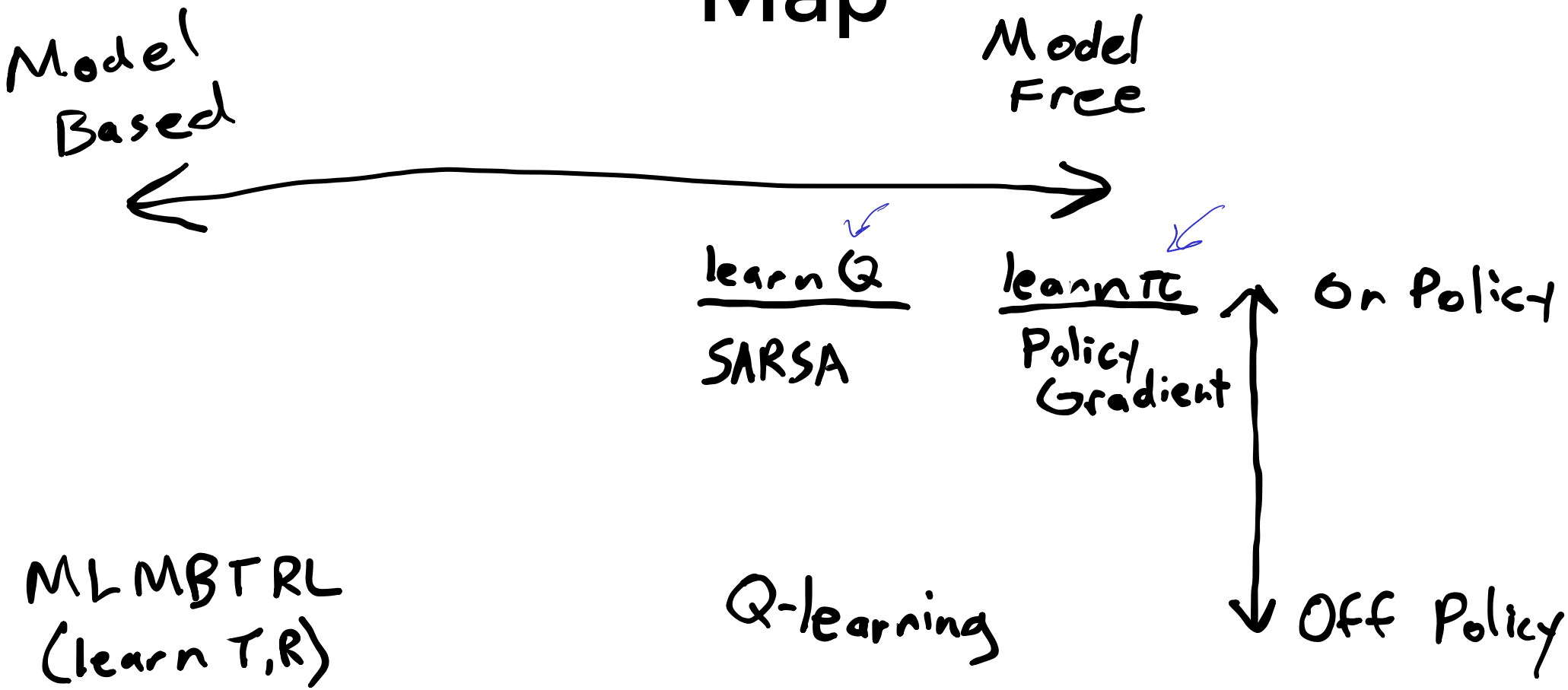
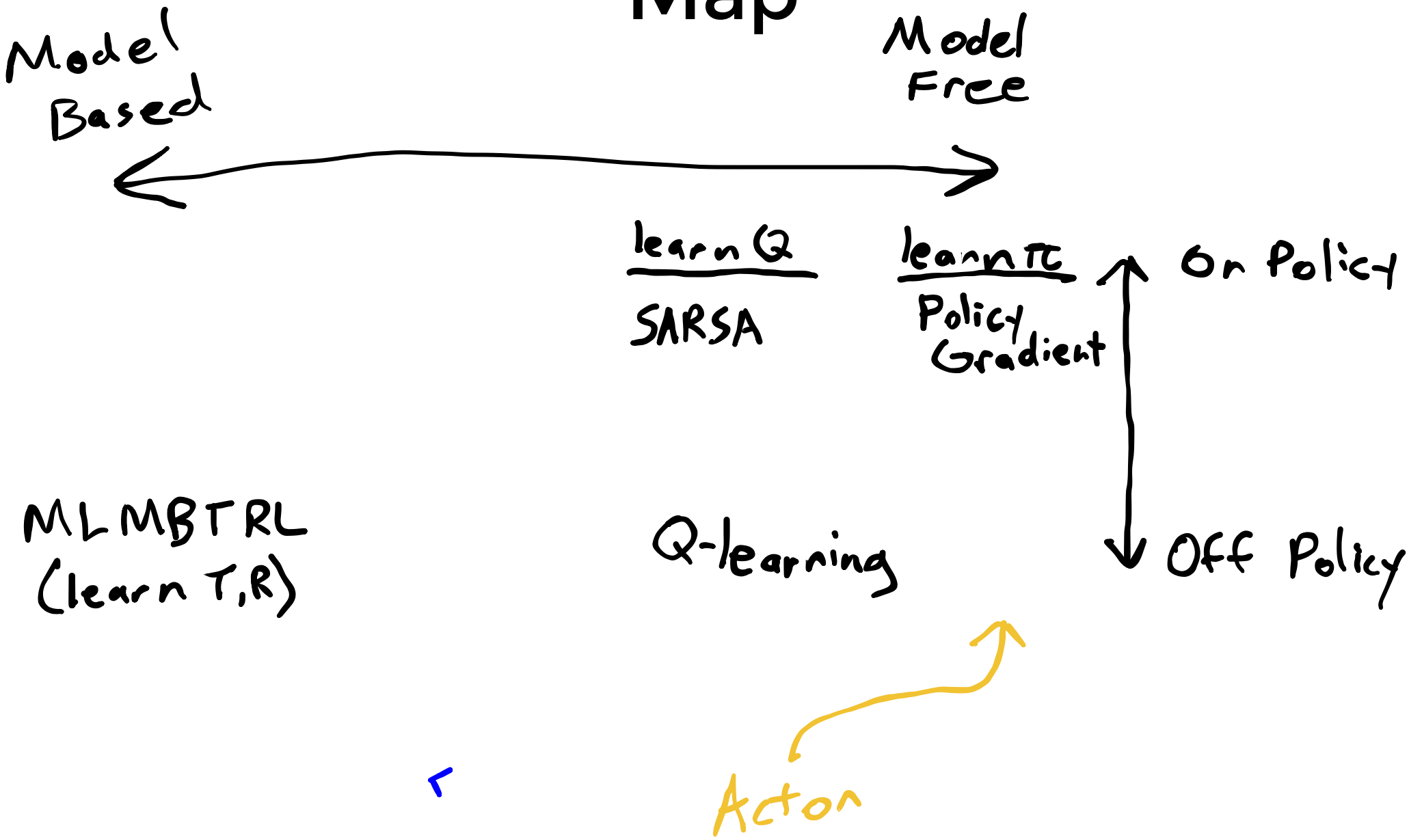
Map

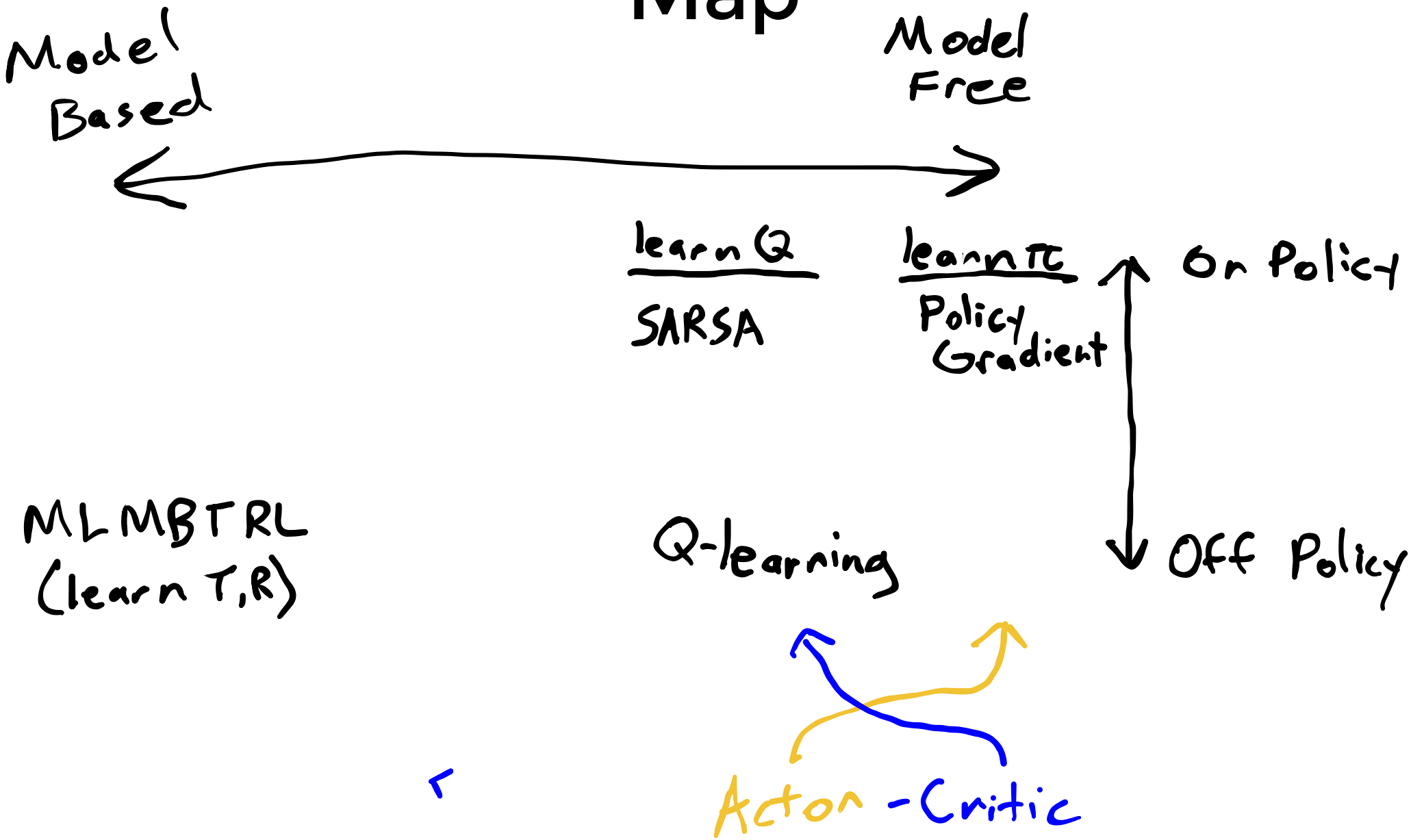
Map


Model Based
Model Free
learn Q
learn π
SARSA
Policy Gradient
On Policy
Off Policy
ML MBTRL (learn T,R)
Q-learning

