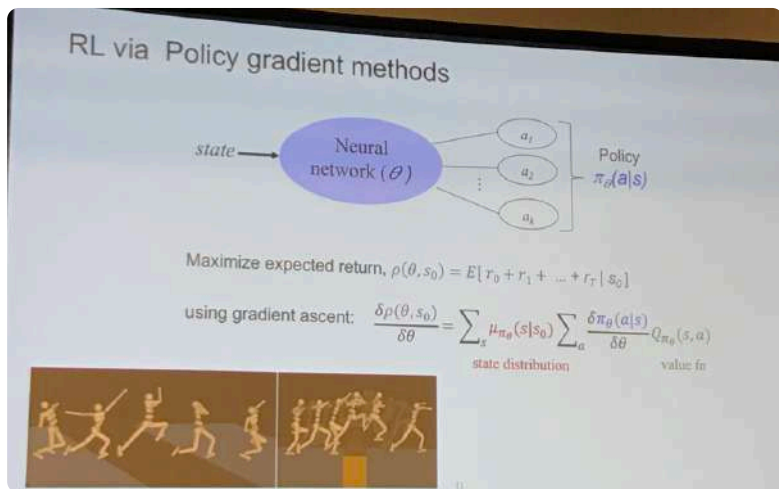


DAY 3 - NIPS 2018 - MAXIN DAY 2

Talk 1 - Joelle Pineau (FAIR) - Reproducible, Reusable and Robust RL

- Reproducibility = ability to duplicate
- Reusability = using same materials
- Robustness = min. necessary addition
- Deep RL that Matters (Attal, don't)

Nature (2016)
452% of researchers
state that there is
a crisis



- Baseline comparison:
 - always logrel

- TRPO, PPO, DDPO, ACKTR

↳ different results for different
domains / for different
implementations /
different network
architectures / random seeds!

- Underfitting the baseline! Overfitting your approach!

- Fair comparisons \Rightarrow robust conclusions



Careful thinking / reporting of
experiments!

influence of
hardware!

add to submission process!

We surveyed 50 RL papers from 2018
(published at NeurIPS, ICML, ICLR)

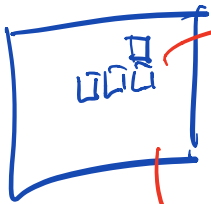
| | Yes: |
|--|------|
| Paper has experiments | 100% |
| Paper uses neural networks | 90% |
| All hyperparams for proposed algorithm are provided | 90% |
| All hyperparams for baselines are provided | 60% |
| Code is linked | 55% |
| Method for choosing hyperparams is specified | 20% |
| Evaluations on some variation of a hold-out test set | 10% |
| Significance testing applied | 5% |

How about a reproducibility checklist?

| | |
|--|--|
| For all algorithms presented, check if you include: | <input type="checkbox"/> A clear description of the algorithm. <input type="checkbox"/> An analysis of the complexity (time, space, sample size) of the algorithm. <input type="checkbox"/> A link to downloadable source code, including all dependencies. |
| For any theoretical claim, check if you include: | <input type="checkbox"/> A statement of the result. <input type="checkbox"/> A clear explanation of any assumptions. <input type="checkbox"/> A complete proof of the claim. |
| For all figures and tables that present empirical results, check if you include: | <input type="checkbox"/> A complete description of the data collection process, including sample size. <input type="checkbox"/> A link to downloadable version of the dataset or simulation environment. <input type="checkbox"/> An explanation of how sample were allocated for training / validation / testing. <input type="checkbox"/> An explanation of any data that was excluded. <input type="checkbox"/> The range of hyperparameters considered, method to select the best hyperparameter configuration, and specification of all hyperparameters used to generate results. <input type="checkbox"/> The exact number of evaluation runs. <input type="checkbox"/> A description of how experiments were run. <input type="checkbox"/> A clear definition of the specific measure or statistics used to report results. <input type="checkbox"/> Clearly defined error bars. <input type="checkbox"/> A description of results including central tendency (e.g. mean) and variation (e.g. std dev). <input type="checkbox"/> The computing infrastructure used. |

also
specify!

