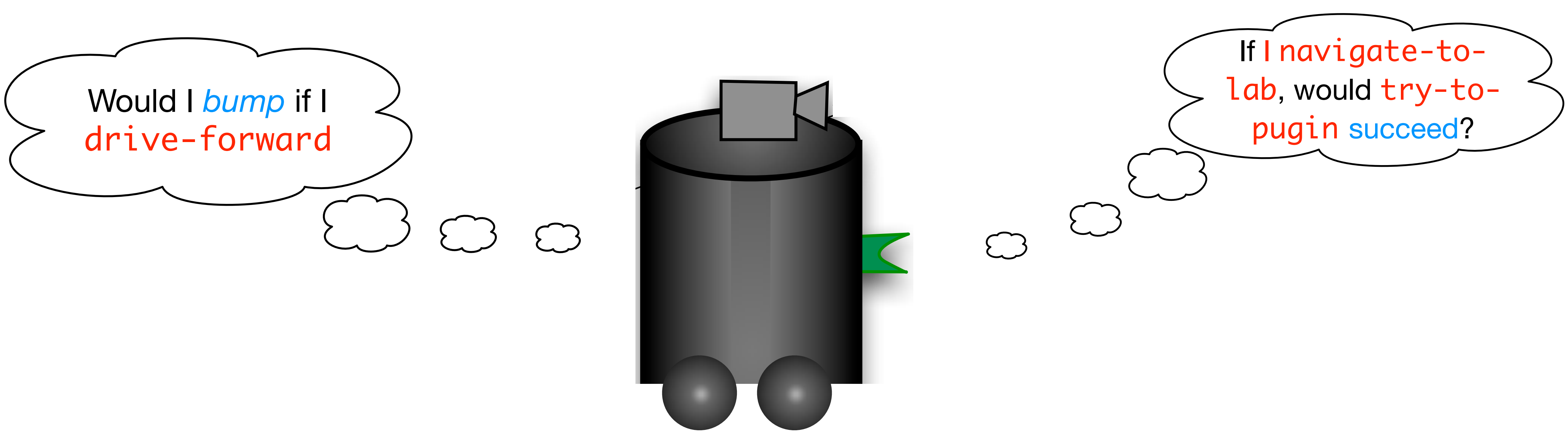
Adapting Behaviour via Intrinsic Rewards to Learn Predictions

Martha White
Assistant Professor
University of Alberta
Joint work with
Cam Linke, Nadia Ady, Thomas Degris and Adam White

Motivation

- Imagine an RL agent is wandering around making many predictions about the world
    - What will happen if I pick up this object?
    - How many steps until I get to the door?
- How should it **act**, to make those **predictions** more **accurate**?



This problem has been studied in many flavours

This problem has been studied in many flavours

- Active learning and optimal experimental design
  - e.g., batch of data, choose most useful subset to label
- Active perception and attention
  - e.g., what part of an image should the agent look at

This problem has been studied in many flavours

- Active learning and optimal experimental design
  - e.g., batch of data, choose most useful subset to label
- Active perception and attention
  - e.g., what part of an image should the agent look at
- How to adapt behaviour to learn **many parallel predictions online** has not been explicitly formalized

This talk is about problem formulation

- It is about **understanding how to formalize** active data gathering for **learning predictions** in parallel
- This talk is **not about**
  - a solution strategy for exploration
  - new algorithms

On arXiv next week:
Adapting Behaviour via Intrinsic Reward: A Survey and Empirical Study

Problem Setting

Problem Setting

- N targets, for which we have N prediction learners
- Online setting: one stream of (nonstationary) experience
- **Goal**: take actions to provide data that makes the N prediction learners as accurate as possible

Problem Setting

- N targets, for which we have N prediction learners
- Online setting: one stream of (nonstationary) experience
- **Goal**: take actions to provide data that makes the N prediction learners as accurate as possible
- **Issues**:
  - unknown world and what data is useful is not obvious
  - different samples are useful for different learners — the behavior needs to balance these needs

How can we formalize this problem?

- **Ideally**, maximize prediction accuracy across time
- **Hard** to specify as a continuous optimization problem
  - action selection indirectly affects prediction accuracy

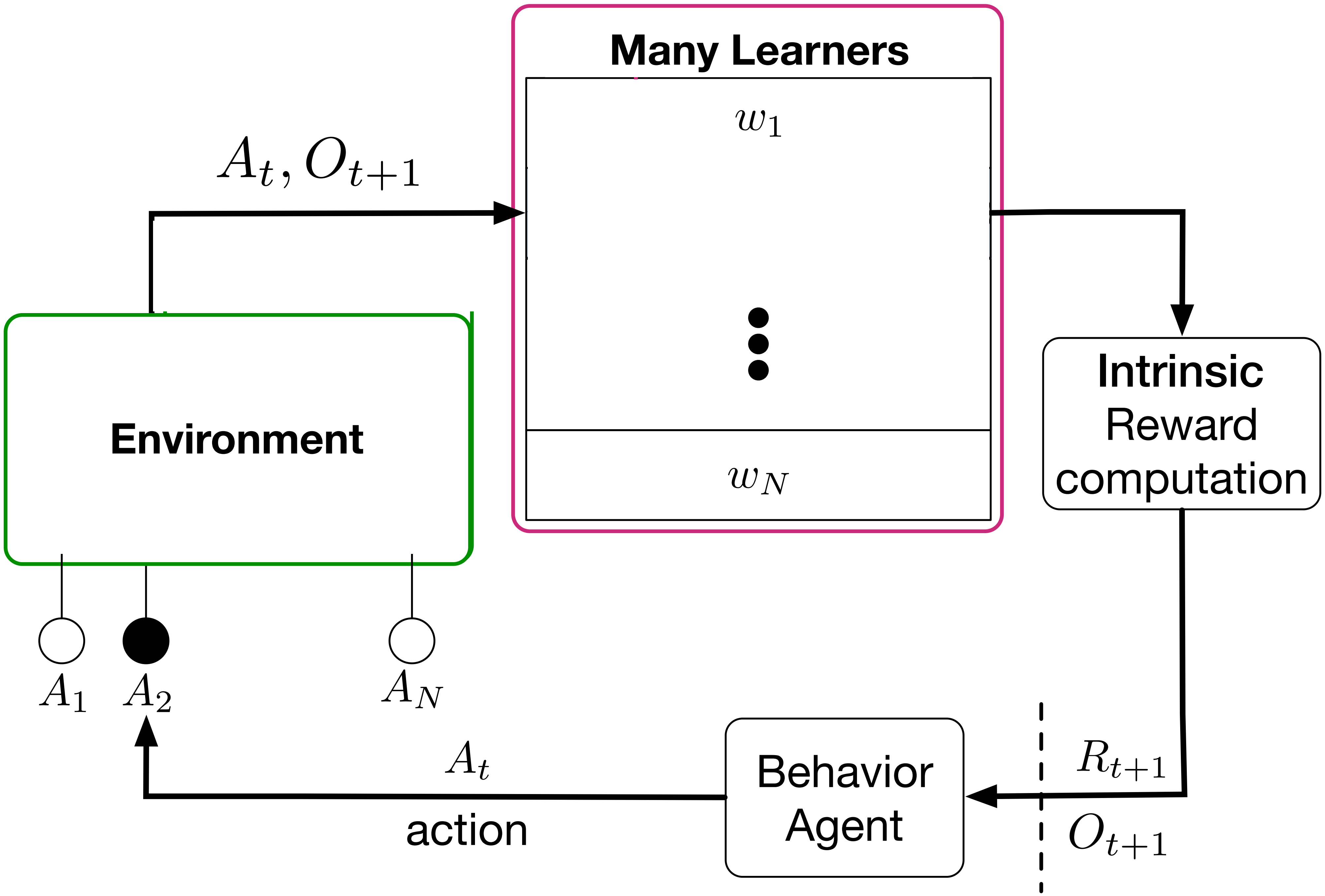
Naturally formulated as an RL problem

- **Reward** the behavior for taking actions that produce data that is "**useful**" for the prediction learners

