UCI University of California, Irvine

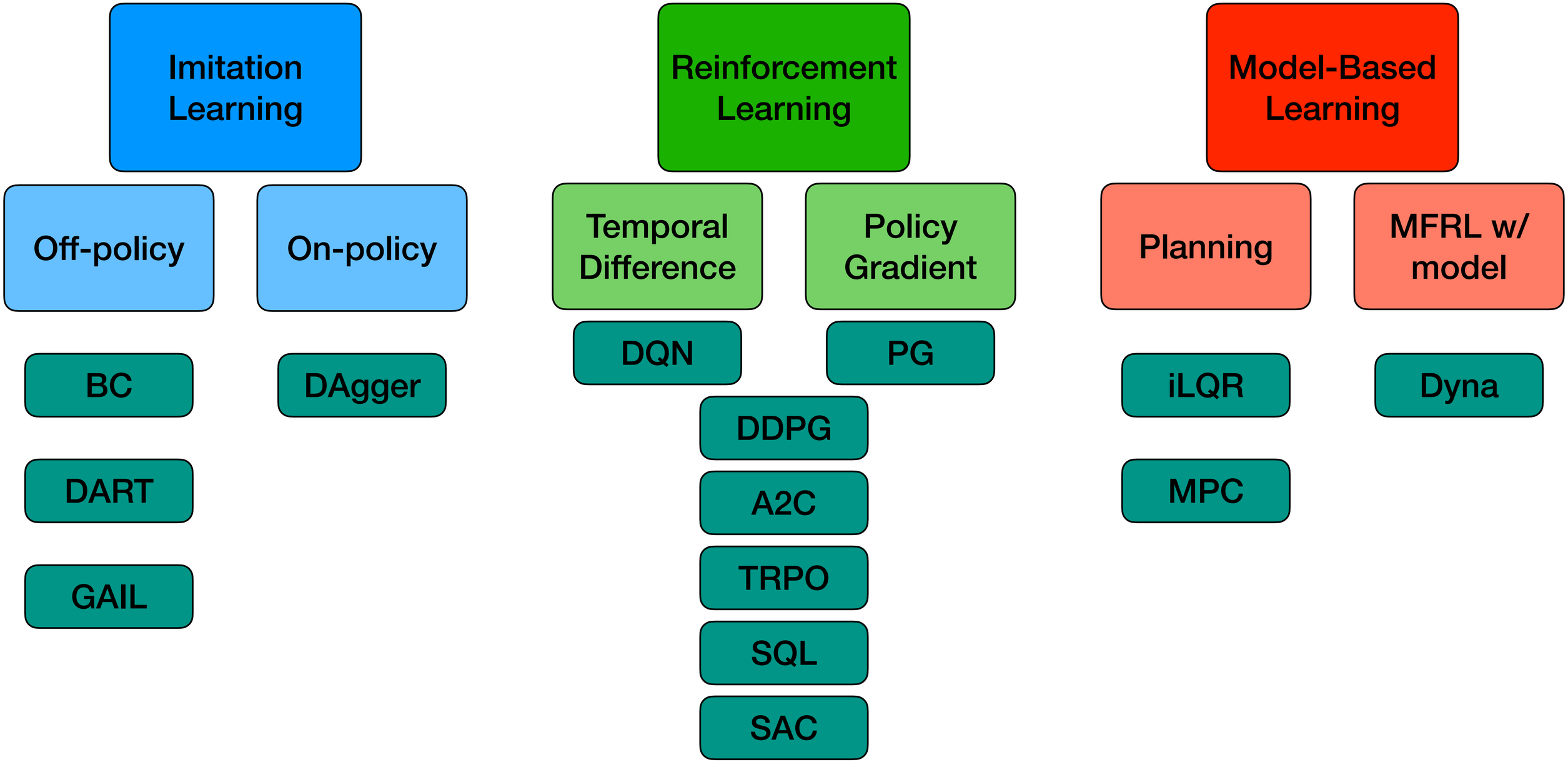
CS 277: Control and Reinforcement Learning

