Sequential Decision Making: Theory and Practice

Lecture 6: Offline RL and Imitation Learning

Yuejie Chi

Yale Statistics & Data Science



Fall 2026

Project update

- My office is moved to Kline 1215.
- Today: second project draft due.
  - The second draft will also be posted in preparation for the presentation next week.
- Next week: in-class presentation of your project proposal, and peer review.
  - An opportunity to know what other people are cooking, and find inspiration!
  - Upload the presentation draft by this Sunday
  - In-class lightning talks: 5 min each, order to be determined/randomized
  - Peer review feedback by Wednesday (10/21), shared to you by Thursday (10/22)
  - Final proposal due next Sunday (10/25)
- Participation in peer review counts toward your course credit, but the reviews will not be used to determine the project grade.

Outline

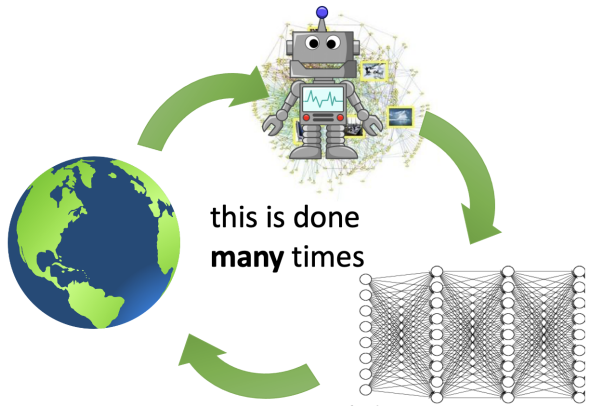
Offline multi-arm bandits

Offline RL

Behavior cloning

The interactive setting: DAgger

Online versus Offline RL

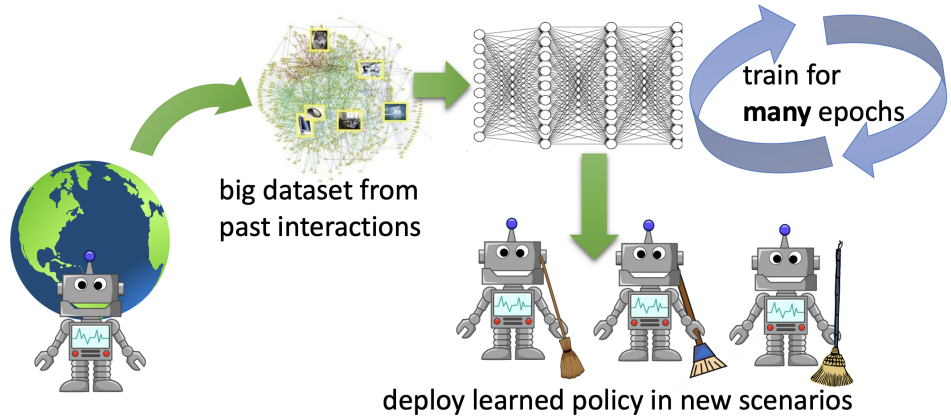


Online RL

- interact with environment
- actively collect new data

Offline/Batch RL

- no interaction
- data is given



Two main types of offline data and approaches

- Expert data: e.g., expert demonstration
  - imitation learning (imitate experts' behavior)
- Uniform coverage data: e.g., generative model / simulator
  - offline RL (e.g., model-based, model-free methods)



Imitation learning (IL)

Imitation is the sincerest form of flattery that mediocrity can pay to greatness.

— *Oscar Wilde*







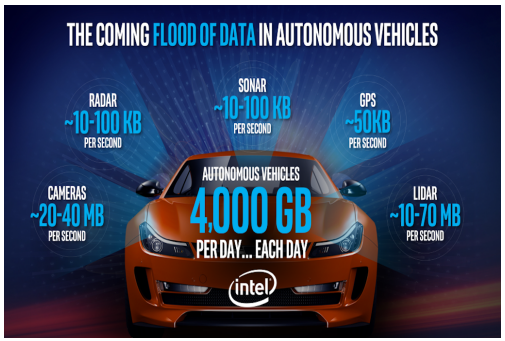
Goal: learn a good policy that mimics the expert demonstration.

Offline RL / Batch RL

- Sometimes we can not explore or generate new data
- But we have already stored tons of historical data



medical records



data of self-driving

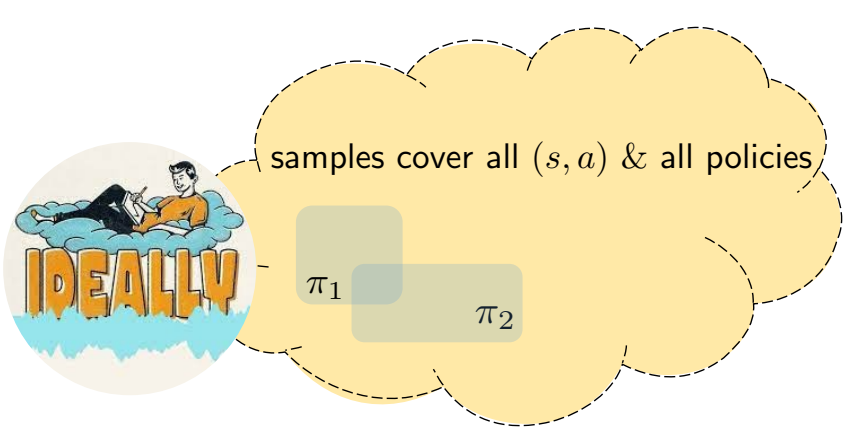


clicking times of ads

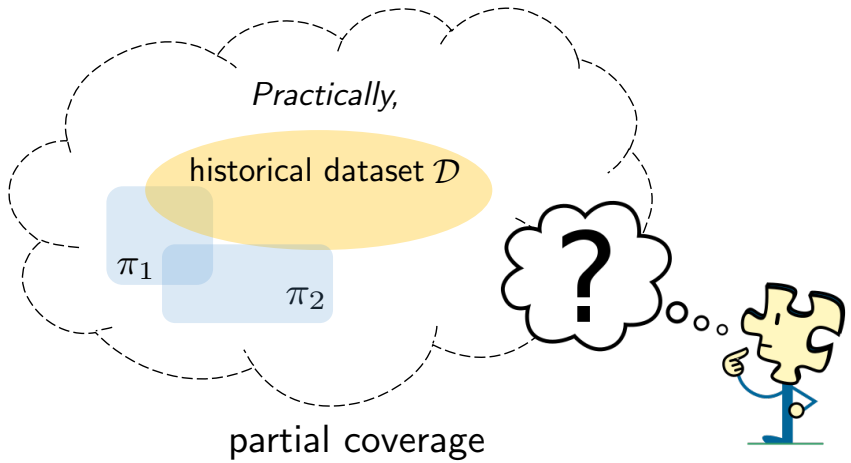
Goal: learn a good policy based on historical data without active exploration, but not through mimicking

Challenges of Offline RL

Partial coverage of state-action space:



uniform coverage over entire space
(sufficiently explored)



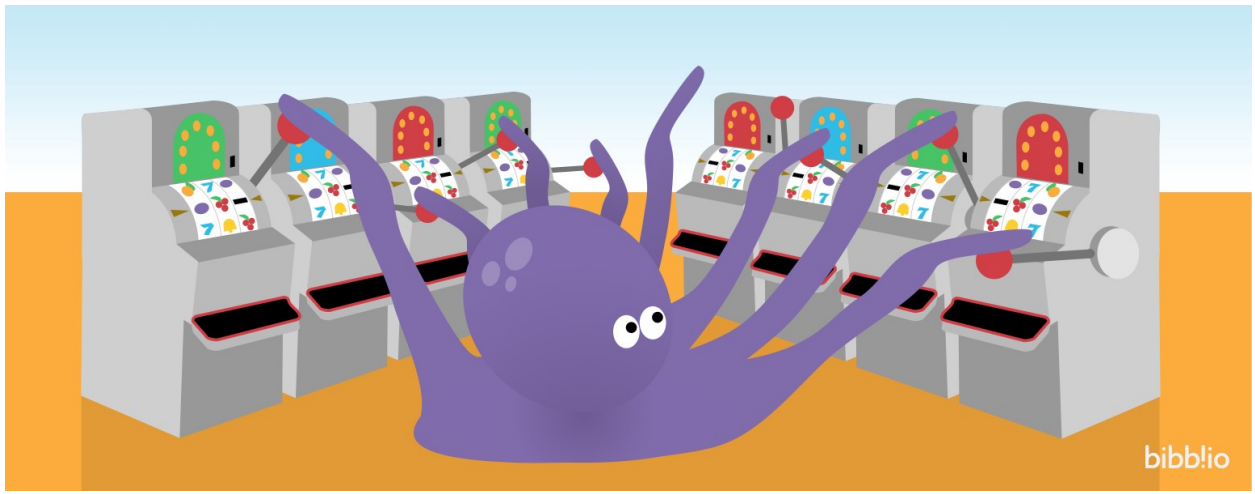
partial coverage
(inadequately explored)