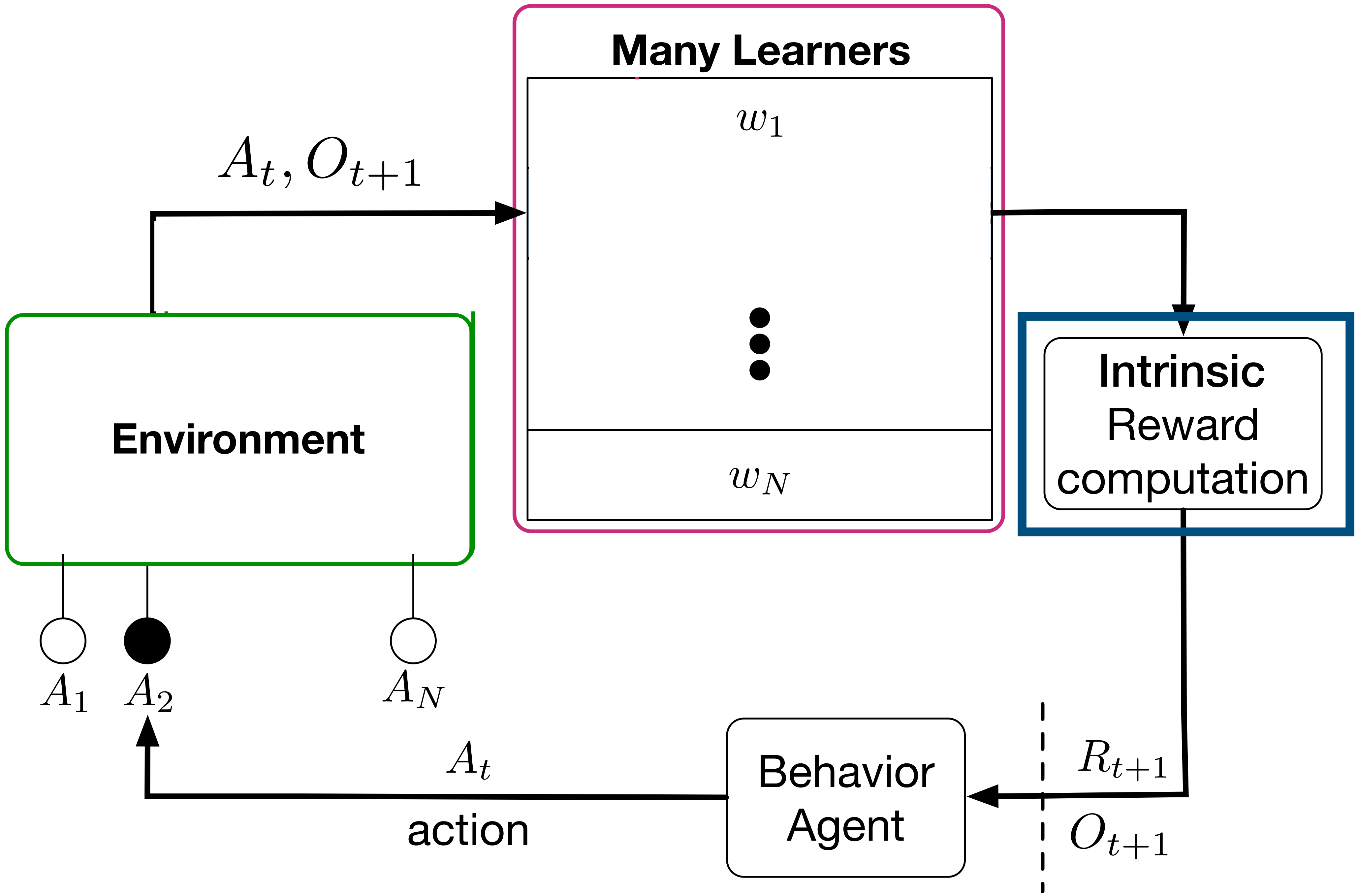
- The **actions** are still the actions in the underlying MDP

- The **states** should be the MDP state and the parameters of the N prediction learners

