Sparser Rewards May Lead to Improved Reinforcement Learning

Franklyn H. Wang
franklyn_wang@college.harvard.edu

Fall 2020

Abstract

In offline reinforcement learning, agents can only interact with the environment through a dataset of transitions. One approach to offline RL is using a learned model and then planning through it. A problem of great interest is thus bounding the *discrepancy* – the difference between the reward under the true dynamics and the learned dynamics. In many cases, these bounds are quite weak to the point of almost being vacuous. Yet in practice, sometimes offline RL works quite well without any modifications. In this work, we use vector decomposition, a tool from spectral graph theory, to suggest that sparse rewards, paired with a new spectral expansion assumption, may allow for finer control of the discrepancy. Our results suggest deliberately shaping rewards to be sparser, which is surprising as sparser rewards tend to make standard tasks in reinforcement learning far more difficult. However, in offline RL we can plan far into the future, which may make sparser rewards better, not worse. We also suggest that existing benchmarks have sparse structures in them, which may make them more amenable to analysis.

1 Introduction

Offline Reinforcement Learning Offline Reinforcement Learning (offline RL) is an increasingly popular field of reinforcement learning where there is no interaction between the agent and the environment. Instead, the agent must learn a policy entirely from a dataset of the dynamics. Then, the agent is deployed into the real world without any further training. Offline reinforcement learning has clear applications to fields like self-driving cars, where collecting data in the real world with an untrained policy can have catastrophic (and sometimes lethal) consequences.

Model-Based Offline Reinforcement Learning In the absence of interaction with the environment, one natural way to obtain more information is to learn a predictive model of the dynamics, as in model-baesd reinforcement learning, and plan through this predicted MDP. However, these types of approaches are vulnerable to *model exploitation* – the agent often learns policies which take advantage of inaccuracies in the model to obtain greater reward.

One way to deal with model exploitation is by creating models that cannot be exploited – for example, by penalizing the agent when it takes uncertain actions. For example,

- In [YTY+20], they construct an MDP where the reward function is simply $\hat{r}(s,a) = r(s,a) - \alpha h(s,a)$ where $h(s,a)$ is the uncertainty in the model on (s,a).

- In [KRNJ20], they construct an MDP where high uncertainty actions are sent to an absorbing state with infinite negative reward.

Each of these papers bounds the difference between the reward of a given policy under the true and the learned dynamics, which we call the *discrepancy*. If the discrepancy gap is small, then simply planning on the learned model is enough to obtain high rewards on the true dynamics. However, the currently known bounds obtained on the discrepancy are essentially vacuous, unless the learned model is quite accurate. In spite of this, the empirical results given in the papers far exceeds the theoretical bounds on the performance provided.

In this paper, we first attempt to improve the bounds at the core of [KRNJ20, YTY+20]. We accomplish this by expressing the existing bounds on the discrepancy in terms of the transition matrix of the MDP. We note that their assumptions are too weak to obtain meaningful results, and thus incorporate an additional spectral expansion assumption. We then show that applying the *vector decomposition technique* on the resulting expression allows for tighter bounds in many cases, and especially when the rewards are sparse.

After that, we explain how many of the tasks in D4RL ([FKN+20]), the most popular offline reinforcement learning dataset, have naturally sparse reward structures within them. This seems to suggest that one reason why discrepancy gaps on these are often observed to be low is because while the rewards may not initially appear to be sparse, the true underlying tasks are sparse.

Our contributions are:

- Providing tighter bounds on the performance of certain MDPs, like the pessimal MDP of [KRNJ20], by incorporating a new spectral expansion assumption on the MDP.
- Utilizing these bounds to suggest ways of shaping rewards in offline RL to be sparser, which may yield lower discrepancy gaps and hence better empirical and theoretical results.

