Sensorimotor Learning
(Spring’23)

Pulkit Agrawal

Lecture 4: Policy Gradients

Feb 16 2023

Lecture Outline

Understand Policy Gradients

Credit Assignment Problem

Variance Reduction Techniques

- Causality
- Discounting
- Baselines
- Use of Critic
  - Generalized Advantage Estimation

Why if Policy Gradients On-Policy?

Asynchronous Methods

$$a_2 \quad a_3$$

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

(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)

$$a_1 \quad a_2 \quad a_3$$

$$a_2 \quad a_3$$

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

(female, 20s, computer-savvy)

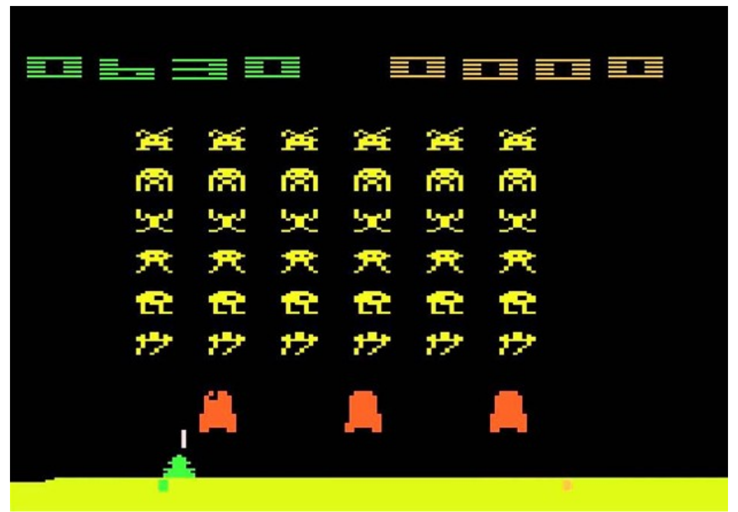
(male, 30s, computer-savvy)

How to use these “features” in decision making?