Distribution shift:

$$\text{distribution}(\mathcal{D}) \neq \text{target distribution under } \pi^\star$$

Warm-up: offline multi-arm bandits

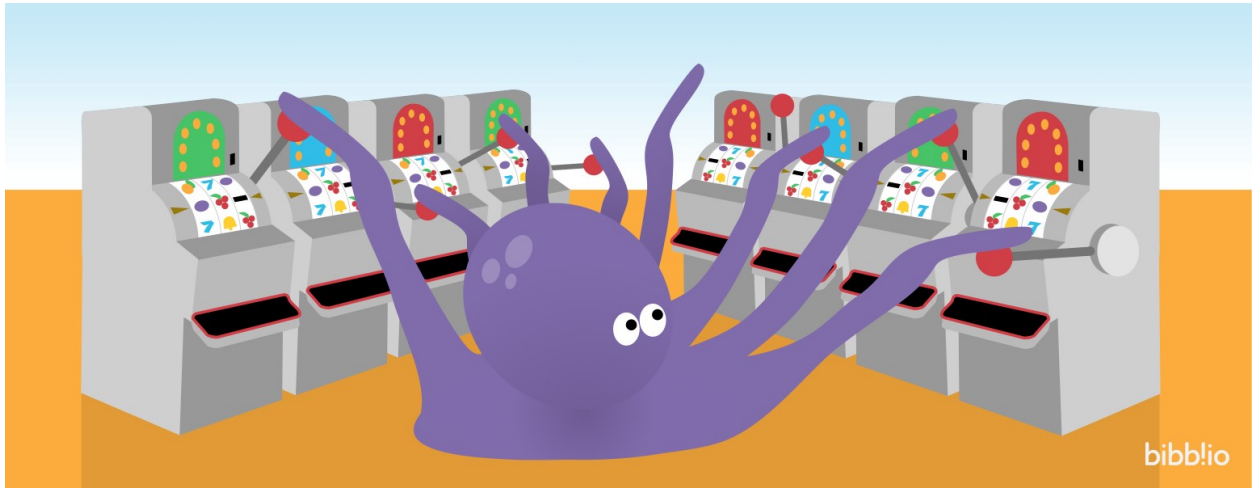
Multi-arm bandit



- Action space: $\mathcal{A} = \{1, 2, \ldots, A\}$
- Reward distributions: $R(\cdot \mid a)$ with mean $r(a)$

— *correspond to MDP with single state and* $\gamma = 0$

Offline learning in multi-arm bandit



Batch dataset: $\mathcal{D} = \{(a_i, r_i)\}_{1 \le i \le N}$, where

$$a_i \sim \mu, \qquad r_i \sim R(\cdot \mid a_i)$$

are collected in an *i.i.d.* manner, where $\mu \in \Delta(\mathcal{A})$ is the **behavior** policy.

Goal: minimize expected sub-optimality based on collected data

$$\mathbb{E}_{\mathcal{D}}[r(a^\star) - r(\widehat{a})],$$

where $a^\star = \arg\max_a r(a)$ is the optimal arm with the highest mean reward.

How to capture the distribution shift?

Single-policy concentrability coefficient [Rashidinejad et al., 2021]

$$C^\star := \max_a \frac{\pi^\star(a)}{\mu(a)} = \frac{1}{\mu(a^\star)}.$$

μ: behavior policy

$\pi^\star$: optimal policy

- When $C^\star = 1$: expert data
- When $C^\star > 1$: behavior policy deviates from the optimal policy
- When μ is uniform (random exploration), $C^\star = A$.
- Partial coverage: $C^\star$ is finite as long as $\mu(a^\star) > 0$.

A natural idea: empirical best arm

A natural idea is to pick the empirical best arm

$$\hat{a} := \arg\max_{a} \hat{r}(a),$$

where $\hat{r}(a)$ is the empirical mean reward of arm a.

However, this fails easily...

Example: Suppose that $\mathcal{A} = \{1, 2\}$, and $r_1 = 1$ and $r_2 = 0$. Assume we have N samples where $a_i = 1$ for all $i = 1, \ldots, N-1$, and $a_N = 2$.

- It follows that $\hat{r}_1 \sim \mathcal{N}(1, \frac{1}{N-1})$ and $\hat{r}_2 \sim \mathcal{N}(0, 1)$, and hence

$$\mathbb{P}(\hat{r}_1 < 1) = 0.5 \qquad \text{and} \qquad \mathbb{P}(\hat{r}_2 \geq 1) \geq 0.15.$$

- Then for any $N \geq 2$, we have

$$\mathbb{E}[r_1 - r_{\hat{a}}] = \mathbb{P}(\hat{r}_2 > \hat{r}_1) \geq \mathbb{P}(\hat{r}_1 < 1)\mathbb{P}(\hat{r}_2 \geq 1) \geq 0.075$$

A natural idea: empirical best arm

A natural idea is to pick the empirical best arm

$$\hat{a} := \arg\max_{a} \hat{r}(a),$$

where $\hat{r}(a)$ is the empirical mean reward of arm a.

Theorem 1 ([Rashidinejad et al., 2021])

For any $\epsilon < 0.05$, $N \geq 500$, there exists a bandit problem with two arms such that for $\hat{a} = \mathrm{argmax}_a \hat{r}(a)$, one has

$$\mathbb{E}_{\mathcal{D}}[r(a^\star) - r(\hat{a})] \geq \epsilon.$$

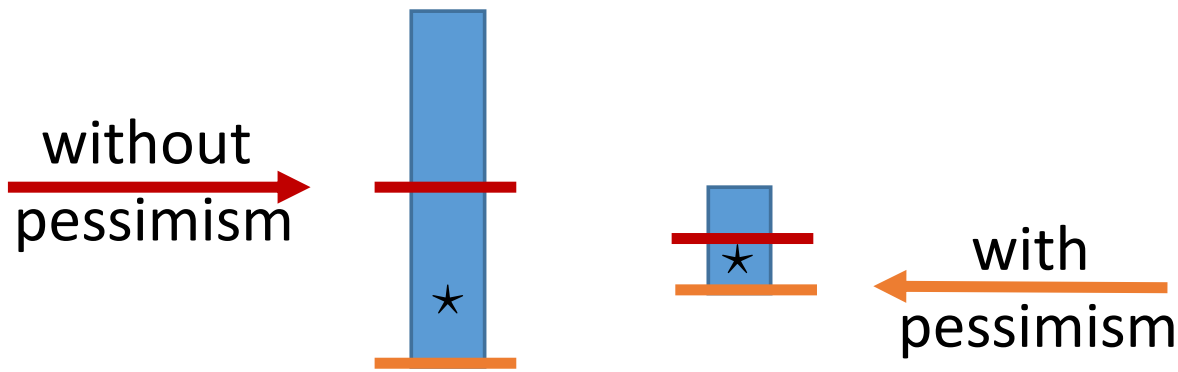
- This happens even when $C^\star$ is close to 1

Empirical best arm is sensitive to arms with few observations!

