# CHALLENGE:


Let us build “expert” on the fly

## 1. Policy Evaluation via (Online Bellman Residual)

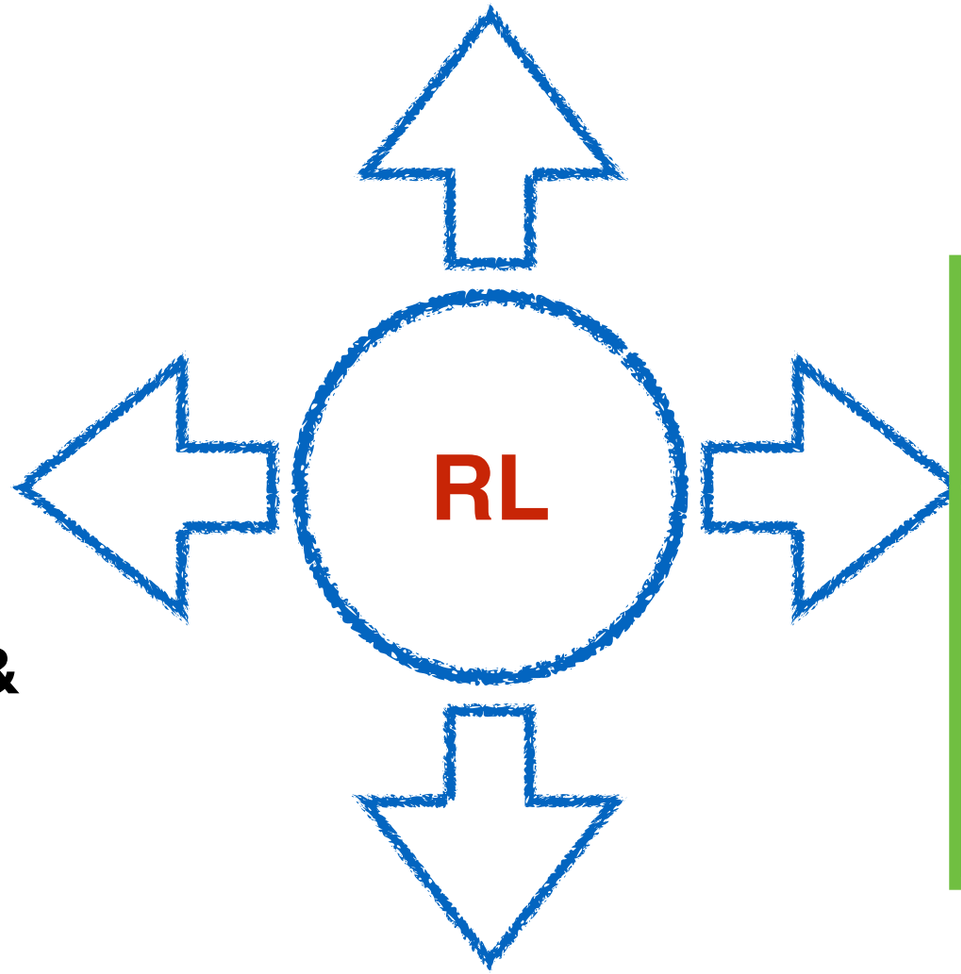
[Sun & Bagnell, 15, UAI (Best Student Paper)]

**Function Approximation**

## 2. RL via Imitation (Imitation Learning)

[Sun et.al 17, ICML; 18, ICLR]

**Function Approximation & Imitation**




## 3. RL via Indirect Imitation (Dual Policy Iteration)

[Sun et.al, 18, submitted to ICML]

**Function Approximation Optimal Control**

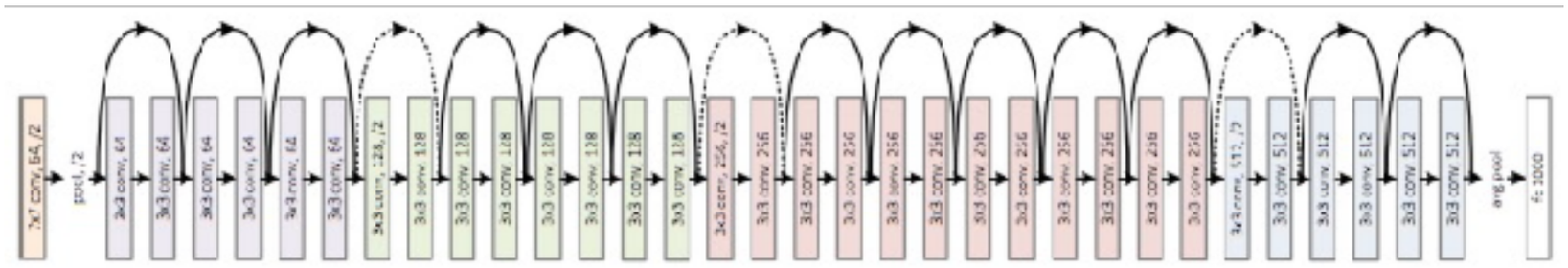
## Proposed Work: Temporal Difference Learning & Apprenticeship Learning

