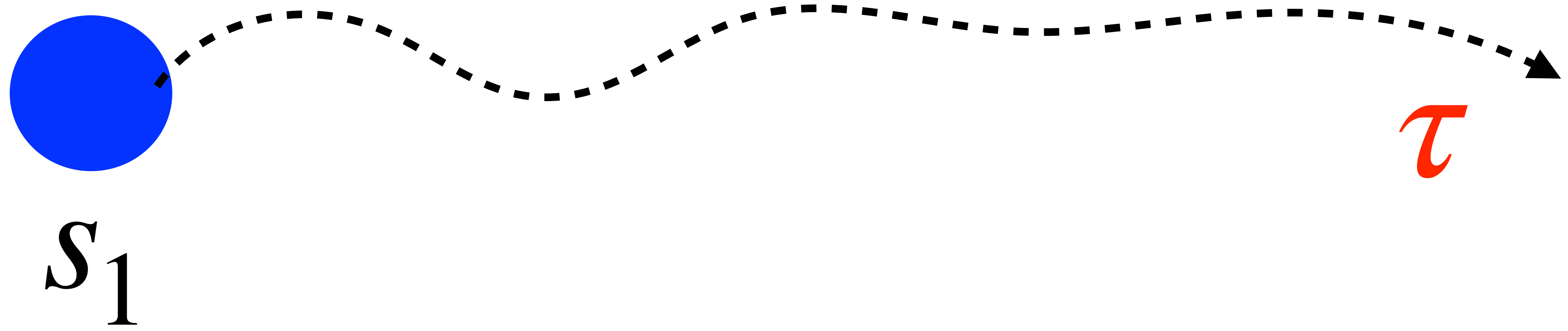
$$\frac{1}{N}\left(\sum_{t=1}^{NT}\left(\nabla_\theta \log \pi_\theta(a_t \mid s_t)\right)R(\tau)\right)$$

One Episode

Why define episodes?



N Episodes



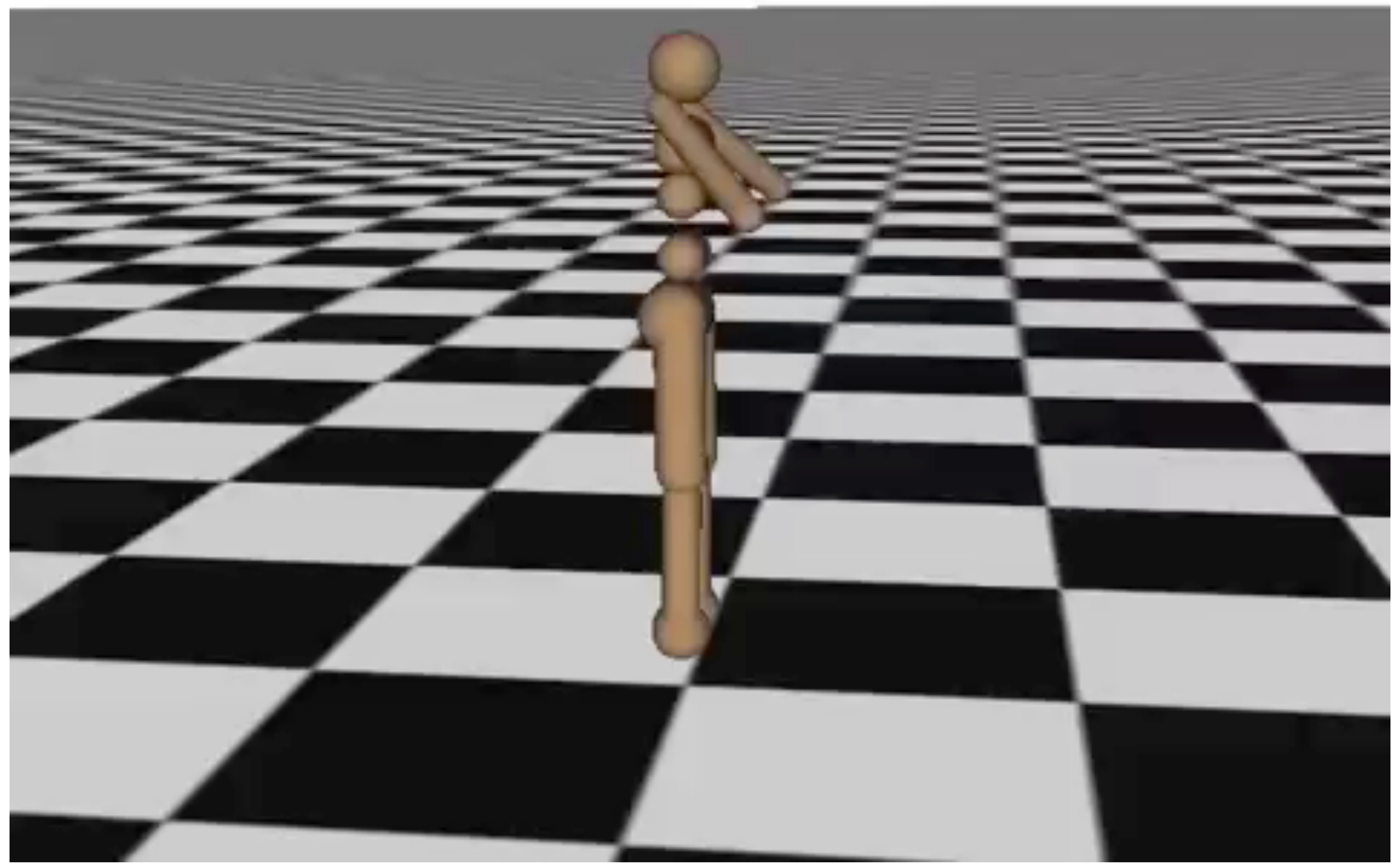
The Idea of Episode

One Episode

$$\frac{1}{N}\sum_{i=1}^{N}\left(\boxed{\sum_{t=1}^{T}\left(\nabla_\theta \log \pi_\theta\left(a_t^i \mid s_t^i\right)\right)R(\tau)}\right)$$

Why define episodes?