2 Preliminaries

Following the standard Markov decision process framework, we let the space of states be given by $\mathcal{S}$ and the action space be given by $\mathcal{A}$. We assume a finite state and action space, but we use this assumption solely for notational convenience, and in many places it can be dropped with no change to the results. The reward function takes state action pairs to real numbers between $-R_{\max}$ and $R_{\max}$, which we state formally as $r(\cdot,\cdot) : \mathcal{S} \times \mathcal{A} \to [-R_{\max}, R_{\max}]$. We consider the transition matrix $(\mathcal{S} \times \mathcal{A}) \to (\mathcal{S} \times \mathcal{A})$ which takes a transition step followed by a policy step. If we let a given state be x, the state distribution of the next state is Tx, and the matrix that turns this into a state-action transition vector is the policy matrix Π, where exactly one entry in each row is positive. Explicitly, the matrices can be written as follows. T is a $|\mathcal{S}| \times (|\mathcal{S}||\mathcal{A}|)$ matrix, Π is a $|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|$ matrix, and x is a $|\mathcal{S}||\mathcal{A}|$-length vector.

$$T = \begin{bmatrix} p(s'=1|s=1,a=1) & \dots \\ p(s'=2|s=1,a=1) & \dots \\ \vdots & \\ p(s'=|\mathcal{S}||s=1,a=1) & \dots \end{bmatrix}$$

$$\Pi = \begin{bmatrix} \pi(a=1|s=1) & & \\ \pi(a=2|s=1) & & \\ \pi(a=3|s=1) & & \\ & \vdots & \\ & & \pi(a=|\mathcal{A}|-2|s=n) \\ & & \pi(a=|\mathcal{A}|-1|s=n) \\ & & \pi(a=|\mathcal{A}||s=n) \end{bmatrix}$$

and it is multiplied by a state vector which looks like

$$x = \begin{bmatrix} p(a=1,s=1) \\ p(a=2,s=1) \\ \vdots \\ p(a=|\mathcal{A}|,s=|\mathcal{S}|) \end{bmatrix}$$

Thus, we get a transition function of $x \mapsto \Pi T x$, where T is the transition function. This can also be written as $x \mapsto Wx$ where $W = \Pi T$, which we call the *state-action transition matrix*. Thus, the expected reward function at each step (say, step t) is $r^\top W^t x$.

The *discrepancy* between a learned and a true MDP is the maximum difference between the expected reward on the learned MDP and the true MDP.

Our horizon is infinite, with discount factor γ, so the *overall value function* is

$$\eta_p[\pi] = \mathbb{E}_\tau \sum \gamma^t r(s_t, a_t)$$

where s_t, a_t are the actions. In general, when working with the true model we will use the notation $\eta[\pi]$ and when working with the learned model we will use the notation $\hat{\eta}[\pi]$.

3 Prior Work

Here we state the guarantees on discrepancy achieved by [KRNJ20] and [YTY+20] in less general form. First, we must define a pessimal MDP. An MDP M' is *pessimal* for an MDP M if the reward of any policy π with a fixed starting distribution is lower on M' than it is on M.

Theorem 3.1 ([YTY+20])**.** *The MDP whose rewards are given by*

$$r'_{s,a} = r_{s,a} - \frac{2R_{max}}{1-\gamma} h(s,a)$$

and whose transitions are given by $\hat{p}(s'|s,a)$ where $h(s,a) \geq D_{TV}(\hat{p}(s'|s,a), p(s'|s,a))$ (in other words $(r', \hat{p})$) is a pessimistic MDP for (r,p).

Proof. Essentially, the idea is to couple the trajectories on the real and learned MDPs together, which lets us show that the total variation distance is at most $\mathbb{E}_{\tau\sim(p,\pi)}[\sum_i D_{TV}(\hat{p}(s'|s_i,a_i), p(s'|s_i,a_i))]$. □

A pessimal MDP can be thought of as having essentially discrepancy gap zero, though it comes at the cost of changing the reward function. In the examples we see next, the reward functions of the learned MDP are the same as those of the true MDP. This will inevitably lead to nonzero discrepancy gaps.

The crucial lemma for [KRNJ20] is simply this result with $h(s,a) = \alpha$, a uniform bound over all total variation distances.

4 A spectral formulation of the bounds

In this section we reproduce the analysis of [KRNJ20], which essentially uses the *simulation lemma* of [KS02] to bound the discrepancy of the MDP. We first explain their argument with couplings, and then we express the argument with linear algebra. The second derivation is a natural stepping stone to our analysis with sparse rewards.

We'd like to understand the difference $\eta[\pi] - \hat{\eta}[\pi]$, where η is the reward under the true dynamics, and $\hat{\eta}$ is the reward under the learned dynamics. We assume a discount factor of γ. For simplicity, assume that the initial state distribution is equal.