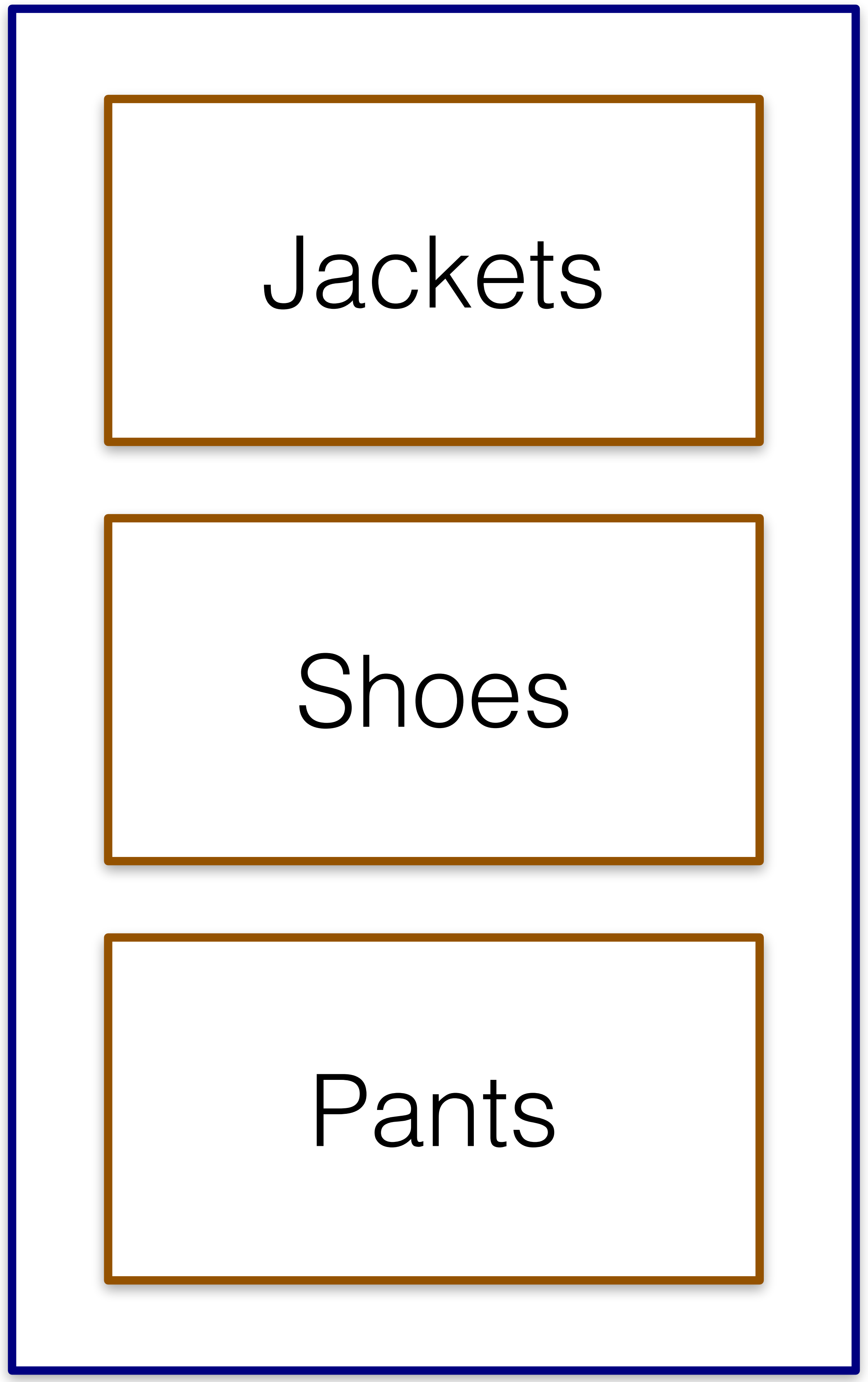
Contextual Bandits

BUT, Actions don’t change future state

Model Free Reinforcement Learning

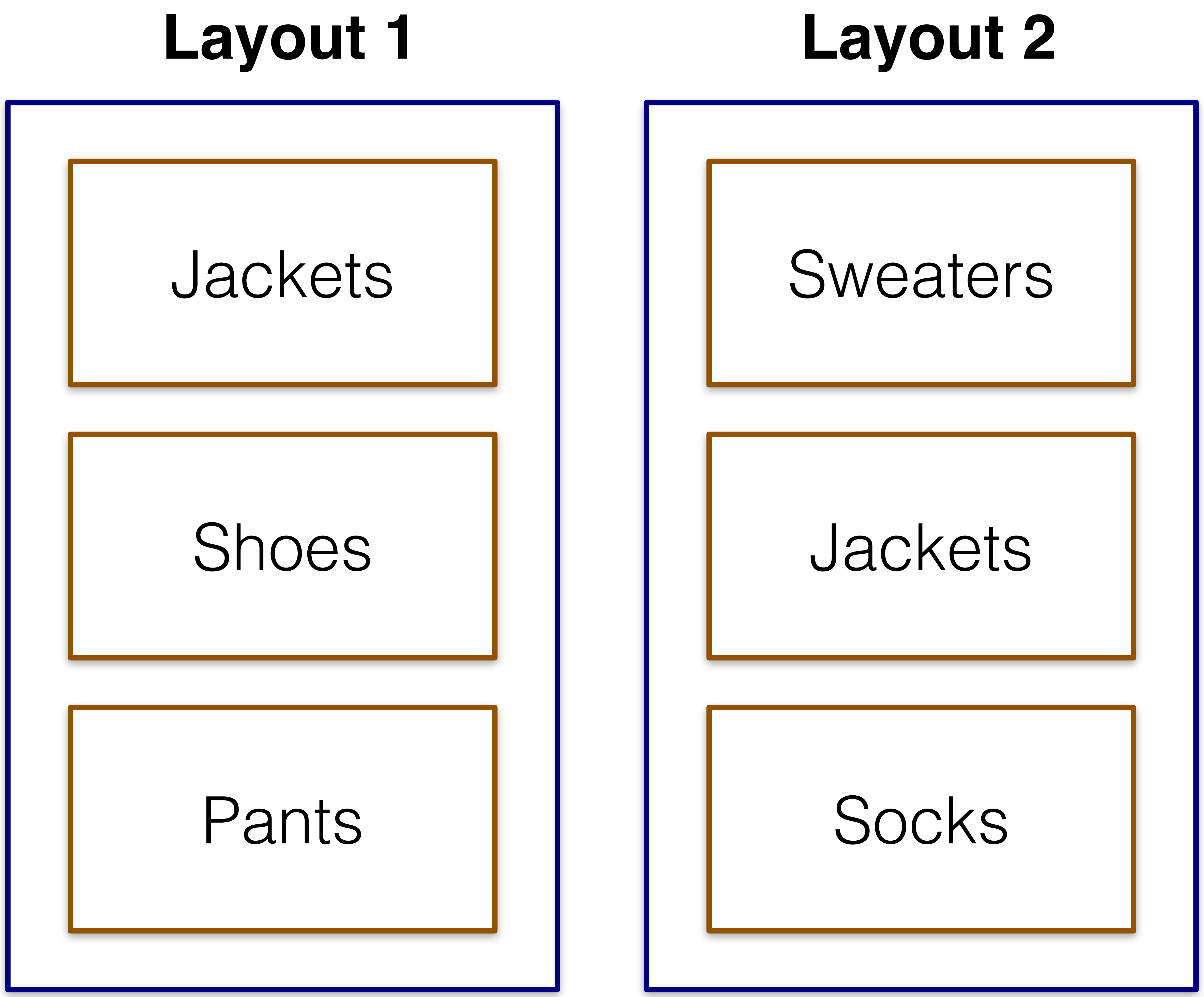


Layout 1



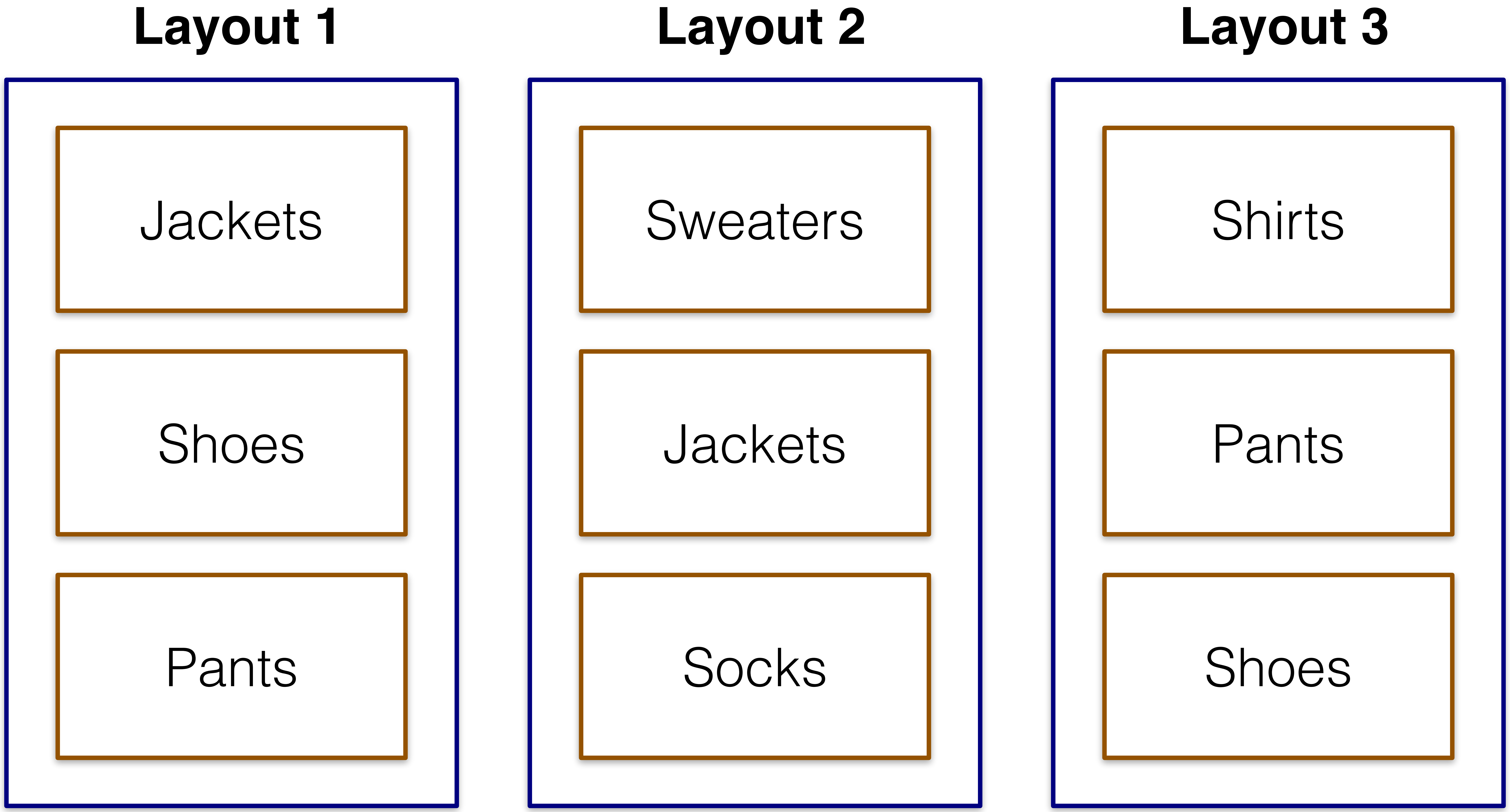
state: x_t x_t: user features

action: a_1



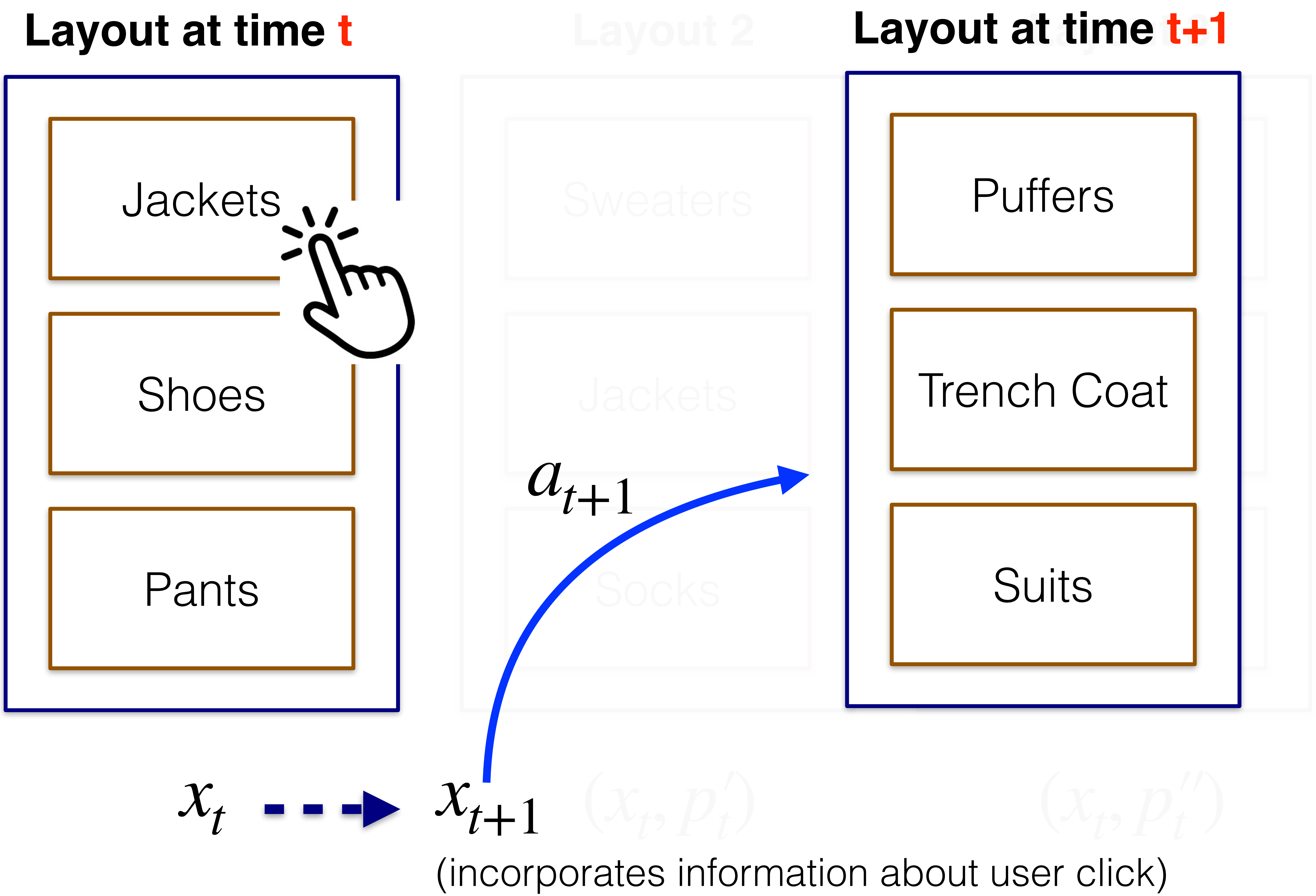
state: x_t

action: a_1 a_2



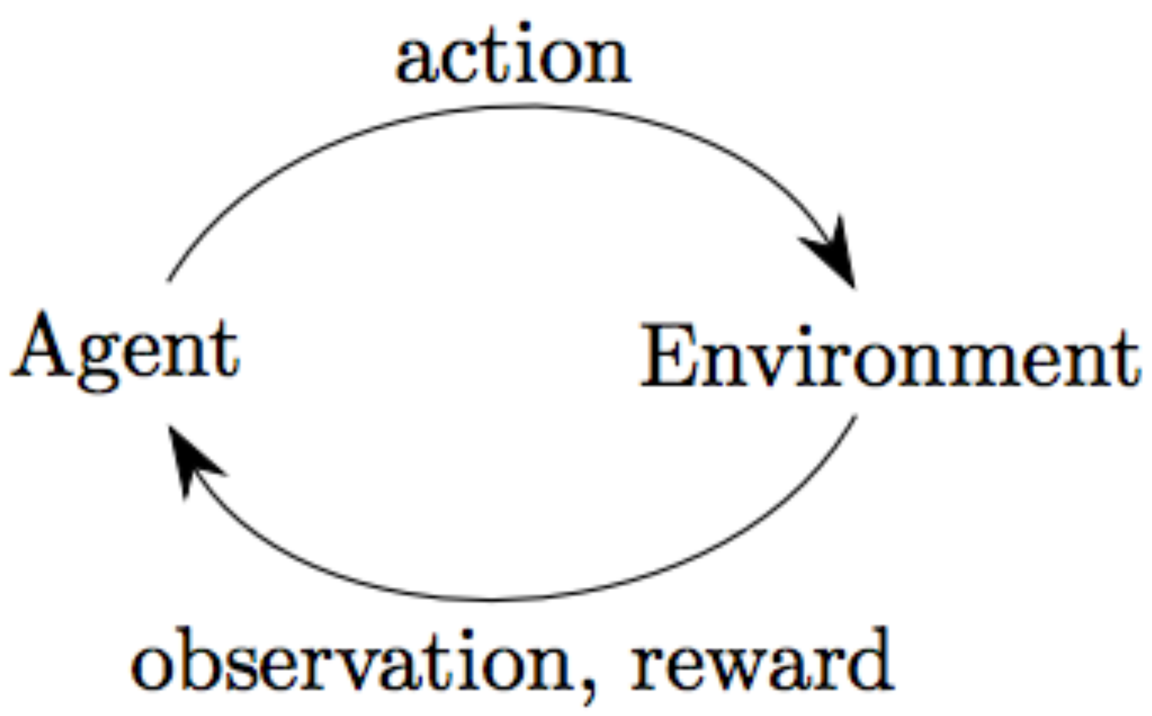
state: x_t

action: a_1 a_2 a_3



State of the system evolves with actions

The problem

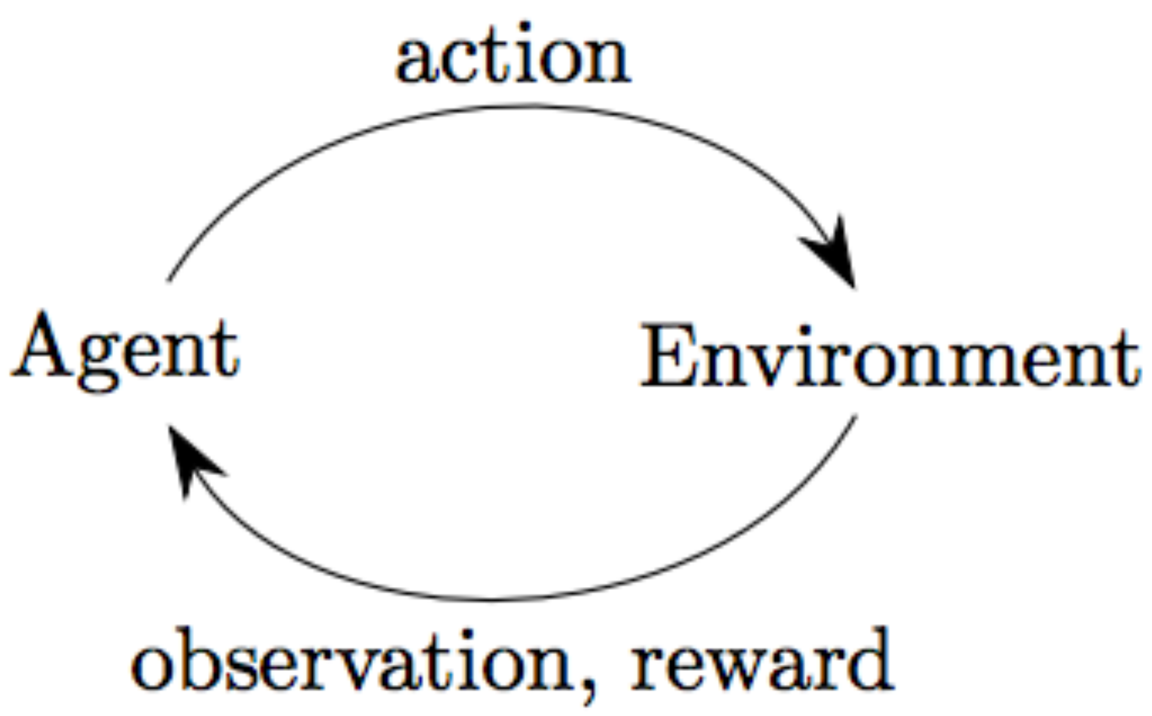


$$s_1, a_1, r_1, s_2, a_2, r_2, s_3, a_3, \ldots .$$

(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

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

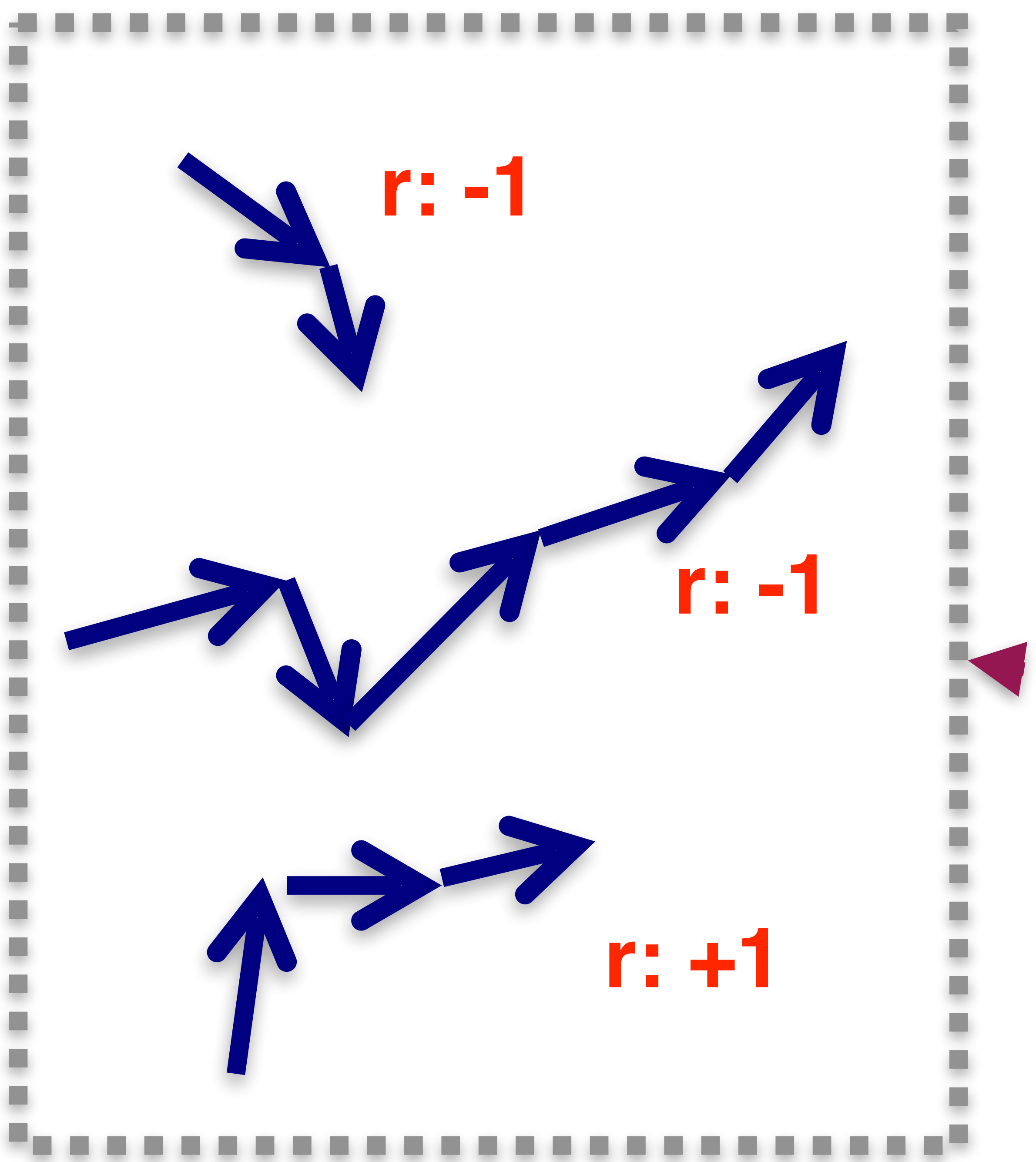
Infinite Time Horizon

$$\sum_t r_t$$

Finite Time Horizon

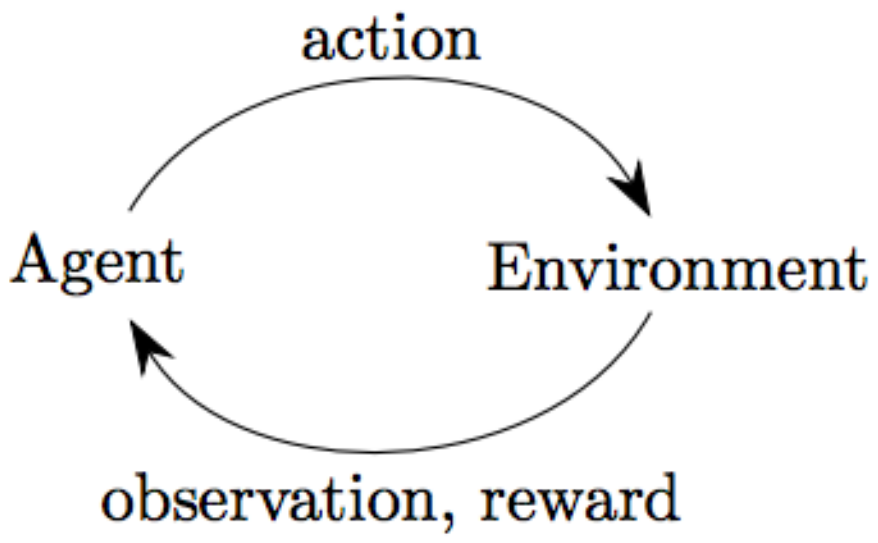
$$\sum_{t=1}^{T} r_t$$

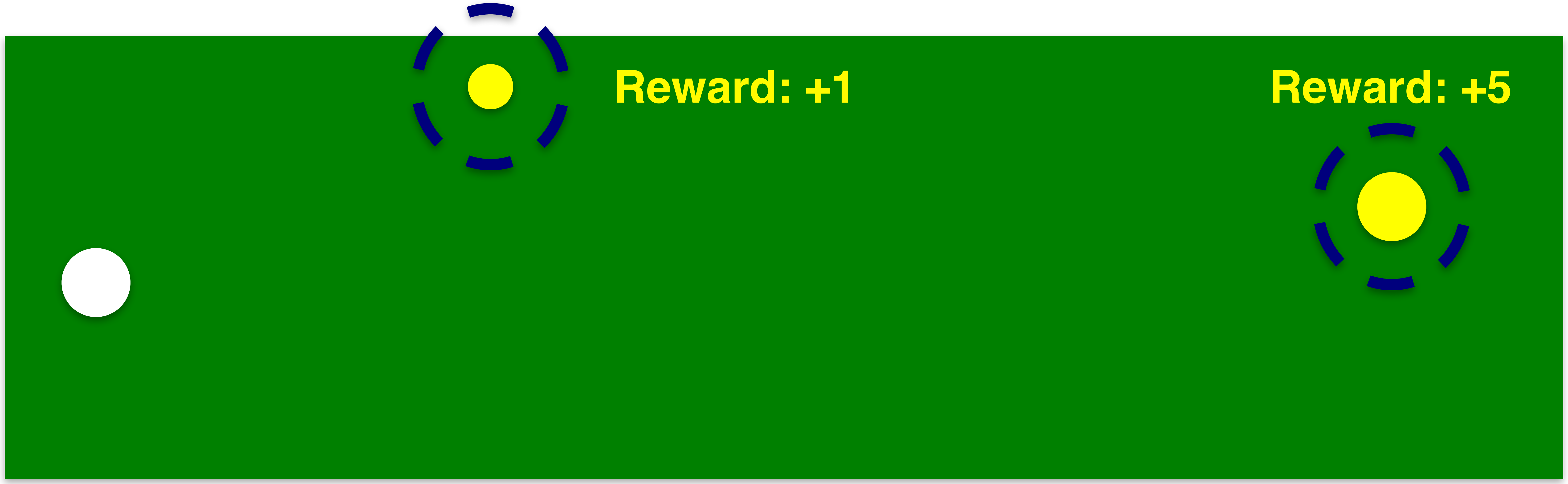
Dataset



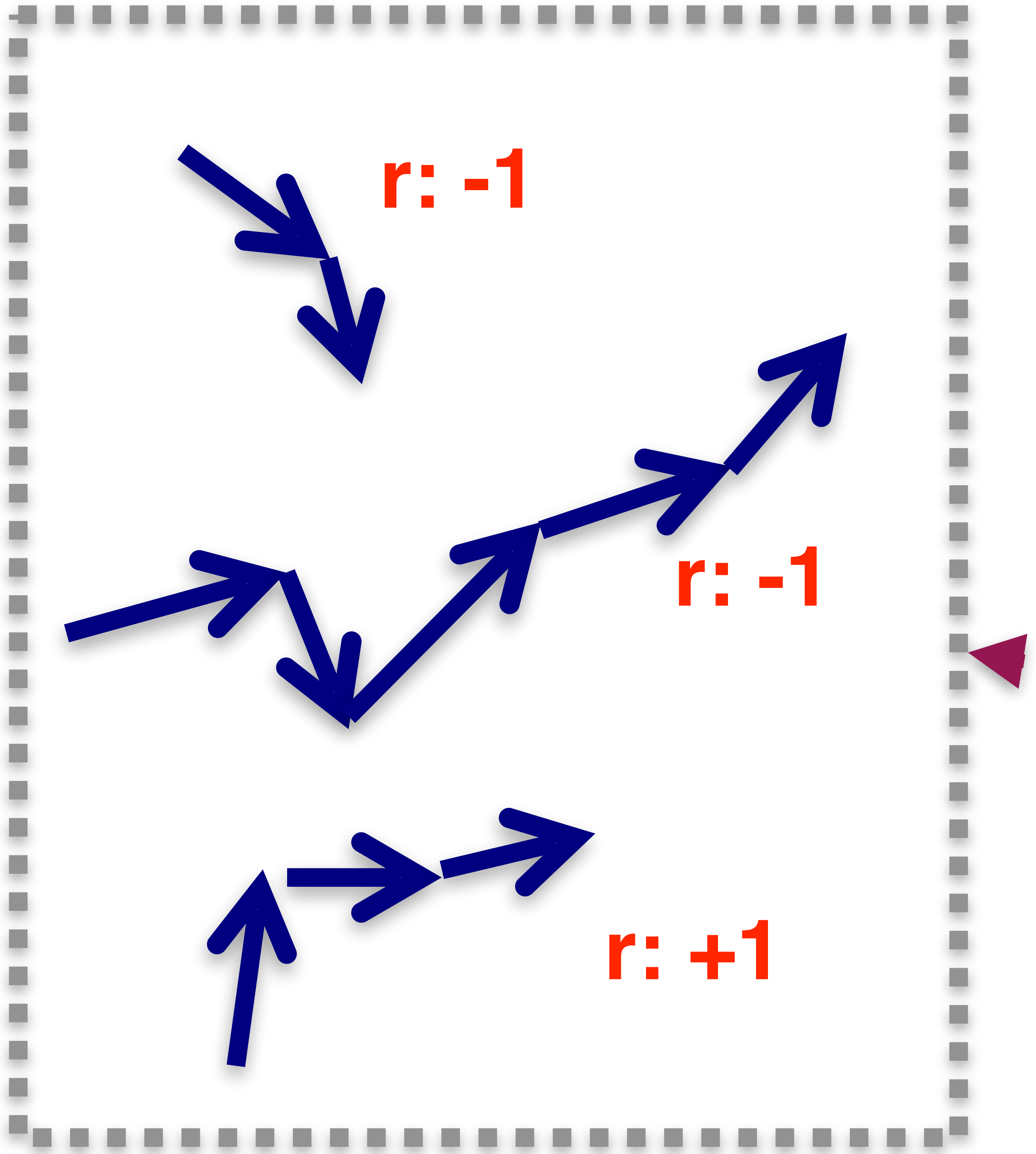
$$\sum_{t=1}^{T=100} r_t$$

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

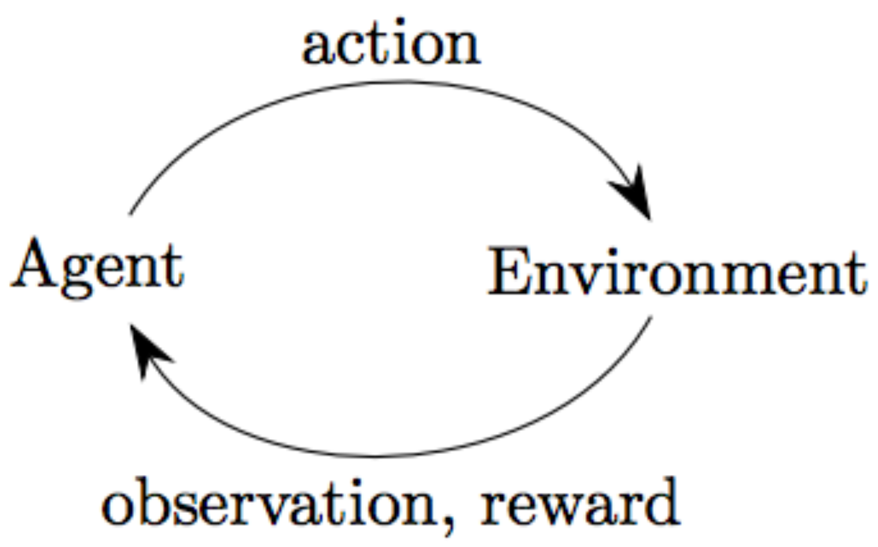


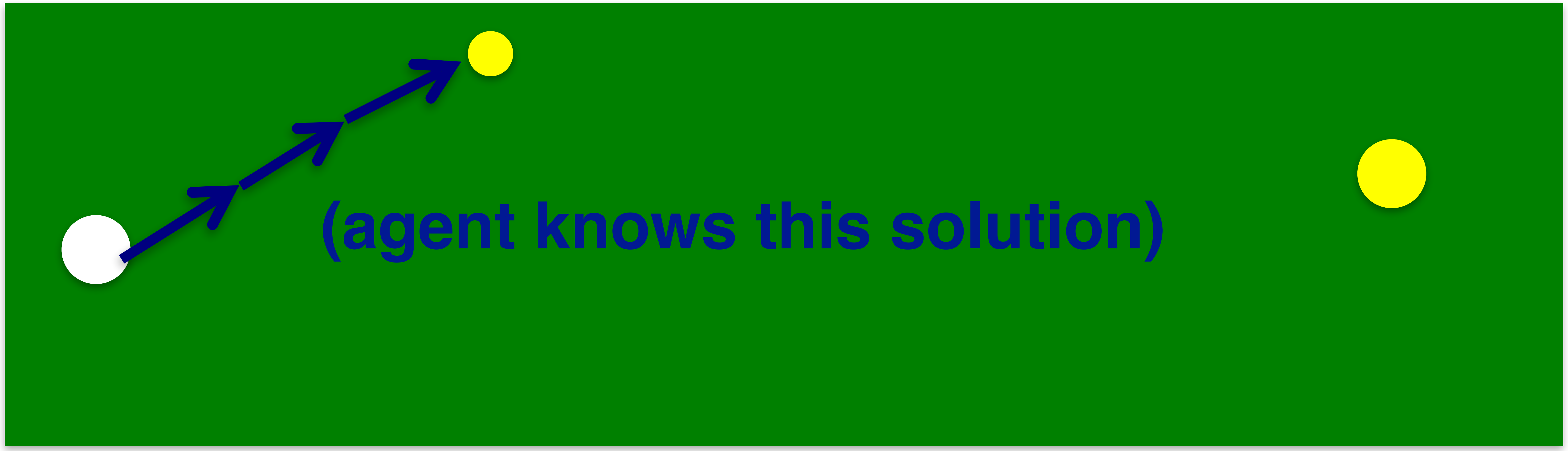


Dataset

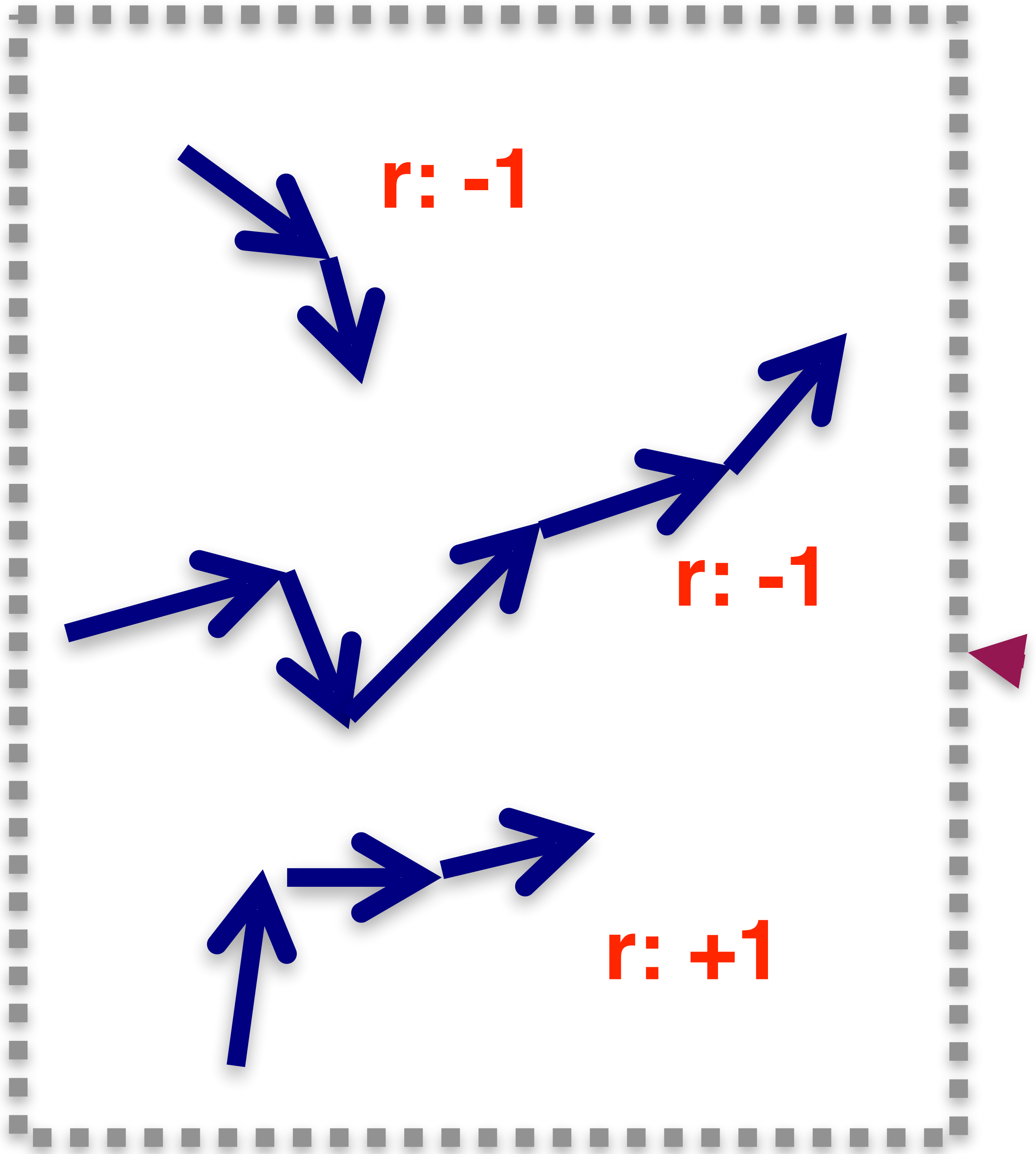


$$\sum_{t=1}^{T=100} r_t \qquad a_t = \pi_\theta(s_{0:t})$$

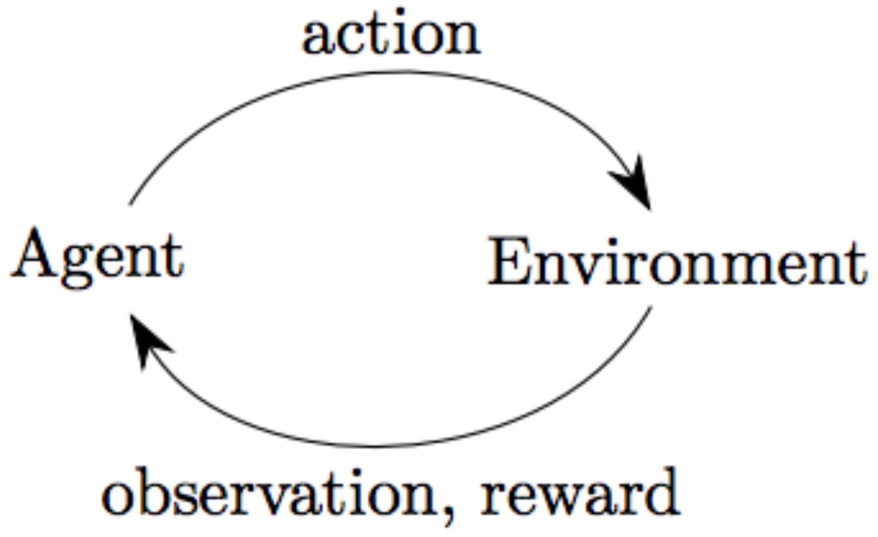




Dataset

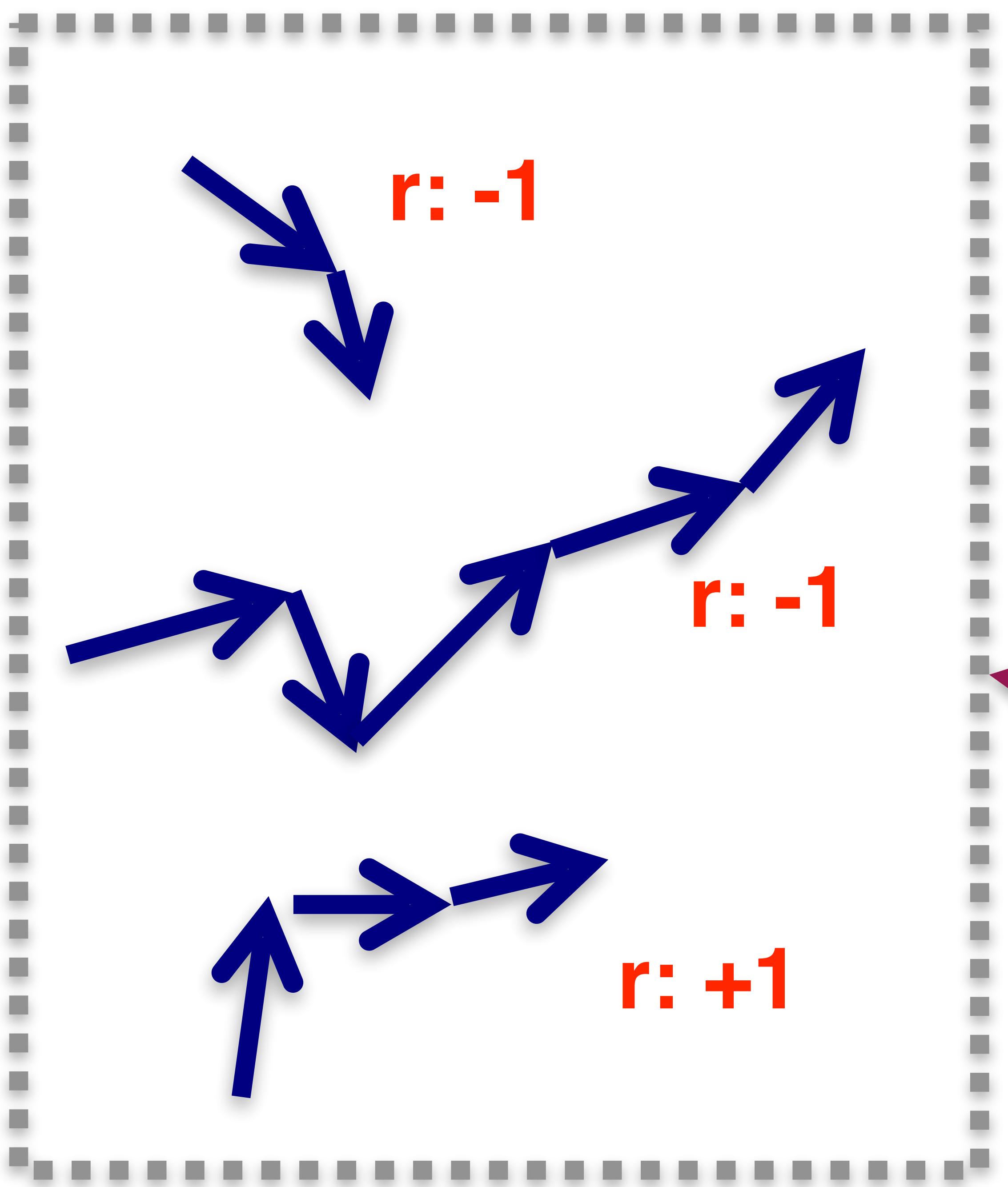


$$\sum_{t=1}^{T=100} r_t \qquad a_t = \pi_\theta(s_{0:t})$$

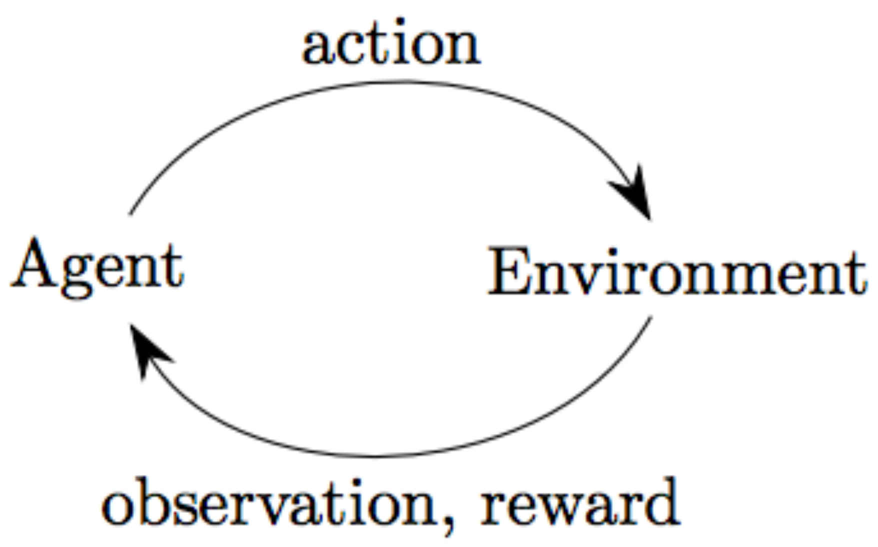


time t

Dataset

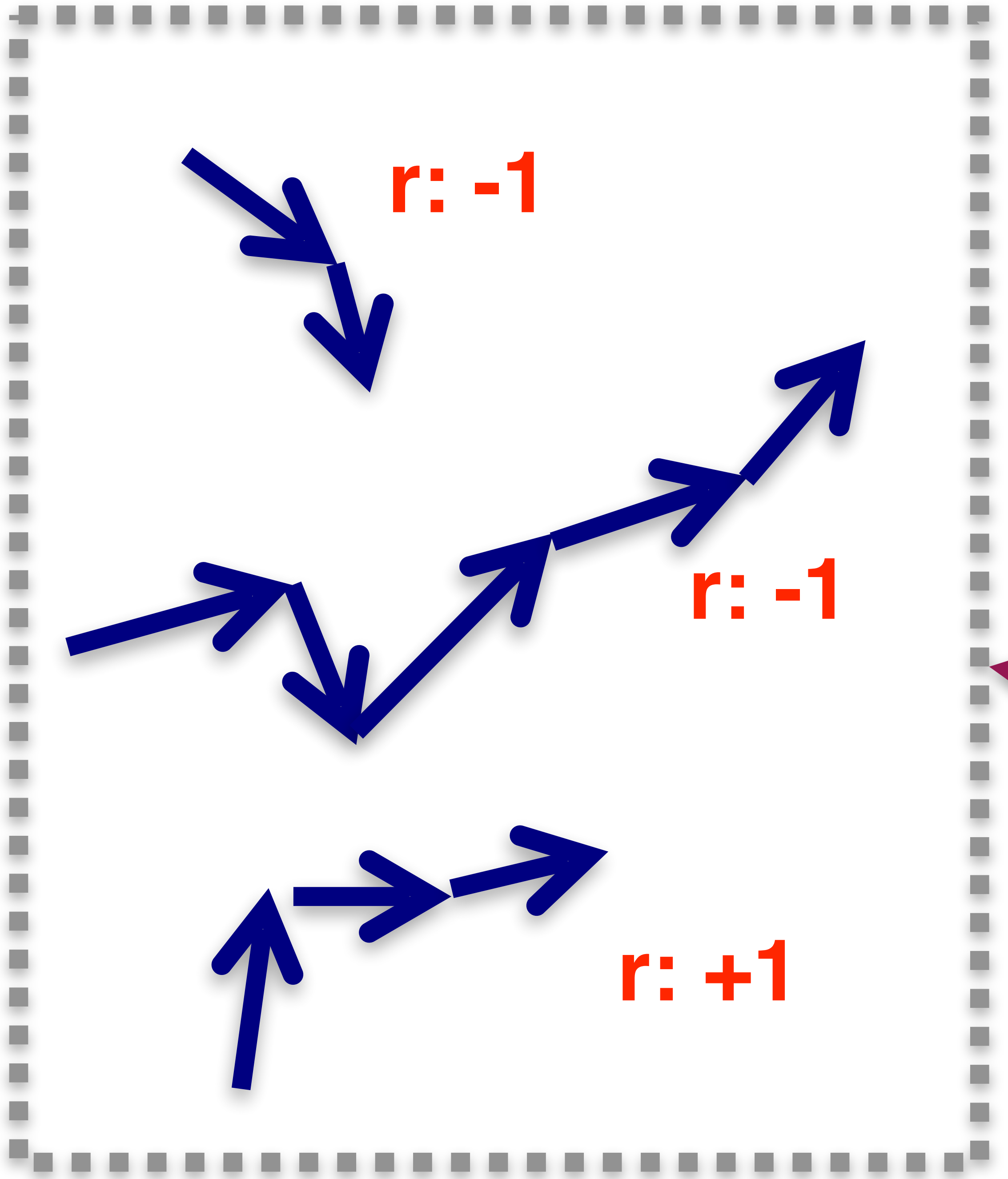


$$\sum_{t=1}^{T=100} r_t \qquad a_t = \pi_\theta(s_{0:t})$$

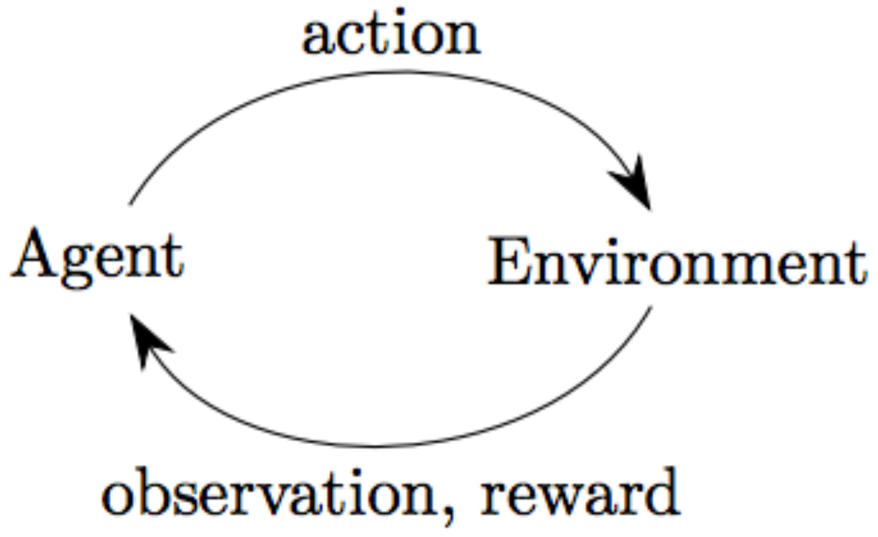


time t t=97

Dataset

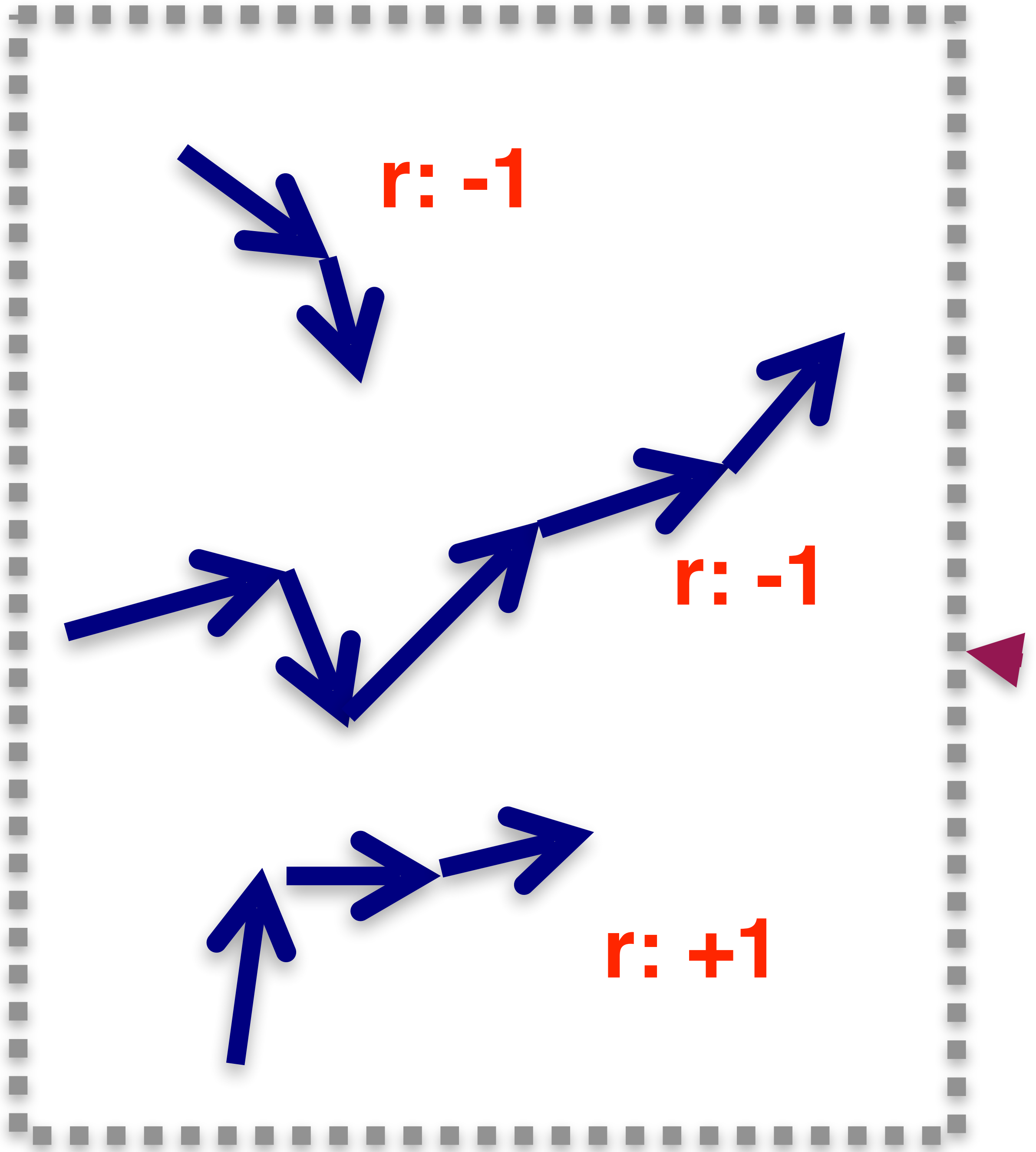


$$\sum_{t=1}^{T=100} r_t \qquad a_t = \pi_\theta(s_{0:t})$$

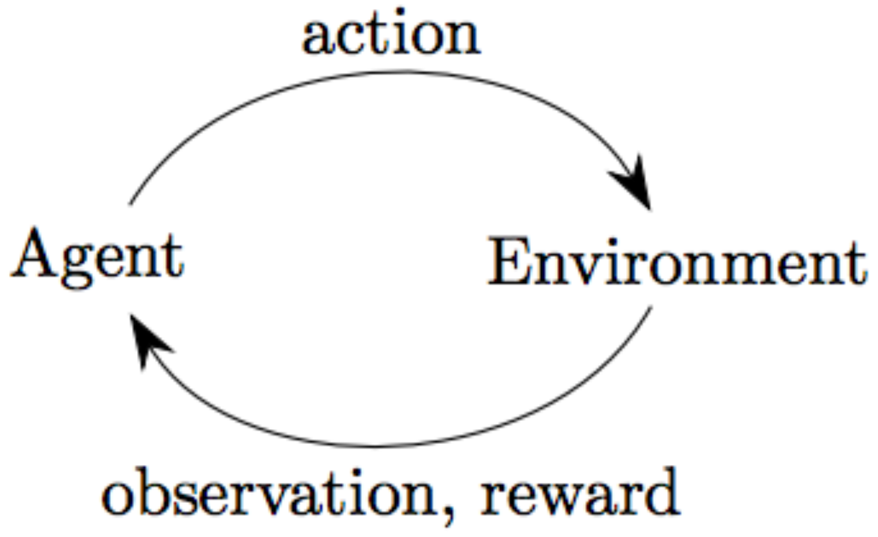


time t **t=97** **t=10**

Dataset



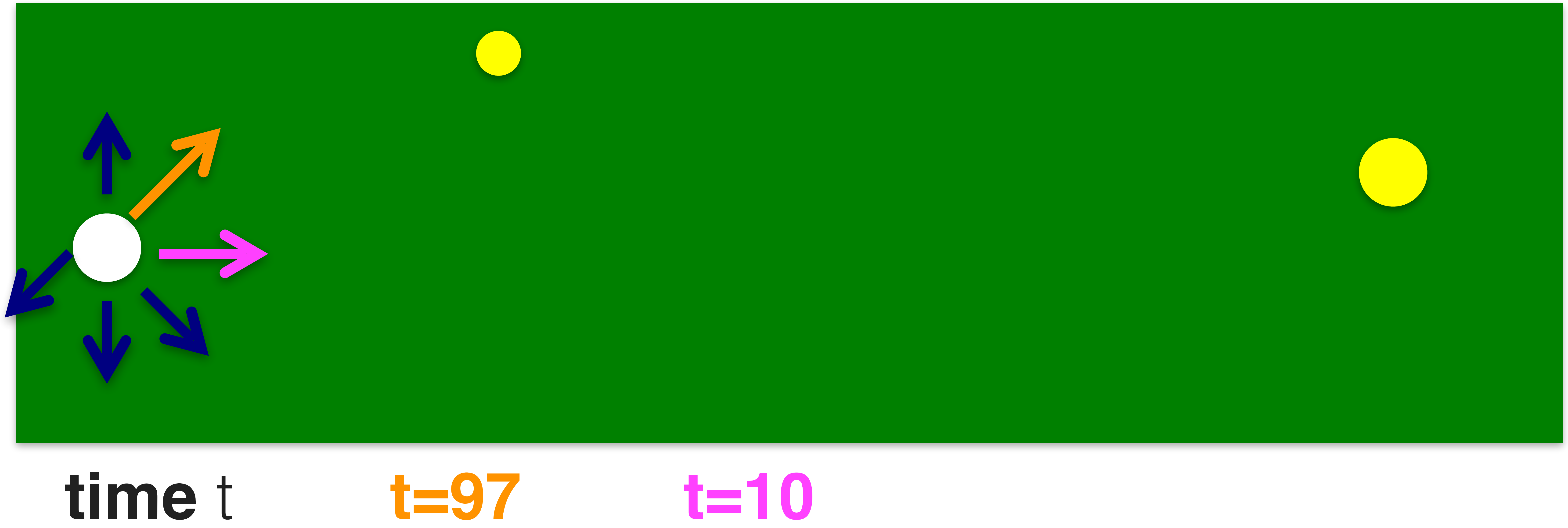
$$\sum_{t=1}^{T=100} r_t$$



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

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

(is this a problem?)



Goal

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

Finite Time Horizon

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

Infinite Time Horizon

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

$$0 < \gamma < 1$$

discount factor

Commonly Used

Maximizing Rewards

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

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

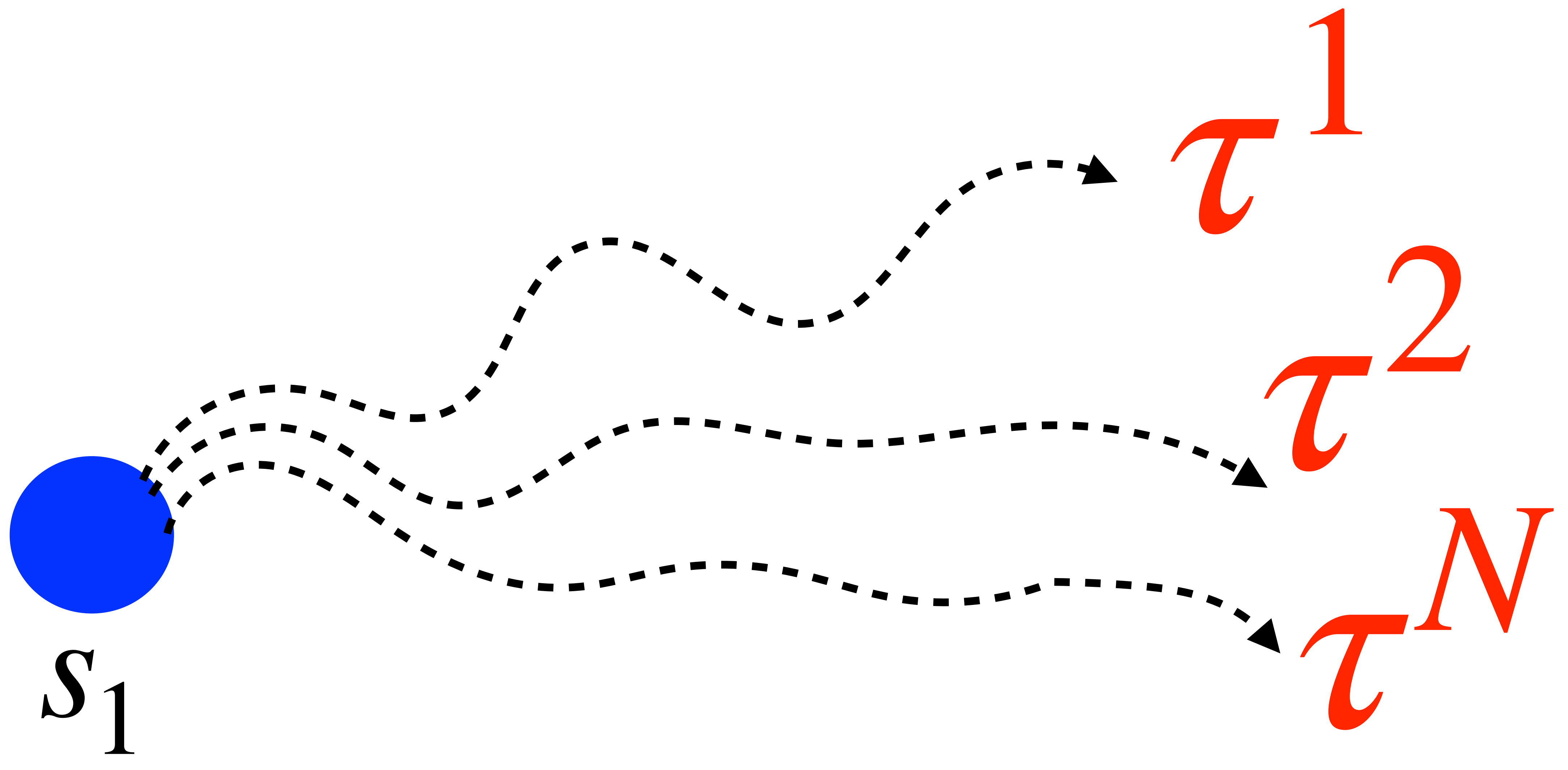
Why do we need probability of a rollout?

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

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

$$\downarrow$$

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

$$\downarrow$$

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

Average reward

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

Maximize Reward **Policy Gradients!**

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

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!

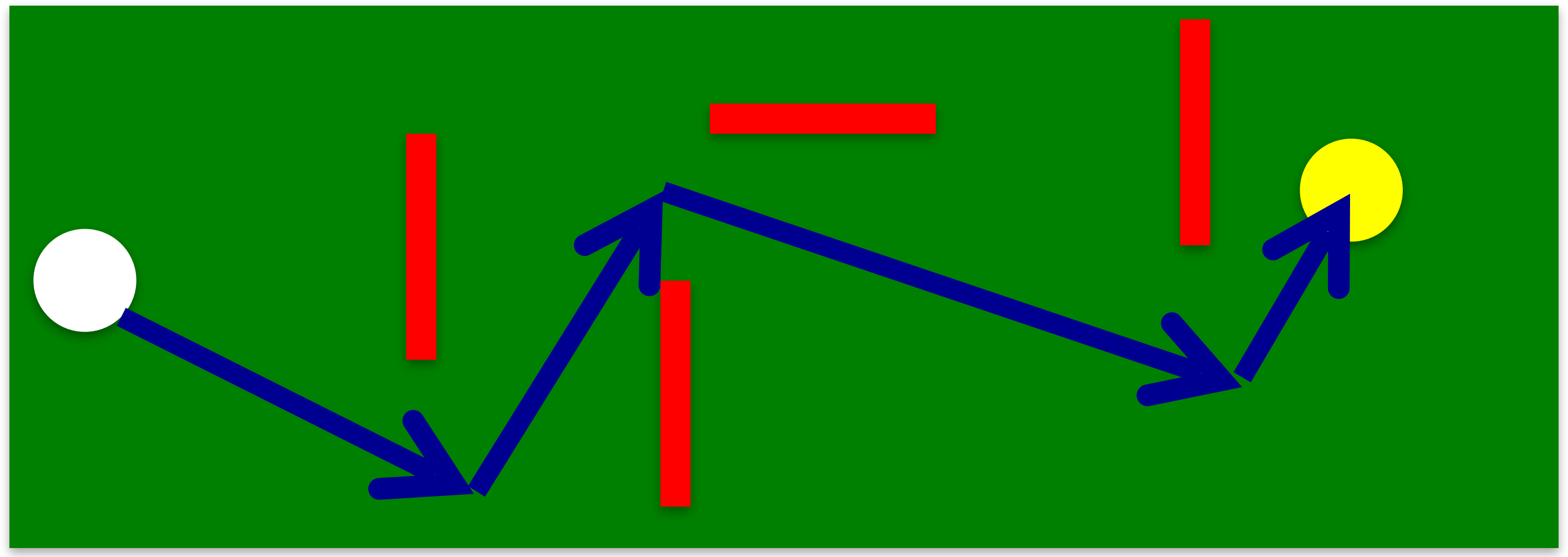
Policy Gradients

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

Intuitive Interpretation

Roll out multiple trajectories

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



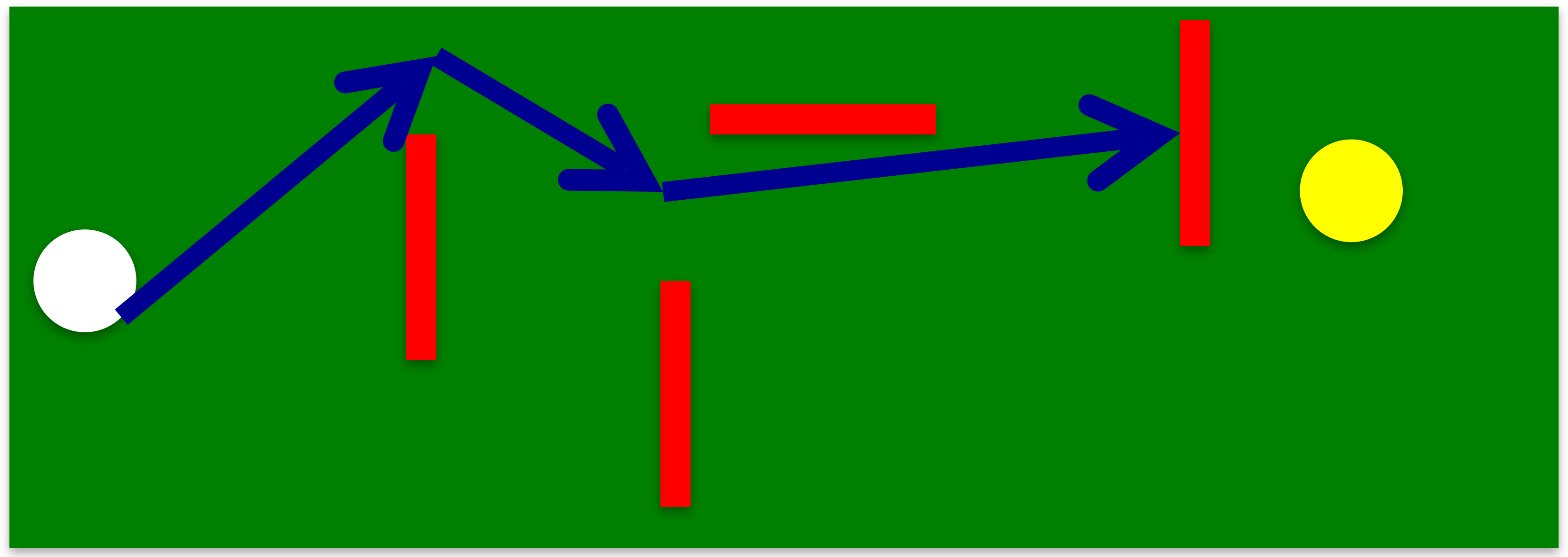
Policy Gradients

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

Intuitive Interpretation

Roll out multiple trajectories

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



↑ Increase log-prob by small amount

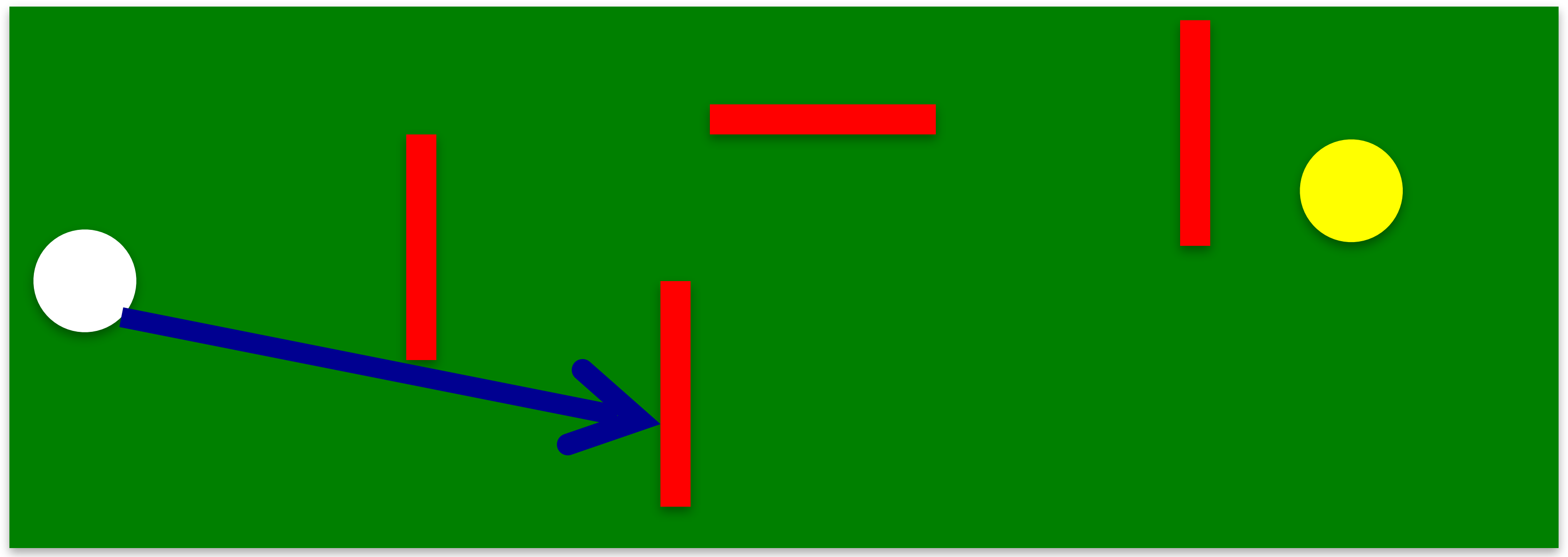
Policy Gradients

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

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 $p\left(s_t \mid s_{1:t-1}, a_{1:t-1}\right)$