Winter 2021

Lecture 19: Open Questions

Roy Fox
Department of Computer Science
Bren School of Information and Computer Sciences
University of California, Irvine

Taxonomy


Imitation Learning
Reinforcement Learning
Model-Based Learning
Off-policy
On-policy
Temporal Difference
Policy Gradient
Planning
MFRL w/ model
BC
DAgger
DQN
PG
iLQR
Dyna
DDPG
DART
A2C
MPC
TRPO
GAIL
SQL
SAC

Bounded RL

- Back to the general case: $\max_\pi \mathbb{E}_{s,a\sim p_\pi}[\beta r(s,a)] - \mathbb{D}[\pi \| \pi_0]$
- Define an entropy-regularized Bellman optimality operator

$$\mathcal{B}[V](s) = \max_\pi \mathbb{E}_{a|s\sim\pi}\left[r(s,a) - \frac{1}{\beta}\log\frac{\pi(a|s)}{\pi_0(a|s)} + \gamma\,\mathbb{E}_{s'|s,a\sim p}[V(s')]\right]$$

  - As in the unbounded case $\beta \to \infty$, this operator is contracting

- Optimal policy:

$$\pi(a|s) \propto \pi_0(a|s)\exp\beta(r(s,a) + \gamma\,\mathbb{E}_{s'|s,a\sim p}[V(s')]) = \pi_0(a|s)\exp\beta Q(s,a)$$

- Optimal value recursion:

$$V(s) = \frac{1}{\beta}\log Z(s) = \frac{1}{\beta}\log\mathbb{E}_{a|s\sim\pi_0}[\exp\beta(r(s,a) + \gamma\,\mathbb{E}_{s'|s,a\sim p}[V(s')])]$$

Soft Q-Learning (SQL)

- TD off-policy algorithm for model-free bounded RL
- With tabular parametrization:

$$\Delta Q(s,a) = r + \frac{\gamma}{\beta} \log \mathbb{E}_{a'|s' \sim \pi_0} [\exp \beta Q(s',a')] - Q(s,a)$$

- With differentiable parametrization:

$$\mathcal{L}_\theta(s,a,r,s') = \left(r + \frac{\gamma}{\beta} \log \mathbb{E}_{a'|s' \sim \pi_0} [\exp \beta Q_{\bar{\theta}}(s',a')] - Q_\theta(s,a) \right)^2$$

- As $\beta \to \infty$, this becomes (Deep) Q-Learning

Soft Actor–Critic (SAC)

- AC off-policy algorithm for model-free bounded RL

- Optimally:

$$\pi(a|s) = \frac{\pi_0(a|s) \exp \beta Q(s,a)}{\exp \beta V(s)} \qquad \forall a: \; V(s) = Q(s,a) - \frac{1}{\beta} \log \frac{\pi(a|s)}{\pi_0(a|s)}$$