Iteration 0



Agent can enter
bad parts of state-space

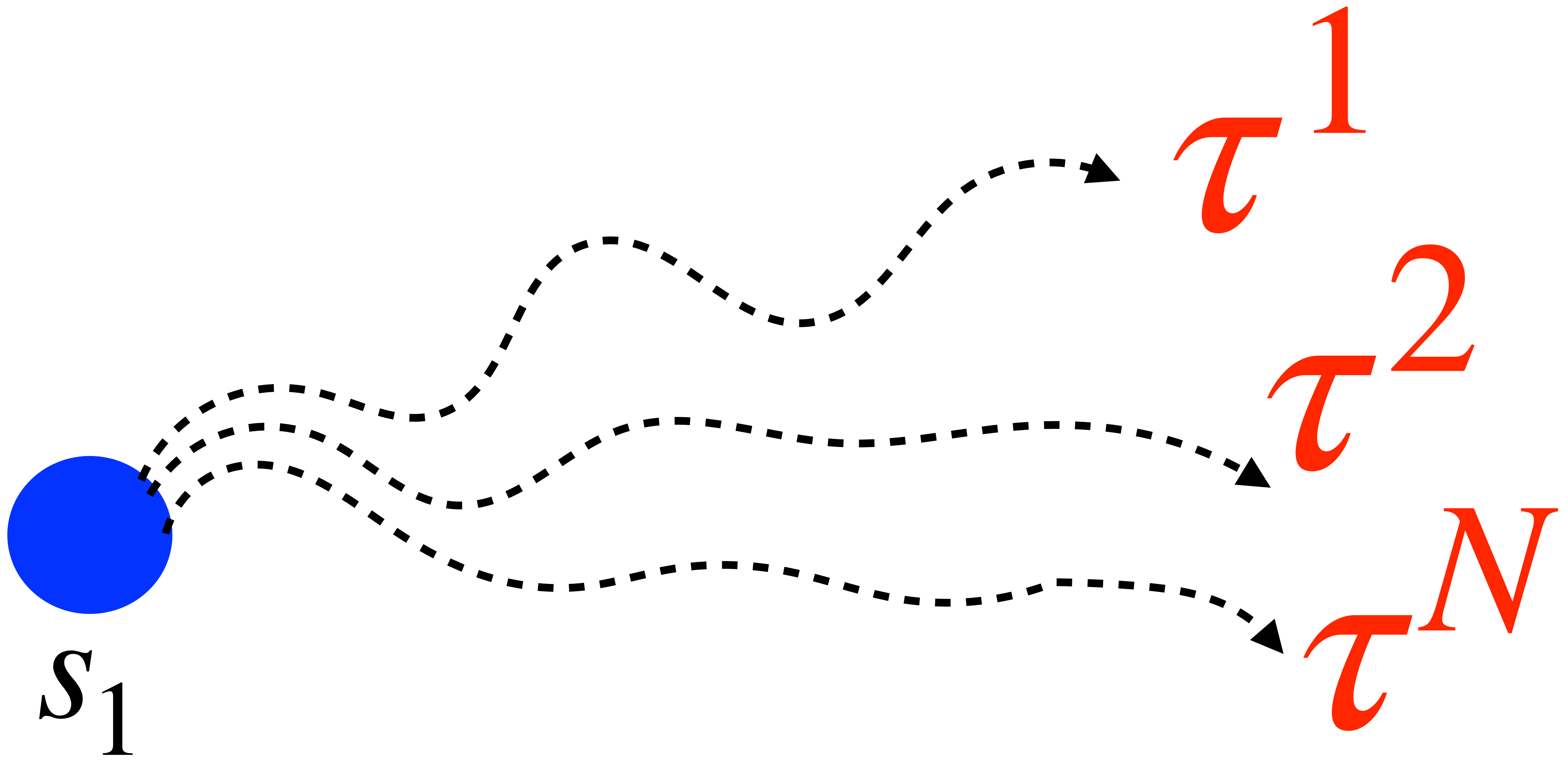
"reset" to good initial state

The Idea of Episode

One Episode

$$\frac{1}{N}\sum_{i=1}^{N}\left(\boxed{\sum_{t=1}^{T}\left(\nabla_\theta \log \pi_\theta(a_t^i \mid s_t^i)\right)R(\tau)}\right)$$

Why define episodes?



Sample multiple trajectories from same initial states

↓

Better monte-carlo estimate

The Idea of Episode

One Episode

$$\frac{1}{N}\sum_{i=1}^{N}\left(\boxed{\sum_{t=1}^{T}\left(\nabla_\theta \log \pi_\theta(a_t^i \mid s_t^i)\right)R(\tau)}\right)$$

Why define episodes?

Some Tasks are Episodic

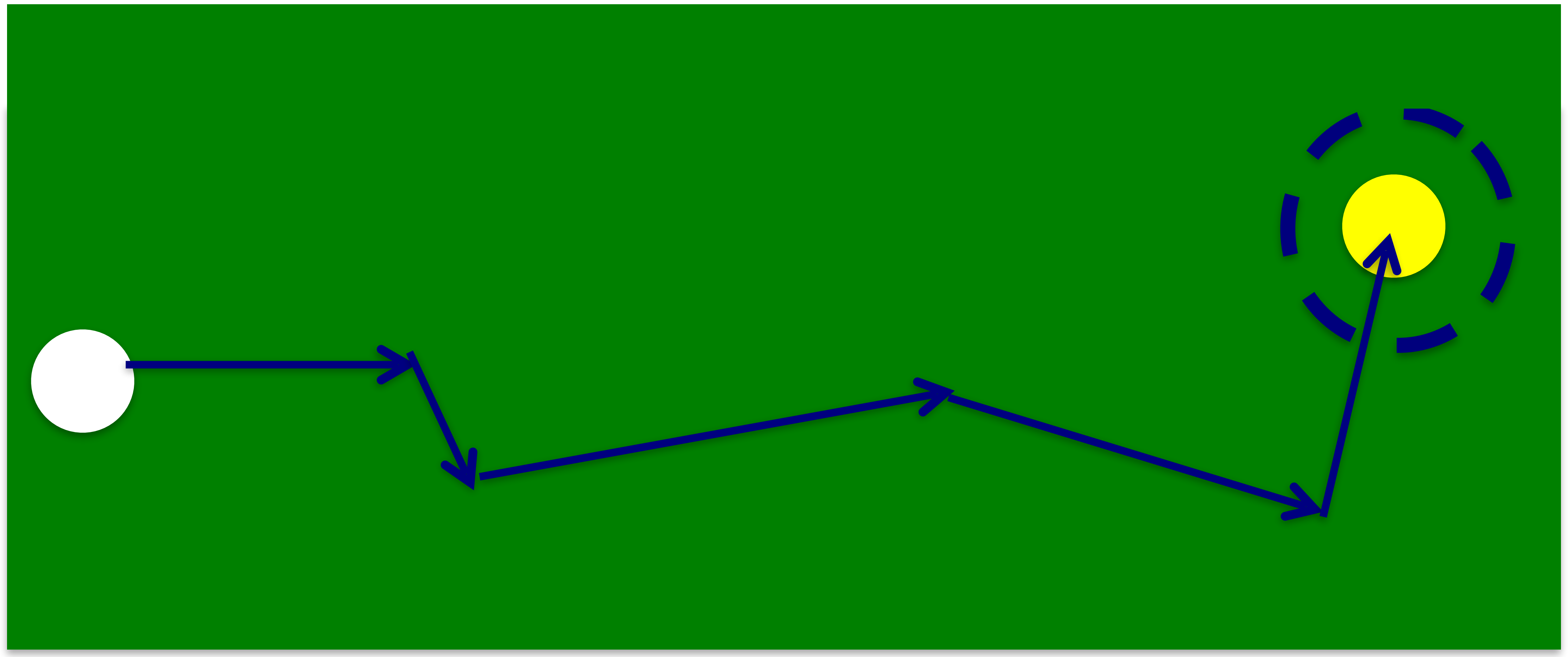
THE CREDIT ASSIGNMENT CHALLENGE

Issue of Credit Assignment

$$E\tau[\nabla_\theta(\log p_\theta(\tau))R(\tau)]$$

Issue of Credit Assignment

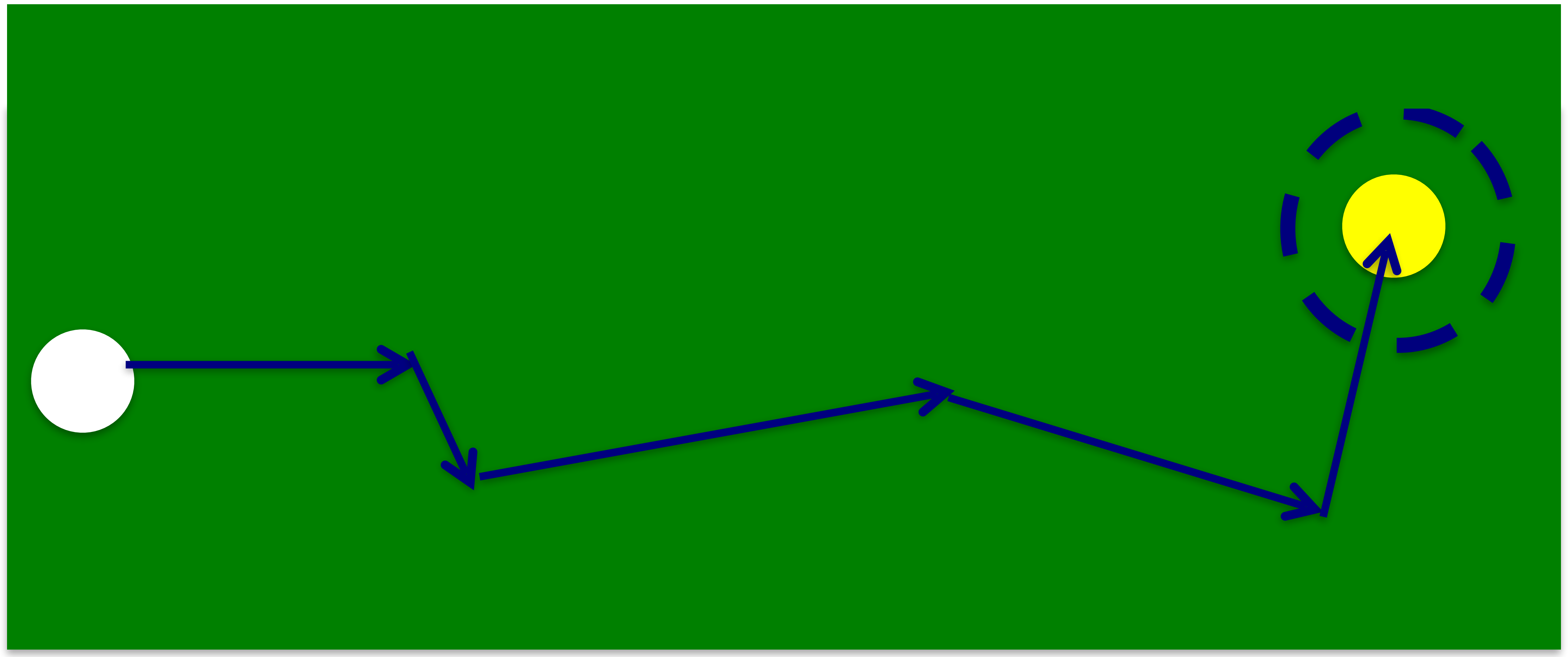
$$E\tau[\nabla_\theta(\log p_\theta(\tau))R(\tau)]$$



