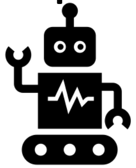
Off-Policy Imitation Learning from Observations

Zhuangdi Zhu, Kaixiang Lin, Jiayu Zhou, Bo Dai, 2020 NuerIPS

Imitation Learning

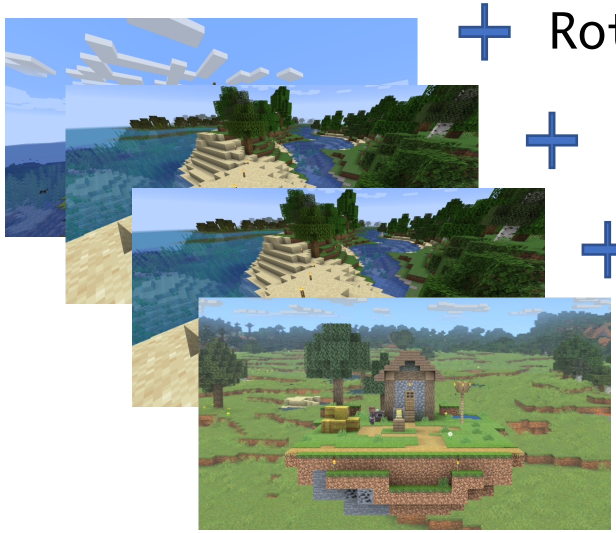
Key idea of Imitation Learning : Learning policy π_θ by imitating samples from an expert policy π_E

Expert:



π_E

States + Actions



+ Rotate left 30°

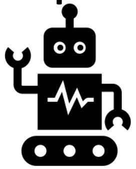
+ forward

+ forward

+ Rotate right 45°

Learning from ***Demonstrations:*** learning agent has access to samples of (state, action) pairs.

Expert:



π_E

States

Learning from ***Observations:*** learning agent has access to samples of state only.

Motivation: Why Learning from Observations

Dispense with the costs of collecting expert actions.

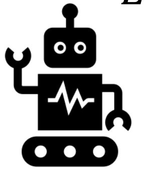
Approach to Human intelligence.

The Goal of Learning from Observations

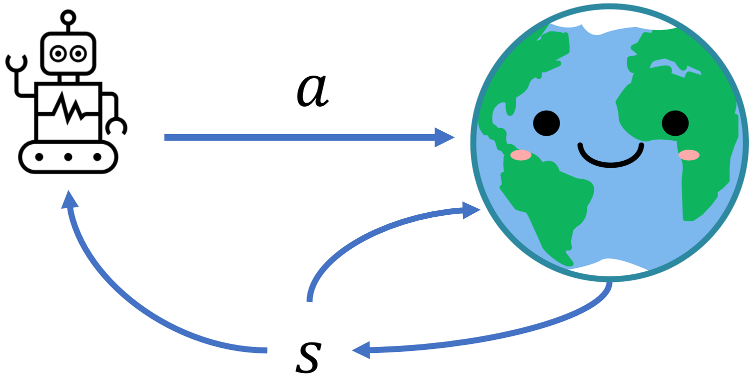
Minimizing the ***footprint*** (state-transition) distribution between the expert and the learning agent

$$\min J_{\mathrm{LfO}}(\pi) := \mathbb{D}_{\mathbf{KL}}[\mu^{\pi}(s,s')||\mu^{E}(s,s')].$$