We now state a lemma from [KRNJ20]:

Theorem 4.1. *Assume that $D_{TV}(\hat{p}(s'|s,a), p(s'|s,a)) \leq \alpha$ for all (s,a) pairs. Then for all policies π, we have*

$$\eta[\pi] - \hat{\eta}[\pi] \leq \frac{2\gamma\alpha R_{max}}{(1-\gamma)^2}$$

The standard proof of this lemma is given as follows:

Proof. Couple the trajectories of distributions under $\hat{p}$ and p in the direct fashion. Then, the total variation distance between the tth state marginal is at most $1-(1-\alpha)^t$. Thus, the total sum is

$$\begin{aligned}
\eta[\pi] - \hat{\eta}[\pi] &= \sum_t \gamma^t (\mathbb{E}_{a\sim\pi(\tau^{(t)})}[R(\tau^{(t)}, a)] - \mathbb{E}_{a\sim\pi(\hat{\tau}^{(t)})}[R(\hat{\tau}^{(t)}, a)]) \\
&\leq 2R_{\max} \sum_t \gamma^t 1_{\tau^{(t)} \neq \hat{\tau}^{(t)}} \\
&\leq 2R_{\max} \sum_t \gamma^t (1-(1-\alpha)^t) \\
&\leq \frac{2\gamma\alpha R_{\max}}{(1-\gamma)^2}
\end{aligned}$$

□

Now, we can write it in terms of random-walk matrices. Let W be the transition matrix by the policy and the true transition matrix and W' be the transition matrix by the policy and the learned transition matrix. Then the total variation distance condition implies the existence of nonnegative matrices $X, \mathcal{E}, \mathcal{E}'$ such that

$$W = (1-\alpha)X + \alpha\mathcal{E} \tag{1}$$
$$W' = (1-\alpha)X + \alpha\mathcal{E}', \tag{2}$$

where $\|X\|_1 \leq 1$, $\|\mathcal{E}'\|_1 \leq 1$ and $\|\mathcal{E}\|_1 \leq 1$. This neatly captures the total variation distance constraint, and is crucial in our second proof of the theorem.

Proof. We begin with a lemma to bound the single-step difference.

Lemma 4.2. *We have that $\left\|W^t - W'^t\right\|_1 \leq 2(1-(1-\alpha)^t)$*

Proof. We have that

$$\begin{aligned}
\left\|W^t - W'^t\right\|_1 &= \left\|\sum_{i=0}^{t} \binom{t}{i} \alpha^i (1-\alpha)^{t-i} X^{t-i} \mathcal{E}^i - \alpha^i (1-\alpha)^{t-i} X^{t-i} \mathcal{E}'^i\right\|_1 \\
&= \left\|\sum_{i=1}^{t} \binom{t}{i} \alpha^i (1-\alpha)^{t-i} X^{t-i} \mathcal{E}^i - \alpha^i (1-\alpha)^{t-i} X^{t-i} \mathcal{E}'^i\right\|_1 \\
&\leq \sum_{i=1}^{t} \left\|\binom{t}{i} \alpha^i (1-\alpha)^{t-i} X^{t-i} \mathcal{E}^i\right\|_1 + \left\|\binom{t}{i} \alpha^i (1-\alpha)^{t-i} X^{t-i} \mathcal{E}'^i\right\|_1 \\
&\leq 2 \sum_{i=1}^{t} \binom{t}{i} \alpha^i (1-\alpha)^{t-i} \\
&= 2(1 - (1-\alpha)^t)
\end{aligned}$$

□

The bound is

$$\begin{aligned}
\eta[\pi] - \hat{\eta}[\pi] &= \sum_t \gamma^t (\mathbb{E}_{a \sim \pi(\tau^{(t)})}[R(\tau^{(t)}, a)] - \mathbb{E}_{a \sim \pi(\hat{\tau}^{(t)})}[R(\hat{\tau}^{(t)}, a)]) \\
&= r^\top \sum_t \gamma^t (W^t - W'^t) x \\
&\leq \|r\|_\infty \left\|\sum_t \gamma^t (W^t - W'^t) x\right\|_1 \\
&\leq \|r\|_\infty \sum_t \gamma^t \left\|(W^t - W'^t) x\right\|_1 \\
&\leq \|r\|_\infty \sum_t \gamma^t \left\|W^t - W'^t\right\|_1 \|x\|_1 \\
&\leq \|r\|_\infty \sum_t 2\gamma^t (1 - (1-\alpha)^t) \\
&\leq \frac{2\gamma\alpha R_{\max}}{(1-\gamma)^2}
\end{aligned}$$