# Example: AlphaGo-Zero

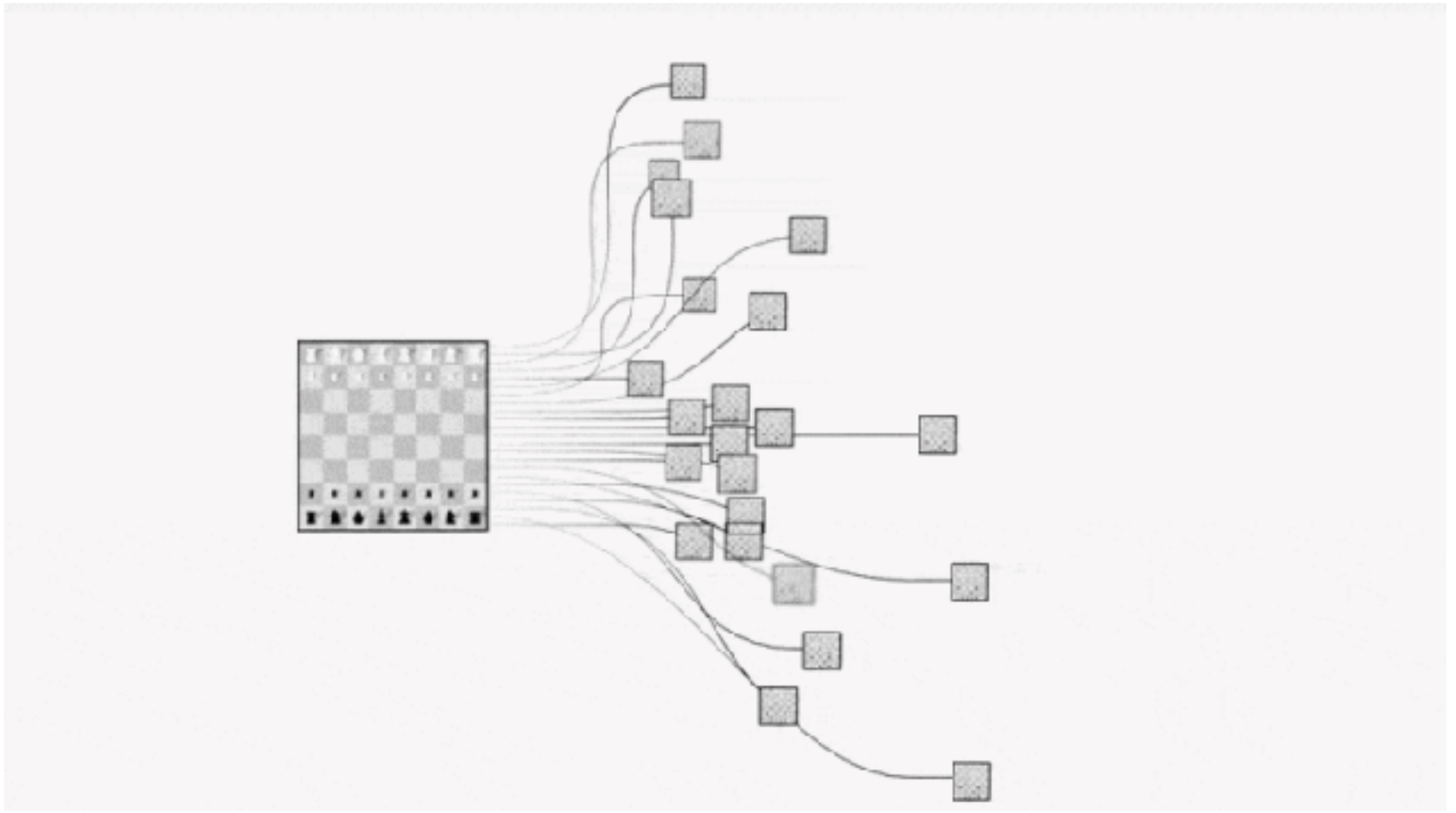
[Silver, et.al, 17, Nature]

**known** & **deterministic** model

A fast Policy $\pi$



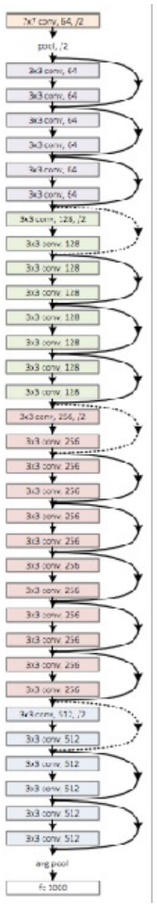
A slow Policy $\eta$ (MCTS)