Pessimism via lower confidence bound

Lessons learned from failure of empirical best arm

- Should not treat arms equally
- Need to be **pessimistic** about arms with few observations



Lower confidence bound (LCB) for bandit: fix some $L > 0$, return

$$\hat{a} := \arg\max_a \quad \hat{r}(a) - \frac{L}{\sqrt{\max\{N(a), 1\}}}$$

$N(a)$: *number of times arm* a *is seen*

A closer look at LCB

Lower confidence bound for bandit: fix some $L > 0$, return

$$\hat{a} := \arg\max_a \quad \hat{r}(a) - \frac{L}{\sqrt{\max\{N(a), 1\}}}$$

$N(a)$: *number of times arm a is seen*

- View $\hat{r}(a) - \frac{L}{\sqrt{\max\{N(a),1\}}}$ as lower confidence bound of $r(a)$
- $\frac{L}{\sqrt{\max\{N(a),1\}}}$ arises from Hoeffding concentration inequality
- $\frac{L}{\sqrt{\max\{N(a),1\}}}$ is large when $N(a)$ is small: discount empirical mean with few observations

Performance guarantees of LCB for bandit

Theorem 2 ([Rashidinejad et al., 2021])

Set $L \asymp \sqrt{\log(AN)}$. *Policy* $\hat{a}$ *returned by LCB algorithm obeys*

$$\mathbb{E}_{\mathcal{D}}[r(a^\star) - r(\hat{a})] \lesssim \sqrt{\frac{C^\star}{N}}$$

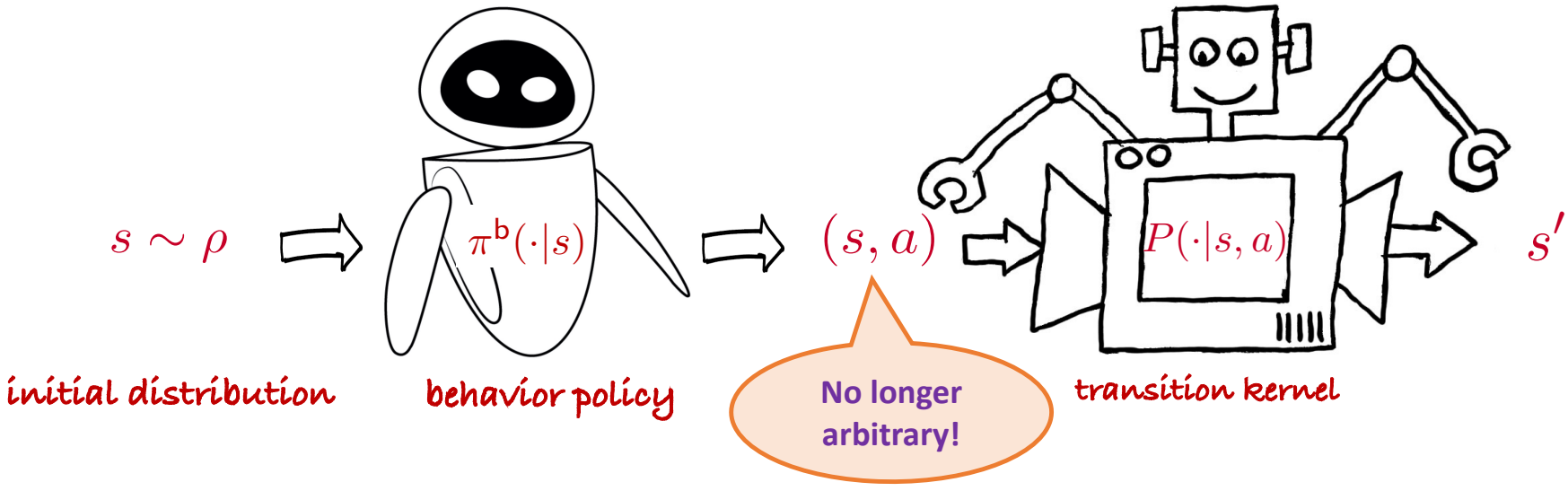
- LCB beats empirical best arm
- To achieve ϵ-optimality, the sample size needs to scale as

$$N \gtrsim \frac{C^\star}{\epsilon^2}.$$

- Performance of LCB degrades gracefully w.r.t. $C^\star$.

Offline RL: mathematical setup

A model of history data from behavior policy



Goal of offline RL: given history data $\mathcal{D} := \{(s_i, a_i, s_i')\}_{i=1}^N$, find an ε-optimal policy $\widehat{\pi}$ obeying

$$V^\star(\rho) - V^{\widehat{\pi}}(\rho) \leq \varepsilon$$

— *in a sample-efficient manner*

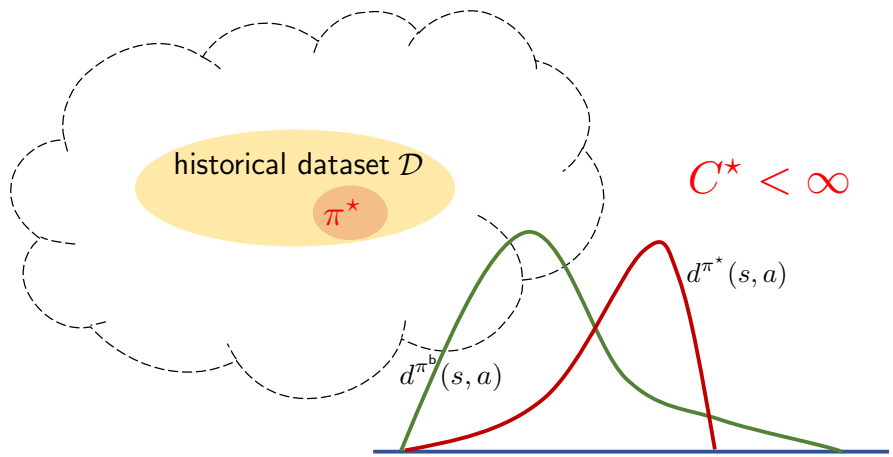
How to capture the distribution shift?

Single-policy concentrability coefficient [Rashidinejad et al., 2021]

$$C^\star := \max_{s,a} \frac{d^{\pi^\star}(s,a)}{d^{\pi^b}(s,a)} \geq 1$$

where $d^\pi(s,a)$ is the discounted state-action occupation density of policy π.

- allows for partial coverage
- Behavior cloning $C^\star = 1$



Minimax lower bound of offline RL

Theorem 3 ([Li et al., 2024])

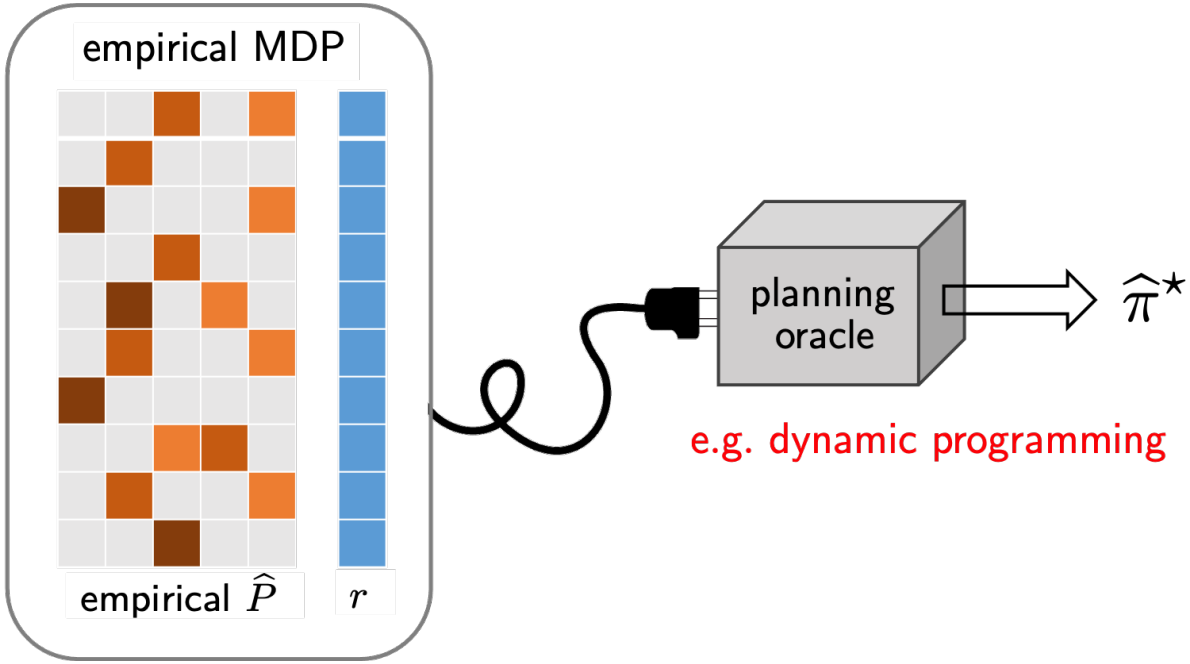
For any $\gamma \in [2/3, 1)$, $S \geq 2$, $C^\star \geq 3$, and $\varepsilon \leq \frac{1}{672(1-\gamma)}$, there exists some MDP and batch dataset such that no algorithm succeeds if the sample size is below

$$\widetilde{\Omega}\left(\frac{SC^\star}{(1-\gamma)^3\varepsilon^2}\right).$$

- The sample complexity becomes higher as $C^\star$ increases.

