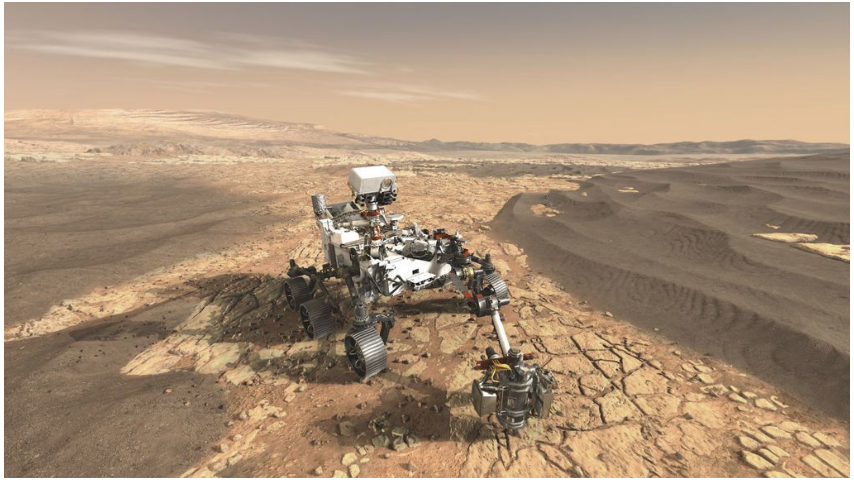
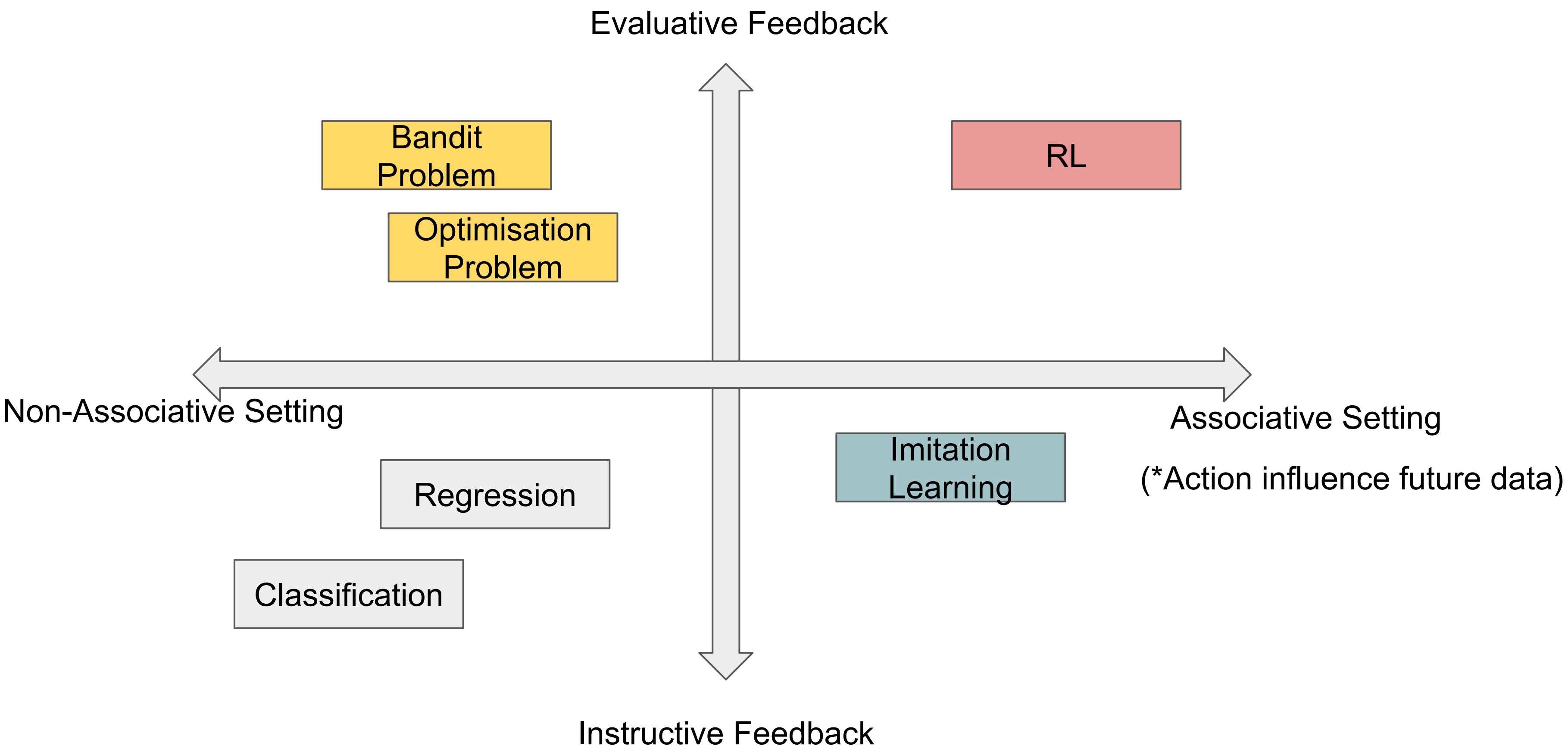
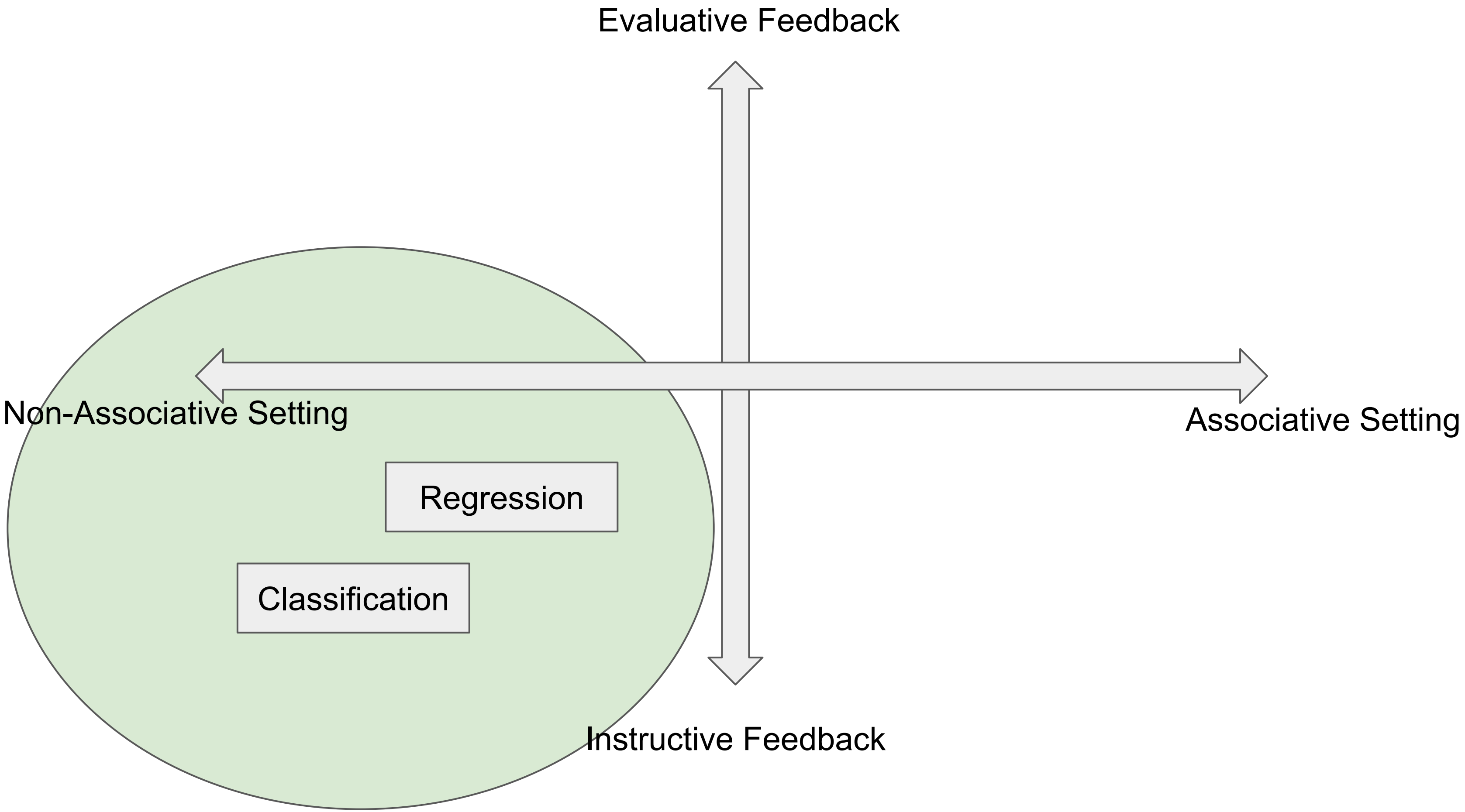
Exploration in RL