□

This bound is not very strong. One can easily show that the maximum reward is $R_{\max} + R_{\max}\gamma + \ldots = \frac{R_{\max}}{1-\gamma}$, so we can obtain a trivial upper bound of $\frac{2R_{\max}}{1-\gamma}$ for the discrepancy. In order for Theorem 4.1 to be stronger than the trivial upper bound, we need $\alpha < (1-\gamma)$ – which is a tough condition to satisfy, since γ is often around 0.99.

5 Analysis with Vector Decomposition

5.1 Motivation

In the above, the assumptions were so broad that obtaining useful bounds was difficult. In many real-world cases, for example, dynamics are deterministic, so any continuous distribution (which is the standard parametrization of dynamics) necessarily has almost vacuous total variation distance from the true dynamics. Setting that aside, in the equality cases for the arguments mentioned above, once the trajectories diverge, they never meet again, and the trajectory induced

by the learned policy only collects maximum rewards and the true model only collects negative maximum rewards always. We call this *the problem of the diverging MDP*.

In order to deal with this, one common assumption that researchers have explored is Lipschitz continuity of the value function [LXL+18]. However, this assumption comes with the downside of requiring explicit representations of the state space (e.g. the coordinate space).

We address the failure mode of "diverging MDPs" by controlling the connectedness of the MDPs, so that even after diverging the MDPs may (in fact, must) meet again. In what follows, we make spectral expansion assumptions on each the MDPs. Our spectral assumptions do not depend on a state representation for the MDPs. Yet, as we show, they might still lead to improved bounds. In order to take advantage of spectral expansion, we will need to use L_2 bounds on the state vector instead of L_1 bounds. Since this requires using the 2-norm of the reward vector, our bounds will become stronger with sparsity assumptions on the reward vector.

5.2 Analysis

In order to improve the bounds from the previous section, we'd like to take advantage of certain expansion assumptions on the MDPs, which necessitates slightly changing the approach. We're trying to bound an expression of the form $r^\top v$. The attempts given in the prior work essentially use

$$r^\top v \leq \|r\|_\infty \|v\|_1 .$$

By contrast, in order to use expansion assumptions on the MDP we use

$$r^\top v \leq \|r\|_2 \|v\|_2 .$$

We note that in a sparse setting, because $\|r\|_2 \approx \|r\|_\infty$, bounding the L2 norm will not cost anything, and it could even turn out to be very helpful if we could control $\|v\|_2$ better than $\|v\|_1$.

In the following, we consider the case where the state-action transition matrices are arbitrary Markov chains, and the stationary distributions are shared, so $\boldsymbol{\pi}_0$ is a stationary distribution of both W and W'. We will also assume that $\left\|W|_{\boldsymbol{\pi}_0^\perp}\right\|_2, \left\|W'|_{\boldsymbol{\pi}_0^\perp}\right\|_2 \leq \omega$, which is how we express the well-connectedness condition.

With those assumptions in mind, we can use vector decomposition via

$$r^\top W^t x = r^{\parallel\top} W^t x^\parallel + r^{\perp\top} W^t x^\perp$$

where $v^\parallel$ represents the parallel component of a vector to the stationary distribution π, and $v^\perp$ represents the perpendicular component.

