Offline Learning for Planning: A Summary

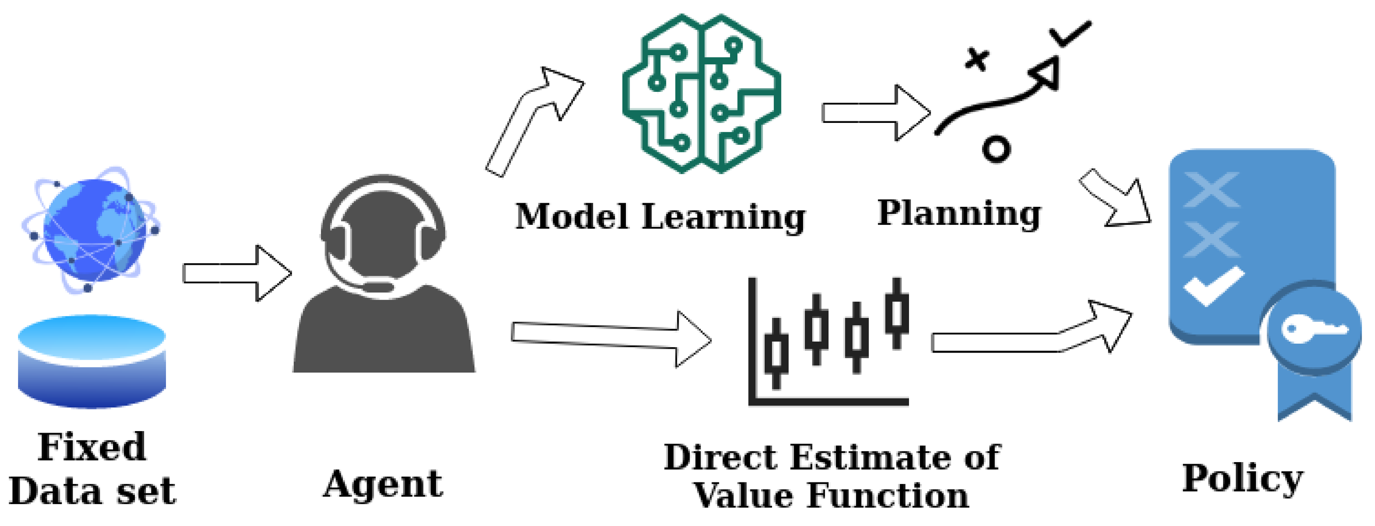
Giorgio ANGELOTTI
Nicolas DROUGARD
Caroline P. C. CHANEL

Context: Learning and Planning with a Fixed Data Set

⚠ **Expensive data acquisition procedure due to:** ① risky interaction with the environment, ② humans or living beings in the loop, ③ time consuming tasks, ④ costly experimental setups.

- **Exploit pre-collected *static* data sets but...** → ⚠ dangerous distributional shift (*data distribution* $\neq$ *environment distribution*)
- **Epistemic uncertainty penalized** Markov Decision Process (MDP) → plan to not perform actions that lead to experiences:
  ① *too* different from the data distribution
  ② for which we are not able to correctly predict the outcomes

Planners and Learners



Model Based Planners infer a MDP model from samples and then plan in it

Model Free Learners estimate the MDP Value function or policies directly from the batch

Hybrid approaches ① infer a generative model, ② use it to generate new data, ③ apply a model-free paradigm on the data augmented data set

Offline Learning: Model Free

Main source of error: evaluation of an approximation of the Q-value for Out Of Distribution Actions (no data → bad fit) leads to an accumulation of error [1]