Expert: π_E

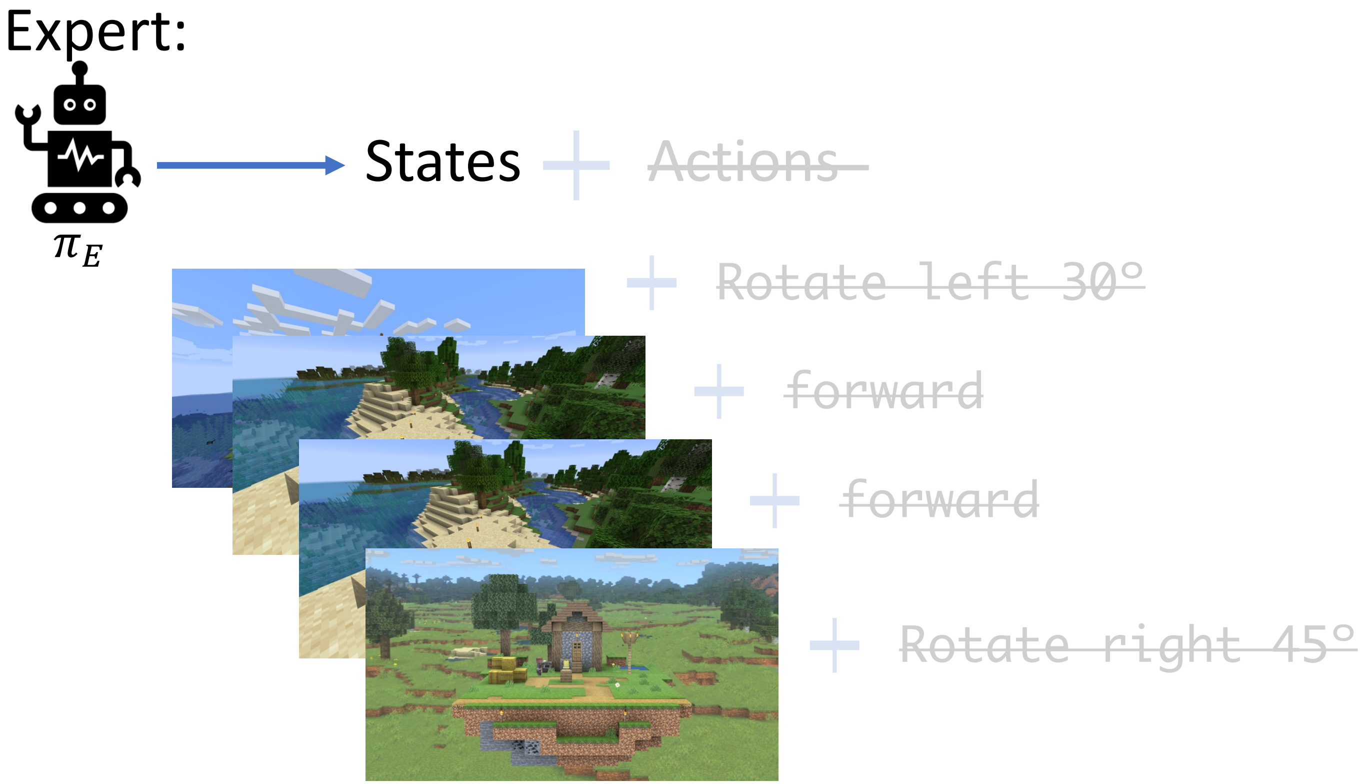


Learning policy: $\pi_\theta(a|s)$



Challenges of Learning from Observations

Lack of action guidance



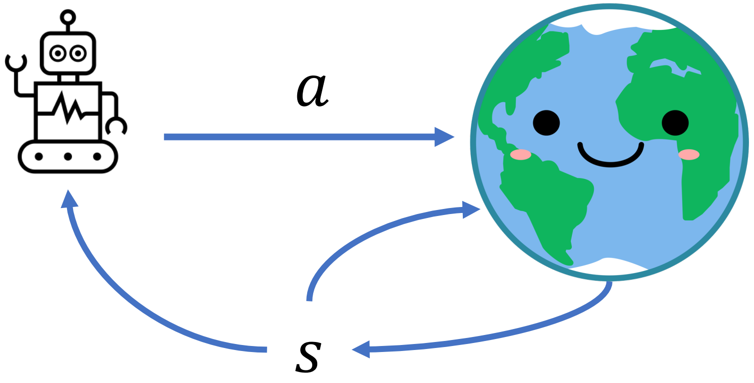
Challenges of Learning from Observations