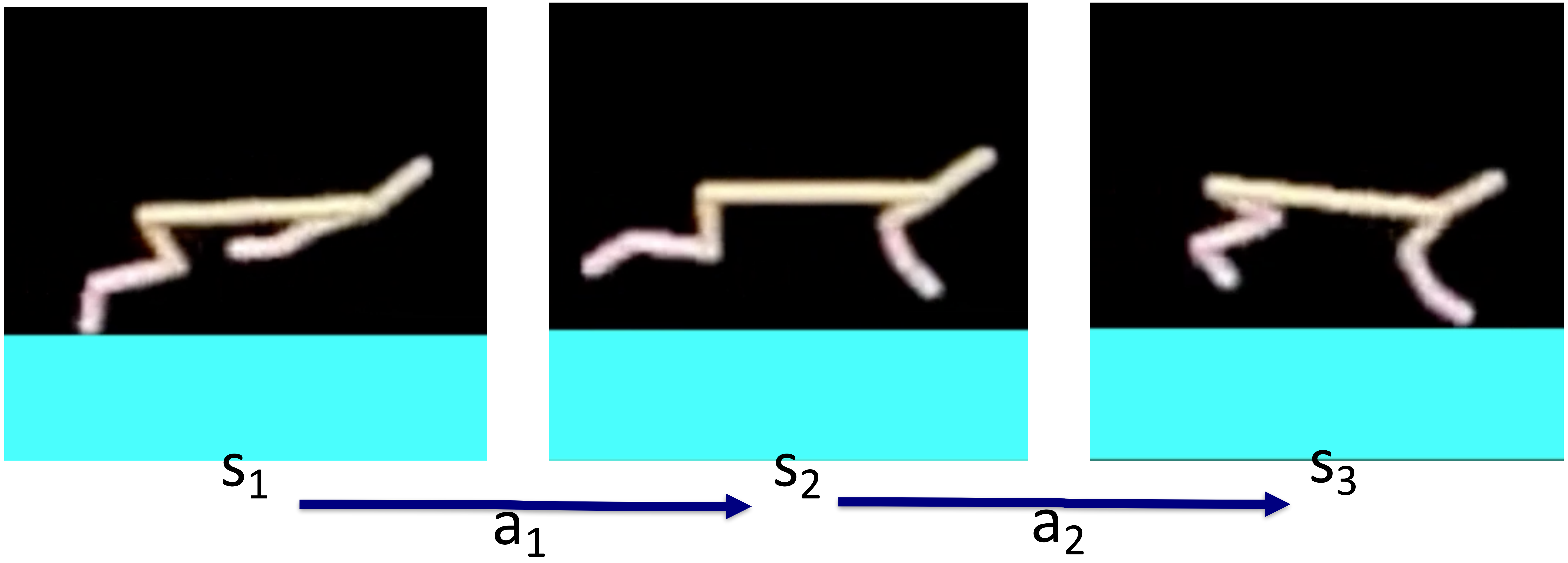
Expanding on Policy Gradients

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

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 …)

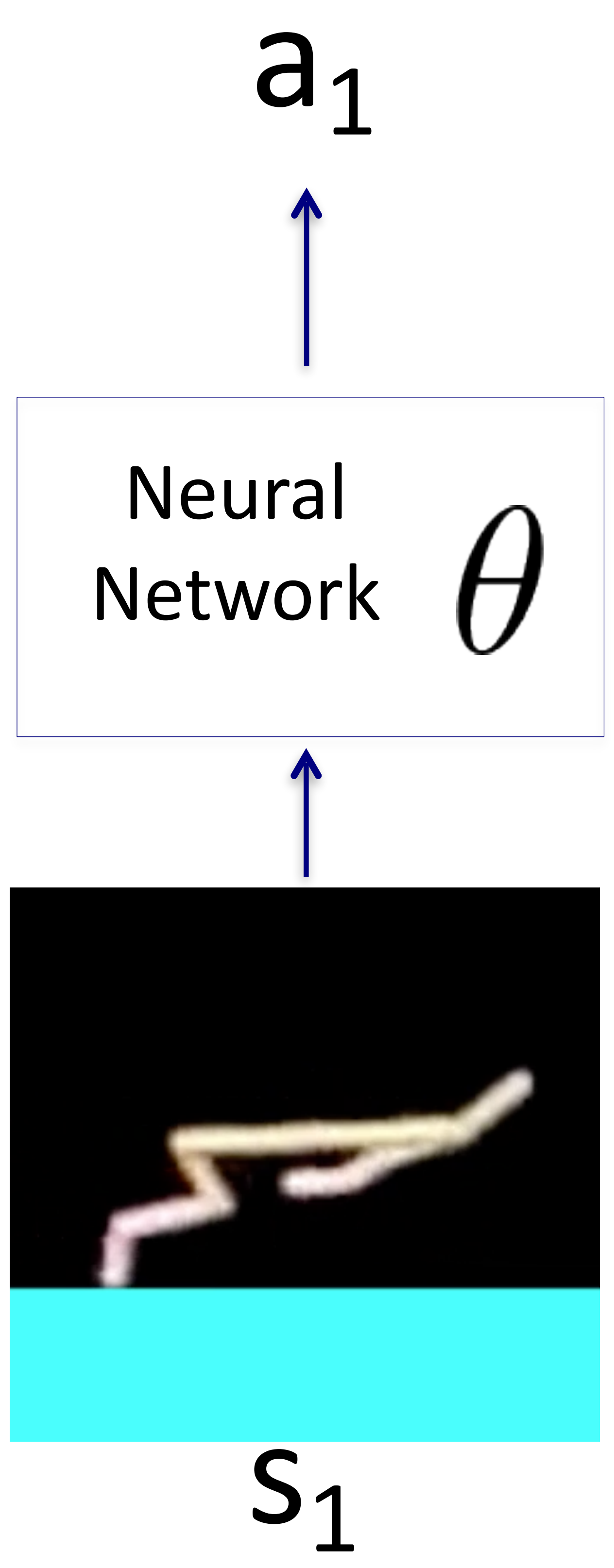
desired joint position

Rewards (r_1, r_2 …)

Speed of the Cheetah

Illustration of Policy Gradients

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



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

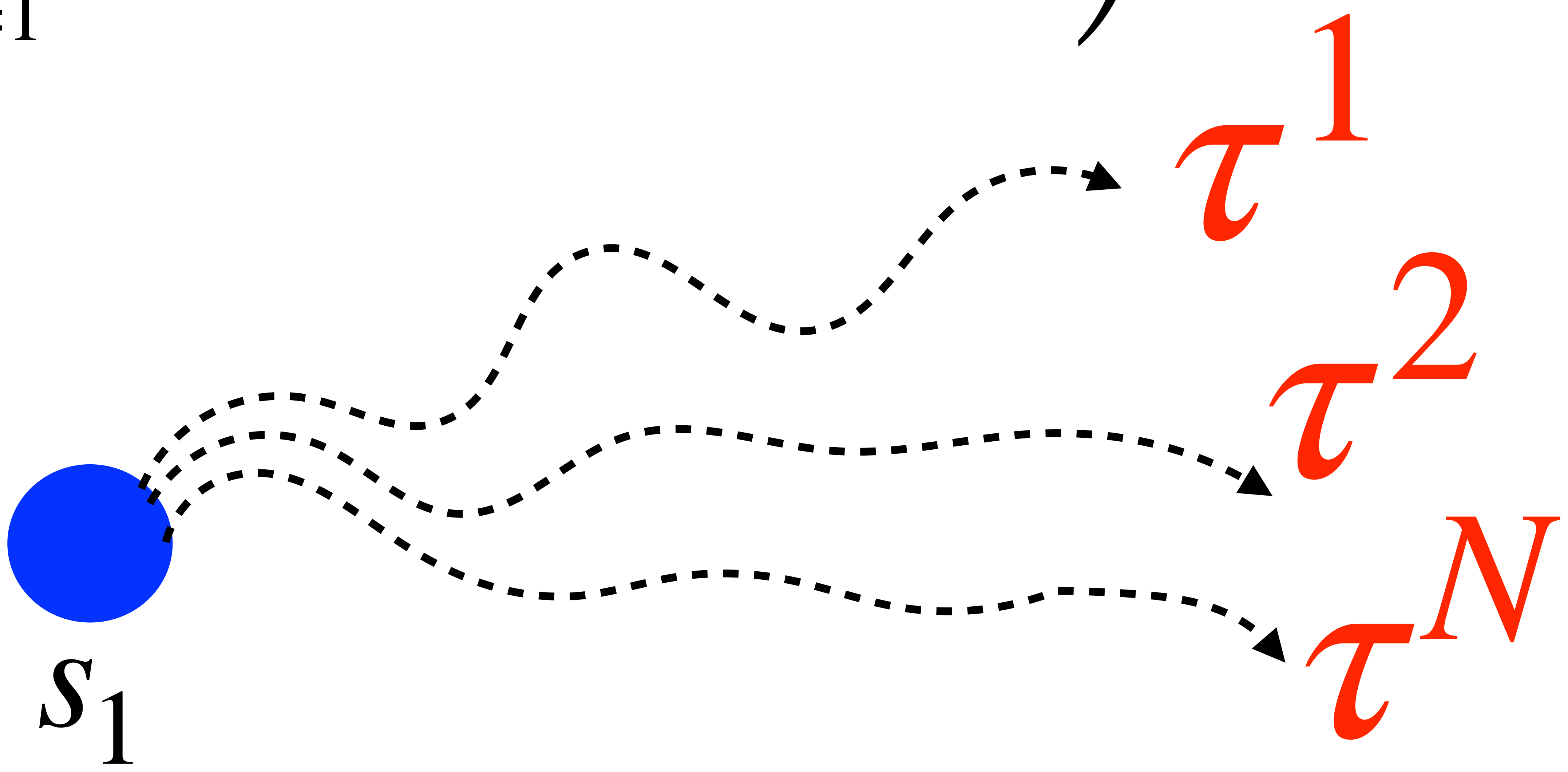
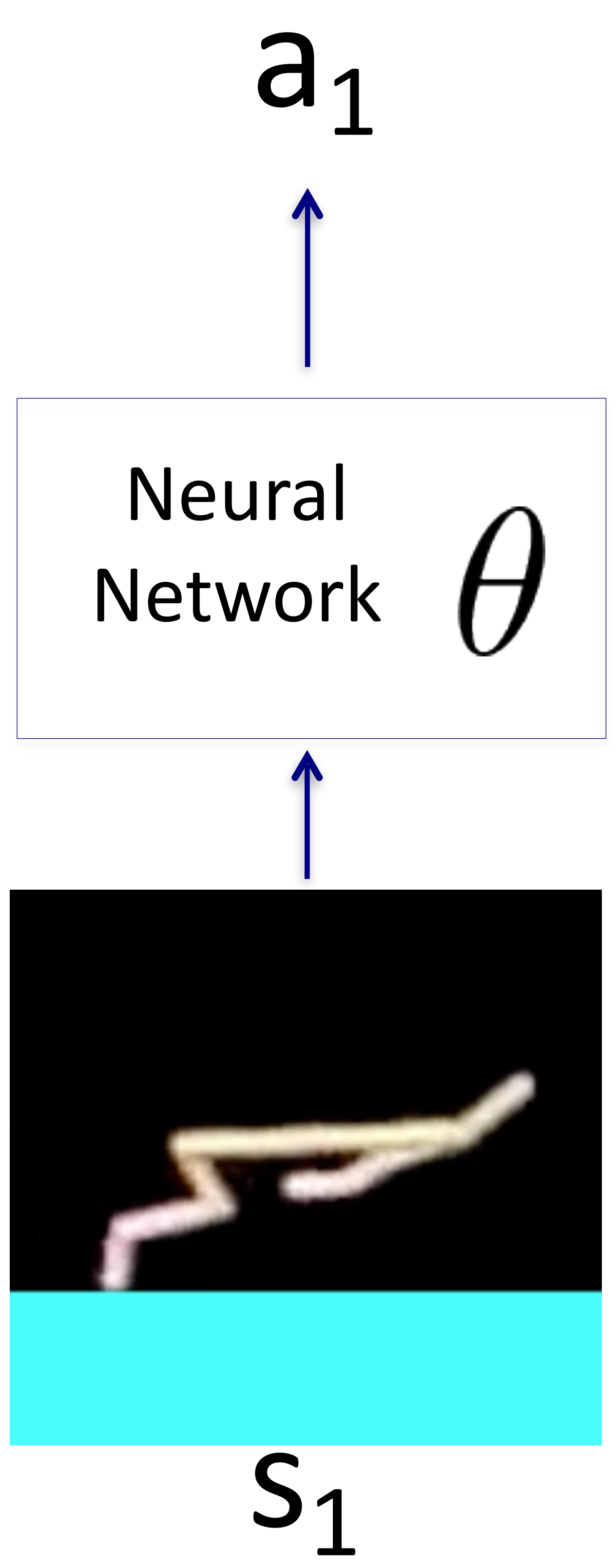


Illustration of Policy Gradients

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



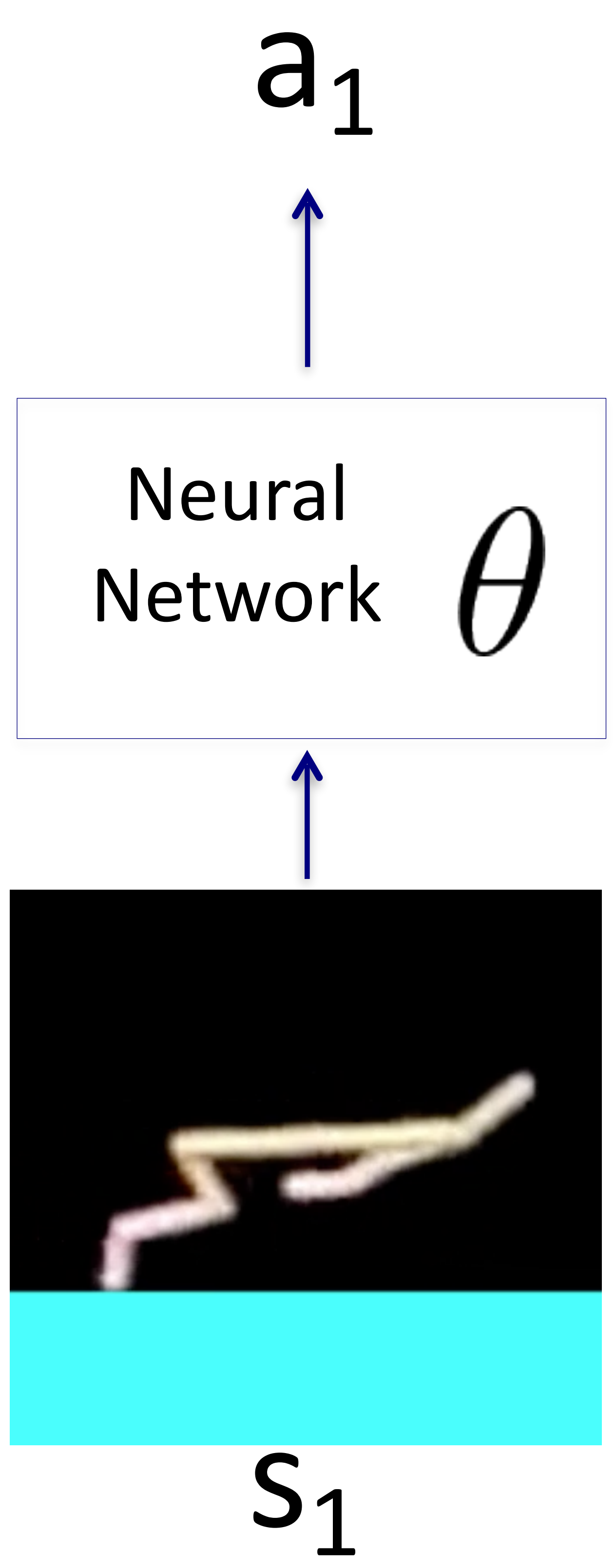
$$\frac{1}{N}\sum_{i=1}^{N}\left(\boxed{\sum_{t=1}^{T}}\left(\nabla_\theta \log \pi_\theta\left(a_t^i \mid s_t^i\right)\right)R(\tau)\right)$$

in practice can't roll out until infinity

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

Illustration of Policy Gradients

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

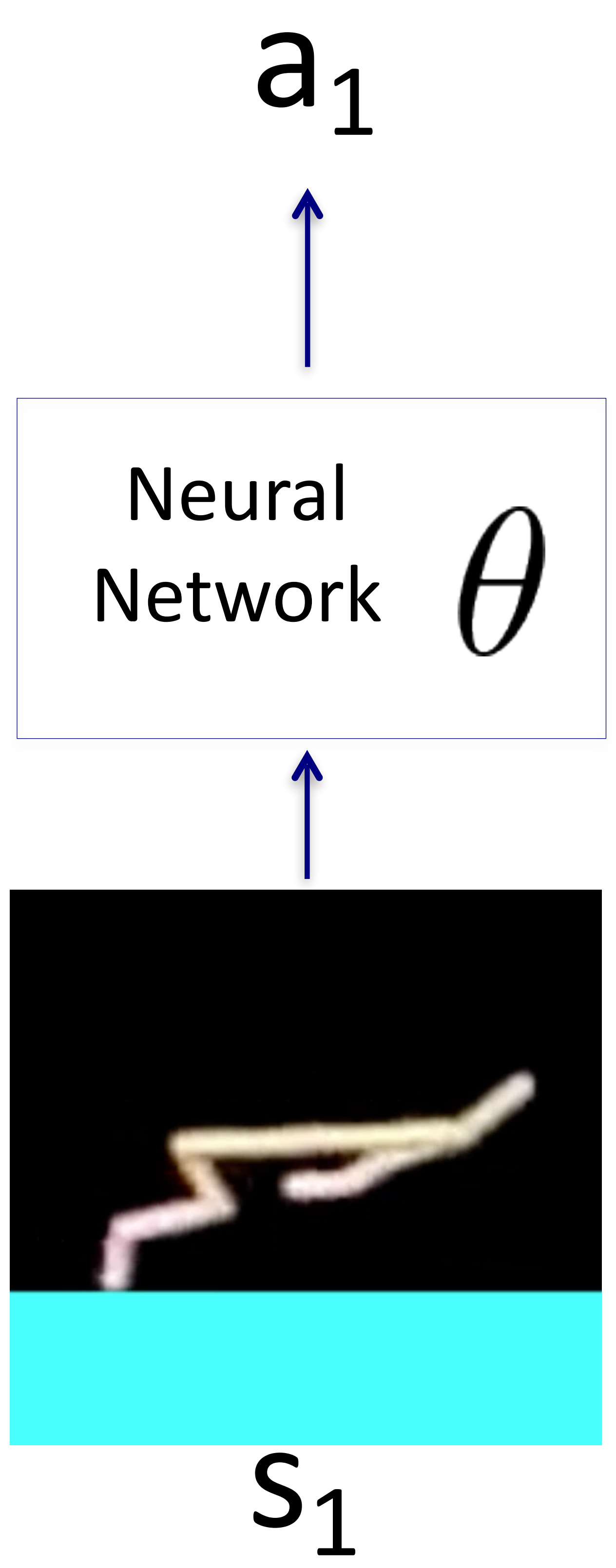


$$\frac{1}{N}\sum_{i=1}^{N}\left(\boxed{\sum_{t=1}^{T}}\left(\nabla_\theta \log \pi_\theta\left(a_t^i \mid s_t^i\right)\right)\boxed{R^\gamma(\tau)}\right)$$

in practice can't roll out until infinity

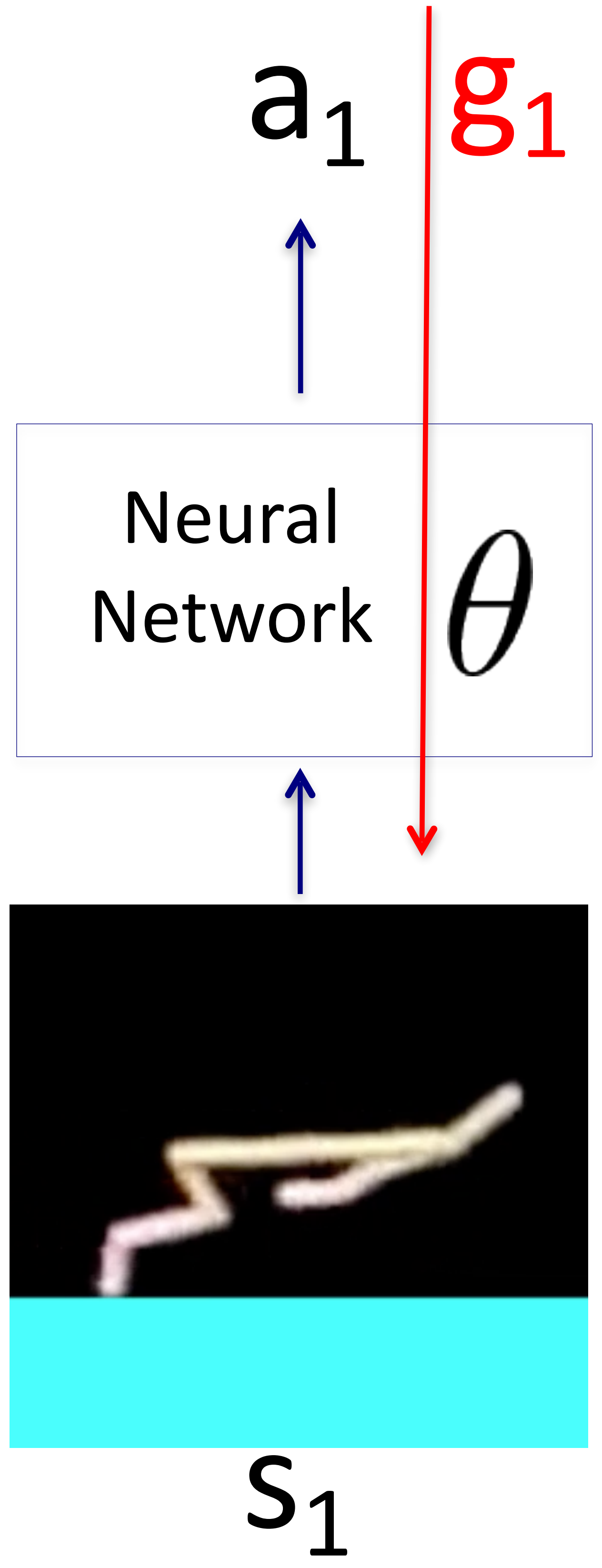
Treat **finite** horizon as **infinite** horizon with discount

Illustration of Policy Gradients



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

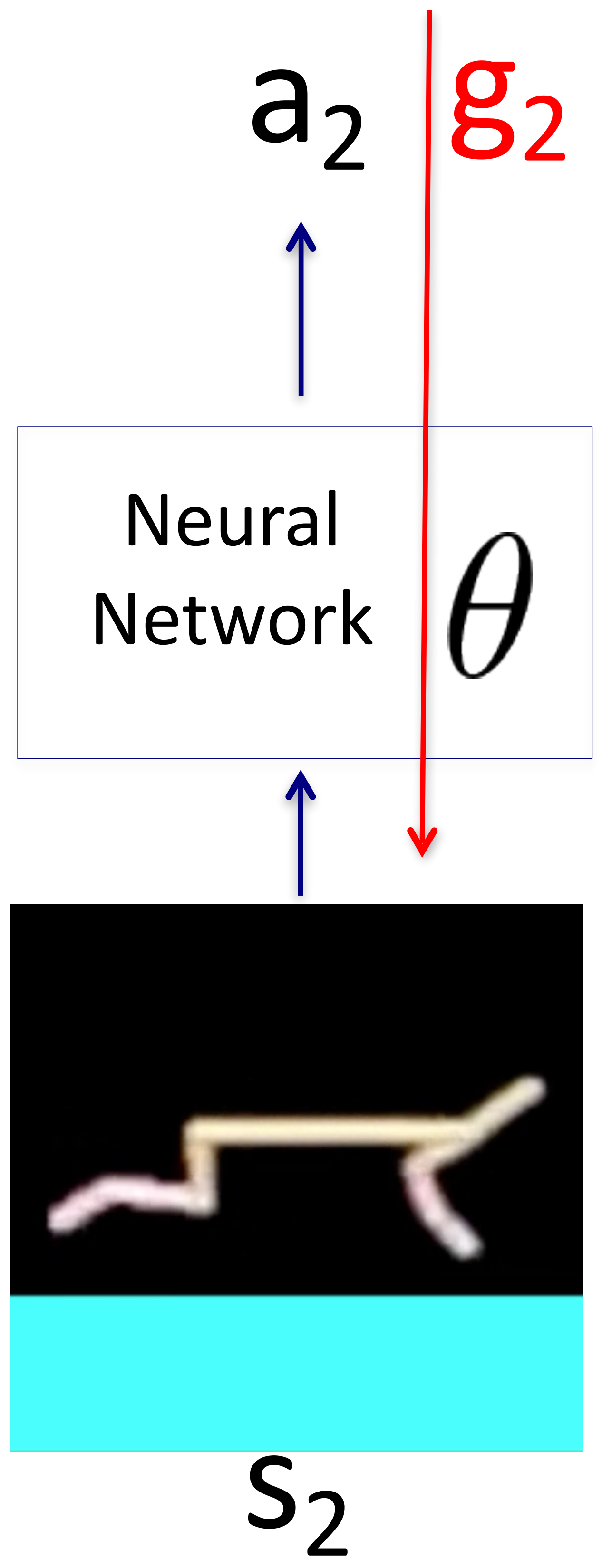
Illustration of Policy Gradients



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

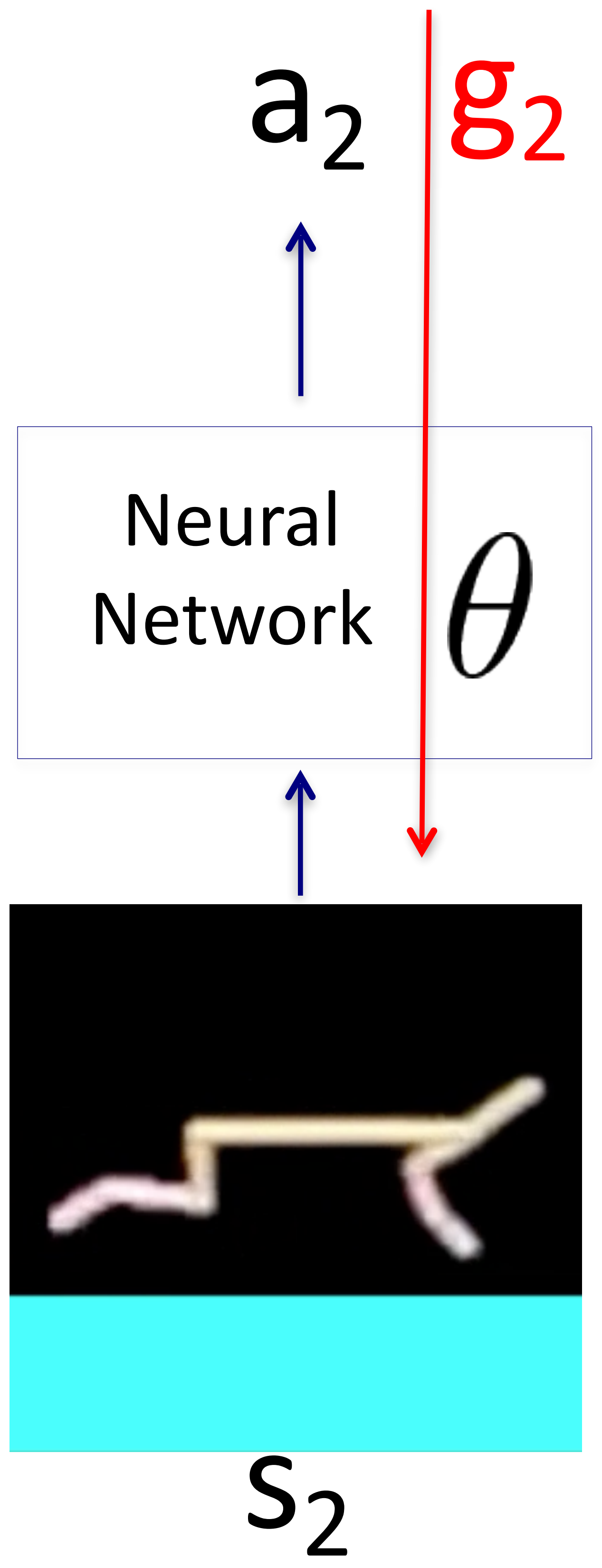
$$g_1 = \nabla_\theta \log \pi_\theta(a_1|s_1)$$

Illustration of Policy Gradients



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

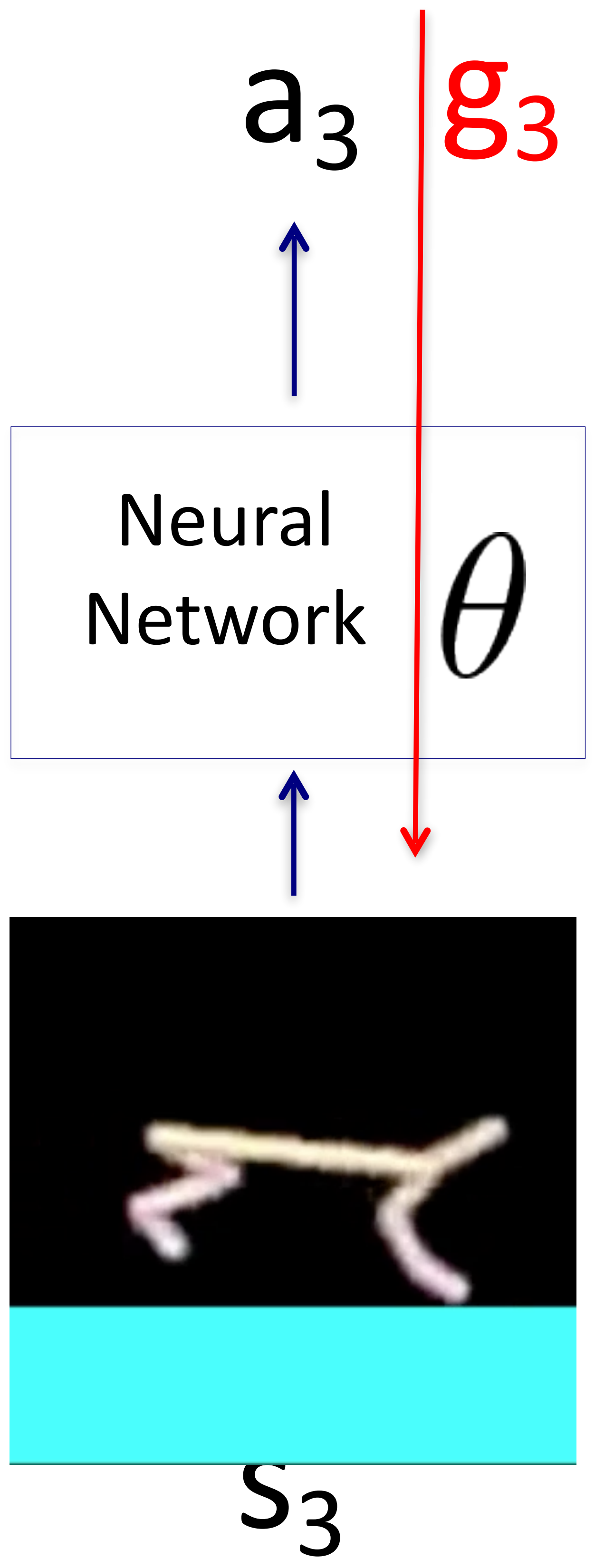
Illustration of Policy Gradients



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

$G = g_1 + g_2$

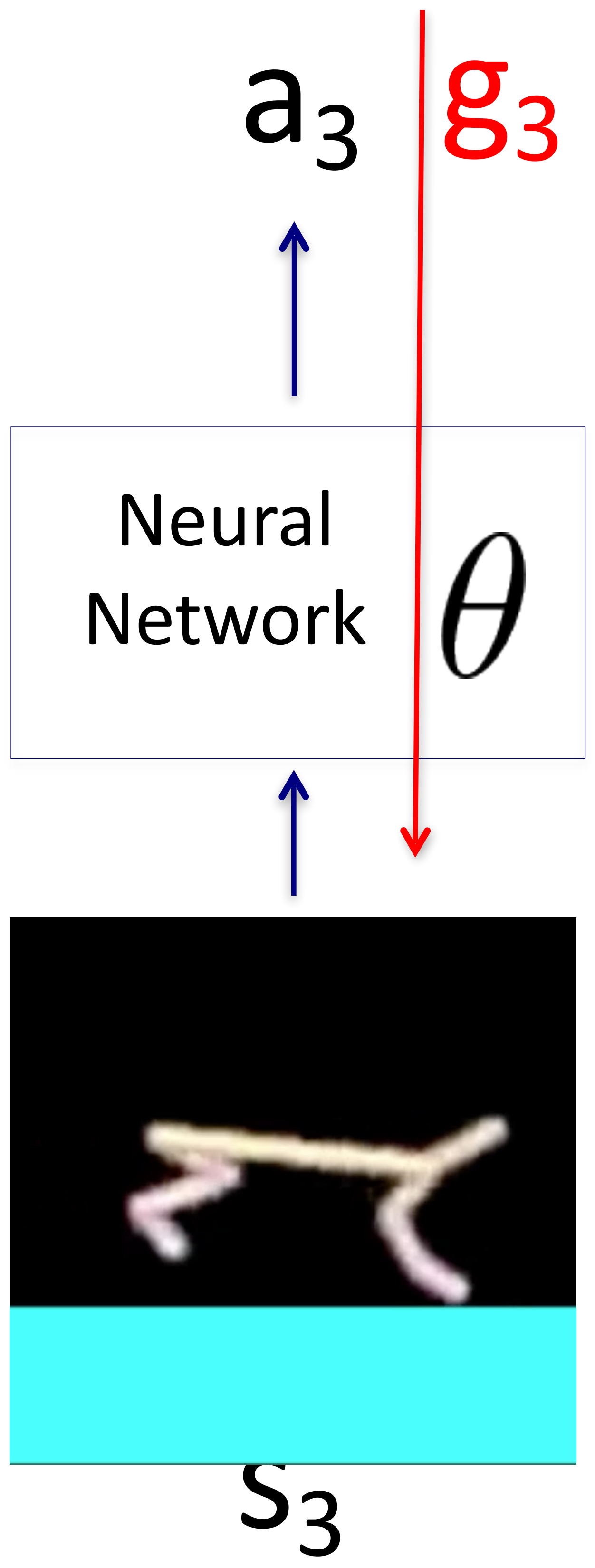
Illustration of Policy Gradients



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

$G = g_1 + g_2 + g_3$

Illustration of Policy Gradients



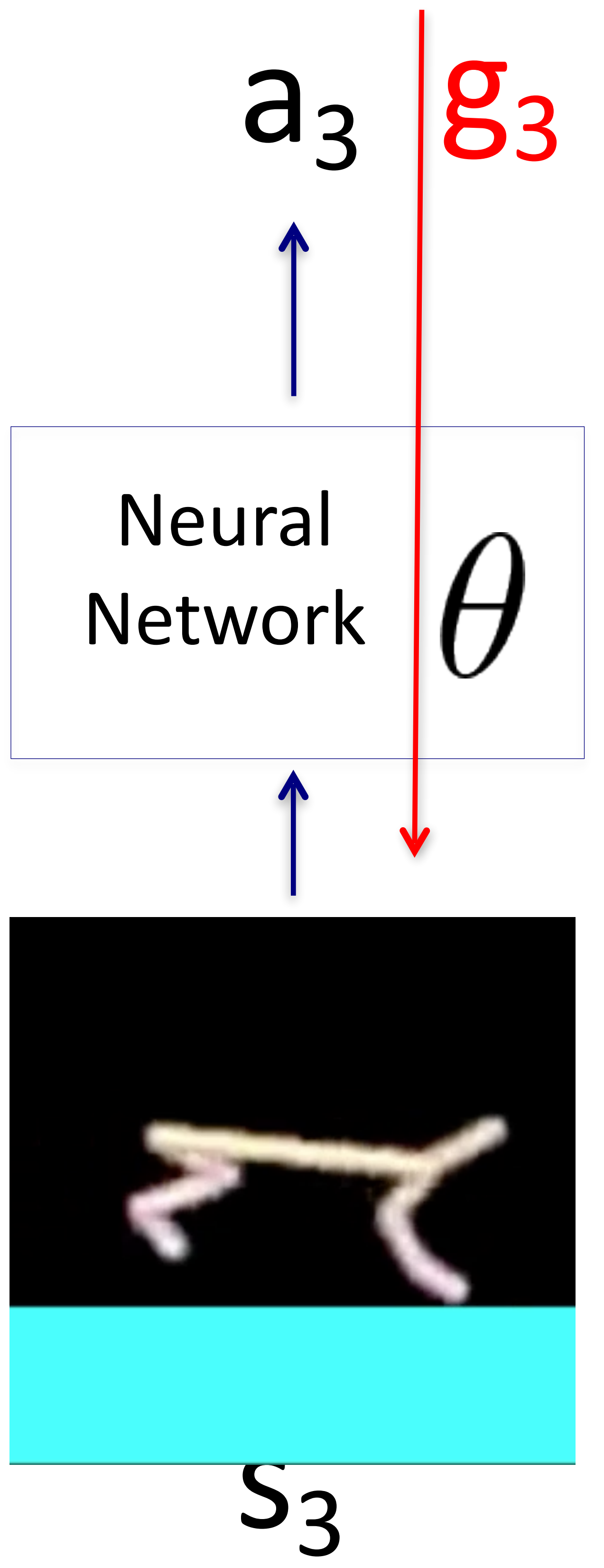
$$\frac{1}{N}\sum_{i=1}^{N}\left(\underbrace{\boxed{\sum_{t=1}^{T}\left(\nabla_\theta \log \pi_\theta(a_t^i \mid s_t^i)\right)}}_{G}\underbrace{\boxed{R^\gamma(\tau)}}_{V}\right)$$