The Parallel Prediction Learning Problem

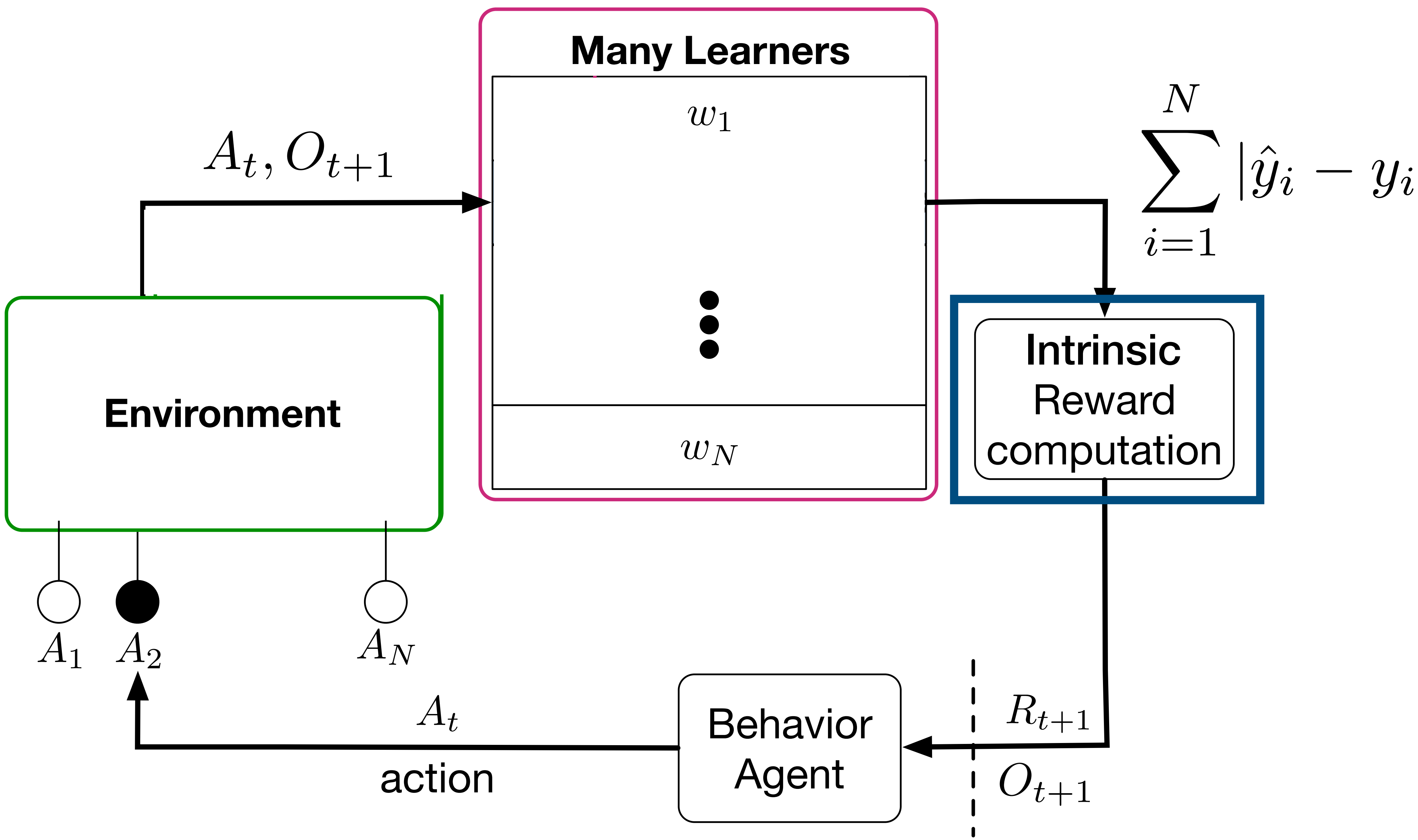


The Parallel Prediction Learning Problem



The key to the formulation is the reward definition

The Parallel Prediction Learning Problem



The key to the formulation is the reward definition

Issue with Error-based rewards

- Prediction Error includes Variance of Targets

$$\sum_{i=1}^{N} |\hat{y}_i - y_i|$$

- What we really want is error to true target

$$\sum_{i=1}^{N} |\hat{y}_i - \mathbb{E}[Y_i]|$$

- …But we do not have the true target

The first step is to understand existing **intrinsic rewards**

Only minor connection to Intrinsic Motivation in RL

- **There the goal is to find an optimal policy**
  - internal reward is added just to encourage exploration
  - accuracy of prediction is secondary, if considered at all
- Example: Count-based methods encourage systematically revisiting all of the space
- Example: Use Model Error to encourage exploration (though predictions from the model are not used)

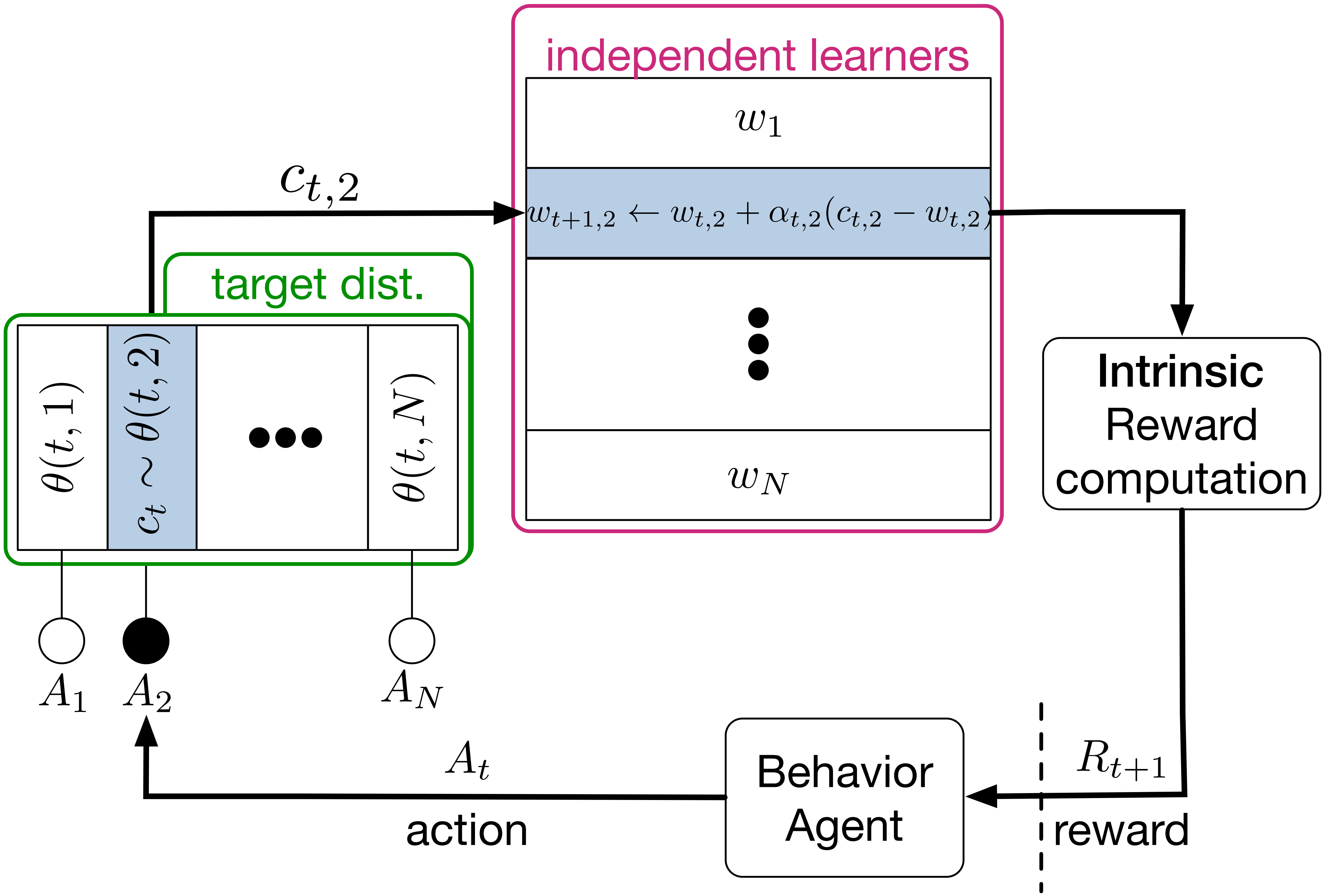
The first step is to understand existing **intrinsic rewards**

using an empirical study

Let's start in the simplest setting

- There is **no context** or state
- Each prediction learner is estimating the **mean** of a **different target** (N independent learners)
- Each **action** only generates **data** for **one** prediction learner
    - there are N actions