Lack of action guidance

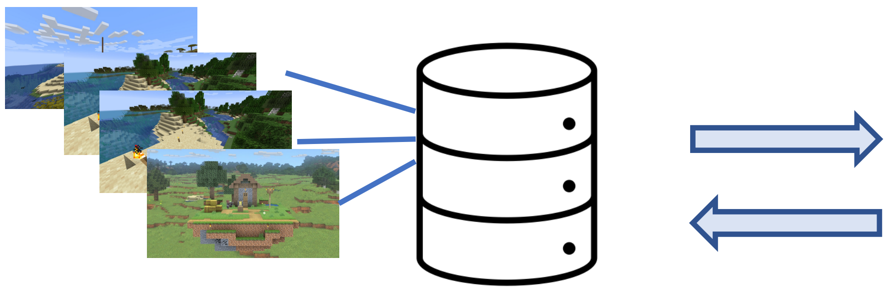
Sample inefficient

On-policy Sampling and Learning

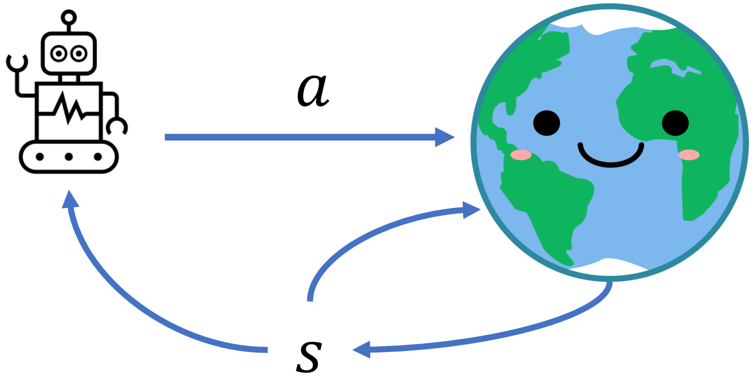
Learning policy: $\pi_\theta(a|s)$



Difference between On-Policy and Off-Policy Learning



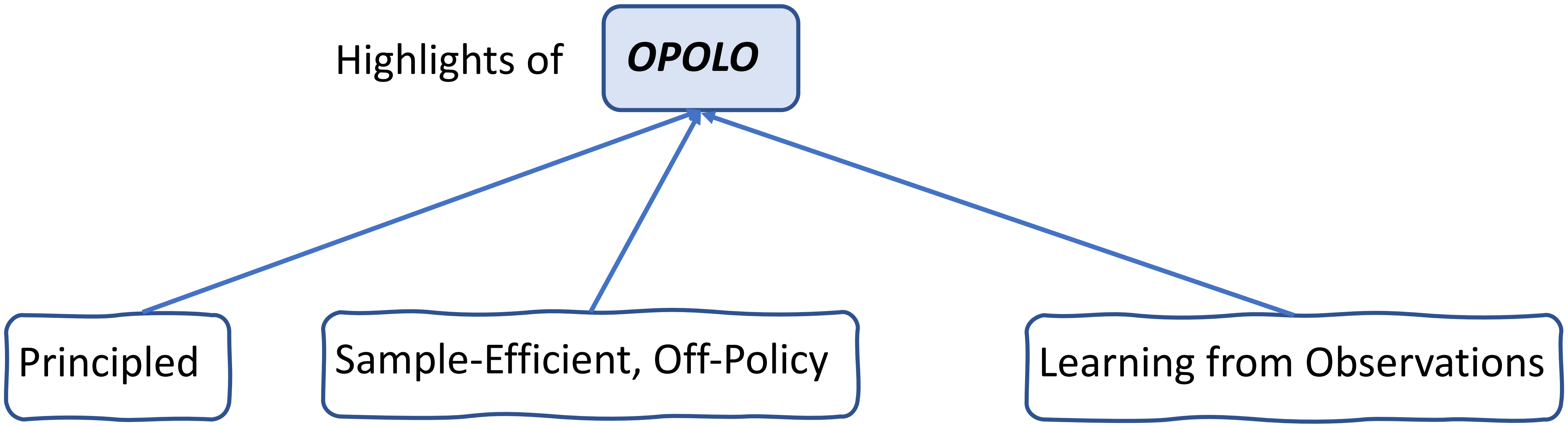
Learning policy: $\pi_\theta(a|s)$



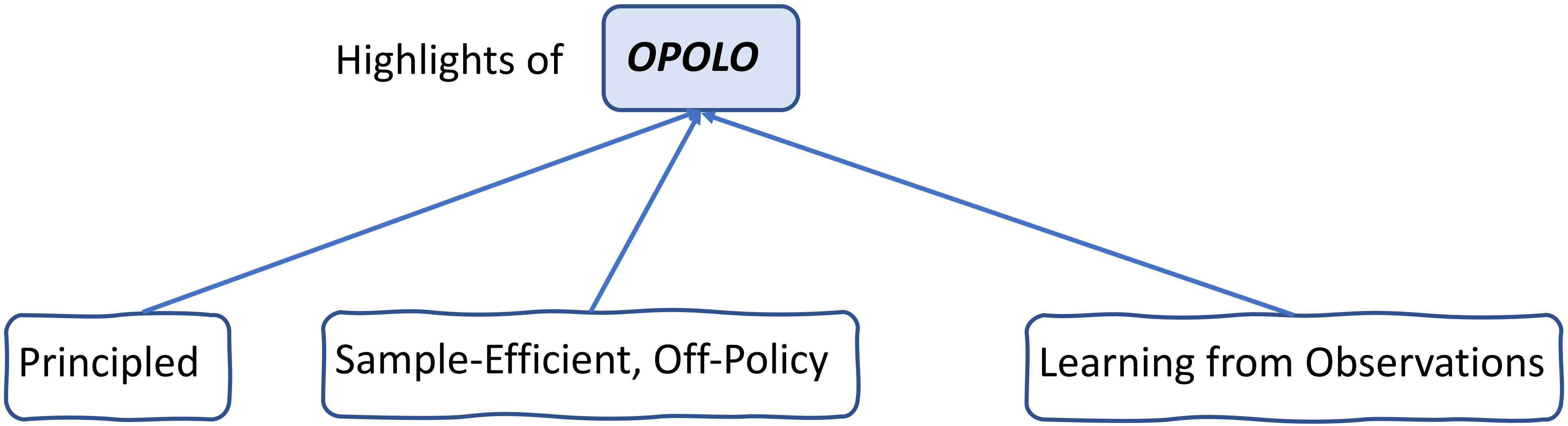
For ***off-policy*** learning, the agent can reuse samples from a replay buffer to speed up learning.