The first term is equal between the terms (as $W^t \boldsymbol{\pi}_0 = \boldsymbol{\pi}_0$), so we can neglect it when calculating the discrepancy. Therefore, we can bound

$$\begin{aligned}
\sum_t \gamma^t (r^\top W^t x - r^\top W'^t x) &= \sum_t \gamma^t (r^{\parallel\top} W^t x^\parallel + r^{\perp\top} W^t x^\perp - r^{\parallel\top} W'^t x^\parallel - r^\top W'^t x) \qquad (3)\\
&\leq \sum_t \gamma^t (r^{\perp\top} W^t x^\perp - r^{\perp\top} W'^t x^\perp)\\
&\leq \left\|r^\perp\right\|_2 \left\|\sum_t \gamma^t (W^t - W'^t) x^\perp\right\|_2 \qquad (4)
\end{aligned}$$

Now, we tackle the second quantity, with two theorems that try and bound this gap. Our first theorem essentially bounds the discrepancy with no assumptions as to how close the MDPs are. Despite this, it already implies nontrivial bounds, which is inviting.

Theorem 5.1. *Assuming that W, W' are symmetric, nonnegative, and share a stationary distribution $\boldsymbol{\pi_0}$, and letting $\omega \leq 1$ be the max of the 2-norm of W, W' on $\boldsymbol{\pi_0}^{\perp}$, we have*

$$\left\| \sum_{t=0}^{\infty} \gamma^t W^t \boldsymbol{x} - \sum_{t=0}^{\infty} \gamma^t W'^t \boldsymbol{x} \right\|_2 \leq \frac{2\gamma\omega \left\| x^{\perp} \right\|_2}{1 - \gamma\omega}.$$

Proof. Note that

$$\begin{aligned}
\left\| \sum_{t=0}^{\infty} \gamma^t W^t \boldsymbol{x} - \sum_{t=0}^{\infty} \gamma^t W'^t \boldsymbol{x} \right\|_2 &= \left\| \sum_{t=0}^{\infty} \gamma^t W^t \boldsymbol{x}^{\perp} - \sum_{t=0}^{\infty} \gamma^t W'^t \boldsymbol{x}^{\perp} \right\|_2 \\
&= \left\| \sum_{t=1}^{\infty} \gamma^t W^t \boldsymbol{x}^{\perp} - \sum_{t=1}^{\infty} \gamma^t W'^t \boldsymbol{x}^{\perp} \right\|_2 \\
&\leq \sum_{t=1}^{\infty} \gamma^t \left\| W^t \boldsymbol{x}^{\perp} \right\|_2 + \sum_{t=1}^{\infty} \gamma^t \left\| W'^t \boldsymbol{x}^{\perp} \right\|_2 \\
&\leq 2 \sum_{t=1}^{\infty} \gamma^t \omega^t \left\| x^{\perp} \right\|_2 \\
&= \frac{2\gamma\omega \left\| x^{\perp} \right\|_2}{1 - \gamma\omega}
\end{aligned}$$

□

As a consequence, we have

Corollary 5.1.1. *Assuming that W, W' are symmetric, nonnegative, and share a stationary distribution $\boldsymbol{\pi_0}$, and letting $\omega \leq 1$ be the max of the 2-norm of W, W' on $\boldsymbol{\pi_0}^{\perp}$, we have*

$$|\eta[\pi] - \hat{\eta}[\pi]| \leq \frac{2\gamma\omega \left\| r^{\perp} \right\|_2 \left\| x^{\perp} \right\|_2}{1 - \gamma\omega}$$

after using Equation 4.

If $1 - \omega$ is multiples larger than $1 - \gamma$ (which is on the order of 0.01), this greatly reduces the value of the discrepancy. Of course, this proof doesn't even use the quality of our learned dynamics, which our next result hopes to remedy.

Theorem 5.2. *Assuming that W, W' are symmetric, nonnegative, and share a stationary distribution $\boldsymbol{\pi_0}$, and letting $\omega \leq 1$ be the max of the 2-norm of W, W' on $\boldsymbol{\pi_0}^{\perp}$, we have*

$$\left\| \sum_{t=0}^{\infty} \gamma^t W^t \boldsymbol{x} - \sum_{t=0}^{\infty} \gamma^t W'^t \boldsymbol{x} \right\|_2 \leq \frac{2\sqrt{\alpha \left\| x^{\perp} \right\|_2}}{(1 - \gamma\sqrt{\omega})^{3/2}}.$$

Proof. Let $y_t = (W^t - W'^t)\boldsymbol{x}$. Then, note that $y_t = (W^t - W'^t)\boldsymbol{x}^{\perp}$ because $W^t \boldsymbol{x}^{\parallel} = W'^t \boldsymbol{x}^{\parallel}$. Now, we make two observations, which we encode in a Lemma.