Outline today




Evaluative Feedback
Non-Associative Setting
Associative Setting
Regression
Classification
Instructive Feedback

1. Exploration in Non-Associative & Instructive Feedback Setting (AKA Active Learning)

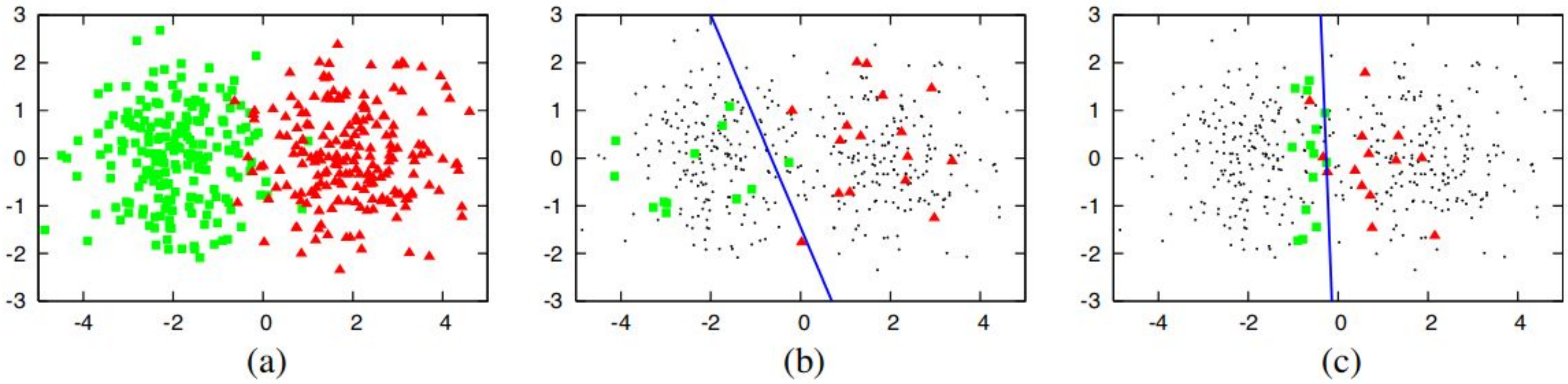


Image taken from
http://burrsettles.com/pub/settles.activelearning.pdf

1. Exploration in Non-Associative & Instructive Feedback Setting (AKA Active Learning)

- We want ``good`` model … but what does it mean?

1. Exploration in Non-Associative & Instructive Feedback Setting (AKA Active Learning)

- We want ``good`` model … but what does it mean?