- We can train the critic off-policy

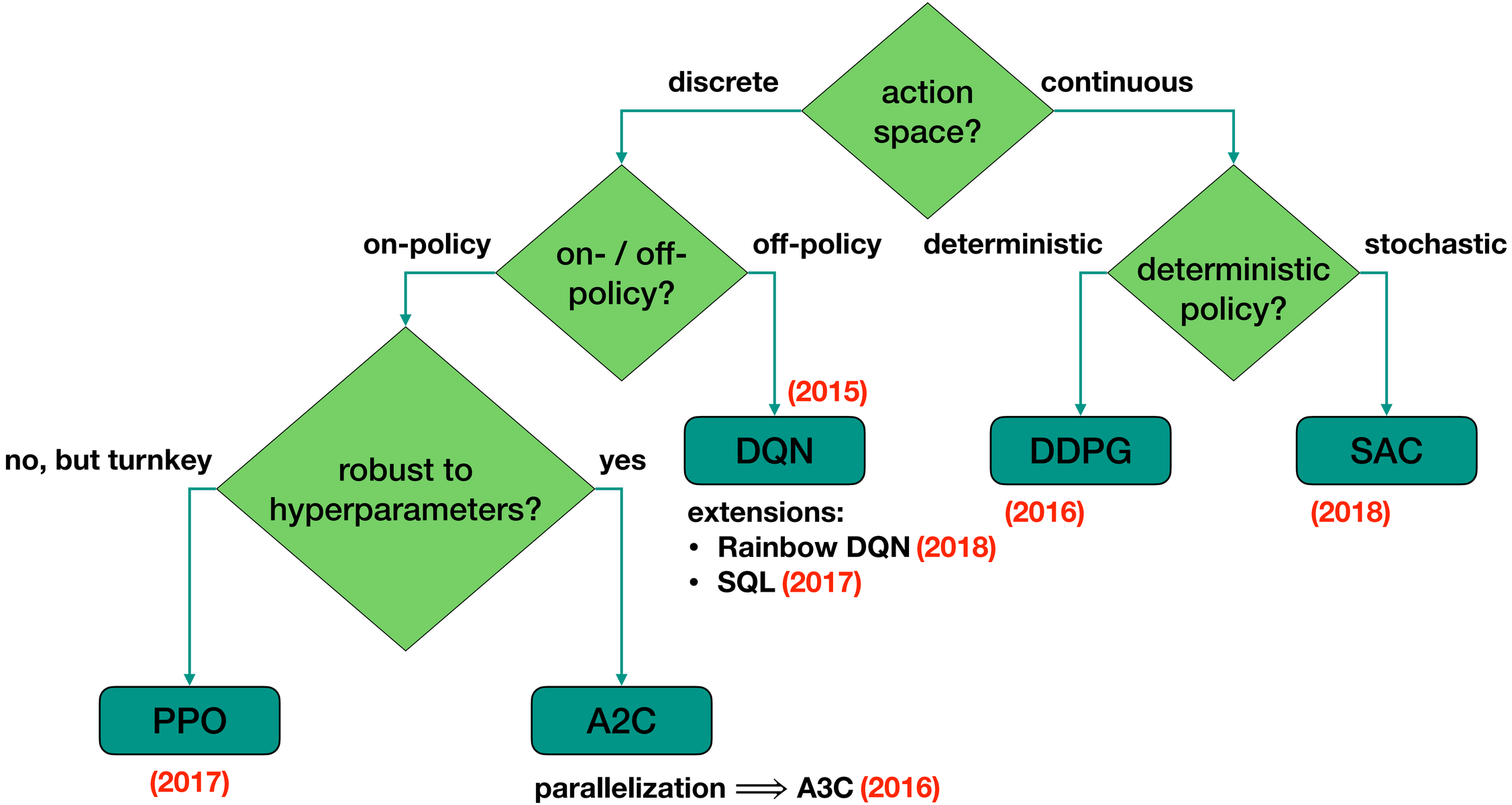
$$\mathcal{L}_\phi(s,a,r,s',a') = \left(r + \gamma \left(Q_{\bar{\phi}}(s',a') - \frac{1}{\beta} \log \frac{\pi_\theta(a'|s')}{\pi_0(a'|s')} \right) - Q_\phi(s,a) \right)^2$$

- And the actor to be soft-greedy = distill / imitate the critic

$$\mathcal{L}_\theta(s) = \mathbb{E}_{a|s\sim\pi_\theta}[\log \pi_\theta(a|s) - \log \pi_0(a|s) - \beta Q_\phi(s,a)]$$

- Allows continuous action spaces

Flowchart: which algorithm to choose?



On- or off-policy data?