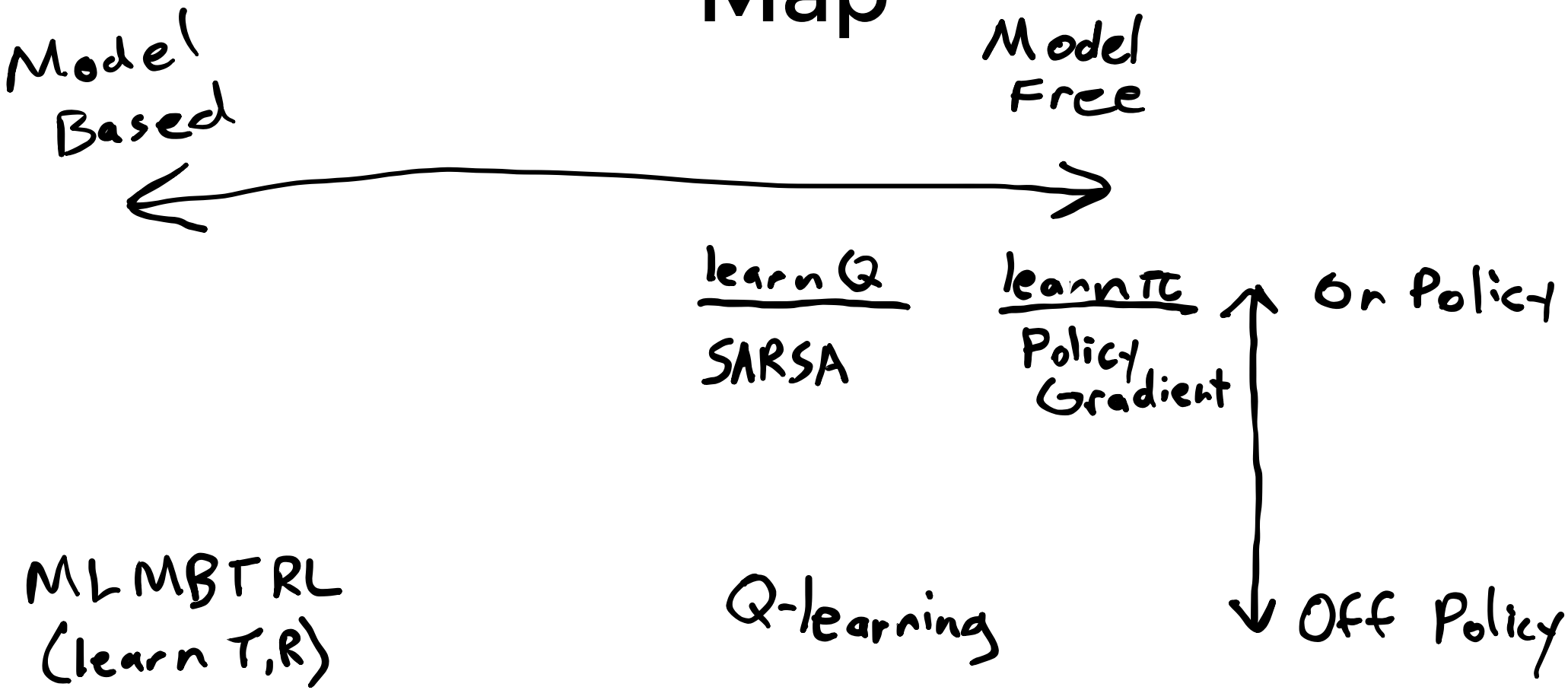
Map


Model Based
Model Free
learn Q
SARSA
learn π
Policy Gradient
On Policy
Off Policy
ML MBTRL
(learn T,R)
Q-learning
Actor

Map


Model Based
Model Free
learn Q
SARSA
learn π
Policy Gradient
On Policy
Off Policy
ML MBTRL
(learn T,R)
Q-learning
Actor-Critic

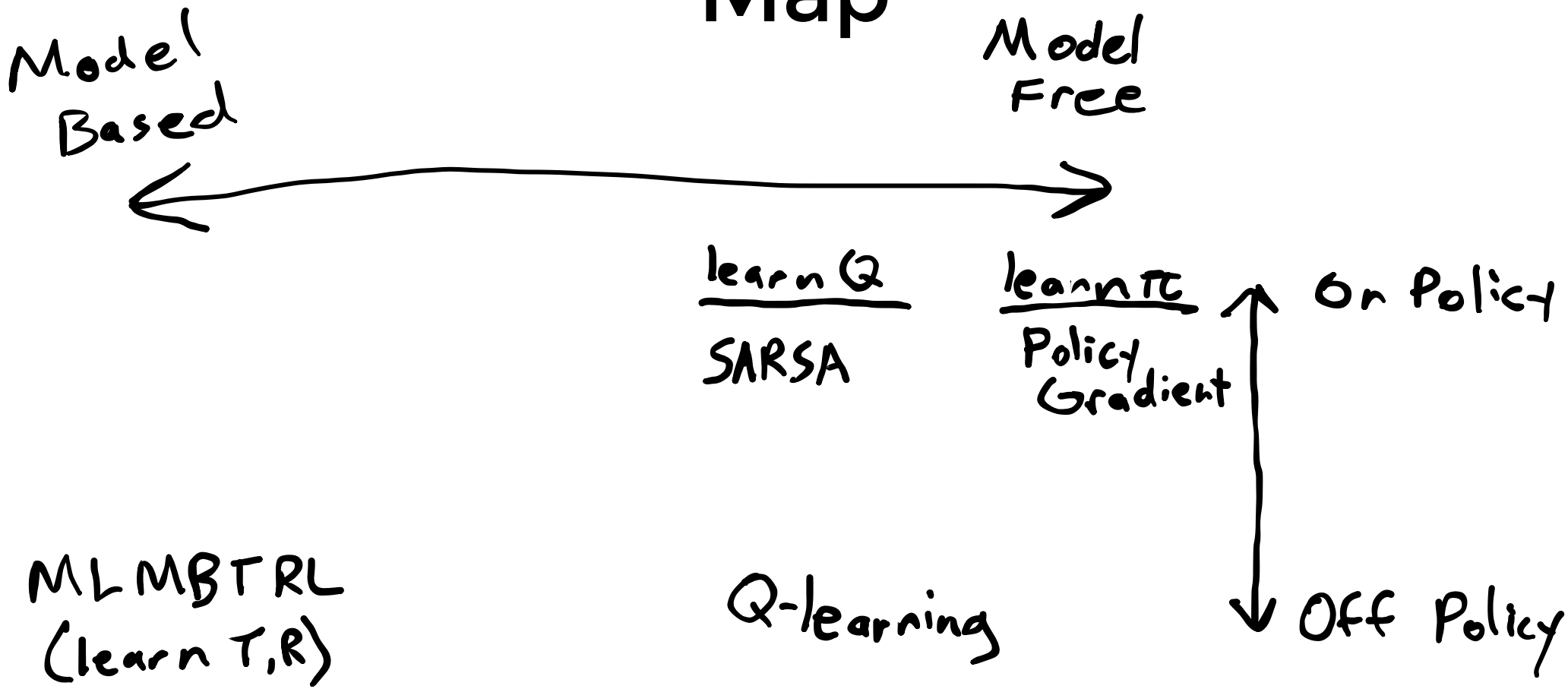
Map



Challenges:

1. Exploration vs Exploitation
2. Credit Assignment
3. Generalization

Map



Challenges:

1. Exploration vs Exploitation ←
2. Credit Assignment
3. Generalization

Is Exploration Important? Montezuma's Revenge

Recognize + track

UCB

Generalization?

Is Exploration Important? Theory

Exploration Bonus

Exploration Bonus

- In General, $R^+(s,a) = R(s,a) + B(s,a)$
- UCB: $B(s,a) = c\sqrt{\frac{\log N(s)}{N(s,a)}}$

Exploration Bonus

Example 1: Learn Pseudocount

Bellemare, et al. 2016 "Unifying Count-Based Exploration..."

Exploration Bonus

Example 1: Learn Pseudocount

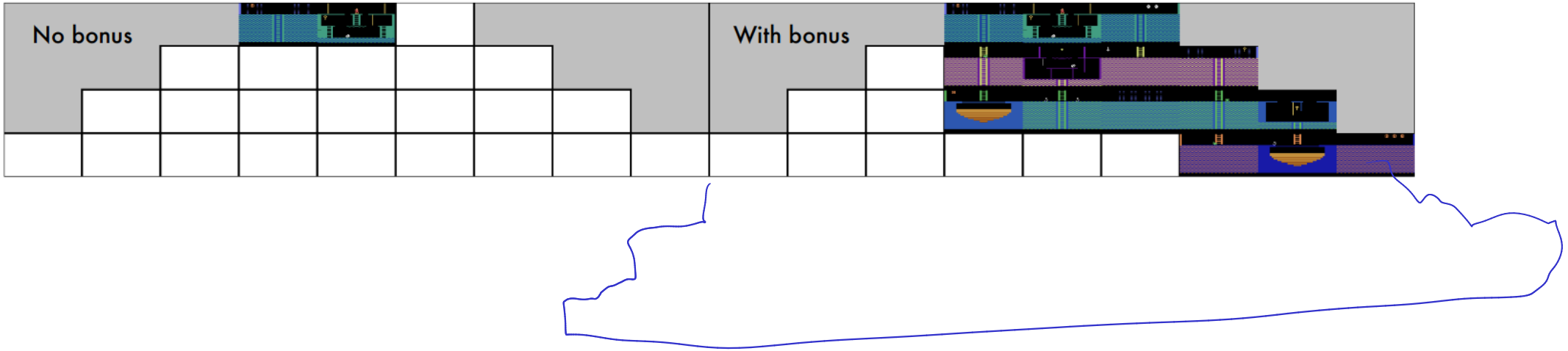
$B(s,a) \approx \frac{1}{\sqrt{\hat{N}(s)}}$ where $\hat{N}(s)$ is a learned function approximation

Bellemare, et al. 2016 "Unifying Count-Based Exploration..."

Exploration Bonus

Example 1: Learn Pseudocount

$B(s,a) \approx \frac{1}{\sqrt{\hat{N}(s)}}$ where $\hat{N}(s)$ is a learned function approximation

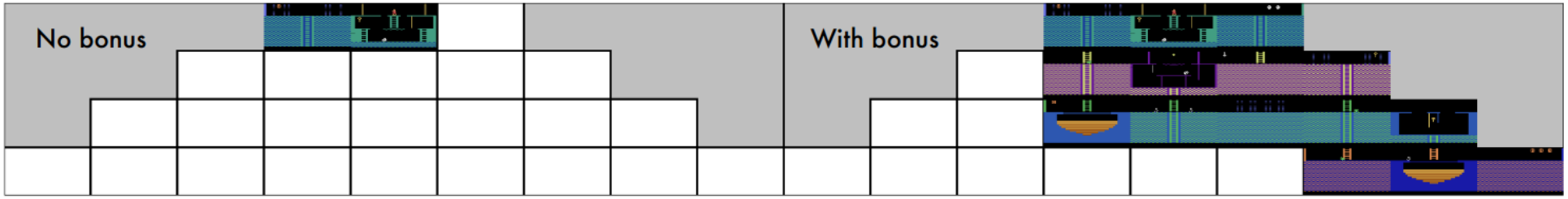


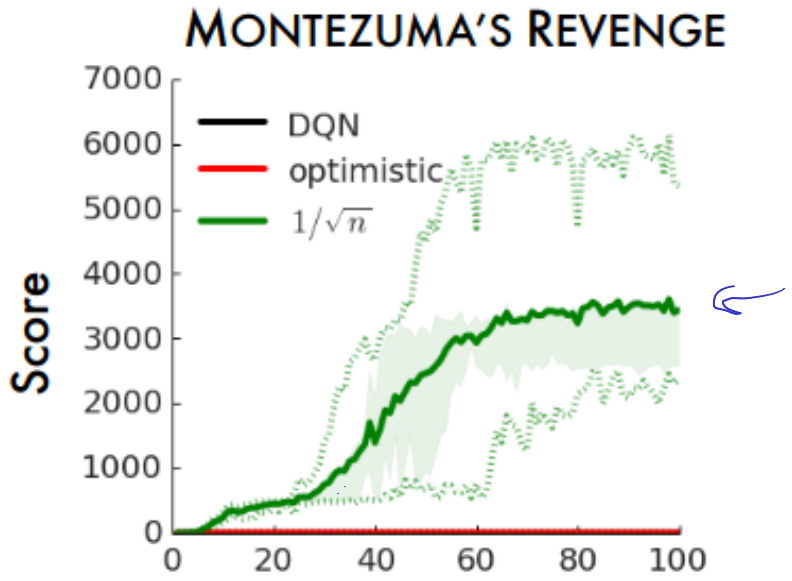
Bellemare, et al. 2016 "Unifying Count-Based Exploration..."