1. It has to explain observed data

2. It has to generalise to unseen data

1. Exploration in Non-Associative & Instructive Feedback Setting (AKA Active Learning)

- Principles :

1. Choose data points that would eliminate as many model hypotheses as possible. (Better explain data)

2. Choose data points that is representative of the class. (Improve generalisation)

1. Exploration in Non-Associative & Instructive Feedback Setting (AKA Active Learning)

- Principles :

1. Choose data points that would eliminate as many model hypotheses as possible. (Better explain data)

2. Choose data points that is representative of the class. (Improve generalisation)

1. Exploration in Non-Associative & Instructive Feedback Setting (AKA Active Learning)

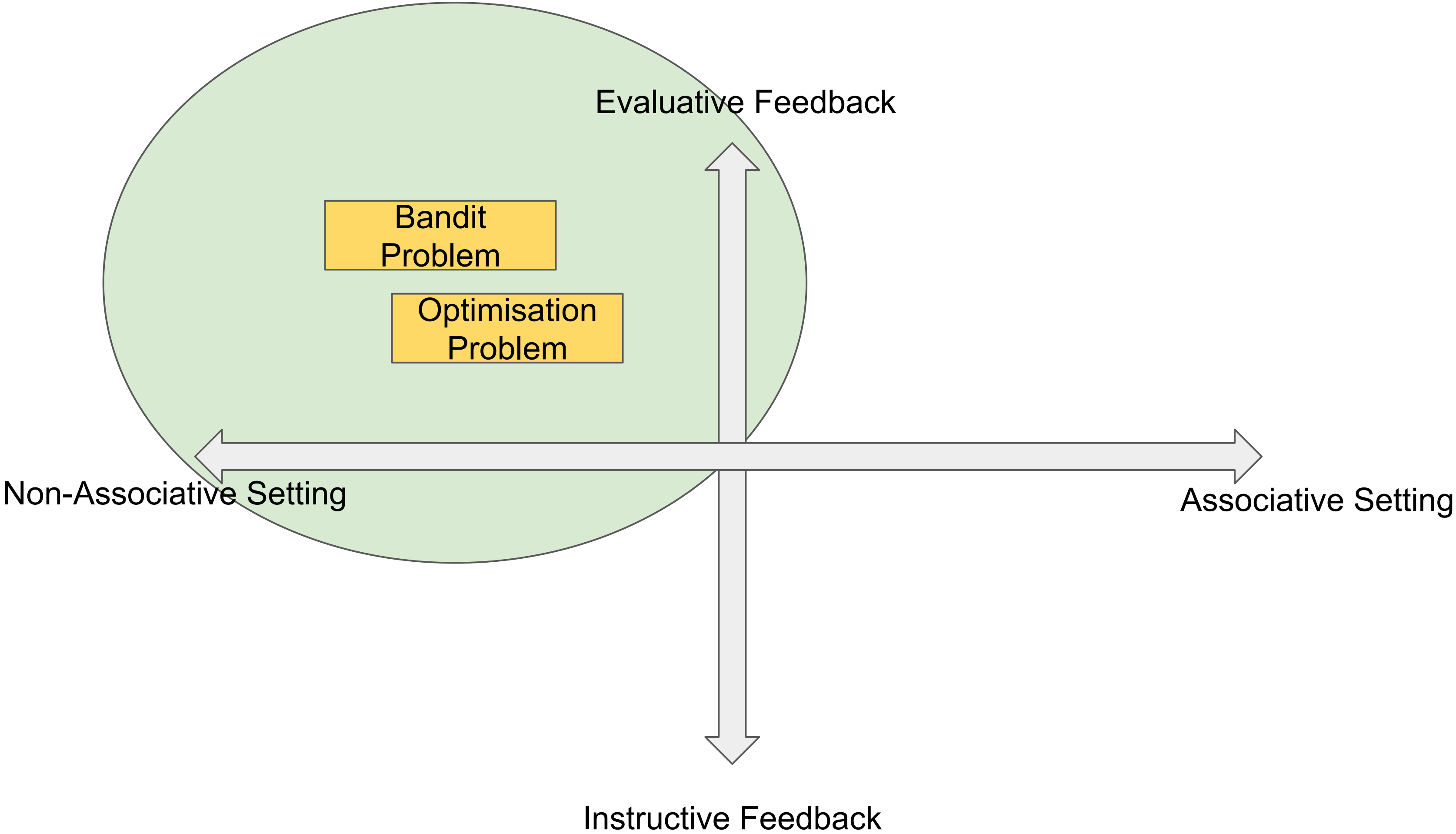
- 1.1 Uncertainty Sampling

1. Exploration in Non-Associative & Instructive Feedback Setting (AKA Active Learning)