Sum of velocities across time

$G = g_1 + g_2 + g_3$

Illustration of Policy Gradients

This is also called the REINFORCE Algorithm

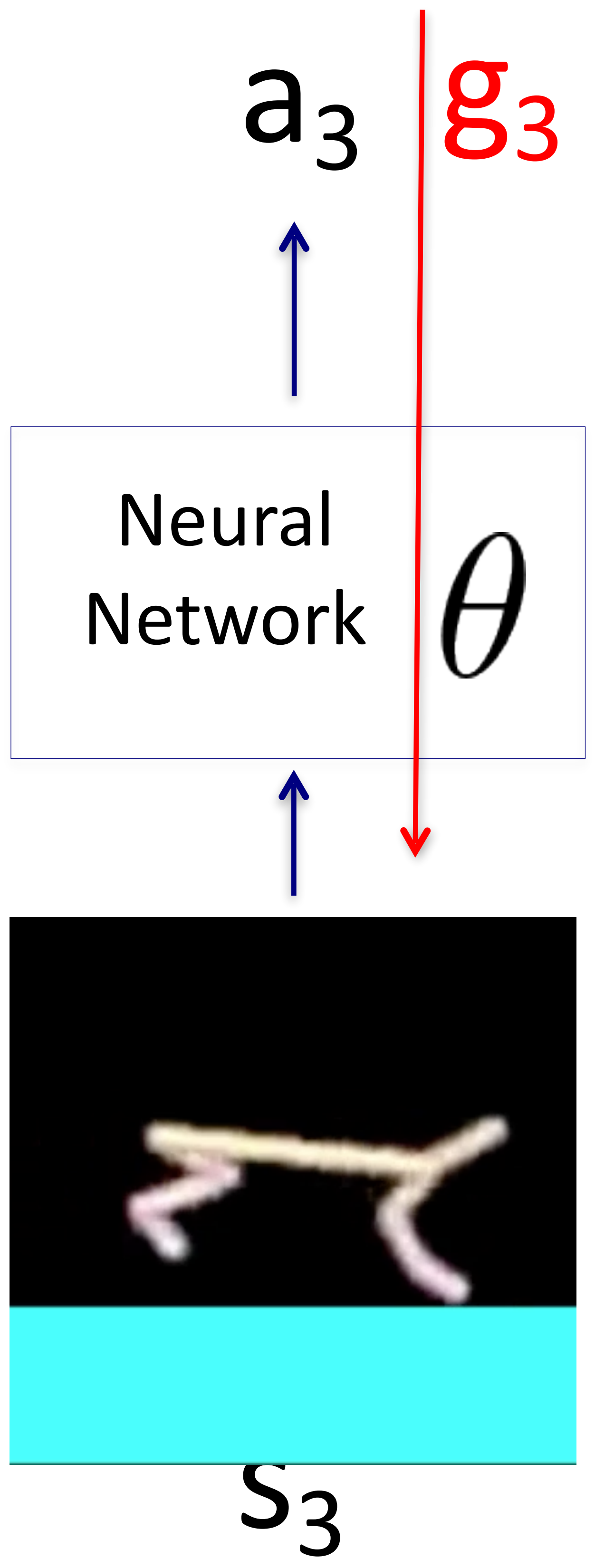


$$\frac{1}{N}\sum_{i=1}^{N}\left(\underbrace{\boxed{\sum_{t=1}^{T}\left(\nabla_\theta \log \pi_\theta\left(a_t^i \mid s_t^i\right)\right)}}_{G}\underbrace{\boxed{R^\gamma(\tau)}}_{v}\right)$$

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

$$\sum_t r_t$$

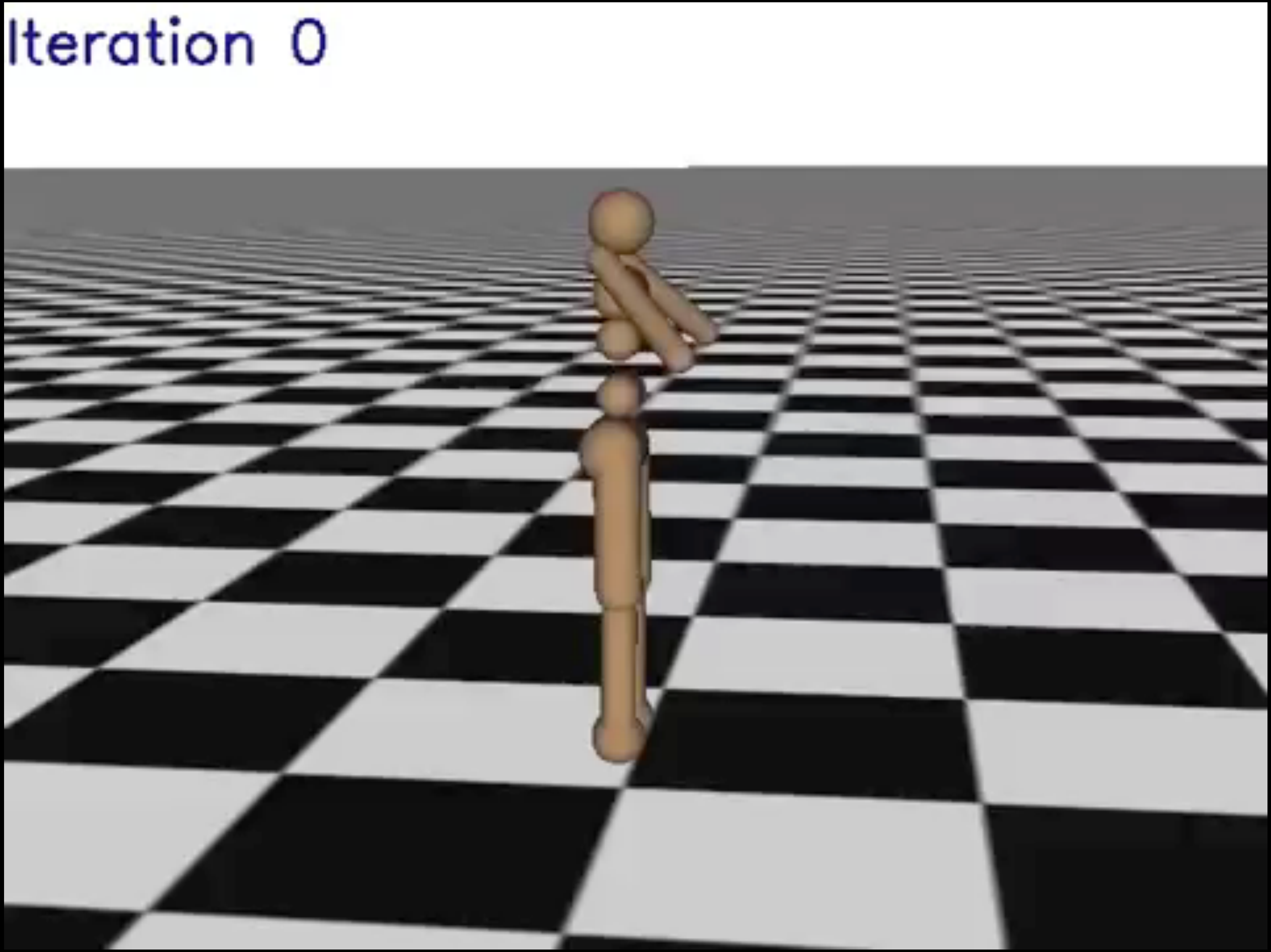
$$\tau^{gt} = (s_1, a_1^{gt}, s_2, a_2^{gt}, \ldots)$$

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

$$E_{\tau^{gt}}[\nabla_\theta \left(\log p_\theta(\tau^{gt})\right)]$$

Policy Gradients

Maximum Likelihood



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

The Idea of Episode

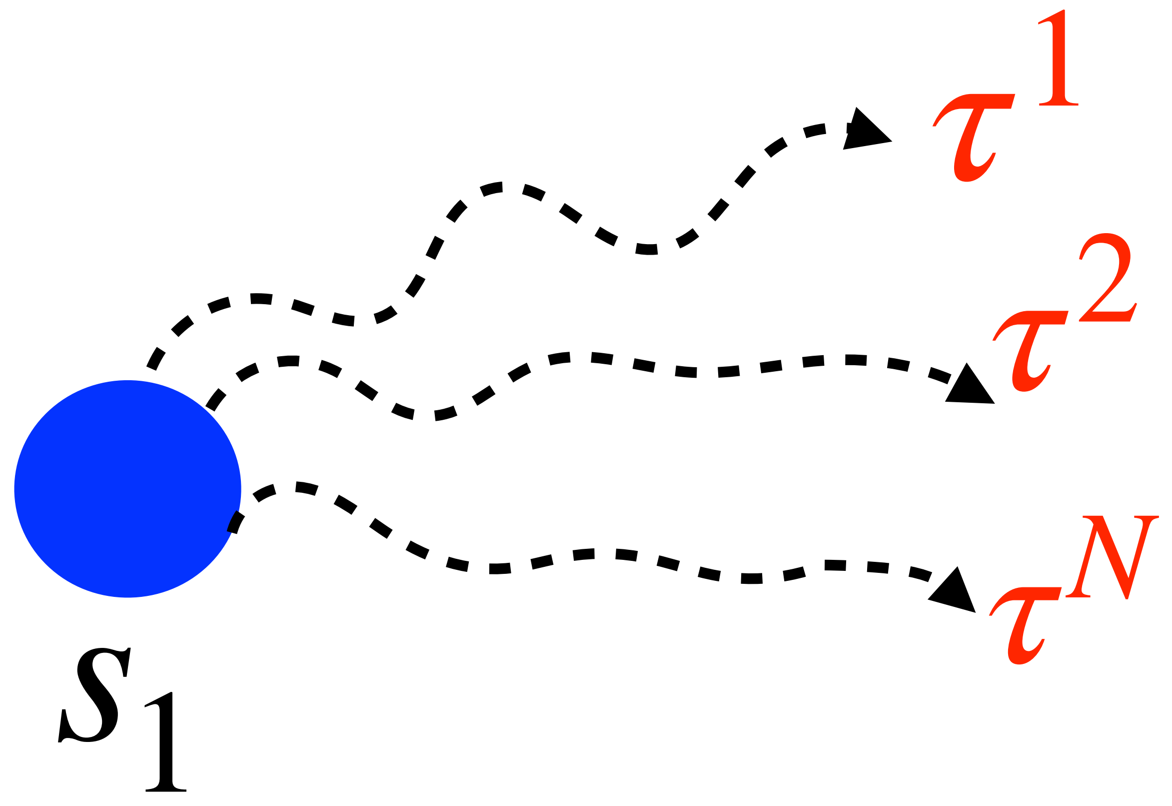
$$E_\tau[\nabla_\theta(\log p_\theta(\tau))R(\tau)]$$

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

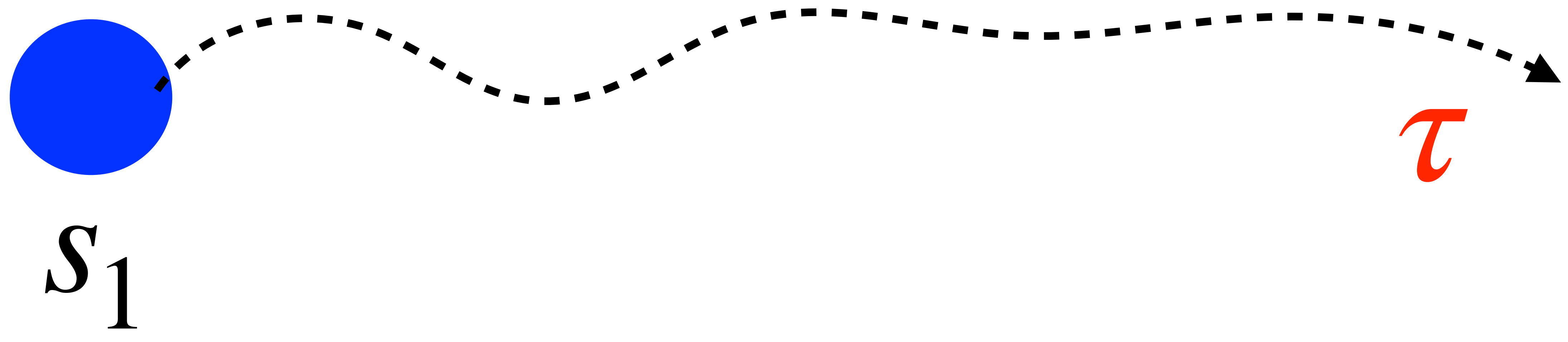
$$\frac{1}{N}\left(\sum_{t=1}^{NT}\left(\nabla_\theta \log \pi_\theta(a_t \mid s_t)\right)R(\tau)\right)$$

One Episode

Why define episodes?



N Episodes



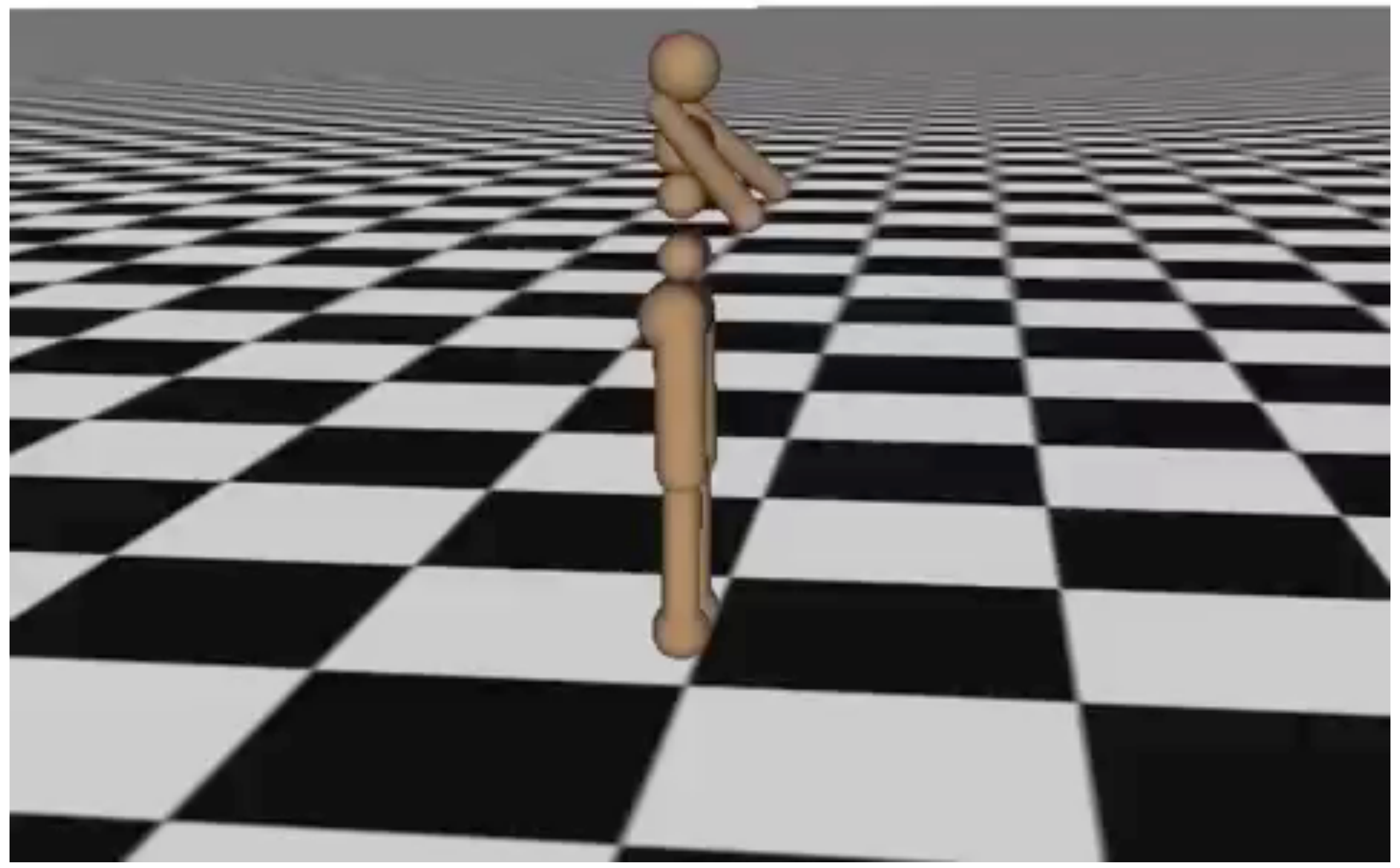
The Idea of Episode

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

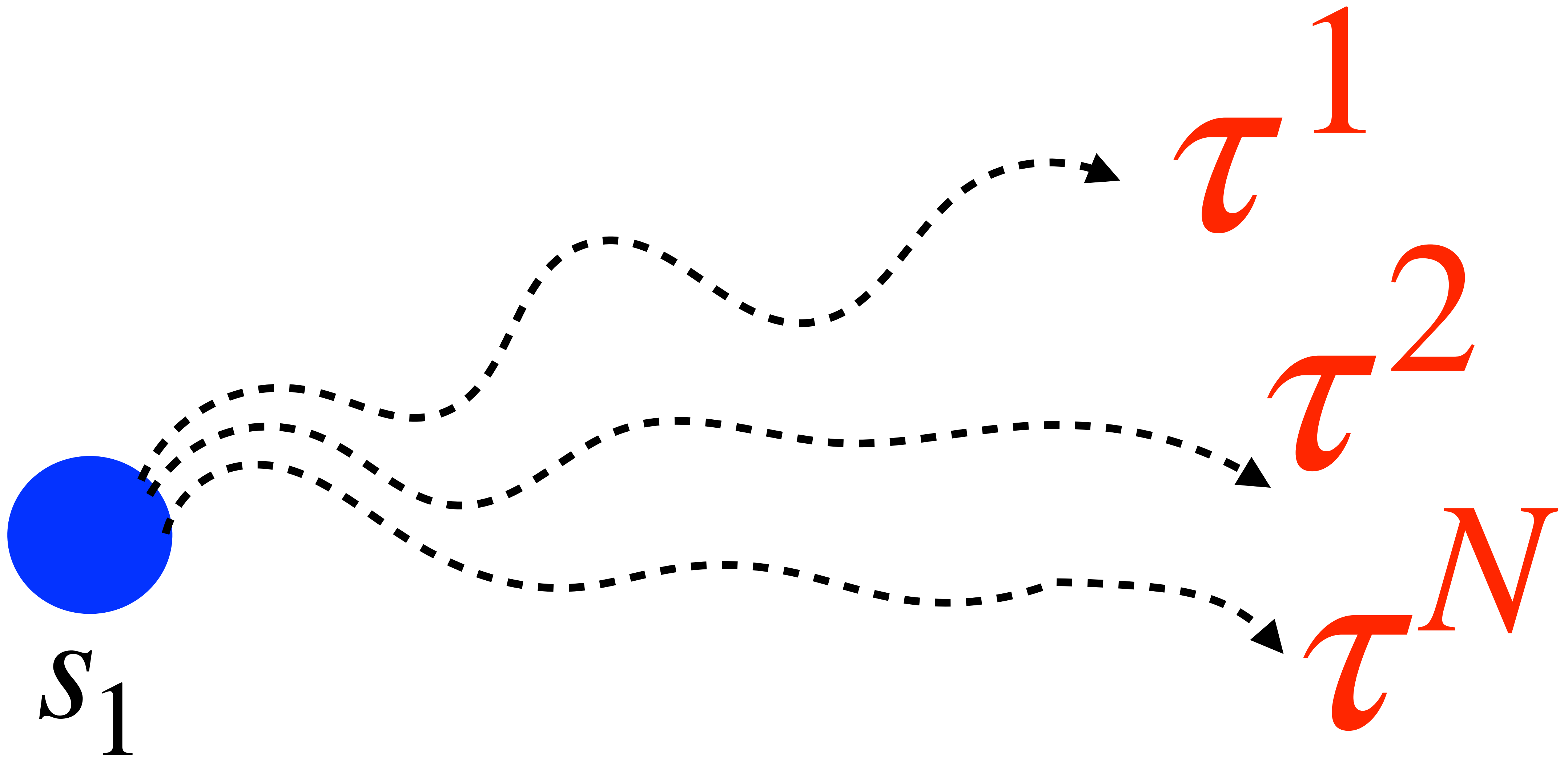
“reset” to good initial state

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

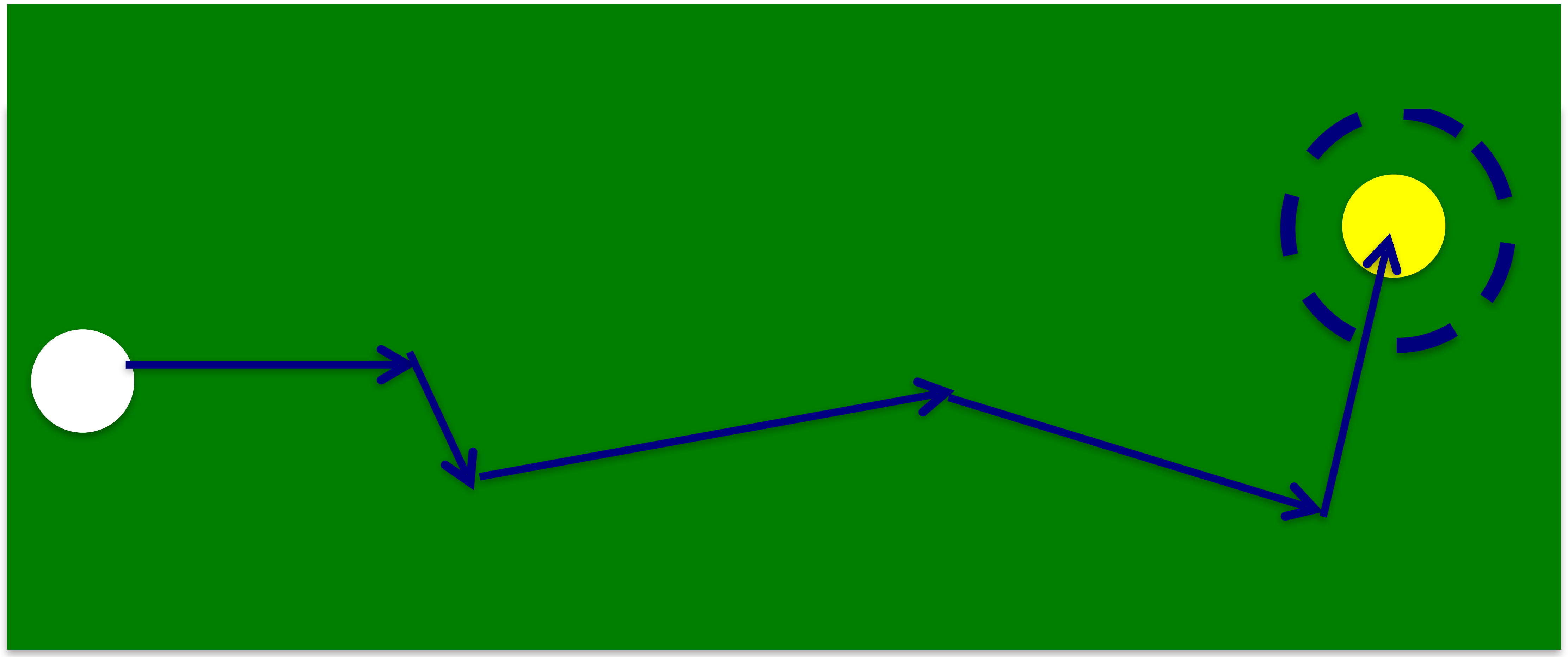
THE CREDIT ASSIGNMENT CHALLENGE

Issue of Credit Assignment

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

Issue of Credit Assignment

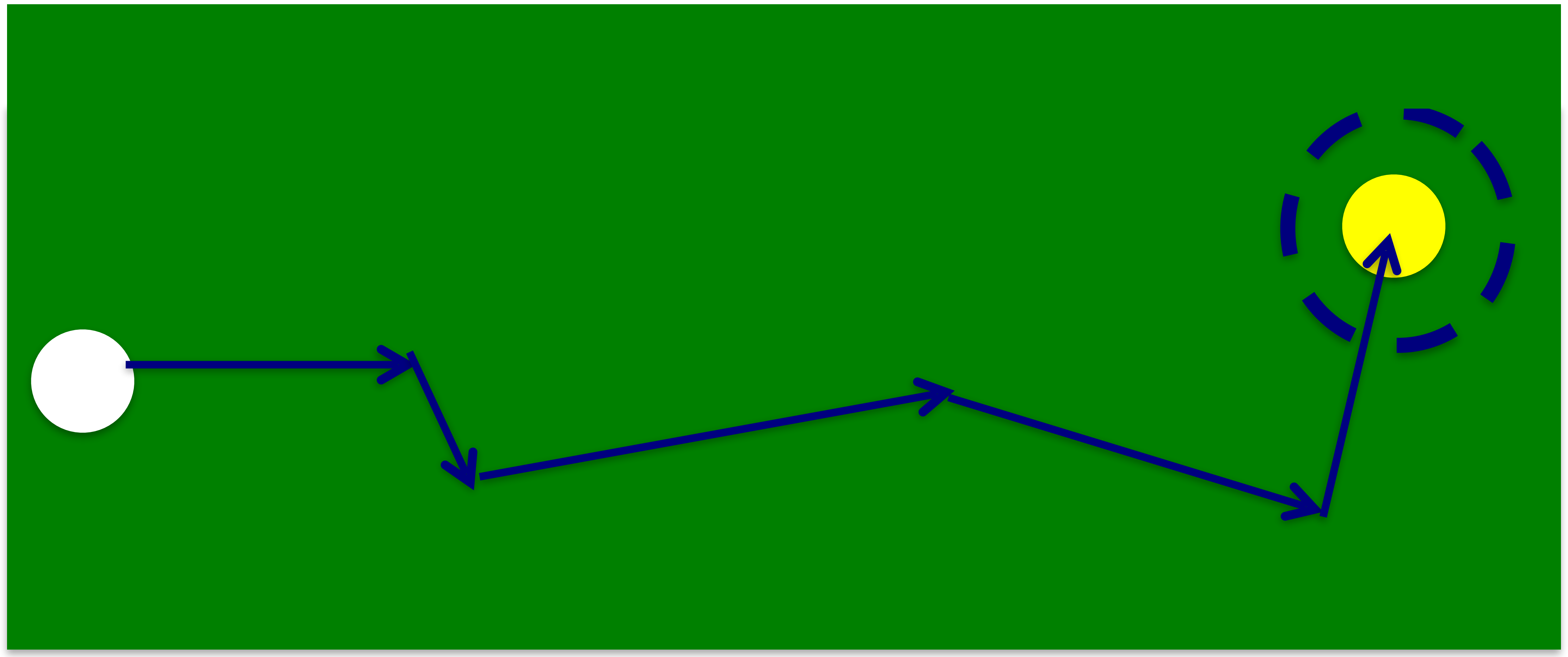
$$E\tau[\nabla_\theta(\log p_\theta(\tau))R(\tau)]$$



Issue of Credit Assignment

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

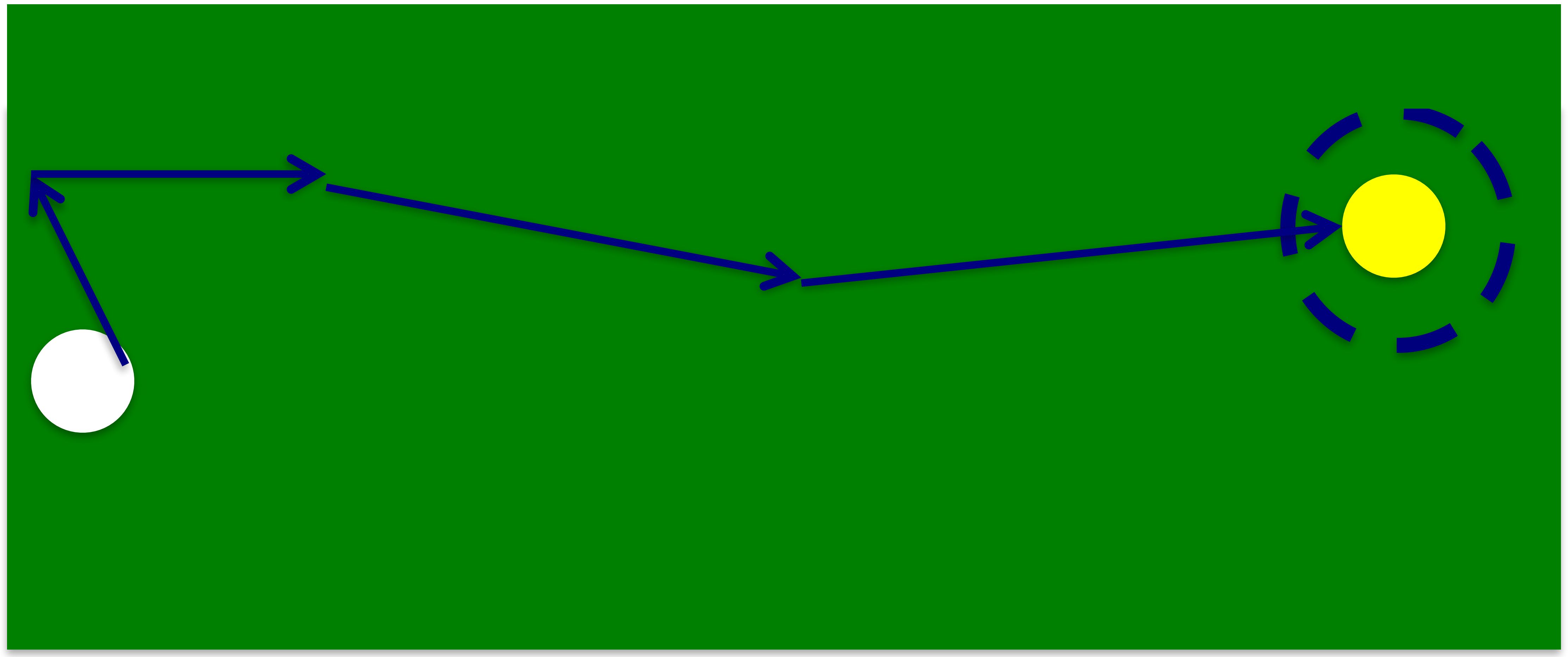
log-prob of each action is increased



Issue of Credit Assignment

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

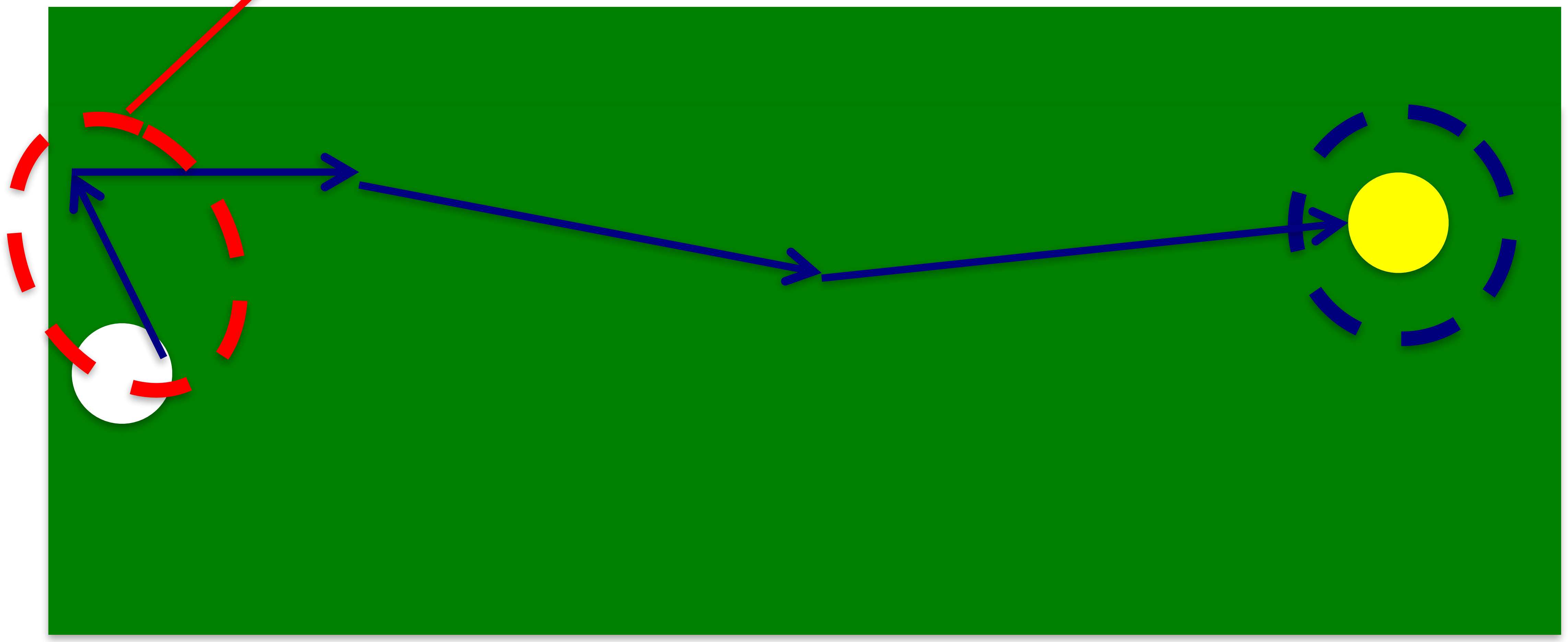
What about in this case?