Model-based approach: value iteration

— [Azar et al., 2013]



Planning (e.g., value iteration) based on the the empirical MDP $\widehat{P}$:

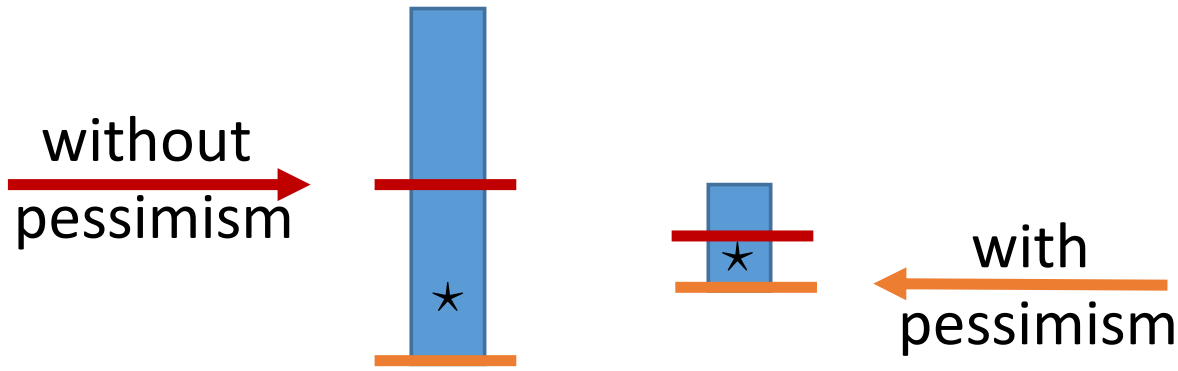
$$\widehat{Q}(s,a) \leftarrow r(s,a) + \gamma \langle \widehat{P}(\cdot \,|\, s,a), \widehat{V} \rangle, \quad \widehat{V}(s) = \max_a \widehat{Q}(s,a).$$

Issue: poor value estimates under partial and poor coverage.

Pessimistic value iteration

Pessimism in the face of uncertainty: penalize value estimate of (s, a) pairs that were poorly visited

— [Jin et al., 2021, Rashidinejad et al., 2021, Li et al., 2024]



Value iteration with lower confidence bound (VI-LCB):

$$\widehat{Q}(s,a) \leftarrow \max\left\{r(s,a) + \gamma\left\langle \widehat{P}(\cdot \,|\, s,a), \widehat{V}\right\rangle - \underbrace{b(s,a;\widehat{V})}_{\text{LCB penalty}}, 0\right\},$$

where $\widehat{V}(s) = \max_a \widehat{Q}(s,a)$.

Sample complexity of model-based offline RL

Theorem 4 ([Li et al., 2024])

For any $0 < \varepsilon \leq \frac{1}{1-\gamma}$, the policy $\widehat{\pi}$ returned by VI-LCB using a Bernstein-style penalty term achieves

$$V^{\star}(\rho) - V^{\widehat{\pi}}(\rho) \leq \varepsilon$$

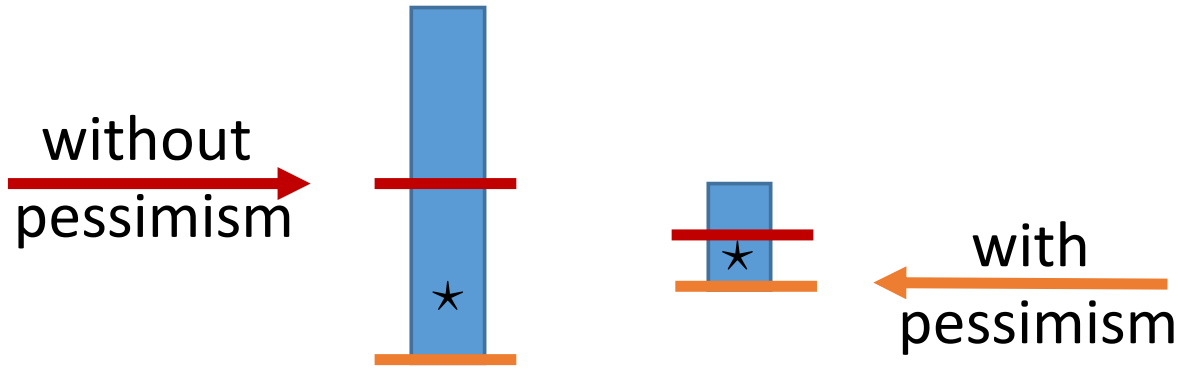
with high prob., with sample complexity at most

$$\widetilde{O}\left(\frac{SC^{\star}}{(1-\gamma)^3\varepsilon^2}\right).$$

- depends on distribution shift (as reflected by $C^{\star}$)
- minimax-optimal in the full ε-range

Model-based RL is minimax optimal with no burn-in cost!

Model-free offline RL: pessimistic Q-learning



— [Shi et al., 2022, Yan et al., 2022]

$$Q_{t+1}(s_t,a_t) \leftarrow \underbrace{(1-\eta_t)Q_t(s_t,a_t)+\eta_t\mathcal{T}_t\left(Q_t\right)(s_t,a_t)}_{\text{classical Q-learning}} - \eta_t \underbrace{b_t(s_t,a_t)}_{\text{LCB penalty}}$$

- $b_t(s,a)$: Hoeffding-style confidence bound

sample size: $\widetilde{O}\left(\frac{SC^\star}{(1-\gamma)^5\varepsilon^2}\right)$ $\Longrightarrow$ sub-optimal by a factor of $\frac{1}{(1-\gamma)^2}$

Issue: *large variability in stochastic update rules*

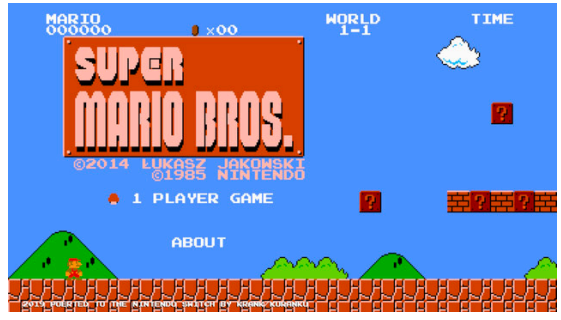
Behavior cloning with expert demonstration

Behavior cloning: IL from offline data

- Reward function is unknown or hard-to-tune in practice.
- Behavior cloning leverages expert demonstration to directly learn a policy without inferring the reward function.







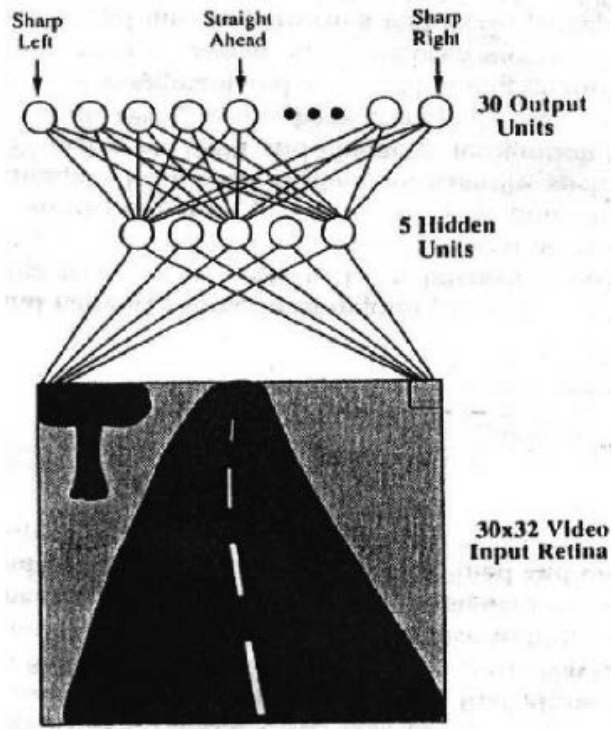
Many successes: robots, autonomous driving, drones, ...