Exploration Bonus

Example 1: Learn Pseudocount

$B(s,a) \approx \frac{1}{\sqrt{\hat{N}(s)}}$ where $\hat{N}(s)$ is a learned function approximation



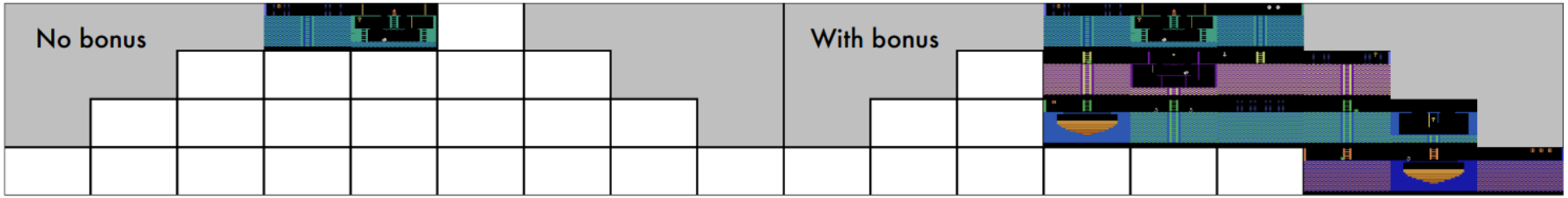


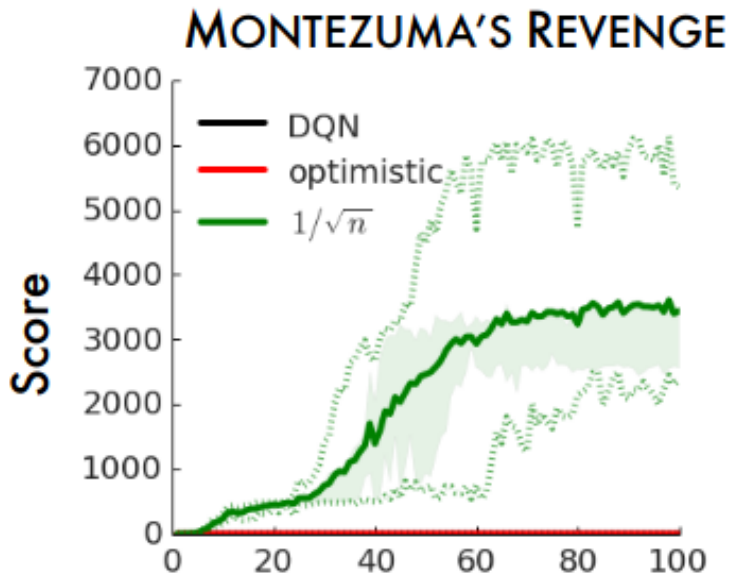
Bellemare, et al. 2016 "Unifying Count-Based Exploration..."

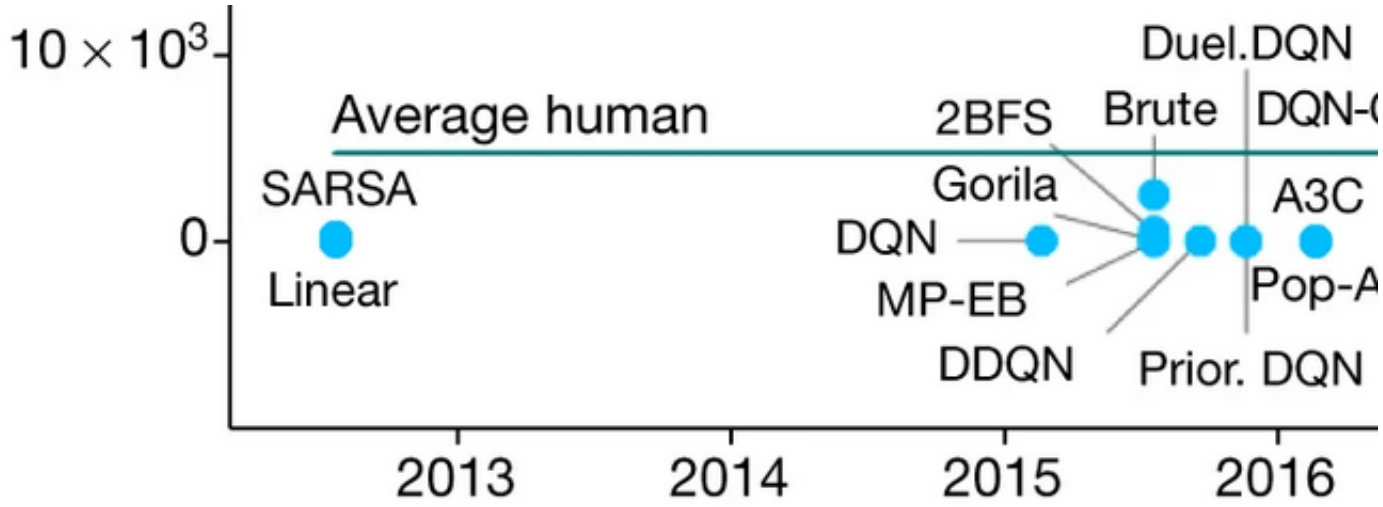
Exploration Bonus

Example 1: Learn Pseudocount

$B(s,a) \approx \frac{1}{\sqrt{\hat{N}(s)}}$ where $\hat{N}(s)$ is a learned function approximation





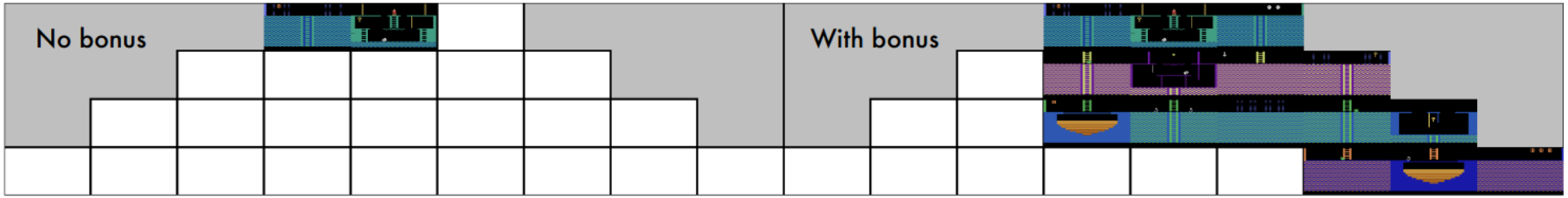


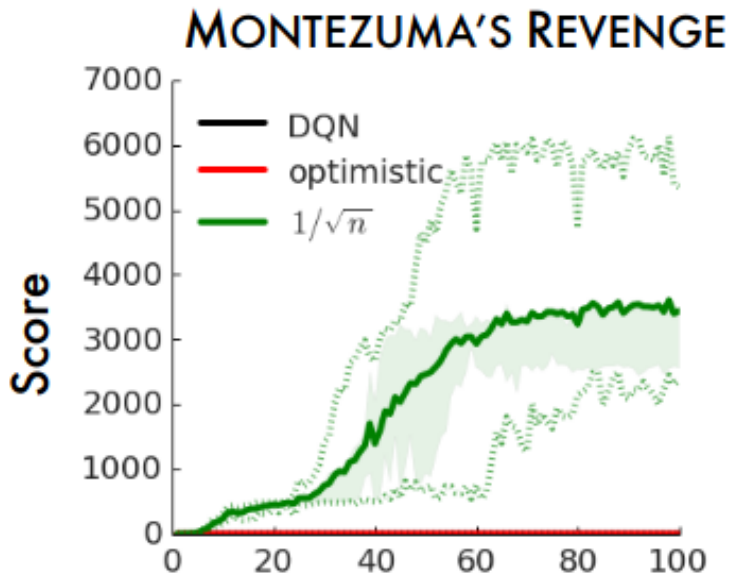
Bellemare, et al. 2016 "Unifying Count-Based Exploration..."

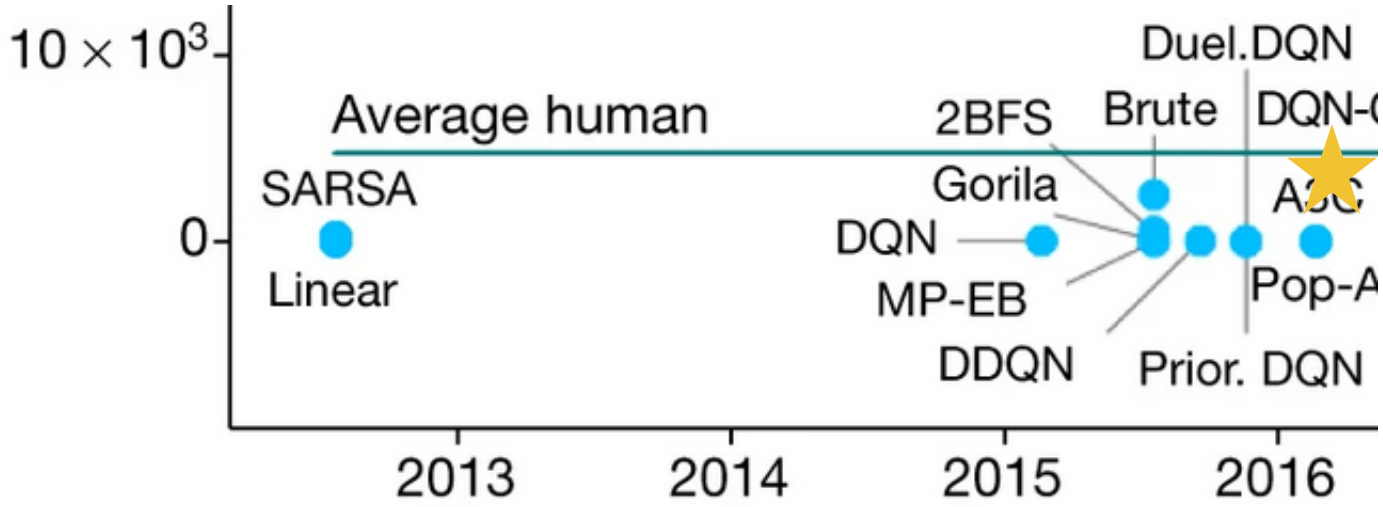
Exploration Bonus

Example 1: Learn Pseudocount

$B(s,a) \approx \frac{1}{\sqrt{\hat{N}(s)}}$ where $\hat{N}(s)$ is a learned function approximation







Bellemare, et al. 2016 "Unifying Count-Based Exploration..."