For ***on-policy*** learning, it requires that the behavior policy = target policy, so only ***fresh*** samples from the current policy can be used for training.

Proposed Approach: *OPOLO*
*O*ff-*Po*licy *L*earning from *O*bservations



*OPOLO: O*ff-*Po*licy *L*earning from *O*bservations



*OPOLO: O*ff-*Po*licy *L*earning from *O*bservations

- Upper-bound of the *Learning-from-Observation (Lf0)* Objective:

$$\mathbb{D}_{\mathbf{KL}}\left[\mu^{\pi}(s,s')||\mu^{E}(s,s')\right] \leq \mathbb{E}_{\mu^{\pi}(s,s')}\left[\log\frac{\mu^{R}(s,s')}{\mu^{E}(s,s')}\right] + \mathbb{D}_{\mathbf{KL}}\left[\mu^{\pi}(s,a)||\mu^{R}(s,a)\right]. \quad (4)$$

- Surrogate Objective:

$\mathbb{D}_{\mathbf{KL}}[P||Q] \leq \mathbb{D}_f[P||Q]$ When $f = \frac{1}{2}x^2$

$$\boxed{\min_{\pi} J_{\text{opolo}}(\pi) := \mathbb{E}_{\mu^{\pi}(s,s')}\left[\log\frac{\mu^{R}(s,s')}{\mu^{E}(s,s')}\right] + \mathbb{D}_f[\mu^{\pi}(s,a)||\mu^{R}(s,a)]. \quad (6)}$$

*OPOLO: O*ff-*Po*licy *L*earning from *O*bservations

- Surrogate Objective:

$$\boxed{\min_{\pi} J_{\text{opolo}}(\pi) := \mathbb{E}_{\mu^{\pi}(s,s')}\left[\log \frac{\mu^R(s,s')}{\mu^E(s,s')}\right] + \mathbb{D}_f[\mu^{\pi}(s,a)||\mu^R(s,a)]. \qquad (6)}$$

*OPOLO: O*ff-*Po*licy *L*earning from *O*bservations

How to enable Off-Policy Optimization ?

- Surrogate Objective:

Still On-Policy Distribution

Even more complicated with the extra D_f divergence

$$\min_{\pi} J_{\text{opolo}}(\pi) := \mathbb{E}_{\boxed{\mu^{\pi}(s,s')}}\left[\log \frac{\mu^{R}(s,s')}{\mu^{E}(s,s')}\right] + \boxed{\mathbb{D}_f[\mu^{\pi}(s,a)||\mu^{R}(s,a)].} \tag{6}$$

*OPOLO: O*ff-*Po*licy *L*earning from *O*bservations

Objective can be off-policy optimized ! 😉

- Surrogate Objective:

$$\min_{\pi} J_{\text{opolo}}(\pi) := \mathbb{E}_{\mu^{\pi}(s,s')}\left[\log \frac{\mu^{R}(s,s')}{\mu^{E}(s,s')}\right] + \mathbb{D}_f[\mu^{\pi}(s,a)||\mu^{R}(s,a)]. \tag{6}$$

$$= (1-\gamma)\mathbb{E}_{s_0\sim p_0, a_0\sim\pi(\cdot|s_0)}[Q(s_0,a_0)] + \mathbb{E}_{(s,a)\sim\mu^{R}(s,a)}[f_*((\mathcal{B}^{\pi}Q - Q)(s,a))].$$