- The faster our simulator $\Longrightarrow$ the faster we can refresh our data
  - And still keep sufficient diversity for training
- Fast enough $\Longrightarrow$ can use on-policy data
  - No need for replay buffer
  - No train$\rightarrow$test distributional mismatch (= covariate shift)
  - Can still use off-policy algorithms with on-policy data
- Extremely slow simulator $\Longrightarrow$ not even off-policy, just offline RL

Topics we covered

- Imitation learning
- Policy evaluation + improvement
  - Monte-Carlo vs. Temporal Difference
  - On- vs. off-policy
- Policy Gradient
  - Advantage estimation, Actor–Critic
- Optimal control
- Planning, model-based learning
- Partial observability
- Exploration
- Inverse RL
- Control as Inference
- Structured control
- Multi-task learning
- Multi-agent RL

Topics we didn't cover

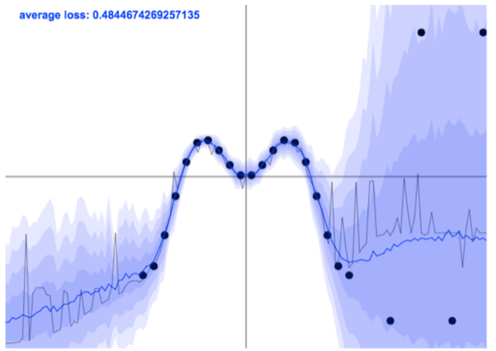
- Hindsight Experience Replay (HER)
- Eligibility traces
- Generalized Value Functions (GVF)
  - Successor representation
- Value Iteration / Prediction Nets (VIN / VPN)
- Natural policy gradient
  - Mirror descent
- Distributional RL
- Bayesian RL
- Hyperparameter tuning
- Distributed RL
- Robot learning
- Safety
- Curiosity + empowerment
- Preference elicitation
- Offline RL
- Meta-learning
- Lifelong learning

Trends and open questions in ML

- Bayesian Deep Learning
- Optimization theory
- Neuro-symbolic AI
- Meta-learning / learning to learn
- Lifelong learning
- Interpretability, explainability
- AI ethics: fairness, safety

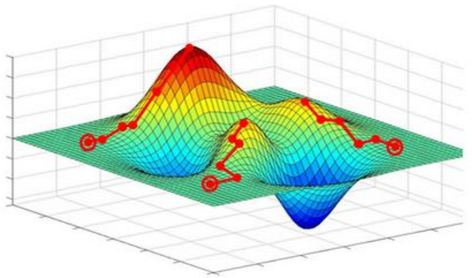
Bayesian RL

- Two kinds of uncertainty
  - Aleatoric = things I haven't seen / haven't happened yet: $p(s_t \mid m_t)$, $p(r_{t+k} \mid m_t)$, ...
  - Epistemic (= model uncertainty) = things I haven't modeled / learned yet: $\hat{p}$, π_θ,...
- Standard RL already considers aleatoric uncertainty
  - “Overtake truck quickly, to reduce time with partial observability, prob of crash”
- Bayesian RL can estimate epistemic uncertainty: $p(\theta \mid \mathcal{D})$
  - Can help improve exploration (cf. Thompson sampling)
  - Can improve learning in bounded agents (uncertain $Q \Longrightarrow$ winner's curse)



Optimization $\Longleftrightarrow$ RL

- Special considerations of optimization $\rightarrow$ RL:
  - Covariate shift
  - Temporal-Difference $\Longrightarrow$ non-stationary loss landscape
  - Saddle points in multi-agent RL
- RL $\rightarrow$ optimization: iterative optimization is a dynamical process
  - Gradient descent = maximize “reward” of descending loss landscape
  - Optimal control concepts (e.g. Langevin dynamics) key in analysis

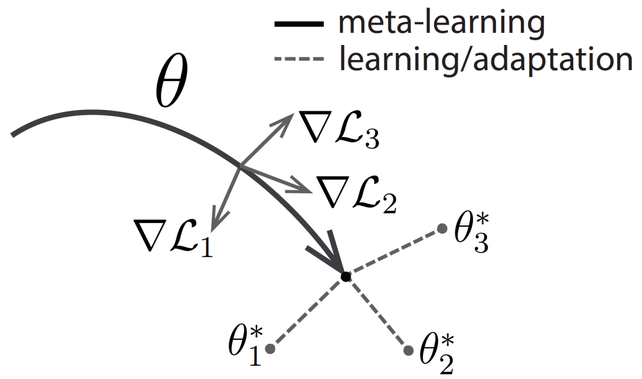


Neuro-symbolic RL

- Is there any benefit to discrete components in gradient-based methods?
  - E.g. modularity
- Structured control = discrete memory components
  - Can help sample efficiency, generalization, transfer, interpretability, ...
- How to learn under given structure?
- How to discover optimal structure?