Issue of Credit Assignment

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

logprob of this action also increases

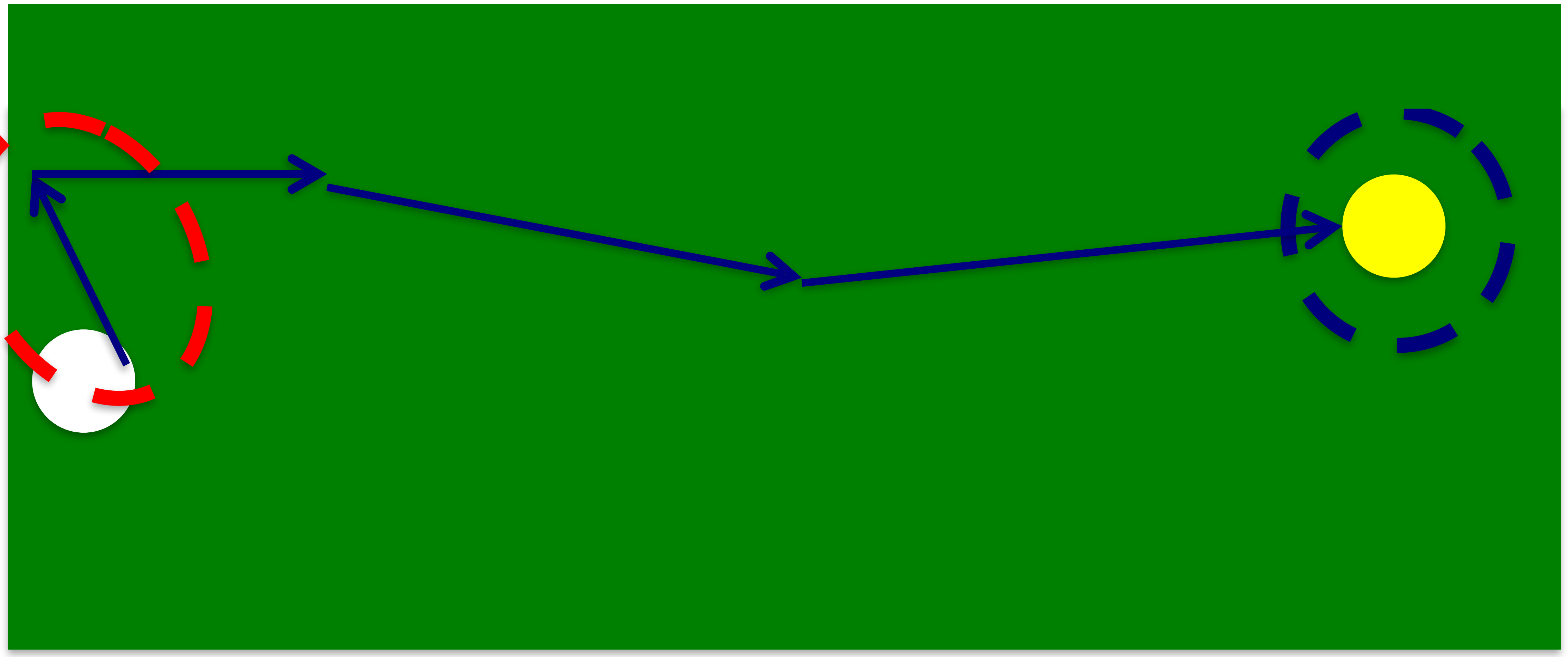


Issue of Credit Assignment

Does this also happen In supervised learning?

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

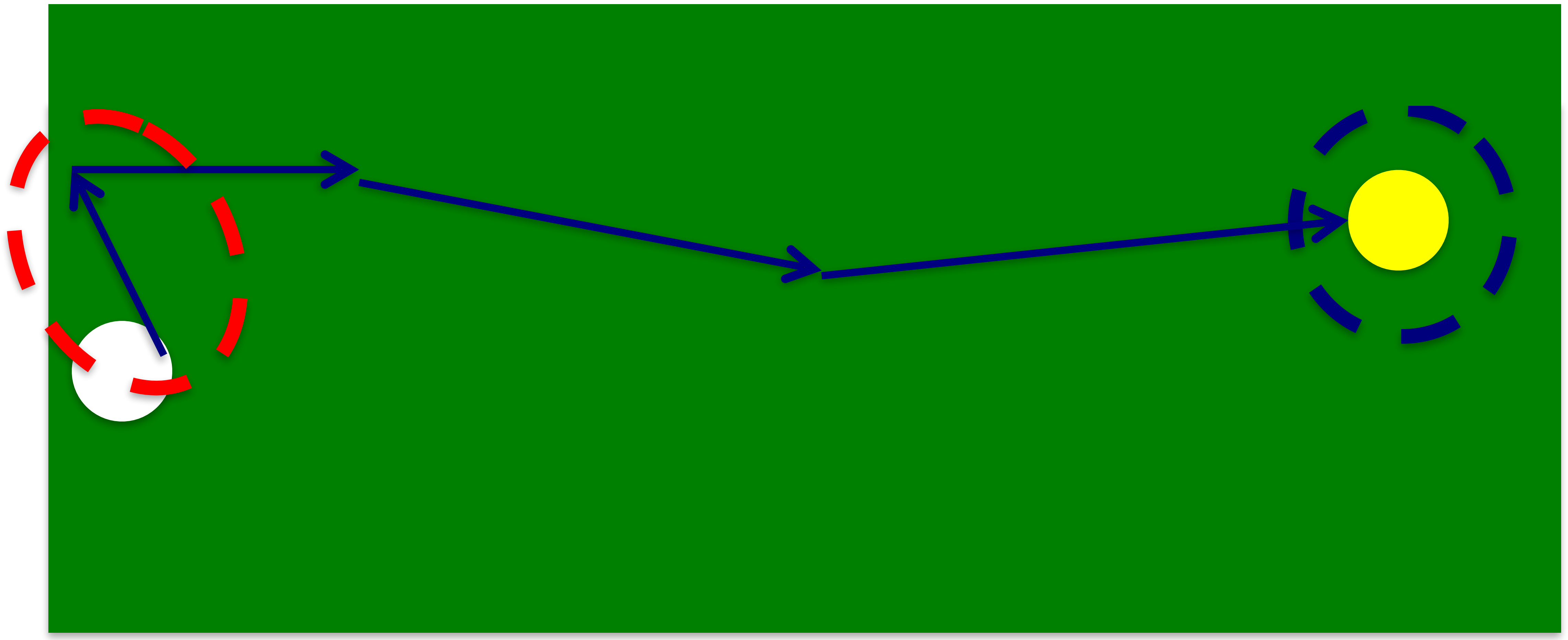
Delayed reward → Ambiguity in which action should be credited



Issue of Credit Assignment

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

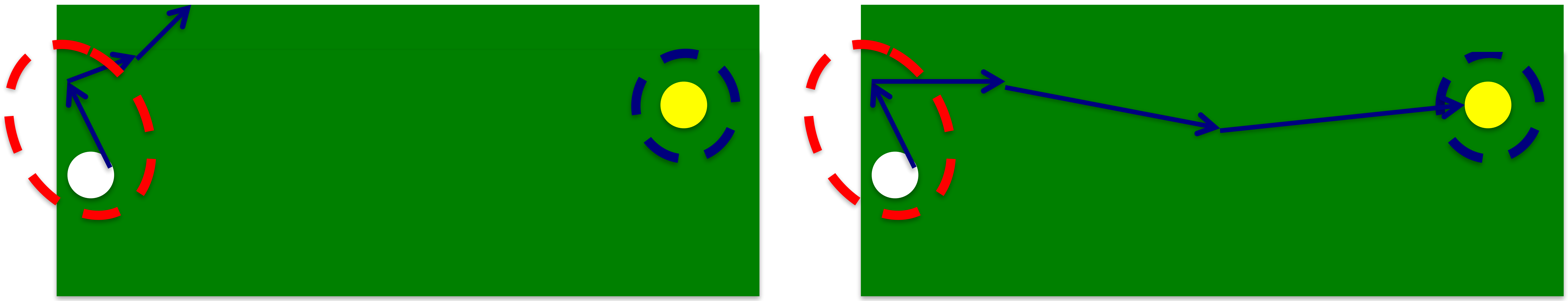
High Variance in gradient estimates



Variance in Policy Gradients

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

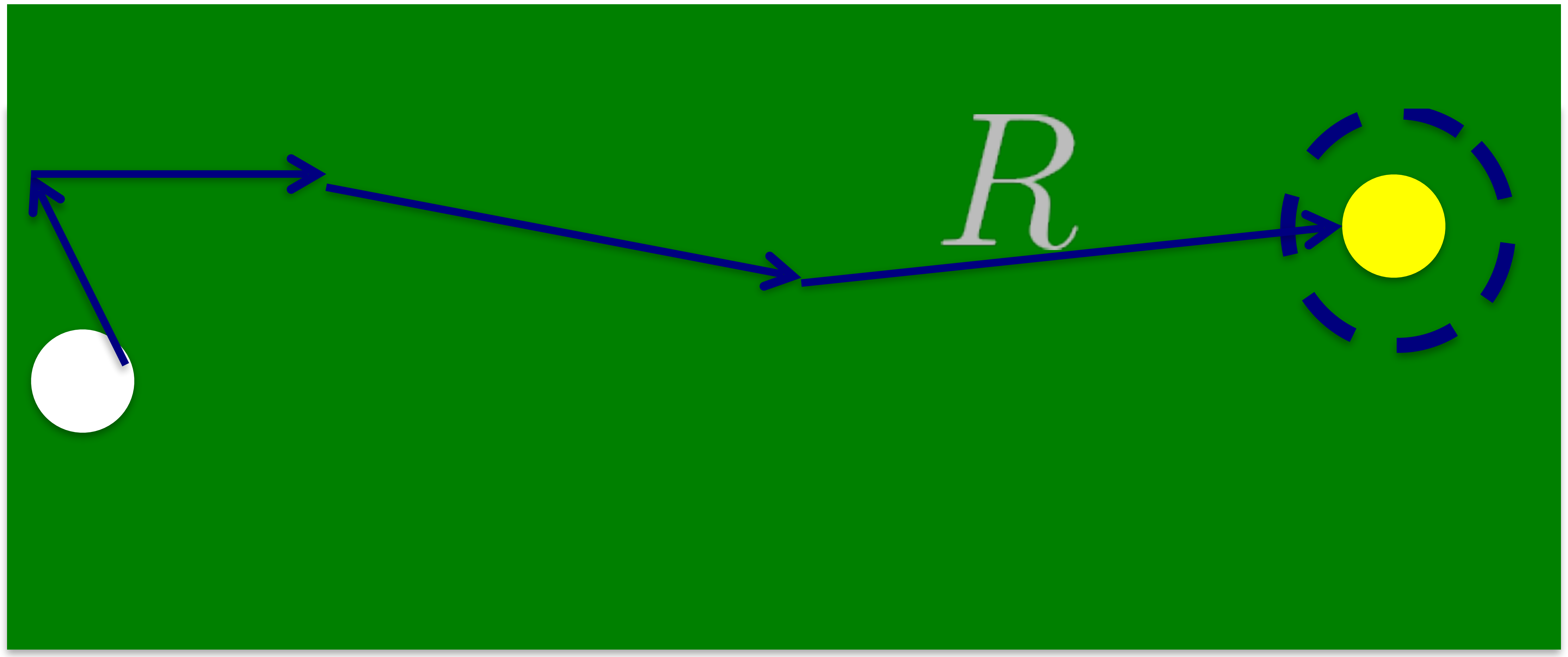
Same action — different trajectory rewards



conflicting gradients: variance

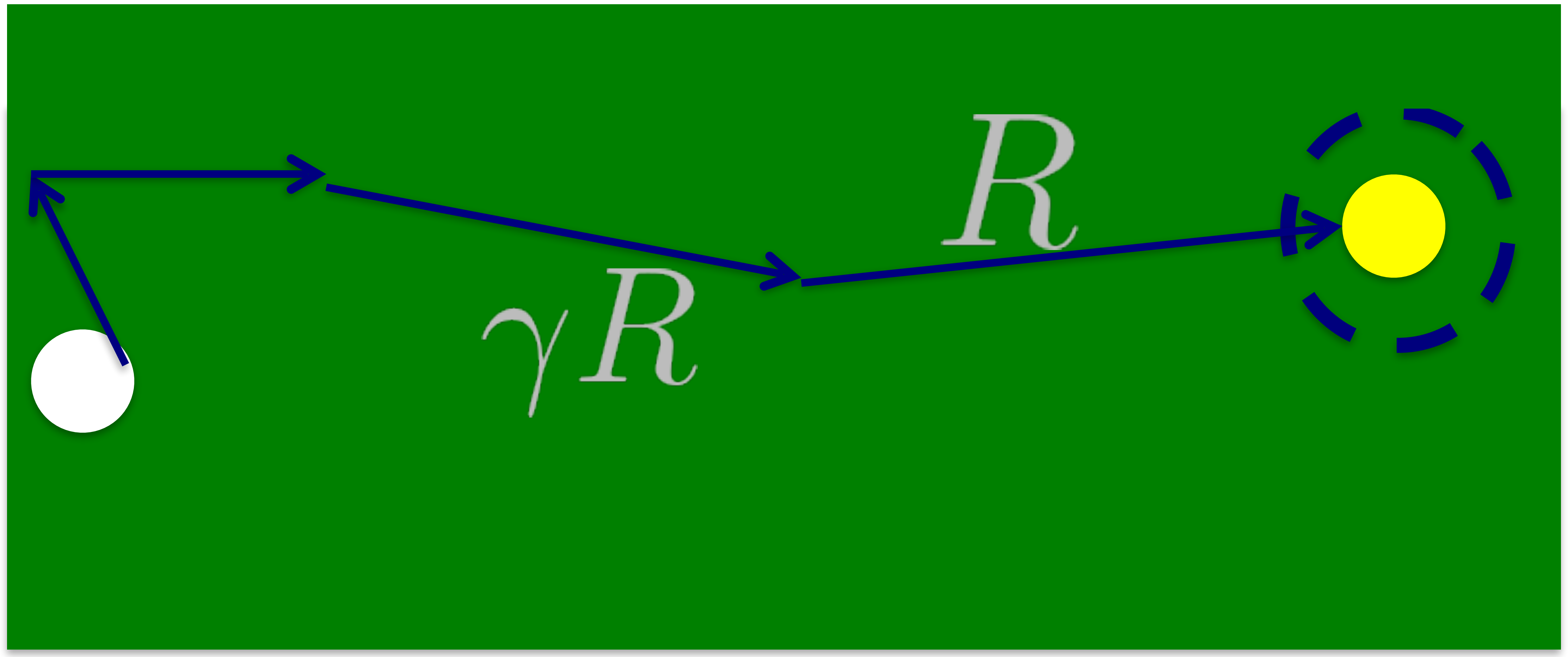
$$\mathrm{Var}\left[\nabla_\theta \log p_\theta(\tau) R(\tau)\right]$$

Variance Reduction Idea -- Discounts



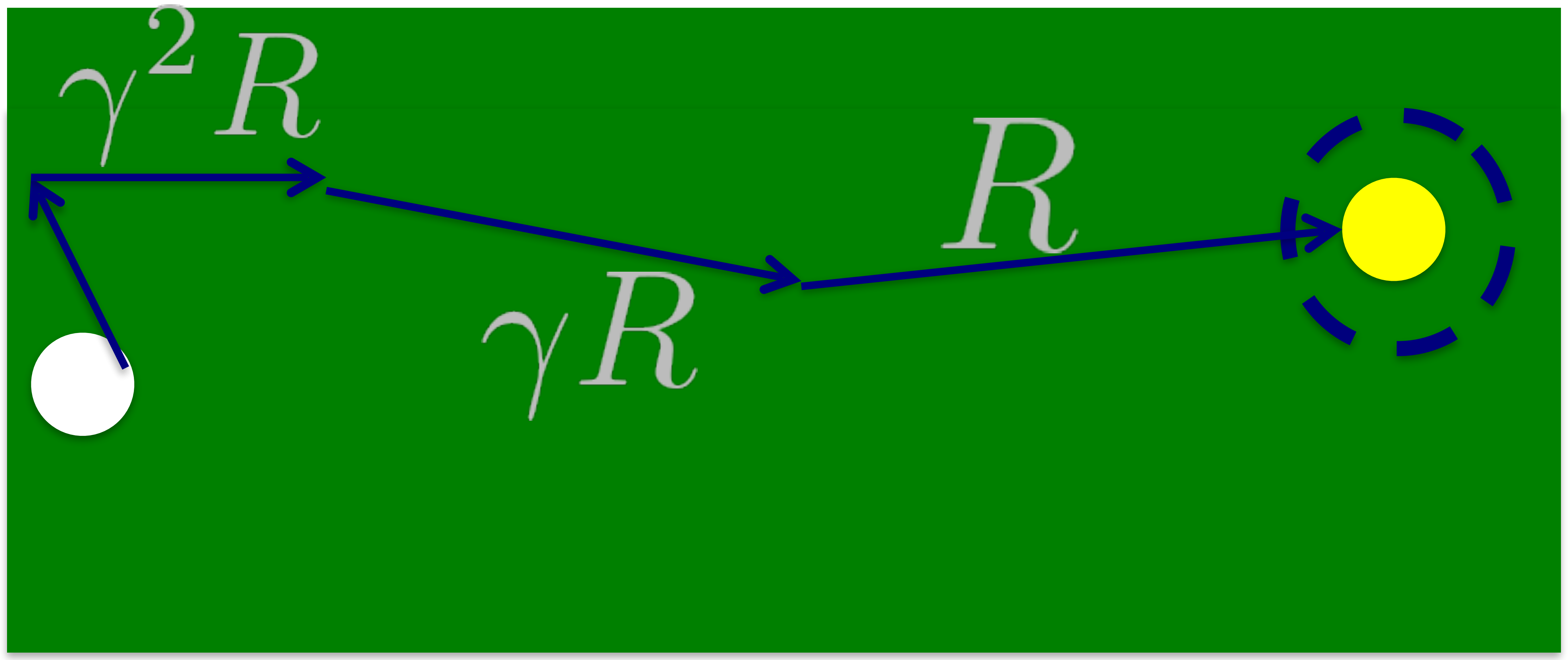
Variance Reduction Idea -- Discounts

$$\gamma < 1$$



Variance Reduction Idea -- Discounts

$$\gamma < 1$$



Variance Reduction with Discount

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

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

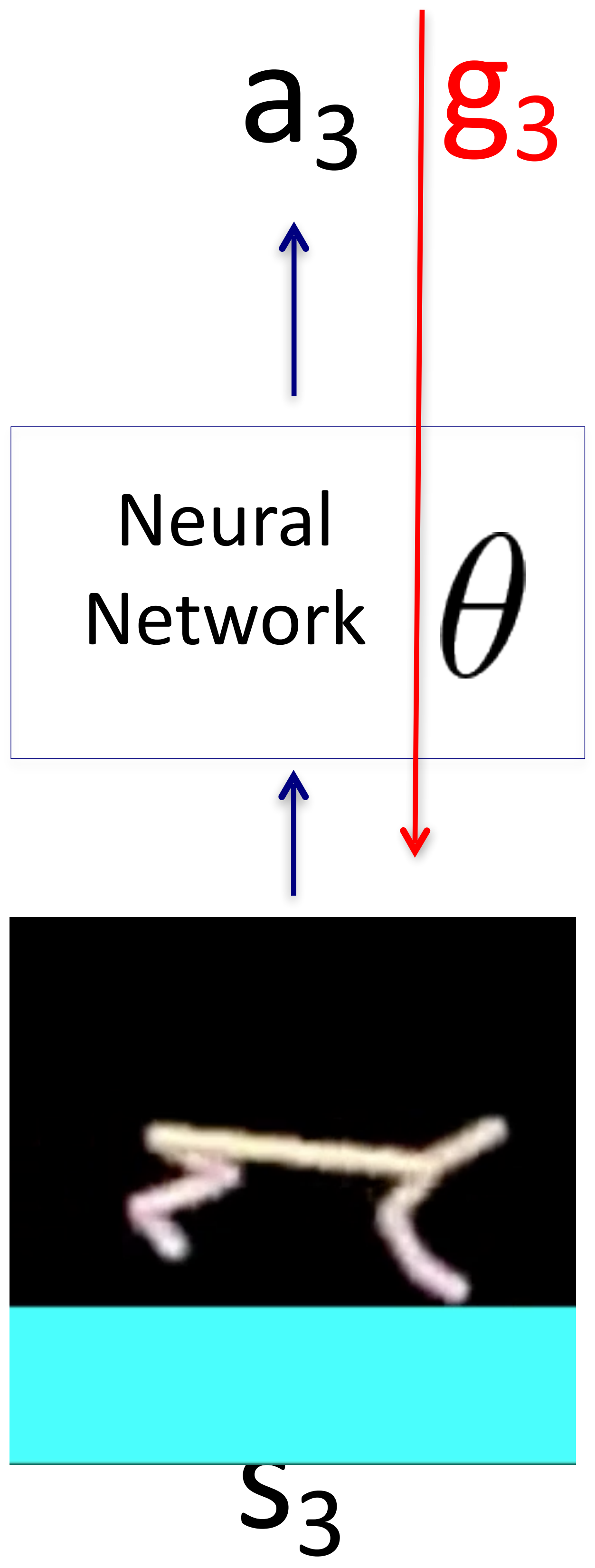
$$R^\gamma(\tau) = \sum_t \gamma^t r_t$$

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!

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

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

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

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

Can we reduce variance?

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

current actions don’t effect past rewards!

Reducing Variance

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

- Discounting
- Causality

Other methods?

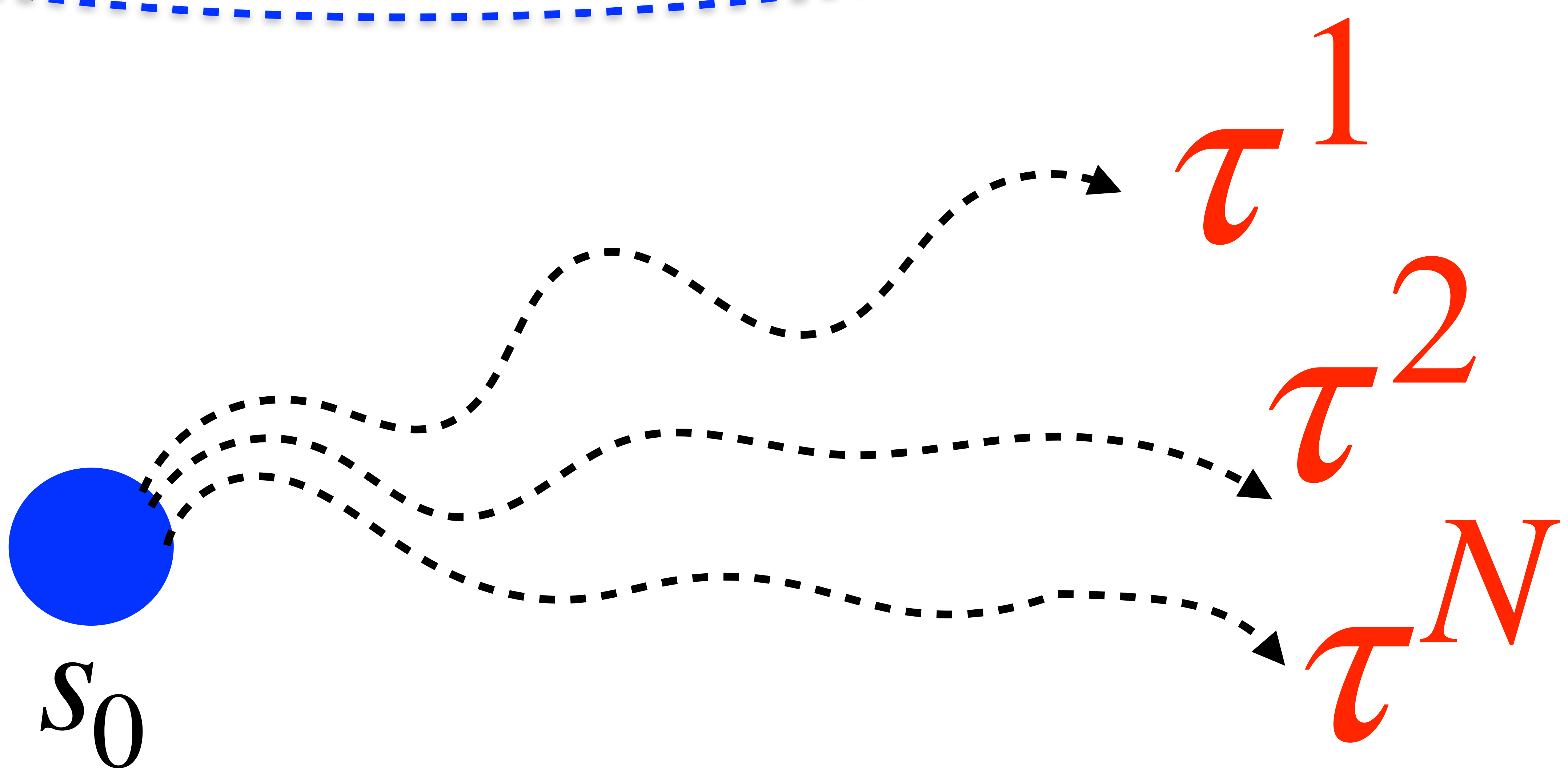
Estimating the rewards

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

Good estimate??

Increase N!

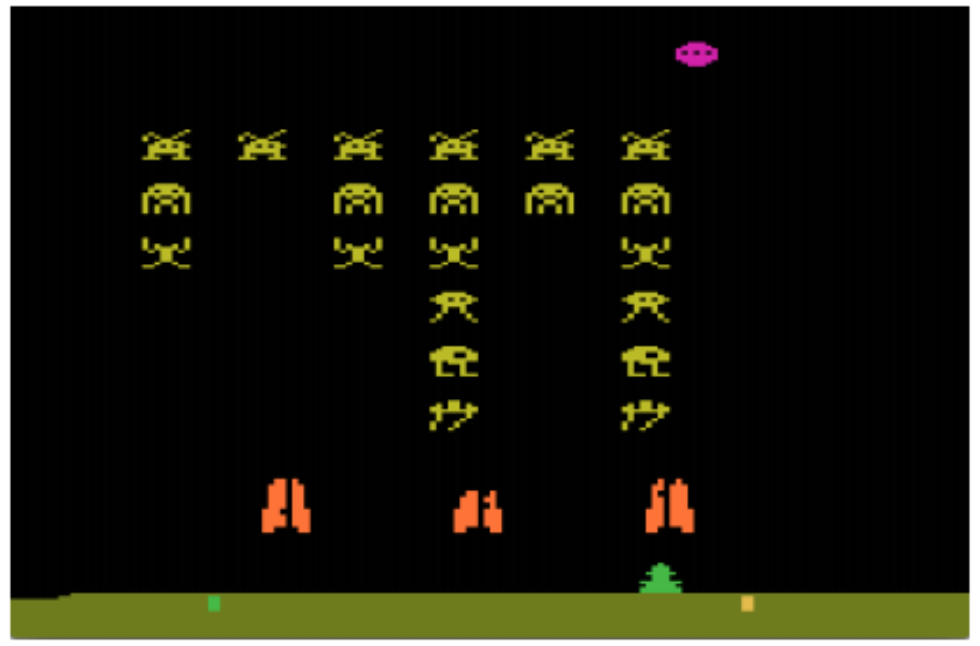
(reduces variance)



Estimating the rewards

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

Good estimate??



Massively parallelize data collection

???

Use an Existing dataset

(say using $\pi_\phi(a \mid s)$)

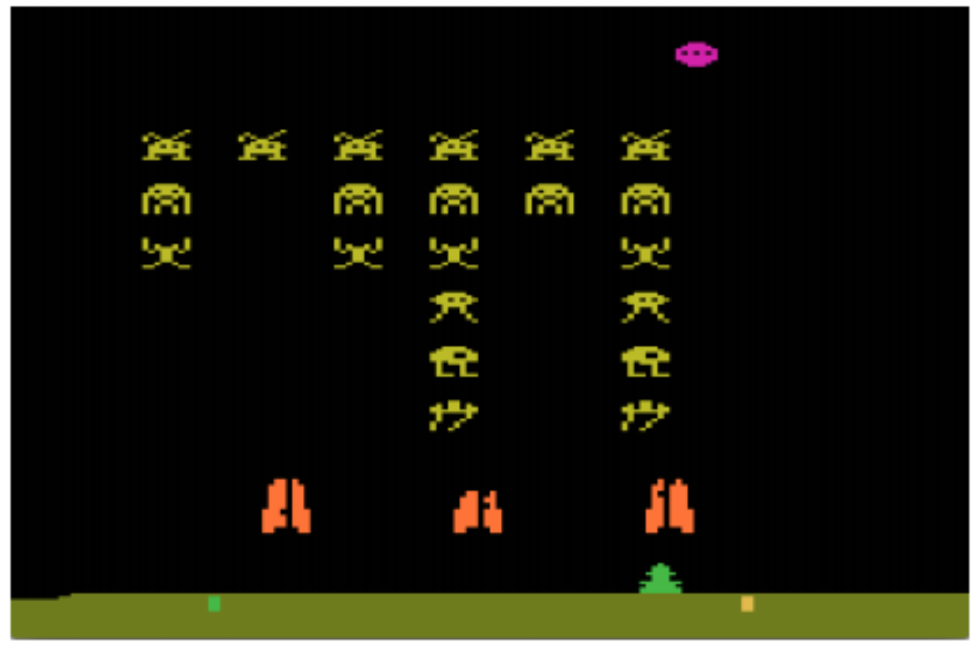
Estimating the rewards

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

Consider

$$\nabla_w f(w)$$

$$\nabla_{w=\theta} f(\theta) \qquad \cancel{\nabla_{w=\theta} f(\phi)}$$



Massively parallelize data collection

???

~~Use an Existing dataset~~

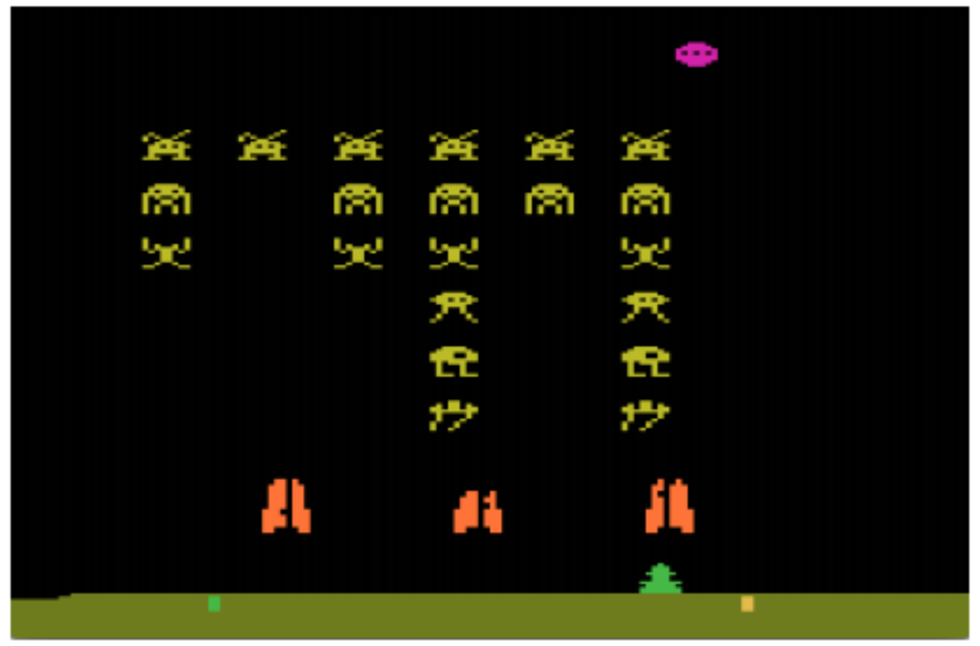
~~(say using $\pi_\phi(a \mid s)$)~~

Estimating the rewards

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

Need **data from current policy!!**

On-Policy Learning
(sample inefficient)



Massively parallelize data collection

???

~~**Use an Existing dataset**~~

~~(say using $\pi_\phi(a|s)$)~~

Estimating the rewards

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

Need **data from current policy!!**

On-Policy Learning
(sample inefficient)

Off-Policy Data

Importance Sampling

Off-Policy Learning

What is the implication on-policy sampling?

Reducing Variance

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

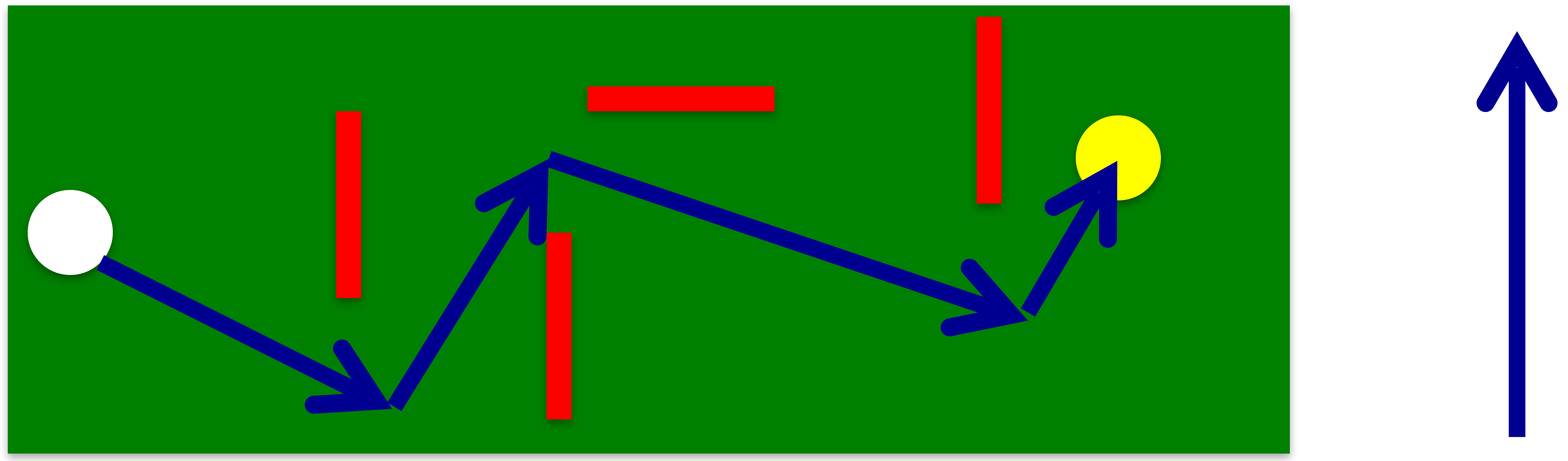
- Discounting
- Causality
- **Collect more data**

Other methods?