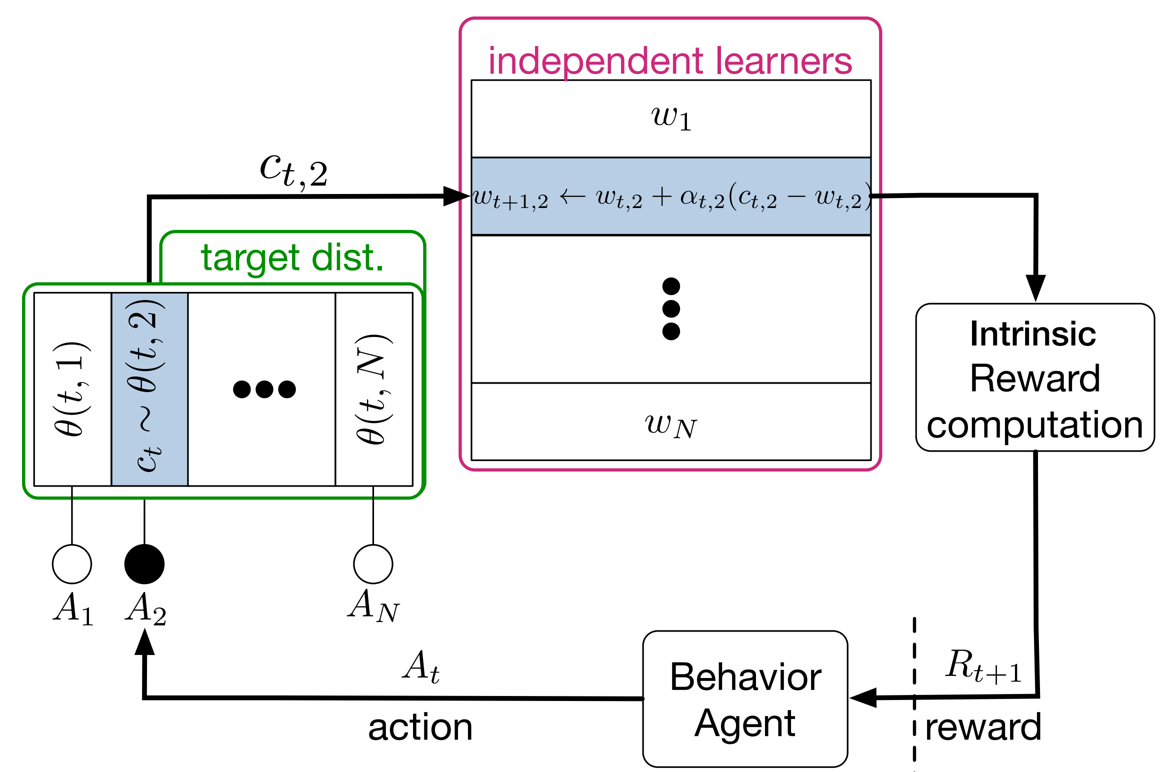
Formalizing a Testbed



$\theta(t, i)$ distribution for ith target on timestep t

This setting still has the key properties of the problem

- Must balance needs of several learners
- Some learners might have harder to estimate targets
- The world is unknown and potentially non-stationary/partially observable
- An appropriate reward for behavior is not obvious