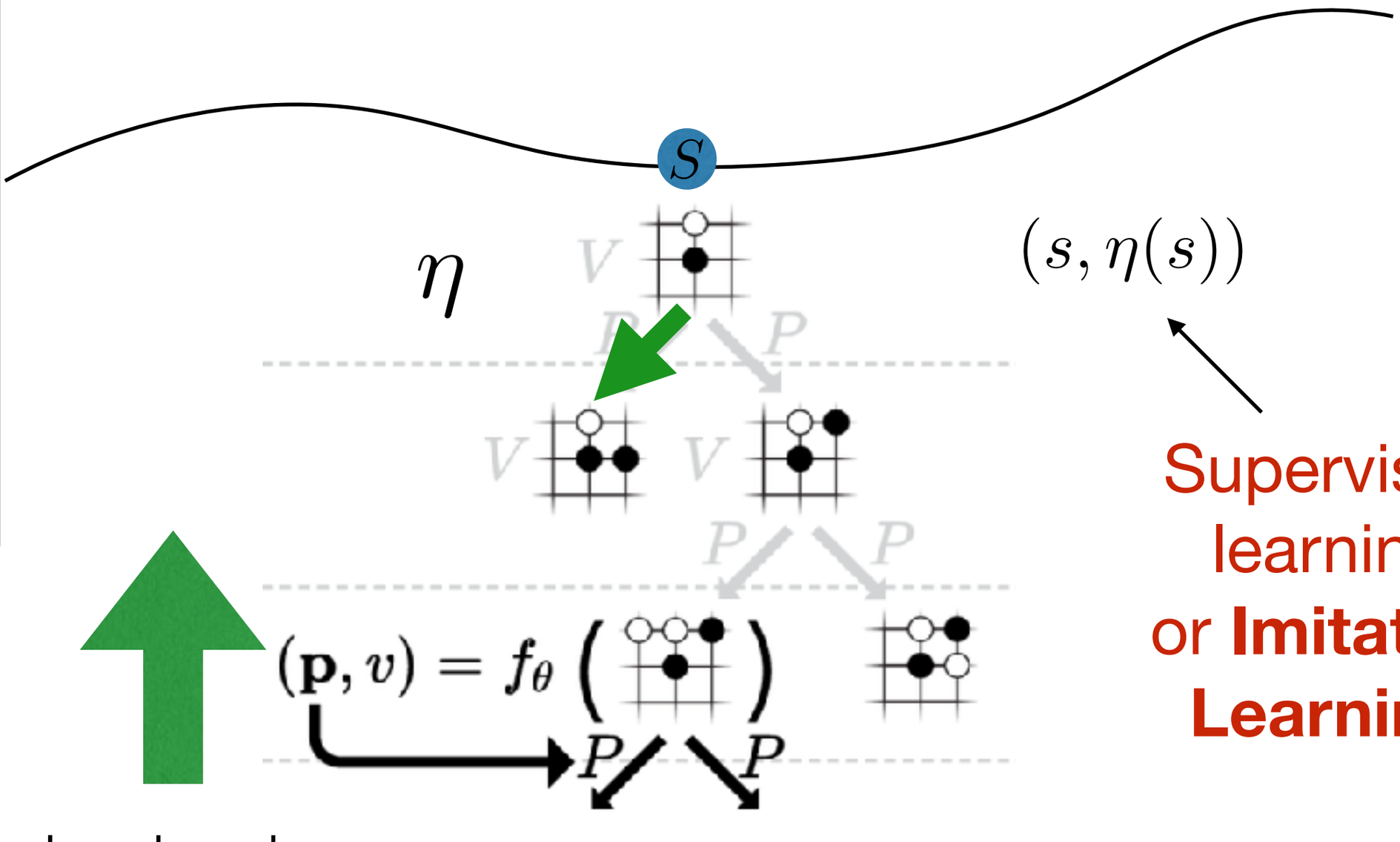
# Example: AlphaGo-Zero

[Silver, et.al, 17, Nature]