Exploration Bonus

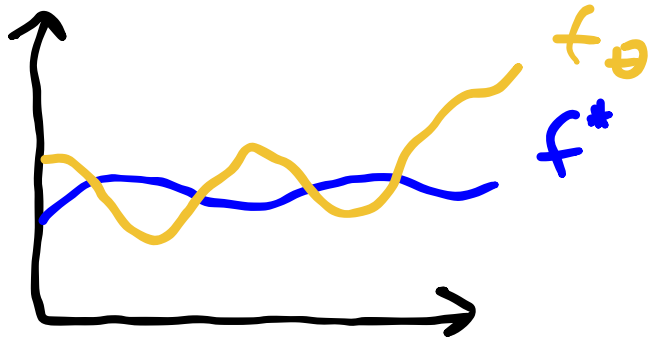
Exploration Bonus

Example 2: Learn a function of the state and action

Exploration Bonus

Example 2: Learn a function of the state and action

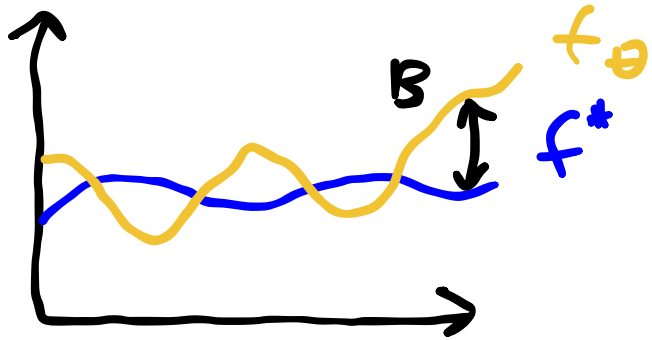
$$B(s,a) = \|\hat{f}_\theta(s,a) - f^*(s,a)\|^2$$



Exploration Bonus

Example 2: Learn a function of the state and action

$$B(s,a) = \|\hat{f}_\theta(s,a) - f^*(s,a)\|^2$$

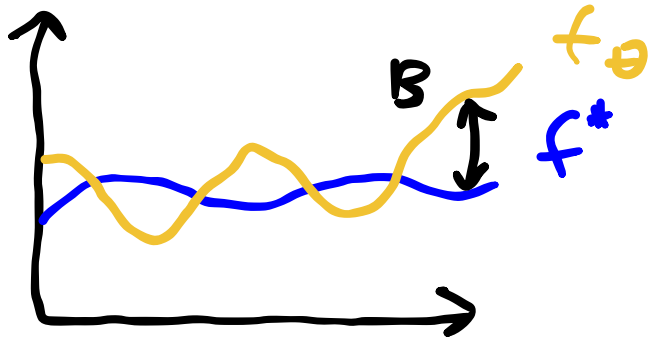


Exploration Bonus

Example 2: Learn a function of the state and action

$$B(s,a) = \|\hat{f}_\theta(s,a) - f^*(s,a)\|^2$$

What should f^* be?

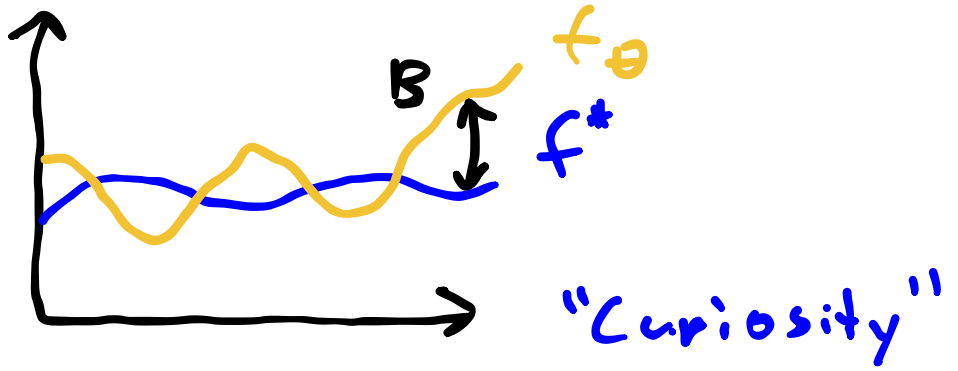


Exploration Bonus

Example 2: Learn a function of the state and action

$$B(s,a) = \|\hat{f}_\theta(s,a) - f^*(s,a)\|^2$$

What should f^* be?



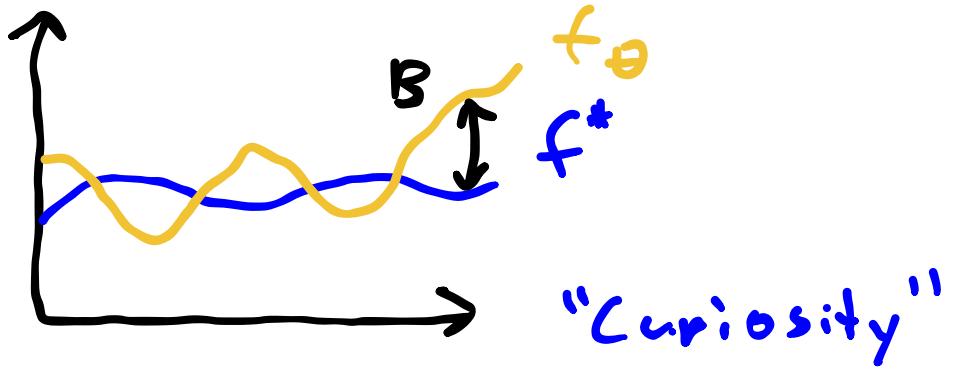
Exploration Bonus

Example 2: Learn a function of the state and action

$$B(s,a) = \|\hat{f}_\theta(s,a) - f^*(s,a)\|^2$$

What should f^* be?

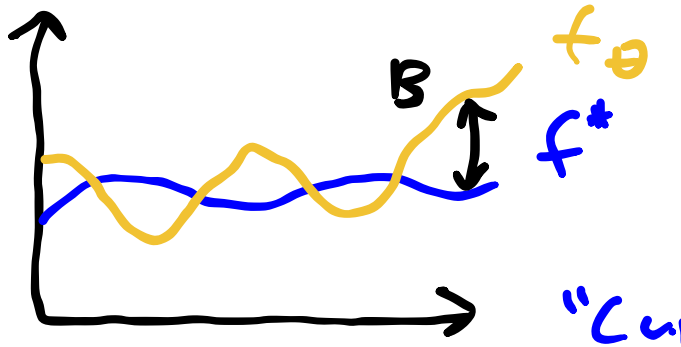
- $f^*(s,a) = s'$ (Next state prediction)



Exploration Bonus

Example 2: Learn a function of the state and action

$$B(s,a) = \|\hat{f}_\theta(s,a) - f^*(s,a)\|^2$$



"Curiosity"

What should f^* be?

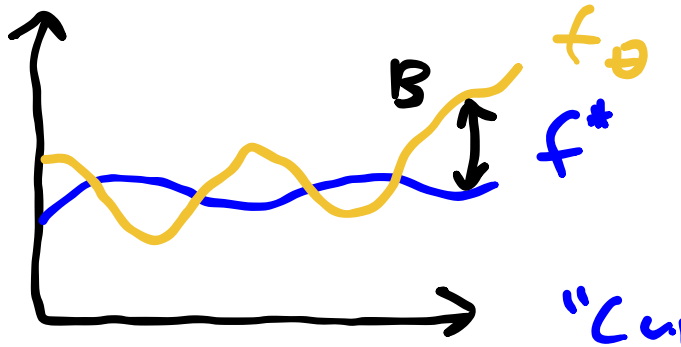
- $f^*(s,a) = s'$ (Next state prediction)
- $f^*(s,a) = f_\phi(s,a)$ where f_ϕ is a random neural network.

"Random Network Distillation"

Exploration Bonus

Example 2: Learn a function of the state and action

$$B(s,a) = \|\hat{f}_\theta(s,a) - f^*(s,a)\|^2$$

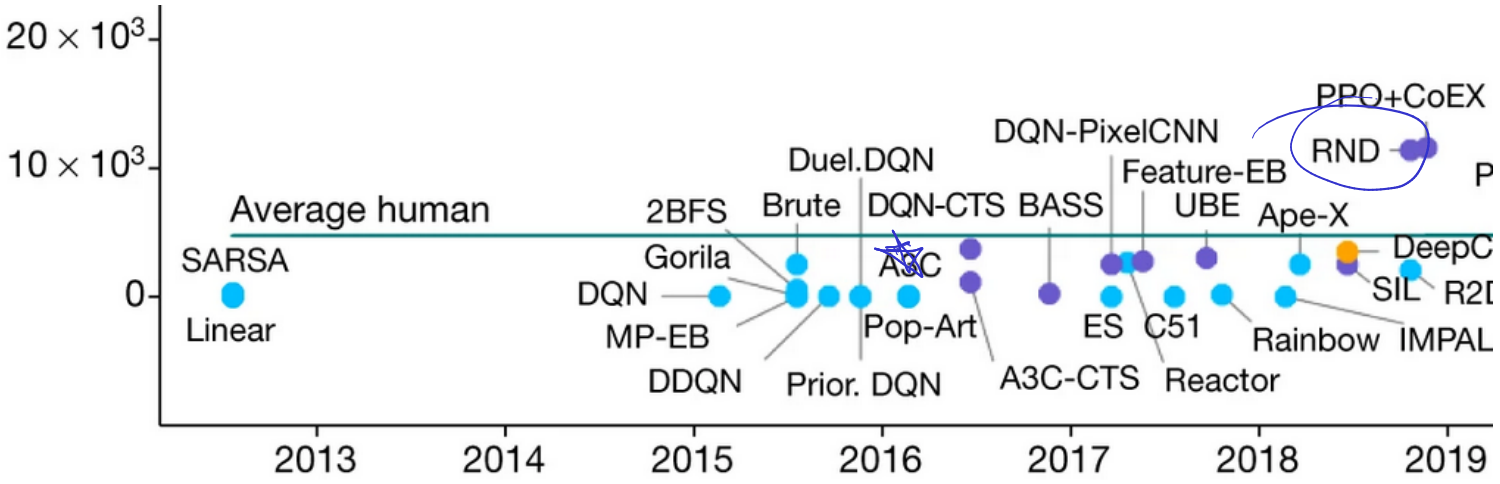


"Curiosity"

What should f^* be?

- $f^*(s,a) = s'$ (Next state prediction)
- $f^*(s,a) = f_\phi(s,a)$ where f_ϕ is a random neural network.