Issue of Credit Assignment

$$E\tau[\nabla_\theta(\log p_\theta(\tau))R(\tau)]$$

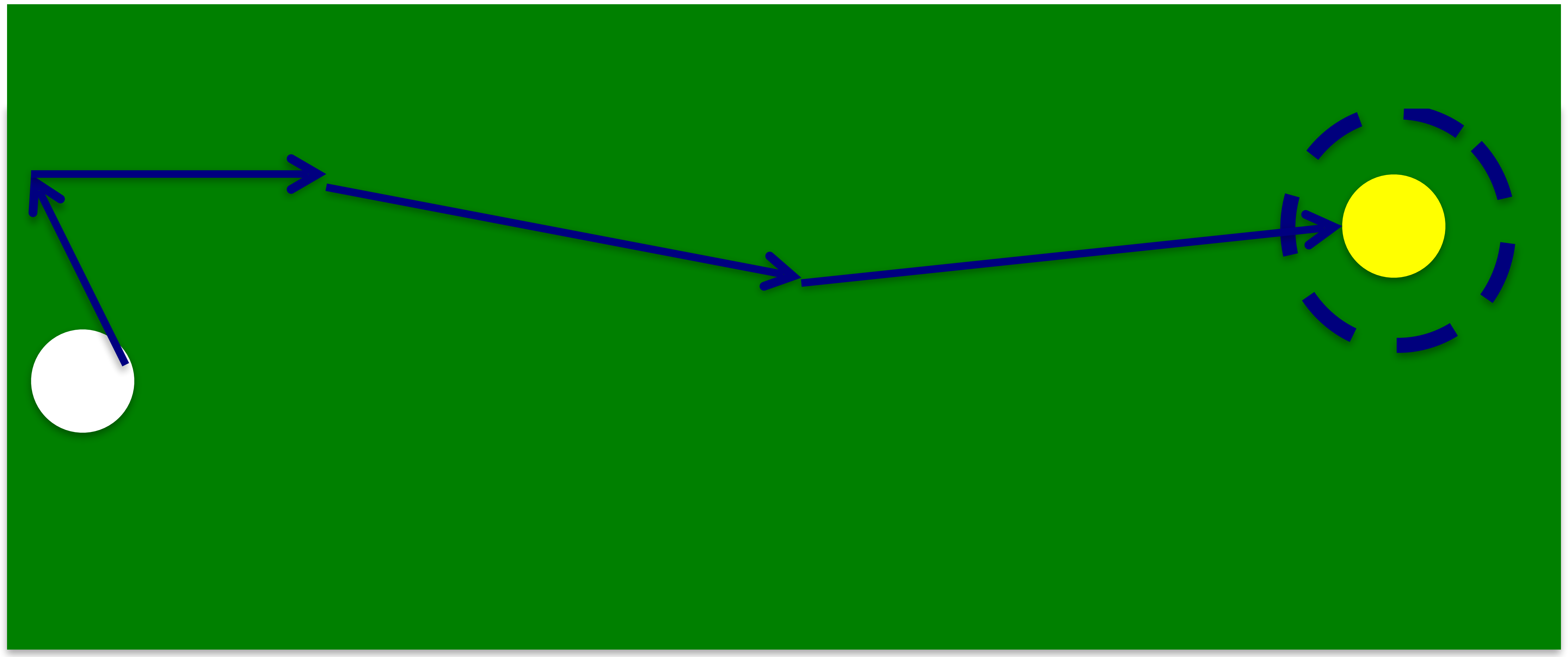
log-prob of each action is increased



Issue of Credit Assignment

$$E\tau[\nabla_\theta(\log p_\theta(\tau))R(\tau)]$$

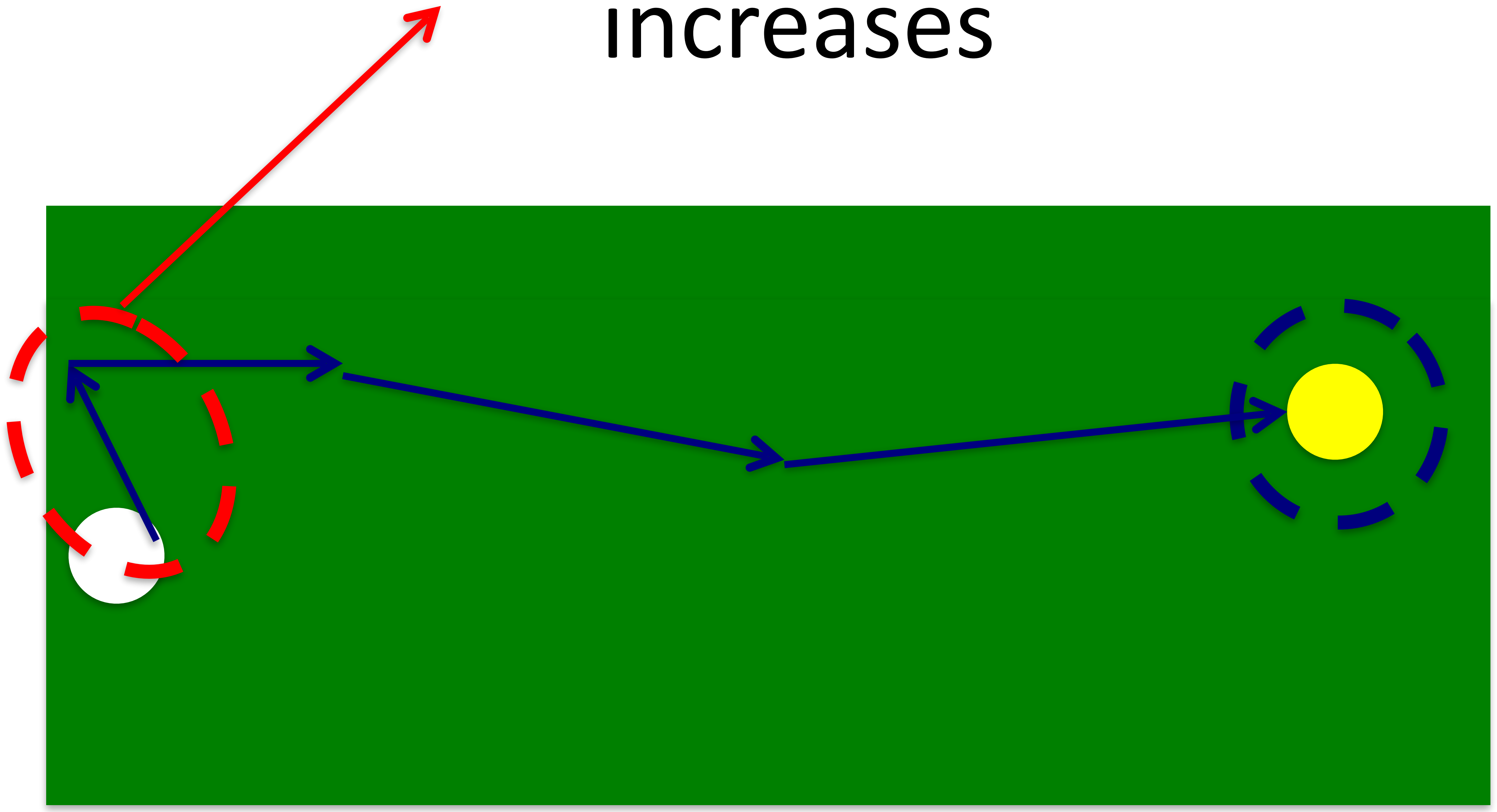
What about in this case?