Meta-learning $\Longleftrightarrow$ RL

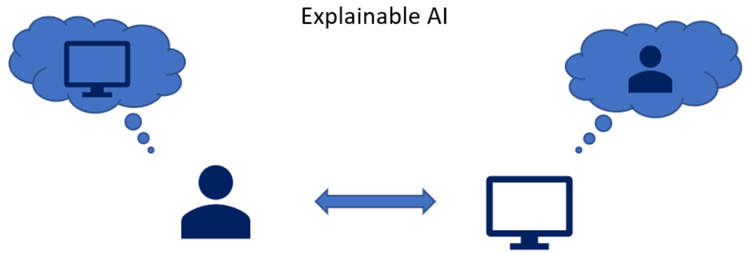
- Multi-task learning = transfer / share learning products between tasks
  - E.g. features, models, policies, skills
- Meta-learning = transfer / share learning of learner components
  - Network architecture = Neural Architecture Search (NAS)
  - Parameter initializations (MAML)
  - Optimizer hyperparameters

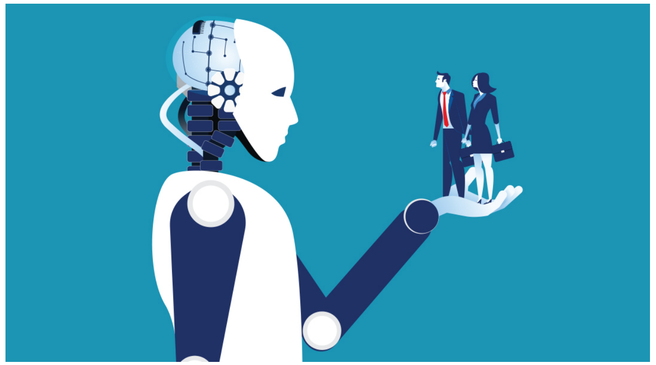


- Learning to perform sequence of tasks = sequential decision making
  - E.g. can use RNNs

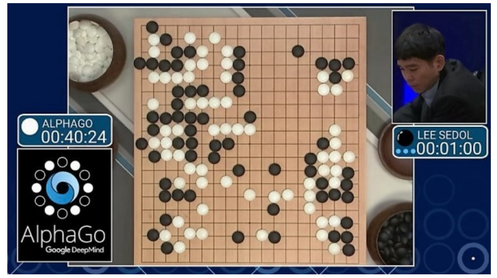
Trends and open questions in ML

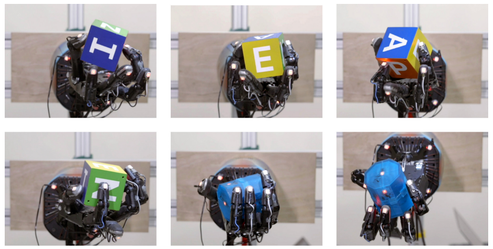
- Bayesian Deep Learning
- Optimization theory
- Neuro-symbolic AI
- Meta-learning / learning to learn
- Lifelong learning
- Interpretability, explainability
- AI ethics: fairness, safety





Reproducibility crisis





- Reinforcement learning has seen immense success
  - But remains largely irreproducible
- Many algorithms are very sensitive to hyperparameters
- Very sensitive to parameter initialization $\Longrightarrow$ need to average over many runs
- Small implementation details may have unexpected effects
- How to go beyond this pre-paradigmatic phase?
  - Better RL theory
  - Build practical RL (and ML) as experimental field

Other open questions

- Imitation learning / inverse RL
  - Discover structure / memory features in teacher demonstrations
- Control as Inference
  - How much “bounded” should the agent be? How to anneal this temperature?
- Structured control
  - Which structures can we discover? Which structures are useful for control?
- Multi-task learning
  - How to discover which tasks are related / unrelated?