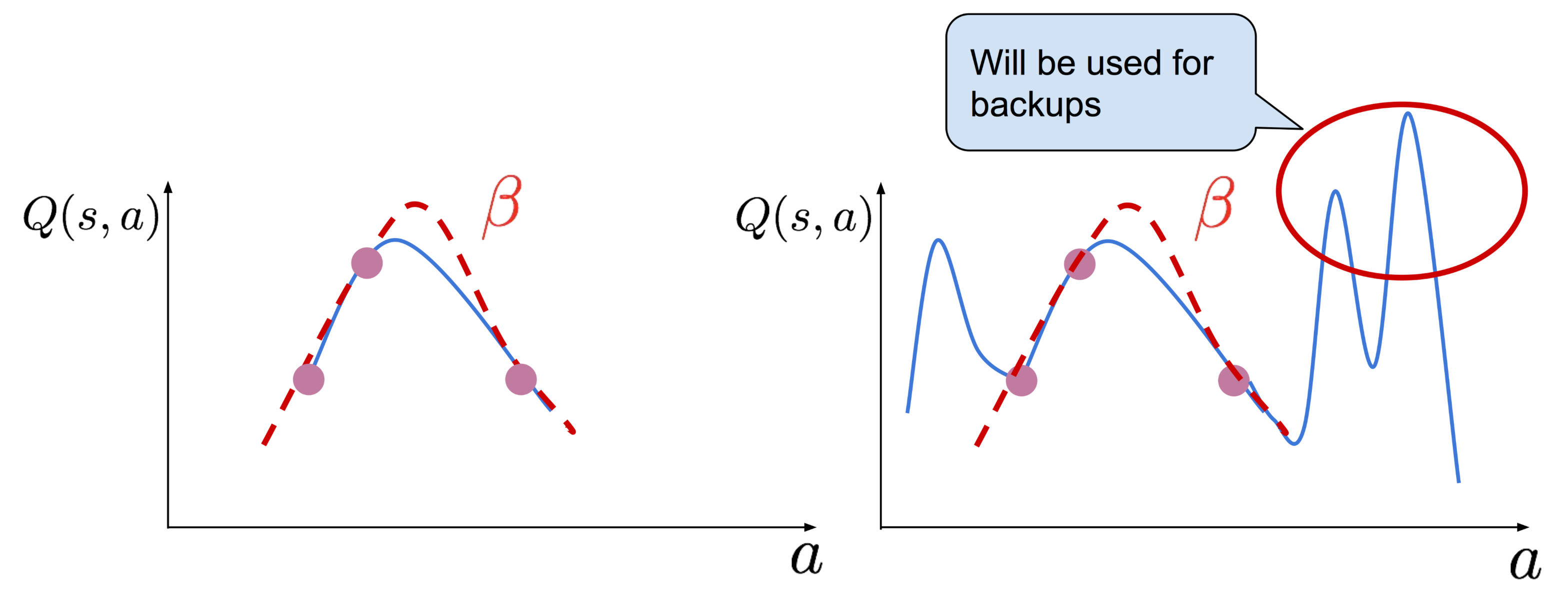


Figure from Kumar A., `https://bair.berkeley.edu/blog/2019/12/05/bear/`

IDEA: penalize Q-Learning by policy constrain [1, 2] ($\pi_{\mathcal{B}}$ generating the batch):