- 1.2 Representative Sampling

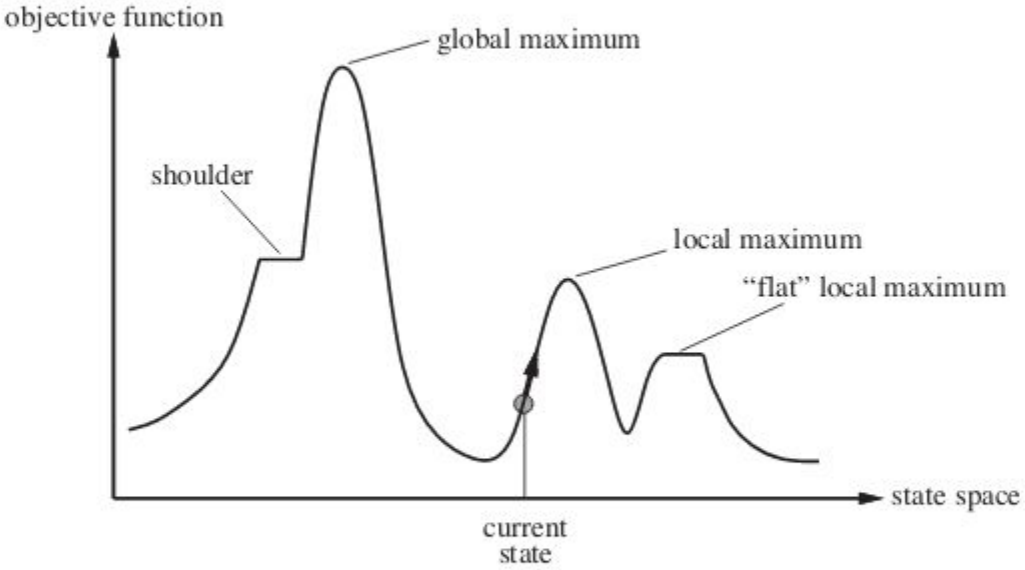
1. Exploration in Non-Associative & Instructive Feedback Setting (AKA Active Learning)

- 1.3 Diversification


Evaluative Feedback
Bandit Problem
Optimisation Problem
Non-Associative Setting
Associative Setting
Instructive Feedback

2. Evaluative Feedback & Non-associative Setting

2.1 Optimisation Problem

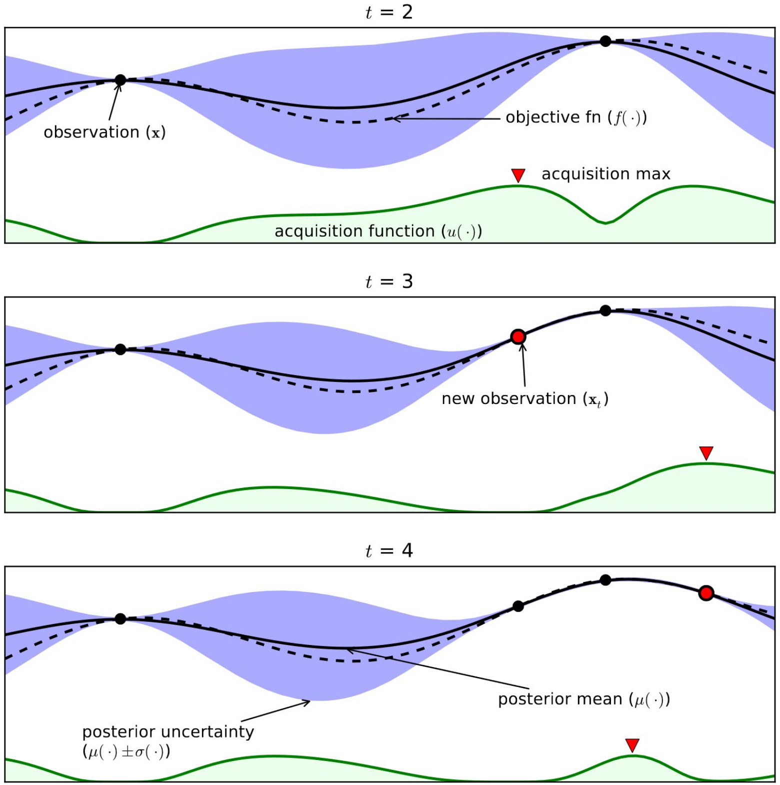


2. Evaluative Feedback & Non-associative Setting

2.1 Optimisation Problem

General idea:

1. Evaluate the posterior measure given data
2. Compute Acquisition function
3. Select the point at the maximum of acquisition function



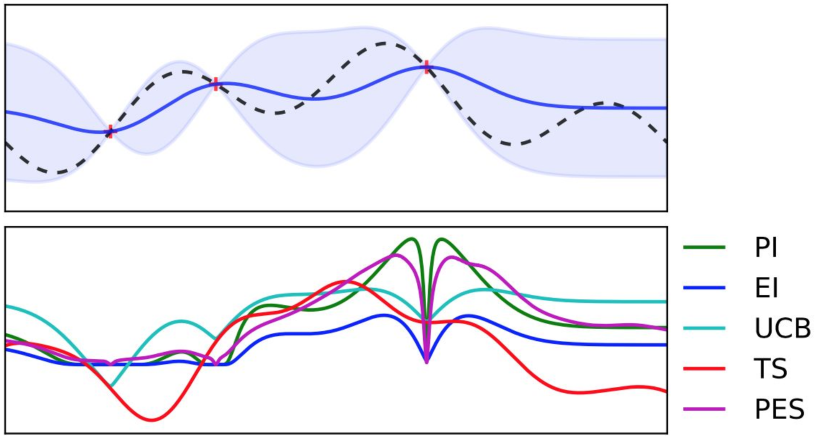
https://www.cs.ox.ac.uk/people/nando.defreitas/publications/BayesOptLoop.pdf

2. Evaluative Feedback & Non-associative Setting

2.1 Optimisation Problem

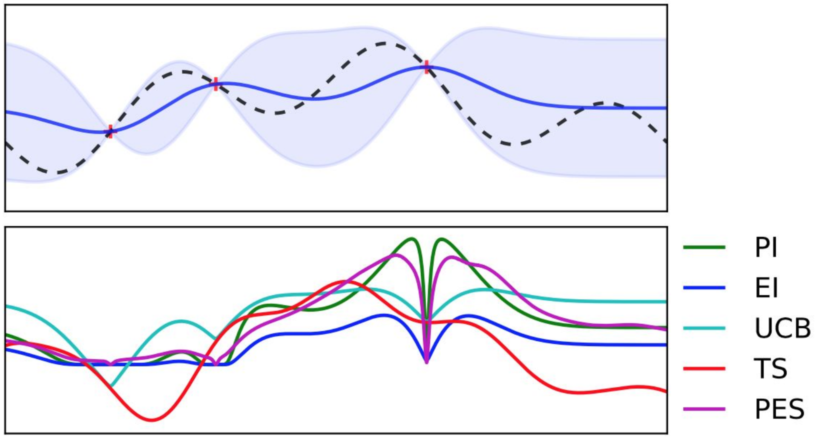
Acquisition Function?