**known** & **deterministic** model



$\pi$




$(s, \eta(s))$

Supervised learning or **Imitation Learning**

At leaf, value backup with $V^\pi$

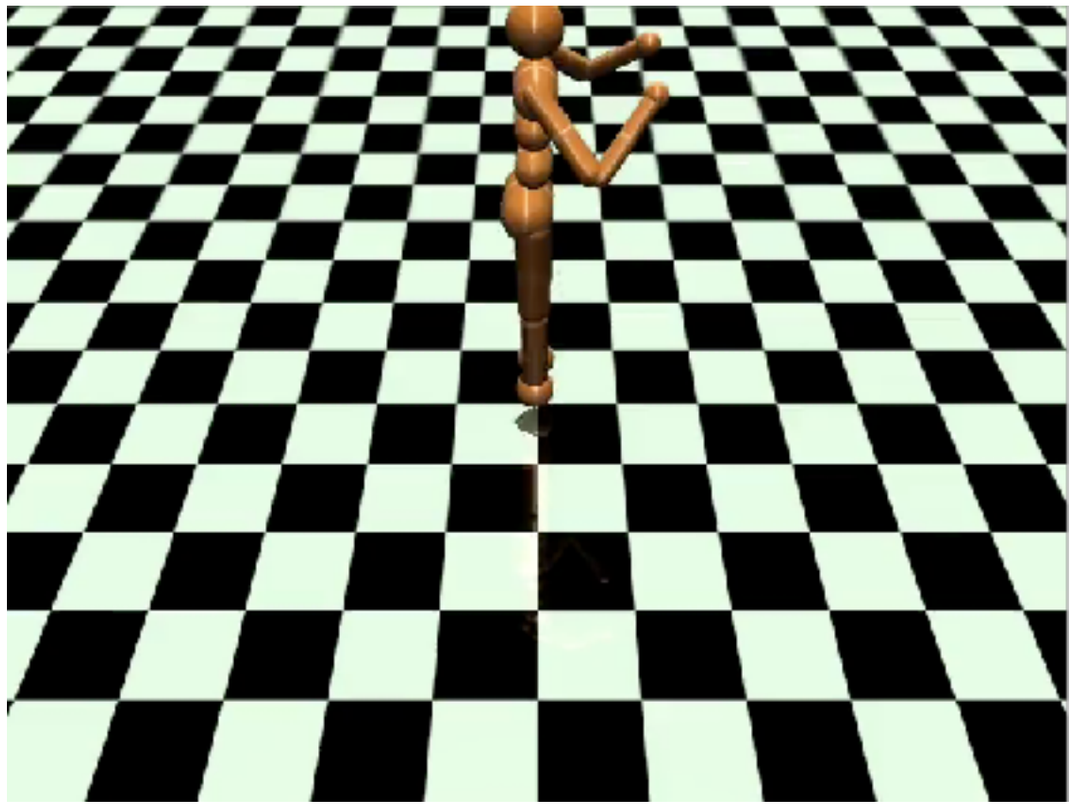
$\eta$: Forward Search + $\pi$

# CHALLENGE:

## What if model is unknown

$$\hat{P}(\cdot|s,a) \approx P(\cdot|s,a), \forall s \in \mathcal{S}, a \in \mathcal{A}$$

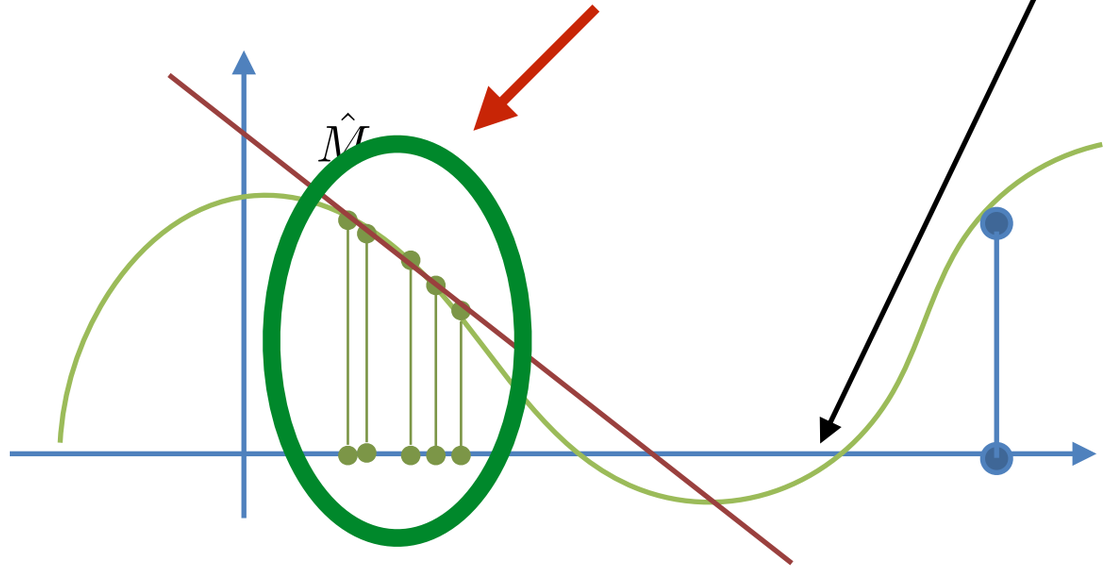
Not realistic



[openAI Gym]

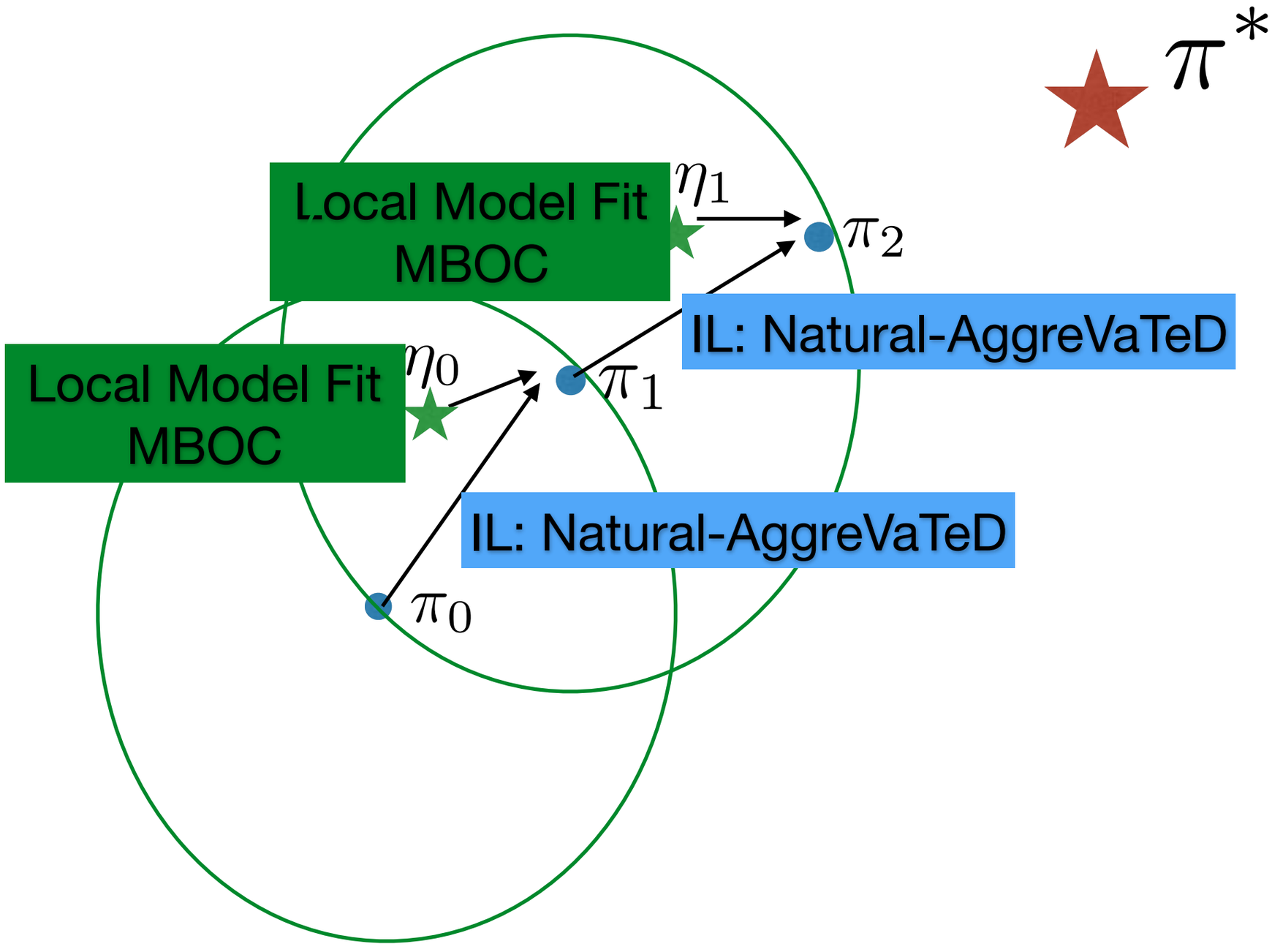
Model-based RL (e.g., iLQR)