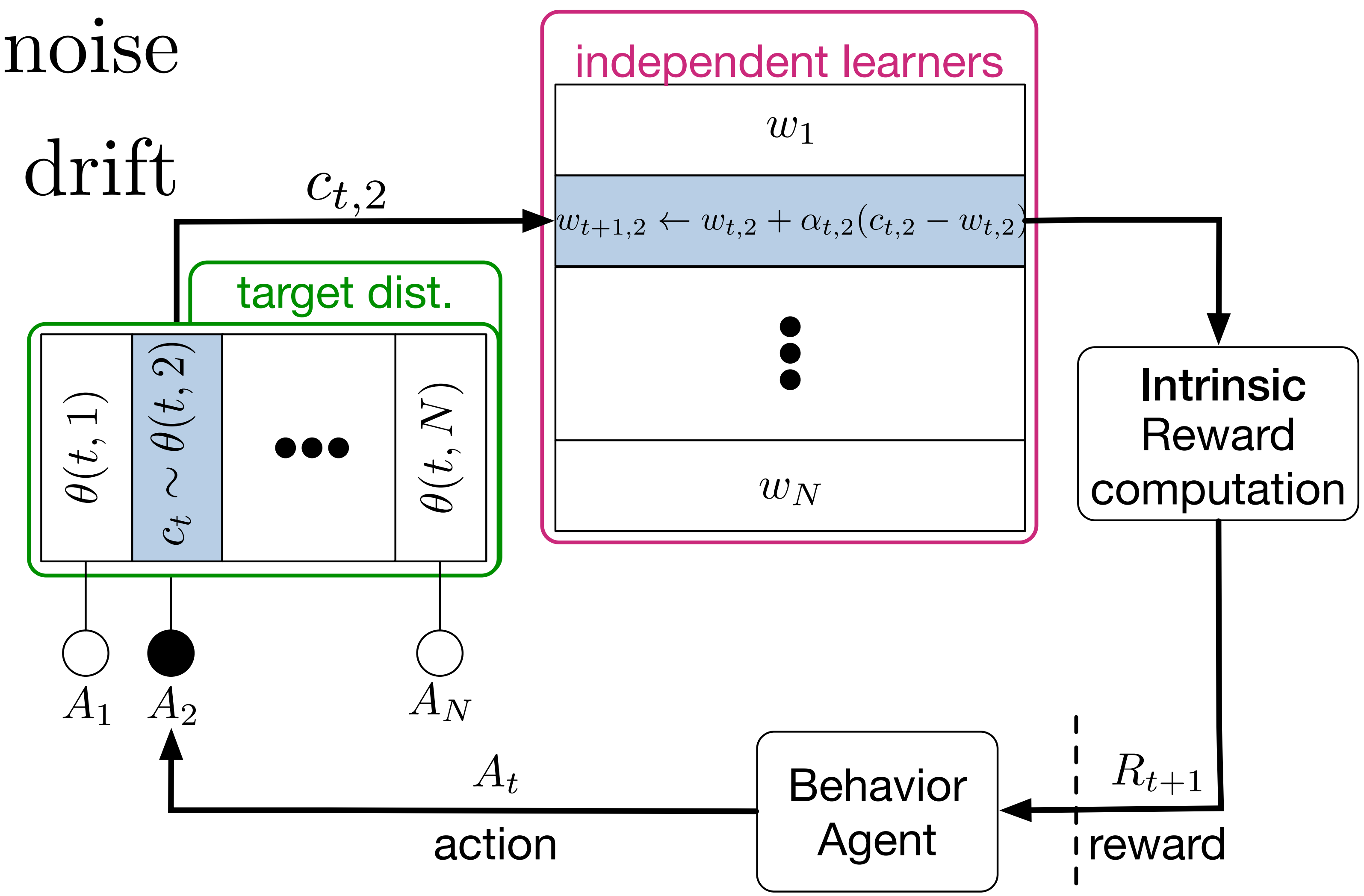
Target distributions

$$\theta(t,i) \stackrel{\text{def}}{=} \mathcal{N}(\mu_{t,i}, \sigma^2_{t,i})$$

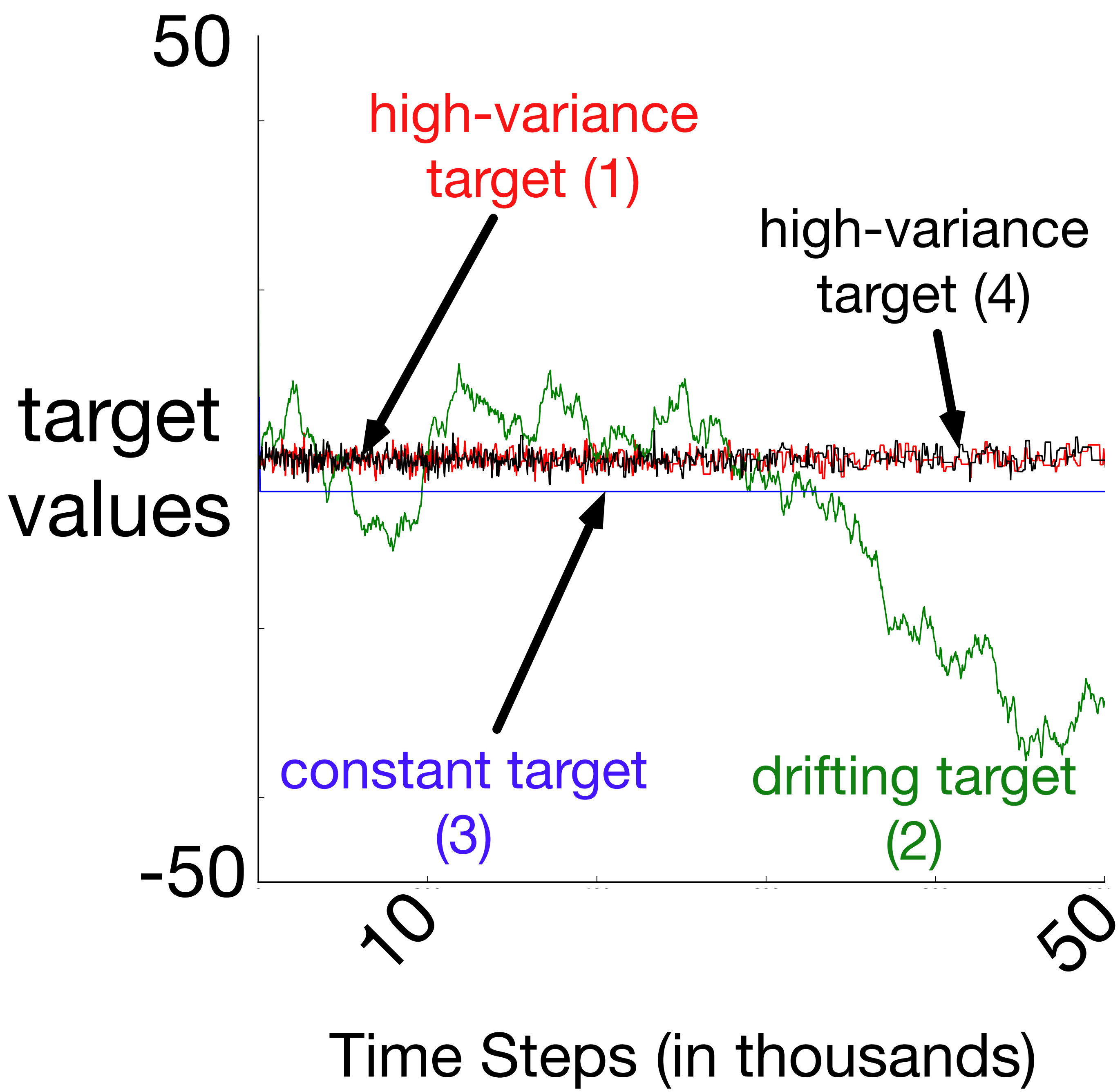
$$\text{for} \quad \mu_{t+1,i} \leftarrow \Gamma_{[-50,50]}\left(\mu_{t,i} + \mathcal{N}(0, \xi^2_{t,i})\right)$$