1. Probability of Improvement (PI)
2. Expected Improvement (EI)
3. Upper Confidence Bound (UCB)
4. Thompson Sampling (TS)
5. Predictive Entropy Search (PES)



Acquisition Function?

1. Probability of Improvement (PI)
2. Expected Improvement (EI)
3. Upper Confidence Bound (UCB)
4. Thompson Sampling (TS)
5. Predictive Entropy Search (PES)

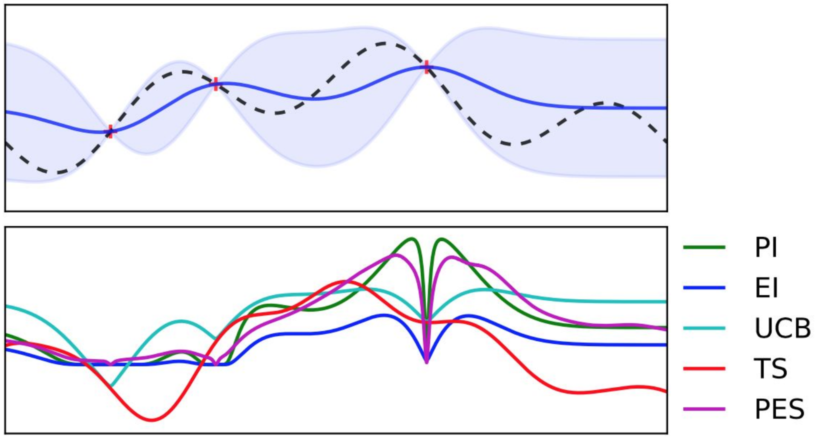


2. Evaluative Feedback & Non-associative Setting

2.1 Optimisation Problem

Acquisition Function?

1. Probability of Improvement (PI)
2. Expected Improvement (EI)
3. Upper Confidence Bound (UCB)
4. **Thompson Sampling (TS)**
5. Predictive Entropy Search (PES)

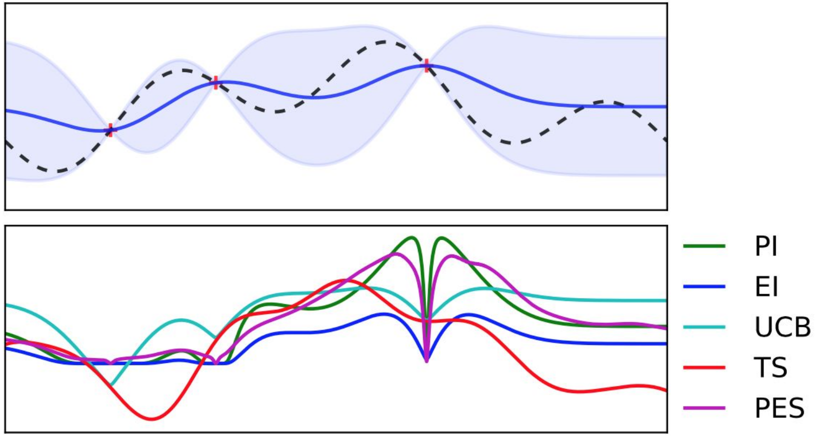


2. Evaluative Feedback & Non-associative Setting

2.1 Optimisation Problem

Acquisition Function?

1. Probability of Improvement (PI)
2. Expected Improvement (EI)
3. Upper Confidence Bound (UCB)
4. Thompson Sampling (TS)
5. Predictive Entropy Search (PES)



2. Evaluative Feedback & Non-associative Setting

2.2 Bandit Problem

$\mathcal{A} = \{\text{pull arm}\}$

$r(\text{pull arm}) = ?$

$\mathcal{A} = \{\text{pull}_1, \text{pull}_2, \dots, \text{pull}_n\}$

$r(a_n) = ?$

assume $r(a_n) \sim p(r|a_n)$

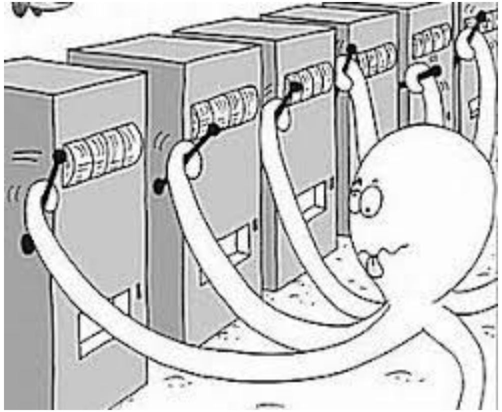
unknown *per-action* reward distribution!

From Sergey Levine’s slide
http://rail.eecs.berkeley.edu/deeprlcourse-fa17/f17docs/lecture_13_exploration.pdf

2. Evaluative Feedback & Non-associative Setting

2.2 Bandit Problem = Stochastic evaluative feedback + online learning

http://iosband.github.io/2015/07/28/Beat-the-bandit.html



In this simple setting, we have many provably \`optimal algorithms'. Although, the empirical performance may vary.