ALVINN

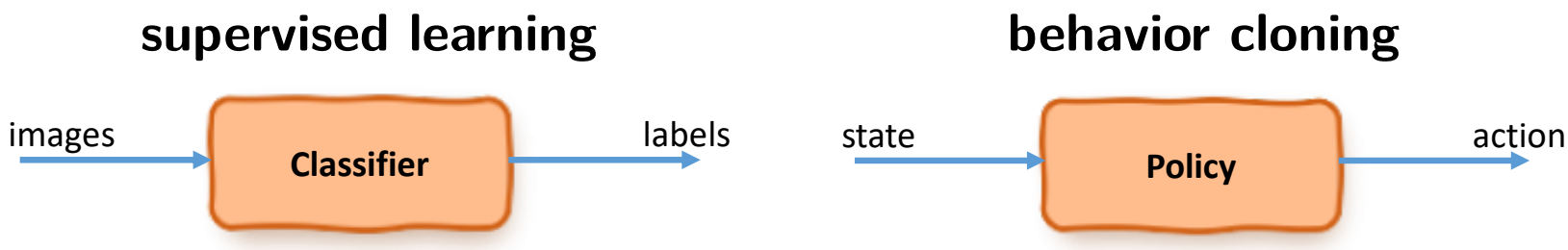
ALVINN (Autonomous Land Vehicle In a Neural Network), the self-driving car from 1988 [Pomerleau, 1988]!



Photo 1: *The NAVLAB autonomous navigation test-bed vehicle and the road used for trial runs.*



Reduction to supervised learning



Invoke supervised learning approaches to learn the policy:
a state-to-action mapping!

- Break down the expert demonstration (e.g., a trajectory) into a training dataset of (state, action) pairs;
- Learn a state-to-action mapping (i.e., policy) from the training data via your favorite supervised learning algorithm.

— *doesn't require knowing the reward!*

Warmup: revisiting the bandit example

Focus on the regime of near-expert data: $C^\star \in [1, 2)$.

- Observe the expert's action selections in N rounds: $\{a_i\}_{i=1}^N$.
- The BC policy is simply picking the action that is mostly selected:

$$\hat{a} = \arg\max_a N(a)$$

- Its error decays exponentially with N:

$$\begin{aligned}
\mathbb{E}[r(a^\star) - r(\hat{a})] &\leq P(\hat{a} \neq a^\star) \\
&\leq P\left(N(a^\star) \leq \frac{N}{2}\right) \\
&\leq \exp\left(-N \cdot \mathsf{KL}\big(\mathsf{Bern}(\frac{1}{2}) \| \mathsf{Bern}(\frac{1}{C^\star})\big)\right),
\end{aligned}$$

where the last line follows from the Chernoff bound for binomial rvs.

Supervised learning

Expert data: suppose we have an i.i.d. dataset

$$\mathcal{D} = \{s_i, a_i\}_{i=1}^N, \qquad \text{with} \quad a_i = \pi^\star(s_i), \;\; s_i \sim d_\rho^{\pi^\star}.$$

Here, $\pi^\star$ is the optimal policy, and $d_\rho^{\pi^\star}$ is the discounted state-visitation distribution induced by $\pi^\star$.

Supervised learning: we learn a policy

$$\widehat{\pi} = \arg\min_{\pi \in \Pi} \sum_{i=1}^N \ell(\pi; s_i, a_i)$$

- Π is the policy class, which is assumed to be finite, and $\pi^\star \in \Pi$
- $\ell(\pi; s_i, a_i)$ is the sample loss, e.g.
  - negative log-likelihood: $-\log \pi(a_i|s_i)$
  - least-squares loss: $\|\pi(s_i) - a_i\|_2^2$
  - hinge loss, 0-1 loss, etc...

Performance guarantee of behavior cloning

Theorem 5 ([Ross and Bagnell, 2010])

Suppose that supervised learning works, where the learned deterministic policy satisfy

$$\mathbb{E}_{s\sim d_\rho^{\pi^\star}} \mathbb{I}\left(\widehat{\pi}(s) \neq \pi^\star(s)\right) \leq \epsilon,$$

then BC returns a policy $\widehat{\pi}$ such that

$$V^\star(\rho) - V^{\hat{\pi}}(\rho) \leq \frac{2}{(1-\gamma)^2}\epsilon.$$

- The error is "amplified" by a quadratic factor of the horizon dependency $\frac{1}{(1-\gamma)^2}$, which is unavoidable in the worst case [Rajaraman et al., 2020].

Key ingredients of the proof

Lemma 6 (Performance difference lemma)

For all policies π, π' and distributions ρ over $\mathcal{S}$,

$$V^{\pi}(\rho) - V^{\pi'}(\rho) = \frac{1}{1-\gamma}\mathbb{E}_{s'\sim d_{\rho}^{\pi}}\mathbb{E}_{a'\sim\pi(\cdot|s')}\left[A^{\pi'}(s',a')\right],$$

where $A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s)$ is the advantage function, and $d_{\rho}^{\pi}(s) = (1-\gamma)\sum_{t\geq 0}\gamma^{t}\mathbb{P}_{\pi}(s_t = s)$ is the discounted state visitation distribution induced by π.

- measures the performance difference for any pairs of policies

Proof of Lemma 6

Let $\mathbb{E}_\pi$ denote expectation over trajectories with $s_0 \sim \rho$, $a_t \sim \pi(\cdot|s_t)$, and $s_{t+1} \sim P(\cdot|s_t, a_t)$. Then

$$
\begin{aligned}
V^\pi(\rho) - V^{\pi'}(\rho) &= \mathbb{E}_\pi\Big[\sum_{t\geq 0} \gamma^t r(s_t, a_t) - V^{\pi'}(s_0)\Big] \\
&= \mathbb{E}_\pi\Big[\sum_{t\geq 0} \gamma^t \big(r(s_t, a_t) + \gamma V^{\pi'}(s_{t+1}) - V^{\pi'}(s_t)\big)\Big] \text{ (telescoping)} \\
&= \mathbb{E}_\pi\Big[\sum_{t\geq 0} \gamma^t \big(Q^{\pi'}(s_t, a_t) - V^{\pi'}(s_t)\big)\Big] \\
&= \mathbb{E}_\pi\Big[\sum_{t\geq 0} \gamma^t A^{\pi'}(s_t, a_t)\Big] \\
&= \frac{1}{1-\gamma} \mathbb{E}_{s\sim d^\pi_\rho} \mathbb{E}_{a\sim\pi(\cdot|s)}\big[A^{\pi'}(s, a)\big].
\end{aligned}
$$

Proof of Theorem 5

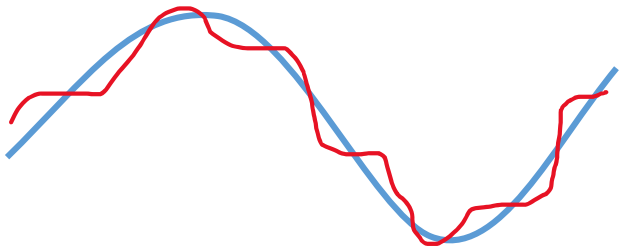
By the performance difference lemma,

$$\begin{aligned} V^{\pi^\star}(\rho) - V^{\hat{\pi}}(\rho) &= \frac{1}{(1-\gamma)} \mathbb{E}_{s \sim d_\rho^\star} A^{\hat{\pi}}(s, \pi^\star(s)) \\ &= \frac{1}{(1-\gamma)} \Big[\mathbb{E}_{s \sim d_\rho^\star} A^{\hat{\pi}}(s, \pi^\star(s)) - \underbrace{\mathbb{E}_{s \sim d_\rho^\star} A^{\hat{\pi}}(s, \widehat{\pi}(s))}_{=:0} \Big] \\ &\leq \frac{2}{(1-\gamma)^2} \mathbb{E}_{s \sim d_\rho^\star} \mathbb{I}\left(\widehat{\pi}(s) \neq \pi^\star(s)\right) \\ &\leq \frac{2\epsilon}{(1-\gamma)^2}, \end{aligned}$$

where the penultimate line used the fact that $-\frac{1}{1-\gamma} \leq A^{\hat{\pi}} \leq \frac{1}{1-\gamma}$, and the last line used $\mathbb{E}_{s \sim d_\rho^{\pi^\star}} \mathbb{I}\left(\widehat{\pi}(s) \neq \pi^\star(s)\right) \leq \epsilon$.