$\sigma_{t,i} \in \mathbb{R}$ for sampling noise

$\xi_{t,i} \in \mathbb{R}$ for the rate of drift

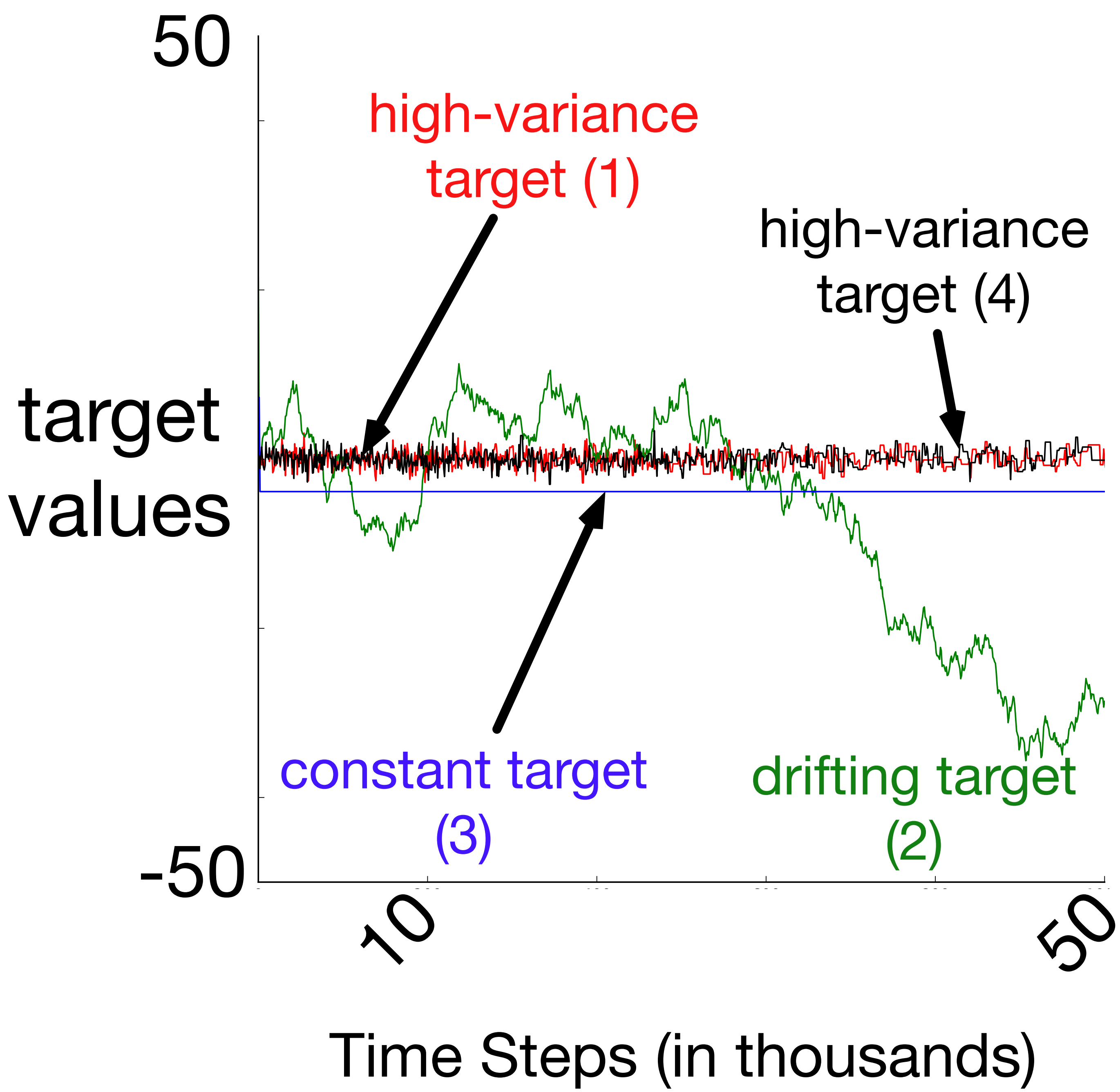


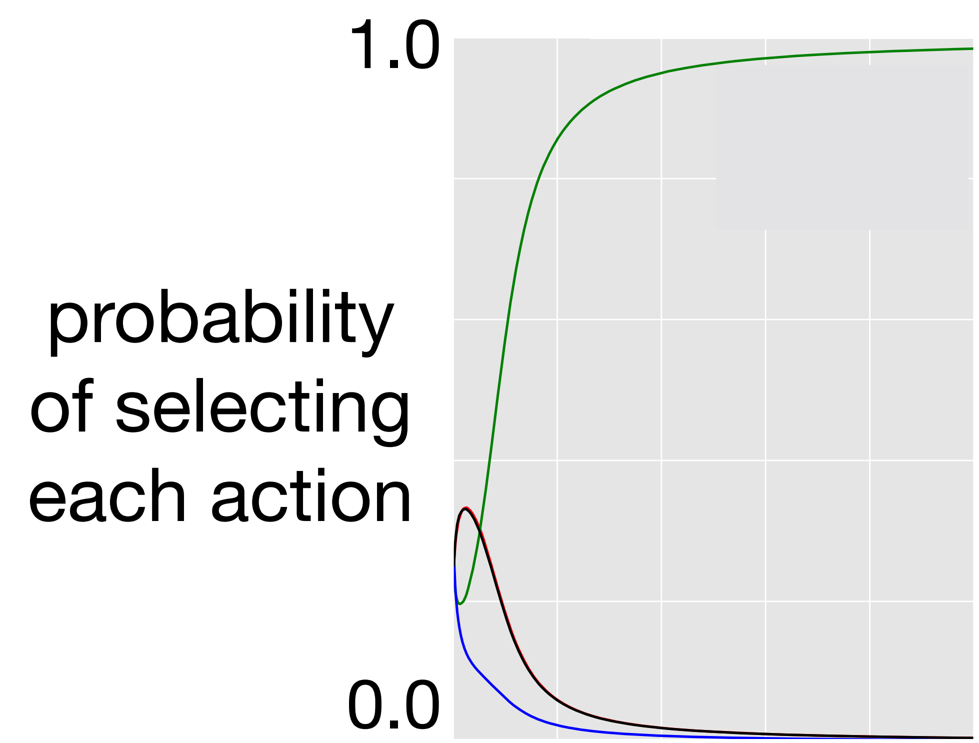
Examples of targets



target type	σ^2	ξ^2
constant	0	0
distractor	1	0
drifter	0	0.1

Example of good behavior





What **intrinsic rewards** give us this good behavior?

A survey of intrinsic rewards

- **Violated Expectations (surprise)**
    - Absolute Error, Squared Error, Expected Error (windowed average)
- Learning Progress (reduction in error)
    - Error Reduction, Error Derivative, Positive Error Part
- Amount of Learning (change in model)
    - Bayesian surprise, Weight Change, Variance of Prediction

Violated Expectations

Absolute Error* $|\delta_{t,i}|$
(Schmidhuber, 1991b)

Squared Error* $\delta_{t,i}^2$
(Gordon and Ahissar, 2011)

A survey of intrinsic rewards

- Violated Expectations (surprise)
    - Absolute Error, Squared Error, Expected Error (windowed average)
- **Learning Progress (reduction in error)**
    - Error Reduction, Error Derivative, Positive Error Part
- Amount of Learning (change in model)
    - Bayesian surprise, Weight Change, Variance of Prediction

Learning Progress

Error Reduction*
(Schmidhuber, 1991a)

$$|\delta_{t-1,i}| - |\delta_{t,i}|$$

Positive Error Part
(Mirolli and Baldassarre, 2013)

$$\max(\delta_{t,i}, 0)$$

Error Derivative
(Oudeyer et al., 2007)

$$\frac{1}{\eta}\sum_{j=0}^{\eta}\delta^2_{t-(j-\tau-\eta),i} - \frac{1}{\eta}\sum_{j=0}^{\eta}\delta^2_{t-(j-\eta),i}$$

A survey of intrinsic rewards

- Violated Expectations (surprise)
  - Absolute Error, Squared Error, Expected Error (windowed average)
- Learning Progress (reduction in error)
  - Error Reduction, Error Derivative, Positive Error Part
- **Amount of Learning (change in model)**
  - Bayesian surprise, Weight Change, Variance of Prediction

Amount of Learning

Weight Change*

$$\|w_{t,i} - w_{t-1,i}\|_1$$

Bayesian Surprise*
(Itti and Baldi, 2006)

$$\mathrm{KL}(p_{w_t} || p_{w_{t-1}})$$

where $p_{w_t}(\theta) = p_{w_{t-1}}(\theta | y_t) = \dfrac{p(y_t|\theta) p_{w_{t-1}}(\theta)}{p_{w_{t-1}}(y_t)}$

Which intrinsic reward is “best”? (And feasible)

- Bayesian surprise with Bayesian prediction learners in expectation equals Information Gain
  - Information Gain = Mutual Information between targets and parameters
- Maximize expected Bayesian surprise = maximize Info Gain
- Bayesian surprise with certain Bayesian or MAP learners corresponds to Weight Change with a MAP learner

Which intrinsic reward is “best”? (And feasible)

- Bayesian surprise with Bayesian prediction learners in expectation equals Information Gain
  - Information Gain = Mutual Information between targets and parameters
- Maximize expected Bayesian surprise = maximize Info Gain
- Bayesian surprise with certain Bayesian or MAP learners corresponds to Weight Change with a MAP learner