From Sergey Levine's slide
http://rail.eecs.berkeley.edu/deeprlcourse-fa17/f17docs/lecture_13_exploration.pdf

2. Evaluative Feedback & Non-associative Setting

2.2.1 Optimistic Exploration

$$a = \arg\max \hat{\mu}_a + C\underbrace{\sigma_a}_{\text{some sort of variance estimate}}$$

Example: UCB

$$a = \arg\max \hat{\mu}_a + \sqrt{\frac{2\ln T}{N(a)}}$$

Number of time we picked this action :
Higher count = more certain
about the outcome

From Sergey Levine’s slide
http://rail.eecs.berkeley.edu/deeprlcourse-fa17/f17docs/lecture_13_exploration.pdf

2. Evaluative Feedback & Non-associative Setting

2.2.2 Thompson Sampling

assume $r(a_i) \sim p_{\theta_i}(r_i)$

idea: sample $\theta_1, \ldots, \theta_n \sim \hat{p}(\theta_1, \ldots, \theta_n)$

pretend the model $\theta_1, \ldots, \theta_n$ is correct

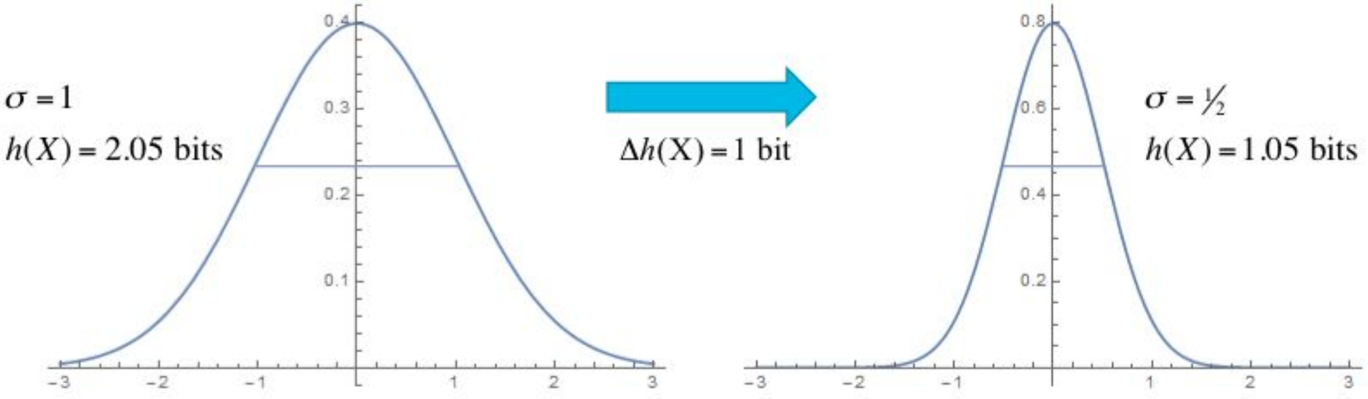
take the optimal action

update the model

2. Evaluative Feedback & Non-associative Setting

2.2.3 Information Gain

- Use entropy measure as a measure of information
- Estimate entropy after observation
- We gain more information if the entropy reduces more!