"Random Network Distillation"



Exploration Bonus

Exploration Bonus

Example 3: Thompson Sampling

Exploration Bonus

Example 3: Thompson Sampling

1. Maintain a distribution over Q

Exploration Bonus

Example 3: Thompson Sampling

1. Maintain a distribution over Q
2. Sample Q

Exploration Bonus

Example 3: Thompson Sampling

1. Maintain a distribution over Q
2. Sample Q
3. Act according to Q sampled Q

Exploration Bonus

Example 3: Thompson Sampling

1. Maintain a distribution over Q ← Hard
2. Sample Q
3. Act according to Q

Exploration Bonus

Example 3: Thompson Sampling

1. Maintain a distribution over Q
2. Sample Q
3. Act according to Q

← Hard

- Bootstrapping with multiple Q networks

Exploration Bonus

Example 3: Thompson Sampling

1. Maintain a distribution over Q ← Hard
2. Sample Q
3. Act according to Q

- Bootstrapping with multiple Q networks
- Dropout

Exploration Bonus

Exploration Bonus

Example 4: Go-Explore

Exploration Bonus

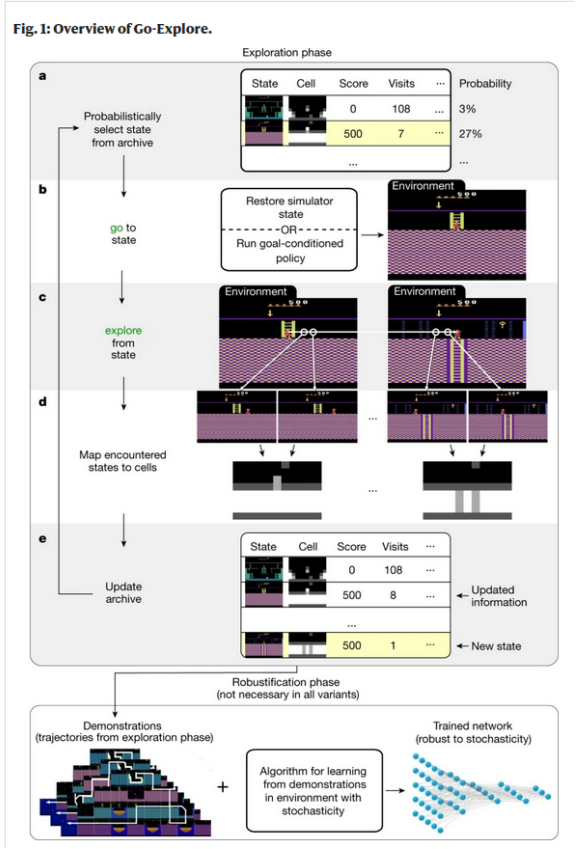
Example 4: Go-Explore

"First return, then explore"

Exploration Bonus

Example 4: Go-Explore

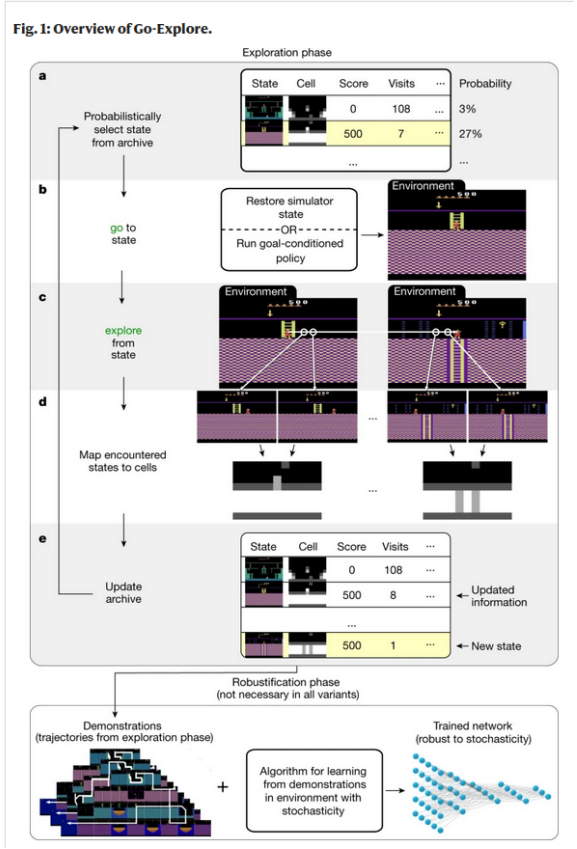
"First return, then explore"

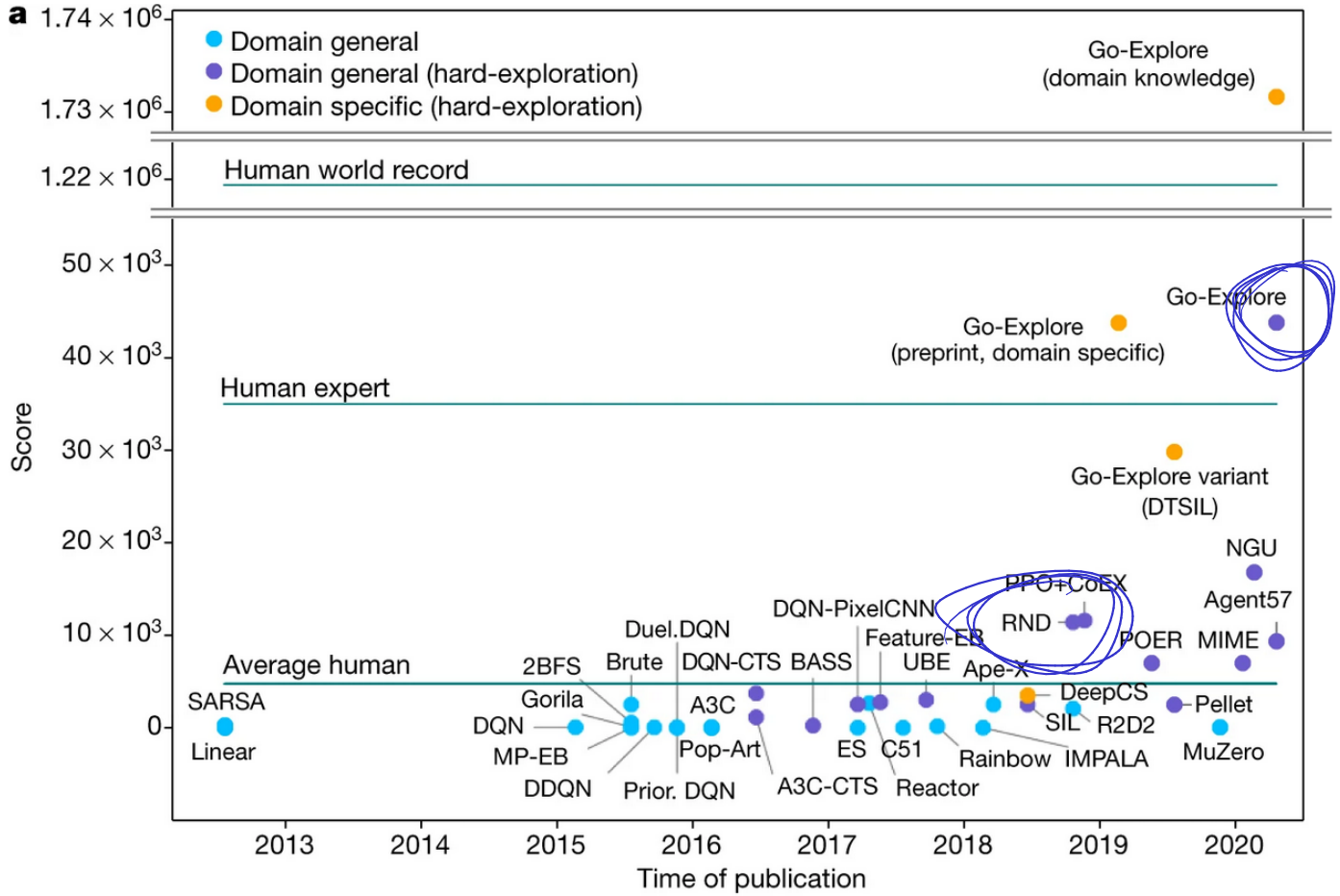


Exploration Bonus

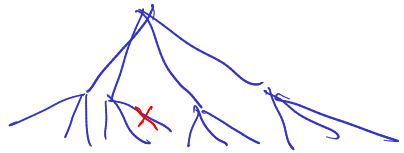
Example 4: Go-Explore

"First return, then explore"



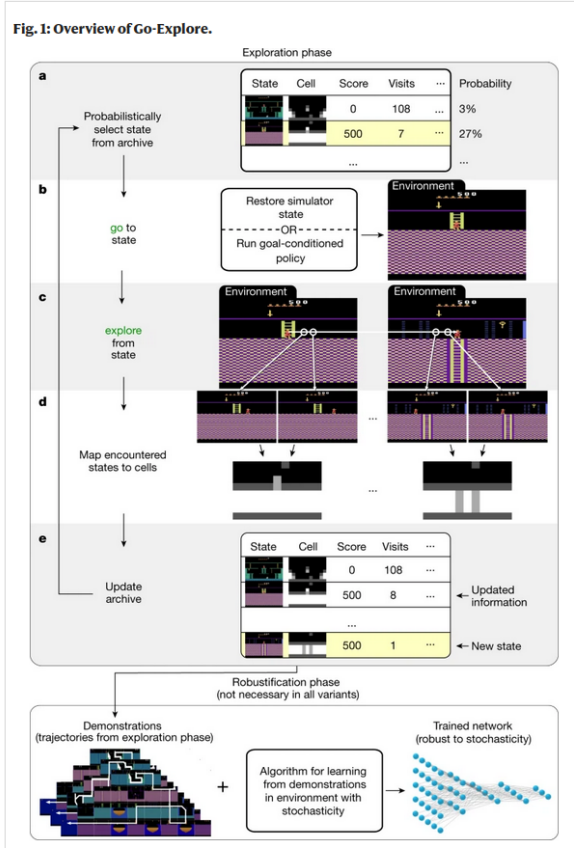


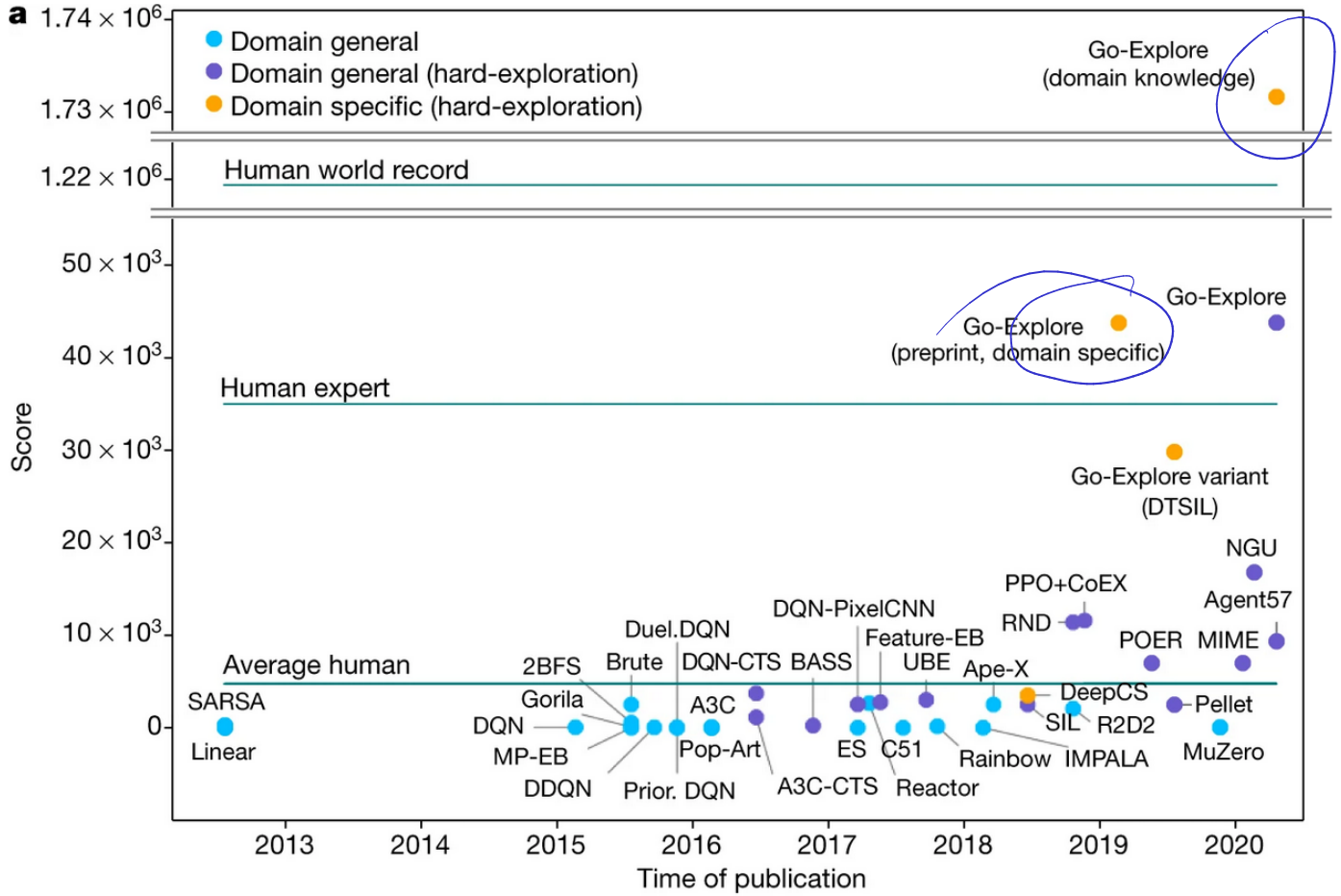
Exploration Bonus



Example 4: Go-Explore

"First return, then explore"





(Uber AI Labs)

Actor-Critic

Actor-Critic

Can we combine value-based and policy-based methods?