within a Trust-Region



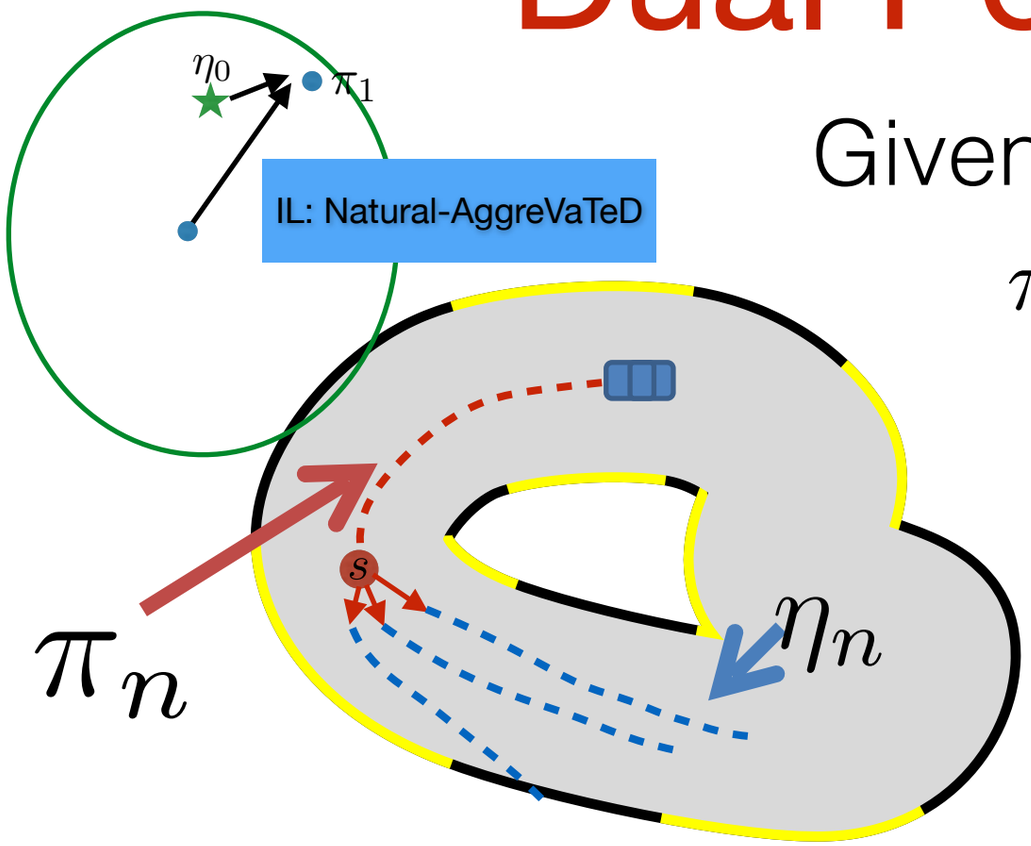
# Dual Policy Iteration

[Sun et.al, 18, submitted to ICML]



[Sun et.al, 18, submitted to ICML]