Recall

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

Intuitive Interpretation

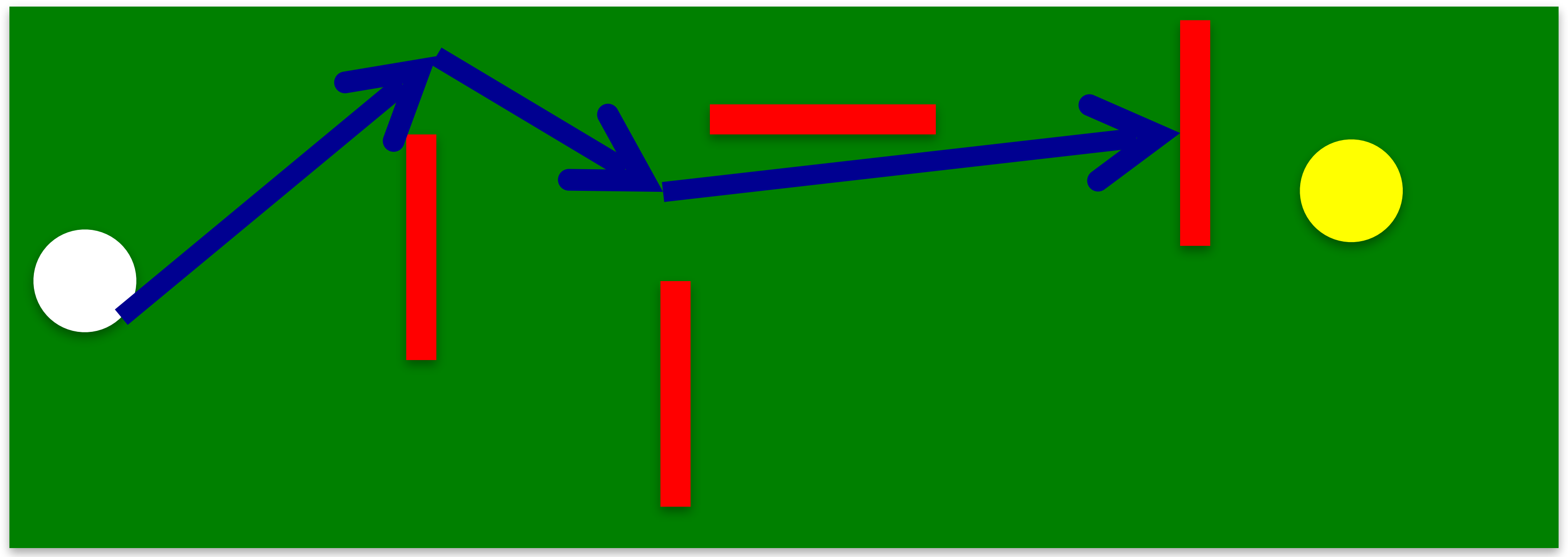


Increase
log-prob

Recall

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

Intuitive Interpretation

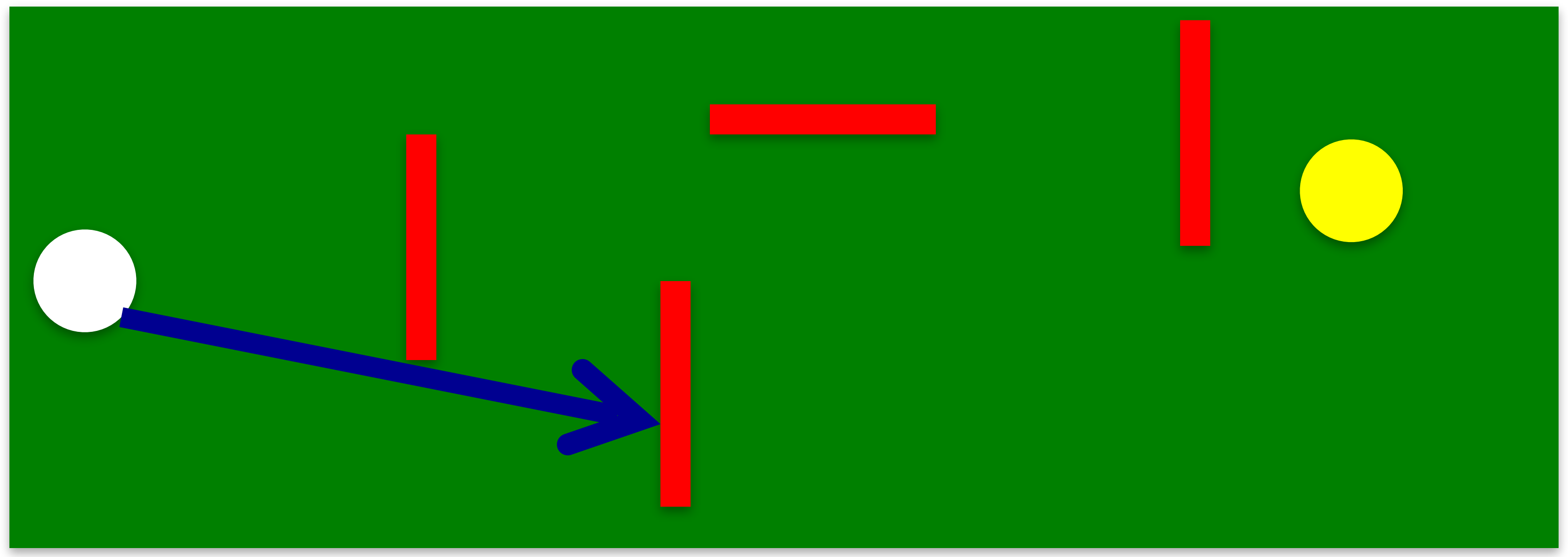


$\uparrow$ Increase log-prob by small amount

Recall

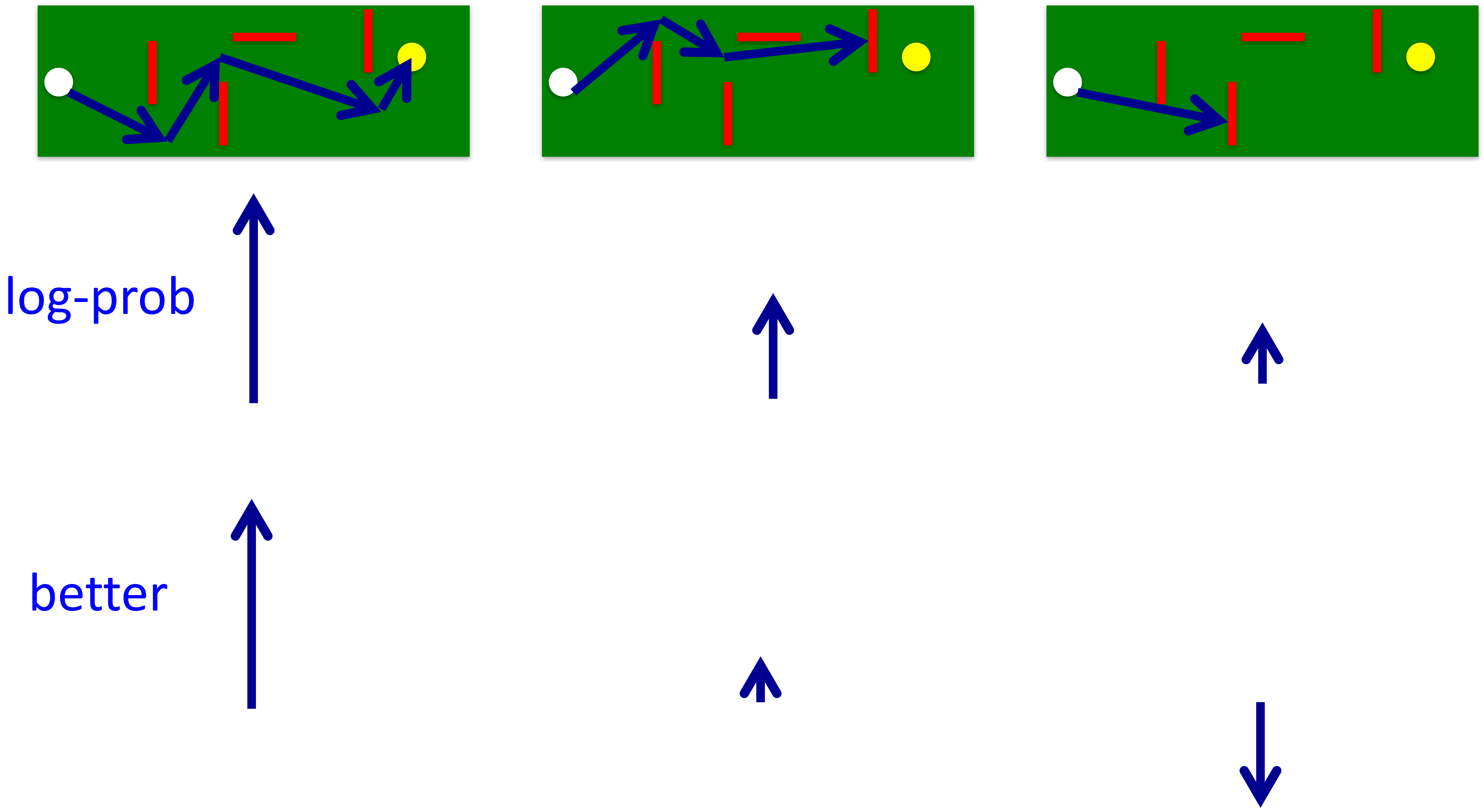
$$E_\tau[\nabla_\theta\left(\log p_\theta(\tau)\right)R(\tau)]$$

Intuitive Interpretation



Increase log-prob by smaller amount

Policy Gradients



reduce variance

only increase log-prob if better than average return

Baselines: Reducing Variance

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

$$\Downarrow$$

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

can we do this?

Yes, if **b** does not depend on θ

Prove in HW!

Why do baselines work?

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

Known: $\mathrm{Var}[x-y] = \mathrm{Var}[x] - 2\mathrm{Cov}[x,y] + \mathrm{Var}[y]$

$$\mathrm{Var}[x-y] \le \mathrm{Var}[x]$$

if

$$2\mathrm{Cov}[x,y] \ge \mathrm{Var}[y]$$

(i.e., if x, y are correlated)

$R(\tau)$, $b = E[R(\tau)]$ are correlated!

Lets try to find an optimal baseline

$$Var[x] = E[x^2] - E[x]^2$$

$$\min_b Var_\tau\left[\underbrace{\nabla_\theta\left(\log p_\theta(\tau)\right)}_{g(\tau)}\left(R(\tau) - b\right)\right]$$

$$b^* = \frac{E\left[g(\tau)^2 R(\tau)\right]}{E\left[g(\tau)^2\right]}$$

weighted
trajectory reward

$$b = E\left[R(\tau)\right] = V(s)$$

Value Function!

not necessarily the best choice,
but works well in practice!

Variance Reduction Techniques for Gradient Estimates in Reinforcement Learning, Greensmith et al., 2004

Putting it all together

How to get this?

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

Monte-Carlo Estimate

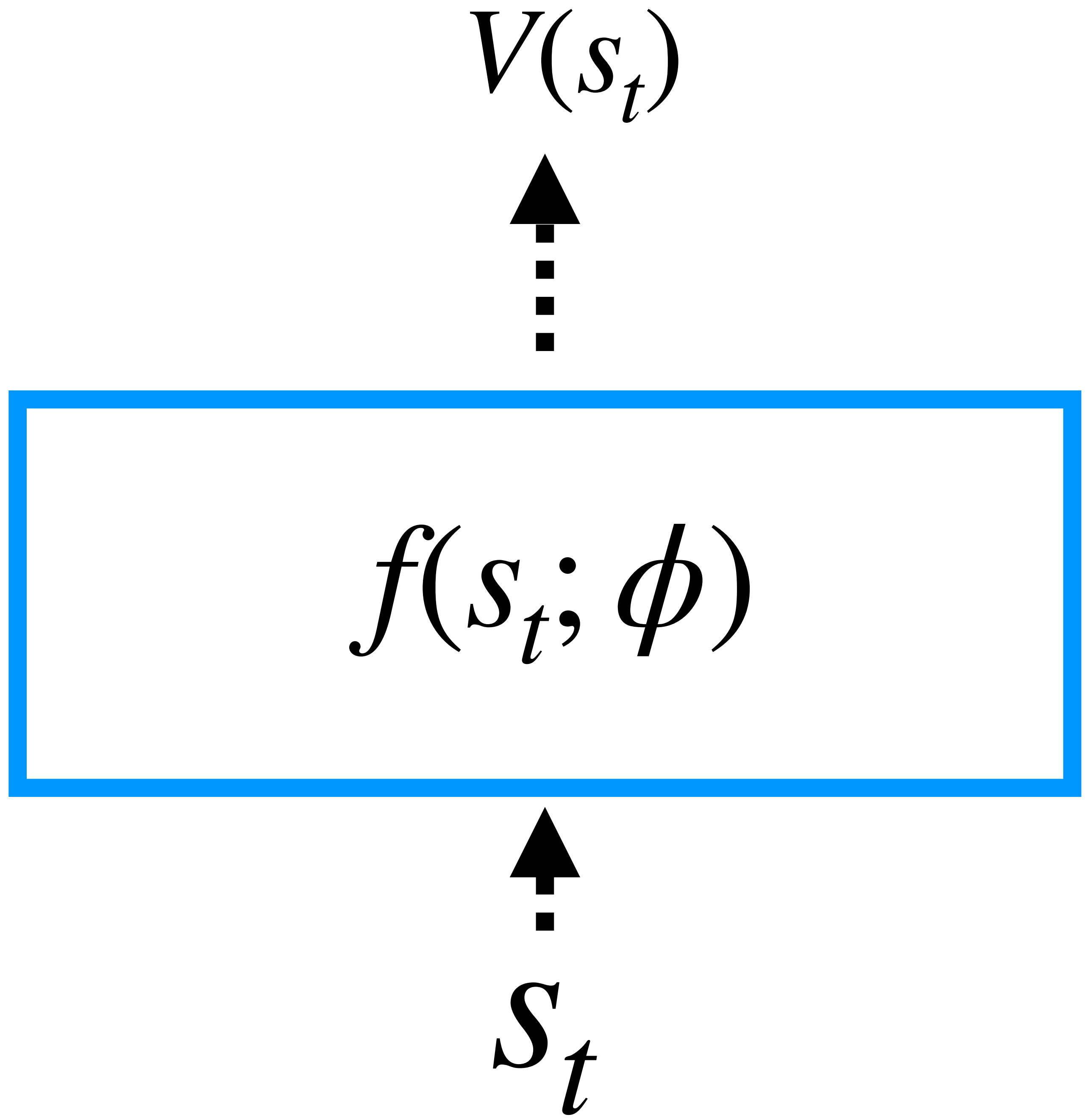
Function Approximation

Using value iteration / Temporal Difference (TD) Learning

Putting it all together

How to get this?

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



new estimate

$$\min_{\phi} \| V^{\phi}(s_t) - \left(r_t + \gamma V^{\phi'}(s_{t+1})\right) \|_2^2$$

Estimate using backup term

Temporal Differencing (TD) Error

Reducing Variance

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

- Discounting
- Causality
- Collect more data
- **Baselines**

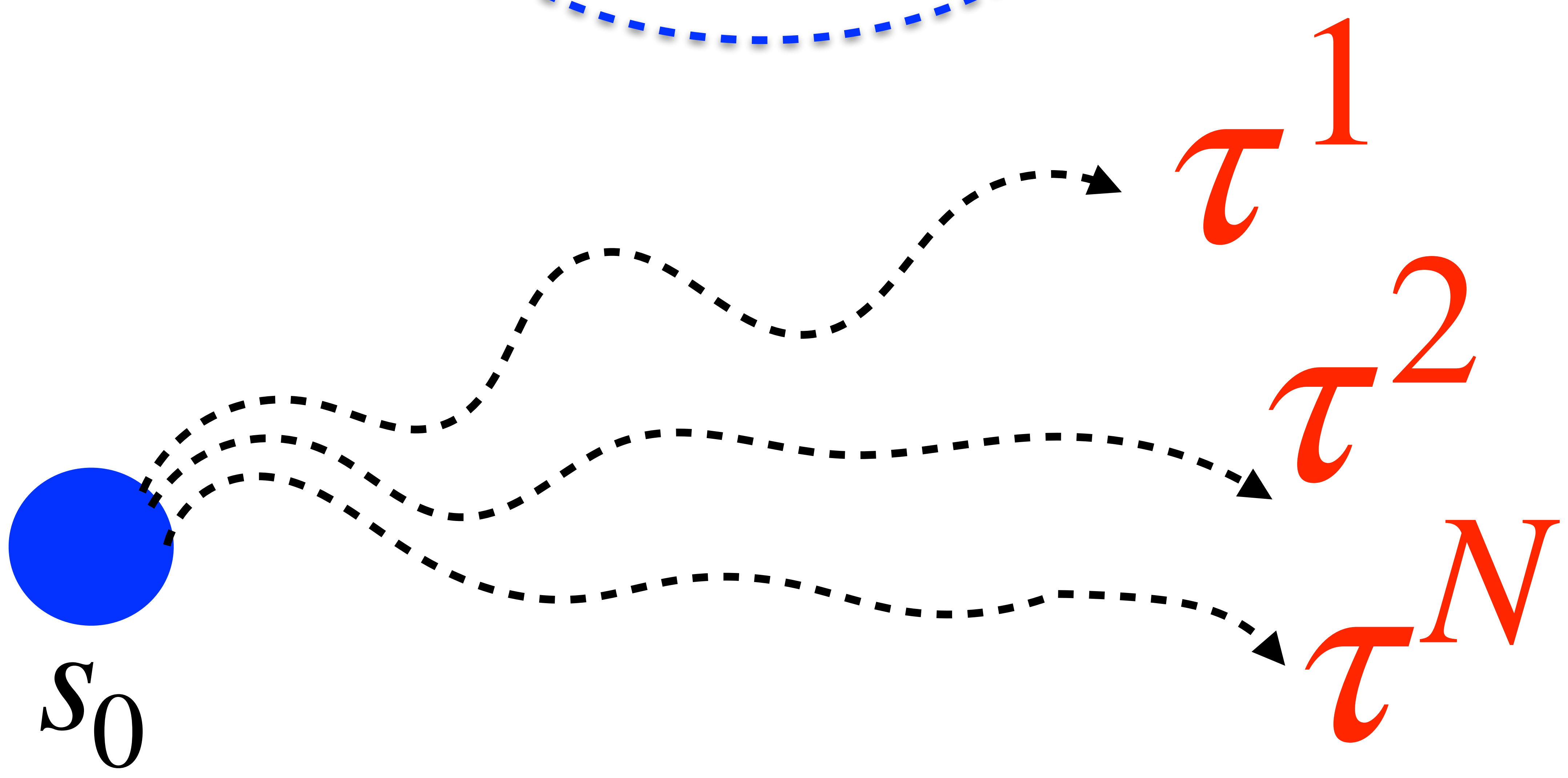
Other methods?

Bias-Variance Tradeoff

ACTOR CRITIC METHODS

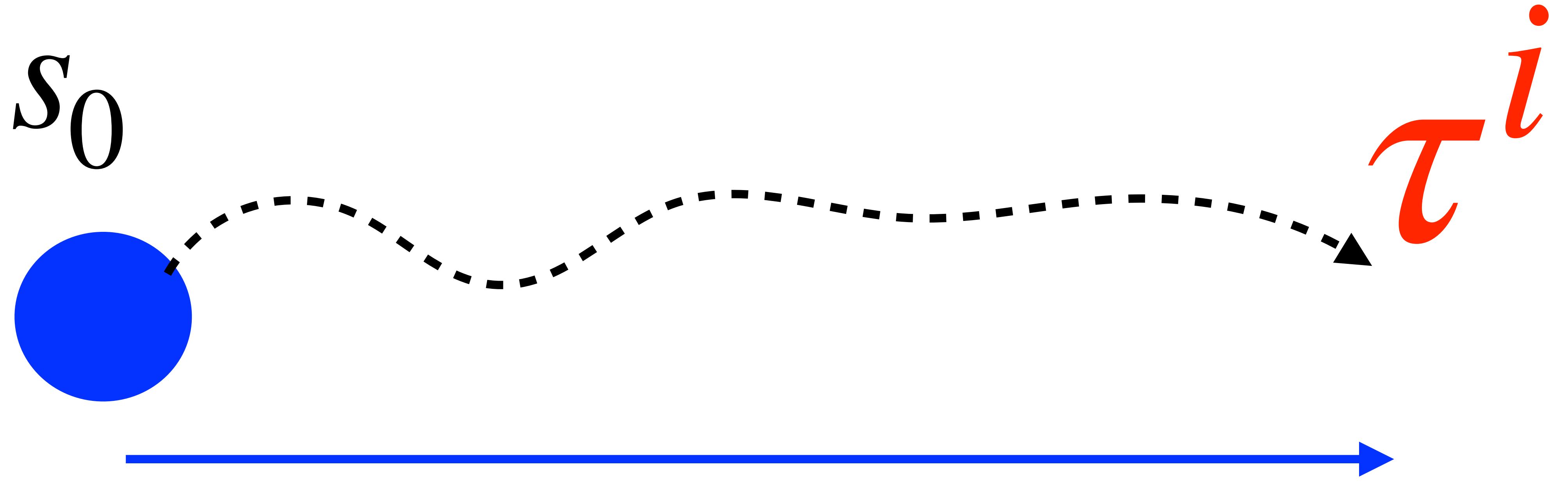
Now consider,

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



Now consider,

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

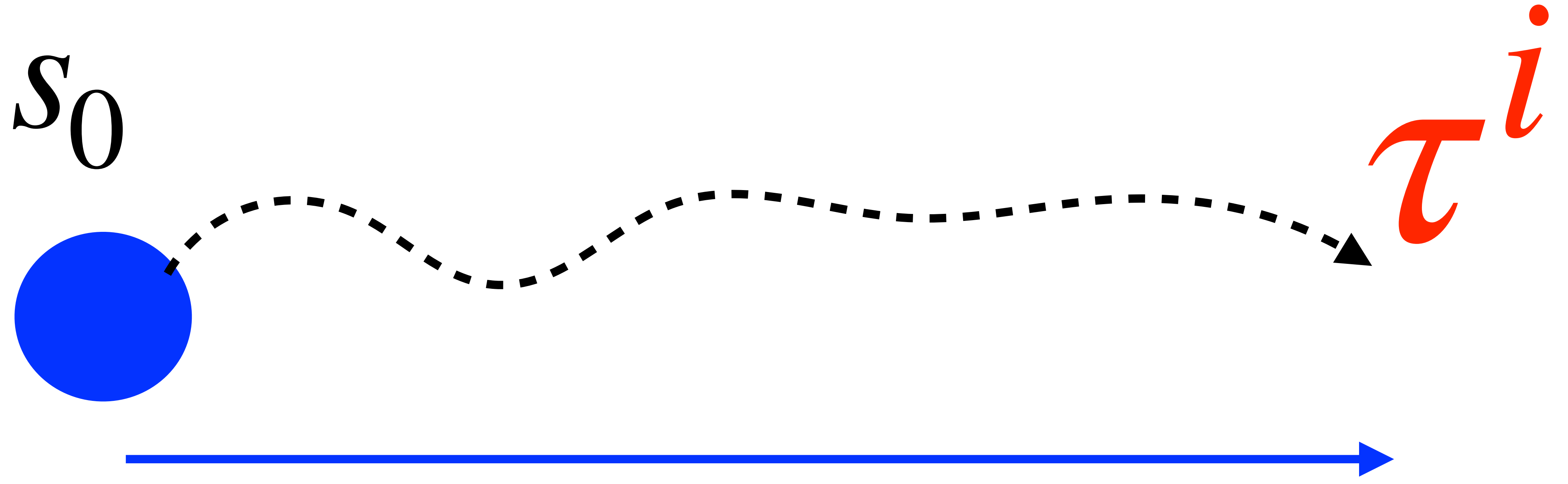


$$\sum_{t'=t}^{T}\gamma^{t'-t} r(s_t^i, a_t^i)$$

Monte-Carlo Estimation

Now consider,

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

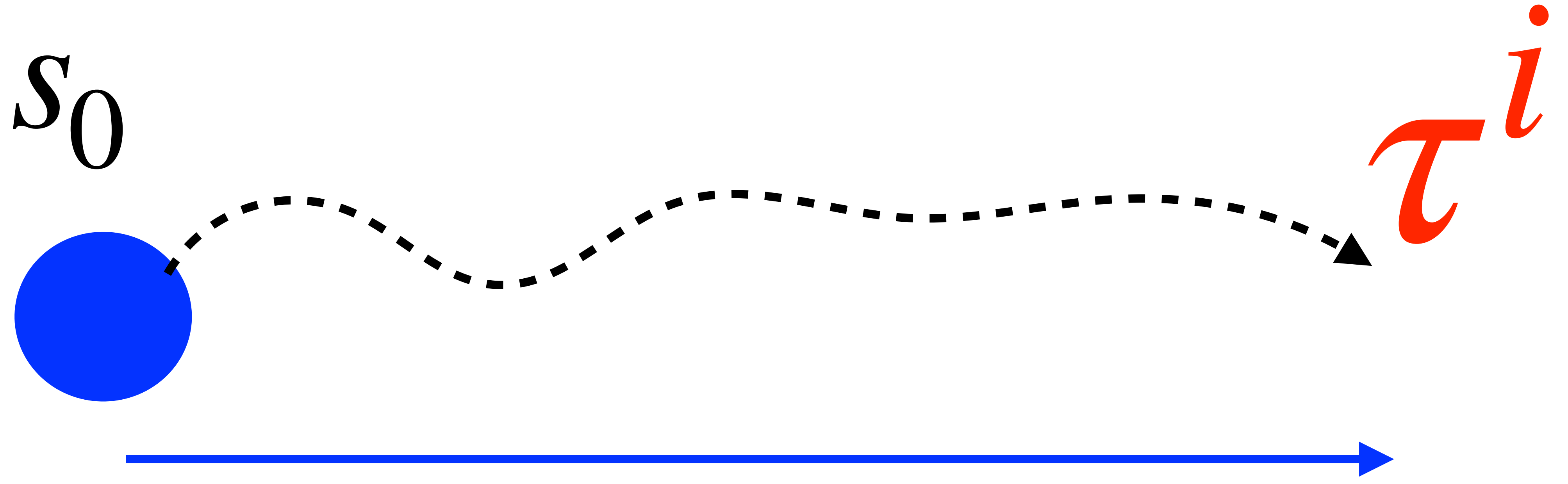


$$\sum_{t'=t}^{T}\gamma^{t'-t}r(s_t^i, a_t^i)$$

$$= r_t + \gamma\sum_{t'=t+1}^{T}\gamma^{t'-(t+1)}r_{t'}^i$$

Now consider,

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

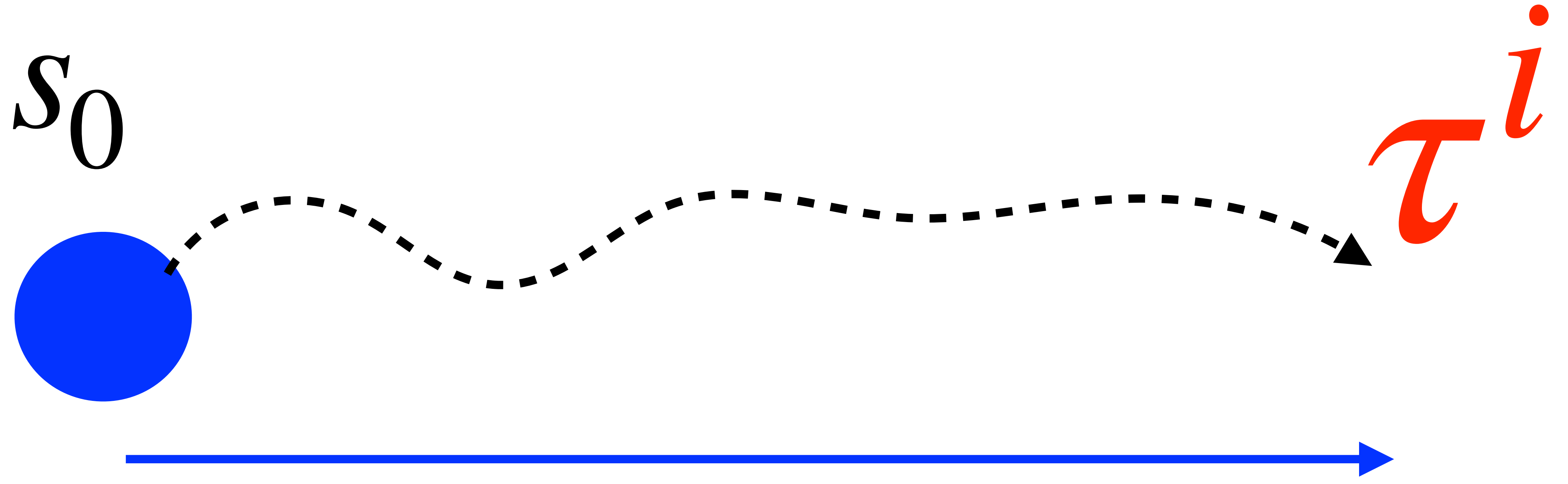


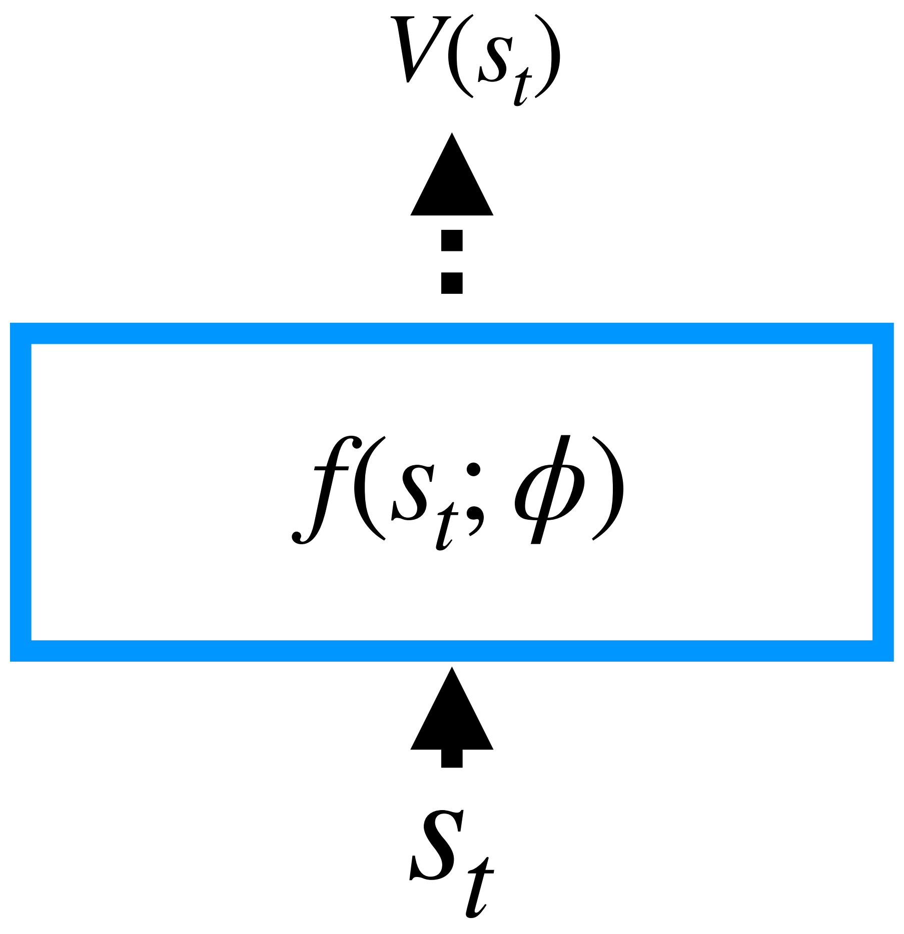
$$V(s_{t'+1})$$

$$= r_t + \gamma \sum_{t'=t+1}^{T}\gamma^{t'-(t+1)} r_{t'}^i$$

Now consider,

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



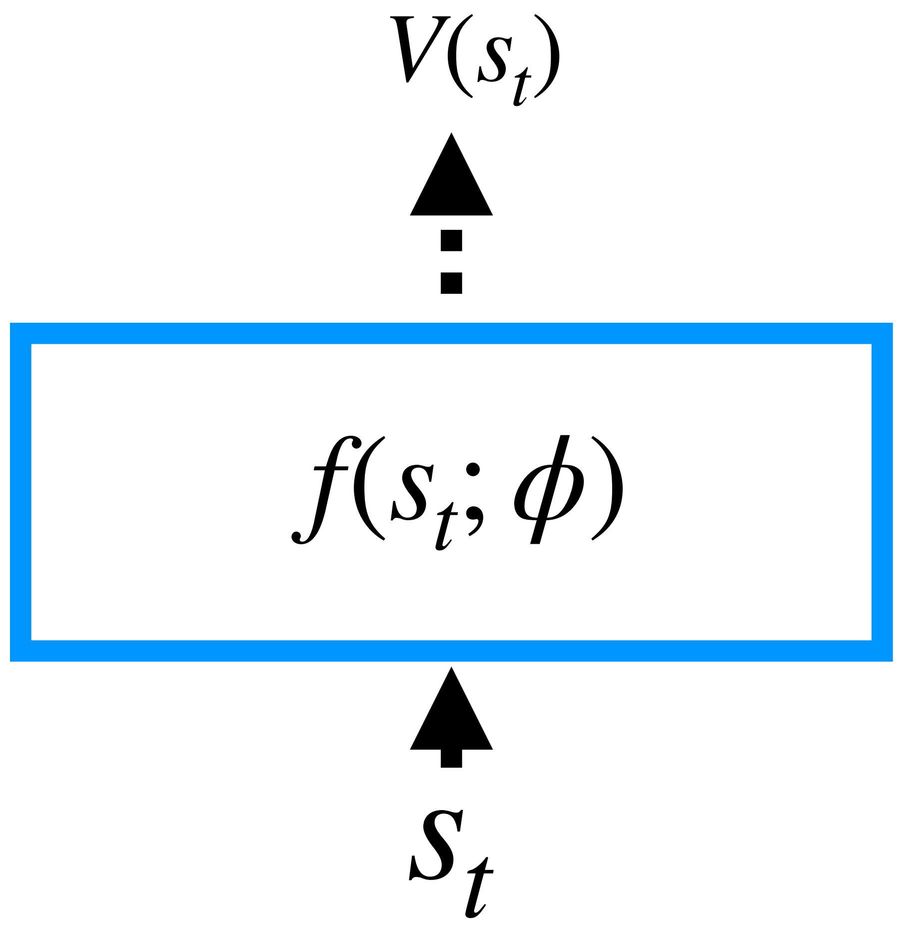


$$\sum_{t'=t}^{T}\gamma^{t'-t} r_{t'} \quad \textbf{v/s} \quad r_{t'} + \gamma V(s_{t'+1})$$

Now consider,

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

Variance **v/s** Bias



$$\sum_{t'=t}^{T}\gamma^{t'-t} r_{t'} \quad \textbf{v/s} \quad r_{t'} + \gamma V(s_{t'+1})$$

Now consider,

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

Variance	v/s	Bias
$\sum_{t'=t}^{T}\gamma^{t'-t} r_{t'}$	v/s	$r_{t'} + \gamma V(s_{t'+1})$

Actor-Critic Method

Critic

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

Actor-Critic Method

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

Q-Value Function

$$\left(Q(s_t^i, a_t^i) - V(s_t^i)\right)$$

Actor-Critic Method

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

$$\left(Q(s_t^i, a_t^i) - V(s_t^i)\right)$$

Advantage Function! $\left(A(s_t^i, a_t^i)\right)$

Advantage Actor-Critic (A2C) Method

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

$$\left(Q(s_t^i, a_t^i) - V(s_t^i)\right)$$

Advantage Function! $\left(A(s_t^i, a_t^i)\right)$

Advantage Actor-Critic (A2C) Method

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

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

Advantage Actor-Critic (A2C) Method

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

compare to

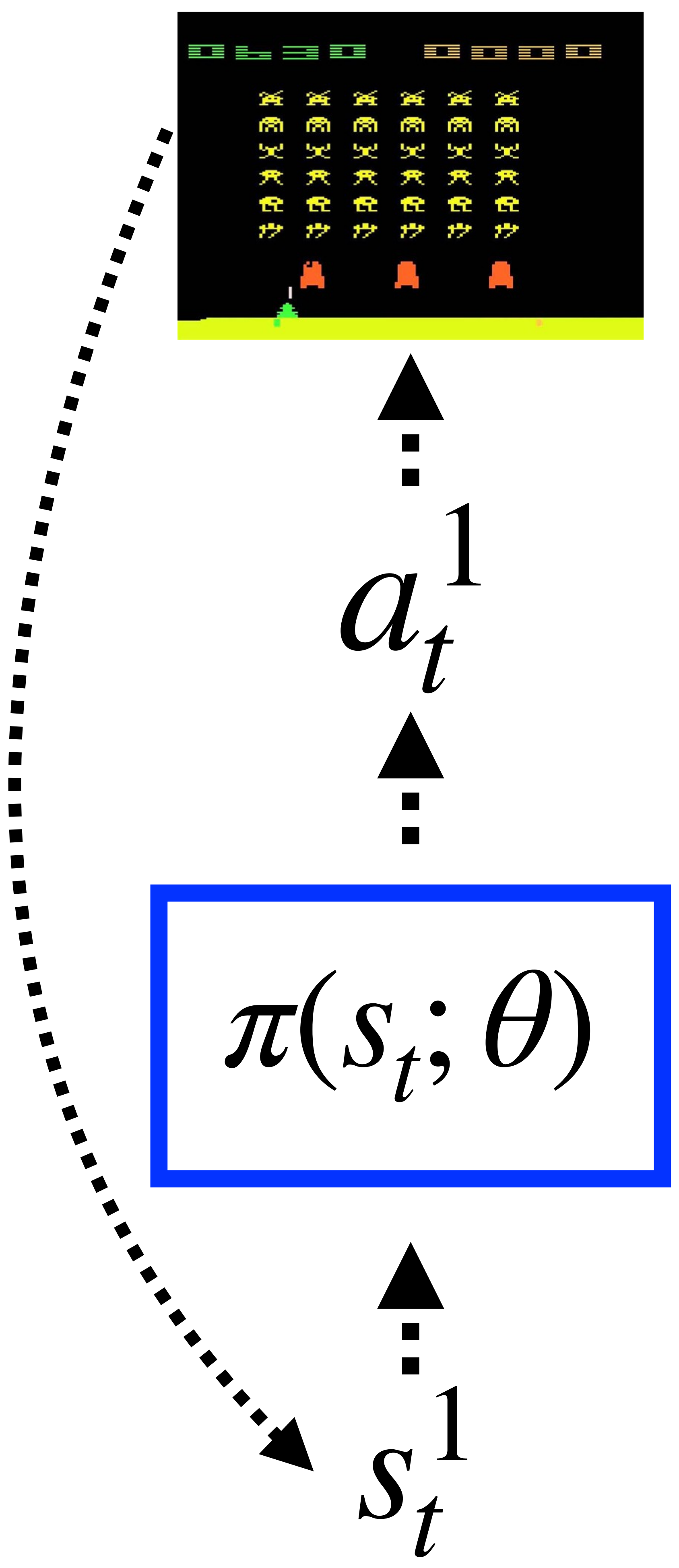
$$E_\tau\left[\nabla_\theta\left(\log p_\theta(\tau)\right)\boxed{R(\tau)}\right]$$

Reducing Variance

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

- Discounting
- Causality
- Collect more data
- Baselines
- **Use of Critic:** *Bias-Variance Tradeoff*

Actor

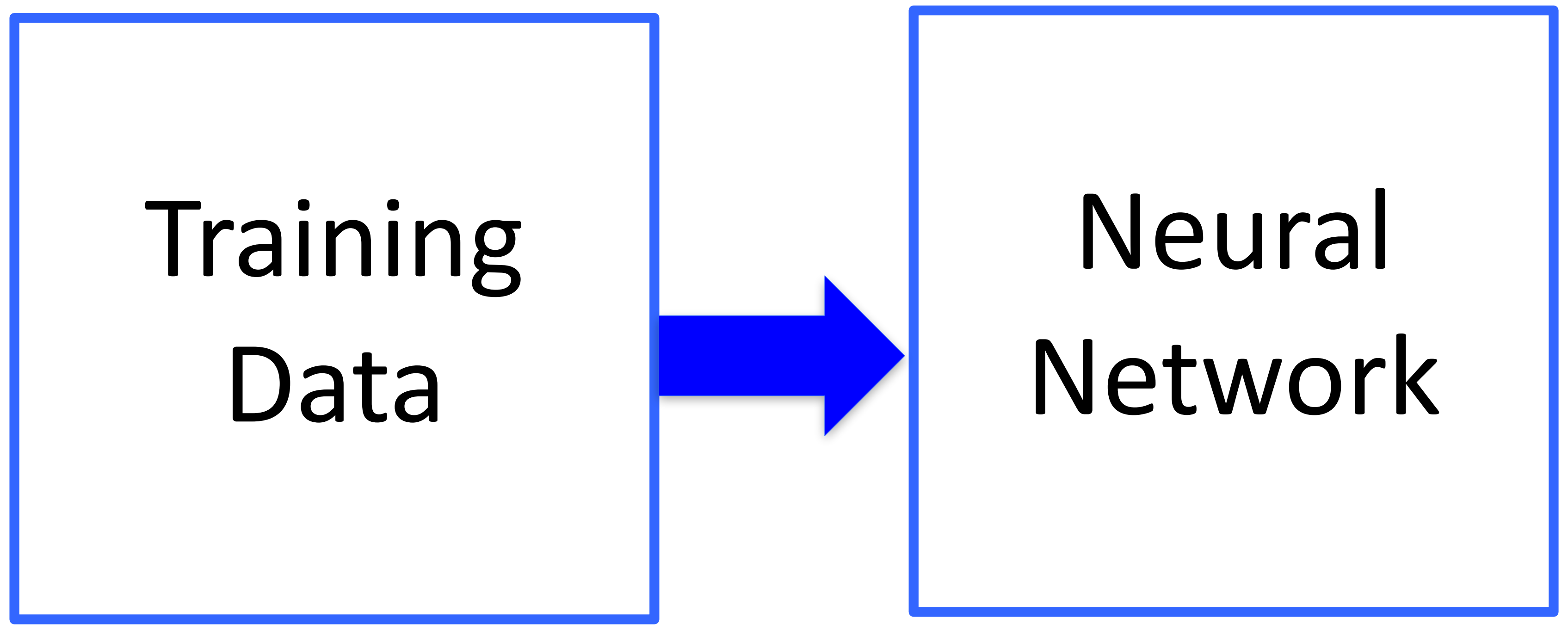


Advantage Actor Critic
(A2C)

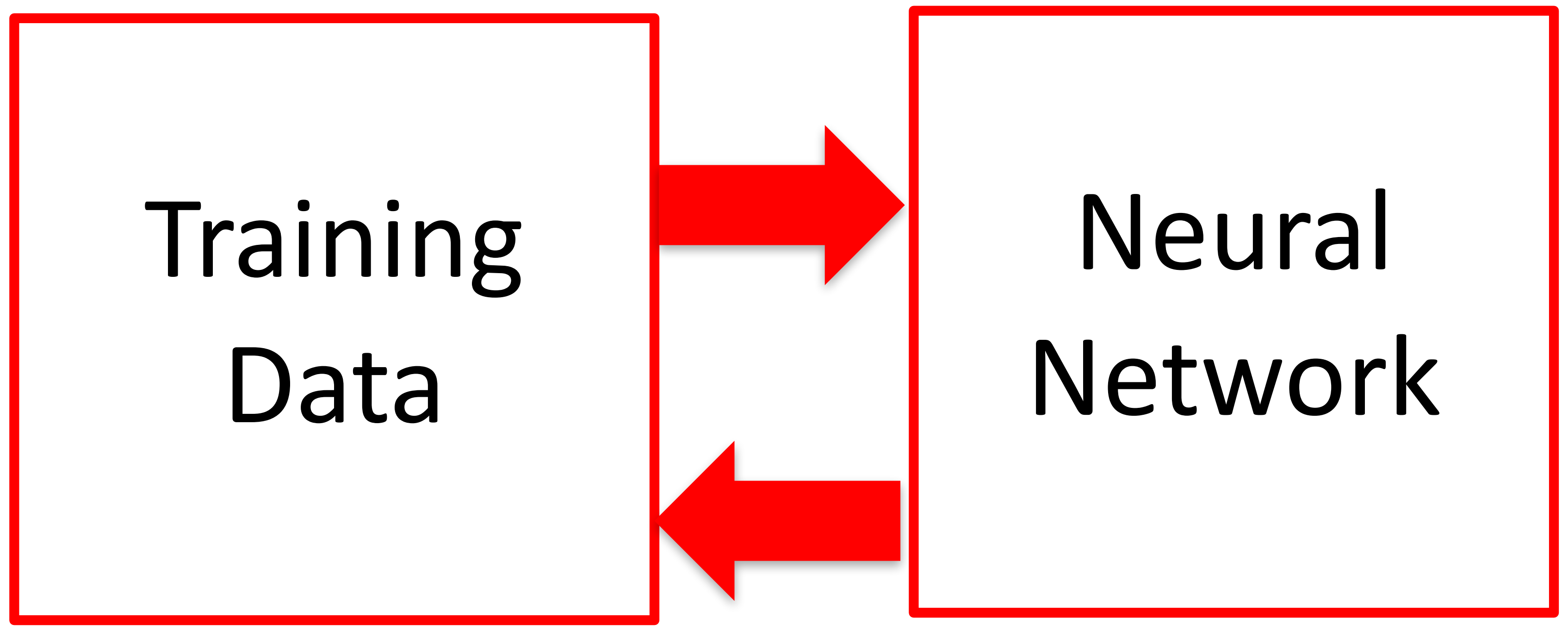
$$\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(a_t^i \mid s_t^i) A(s_t^i, a_t^i)\right)$$

Does this work well in practice?