- RL: train ⊕ test on same env / dataset set → generalization ⇒ AGI
 - ↳ somewhere in between: use different seeds ↳ different tests for train / test!
 - ↳ simulators never can depict complexity of real world!
- simulator is convenient for reproducibility!
- Zhang, Rollins, Pincus (2018) ⇒ active defense!



RL agent never around things until sure, the inputs classifiable ⇒ correct? : get reward → desired

train on images / test on others!

→ Video background for AIE envs, Photo realistic simulators



- Finer works on neural netw. with help of RL
 - ↳ Go into real world!
 - ↳ No exploration usually possible!
 - ↳ very slow data acquisition

• ICLR reproducibility challenge

1st Oral / Spotlight Session - RL and Neuroscience

Why is exploration more challenging with a misspecified state space?

- All existing methods known to **efficiently** balance exploration and exploitation in RL with **theoretical guarantees** rely on the **optimism in the face of uncertainty** principle
- All such methods **fail to learn** when the state space is **misspecified**

$a_0, r_0 = 0$

$a_0, r_0 = 1 = r_{\max}$

$a_1, r_1 = \frac{1}{2}$

Not reachable from s

Optimism

Example 1 of Ortner [2008]

a_0, a_1 bef. time \Rightarrow linear regret!

Exploration-exploitation in RL with Misspecified State Space - R. Fruit

Oral 1: Exploration in Structured RL

\rightarrow structure via ϕ -approx \Rightarrow imposed on \mathcal{Q} - ϕ - fct. \rightarrow optimally exploits!

This Work: Analysis on Structured RL

- How can we **optimally exploit** a known structure?
 - Minimizing regret (accumulated reward loss due to the need of learning)

Theorem 1. Regret fundamental limits for a given arbitrary structure
Semi-infinite linear programming (LP) yields the minimal exploration rates η^*

Theorem 2. Potential regret reduction when leveraging the structure
Lipchitz regret $O(S_{\text{Lip}} A_{\text{Lip}} \log T)$, $S_{\text{Lip}} A_{\text{Lip}} = O(1)$ in reasonable scenarios

Theorem 3. An algorithm achieving minimal regret
Directed Exploration Learning (DEL); drives the exploration rates towards η^*

\rightarrow structured MDP

$$\phi \in \mathbb{F}$$

MDP

structure

continuously reward/
transition

\Rightarrow E.g. Lipschitz left

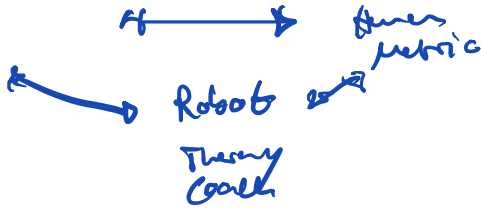
\hookrightarrow How to exploit the structure?!

\hookrightarrow convex under

Table 2 - A. Howard - Human-x1 Trust Phenomenon

- Trust \rightarrow belief truster that trustee mitigates truster's risk
 - \rightarrow Human-robot interaction: proximity \Rightarrow physiotherapy
 - \rightarrow Cognitive attention of learner/child \Rightarrow faster in!
 - \hookrightarrow bonding / emotional engagement \Rightarrow not eyes but sensors

Gamified
Therapy



- Goal: Build robot for effective physiotherapy sessions

\hookrightarrow Humans trust robots often more than humans!