# Dual Policy Iteration




Given "expert" $\eta_n$ :

$$\pi_n \rightarrow \pi_{n+1}$$

$$\left\{ s, \begin{bmatrix} Q^{\eta_n}(s, a_1) \\ \dots \\ Q^{\eta_n}(s, a_A) \end{bmatrix} \right\}_N$$

$$\pi_{n+1} = \arg\min_{\pi \in \Pi} \sum_a \pi(a|s) Q^{\eta_n}(s, a)$$

$$s.t., \sum_s \|\pi(\cdot|s) - \pi_n(\cdot|s)\|_1 \leq \beta$$

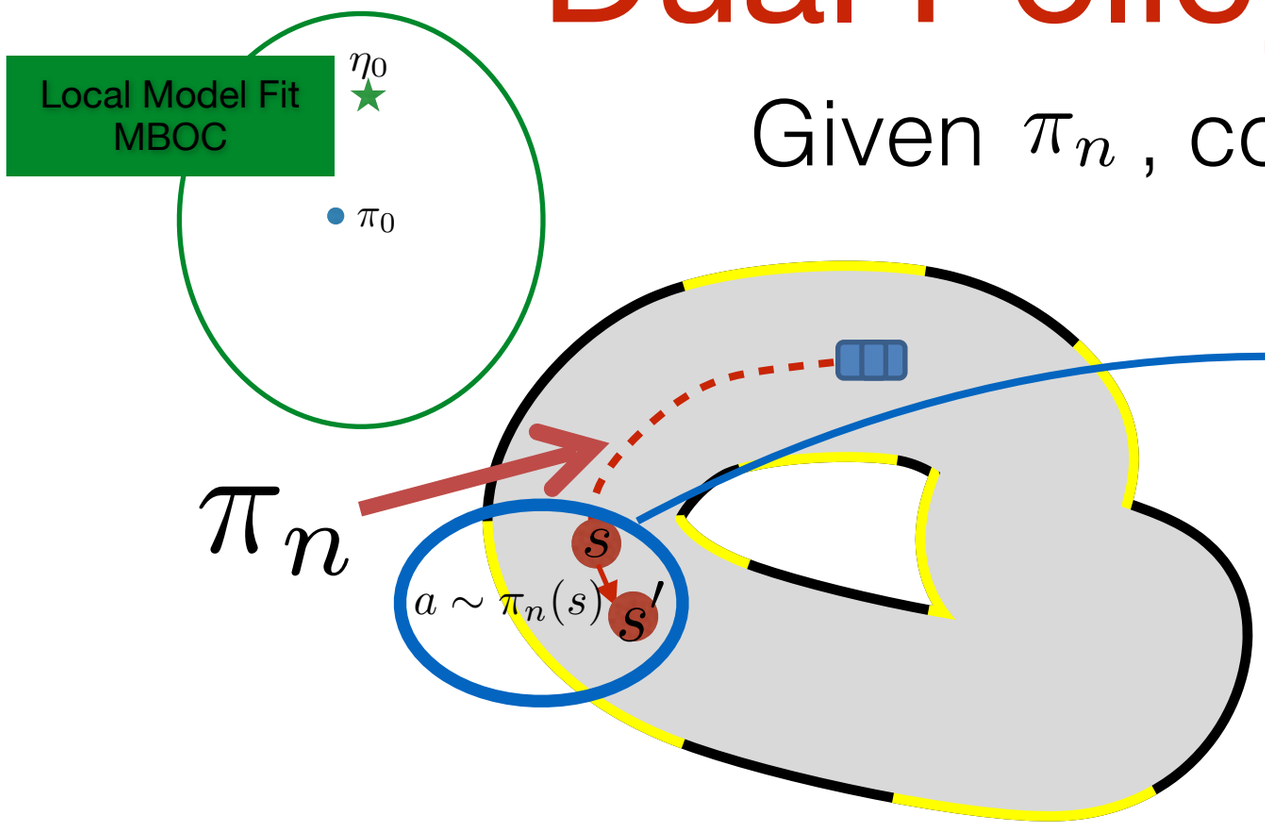
Cost-sensitive Classification (s.t. L1 constraint)

AggreVaTeD-NG [Sun et.al, 17, ICML]
Frank-Wolf Update (CPI) [KL02,ICML]

[Sun et.al, 18, submitted to ICML]