*OPOLO: O*ff-*Po*licy *L*earning from *O*bservations

Objective can be off-policy optimized ! ☺

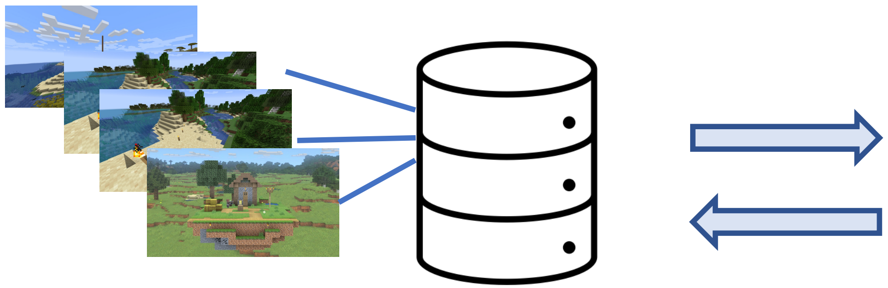
- Surrogate Objective:

$$\min_{\pi} J_{\text{opolo}}(\pi) := \mathbb{E}_{\mu^{\pi}(s,s')}\left[\log \frac{\mu^{R}(s,s')}{\mu^{E}(s,s')}\right] + \mathbb{D}_f[\mu^{\pi}(s,a) || \mu^{R}(s,a)]. \quad (6)$$

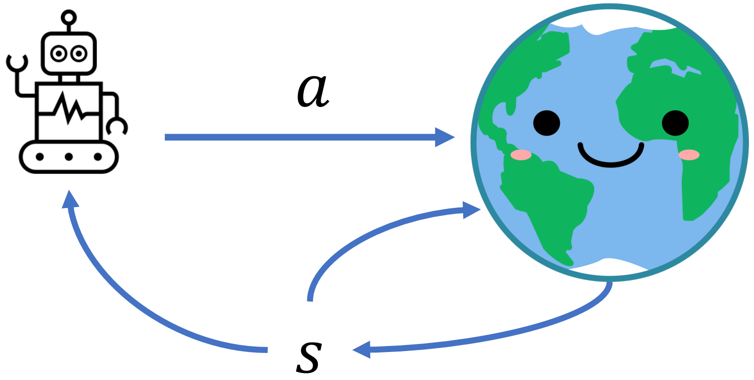
$$= (1-\gamma)\underbrace{\boxed{\mathbb{E}_{s_0\sim p_0, a_0\sim\pi(\cdot|s_0)}}}_{\text{Initial Distribution}}[Q(s_0,a_0)] + \underbrace{\boxed{\mathbb{E}_{(s,a)\sim\mu^{R}(s,a)}}}_{\text{Off-policy Distribution}}[f_*((\mathcal{B}^{\pi}Q - Q)(s,a))].$$

*OPOLO: O*ff-*Po*licy *L*earning from *O*bservations

Objective can be off-policy optimized !



Learning policy: $\pi_\theta(a|s)$



For ***off-policy*** learning, the agent can reuse samples from a replay buffer to speed up learning.

For ***on-policy*** learning, it requires that the behavior policy = target policy, so only ***fresh*** samples from the current policy can be used for training.

To Learn Even Faster: Policy Regularization as Forward Distribution Matching

Difference between ***inverse*** and ***forward*** imitation learning by distribution matching:

teacher distribution

$$\min_{\pi} D_{KL}\left[\boldsymbol{\mu}^{\boldsymbol{E}}(\cdot)\|\mu^{\pi}(\cdot)\right]$$

forward matching:

learning agent distribution

$$\min_{\pi} D_{KL}\left[\mu^{\pi}(\cdot)\ \|\boldsymbol{\mu}^{\boldsymbol{E}}(\cdot)\right]$$

Inverse matching:

The proposed objective optimizes (an upper-bound of) the ***inverse*** KL-divergence:

$$\min_{\pi} J_{\text{opolo}}(\pi) := \mathbb{D}_{\mathbf{KL}}\left[\mu^{\pi}(s,s')||\mu^{E}(s,s')\right]$$

To Learn Even Faster: Policy Regularization as Forward Distribution Matching

The proposed objective optimizes (an upper-bound of) the ***inverse*** KL-divergence:

$$\min_{\pi} J_{\text{opolo}}(\pi) := \mathbb{D}_{\mathbf{KL}}\left[\mu^{\pi}(s, s') || \mu^{E}(s, s')\right]$$

We can combine it with a ***forward*** distribution matching objective to speed up learning:

$$\mathbb{D}_{\mathbf{KL}}[\pi_E(a|s)||\pi(a|s)] = \mathbb{D}_{\mathbf{KL}}[\mu^E(s'|s)||\mu^{\pi}(s'|s)] + \mathbb{D}_{\mathbf{KL}}[\mu^E(a|s,s')||\mu^{\pi}(a|s,s')],$$

OPOLO In A Nutshell