Issue of Credit Assignment

$$E\tau[\nabla_\theta(\log p_\theta(\tau))R(\tau)]$$

logprob of this action also increases

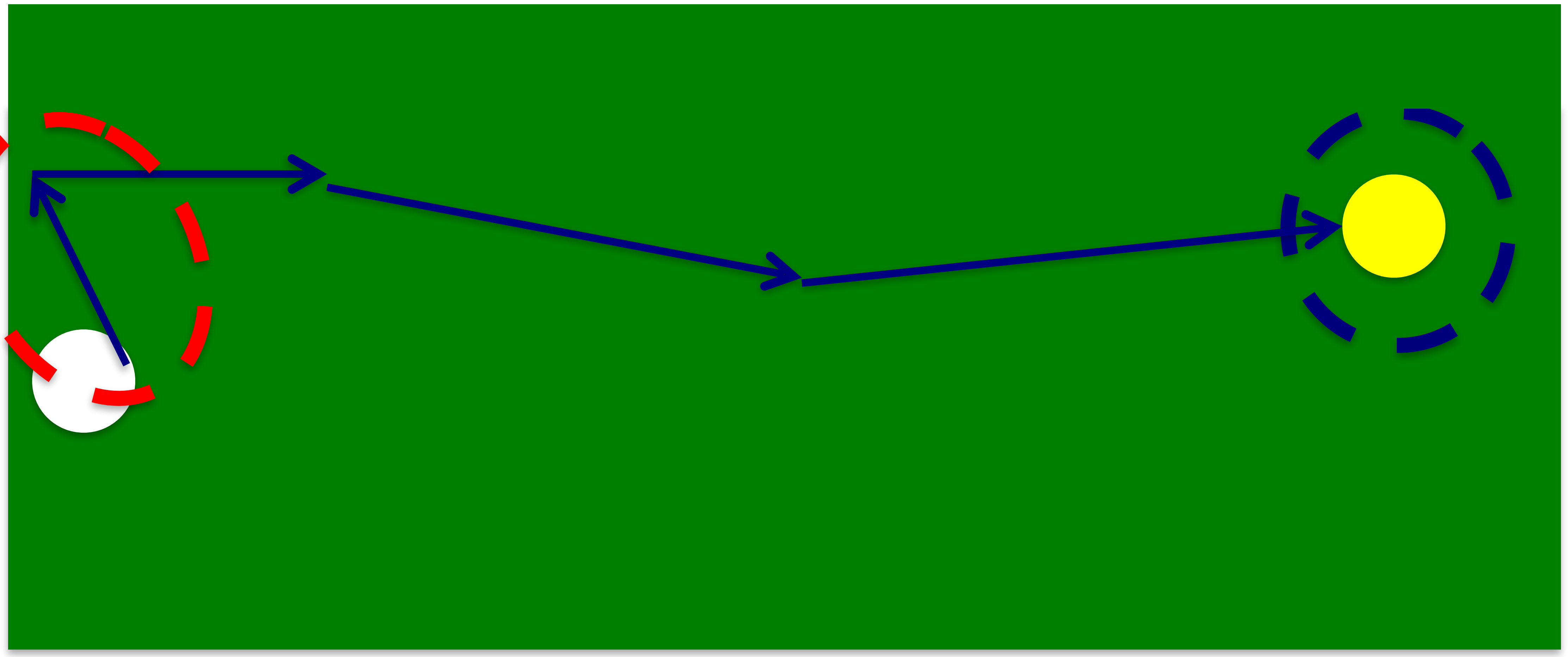


Issue of Credit Assignment

Does this also happen In supervised learning?

$$E\tau[\nabla_\theta(\log p_\theta(\tau))R(\tau)]$$

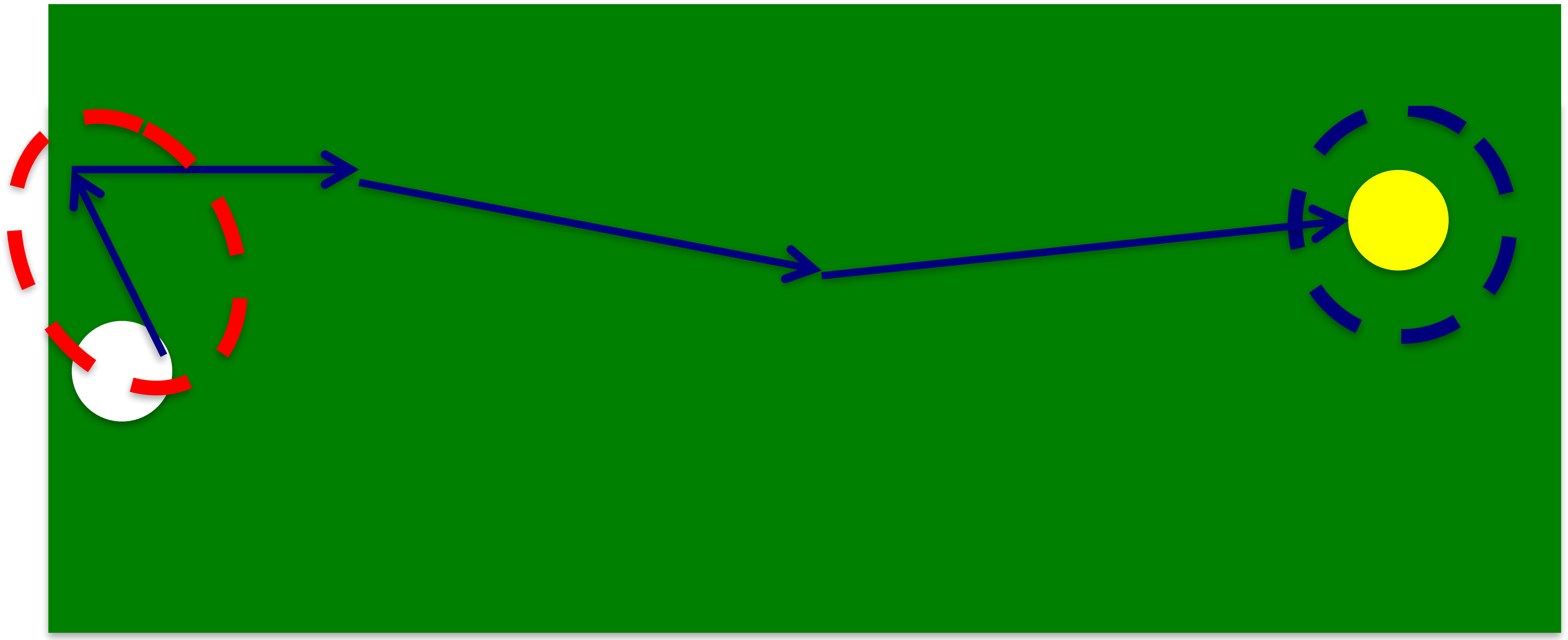
Delayed reward → Ambiguity in which action should be credited