Actor-Critic

$$\nabla U(\theta) = E_\tau \left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \gamma^k (r_{k,\text{to-go}} - r_{\text{base}}(s_k)) \right]$$

Actor-Critic

$$\nabla U(\theta) = E_\tau \left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \gamma^k (r_{k,\text{to-go}} - r_{\text{base}}(s_k)) \right]$$

$r_{k,\text{to-go}}$: $Q(s,a)$

$r_{\text{base}}(s_k)$: $V(s)$

Actor-Critic

$$\nabla U(\theta) = E_\tau \left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \gamma^k (r_{k,\text{to-go}} - r_{\text{base}}(s_k)) \right]$$

$r_{k,\text{to-go}}$: $Q(s,a)$

$r_{\text{base}}(s_k)$: $V(s)$

Advantage Function: $A(s,a) = Q(s,a) - V(s)$

$\max_a Q(s,a) = V(s)$

Actor-Critic

$$\nabla U(\theta) = E_\tau \left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \gamma^k (r_{k,\text{to-go}} - r_{\text{base}}(s_k)) \right]$$

$Q(s,a) - V(s)$

$A(s,a)$

Advantage Function: $A(s,a) = Q(s,a) - V(s)$

Actor-Critic

$$\nabla U(\theta) = E_\tau \left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \gamma^k (r_{k,\text{to-go}} - r_{\text{base}}(s_k)) \right]$$

$r_{k,\text{to-go}}$: $Q(s,a)$

$r_{\text{base}}(s_k)$: $V(s)$

$r_{k,\text{to-go}} - r_{\text{base}}(s_k)$: $A(s,a)$

Advantage Function: $A(s,a) = Q(s,a) - V(s)$

- Actor: π_θ

Actor-Critic

$$\nabla U(\theta) = E_\tau \left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \gamma^k (r_{k,\text{to-go}} - r_{\text{base}}(s_k)) \right]$$

$Q(s,a) - V(s)$

$A(s,a)$

Advantage Function: $A(s,a) = Q(s,a) - V(s)$

- Actor: π_θ
- Critic: Q_ϕ and/or A_ϕ and/or V_ϕ

Actor-Critic

$$\nabla U(\theta) = E_\tau \left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \gamma^k (r_{k,\text{to-go}} - r_{\text{base}}(s_k)) \right]$$

$Q(s,a) - V(s)$

$A(s,a)$

Advantage Function: $A(s,a) = Q(s,a) - V(s)$

- Actor: π_θ
- Critic: Q_ϕ and/or A_ϕ and/or V_ϕ

Alternate between training Actor and Critic

Actor-Critic

$$\nabla U(\theta) = E_\tau \left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \gamma^k (r_{k,\text{to-go}} - r_{\text{base}}(s_k)) \right]$$

$Q(s,a)$ $V(s)$

$A(s,a)$

Advantage Function: $A(s,a) = Q(s,a) - V(s)$

- Actor: π_θ
- Critic: Q_ϕ and/or A_ϕ and/or V_ϕ

Alternate between training Actor and Critic

Problem: Instability

Actor-Critic

Which should we learn? A, Q, or V?

$Q(s,a)$

$V(s)$

Actor-Critic

Which should we learn? A, Q, or V?

$$\nabla U(\theta) = E_\tau \left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \gamma^k (r_k + \gamma V_\phi(s_{k+1}) - V_\phi(s_k)) \right]$$

Actor-Critic

Which should we learn? A, Q, or V?

$$\nabla U(\theta) = E_\tau \left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \gamma^k (r_k + \gamma V_\phi(s_{k+1}) - V_\phi(s_k)) \right]$$

temporal difference residual

Actor-Critic

Which should we learn? A, Q, or V?

$$\nabla U(\theta) = E_\tau \left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \gamma^k (r_k + \gamma V_\phi(s_{k+1}) - V_\phi(s_k)) \right]$$

temporal difference residual

$$l(\phi) = E\left[\left(V_\phi(s) - V^{\pi_\theta}(s) \right)^2 \right]$$

Actor-Critic

Which should we learn? A, Q, or V?

$$\nabla U(\theta) = E_\tau \left[\sum_{k=0}^{d} \nabla_\theta \log \pi_\theta(a_k \mid s_k) \gamma^k (r_k + \gamma V_\phi(s_{k+1}) - V_\phi(s_k)) \right]$$

temporal difference residual

$$l(\phi) = E\left[\left(V_\phi(s) - V^{\pi_\theta}(s) \right)^2 \right]$$

↑ estimate with reward to go from sims

Generalized Advantage Estimation

Generalized Advantage Estimation

$$A(s_k, a_k) \approx r_k + \gamma V_\phi(s_{k+1}) - V_\phi(s_k)$$

Generalized Advantage Estimation

$$A(s_k, a_k) \approx r_k + \gamma V_\phi(s_{k+1}) - V_\phi(s_k)$$

$$A(s_k, a_k) \approx \sum_{t=k}^{\infty} \gamma^{t-k} r_t$$

Generalized Advantage Estimation

V^{π_θ}

$$A(s_k, a_k) \approx r_k + \gamma V_\phi(s_{k+1}) - V_\phi(s_k)$$ ← High Bias

$$A(s_k, a_k) \approx \sum_{t=k}^{\infty} \gamma^{t-k} r_t$$ ← High Variance

r_{togo}

Generalized Advantage Estimation

$$A(s_k, a_k) \approx r_k + \gamma V_\phi(s_{k+1}) - V_\phi(s_k)$$ ← High Bias

$$A(s_k, a_k) \approx \sum_{t=k}^{\infty} \gamma^{t-k} r_t$$ ← High Variance

$$A(s_k, a_k) \approx \sum_{t=k}^{d-1} \gamma^{t-k} r_t + \gamma^{d-k} r_d + \gamma V_\phi(s_{d+1}) - V_\phi(s_d)$$

Generalized Advantage Estimation

$$A(s_k, a_k) \approx r_k + \gamma V_\phi(s_{k+1}) - V_\phi(s_k)$$ ← High Bias

$$A(s_k, a_k) \approx \sum_{t=k}^{\infty} \gamma^{t-k} r_t$$ ← High Variance

$$A(s_k, a_k) \approx \sum_{t=k}^{d-1} \gamma^{t-k} r_t + \gamma^{d-k} r_d + \gamma V_\phi(s_{d+1}) - V_\phi(s_d)$$

↖ When should we stop?

Generalized Advantage Estimation

$$A(s_k, a_k) \approx r_k + \gamma V_\phi(s_{k+1}) - V_\phi(s_k)$$ ← High Bias

$$A(s_k, a_k) \approx \sum_{t=k}^{\infty} \gamma^{t-k} r_t$$ ← High Variance

$$A(s_k, a_k) \approx \sum_{t=k}^{d-1} \gamma^{t-k} r_t + \gamma^{d-k} r_d + \gamma V_\phi(s_{d+1}) - V_\phi(s_d)$$

↖ When should we stop?

let $\delta_t = r_t + \gamma V_\phi(s_{t+1}) - V_\phi(s_t)$

Generalized Advantage Estimation

$$A(s_k, a_k) \approx r_k + \gamma V_\phi(s_{k+1}) - V_\phi(s_k)$$ ← High Bias

$$A(s_k, a_k) \approx \sum_{t=k}^{\infty} \gamma^{t-k} r_t$$ ← High Variance

$$A(s_k, a_k) \approx \sum_{t=k}^{d-1} \gamma^{t-k} r_t + \gamma^{d-k} r_d + \gamma V_\phi(s_{d+1}) - V_\phi(s_d)$$

↖ When should we stop?

let $\delta_t = r_t + \gamma V_\phi(s_{t+1}) - V_\phi(s_t)$

$$A_{\text{GAE}}(s_k, a_k) \approx \sum_{t=k}^{\infty} (\gamma\lambda)^{t-k} \delta_t$$

Generalized Advantage Estimation

$$A(s_k, a_k) \approx r_k + \gamma V_\phi(s_{k+1}) - V_\phi(s_k)$$ ← High Bias

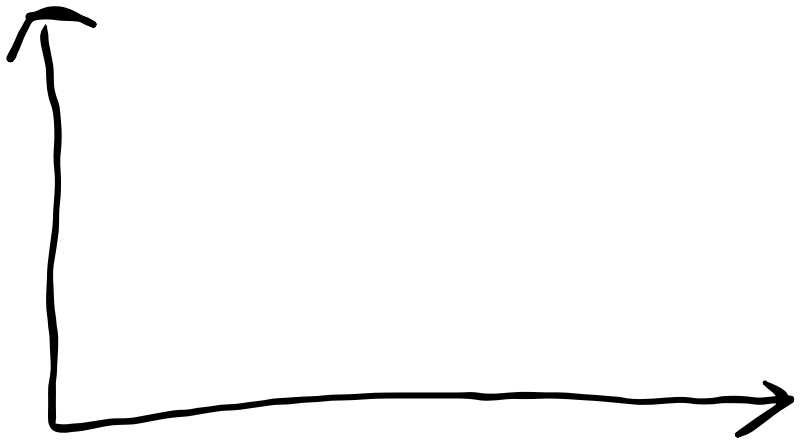
$$A(s_k, a_k) \approx \sum_{t=k}^{\infty} \gamma^{t-k} r_t$$ ← High Variance

$$A(s_k, a_k) \approx \sum_{t=k}^{d-1} \gamma^{t-k} r_t + \gamma^{d-k} r_d + \gamma V_\phi(s_{d+1}) - V_\phi(s_d)$$

↖ When should we stop?

let $\delta_t = r_t + \gamma V_\phi(s_{t+1}) - V_\phi(s_t)$

$$A_{\text{GAE}}(s_k, a_k) \approx \sum_{t=k}^{\infty} (\gamma\lambda)^{t-k} \delta_t$$



Generalized Advantage Estimation

$$A(s_k, a_k) \approx r_k + \gamma V_\phi(s_{k+1}) - V_\phi(s_k)$$ ← High Bias

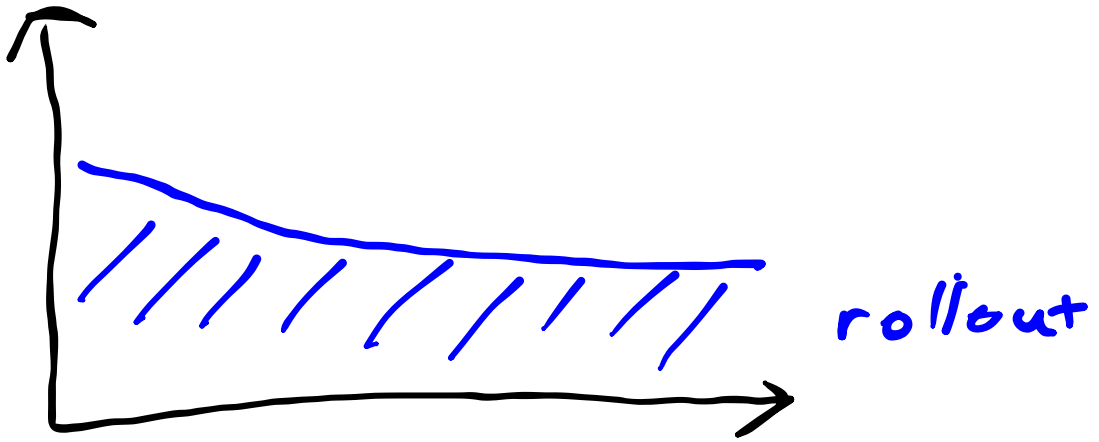
$$A(s_k, a_k) \approx \sum_{t=k}^{\infty} \gamma^{t-k} r_t$$ ← High Variance

$$A(s_k, a_k) \approx \sum_{t=k}^{d-1} \gamma^{t-k} r_t + \gamma^{d-k} r_d + \gamma V_\phi(s_{d+1}) - V_\phi(s_d)$$