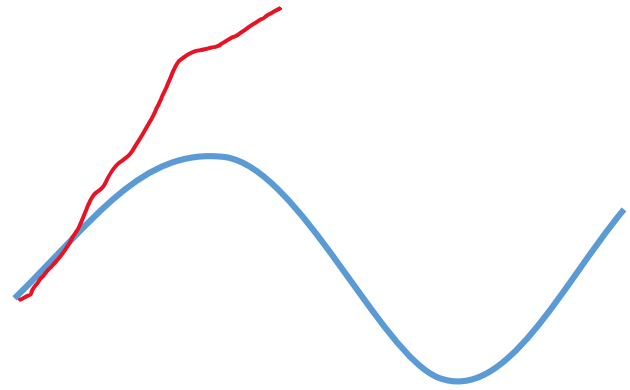
Why does error compounding occur?

Consider prediction over a horizon H: error at step t is ϵ.



supervised learning
independent error over t

$$\epsilon + \ldots + \epsilon = H\epsilon$$



behavior cloning
error propagates into the future

$$H\epsilon + (H-1)\epsilon + \ldots + \epsilon \sim H^2\epsilon$$

The BC error is quadratic in H instead of linear in H!

Error rate of supervised learning?

Taking Maximum likelihood estimate (MLE) as an example...

$$\widehat{\pi}_{\mathsf{MLE}} = \arg\max_{\pi \in \Pi} \sum_{i=1}^{N} \log \pi(a_i|s_i)$$

Theorem 7 ([Agarwal et al., 2019])

With probability at least $1-\delta$*, we have*

$$\mathbb{E}_{s \sim d_\rho^{\pi^\star}} \left\| \widehat{\pi}(\cdot|s) - \pi^\star(\cdot|s) \right\|_{\mathsf{TV}}^2 \leq \frac{2\log(|\Pi|/\delta)}{N}$$

- The error depends on $\log|\Pi|$, allowing rich policy class.

Performance guarantees of BC with MLE

Theorem 8 ([Agarwal et al., 2019])

With probability at least $1-\delta$*, BC returns a policy* $\widehat{\pi}_{\mathsf{MLE}}$ *such that*

$$V^{\star}(\rho) - V^{\hat{\pi}_{\mathsf{MLE}}}(\rho) \leq \frac{3}{(1-\gamma)^2}\sqrt{\frac{\log(|\Pi|/\delta)}{N}}.$$

- To achieve ϵ-accuracy, the sample size needs to be at least

$$N \gtrsim \frac{\log|\Pi|}{(1-\gamma)^4\epsilon^2}.$$

Distribution shift

$$\widehat{\pi} = \arg\min_{\pi \in \Pi} \mathbb{E}_{s \sim d^{\pi^\star}} \left[\ell(\pi; s, \pi^\star(s)) \right]$$

In supervised learning, a learner's prediction does not influence the distribution of examples upon which it will be tested.

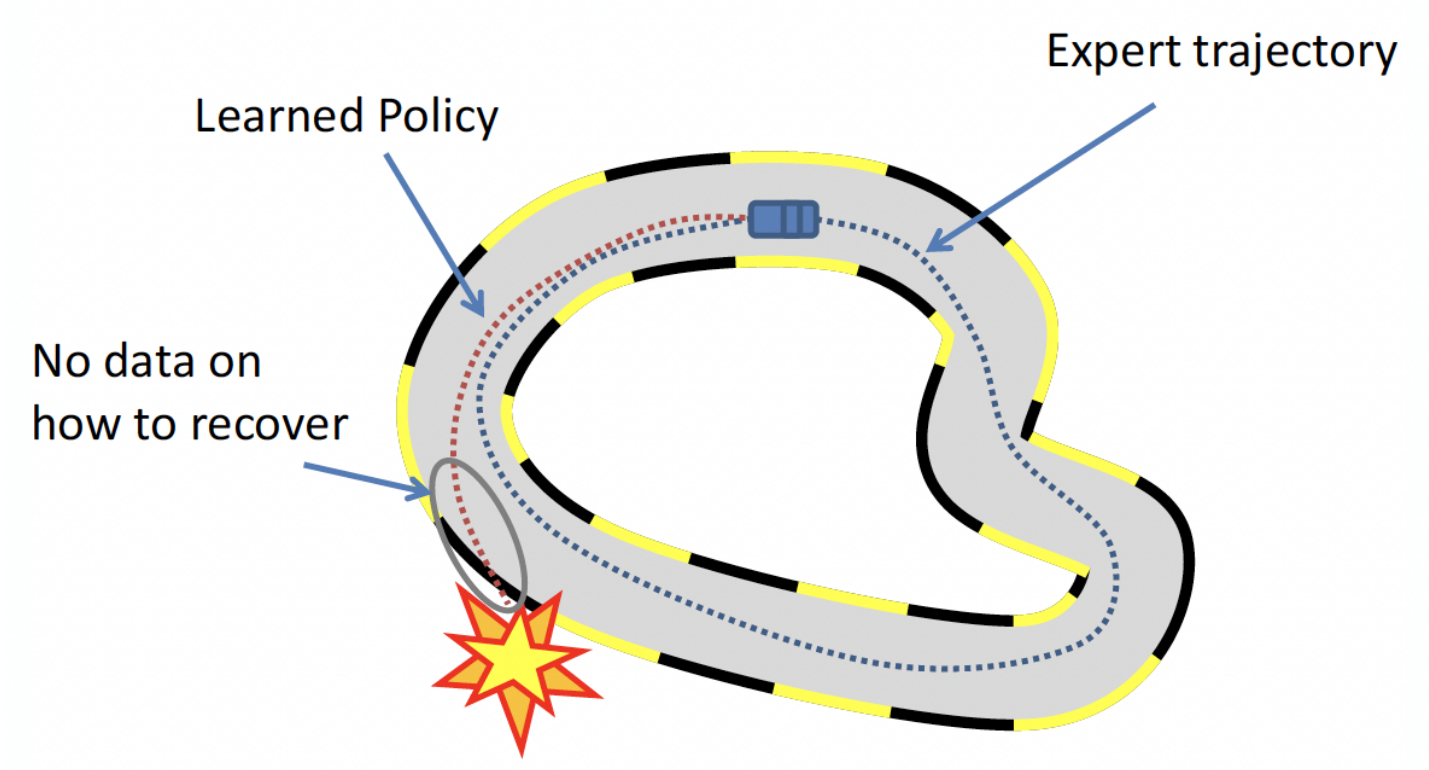
training: $s \sim d^{\pi^\star}$ $=$ test: $s \sim d^{\pi^\star}$

This is not the case with sequential decision making: Predictions affect future inputs/observations!

training: $s \sim d^{\pi^\star}$ $\neq$ test: $s \sim d^{\hat{\pi}}$

Distribution shift

training: $s \sim d^{\pi^\star}$ $\neq$ test: $s \sim d^{\hat{\pi}}$



ALVINN [Pomerleau, 1988]: "when driving for itself, the network may occasionally stray from the road center, so it must be prepared to recover by steering the vehicle back to the center of the road."