Potential Proposal: Weight Change for an approximate MAP learner

Let’s give this a name

Introspective prediction learners

increase their rate of learning when progress is possible

and decrease when it is not

An undesirable situation

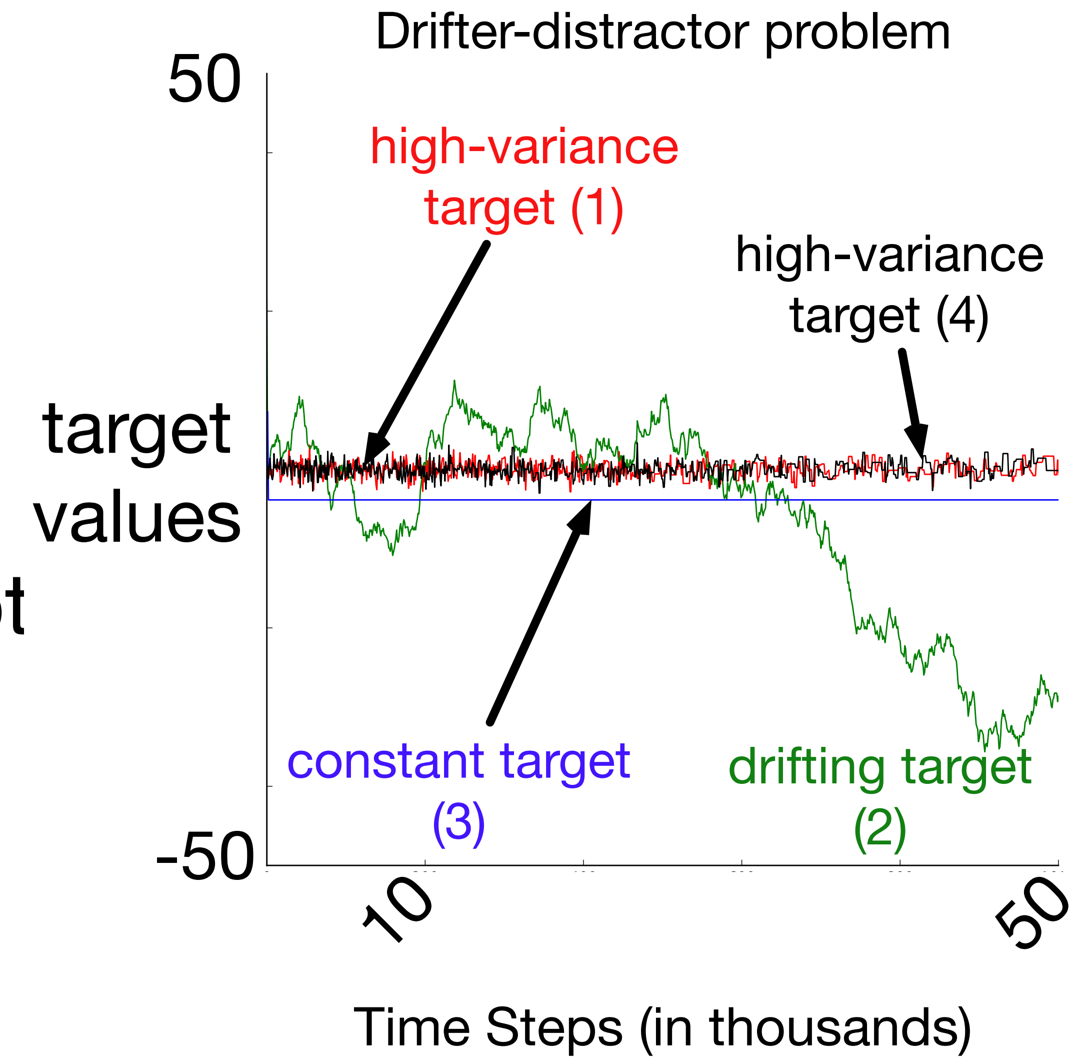
- Imagine an SGD learner with a fixed stepsize
    - a non-introspective learner
- Assume the target is drawn from a standard normal
- This learner chases noise
    - the weights will always change, because of stochasticity in targets
- Weight Change is not meaningful for such a learner

Can't the intrinsic reward account for bad learners?

- If a prediction learner is chasing noise, then intrinsic reward could simply be set to zero to stop wasted effort
- If this can be recognized to modify intrinsic reward, then the prediction learner can recognize it too
- Separation of responsibilities:
  - Prediction learner modulates learning
  - Behavior trusts prediction learners to modulate learning, focuses on exploring and balancing needs across learners

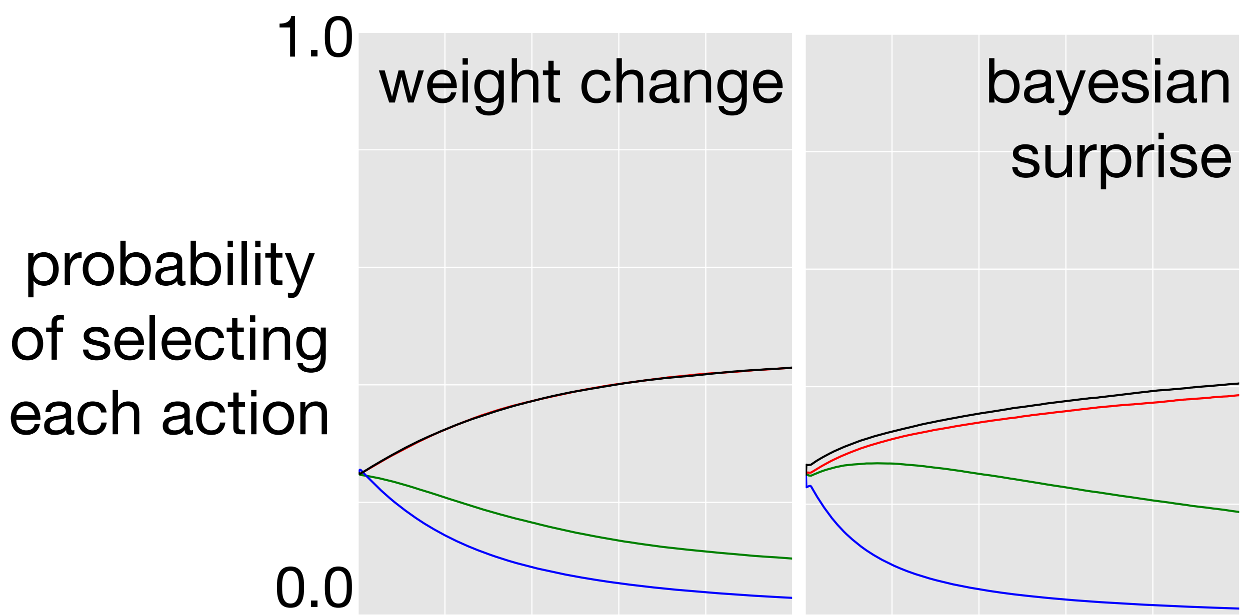
Experiment

- Compared 15 intrinsic rewards in Drifter-Distractor
- Behavior is a stochastic policy over 4 actions, learned with a Gradient Bandit algorithm
- Prediction learner uses
  - fixed stepsizes (non-introspective)
  - adaptive stepsizes (introspective)
- 200 runs, ~14000 parameters swept
- Performed other experiments with sudden changes, more targets