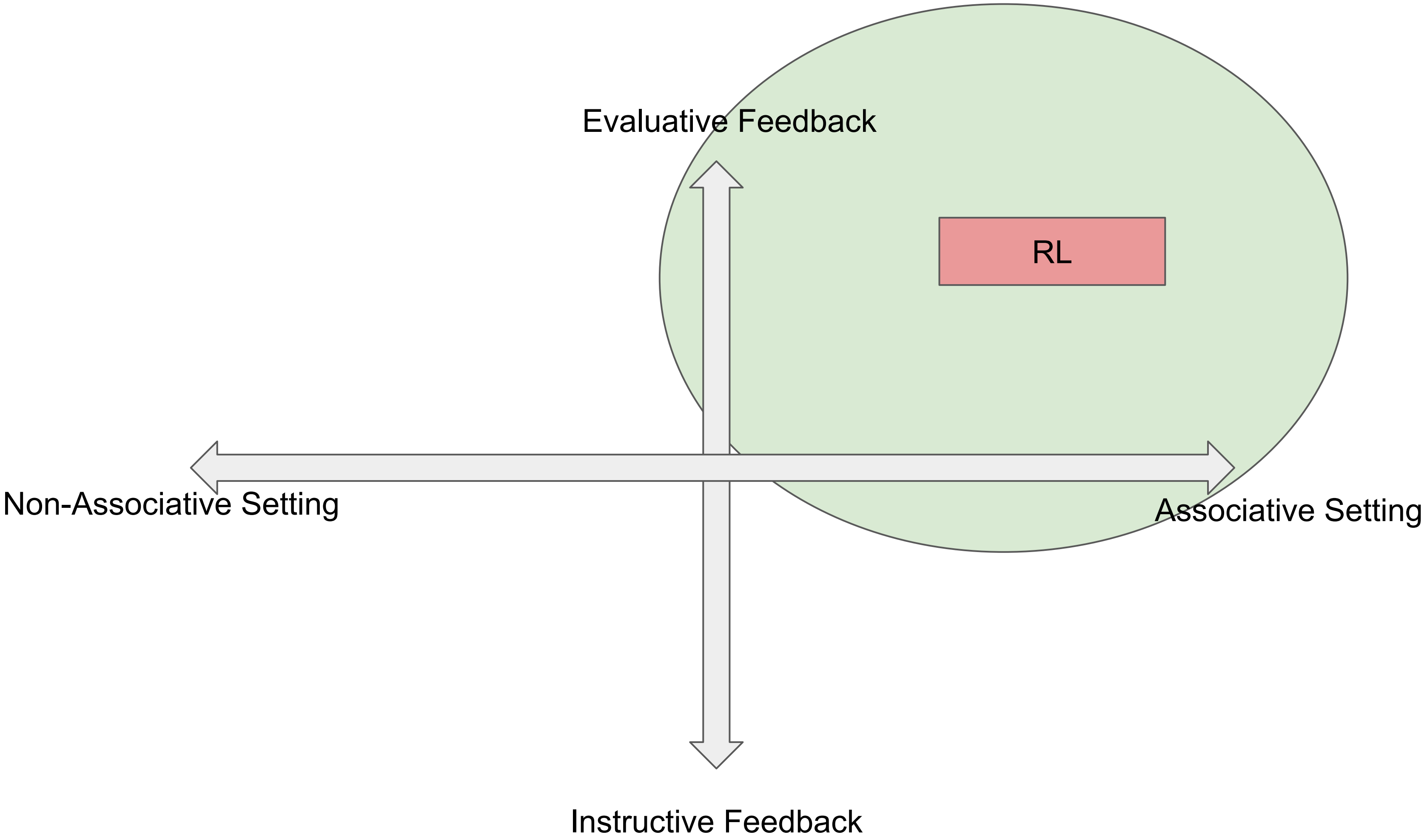


2. Evaluative Feedback & Non-associative Setting

General Principle

1. Require some kind of uncertainty estimation
2. Assume some value to new information

Would these ideas work in RL setting?


Evaluative Feedback
RL
Non-Associative Setting
Associative Setting
Instructive Feedback

Exploration in RL (outline)

1. Naive random exploration + Structured random exploration
2. Optimistic exploration
3. Thompson sampling style
4. Information gain style
5. Intrinsic Motivation style

An exploratory action can affect the future exploratory states !!

Naive Random Exploration

- E-Greedy

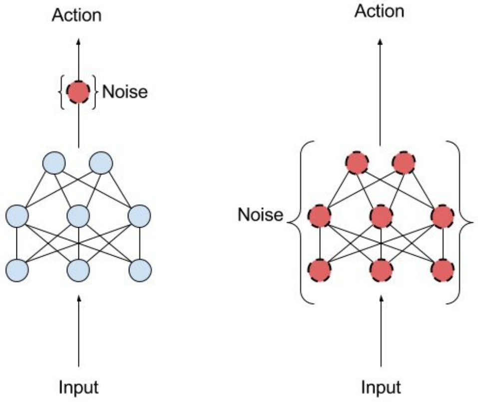
- Boltzmann / Softmax exploration

$$P(a) = \frac{e^{f(a)\cdot\theta^{-1}}}{\sum_{a' \in \text{Actions}} e^{f(a')\cdot\theta^{-1}}}$$

- Entropy Bonus

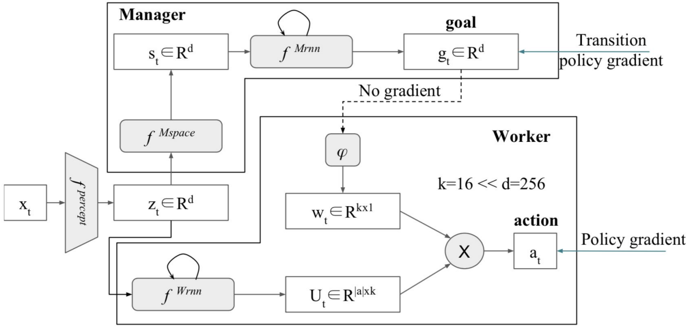
Structured Random Exploration

- **Inject noise at the policy level**



See: https://openai.com/blog/better-exploration-with-parameter-noise

- **Hierarchical RL with exploration on meta-controller**



Veznevets et al. 2017, https://arxiv.org/abs/1703.01161

Optimistic Exploration

- Pseudo-Count (Bellemare et al. 2016; https://arxiv.org/pdf/1606.01868.pdf)
  - Add reward bonus: $$R_n^+(x, a) := \beta(\hat{N}_n(x) + 0.01)^{-1/2}$$
  - Pseudo-count can be estimated with density model:

Optimistic Exploration

- Pseudo-Count (Bellemare et al. 2016; https://arxiv.org/pdf/1606.01868.pdf)

$$P(\mathbf{s}) = \frac{N(\mathbf{s})}{n}$$

count

total states visited

probability/density

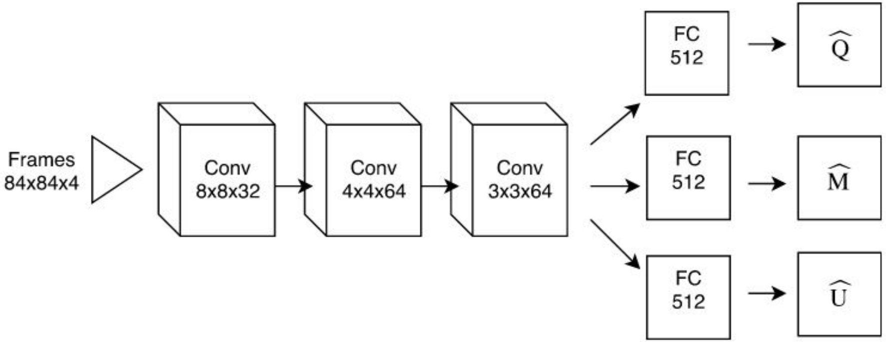
$$P'(\mathbf{s}) = \frac{N(\mathbf{s}) + 1}{n + 1}$$

$$\hat{n} = \frac{1 - p_{\theta'}(\mathbf{s}_i)}{p_{\theta'}(\mathbf{s}_i) - p_{\theta}(\mathbf{s}_i)} p_{\theta}(\mathbf{s}_i)$$

Optimistic Exploration

- Risk-Seeking Exploration (Dilokthanakul and Shanahan 2018)
    - Use variance of return as reward bonus

$$\mathrm{Var}[G] = \mathbb{E}[G^2] - \mathbb{E}[G]^2$$



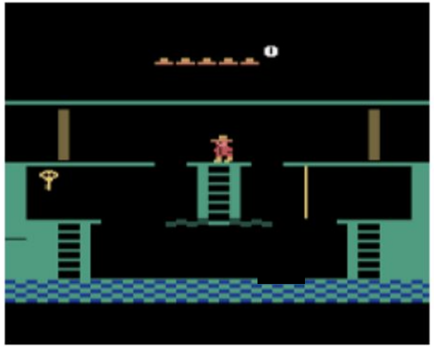
Optimistic Exploration

- Risk-Seeking Exploration (Dilokthanakul and Shanahan 2018)
  - Two types of uncertainty