Kinematic Model

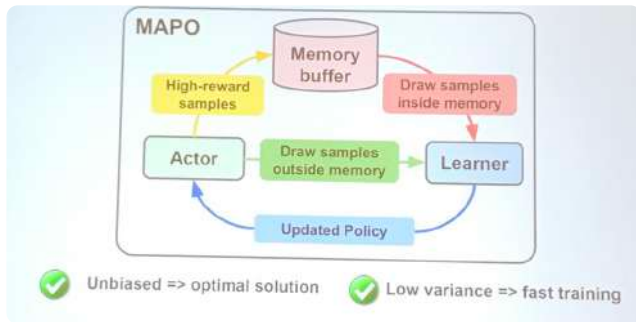
- Require a baseline for comparing measures with respect to a norm. We construct a **4 DOF model** that mimics the kinematics of the human arm.
- Generates an **optimal path** between two points in space as a function of:
 - User's arm's link lengths.
 - User's arm's initial pose.
 - Position of the target.
- Resulting trajectory is a curve that matches the **structure of the curve** generated by an individual's movements. [Morasso et al. 1981]

García-Vergas, Serrano, Chen, Howard, "Developing a Baseline for Upper-Body Motor Skill Assessment using a Robotic Kinematic Model," IEEE Ra-man, 2014.

Georgia Tech School of Interactive Computing

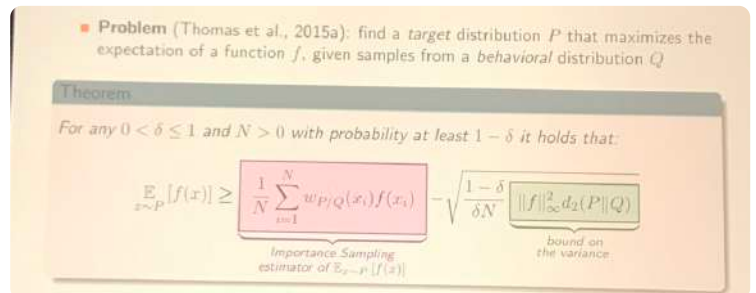
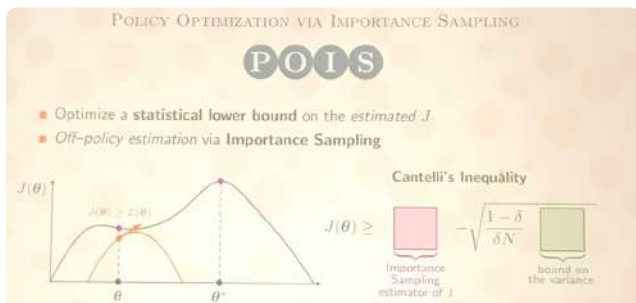
2nd oral / spotlight session - RL and Neuroscience

- Off-policy Evaluation \Rightarrow importance sampling
- Goal-dir. Molecular Graph generation \Rightarrow graph representation + RL
 - ↳ Graph-Conv. Nets
 - ↳ state: incomplete graph
 - ↳ Optimization via PPO
- Memory augmented PO \Rightarrow Semantic Policy / program synthesis
 - variance reduction
 - optimization + a div. ex. realistic
 - TASKS:
 - Generation (full)
 - Completion (partially)
 - program synthesis \rightarrow treat as latent



- variance reduction:
 - enumeration (inside memory)
 - ⊕ sample outside memory
- slower converge but better result

- Meta-RL for Structured Exploration Strategies
 - \rightarrow prior exposure should guide exploration \Rightarrow structured stochasticity!
 - \rightarrow latent conditioned policy
 - ↳ capture prior task distr.
 - Meta-learning in latent space
 - ↳ more conditioning \Rightarrow !
- Policy optimization via importance sampling (POIS)
 - \rightarrow policy search $\Rightarrow \theta^*$ \Rightarrow PPO - reasonable neighborhood
 - \rightarrow PPO \Leftrightarrow surrogate loss
 - \rightarrow statistical lower bound off-policy eval



- action vs. parameter-based → ^{demonstrations in experiments} → ^{robustness} ^{reliability}
- offline and optimization of network!
- bad performance in sparse rewards envs → ^{the trajectory} ^{dataset!}

• Bayesian Generative Adversarial Imitation Learning

Generative Adversarial Imitation Learning (GAIL)

- Use **generative adversarial networks (GANs)** for imitation learning:

$$\min_{\pi} \max_D \mathbb{E}_{\pi} \left[\sum_{t=1}^T \log D(s_t, a_t) \right] + \mathbb{E}_{\pi_E} \left[\sum_{t=1}^T \log(1 - D(s_t, a_t)) \right]$$

- Sample trajectories by using $\pi(a|s)$ and $\pi_E(a|s)$ (expert demonstrations).
- Train discriminator.
- Update policy $\pi(a|s)$ by using reinforcement learning (RL), e.g., TRPO, PPO.