Issue of Credit Assignment

$$E\tau[\nabla_\theta(\log p_\theta(\tau))R(\tau)]$$

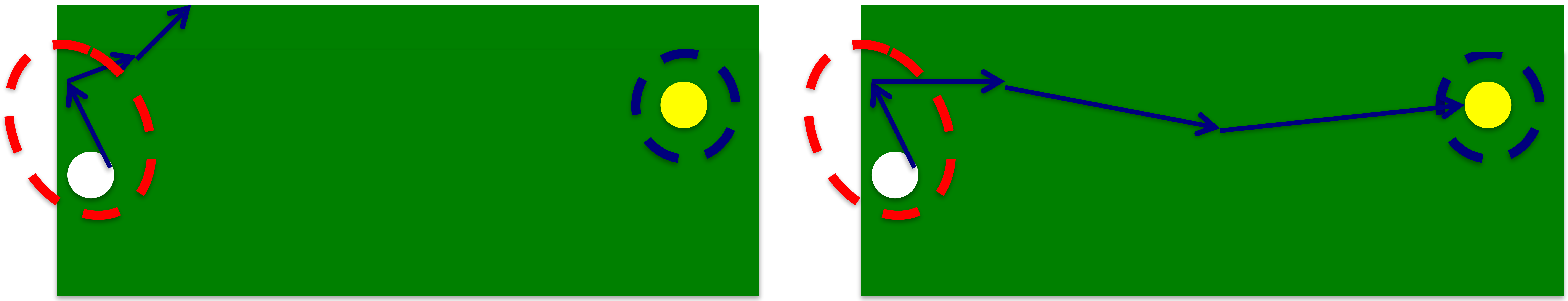
High Variance in gradient estimates



Variance in Policy Gradients

$$E_\tau[\nabla_\theta\left(\log p_\theta(\tau)\right)R(\tau)]$$

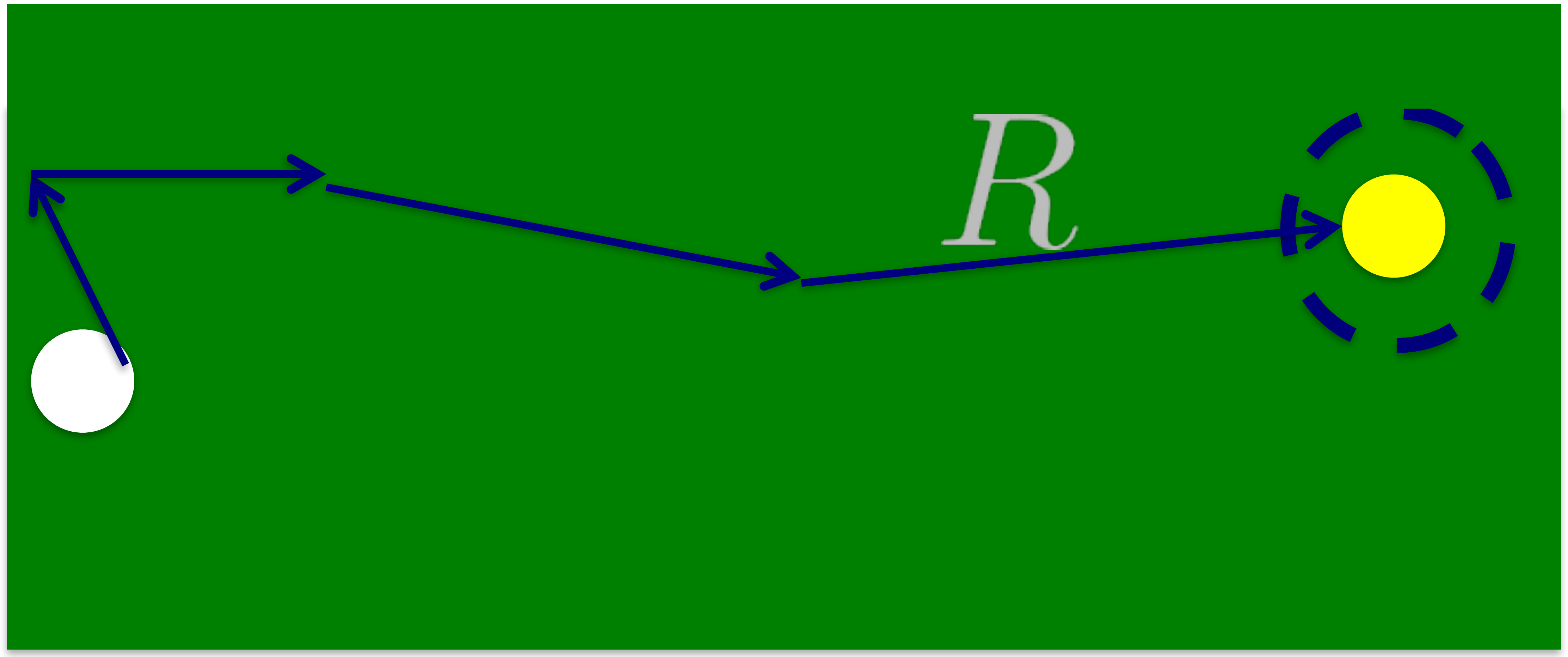
Same action — different trajectory rewards



conflicting gradients: variance

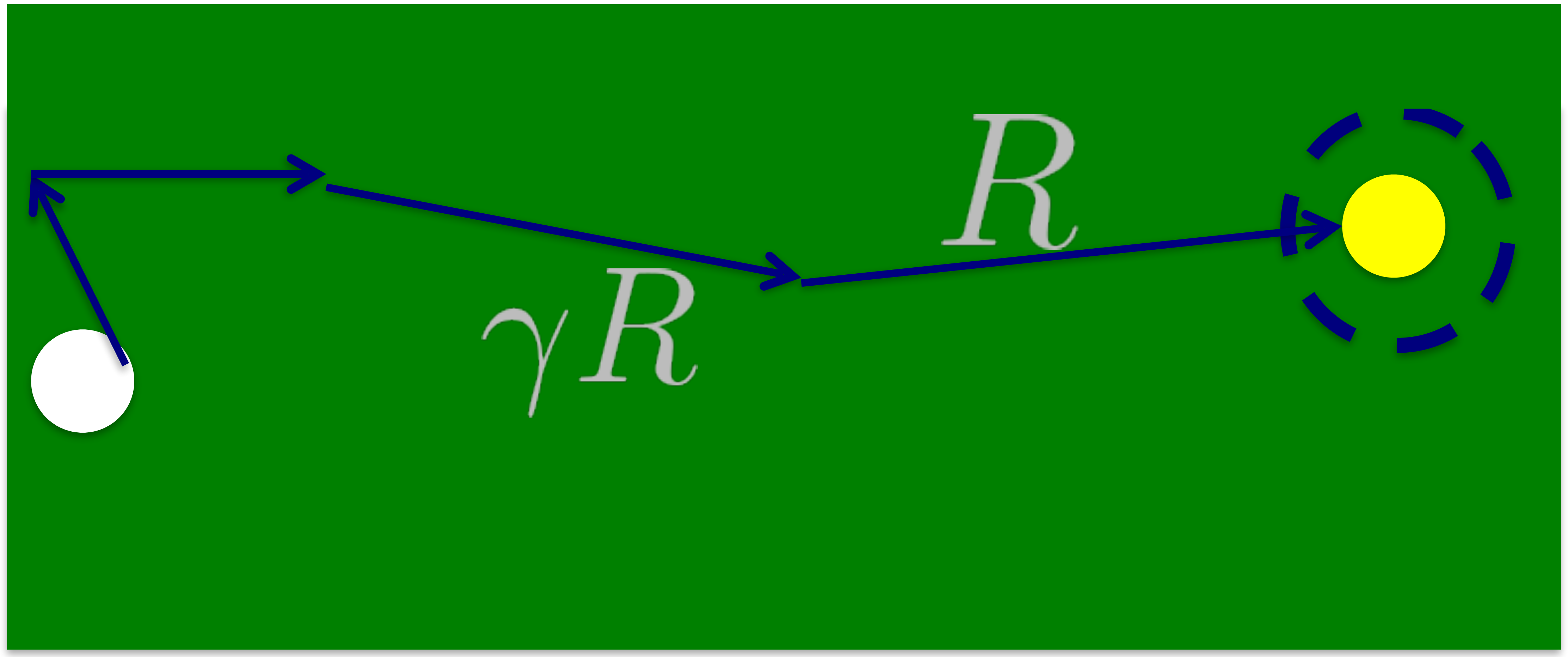
$$\mathrm{Var}\left[\nabla_\theta \log p_\theta(\tau) R(\tau)\right]$$