    1. Epistemic Uncertainty

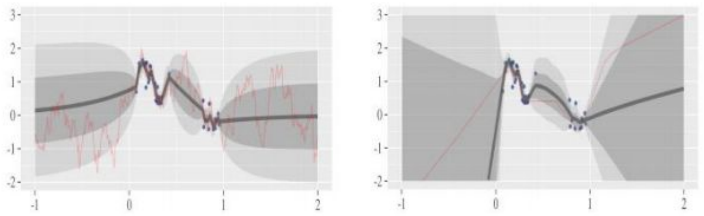
    2. Inherent Uncertainty

State Aliasing Effect

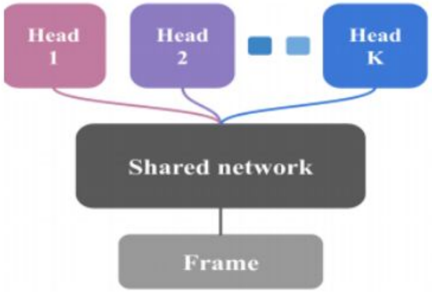


Thompson Sampling Style

- Bootstrap DQN (Osband et al. 2016)



(b) Gaussian process posterior (c) Bootstrapped neural nets



Information Gain Style

- VIME (Houthooft et al. 2016)

IG can be equivalently written as $D_{\mathrm{KL}}(p(\theta|h, \underline{s_t, a_t, s_{t+1}})\|p(\theta|h))$

model parameters for $p_\theta(s_{t+1}|s_t, a_t)$

newly observed transition

history of all prior transitions

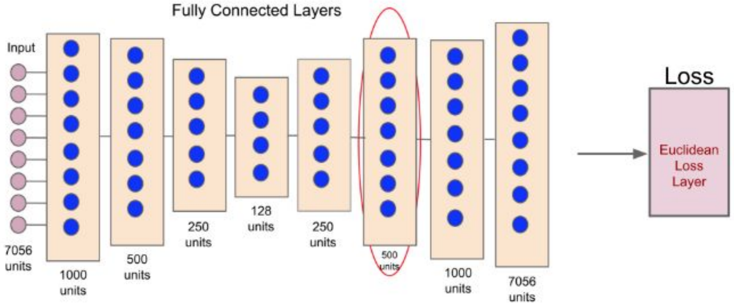
Amount of information gained in model after an observation!

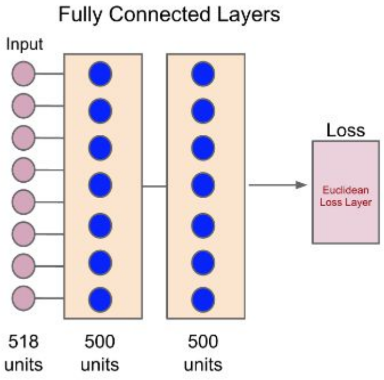
Information Gain Style

- Use model error as proxy (Stadie et al. 2015; https://arxiv.org/abs/1507.00814)

Build a forward predictive model

- If the predictive model predict wrongly then the state contain more information





Intrinsic motivation exploration

- Empowerment Exploration (Shakir Mohamed and Danilo Rezende, 2015)

$$\mathcal{E}(\mathbf{s}) \quad = \quad \max_{\omega} \mathcal{I}^{\omega}(\mathbf{a}, \mathbf{s}'|\mathbf{s}) = \max_{\omega} \mathbb{E}_{p(s'|a,s)\omega(a|s)} \left[\log \left(\frac{p(\mathbf{a}, \mathbf{s}'|\mathbf{s})}{\omega(\mathbf{a}|\mathbf{s}) p(\mathbf{s}'|\mathbf{s})} \right) \right],$$

Intuition 1: We want to go to the state with maximum influence!

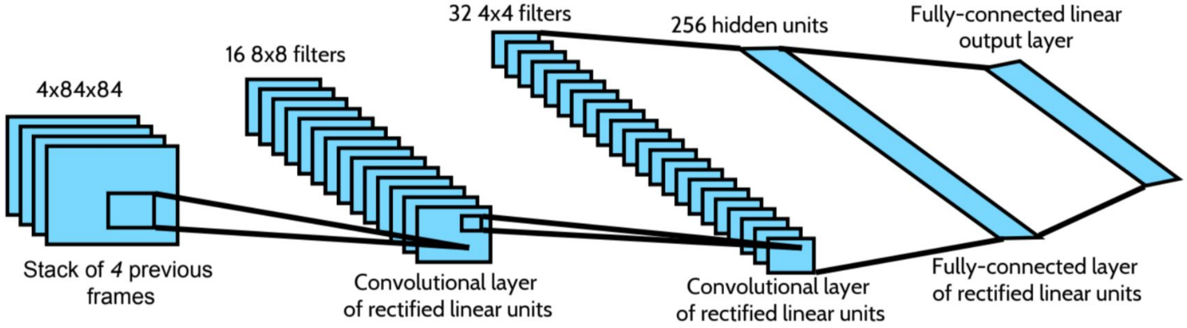
Intuition 2: We want our action to really affect the outcome!

Intrinsic motivation exploration

- Feature-Control as Intrinsic Motivation (Dilokthanakul et al. 2019)

Main idea: Ability to control aspects of the environment is a good skill to have!

Pixel-Control



Feature-Control