Lemma 5.3. $\left\| (W^t - W'^t)\boldsymbol{x}^{\perp} \right\|_2 \leq 2 \min(\omega^t \left\| \boldsymbol{x}^{\perp} \right\|_2, 1 - (1 - \alpha)^t)$

Proof. The first part follows directly from the expansion condition. The second part follows from

$$\left\|(W^t - W'^t)\boldsymbol{x}^\perp\right\|_2 \leq \left\|(W^t - W'^t)\boldsymbol{x}^\perp\right\|_1 \leq \left\|(W^t - W'^t)\right\|_1 \left\|\boldsymbol{x}^\perp\right\|_1$$

and Lemma 4.2. □

Lemma 5.4. *We use the following elementary facts*

1. $$\frac{1}{(1-x)^{3/2}} = \sum_{i=0}^{\infty} \binom{i+1/2}{1/2} x^i$$

2. *For integers t,*
$$\binom{t+1/2}{1/2} \geq \sqrt{t}$$

Proof. 1. This is well known.

2. Consider the quantities
$$A = \binom{t+1/2}{1/2} = \frac{t+1/2}{t} \cdot \frac{t-1/2}{t-1} \cdots \frac{3/2}{1}$$
and
$$B = \frac{t}{t-1/2} \cdot \frac{t-1}{t-3/2} \cdots \frac{2}{3/2}.$$
Note that $A > B$ and that $AB = t + 1/2$ – hence, $A \geq \sqrt{t}$.

□

Lemma 5.5.
$$2\sum_{t=0}^{\infty} \gamma^t \min(c\omega^t, 1-(1-\alpha)^t) \leq 2\frac{\sqrt{c\alpha}}{(1-\gamma\sqrt{\omega})^{3/2}}$$

Proof. We have

$$\begin{aligned}
2\sum_{t=0}^{\infty} \gamma^t \min(c\omega^t, 1-(1-\alpha)^t) &= 2\sum_{t=1}^{\infty} \gamma^t \min(c\omega^t, 1-(1-\alpha)^t) \\
&\leq 2\sum_{t=1}^{\infty} \gamma^t \min(c\omega^t, \alpha t) \\
&\leq 2\sum_{t=1}^{\infty} \gamma^t (\sqrt{t}\sqrt{\alpha c}\omega^{t/2}) \\
&= 2\sqrt{\alpha c}\sum_{t=1}^{\infty} \sqrt{t}(\gamma\sqrt{\omega})^t \\
&\leq 2\sqrt{\alpha c}\sum_{t=1}^{\infty} \binom{t+1/2}{1/2}(\gamma\sqrt{\omega})^t \\
&= \frac{2\sqrt{\alpha c}}{(1-\gamma\sqrt{\omega})^{3/2}}
\end{aligned}$$

where the second inequality uses the fact that the geometric mean is larger than the minimum, the third inequality follows from Lemma 5.4.2 and the last equality follows from Lemma 5.4.1 □

To conclude, we note that

$$\begin{aligned}\left\|\sum_{t=0}^{\infty}\gamma^t W^t \boldsymbol{x} - \sum_{t=0}^{\infty}\gamma^t W'^t \boldsymbol{x}\right\|_2 &\leq \sum_{t=0}^{\infty}\gamma^t \left\|(W^t - W'^t)x\right\|_2 \\ &= \sum_{t=0}^{\infty}\gamma^t \left\|(W^t - W'^t)x^{\perp}\right\|_2 \\ &= \sum_{t=0}^{\infty} 2\gamma^t \min(\omega^t \left\|x^{\perp}\right\|_2, 1-(1-\alpha)^t) \\ &\leq 2\frac{\sqrt{\alpha \left\|x^{\perp}\right\|_2}}{(1-\gamma\sqrt{\omega})^{3/2}}\end{aligned}$$

□

Likewise to before, we can obtain

Corollary 5.5.1. *Assuming that W, W' are symmetric, nonnegative, and share a stationary distribution π_0, and letting $\omega \leq 1$ be the max of the 2-norm of W, W' on $\pi_0^{\perp}$, we have*

$$\eta[\pi] - \hat{\eta}[\pi] \leq \frac{2\sqrt{\alpha \left\|x^{\perp}\right\|_2}\gamma \left\|r^{\perp}\right\|_2}{(1-\gamma\sqrt{\omega})^{3/2}}$$

after using Equation 4.