Variance Reduction Idea -- Discounts



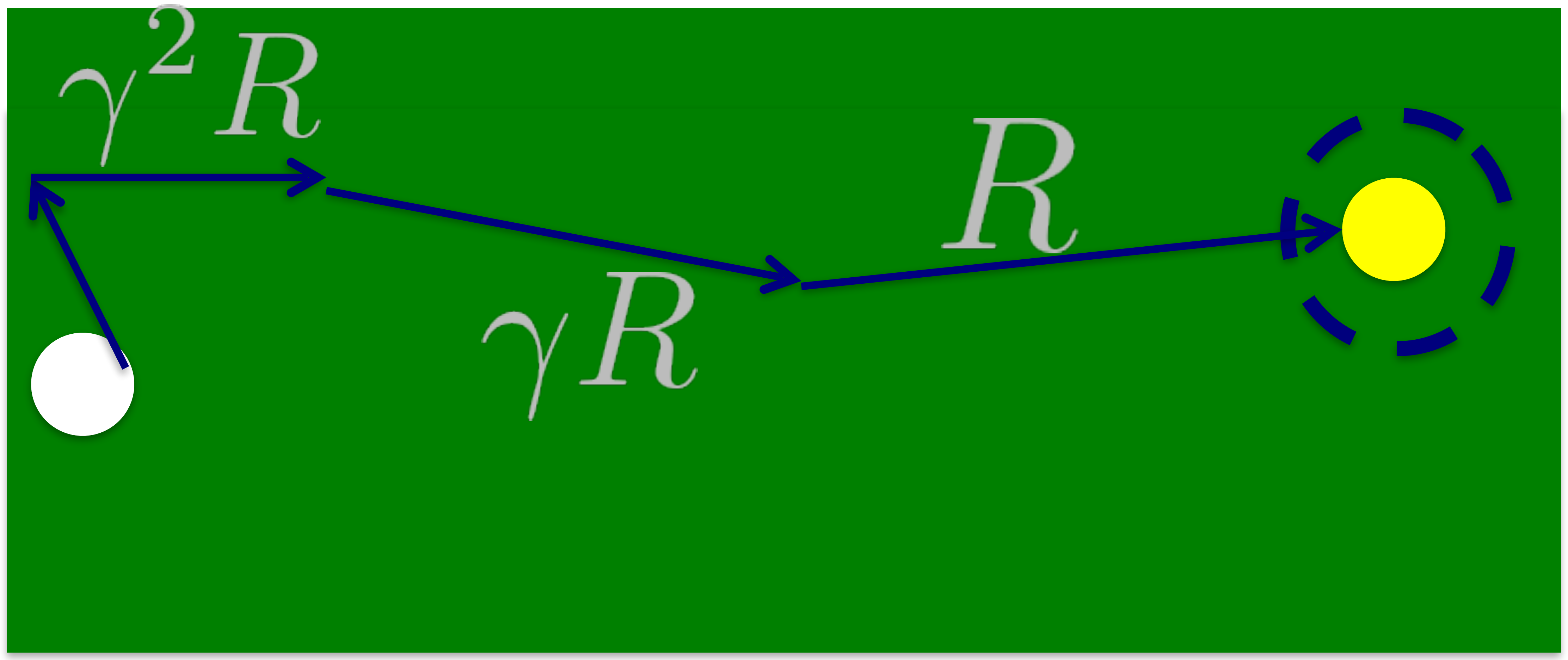
Variance Reduction Idea -- Discounts

$$\gamma < 1$$



Variance Reduction Idea -- Discounts

$$\gamma < 1$$



Variance Reduction with Discount

$$E_\tau[\nabla_\theta \left(\log p_\theta(\tau)\right) R(\tau)]$$

$$E_\tau[\nabla_\theta \left(\log p_\theta(\tau)\right) R^\gamma(\tau)]$$

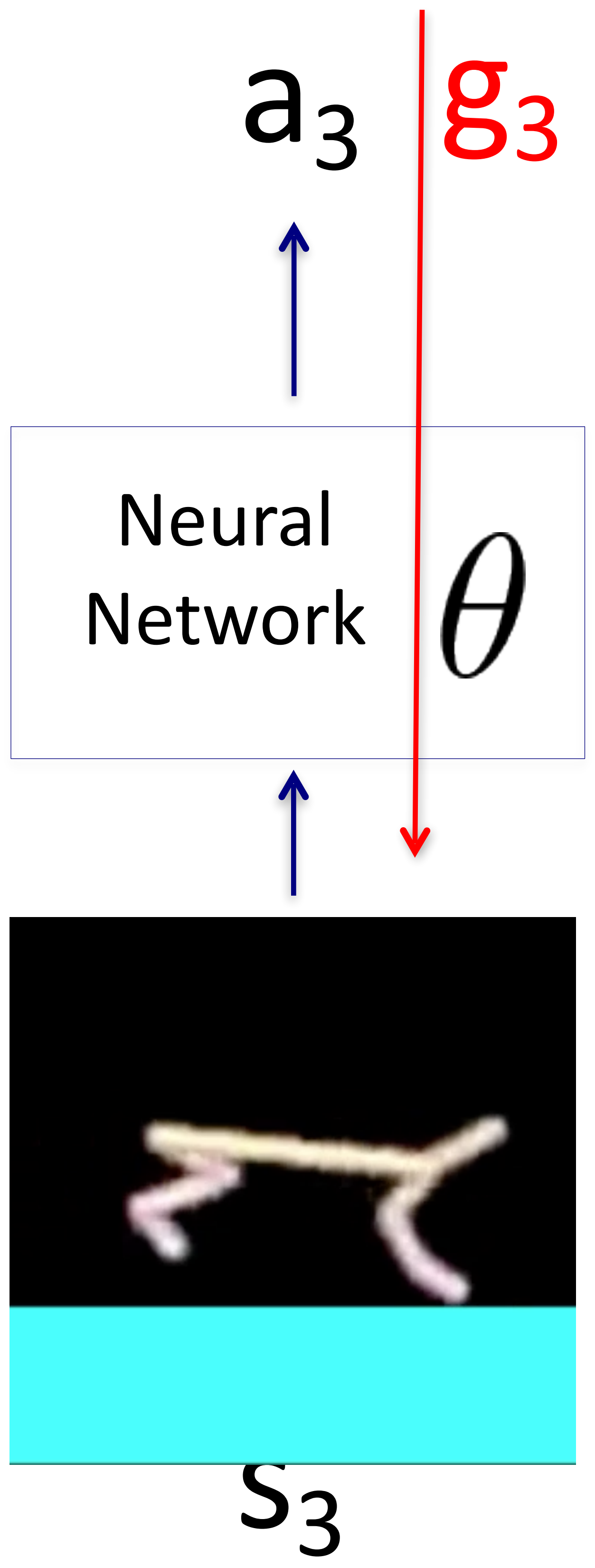
$$R^\gamma(\tau) = \sum_t \gamma^t r_t$$

Faster Convergence

Bias

Makes infinite time horizon work

Bias resulting from discount



If gamma is small, what might happen?

Move fast now
BUT
CAN Fall later!

$$\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\nabla_{\theta}\log\pi(a_t^i\,|\,s_t^i)\sum_{t'=1}^{T}\gamma^{t'}r(s_{t'}^i,a_{t'}^i)\right)$$

This is the BIAS!!

Expanding on Policy Gradients

$$E_\tau\left[\sum_{t=1}^{T}\left(\nabla_\theta \log \pi_\theta(a_t \mid s_t)\right)R(\tau)\right]$$

$$\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta\left(a_t^i \mid s_t^i\right)\left(\sum_{t=1}^{T} r(s_t^i, a_t^i)\right)\right)$$

Expanding on Policy Gradients

$$E_\tau\left[\sum_{t=1}^{T}\left(\nabla_\theta \log \pi_\theta(a_t\,|\,s_t)\right)R(\tau)\right]$$

$$\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\nabla_\theta\log\pi_\theta(a_t^i\,|\,s_t^i)\left(\sum_{t=1}^{T} r(s_t^i,a_t^i)\right)\right)$$

Can we reduce variance?

$$\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\nabla_\theta\log\pi_\theta(a_t^i\,|\,s_t^i)\left(\sum_{t'=t}^{T} r(s_{t'}^i,a_{t'}^i)\right)\right)$$

current actions don't effect past rewards!