The interactive setting: DAgger

Revisit the supervised learning reduction

$$\widehat{\pi} = \arg\min_{\pi \in \Pi} \mathbb{E}_{s \sim d^{\pi}} \left[\ell(\pi; s, \pi^{\star}(s)) \right]$$

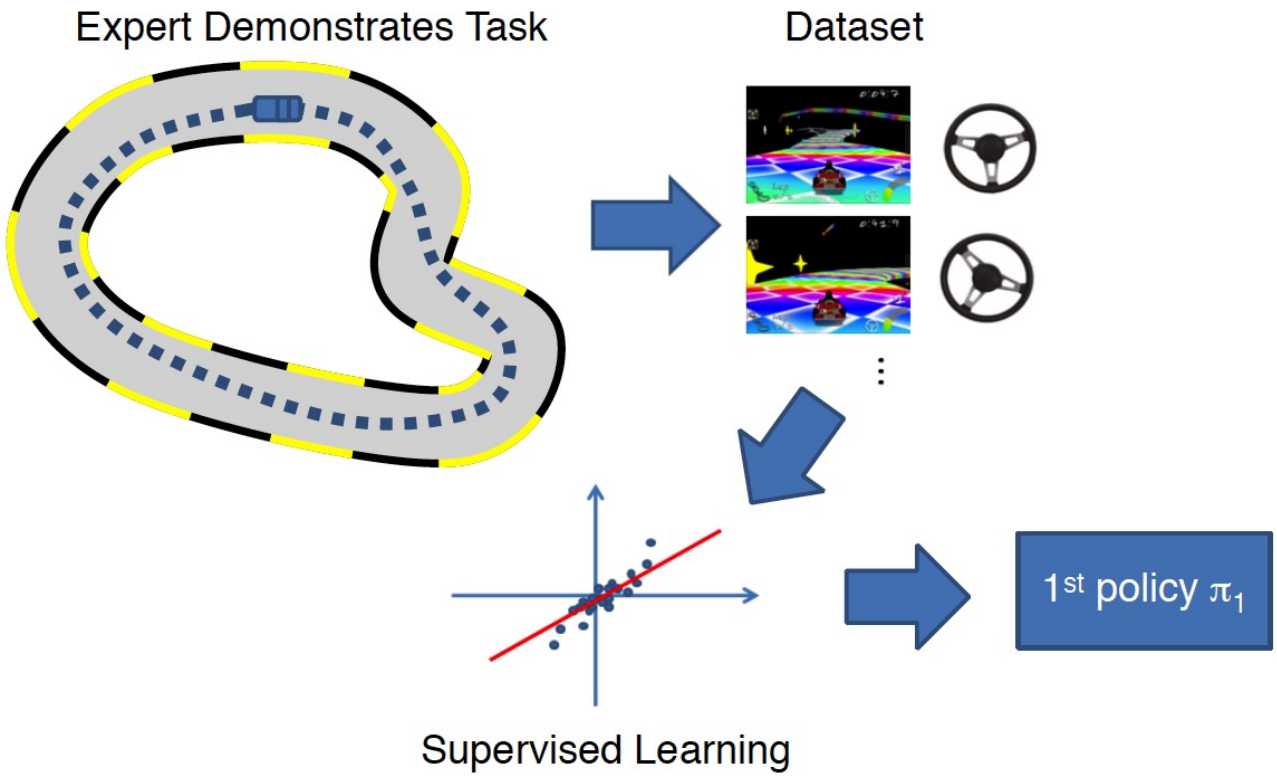
We want to optimize the performance when the state is drawn from d^{π}, which also need to be learned.

- non-i.i.d. supervised learning problem
- much more challenging!

[Ross et al., 2011] proposed an algorithm called `DAgger` (dataset aggregation) that allows sampling new expert data on the learned policy.

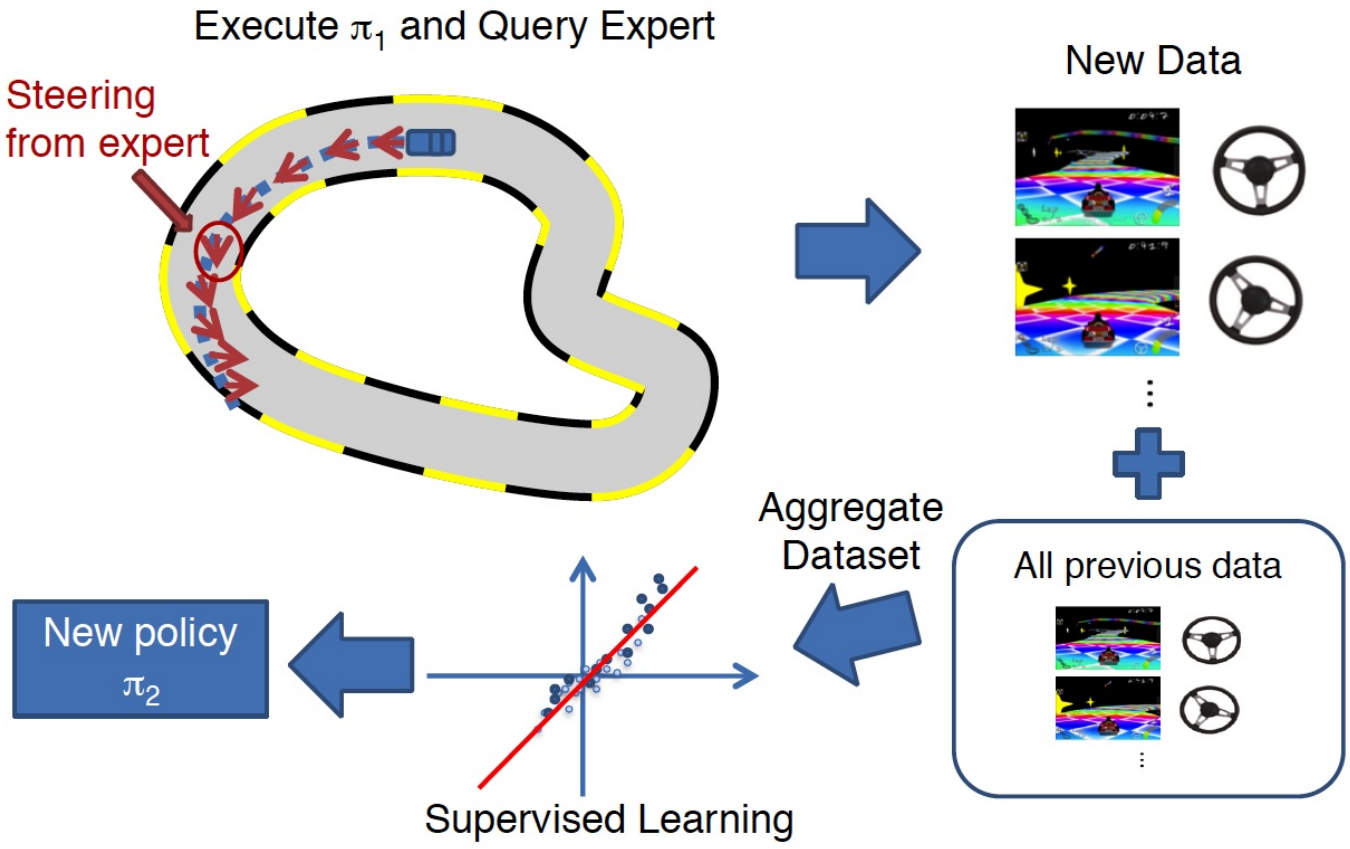
DAgger: dataset aggregation

[Ross et al., 2011]: train on a mixture of expert and behavior policy via reduction to no-regret online learning