(okayish …)



Supervised Learning



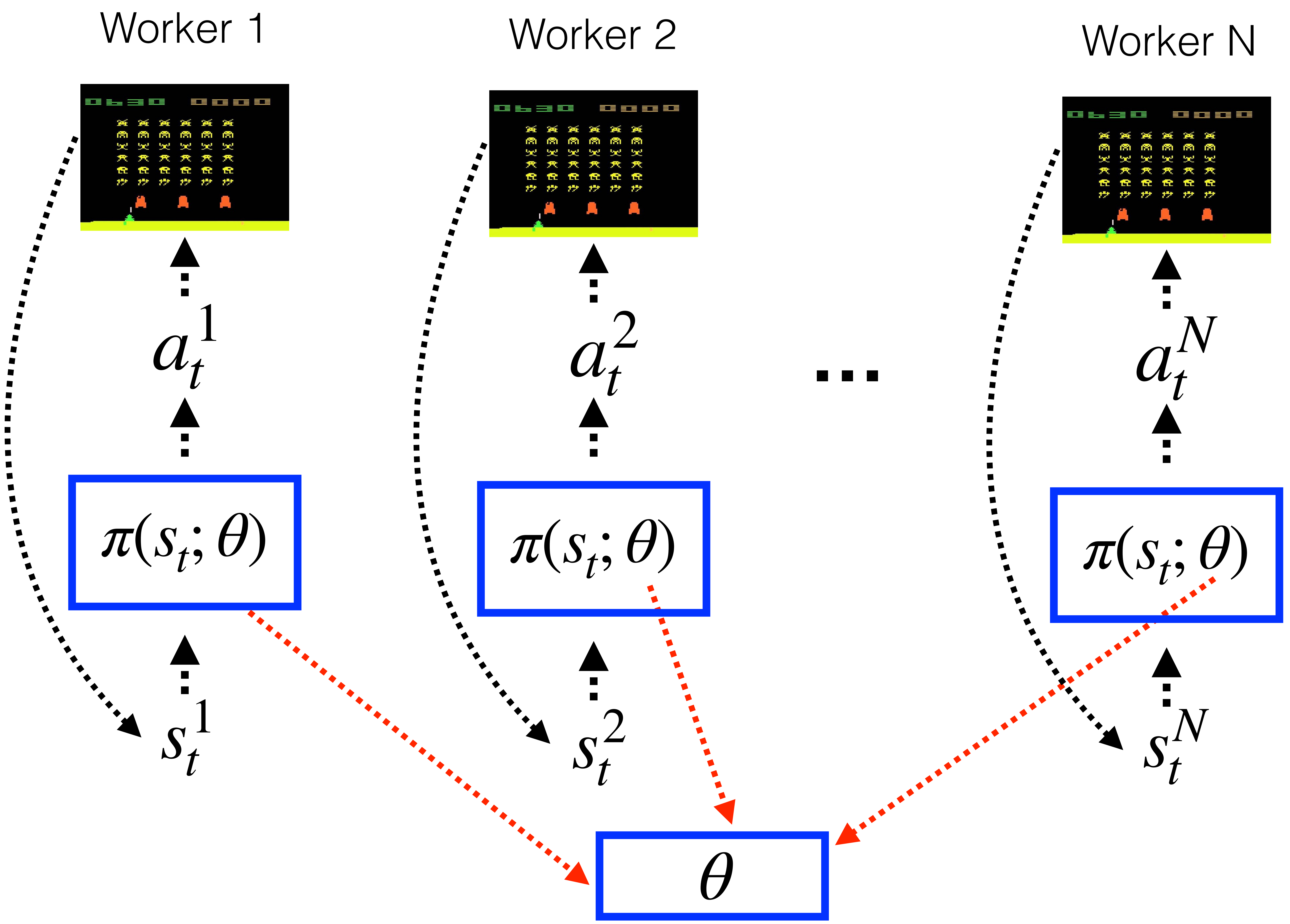
Reinforcement Learning

Stumble into a local minima → Training data collected near this minima

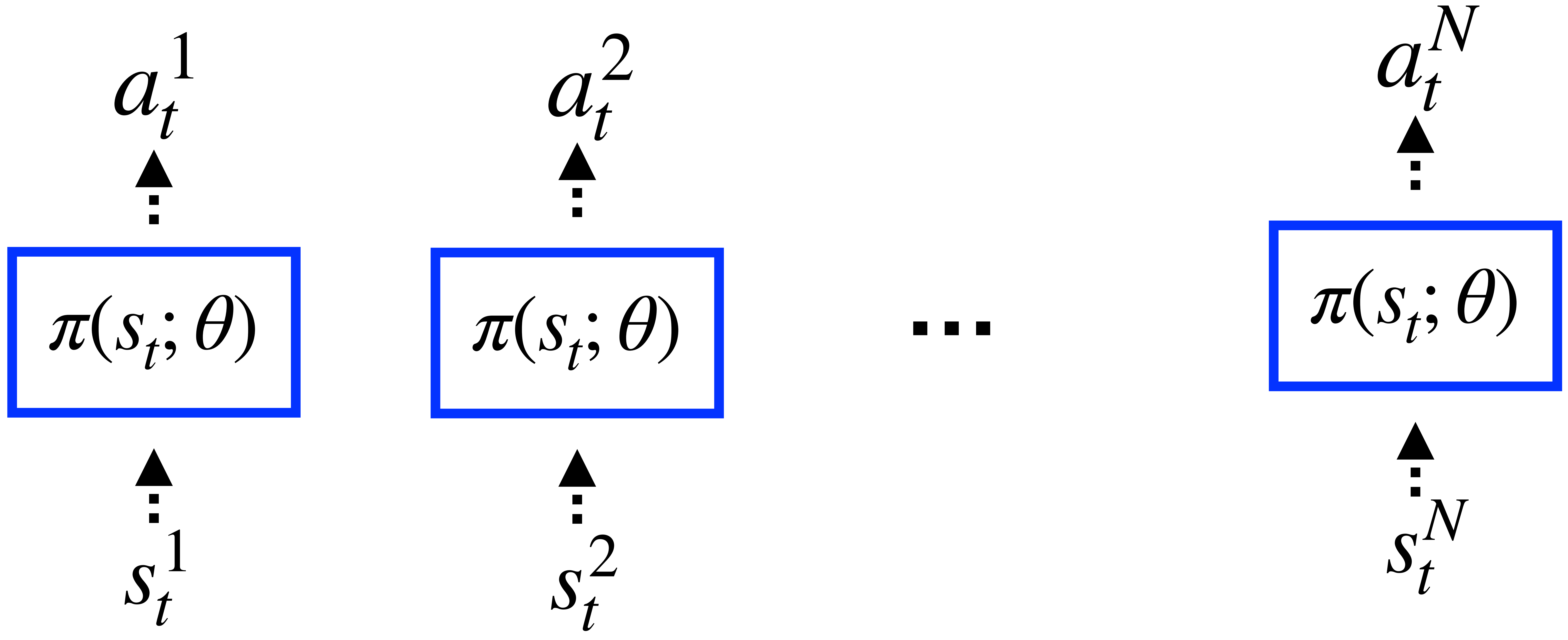
Vicious Cycle ☹

How to
Overcome this problem?

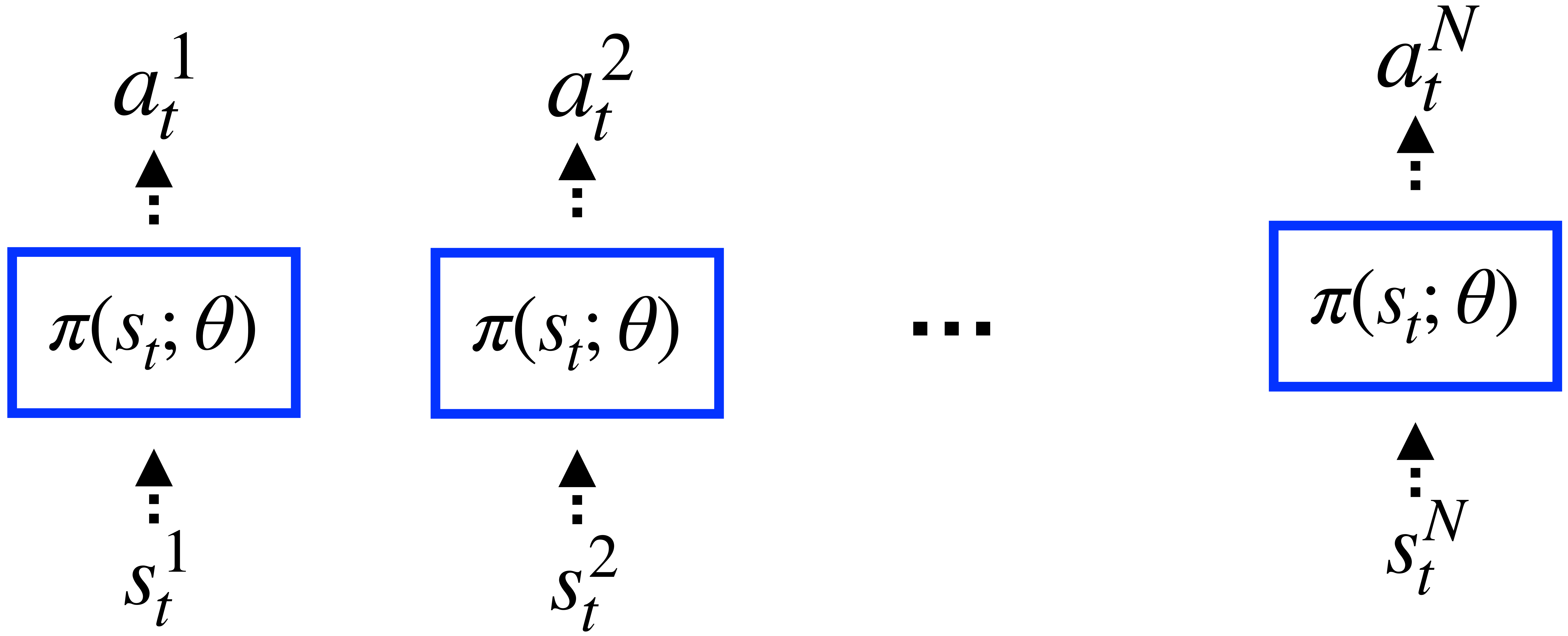
Maintain data-diversity!



(Shared parameters, updated asynchronously; e.g. HogWild)



What's the advantage of N workers?



What's the advantage of N workers?

Each worker has different exploration: more diversity!

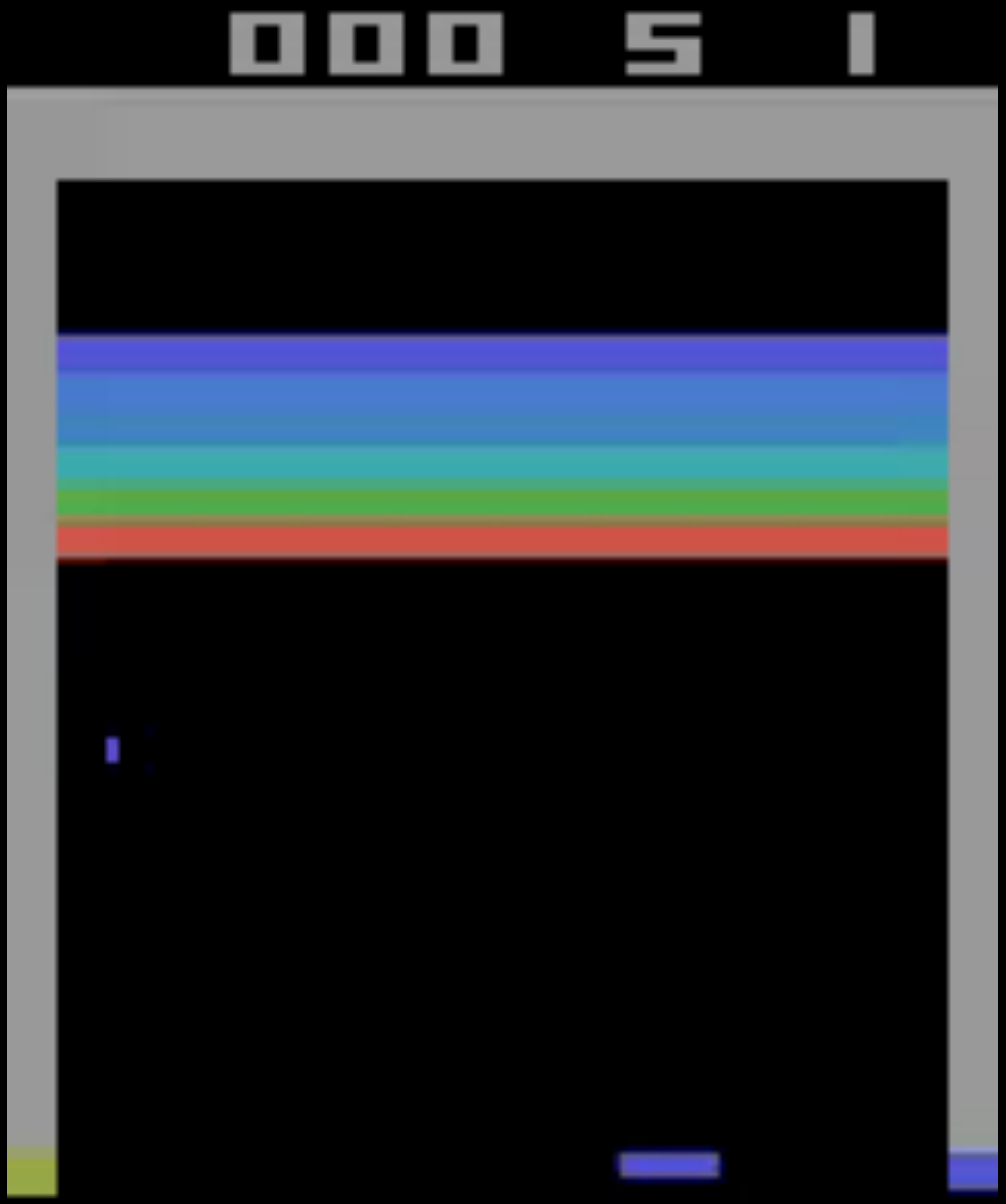
Increase it even more by encouraging **high-entropy in actions**

$$\nabla_\theta \log \pi_\theta(a_t \mid s_t) A(s_t, a_t) + \boxed{\beta \nabla_\theta H\left(\pi_\theta(a_t \mid s_t)\right)}$$

Asynchronous Advantage Actor Critic (A3C)

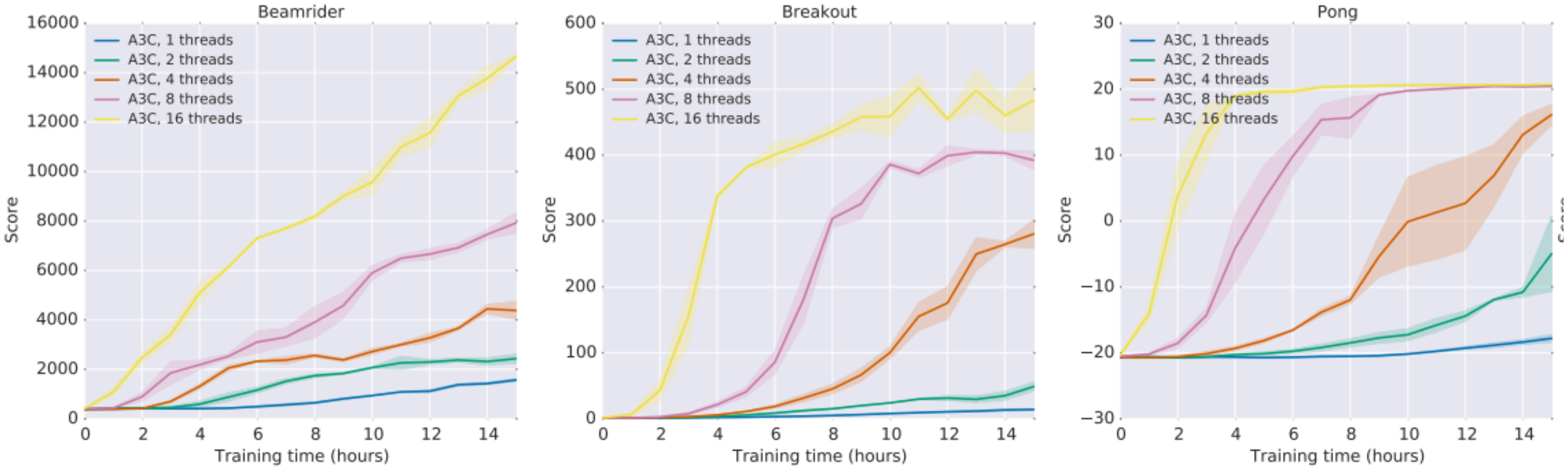
Applying to ATARI Games

Breakout

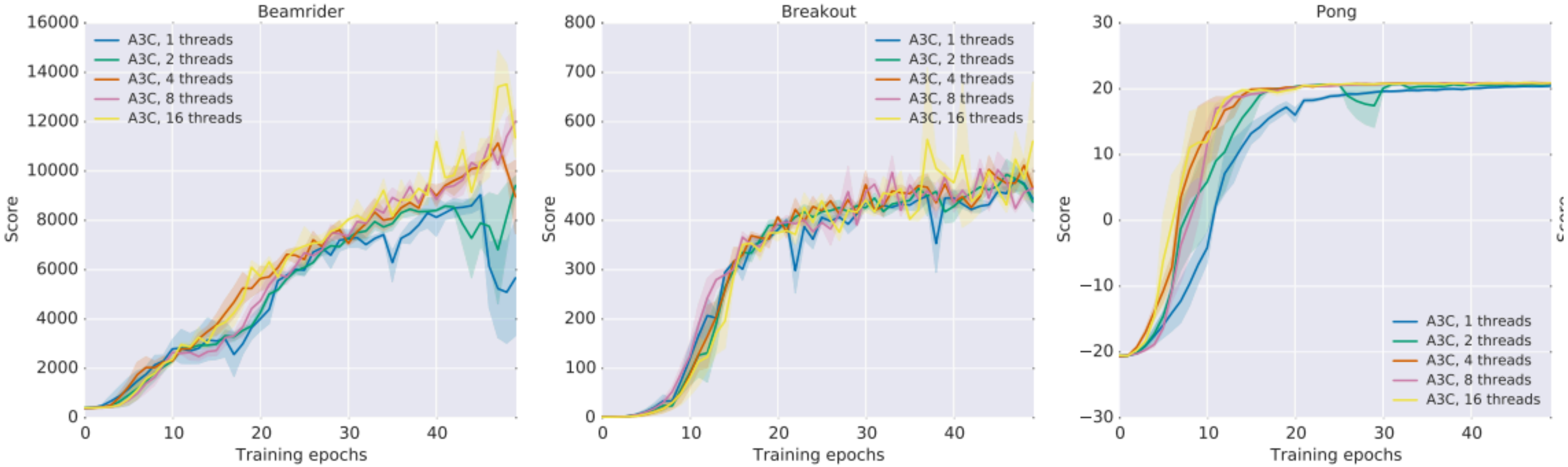


Beamrider

Training speed improvements



Not Data Efficiency



Reducing Variance

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

- Discounting
- Causality
- Collect more data
- Baselines
- **Use of Critic**

Can we better tradeoff bias and variance?

Recall

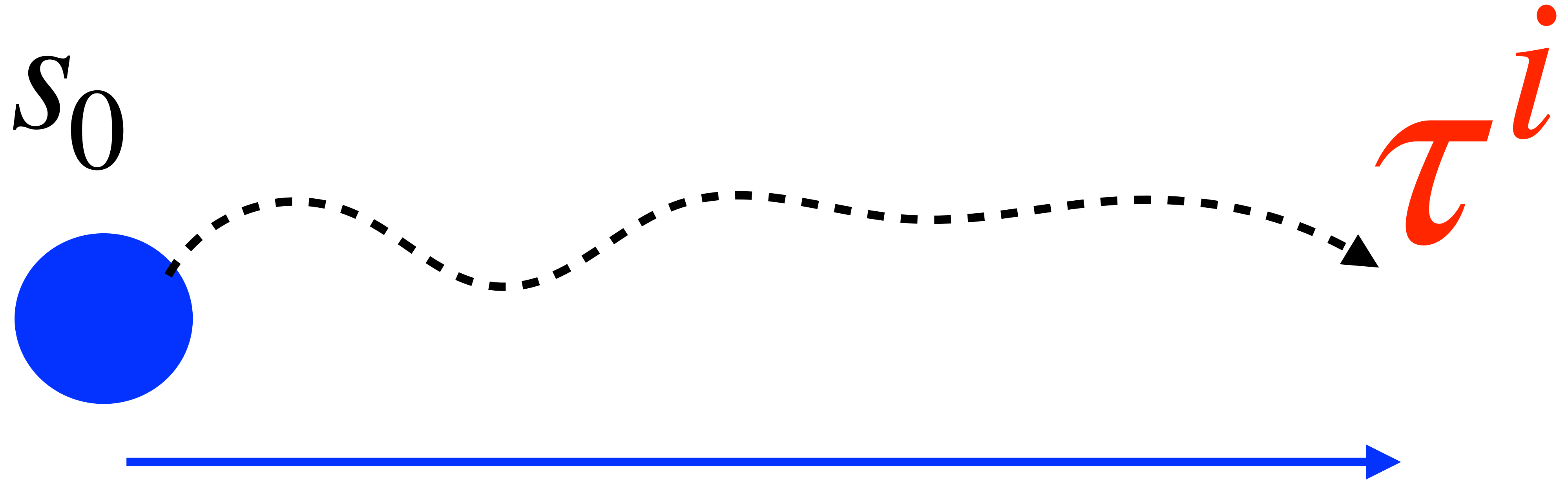
$$\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(a_t^i \mid s_t^i)\left(\left(\sum_{t'=t}^{T}\gamma^{t'-t} r(s_{t'}^i, a_{t'}^i)\right) - V(s_{t'})\right)\right)$$

Variance **v/s** Bias

$$\sum_{t'=t}^{T}\gamma^{t'-t} r_{t'} \quad \textbf{v/s} \quad r_{t'} + \gamma V(s_{t'+1})$$

Trade off variance with bias ..

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



$= r_{t'} + \gamma V(s_{t'+1})$ $\quad R_1$

$= r_{t'} + \gamma r_{t'+1} + \ldots + \gamma^k r_{t'+k} + \gamma^{k+1} V(s_{t'+k+1})$

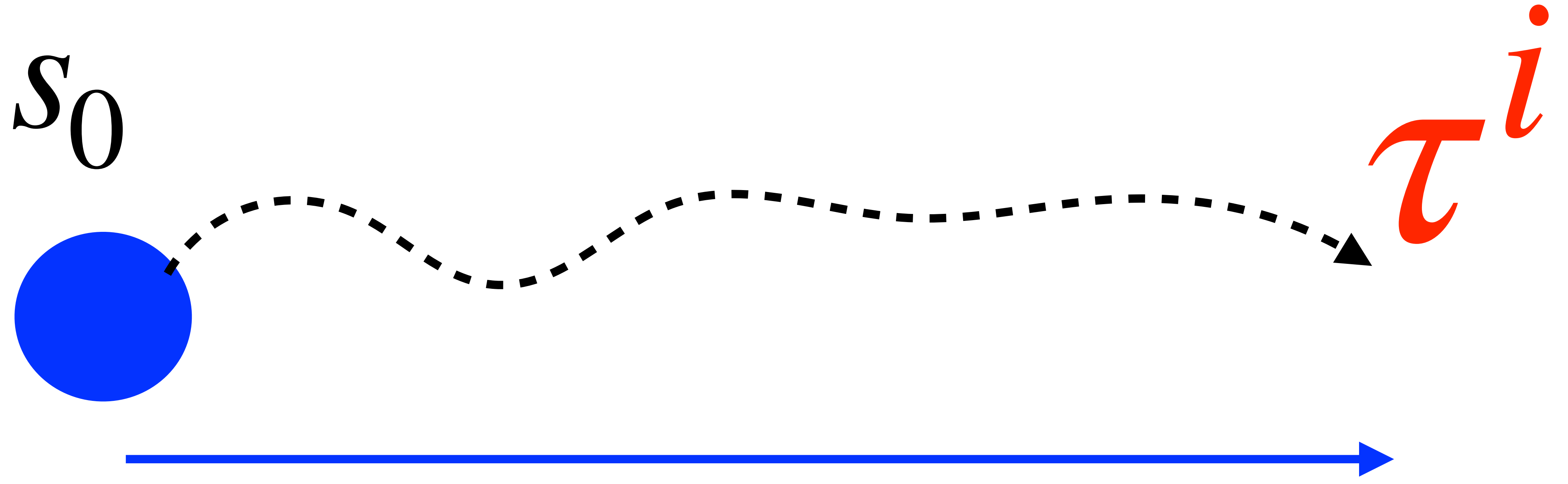
R_k

$$\frac{1}{T}\sum_{k=1}^{T}\lambda_k R_k$$

Trade off variance with bias ..

Generalized Advantage Estimation

$$\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(a_t^i \mid s_t^i)\left(\frac{1}{T}\sum_{k=1}^{T}\lambda_k R_k - V(s_{t'})\right)\right)$$



$$= r_{t'} + \gamma V(s_{t'+1}) \qquad R_1$$

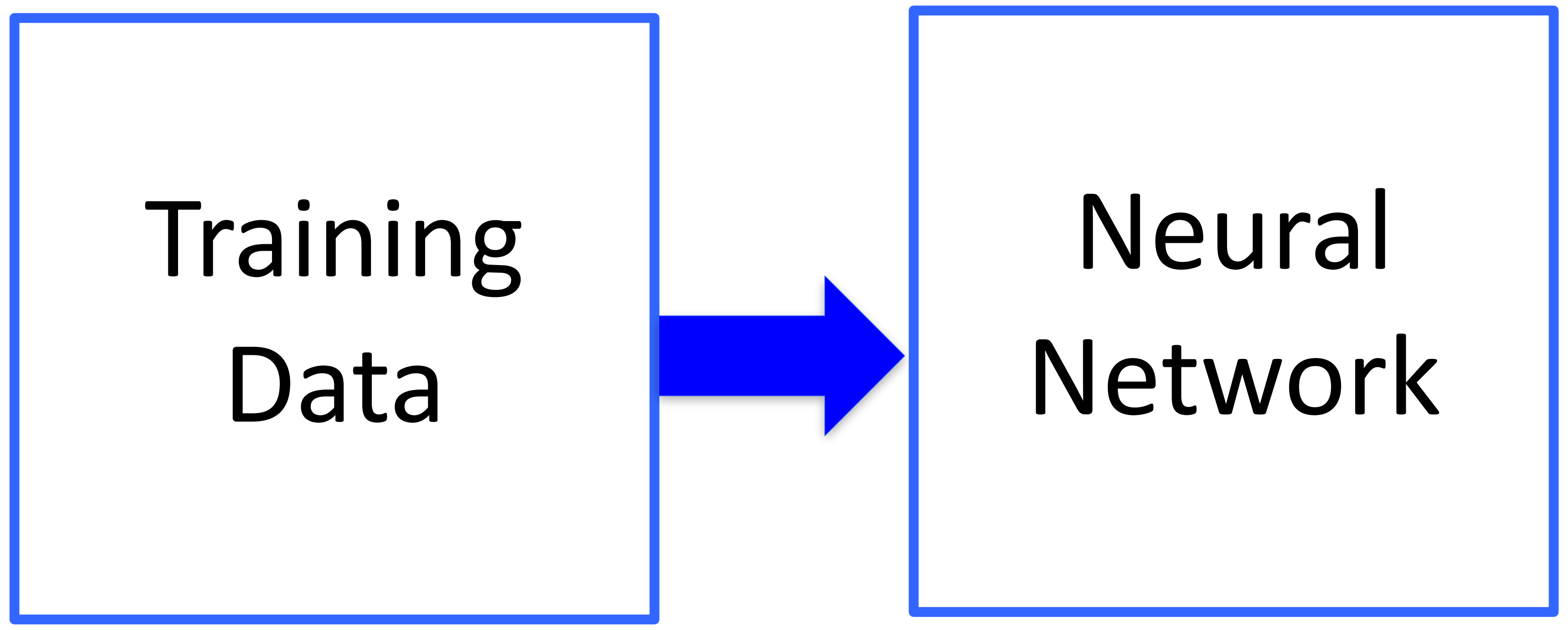
$$= r_{t'} + \gamma r_{t'+1} + \ldots + \gamma^k r_{t'+k} + \gamma^{k+1} V(s_{t'+k+1})$$

$$R_k$$

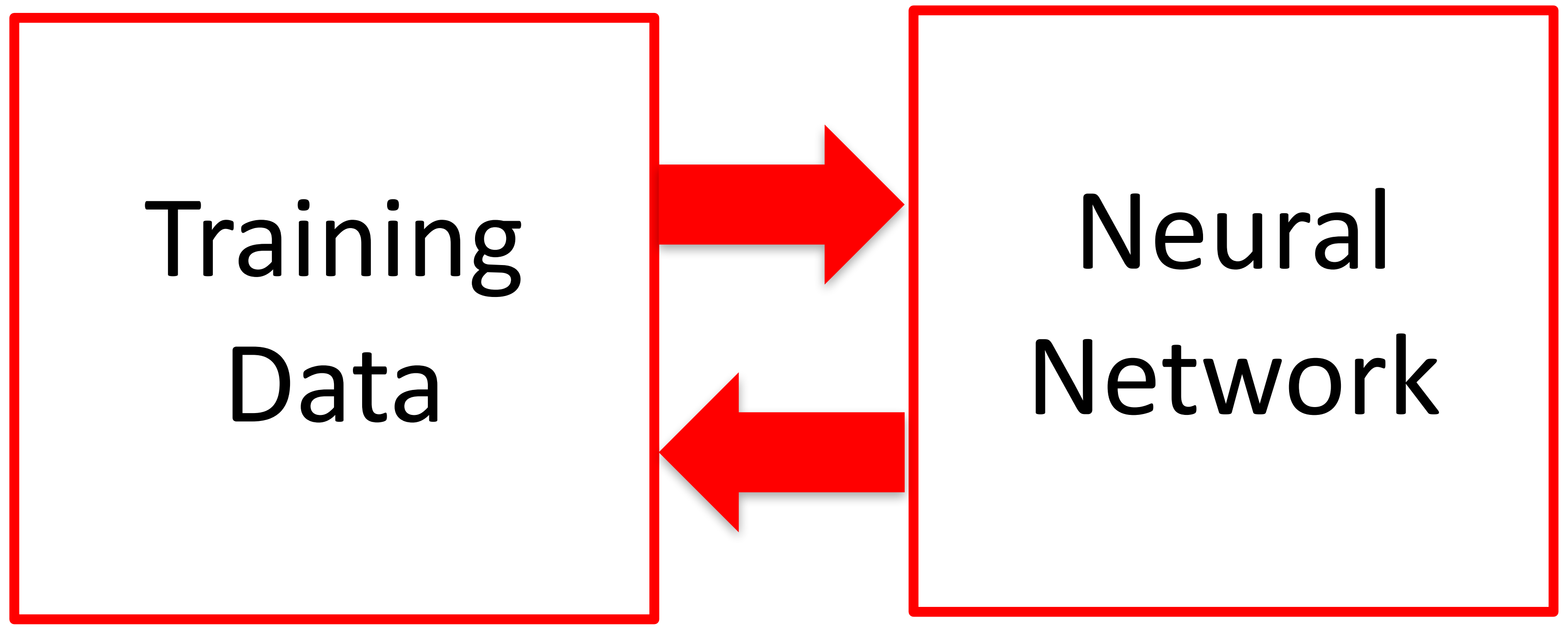
Reducing Variance

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

- Discounting
- Causality
- Collect more data
- Baselines
- Use of Critic
  - **Generalized Advantage Estimation**