# Dual Policy Iteration

Local Model Fit
MBOC




Given $\pi_n$ , compute $\eta_n$

$$\{s, a, s'\}_N$$

$$\pi_n \rightarrow (s, a), s' \sim P(\cdot|s, a)$$

$$\min_{\hat{P}} \sum_{s,a,s'} -\log(\hat{P}(s'|s, a))$$

Local Model Fit

=

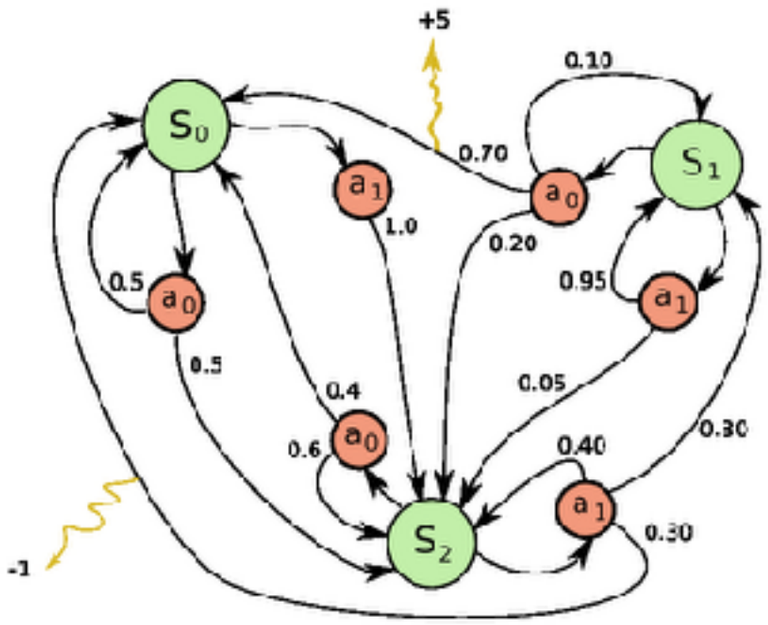
Maximum Likelihood Estimation

$$\eta_n = \text{MBOC Solver}(\hat{P})$$

$$s.t., \sum_s \|\pi_n(\cdot|s) - \eta(\cdot|s)\|_1 \leq \alpha$$

# Experiments

## Synthetic Discrete MDPs

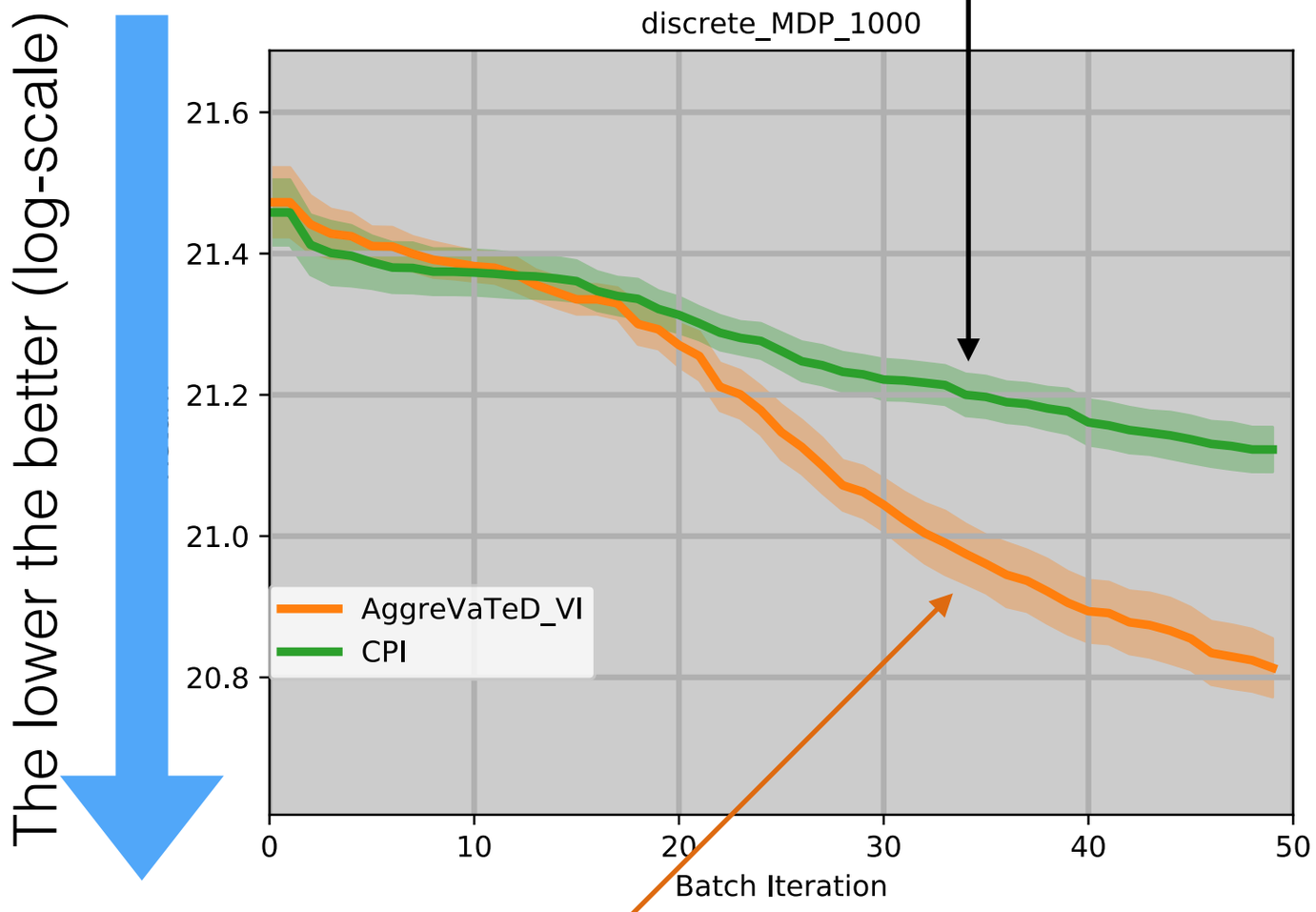



## Garnet Problems

[Scherrer 14, ICML]

## Conservative Policy Iteration

[Kakade&Langford, 02]