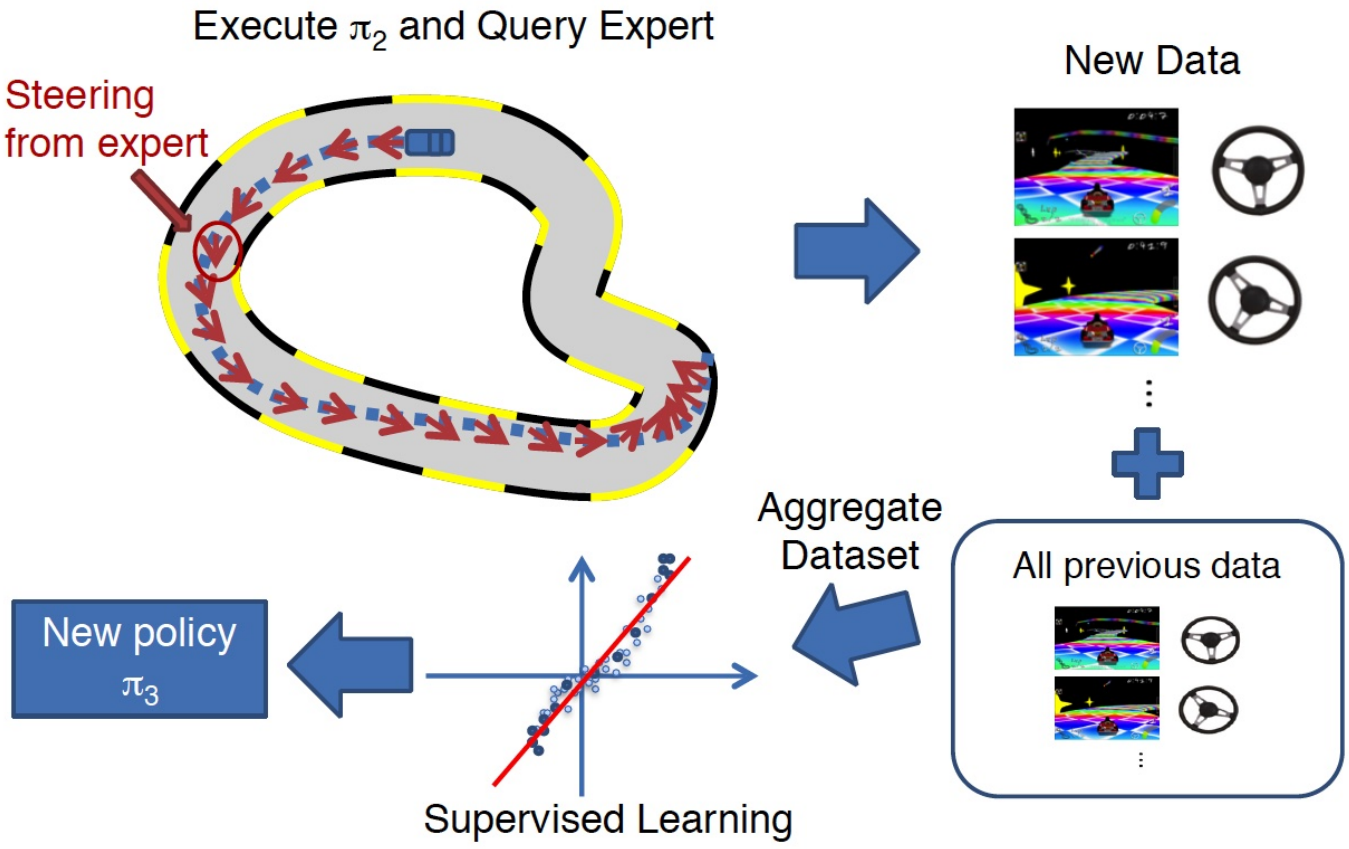
DAgger: dataset aggregation

[Ross et al., 2011]: train on a mixture of expert and behavior policy via reduction to no-regret online learning



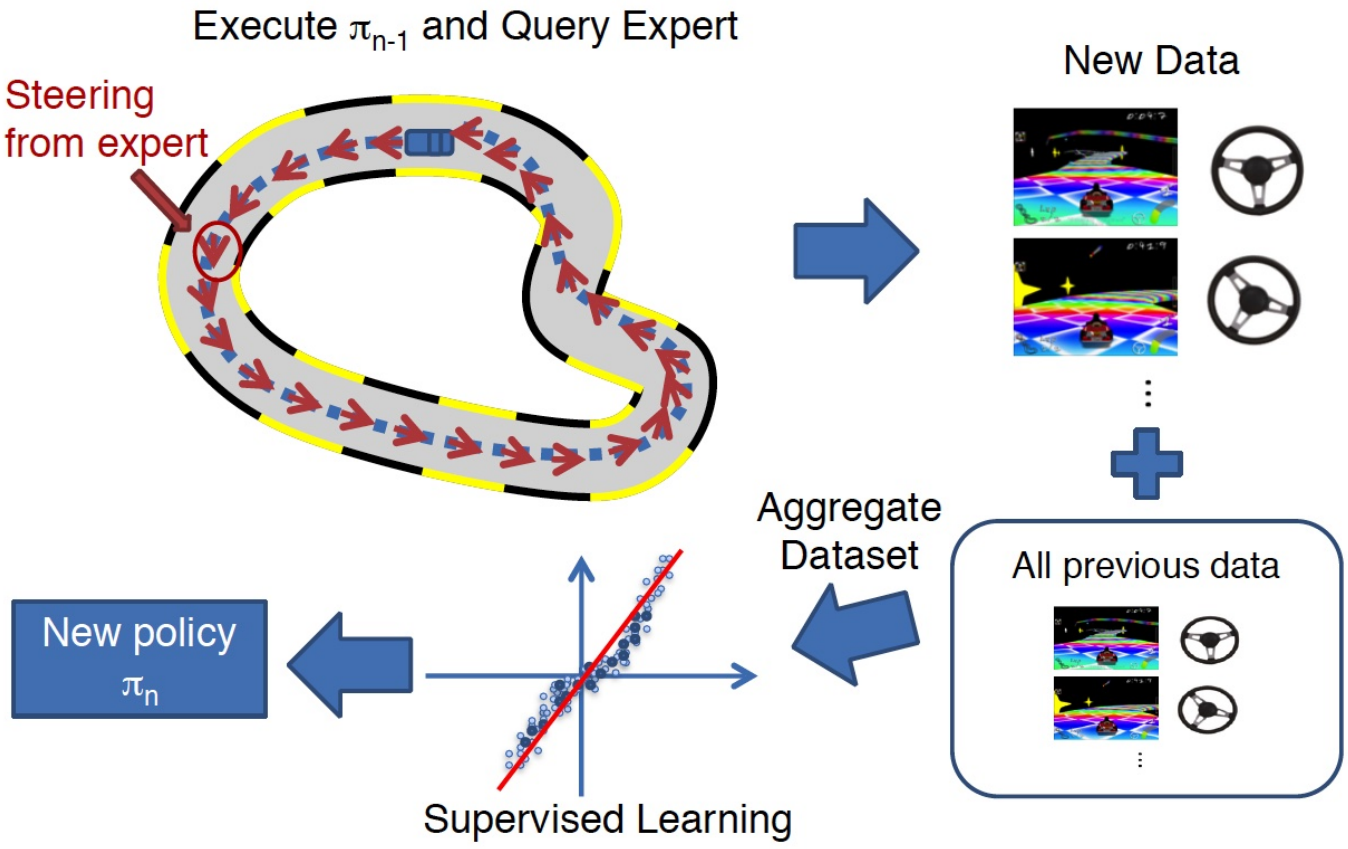
DAgger: dataset aggregation

[Ross et al., 2011]: train on a mixture of expert and behavior policy via reduction to no-regret online learning



DAgger: dataset aggregation

[Ross et al., 2011]: train on a mixture of expert and behavior policy via reduction to no-regret online learning



References I


Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. (2019).
Reinforcement learning: Theory and algorithms.


Azar, M. G., Munos, R., and Kappen, H. J. (2013).
Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model.
Machine learning, 91(3):325–349.


Jin, Y., Yang, Z., and Wang, Z. (2021).
Is pessimism provably efficient for offline RL?
In *International Conference on Machine Learning*, pages 5084–5096.


Li, G., Shi, L., Chen, Y., Chi, Y., and Wei, Y. (2024).
Settling the sample complexity of model-based offline reinforcement learning.
The Annals of Statistics, 52(1):233–260.


Pomerleau, D. A. (1988).
ALVINN: An autonomous land vehicle in a neural network.
Advances in neural information processing systems, 1.


Rajaraman, N., Yang, L., Jiao, J., and Ramchandran, K. (2020).
Toward the fundamental limits of imitation learning.
Advances in Neural Information Processing Systems, 33:2914–2924.

References II



Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. (2021).
Bridging offline reinforcement learning and imitation learning: A tale of pessimism.
Neural Information Processing Systems (NeurIPS).



Ross, S. and Bagnell, D. (2010).
Efficient reductions for imitation learning.
In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 661–668. JMLR Workshop and Conference Proceedings.



Ross, S., Gordon, G., and Bagnell, D. (2011).
A reduction of imitation learning and structured prediction to no-regret online learning.
In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings.



Shi, L., Li, G., Wei, Y., Chen, Y., and Chi, Y. (2022).
Pessimistic Q-learning for offline reinforcement learning: Towards optimal sample complexity.
In *International Conference on Machine Learning*, pages 19967–20025. PMLR.



Yan, Y., Li, G., Chen, Y., and Fan, J. (2022).
The efficacy of pessimism in asynchronous Q-learning.
arXiv preprint arXiv:2203.07368.