CRIStAL
Centre de Recherche en Informatique,
Signal et Automatique de Lille

Take-home messages

Odalric-Ambrym Maillard

2021 – 2022

Inria Scool

Control theory

- Definition of Markov Decision Process
- Markov property
- Discount factor
- Discounted value, Finite time horizon value.
- Bellman operator, Bellman optimal operator
- Dynamic Programming principle
- Policy Evaluation: Direct computation, Iteration, Monte-Carlo
- Contraction of Bellman operator, of Bellman optimal operator.
- Value Iteration
- Policy Iteration
- Modified Policy Iteration
- Quality function, Advantage function
- Bellman Q-operator.
- Incremental Monte-carlo updates: Temporal Difference, $TD(\lambda)$
- Q-temporal difference, Q-learning.
- Function approximation for V: Least-squares TD and Q: Fitted Q-iteration.
- Projection vs Contraction.

Multi-armed bandits

▷ The notion of Regret, of optimality gap Δ_a.

▷ What is Exploration? What is Exploitation?

▷ Exploration-Exploitation trade-off.

▷ Follow the leader, Explore then Commit strategies.

▷ The optimism in face of uncertainty principle.

▷ Hoeffding inequality for finite samples

▷ Handling random number of samples with Union bound.

▷ The Upper Confidence bound (UCB) strategy

▷ The Thompson sampling strategy

▷ Problem dependent regret lower bound: scaling in T, Kullback-Leibler.

▷ Most-confusing instance (e.g. for Bernoulli rewards)

▷ Problem-free (minimax) regret lower bound: scaling in T, A.

▷ KL-UCB strategy lower-bound approach.

▷ IMED strategy.

Structured Multi-armed bandits

- What is Unimodal structure? Lipschitz structure? Linear structure?
- Graph seen as a linear structure.
- Lower-bound for structured bandits: optimization problem.
- Most confusing instance for Lipshitz bandits.
- IMED for Lipschitz bandits.
- Linear regression setup.
- Sub-Gaussian noise assumption.
- Least-squares estimate.
- Optimistic principle for linear bandits.
- Information gain

Model-based MDPs

- Average gain criterion
- Poisson equation (gain and bias).
- Diameter of an MDP
- Value Iteration convergence issues.
- The span semi-norm
- Intrinsic contraction in span semi-norm.
- Stopping criterion for Value Iteration
- Exploration-Exploitation in MDPs
- UCB for MDPs: UCRL
- Building blocks of UCRL: Episode, EVI.
- What is an Extended MDP in EVI?
- What is guaranteed when EVI stops?

Planning

- Monte Carlo Tree Search
- What are the 4 main steps of of MCTS strategy?
- UCT rule for the value of each node.
- What is a Generative model?
- KL-OLOP combines two main algorithms: which ones?
- What is Best-armed identification (BAI) objective?
- What is Simple regret?
- Fixed-budget objective vs Fixed-confidence objective
- Reduction from cumulative to simple regret
- Sequential Halving
- What do we track in Track-and-stop?
- What is forced exploration?
- UCT rule in max node, versus UCT rule in min node.
- Monte-Carlo Graph Search idea
- When to rather use MGTS? When to rather use MCTS?

Deep Reinforcement Learning

- Model-based vs Model-free
- Critic algorithm, Actor algorithm, Actor-critic algorithm.
- Example of Critic, Actor, algorithms?
- Q-learning idea.
- What is slow/fast network updates? Why was it introduced?
- What is experience replay?
- What is prioritized experience replay?
- Double DQN.
- Policy gradient theorem.
- Idea behind Reinforce strategy.
- Natural gradient
- TRPO (name, principle)
- PPO (name, principle)

“*The more applied you go, the stronger theory you need*”

MERCI

odalricambrym.maillard@inria.fr

odalricambrymmaillard.wordpress.com