This could yield a significant improvement on Theorem 4.1.

6 Reward Shaping: Sparsification

6.1 A progress check

Here, we take stock of the progess made in model-based offline RL, and where there is still opportunity. In some sense, the results of [YTY+20] (Theorem 3.1) already have "solved model-based offline RL", in that if the uncertainty could be quantified exactly, their solution is provably optimal. Where does that leave us? One way to close the loop, so to speak, is to solve uncertainty quantification more accurately. However, this is a task with relatively few degrees of freedom, so we will actually choose not to work on this particular task.

One route that we'll take instead is, inspired by our results like Corollary 5.1.1 and Corollary 5.5.1, to try reward shaping. Let's try to change the overall shape of the MDP, instead of just the specific uncertainty measures, so that we can reduce the value of $|r^{\perp}|_2$. Specifically, our results suggest that discrepancies may be lowered by sparsifying the rewards, so in what follows we explain how we might do this on various tasks, and how certain tasks have rewards which are naturally sparse.

6.2 Common RL Benchmark Tasks

We have now seen that reward shaping can potentially drastically reduce our bounds on these convergences. Here we argue that some well-studied tasks in openAI gym and D4RL [FKN+20], the primary offline RL benchmark dataset, can be naturally phrased as sparse reward tasks. Sparse reward tasks are often very difficult for reinforcement learning algorithms.

6.2.1 A sparse reward example

Consider this MDP, where a reward is given at exactly one leaf after $|H|$ steps. This test case, from [KAL16] and others, is known to be very difficult for standard (online) reinforcement learning algorithms. If they never explore the leaf node, they will never be able to learn the ideal path. Since they have no signal as to the correct path without finding it exactly, this case is very difficult, and is frequently used as an example for why sparse rewards make problems more difficult, not easier.

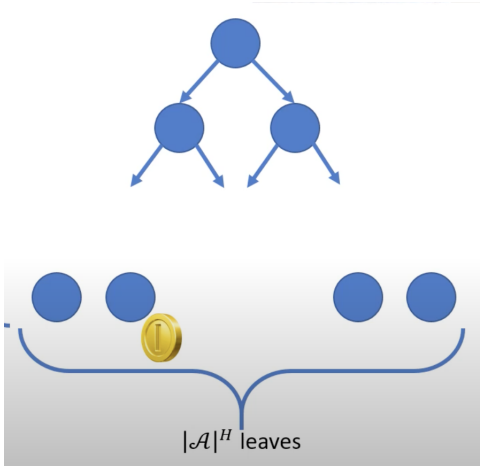

Figure 1: A challenging MDP.

This task is even more daunting for an offline RL algorithm, in which case the data collection process may not find any rewards at all. If no rewards are found, no approach can work, because the dataset of transitions collected becomes indistinguishable from one collected from a dataset of transitions drawn from a tree with no rewards.

However, model-based offline RL algorithms may have a better shot on this particular problem, because they are capable of planning arbitrarily far into the future. The reason this particular task is difficult for a standard reinforcement learning algorithm is because those algorithms are incapable of planning far into the future, a drawback which does not exist for offline planning. This is also where the possible incorporation of prior knowledge may be very useful in reducing the size of the search space.

6.2.2 RL Benchmark Tasks

Here we discuss the tasks that offline RL algorithms are generally evaluated on, and discuss how some already have a naturally sparse reward structure or are close to having a sparse reward structure. This may make them more amenable to analysis. Furthermore, explicit examples allow us to clarify the upshots of our analysis.

Cartpole This is a relatively basic reinforcement learning task where the idea is to keep a pole balanced as long as possible. One point of reward is given for each timestep where the pole is still upright. The rewards become zero once the cartpole falls over.