PROOF OF WHY POLICY GRADIENT IS MODEL FREE

Policy Gradients

$$E\tau[\nabla_\theta \log p_\theta(\tau)(R(\tau) - b)]$$

where,

b: baseline

$$b = E\tau[R(\tau)]$$

Policy Gradients

$$E\tau[\nabla_\theta \log \boxed{p_\theta(\tau)}(R(\tau) - b)]$$

$$p_\theta(\tau) \quad = \quad p_\theta(s_1, a_1, r_1, s_2, a_2, r_2, ..., s_{t-1}, a_{t-1}, r_{t-1}, s_t)$$

Policy Gradients

$$E\tau[\nabla_\theta \log \boxed{p_\theta(\tau)}(R(\tau) - b)]$$

$$\begin{aligned} p_\theta(\tau) &= p_\theta(s_1, a_1, r_1, s_2, a_2, r_2, ..., s_{t-1}, a_{t-1}, \boxed{r_{t-1}, s_t)} \\ &= p_\theta(r_{t-1}, s_t | s_1, a_1, r_1, s_2, a_2, r_2, ..., s_{t-1}, a_{t-1}) p_\theta(s_1, a_1, r_1, ..., s_{t-1}, a_{t-1}) \end{aligned}$$

Policy Gradients

$$E\tau[\nabla_\theta \log \boxed{p_\theta(\tau)}(R(\tau) - b)]$$

$$\begin{aligned}
p_\theta(\tau) &= p_\theta(s_1, a_1, r_1, s_2, a_2, r_2, ..., s_{t-1}, a_{t-1}, r_{t-1}, s_t) \\
&= p_\theta(r_{t-1}, s_t | s_1, a_1, r_1, s_2, a_2, r_2, ..., \boxed{s_{t-1}, a_{t-1}}) p_\theta(s_1, a_1, r_1, ..., s_{t-1}, a_{t-1}) \\
&= p_\theta(r_{t-1}, s_t | s_{t-1}, a_{t-1}) p_\theta(s_1, a_1, r_1, ..., s_{t-1}, a_{t-1})
\end{aligned}$$

Policy Gradients

$$E\tau[\nabla_\theta \log \boxed{p_\theta(\tau)}(R(\tau) - b)]$$

$$\begin{aligned}
p_\theta(\tau) &= p_\theta(s_1, a_1, r_1, s_2, a_2, r_2, ..., s_{t-1}, a_{t-1}, r_{t-1}, s_t) \\
&= p_\theta(r_{t-1}, s_t | s_1, a_1, r_1, s_2, a_2, r_2, ..., s_{t-1}, a_{t-1}) p_\theta(s_1, a_1, r_1, ..., s_{t-1}, a_{t-1}) \\
&= \boxed{p_\theta}(r_{t-1}, s_t | s_{t-1}, a_{t-1}) p_\theta(s_1, a_1, r_1, ..., s_{t-1}, a_{t-1}) \\
&= p(r_{t-1}, s_t | s_{t-1}, a_{t-1}) p_\theta(s_1, a_1, r_1, ..., s_{t-1}, a_{t-1})
\end{aligned}$$

Policy Gradients

$$E\tau[\nabla_\theta \log \boxed{p_\theta(\tau)}(R(\tau) - b)]$$

$$\begin{aligned}
p_\theta(\tau) &= p_\theta(s_1, a_1, r_1, s_2, a_2, r_2, ..., s_{t-1}, a_{t-1}, r_{t-1}, s_t) \\
&= p_\theta(r_{t-1}, s_t | s_1, a_1, r_1, s_2, a_2, r_2, ..., s_{t-1}, a_{t-1}) p_\theta(s_1, a_1, r_1, ..., s_{t-1}, a_{t-1}) \\
&= p_\theta(r_{t-1}, s_t | s_{t-1}, a_{t-1}) p_\theta(s_1, a_1, r_1, ..., s_{t-1}, a_{t-1}) \\
&= p(r_{t-1}, s_t | s_{t-1}, a_{t-1}) p_\theta(s_1, a_1, r_1, ..., s_{t-1}, \boxed{a_{t-1}}) \\
&= p(r_{t-1}, s_t | s_{t-1}, a_{t-1}) p_\theta(a_{t-1} | s_1, a_1, r_1, ..., s_{t-1}) p_\theta(s_1, a_1, r_1, ..., s_{t-1})
\end{aligned}$$

Policy Gradients

$$E\tau[\nabla_\theta \log \boxed{p_\theta(\tau)}(R(\tau) - b)]$$

$$\begin{aligned}
p_\theta(\tau) &= p_\theta(s_1, a_1, r_1, s_2, a_2, r_2, ..., s_{t-1}, a_{t-1}, r_{t-1}, s_t) \\
&= p_\theta(r_{t-1}, s_t | s_1, a_1, r_1, s_2, a_2, r_2, ..., s_{t-1}, a_{t-1}) p_\theta(s_1, a_1, r_1, ..., s_{t-1}, a_{t-1}) \\
&= p_\theta(r_{t-1}, s_t | s_{t-1}, a_{t-1}) p_\theta(s_1, a_1, r_1, ..., s_{t-1}, a_{t-1}) \\
&= p(r_{t-1}, s_t | s_{t-1}, a_{t-1}) p_\theta(s_1, a_1, r_1, ..., s_{t-1}, a_{t-1}) \\
&= p(r_{t-1}, s_t | s_{t-1}, a_{t-1}) p_\theta(a_{t-1} | s_1, a_1, r_1, ..., \boxed{s_{t-1}}) p_\theta(s_1, a_1, r_1, ..., s_{t-1}) \\
&= p(r_{t-1}, s_t | s_{t-1}, a_{t-1}) p_\theta(a_{t-1} | s_{t-1}) p_\theta(s_1, a_1, r_1, ..., s_{t-1})
\end{aligned}$$

Policy Gradients

$$E\tau[\nabla_\theta \log \boxed{p_\theta(\tau)}(R(\tau) - b)]$$

$$\begin{aligned}
p_\theta(\tau) &= p_\theta(s_1, a_1, r_1, s_2, a_2, r_2, ..., s_{t-1}, a_{t-1}, r_{t-1}, s_t) \\
&= p_\theta(r_{t-1}, s_t | s_1, a_1, r_1, s_2, a_2, r_2, ..., s_{t-1}, a_{t-1}) p_\theta(s_1, a_1, r_1, ..., s_{t-1}, a_{t-1}) \\
&= p_\theta(r_{t-1}, s_t | s_{t-1}, a_{t-1}) p_\theta(s_1, a_1, r_1, ..., s_{t-1}, a_{t-1}) \\
&= p(r_{t-1}, s_t | s_{t-1}, a_{t-1}) p_\theta(s_1, a_1, r_1, ..., s_{t-1}, a_{t-1}) \\
&= p(r_{t-1}, s_t | s_{t-1}, a_{t-1}) p_\theta(a_{t-1} | s_1, a_1, r_1, ..., s_{t-1}) p_\theta(s_1, a_1, r_1, ..., s_{t-1}) \\
&= p(r_{t-1}, s_t | s_{t-1}, a_{t-1}) \boxed{p_\theta}(a_{t-1} | s_{t-1}) p_\theta(s_1, a_1, r_1, ..., s_{t-1}) \\
&= p(r_{t-1}, s_t | s_{t-1}, a_{t-1}) \pi_\theta(a_{t-1} | s_{t-1}) p_\theta(s_1, a_1, r_1, ..., s_{t-1})
\end{aligned}$$