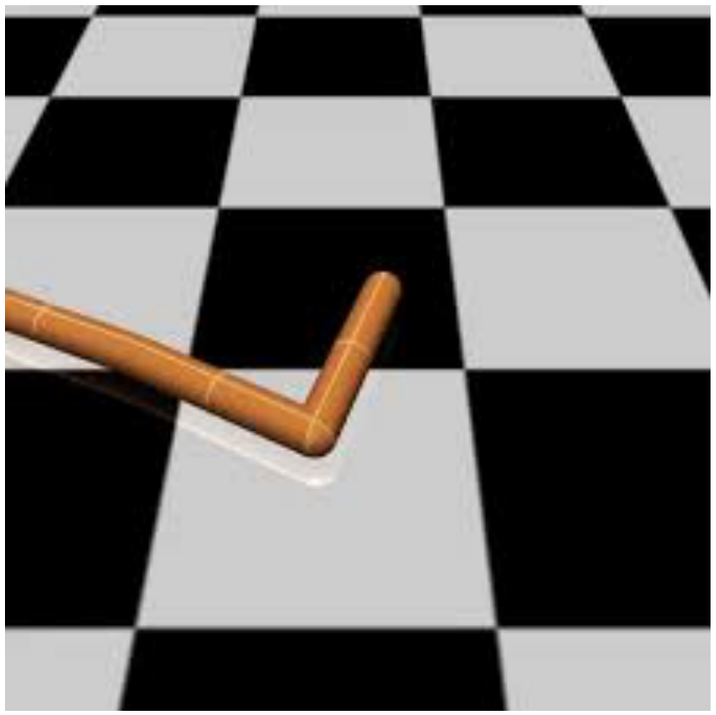



Our Approach: Frank-Wolf + VI

# Experiments

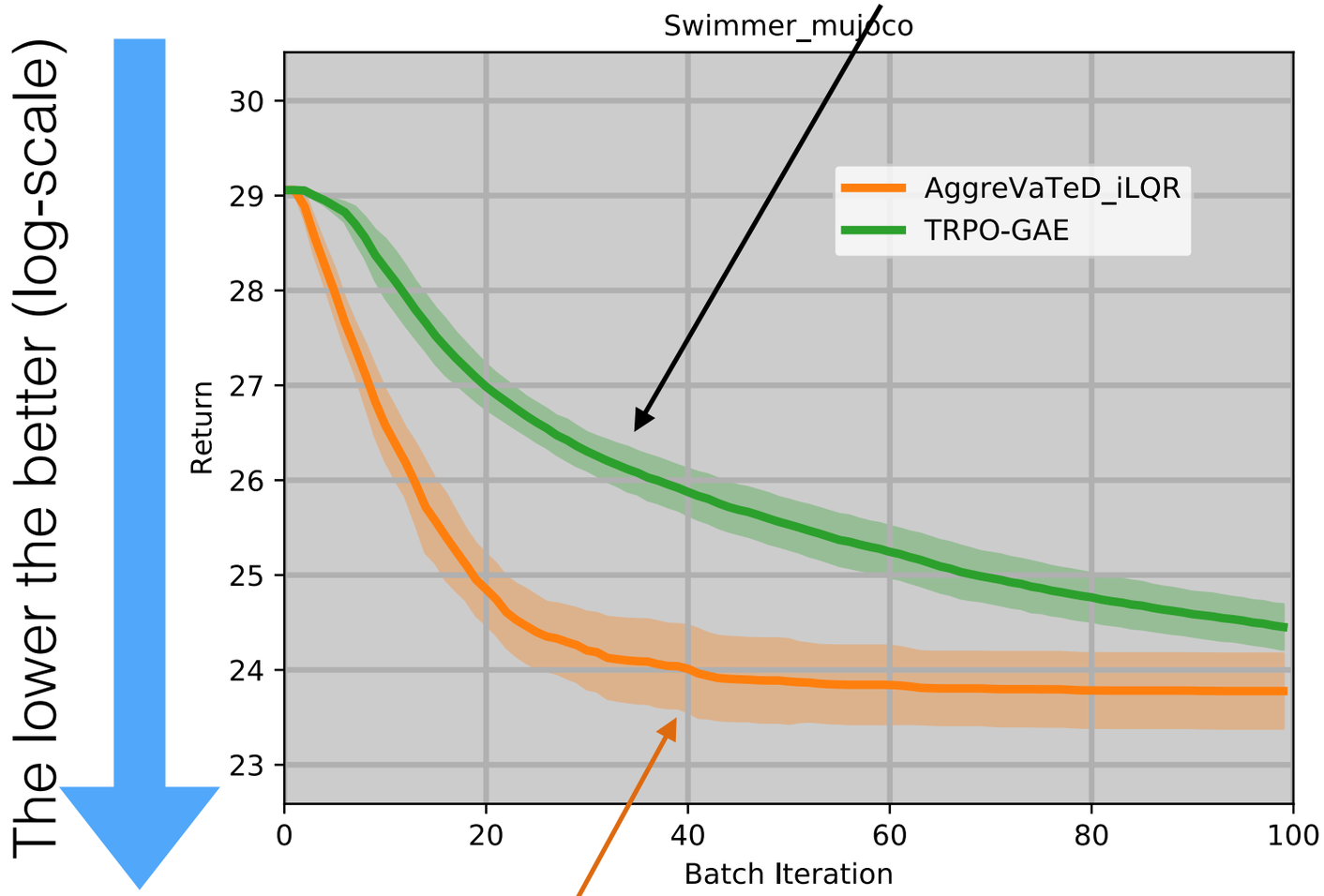
## Swimmer from MuJoCo

[Todorov et.al, 12]



## TRPO-GAE (Actor-Critic)

[Schulman et.al, 15]




## Our Approach (Natural Gradient +iLQG)

[Li & Todorov, 05]

# Summary

Keep Imitating

Easy Credit Assignment
Direct Supervised Learning
Less Random Exploration

AND

Tractable Local Model Fit
Efficient MBOC solver

Build Experts