↖ When should we stop?

let $\delta_t = r_t + \gamma V_\phi(s_{t+1}) - V_\phi(s_t)$

$$A_{\text{GAE}}(s_k, a_k) \approx \sum_{t=k}^{\infty} (\gamma\lambda)^{t-k} \delta_t$$



Generalized Advantage Estimation

V^{π_θ}

$$A(s_k, a_k) \approx r_k + \gamma V_\phi(s_{k+1}) - V_\phi(s_k)$$ ← High Bias

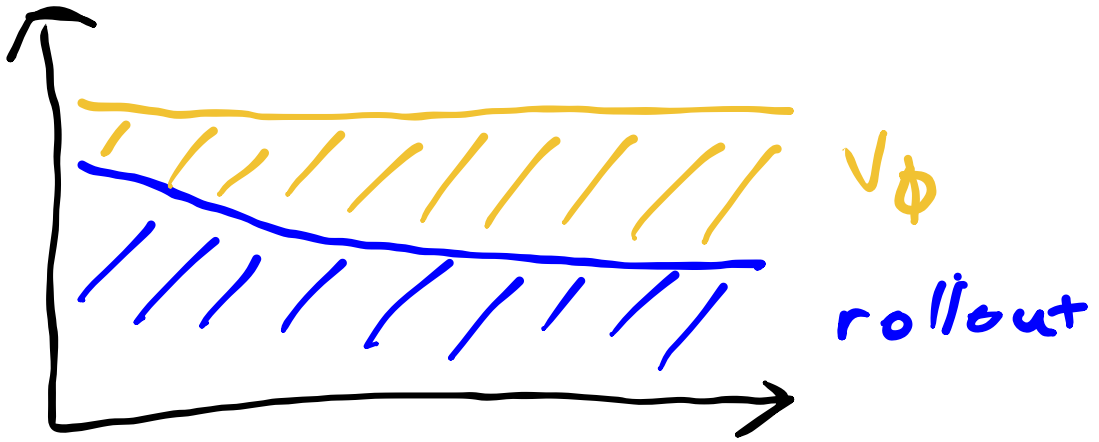
$$A(s_k, a_k) \approx \sum_{t=k}^{\infty} \gamma^{t-k} r_t$$ ← High Variance

$$A(s_k, a_k) \approx \sum_{t=k}^{d-1} \gamma^{t-k} r_t + \gamma^{d-k} r_d + \gamma V_\phi(s_{d+1}) - V_\phi(s_d)$$

↖ When should we stop?

let $\delta_t = r_t + \gamma V_\phi(s_{t+1}) - V_\phi(s_t)$

$$A_{\text{GAE}}(s_k, a_k) \approx \sum_{t=k}^{\infty} (\gamma\lambda)^{t-k} \delta_t$$



Continuous Actions: Deep Deterministic Policy Gradient

$Q_\phi(s,a)$

$\text{argmax}\ Q_\phi$

$\pi_\theta(s) = \underset{a \in A}{\text{argmax}}\ Q_\phi(s,a)$

DDPG – fragile

$$l(\phi) = E\left[\left(r + \gamma Q_\phi(s', \pi_\theta(s')) - Q_\phi(s,a)\right)^2\right]$$

$$U(\theta) = E\left[Q_\phi(s, \pi_\theta(s))\right]$$

$$\nabla U(\theta) = E\left[\nabla_\theta Q_\phi(s, \pi_\theta(s))\right]$$

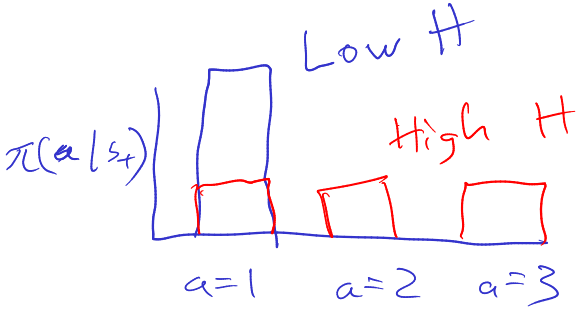
$$= E\left[\nabla_\theta \pi_\theta(s)\ \nabla_a Q_\phi(s,a)\big|_{a=\pi_\theta(s)}\right]$$

Soft Actor Critic: Entropy Regularization

Soft Actor Critic: Entropy Regularization

$$J(\pi) = E\left[\sum_{t=0}^{\infty} \gamma^t \left(r_t + \alpha \mathcal{H}(\pi(\cdot \mid s_t))\right)\right]$$

temp



Soft Actor Critic: Entropy Regularization

$$J(\pi) = E\left[\sum_{t=0}^{\infty} \gamma^t \left(r_t + \alpha \mathcal{H}(\pi(\cdot \mid s_t))\right)\right]$$

$$V(\mathbf{s}_t) = \mathbb{E}_{\mathbf{a}_t \sim \pi}\left[Q(\mathbf{s}_t, \mathbf{a}_t) - \log \pi(\mathbf{a}_t|\mathbf{s}_t)\right]$$

Soft - Value

Soft Actor Critic: Entropy Regularization

$$J(\pi) = E\left[\sum_{t=0}^{\infty} \gamma^t \left(r_t + \alpha \mathcal{H}(\pi(\cdot \mid s_t))\right)\right]$$

$$V(\mathbf{s}_t) = \mathbb{E}_{\mathbf{a}_t \sim \pi}\left[Q(\mathbf{s}_t, \mathbf{a}_t) - \log \pi(\mathbf{a}_t|\mathbf{s}_t)\right]$$

$$\mathcal{T}^{\pi} Q(\mathbf{s}_t, \mathbf{a}_t) \triangleq r(\mathbf{s}_t, \mathbf{a}_t) + \gamma\, \mathbb{E}_{\mathbf{s}_{t+1} \sim p}\left[V(\mathbf{s}_{t+1})\right]$$

Soft Actor Critic: Entropy Regularization

$$J(\pi) = E\left[\sum_{t=0}^{\infty} \gamma^t \left(r_t + \alpha \mathcal{H}(\pi(\cdot \mid s_t))\right)\right]$$

Soft Policy Iteration

$$V(\mathbf{s}_t) = \mathbb{E}_{\mathbf{a}_t \sim \pi}\left[Q(\mathbf{s}_t, \mathbf{a}_t) - \log \pi(\mathbf{a}_t|\mathbf{s}_t)\right]$$

$$\mathcal{T}^{\pi} Q(\mathbf{s}_t, \mathbf{a}_t) \triangleq r(\mathbf{s}_t, \mathbf{a}_t) + \gamma\, \mathbb{E}_{\mathbf{s}_{t+1} \sim p}\left[V(\mathbf{s}_{t+1})\right]$$

$$\pi_{\text{new}} = \arg\min_{\pi' \in \Pi} \mathrm{D}_{\mathrm{KL}}\left(\pi'(\cdot|\mathbf{s}_t) \,\middle\|\, \frac{\exp\left(Q^{\pi_{\text{old}}}(\mathbf{s}_t, \cdot)\right)}{Z^{\pi_{\text{old}}}(\mathbf{s}_t)}\right)$$

$\frac{e^{\lambda Q}}{\sum e^{\lambda Q}}$

Soft Actor Critic

Soft Actor Critic

Algorithm 1 Soft Actor-Critic

Initialize parameter vectors ψ, $\bar{\psi}$, θ, ϕ.
for each iteration **do**
  for each environment step **do**
    $\mathbf{a}_t \sim \pi_\phi(\mathbf{a}_t|\mathbf{s}_t)$
    $\mathbf{s}_{t+1} \sim p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathbf{s}_t, \mathbf{a}_t, r(\mathbf{s}_t, \mathbf{a}_t), \mathbf{s}_{t+1})\}$
  end for
  for each gradient step **do**
    $\psi \leftarrow \psi - \lambda_V \hat{\nabla}_\psi J_V(\psi)$
    $\theta_i \leftarrow \theta_i - \lambda_Q \hat{\nabla}_{\theta_i} J_Q(\theta_i)$ for $i \in \{1, 2\}$
    $\phi \leftarrow \phi - \lambda_\pi \hat{\nabla}_\phi J_\pi(\phi)$
    $\bar{\psi} \leftarrow \tau\psi + (1-\tau)\bar{\psi}$
  end for
end for

Soft Actor Critic

Algorithm 1 Soft Actor-Critic

Initialize parameter vectors ψ, $\bar{\psi}$, θ, ϕ.
for each iteration **do**
  for each environment step **do**
    $\mathbf{a}_t \sim \pi_\phi(\mathbf{a}_t|\mathbf{s}_t)$
    $\mathbf{s}_{t+1} \sim p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathbf{s}_t, \mathbf{a}_t, r(\mathbf{s}_t, \mathbf{a}_t), \mathbf{s}_{t+1})\}$
  end for
  for each gradient step **do**
    $\psi \leftarrow \psi - \lambda_V \hat{\nabla}_\psi J_V(\psi)$
    $\theta_i \leftarrow \theta_i - \lambda_Q \hat{\nabla}_{\theta_i} J_Q(\theta_i)$ for $i \in \{1, 2\}$
    $\phi \leftarrow \phi - \lambda_\pi \hat{\nabla}_\phi J_\pi(\phi)$
    $\bar{\psi} \leftarrow \tau\psi + (1-\tau)\bar{\psi}$
  end for
end for

$$J_V(\psi) = \mathbb{E}_{\mathbf{s}_t \sim \mathcal{D}}\left[\frac{1}{2}\left(V_\psi(\mathbf{s}_t) - \mathbb{E}_{\mathbf{a}_t \sim \pi_\phi}\left[Q_\theta(\mathbf{s}_t, \mathbf{a}_t) - \log \pi_\phi(\mathbf{a}_t|\mathbf{s}_t)\right]\right)^2\right]$$

Soft Actor Critic

Algorithm 1 Soft Actor-Critic

Initialize parameter vectors ψ, $\bar{\psi}$, θ, ϕ.
for each iteration **do**
  for each environment step **do**
    $\mathbf{a}_t \sim \pi_\phi(\mathbf{a}_t|\mathbf{s}_t)$
    $\mathbf{s}_{t+1} \sim p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathbf{s}_t, \mathbf{a}_t, r(\mathbf{s}_t, \mathbf{a}_t), \mathbf{s}_{t+1})\}$
  end for
  for each gradient step **do**
    $\psi \leftarrow \psi - \lambda_V \hat{\nabla}_\psi J_V(\psi)$
    $\theta_i \leftarrow \theta_i - \lambda_Q \hat{\nabla}_{\theta_i} J_Q(\theta_i)$ for $i \in \{1, 2\}$
    $\phi \leftarrow \phi - \lambda_\pi \hat{\nabla}_\phi J_\pi(\phi)$
    $\bar{\psi} \leftarrow \tau\psi + (1-\tau)\bar{\psi}$
  end for
end for

entropy Bonus

$$J_V(\psi) = \mathbb{E}_{\mathbf{s}_t \sim \mathcal{D}}\left[\frac{1}{2}\left(V_\psi(\mathbf{s}_t) - \mathbb{E}_{\mathbf{a}_t \sim \pi_\phi}\left[Q_\theta(\mathbf{s}_t, \mathbf{a}_t) - \log \pi_\phi(\mathbf{a}_t|\mathbf{s}_t)\right]\right)^2\right]$$

$$J_Q(\theta) = \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \mathcal{D}}\left[\frac{1}{2}\left(Q_\theta(\mathbf{s}_t, \mathbf{a}_t) - \hat{Q}(\mathbf{s}_t, \mathbf{a}_t)\right)^2\right]$$

