CHALLENGE:

EXPERT

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

\( \eta \): Forward Search + \( \pi \)

At leaf, value backup with \( V^\pi \)

Supervised learning or Imitation Learning
CHALLENGE:
What if model is unknown

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

Not realistic

Model-based RL (e.g., iLQR) within a Trust-Region
Dual Policy Iteration

[Sun et.al, 18, submitted to ICML]
Dual Policy Iteration

Given “expert” $\eta_n$:

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

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

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

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

Given $\pi_n$, compute $\eta_n$

\[ \{ s, a, s' \} \sim \mathcal{N} \]

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

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

\[ \text{s.t., } \sum_s ||\pi_n(\cdot|s) - \eta(\cdot|s)||_1 \leq \alpha \]

\[ \min \hat{P} \sum_{s,a,s'} - \log(\hat{P}(s'|s, a)) \]

Local Model Fit

Maximum Likelihood Estimation
Experiments

Synthetic Discrete MDPs

Garnet Problems
[Scherrer 14, ICML]

Conservative Policy Iteration
[Kakade&Langford, 02]

The lower the better (log-scale)

Our Approach: Frank-Wolf + VI
Experiments

Swimmer from MuJoCo
[Todorov et.al, 12]

Our Approach
(Natural Gradient + iLQG)

[Li & Todorov, 05]
Summary

Keep Imitating AND Build Experts

Easy Credit Assignment
Direct Supervised Learning
Less Random Exploration

Tractable Local Model Fit
Efficient MBOC solver