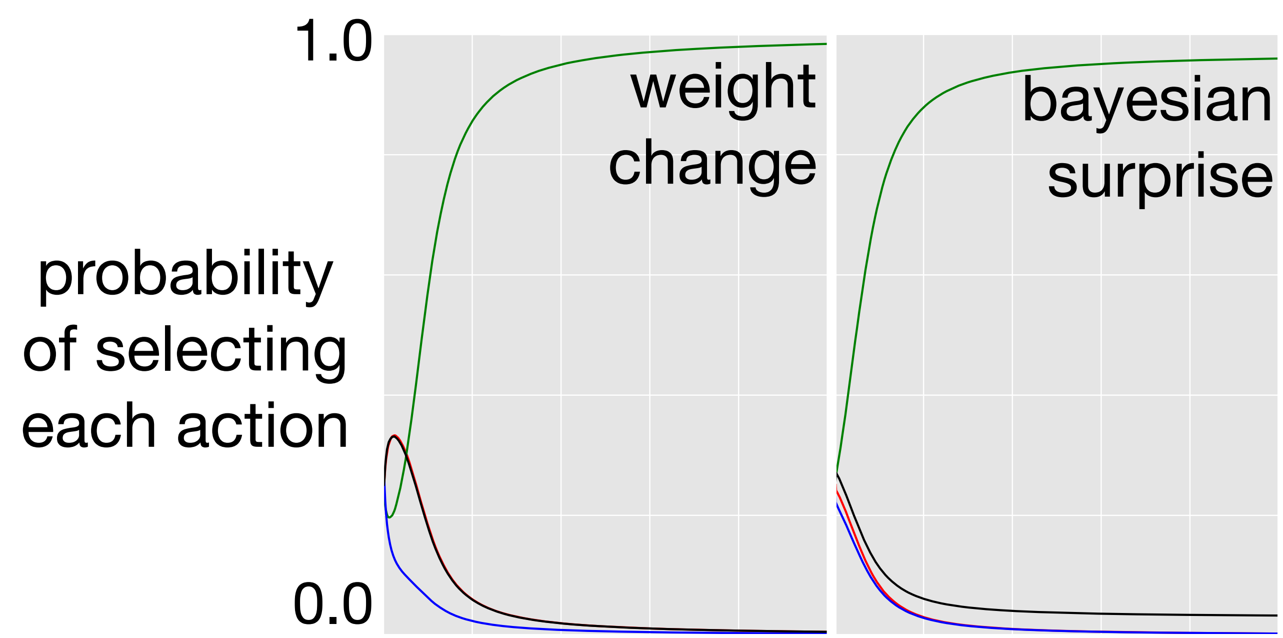


Contrasting learners

Constant
Drifting
Distractors



Non-introspective learners

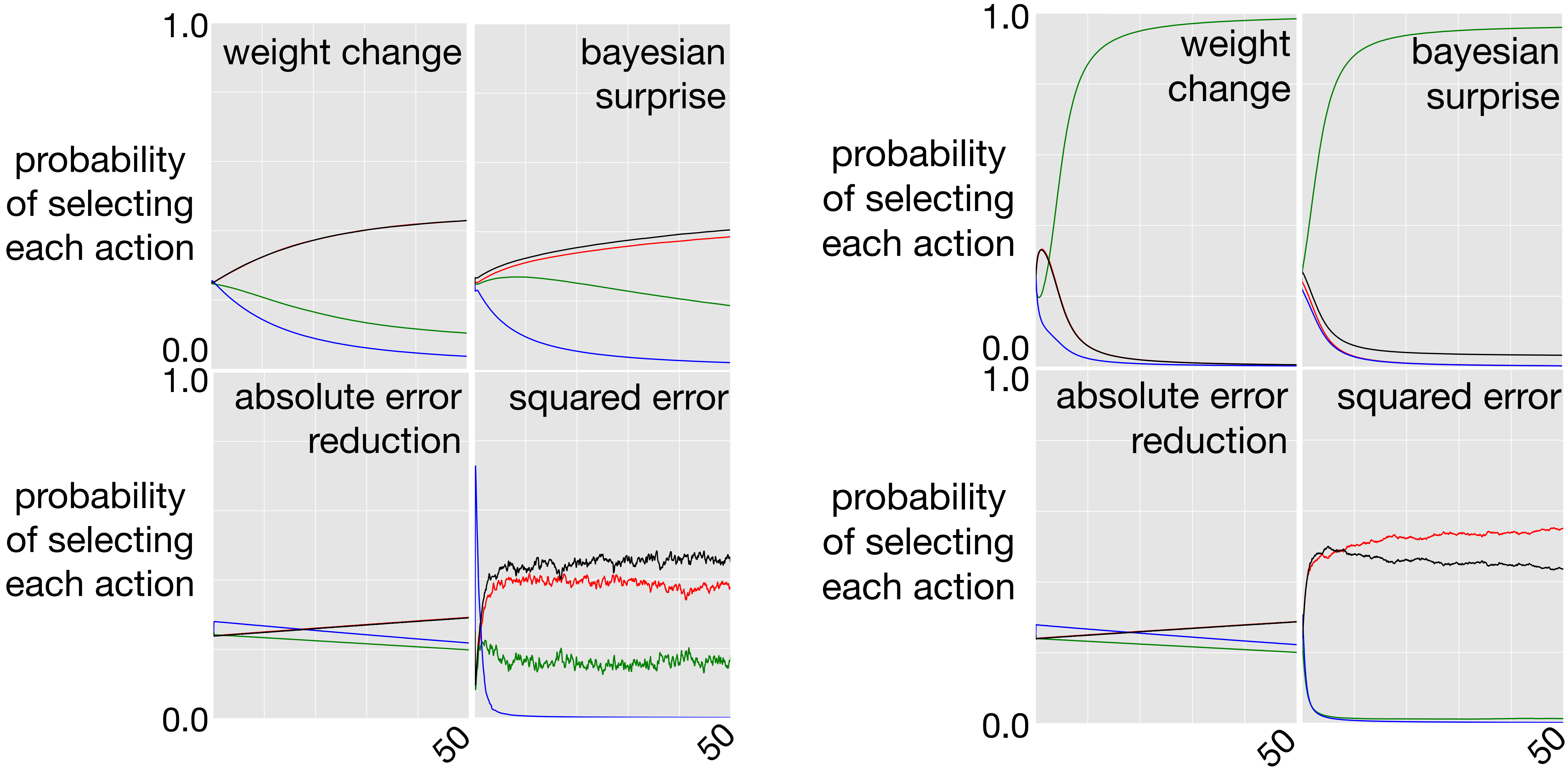


Introspective learners

Constant can be learned fast, should stop selecting
Distractors take longer (due to stochasticity), but also should stop being selected
Drifting needs to be selected forever (non-stationary)

Violated Expectations and Learning Progress do Poorly

Constant
Drifting
Distractors



Non-introspective learners

Introspective learners

Key take-away

Intrinsic rewards based on the **amount of learning**
can generate useful behaviour
if each individual learner is **introspective**.

Scaling to the general RL setting

- The strategies scale to the general RL setting
    - Weight change can easily be calculated for parameterized functions
    - Stepsize adaptation methods can still provide introspective learners
- There are some important differences
    - Learners would no longer be independent, as data for one can provide information for others
    - We will need to use off-policy methods
- It remains to be seen whether the conclusions scale

Open Questions

- What **other intrinsic rewards** could we consider?
- Can we **prove** that maximizing expected intrinsic reward provides guarantees on prediction accuracy?
- **What** behaviour do these intrinsic rewards induce in MDPs?
- Is this an **easier exploration** problem, than maximizing (sparse) environment rewards?
- Do we need **smarter exploration** strategies in MDPs?

Open Questions

- What **other intrinsic rewards** could we consider?
- Can we **prove** that maximizing expected intrinsic reward provides guarantees on prediction accuracy?
- **What** behaviour do these intrinsic rewards induce in MDPs?
- Is this an **easier exploration** problem, than maximizing (sparse) environment rewards?
- Do we need **smarter exploration** strategies in MDPs?

Thank you!