While the rewards in this case are appear to be the opposite of dense (as they are given almost every timestep), our use of the vector decomposition technique shows that analyzing this particular problem is essentially equivalent to analyzing the problem where there is one point of

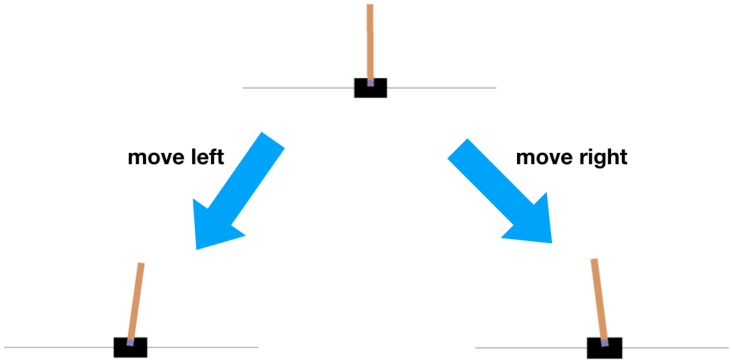

Figure 2: The Cartpole task

punishment when the cartpole falls over, which is clear. However, this new problem may have bounds that are easier to analyze, because the rewards have become sparser.

halfcheetah, hopper, walker2d These are standard tasks where the rewards are proportional to the velocity of the object. There is also a natural sparsification of these rewards, in which the agent gets one unit of reward for crossing certain milestones.

Atari This consists of games similar to space invaders and Offline RL is known to work particularly well in these cases, needing almost no modifications to standard off-policy algorithms. Some Atari games have particularly sparse rewards. The two sparsest games in the Atari 60 games dataset are Venture and Montezuma's revenge. According to [ASN20], the strongest results for the offline learner are on Venture, and the weakest are on Montezuma's revenge.

7 Conclusion

In this paper, we have proven new upper bounds on the discrepancy gap which decrease when we sparsify the rewards. We stated reasons for why sparse rewards may not be as harmful to offline reinforcement learning algorithms due to their ability to see farther in the future and be less dependent on intermediate rewards. Finally, we explained natural sparsity structures in many common RL tasks.

8 Outlook

A pleasantly surprising facet of the machine learning world is that systems work far better than they ought to. While this is an incredibly useful property for practitioners, it often creates frustration with these algorithms when proving results about it. Offline RL seems like a particularly ripe area for this. Sometimes direct approaches work on offline RL, like in Atari, [ASN20]. Other times, it's very hard to beat even simple tasks like Cheetah or Half-walker, where offline RL approaches are far less effective [YTY+20].

The problem of low observed discrepancies is a strange one, and it seems to be inconsistent in practice. By casting this problem spectrally, we hope that we can clarify just why exactly these

algorithms can do well, and measure whether tools like sparsifying rewards can be helpful. Offline RL has great potential to make RL safer, by allowing RL to be simply run offline and then deployed. We hope to be part of this effort.

9 Future Work and Questions

We've shown bounds for the discrepancy that become lower than some known bounds when the rewards are particularly sparse. This suggests several future directions.

1. We could attempt to show stronger discrepancy bounds on the RL benchmark tasks, by showing that the rewards have a fundamentally sparse structure even when they appear to be dense.
2. We could also attempt to plan through MDPs using sparser reward functions, and see if that could lead to improved performance.
3. Finally, and most ambitiously, our results suggest that even if spectral expansion is not already present in the problem, we may want to induce it in the MDP by creating an auxiliary MDP which does have higher expansion, and solving that problem instead.

Q: I'd also like to ask for guidance as to strengthening the bounds in Theorem 5.2.

10 Acknowledgements

We'd like to first and foremost thank Salil Vadhan for his very helpful discussions and patience in explaining concepts, especially suggesting the total variation trick and the use of vector decomposition. We would also like to thank Yang Liu for helpful discussions.

References

[ASN20] Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, 2020.

[FKN+20] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

[KAL16] Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Contextual-mdps for pacreinforcement learning with rich observations. *arXiv preprint arXiv:1602.02722*, 2016.

[KRNJ20] Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. *NeurIPS*, 2020.

[KS02] Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2-3):209–232, 2002.

[LXL+18] Yuping Luo, Huazhe Xu, Yuanzhi Li, Yuandong Tian, Trevor Darrell, and Tengyu Ma. Algorithmic framework for model-based deep reinforcement learning with theoretical guarantees. In *International Conference on Learning Representations*, 2018.

[YTY+20] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *NeurIPS*, 2020.