- requires model-free inner loops
- ↔ sample inefficient
- ↳ train discriminator with Bayesian classification!

Bayesian Framework for GAIL

- Probabilistic model for trajectories
- For each trajectories $\tau = (s_1, a_1, s_2, a_2, \dots, s_T, a_T)$, a sequence of state-action pairs satisfies Markov property:

$$p(s_1, a_1) = p(s_1) \pi(a_1 | s_1),$$

$$p(s_{t+1}, a_{t+1} | s_t, a_t) = P_T(s_{t+1} | s_t, a_t) \pi(a_{t+1} | s_{t+1})$$
- Two policies: agent's policy $\pi_A(a|s)$, expert's policy $\pi_E(a|s)$

$z = (s, a)$

agent's trajectory: τ^A expert's trajectory: τ^E

Reinforcement Learning

- "Sequential decision making under uncertainty."
- Three necessary building blocks:
 - Generalization
 - Exploration vs. Exploitation
 - Credit assignment
- As a field, we are pretty good at combining any 2 of these 3 but we need practical solutions that combine them all.

We need effective uncertainty estimates for Deep RL.

Diagram illustrating the relationship between Data & Estimation, Supervised Learning, + partial feedback, Multi-armed Bandit, + delayed consequences, and Reinforcement Learning.

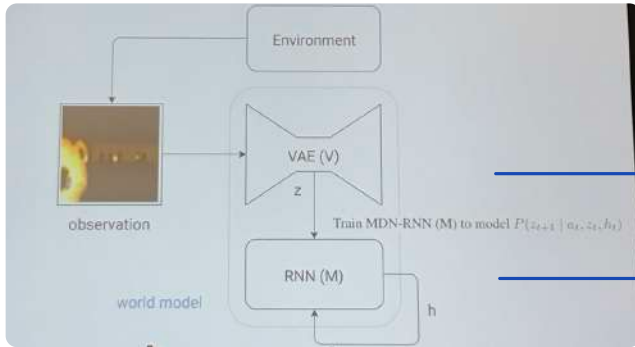
- Visual RL with Imagined Goals
 - autonomous goal setting → only visual input
 - latent representation → VAE → sample new latents used as goals
 - ↳ reward fct.: distance in latent space!
- Regularized prior fct. for DRL
 - address shortcoming

Estimating uncertainty in deep RL

| Dropout sampling | Variational inference | Distributional RL | Count-based density | Bootstrap ensemble |
|---|--|--|--|--|
| "Dropout sample + posterior sample" (Gal & Ghahramani 2015) | Apply VI to Bellman error as if it was an IID supervised loss | Model Q-value as a distribution, rather than point estimate | Estimate number of "visit counts" to state, with bonus | Train ensemble on noisy data - classic statistical procedure |
| Dropout rate does not concentrate with the data | Bellman error $D(s) = r + \gamma \max_{a'} Q(s, a') - Q(s, a)$ | This distribution is posterior uncertainty | The "density model" has nothing to do with the actual task | No explicit "prior" mechanism for "intrinsic motivation" |
| Even "concrete" dropout not necessarily right rate | VI on Q model does not propagate uncertainty | Asymptotic vs. Bayesian - is not the right thing for exploration | With generalization, state "visit counts" is uncertainty | If you've never seen a reward, why would the agent explore? |

- Playing hard exploration games by watching Ya-Tile
→ Sand in Atari games! → input signal!
- Recent World Models Facilitate Policy Evaluation
→ VAE: capturing \rightarrow density Estimator via RNN

\Rightarrow No end-to-end but separate training!



\hat{z}
↓
Feed into controller module \rightarrow agent action