Policy Gradients

$$E\tau[\nabla_\theta \log \boxed{p_\theta(\tau)}(R(\tau) - b)]$$

$$\begin{aligned}
p_\theta(\tau) &= p_\theta(s_1, a_1, r_1, s_2, a_2, r_2, ..., s_{t-1}, a_{t-1}, r_{t-1}, s_t) \\
&= p_\theta(r_{t-1}, s_t | s_1, a_1, r_1, s_2, a_2, r_2, ..., s_{t-1}, a_{t-1}) p_\theta(s_1, a_1, r_1, ..., s_{t-1}, a_{t-1}) \\
&= p_\theta(r_{t-1}, s_t | s_{t-1}, a_{t-1}) p_\theta(s_1, a_1, r_1, ..., s_{t-1}, a_{t-1}) \\
&= p(r_{t-1}, s_t | s_{t-1}, a_{t-1}) p_\theta(s_1, a_1, r_1, ..., s_{t-1}, a_{t-1}) \\
&= p(r_{t-1}, s_t | s_{t-1}, a_{t-1}) p_\theta(a_{t-1} | s_1, a_1, r_1, ..., s_{t-1}) p_\theta(s_1, a_1, r_1, ..., s_{t-1}) \\
&= p(r_{t-1}, s_t | s_{t-1}, a_{t-1}) p_\theta(a_{t-1} | s_{t-1}) p_\theta(s_1, a_1, r_1, ..., s_{t-1}) \\
&= p(r_{t-1}, s_t | s_{t-1}, a_{t-1}) \pi_\theta(a_{t-1} | s_{t-1}) p_\theta(s_1, a_1, r_1, ..., s_{t-1}) \\
&= \prod_{i=1}^{t} p(r_{t-i}, s_{t-i+1} | s_{t-i}, a_{t-i}) \pi_\theta(a_{t-i} | s_{t-i})
\end{aligned}$$

Policy Gradients

$$E\tau[\boxed{\nabla_\theta \log p_\theta(\tau)}(R(\tau) - b)]$$

$$p_\theta(\tau) \quad = \quad \prod_{i=1}^{t} p(r_{t-i}, s_{t-i+1} | s_{t-i}, a_{t-i}) \pi_\theta(a_{t-i} | s_{t-i})$$

Policy Gradients

$$E\tau[\boxed{\nabla_\theta \log p_\theta(\tau)}(R(\tau) - b)]$$

$$\begin{aligned} p_\theta(\tau) &= \prod_{i=1}^{t} p(r_{t-i}, s_{t-i+1} | s_{t-i}, a_{t-i}) \pi_\theta(a_{t-i} | s_{t-i}) \\ \Rightarrow \log p_\theta(\tau) &= \sum_{i=1}^{t} \log p(r_{t-i}, s_{t-i+1} | s_{t-i}, a_{t-i}) + \sum_{i=1}^{t} \log \pi_\theta(a_{t-i} | s_{t-i}) \end{aligned}$$

Policy Gradients

$$E\tau[\boxed{\nabla_\theta \log p_\theta(\tau)}(R(\tau) - b)]$$

$$\begin{aligned}
p_\theta(\tau) &= \prod_{i=1}^{t} p(r_{t-i}, s_{t-i+1}|s_{t-i}, a_{t-i})\pi_\theta(a_{t-i}|s_{t-i}) \\
\Rightarrow \log p_\theta(\tau) &= \sum_{i=1}^{t} \log p(r_{t-i}, s_{t-i+1}|s_{t-i}, a_{t-i}) + \sum_{i=1}^{t} \log \pi_\theta(a_{t-i}|s_{t-i}) \\
\Rightarrow \nabla_\theta \log p_\theta(\tau) &= \sum_{i=1}^{t} \nabla_\theta \log p(r_{t-i}, s_{t-i+1}|s_{t-i}, a_{t-i}) + \sum_{i=1}^{t} \nabla_\theta \log \pi_\theta(a_{t-i}|s_{t-i})
\end{aligned}$$

Policy Gradients

$$E\tau[\boxed{\nabla_\theta \log p_\theta(\tau)}(R(\tau) - b)]$$

$$\begin{aligned}
p_\theta(\tau) &= \prod_{i=1}^{t} p(r_{t-i}, s_{t-i+1} | s_{t-i}, a_{t-i}) \pi_\theta(a_{t-i} | s_{t-i}) \\
\Rightarrow \log p_\theta(\tau) &= \sum_{i=1}^{t} \log p(r_{t-i}, s_{t-i+1} | s_{t-i}, a_{t-i}) + \sum_{i=1}^{t} \log \pi_\theta(a_{t-i} | s_{t-i}) \\
\Rightarrow \nabla_\theta \log p_\theta(\tau) &= \sum_{i=1}^{t} \boxed{\nabla_\theta \log p(r_{t-i}, s_{t-i+1} | s_{t-i}, a_{t-i})} + \sum_{i=1}^{t} \nabla_\theta \log \pi_\theta(a_{t-i} | s_{t-i})
\end{aligned}$$

Policy Gradients

$$E\tau[\boxed{\nabla_\theta \log p_\theta(\tau)}(R(\tau) - b)]$$

$$\begin{aligned}
p_\theta(\tau) &= \prod_{i=1}^{t} p(r_{t-i}, s_{t-i+1} | s_{t-i}, a_{t-i}) \pi_\theta(a_{t-i} | s_{t-i}) \\
\Rightarrow \log p_\theta(\tau) &= \sum_{i=1}^{t} \log p(r_{t-i}, s_{t-i+1} | s_{t-i}, a_{t-i}) + \sum_{i=1}^{t} \log \pi_\theta(a_{t-i} | s_{t-i}) \\
\Rightarrow \nabla_\theta \log p_\theta(\tau) &= \sum_{i=1}^{t} \boxed{\cancel{\nabla_\theta \log p(r_{t-i}, s_{t-i+1} | s_{t-i}, a_{t-i})}} + \sum_{i=1}^{t} \nabla_\theta \log \pi_\theta(a_{t-i} | s_{t-i})
\end{aligned}$$

0

Policy Gradients

$$E\tau[\boxed{\nabla_\theta \log p_\theta(\tau)}(R(\tau) - b)]$$

$$\begin{aligned}
p_\theta(\tau) &= \prod_{i=1}^{t} p(r_{t-i}, s_{t-i+1} | s_{t-i}, a_{t-i}) \pi_\theta(a_{t-i}|s_{t-i}) \\
\Rightarrow \log p_\theta(\tau) &= \sum_{i=1}^{t} \log p(r_{t-i}, s_{t-i+1} | s_{t-i}, a_{t-i}) + \sum_{i=1}^{t} \log \pi_\theta(a_{t-i}|s_{t-i}) \\
\Rightarrow \nabla_\theta \log p_\theta(\tau) &= \sum_{i=1}^{t} \nabla_\theta \log p(r_{t-i}, s_{t-i+1} | s_{t-i}, a_{t-i}) + \sum_{i=1}^{t} \nabla_\theta \log \pi_\theta(a_{t-i}|s_{t-i}) \\
&= \sum_{i=1}^{t} \nabla_\theta \log \pi_\theta(a_{t-i}|s_{t-i})
\end{aligned}$$

Independent of the environment dynamics !!

Policy Gradients

$$E\tau[\boxed{\nabla_\theta \log p_\theta(\tau)}(R(\tau) - b)]$$

$$\begin{aligned}
p_\theta(\tau) &= \prod_{i=1}^{t} p(r_{t-i}, s_{t-i+1}|s_{t-i}, a_{t-i})\pi_\theta(a_{t-i}|s_{t-i}) \\
\Rightarrow \log p_\theta(\tau) &= \sum_{i=1}^{t} \log p(r_{t-i}, s_{t-i+1}|s_{t-i}, a_{t-i}) + \sum_{i=1}^{t} \log \pi_\theta(a_{t-i}|s_{t-i}) \\
\Rightarrow \nabla_\theta \log p_\theta(\tau) &= \sum_{i=1}^{t} \nabla_\theta \log p(r_{t-i}, s_{t-i+1}|s_{t-i}, a_{t-i}) + \sum_{i=1}^{t} \nabla_\theta \log \pi_\theta(a_{t-i}|s_{t-i}) \\
&= \sum_{i=1}^{t} \nabla_\theta \log \pi_\theta(a_{t-i}|s_{t-i}) \\
&= \sum_{i=0}^{t-1} \nabla_\theta \log \pi_\theta(a_i|s_i)
\end{aligned}$$

Independent of the environment dynamics !!