Supervised Learning



Reinforcement Learning

Supervised Learning & RL

FORWARD AND REVERSE KL

Recall Comparison with Supervised Learning

RL

Supervised Learning

$$\sum_t r_t$$

$$\tau^{gt} = (s_1, a_1^{gt}, s_2, a_2^{gt}, \ldots)$$

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

Policy Gradients

$$E_{\tau^{gt}}[\nabla_\theta \left(\log p_\theta(\tau^{gt})\right)]$$

Maximum Likelihood

Forward KL

$$\min_{\theta} D_{KL}(P \,||\, Q_{\theta})$$

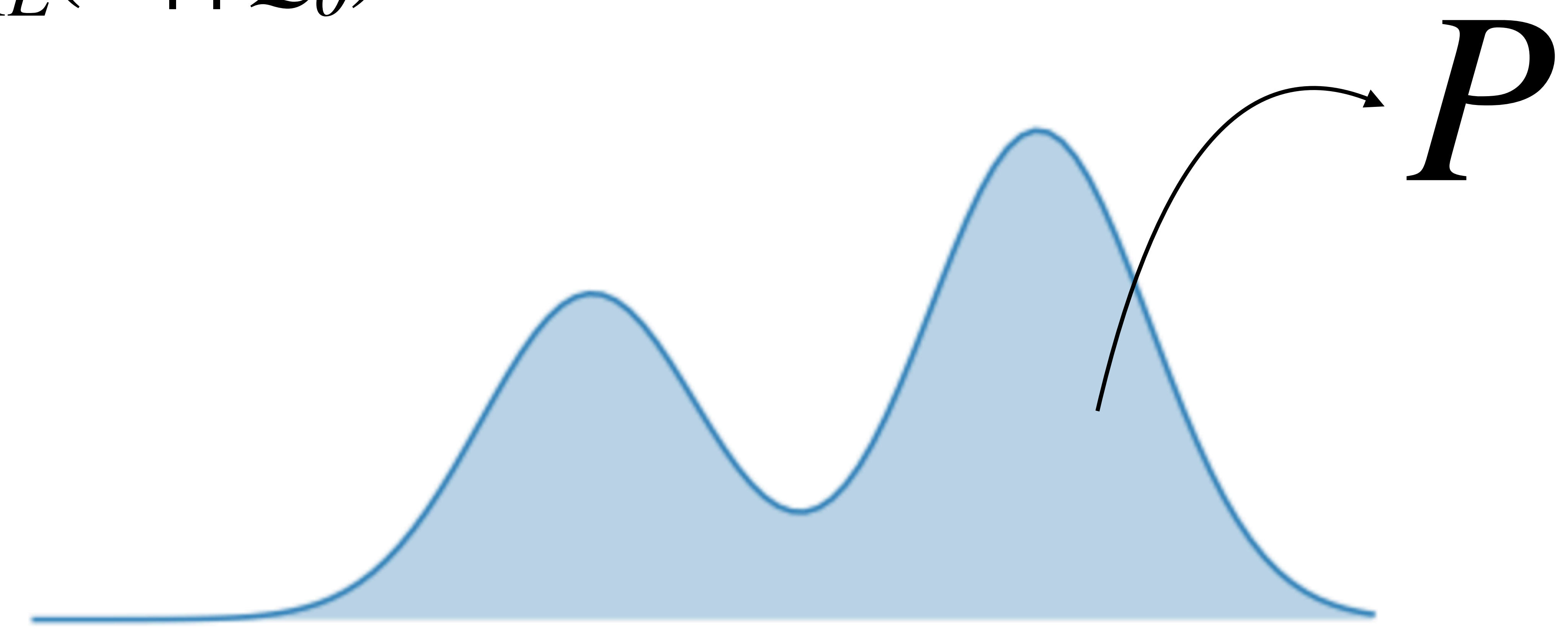
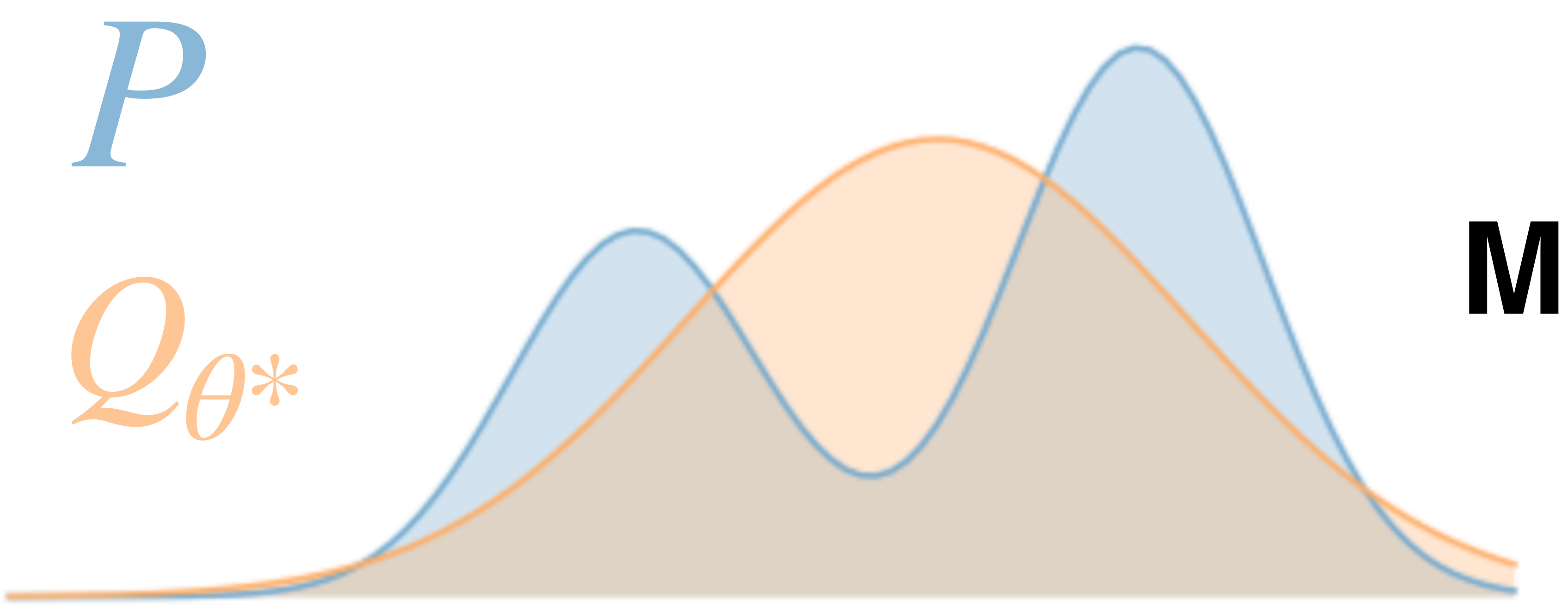


Figure Credits: Dibya Ghosh

Forward KL

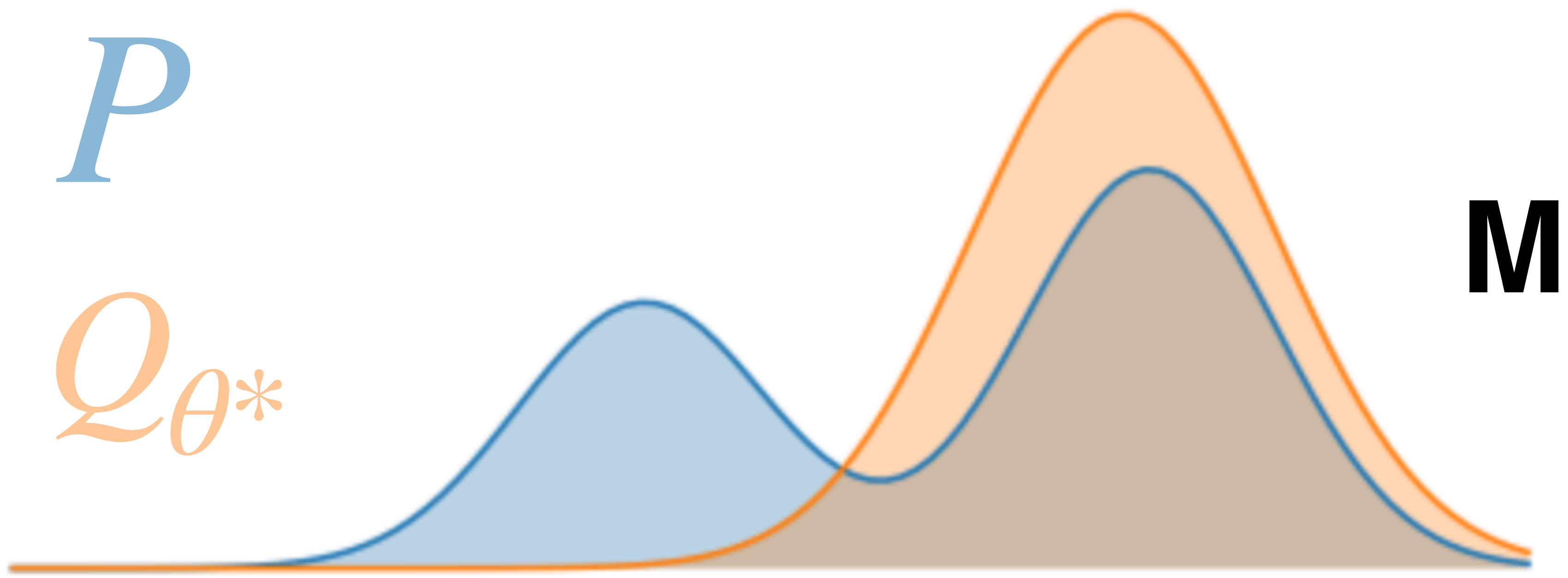
$$\min_{\theta} D_{KL}(P\,||\,Q_\theta)$$



Mean Matching

Reverse KL

$$\min_{\theta} D_{KL}(Q_\theta\,||\,P)$$

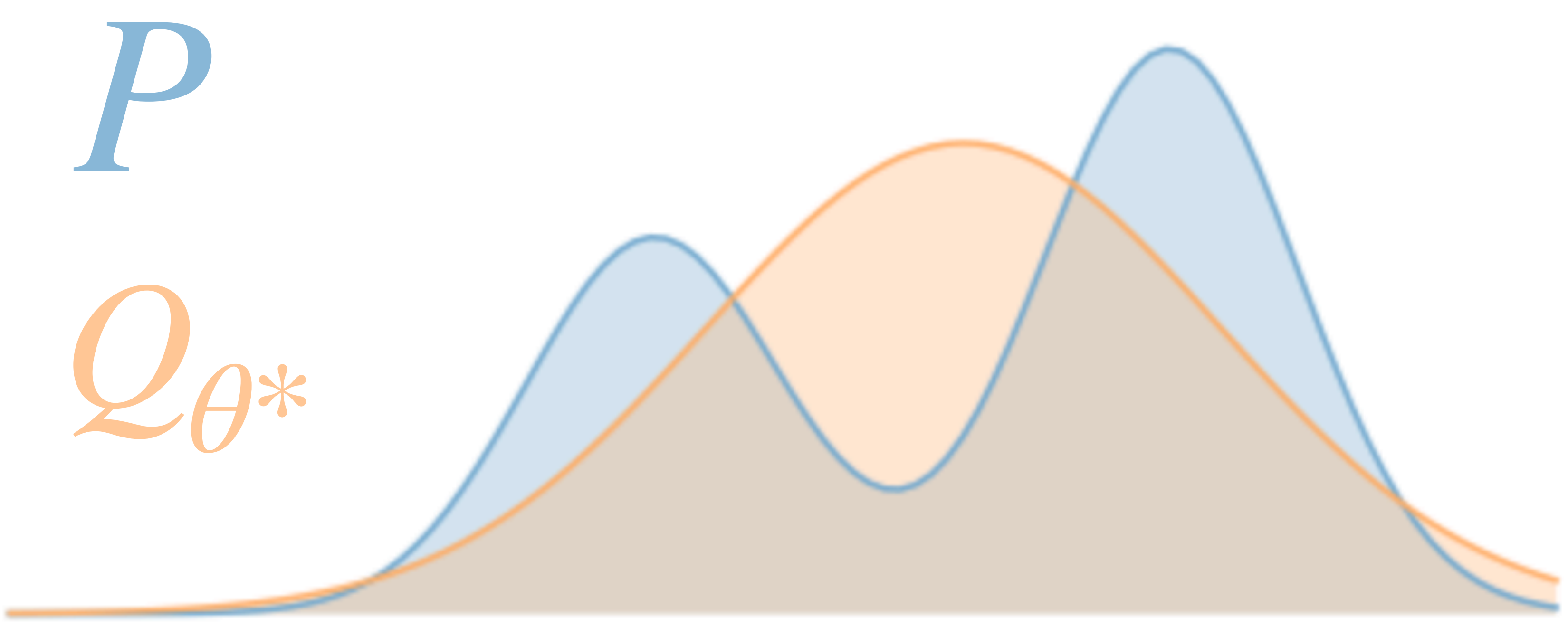


Mode Matching

Figure Credits: Dibya Ghosh

Forward KL

$$\min_{\theta} D_{KL}(P\,||\,Q_\theta) \qquad \nabla_\theta E_{P(x)} \log \frac{P(x)}{Q_\theta(x)}$$



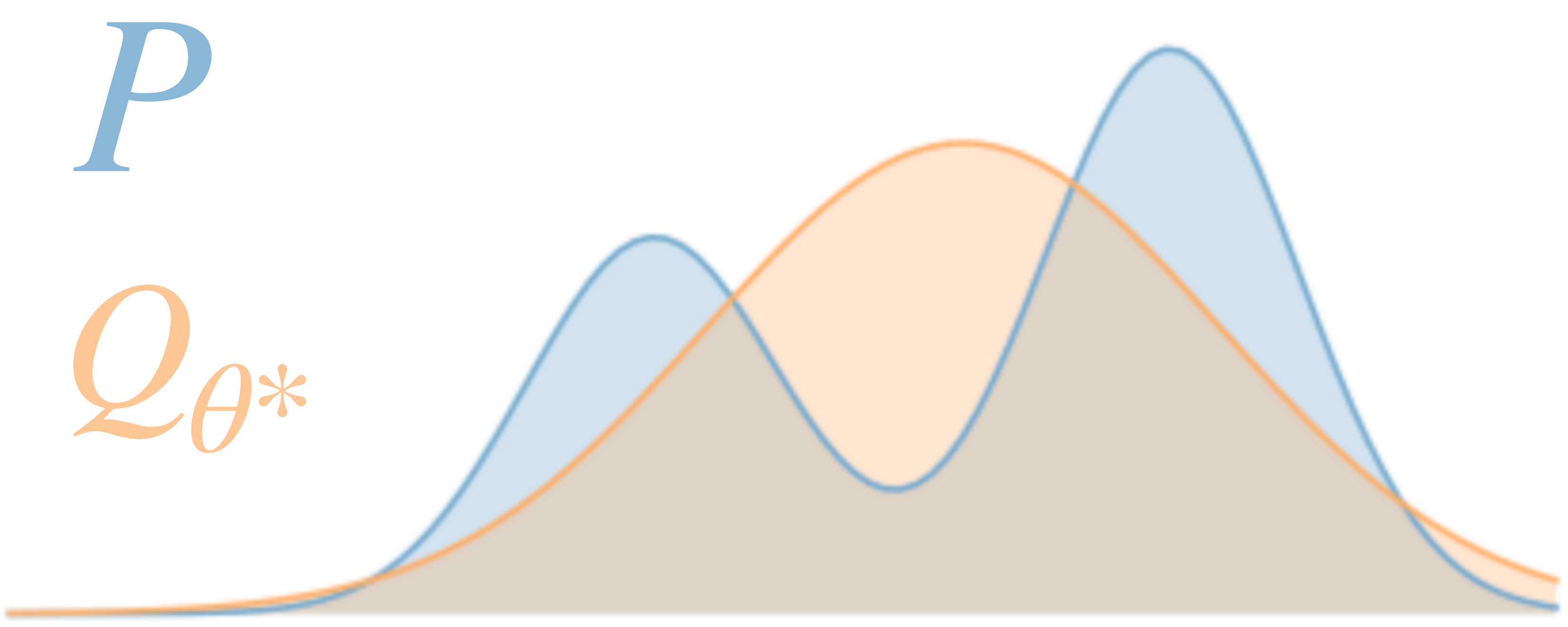
Reverse KL

$$\min_{\theta} D_{KL}(Q_\theta\,||\,P)$$

Figure Credits: Dibya Ghosh

Forward KL

$$\min_{\theta} D_{KL}(P || Q_{\theta}) \quad \nabla_{\theta} E_{P(x)} \log \frac{P(x)}{Q_{\theta}(x)} = - E_{P(x)} \nabla_{\theta} \log Q_{\theta}(x)$$



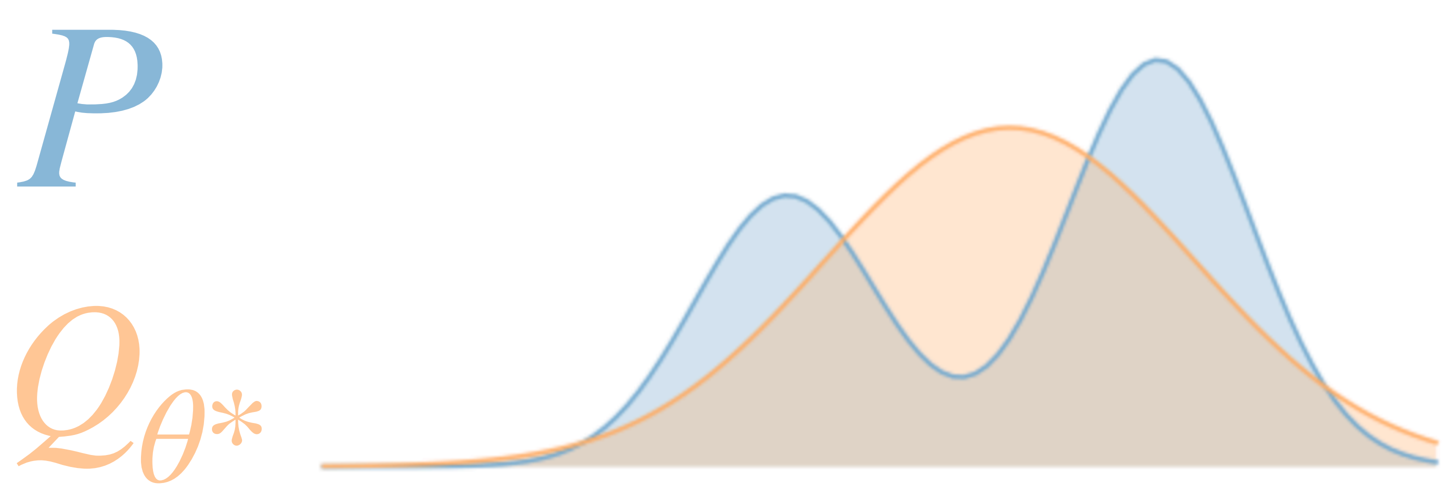
Reverse KL

$$\min_{\theta} D_{KL}(Q_{\theta} || P)$$

Figure Credits: Dibya Ghosh

Forward KL

$$\min_\theta D_{KL}(P\,||\,Q_\theta) \quad \nabla_\theta E_{P(x)} \log \frac{P(x)}{Q_\theta(x)} \;= -\, E_{P(x)} \nabla_\theta \log Q_\theta(x)$$



$$E_{\tau^{gt}}[\,\nabla_\theta\big(\log Q_\theta(\tau^{gt})\big)\,]$$

Supervised Learning

Reverse KL

$$\min_\theta D_{KL}(Q_\theta\,||\,P)$$

Figure Credits: Dibya Ghosh

Forward KL

$$\min_{\theta} D_{KL}(P\,||\,Q_\theta) \quad \nabla_\theta E_{P(x)} \log \frac{P(x)}{Q_\theta(x)} \quad = - E_{P(x)} \nabla_\theta \log Q_\theta(x)$$

$$E_{\tau^{gt}}[\nabla_\theta \left(\log Q_\theta(\tau^{gt})\right)]$$

Supervised Learning

Reverse KL

$$\min_{\theta} D_{KL}(Q_\theta\,||\,P) \quad = E_{Q_\theta(x)} \log \frac{Q_\theta(x)}{P(x)} \quad = - E_{Q_\theta(x)} \log P(x) + E_{Q_\theta(x)} \log Q_\theta(x)$$

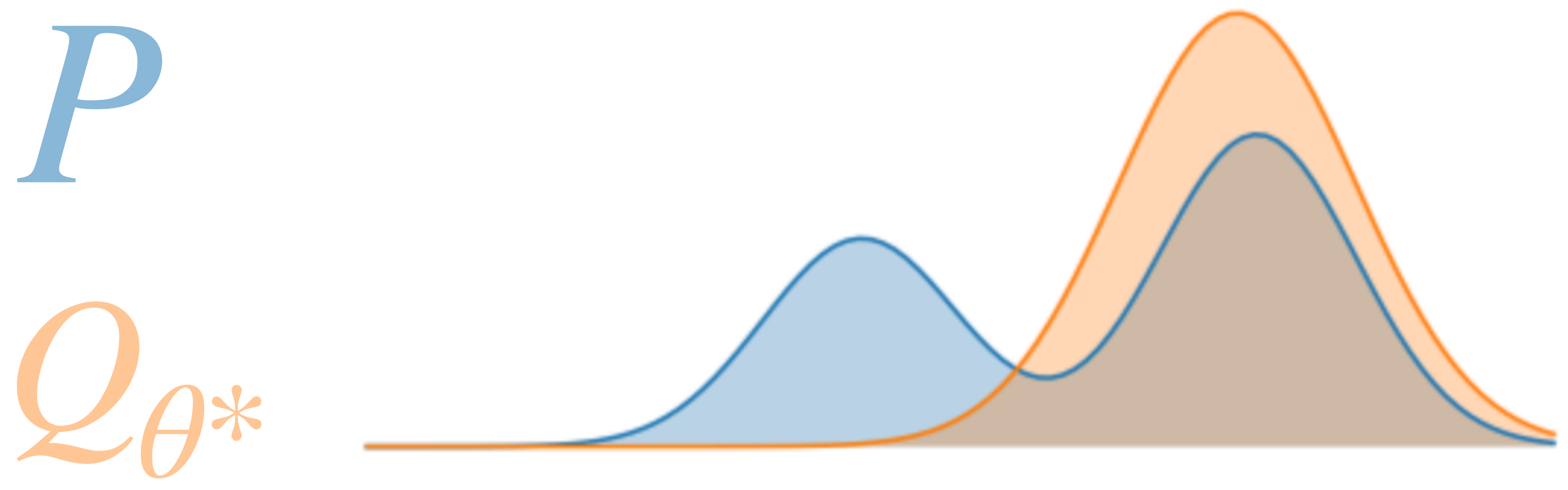


Figure Credits: Dibya Ghosh

Forward KL

$$\min_\theta D_{KL}(P||Q_\theta) \quad \nabla_\theta E_{P(x)} \log \frac{P(x)}{Q_\theta(x)} = -E_{P(x)} \nabla_\theta \log Q_\theta(x)$$

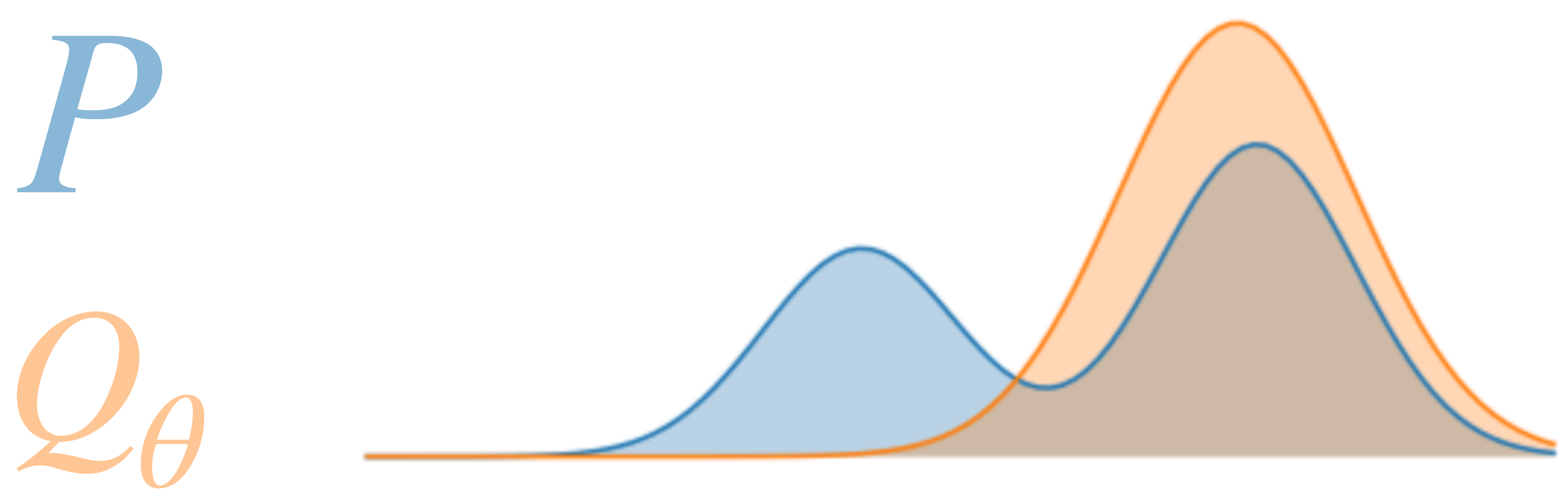
$$E_{\tau^{gt}}[\nabla_\theta(\log Q_\theta(\tau^{gt}))]$$

Supervised Learning

Reverse KL

$$\min_\theta D_{KL}(Q_\theta||P) = E_{Q_\theta(x)} \log \frac{Q_\theta(x)}{P(x)} = -E_{Q_\theta(x)} \log P(x) + E_{Q_\theta(x)} \log Q_\theta(x)$$

$$\max_\theta E_{Q_\theta(x)} \log P(x) - E_{Q_\theta(x)} \log Q_\theta(x)$$



Consider $P(x) \sim e^{R(\tau)}$

Max-Entropy RL Objective

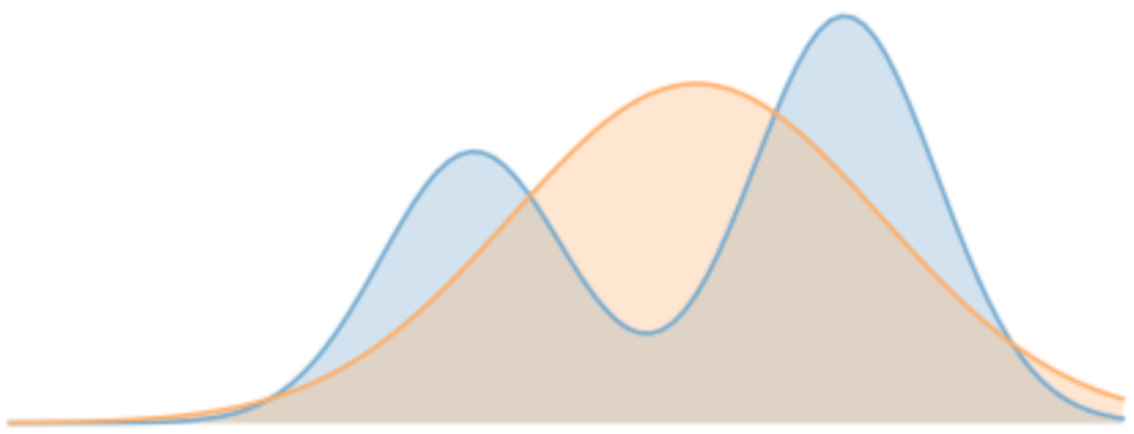
$$\max_\theta E_{\tau:Q_{\theta(x)}}[R(\tau)] + \mathscr{H}\left(Q_\theta(x)\right)$$

Forward KL

$$\min_{\theta} D_{KL}(P\,||\,Q_\theta)$$

P

Q_θ



$$E_{\tau^{gt}}[\,\nabla_\theta \left(\log Q_\theta(\tau^{gt})\right)]$$

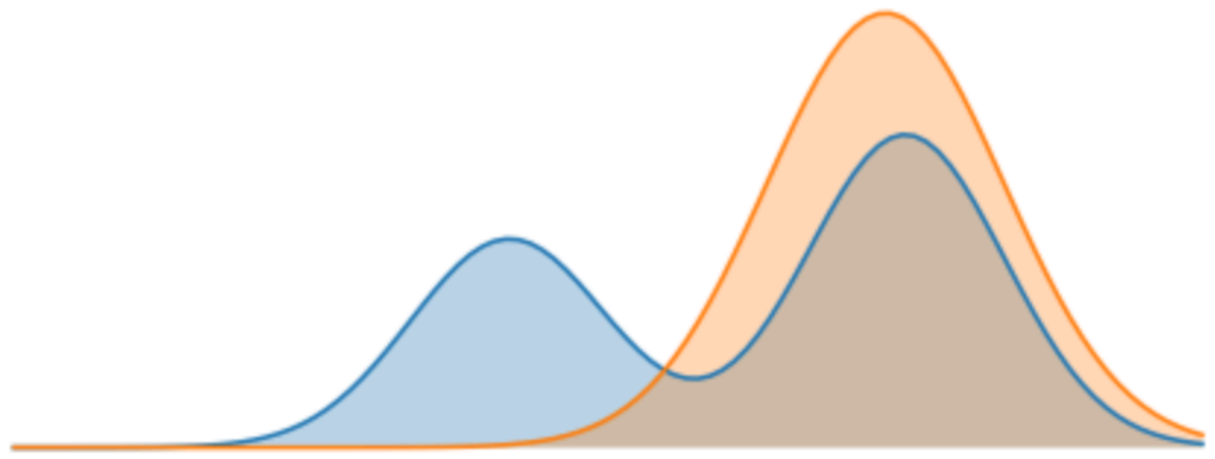
Supervised Learning

Reverse KL

$$\min_{\theta} D_{KL}(Q_\theta\,||\,P)$$

P

Q_θ



$$\max_{\theta} E_{\tau:Q_{\theta(x)}}[R(\tau)] + \mathscr{H}\left(Q_\theta(x)\right)$$

Max-Entropy RL Objective

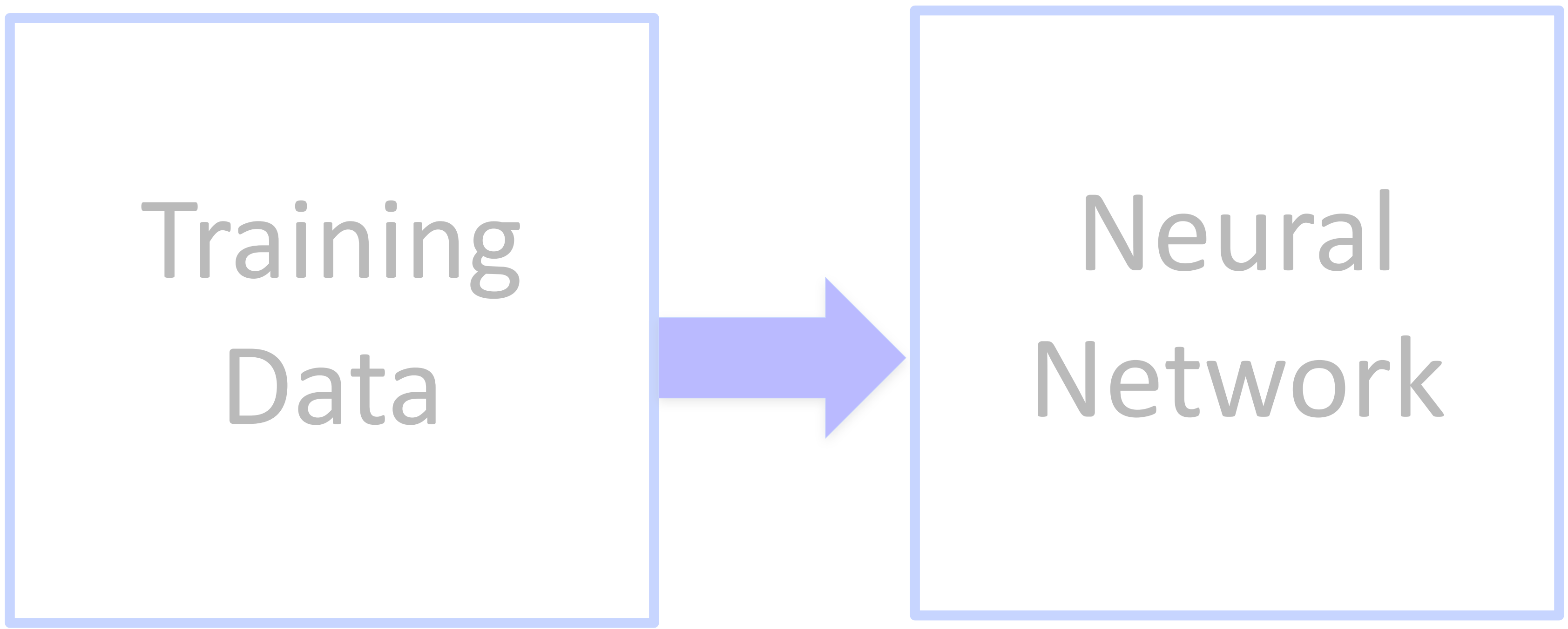
Figure Credits: Dibya Ghosh

End of Variance Reduction

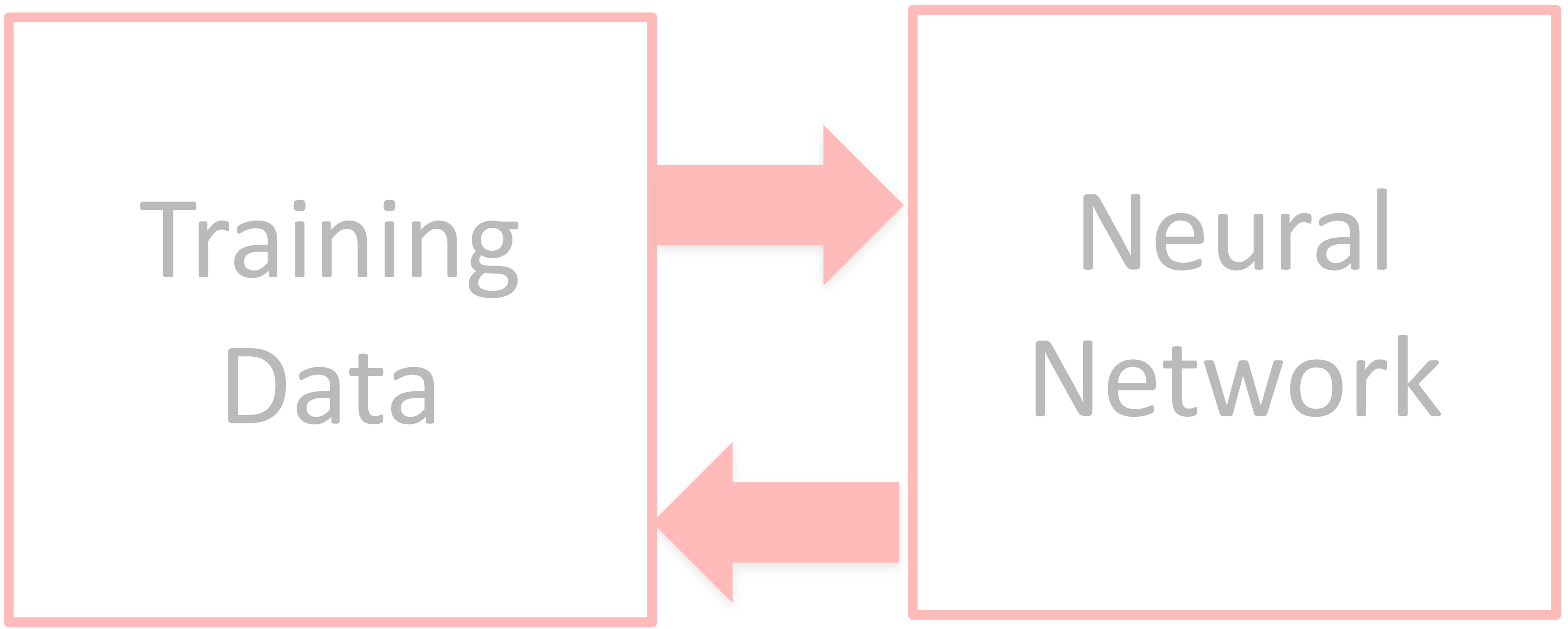
$$\text{Var}_\tau\left[\nabla_\theta\left(\log p_\theta(\tau)\right)R(\tau)\right]$$

- Discounting
- Causality
- Collect more data
  - **Data parallelization**
- Baselines
- Use of Critic
  - *Generalized Advantage Estimation*

What are other ways to learn a better policy?



Supervised Learning



Reinforcement Learning

Stumble into a local minima → Training data collected near this minima

Vicious Cycle ☹

How to
Overcome this problem?

Maintain data-diversity!

How much to update?

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

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

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

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

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

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

Policy Gradients

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

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

Policy Gradients

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

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

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 !!