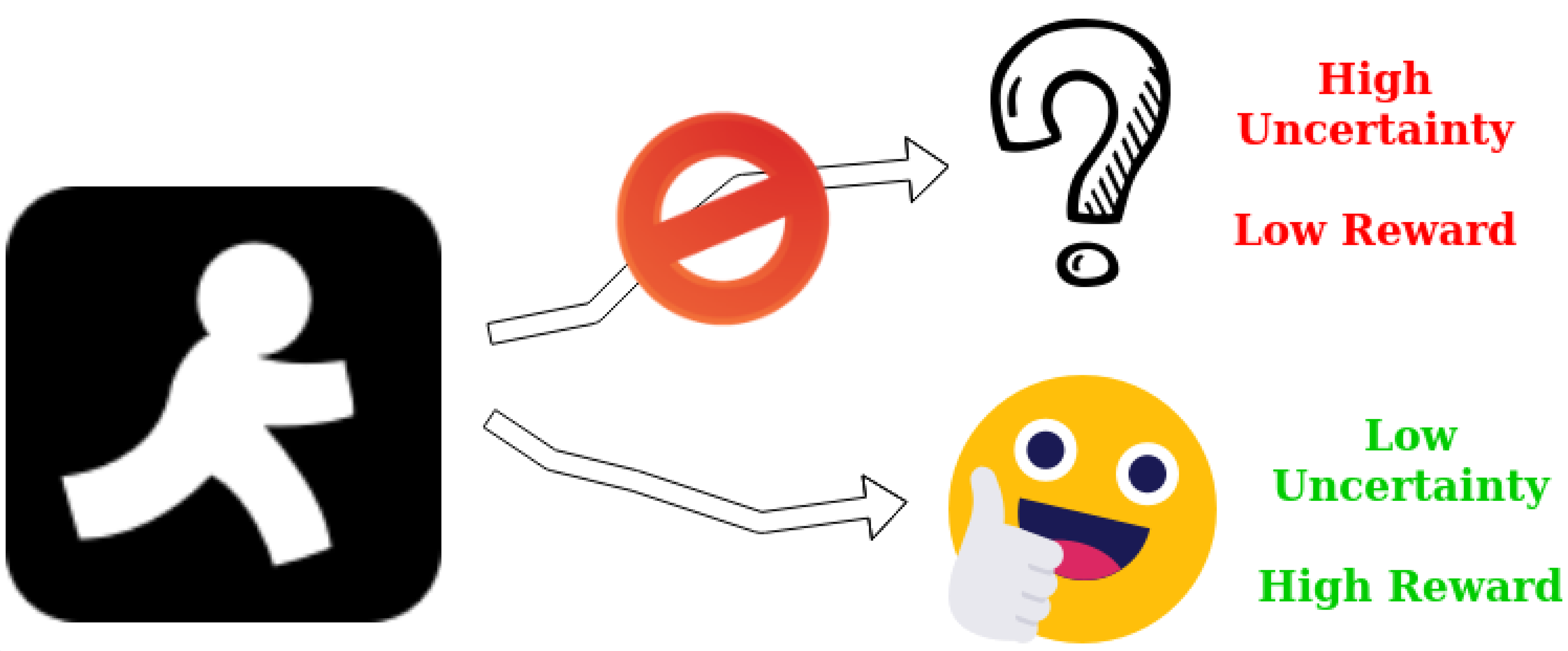
$$\mathop{\mathbb{E}}_{(s,a,r,s')\sim\mathcal{B}}\left[\mathop{\mathbb{E}}_{a\sim\pi(\cdot|s)}\left[Q(s,a)\right]-\alpha D\Big(\pi(\cdot|s),\pi_{\mathcal{B}}(\cdot|s)\Big)\right]$$

Offline Learning: Model Based

Main source of error: *uncertainty* in the estimate of the transition function $\hat{T}$ → model error accumulated during planning

- **IDEA 1:** create an MDP with a great penalized absorbing state if the confidence in the estimate is above a given threshold [4]
- **IDEA 2:** penalize the reward function ← estimate of the model error [3]

$$\tilde{R}(s,a) = R(s,a) - V_{max}\frac{\gamma}{1-\gamma}D\left[T(\cdot|s,a),\hat{T}(\cdot|s,a)\right]$$



Future Perspectives

NEED FOR:

① Better estimates of the distributional shift:
- Robust and Low Variance Importance Sampling estimators

② Better distribution learner:
Generative Adversarial Network
- generate new samples from a distribution similar to the one of the batch.
- no need to specify a prior.

③ Better offline learning for planning paradigms
⚠ always penalizing = too much?

References

[1] Kumar et al. *Stabilizing off-policy q-learning via bootstrapping error reduction*. In Advances in NIPS, 2019
[2] Wu et al. *Behavior regularized offline reinforcement learning*. ArXiv preprint, 2019
[3] Yu et al. *Mopo: Model-based offline policy optimization*, ArXiv preprint, 2020
[4] Kidambi et al. *MOReL : Model-Based Offline Reinforcement Learning*, ArXiv preprint, 2020

isae SUPAERO Institut Supérieur de l'Aéronautique et de l'Espace
ANITI Artificial & Natural Intelligence Toulouse Institute
Université Fédérale Toulouse Midi-Pyrénées