Soft Actor Critic

Algorithm 1 Soft Actor-Critic

Initialize parameter vectors ψ, $\bar{\psi}$, θ, ϕ.
for each iteration **do**
  for each environment step **do**
    $\mathbf{a}_t \sim \pi_\phi(\mathbf{a}_t|\mathbf{s}_t)$
    $\mathbf{s}_{t+1} \sim p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathbf{s}_t, \mathbf{a}_t, r(\mathbf{s}_t, \mathbf{a}_t), \mathbf{s}_{t+1})\}$
  end for
  for each gradient step **do**
    $\psi \leftarrow \psi - \lambda_V \hat{\nabla}_\psi J_V(\psi)$
    $\theta_i \leftarrow \theta_i - \lambda_Q \hat{\nabla}_{\theta_i} J_Q(\theta_i)$ for $i \in \{1, 2\}$
    $\phi \leftarrow \phi - \lambda_\pi \hat{\nabla}_\phi J_\pi(\phi)$
    $\bar{\psi} \leftarrow \tau\psi + (1-\tau)\bar{\psi}$
  end for
end for

$$J_V(\psi) = \mathbb{E}_{\mathbf{s}_t \sim \mathcal{D}}\left[\frac{1}{2}\left(V_\psi(\mathbf{s}_t) - \mathbb{E}_{\mathbf{a}_t \sim \pi_\phi}\left[Q_\theta(\mathbf{s}_t, \mathbf{a}_t) - \log \pi_\phi(\mathbf{a}_t|\mathbf{s}_t)\right]\right)^2\right]$$

$$J_Q(\theta) = \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \mathcal{D}}\left[\frac{1}{2}\left(Q_\theta(\mathbf{s}_t, \mathbf{a}_t) - \hat{Q}(\mathbf{s}_t, \mathbf{a}_t)\right)^2\right]$$

$$\hat{Q}(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + \gamma\, \mathbb{E}_{\mathbf{s}_{t+1} \sim p}\left[V_{\bar{\psi}}(\mathbf{s}_{t+1})\right]$$

Soft Actor Critic

Algorithm 1 Soft Actor-Critic

Initialize parameter vectors ψ, $\bar{\psi}$, θ, ϕ.
for each iteration **do**
  for each environment step **do**
    $\mathbf{a}_t \sim \pi_\phi(\mathbf{a}_t|\mathbf{s}_t)$
    $\mathbf{s}_{t+1} \sim p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathbf{s}_t, \mathbf{a}_t, r(\mathbf{s}_t, \mathbf{a}_t), \mathbf{s}_{t+1})\}$
  end for
  for each gradient step **do**
    $\psi \leftarrow \psi - \lambda_V \hat{\nabla}_\psi J_V(\psi)$
    $\theta_i \leftarrow \theta_i - \lambda_Q \hat{\nabla}_{\theta_i} J_Q(\theta_i)$ for $i \in \{1, 2\}$
    $\phi \leftarrow \phi - \lambda_\pi \hat{\nabla}_\phi J_\pi(\phi)$
    $\bar{\psi} \leftarrow \tau\psi + (1-\tau)\bar{\psi}$
  end for
end for

$$J_V(\psi) = \mathbb{E}_{\mathbf{s}_t \sim \mathcal{D}}\left[\frac{1}{2}\left(V_\psi(\mathbf{s}_t) - \mathbb{E}_{\mathbf{a}_t \sim \pi_\phi}\left[Q_\theta(\mathbf{s}_t, \mathbf{a}_t) - \log \pi_\phi(\mathbf{a}_t|\mathbf{s}_t)\right]\right)^2\right]$$

$$J_Q(\theta) = \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \mathcal{D}}\left[\frac{1}{2}\left(Q_\theta(\mathbf{s}_t, \mathbf{a}_t) - \hat{Q}(\mathbf{s}_t, \mathbf{a}_t)\right)^2\right]$$

$$\hat{Q}(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + \gamma\, \mathbb{E}_{\mathbf{s}_{t+1} \sim p}\left[V_{\bar{\psi}}(\mathbf{s}_{t+1})\right]$$

$$J_\pi(\phi) = \mathbb{E}_{\mathbf{s}_t \sim \mathcal{D}}\left[\mathrm{D}_{\mathrm{KL}}\left(\pi_\phi(\cdot|\mathbf{s}_t) \,\middle\|\, \frac{\exp\left(Q_\theta(\mathbf{s}_t, \cdot)\right)}{Z_\theta(\mathbf{s}_t)}\right)\right]$$

Soft Actor Critic

Advantages:

Soft Actor Critic

Advantages:

- Stable

Soft Actor Critic

Advantages:

- Stable
- Learns many near-optimal policies

Soft Actor Critic

Advantages:

- Stable
- Learns many near-optimal policies
- Exploration

Soft Actor Critic

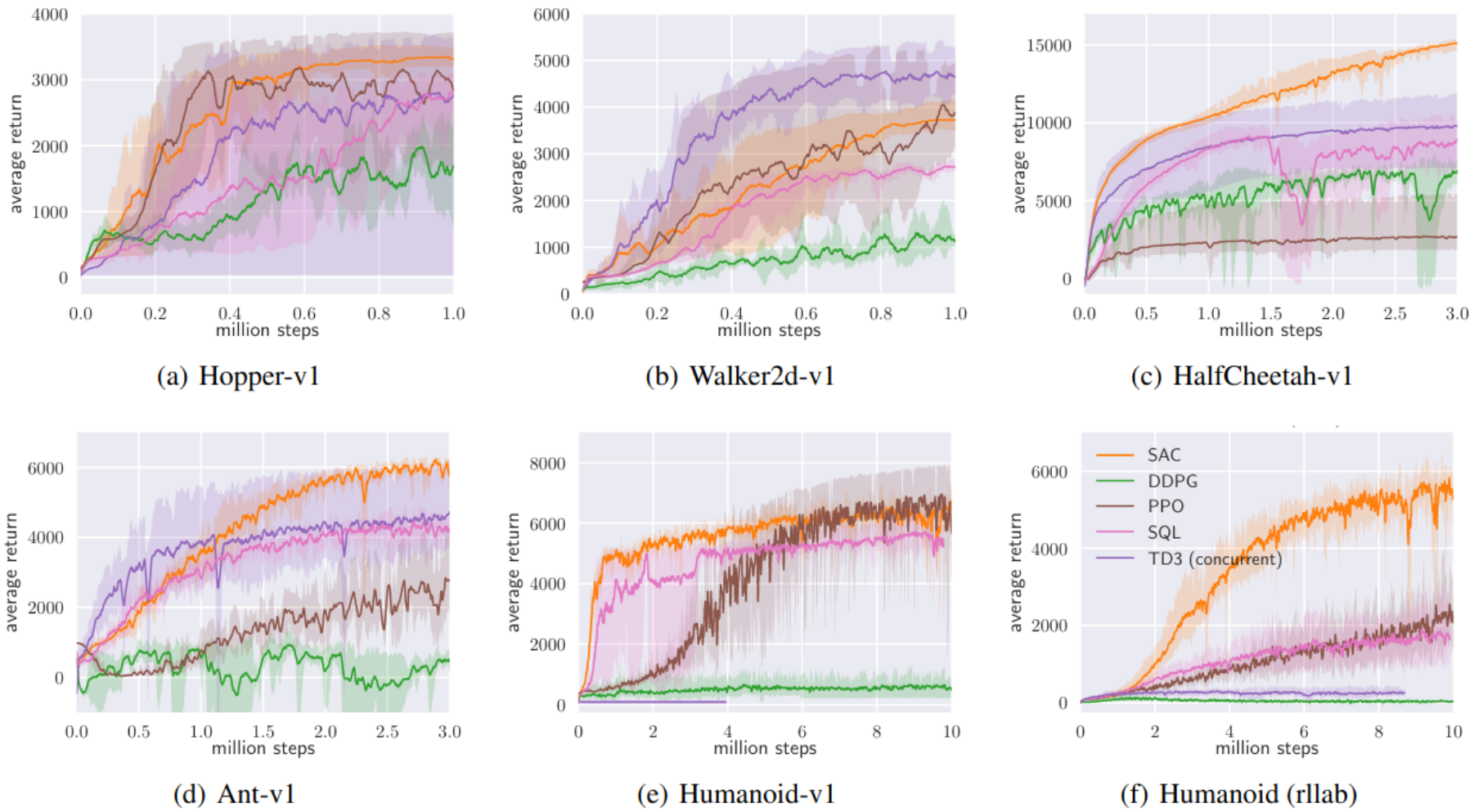
Advantages:

- Stable
- Learns many near-optimal policies
- Exploration
- Insensitivity to hyperparameters

Soft Actor Critic

Advantages:

- Stable
- Learns
- Explora
- Insensit



(a) Hopper-v1 (b) Walker2d-v1 (c) HalfCheetah-v1

(d) Ant-v1 (e) Humanoid-v1 (f) Humanoid (rllab)

Soft Actor Critic

Advantages:

- Stable
- Learns many near-optimal policies
- Exploration
- Insensitivity to hyperparameters

Disadvantages

- Sensitive to α Solution = Entropy *constraint* and adjust α

Soft Actor Critic

maximize $E[\sum_t r_t]$
s.t. $H(\pi(\cdot|s)) \geq h$

Advantages:

- Sta
- Le
- Ex
- Ins

Disadvantages

Algorithm 1 Soft Actor-Critic

Input: θ_1, θ_2, ϕ ▷ Initial parameters
$\bar{\theta}_1 \leftarrow \theta_1, \bar{\theta}_2 \leftarrow \theta_2$ ▷ Initialize target network weights
$\mathcal{D} \leftarrow \emptyset$ ▷ Initialize an empty replay pool
for each iteration **do**
  for each environment step **do**
    $\mathbf{a}_t \sim \pi_\phi(\mathbf{a}_t|\mathbf{s}_t)$ ▷ Sample action from the policy
    $\mathbf{s}_{t+1} \sim p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$ ▷ Sample transition from the environment
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathbf{s}_t, \mathbf{a}_t, r(\mathbf{s}_t, \mathbf{a}_t), \mathbf{s}_{t+1})\}$ ▷ Store the transition in the replay pool
  end for
  for each gradient step **do**
    $\theta_i \leftarrow \theta_i - \lambda_Q \hat{\nabla}_{\theta_i} J_Q(\theta_i)$ for $i \in \{1, 2\}$ ▷ Update the Q-function parameters
    $\phi \leftarrow \phi - \lambda_\pi \hat{\nabla}_\phi J_\pi(\phi)$ ▷ Update policy weights
    $\alpha \leftarrow \alpha - \lambda \hat{\nabla}_\alpha J(\alpha)$ ▷ Adjust temperature
    $\bar{\theta}_i \leftarrow \tau\theta_i + (1-\tau)\bar{\theta}_i$ for $i \in \{1, 2\}$ ▷ Update target network weights
  end for
end for
Output: θ_1, θ_2, ϕ ▷ Optimized parameters

Wisdom

- Brittleness
- Sample Complexity
- Applications (alpha fold)