A COMPARISON OF MODEL-FREE AND MODEL-BASED REINFORCEMENT LEARNING IN THE OFFLINE SETTING

JAMES WIDDOWSON

Abstract. We consider the Offline Reinforcement Learning (RL) problem in a general-purpose (non-tabular), model-based setting through the lens of a novel algorithm based on the log-loss. We obtain a suite of (nearly) Horizon-free regret bounds and compare these results to those achieved in the model-free setting; specifically, we focus on the natural analogue of our algorithm within the model-free case: offline Imitation Learning (IL). Through this, we show that model-based algorithms can offer a viable alternative to the model-free approach when the true MDP has some low-dimensional structure that can be exploited. It remains an open problem to fully extend our results to the unknown-rewards setting, although we present some progress in this direction alongside further suggestions for future work.

1. Introduction

Modern Reinforcement Learning approaches are generally characterised as being either Model-based or Model-free. The difference stems from how the agent approaches planning; in the Model-based setting, algorithms construct models of the environment to help guide action selection. In contrast, Model-free algorithms typically forego building models of the environment and instead opt to model the Value function/actions directly. Whilst both methods have seen empirical success [1], a rigorous understanding of how the sample complexity compares between the two is still not yet fully understood; in the offline case, this issue is further exacerbated since most of the prior work focuses on a comparison within the online setting ([1], [2]). We note that a concrete answer to the sample-complexity question posed above would serve two valuable purposes: 1) it would help direct practical implementations towards the best approach based on the specific problem, and; 2) it would serve as a baseline for a more complete characterisation of the differences between both model-free and model-based methods. Furthermore, we note that, even when existing results do consider the offline case, not all extend beyond the tabular framework in which the State and Action space are small; since practically *all* modern RL algorithms employ some form of function approximation [3], a greater theoretical understanding of general-purpose RL algorithms is thus much desired. In this paper, we will attempt to tackle both issues at once, proving results that hold in the offline, model-based setting for (almost) arbitrary transition-function classes and concluding with a comparison to the model-free analogues obtained in [4].

1.1. Our contributions.

In this paper, we present, to the best of our knowledge, one of the first theoretical comparisons between model-based and model-free approaches for the offline, general-purpose (non-tabular) setting based on the log-loss. We demonstrate theoretically that model-based methods can *potentially* match the results guaranteed by the model-free analogues in [4]; importantly, we show that the regret for the model-based approach scales with the log-complexity of the transition function class, as opposed to the log-complexity of the policy class proven for Imitation Learning. The importance of such a result stems from the fact that, in practice, there often exists some lower-dimensional structure for the transition class which is not shared when considering the corresponding policy class ([1], [2]); as a consequence, we demonstrate that certain problems will potentially benefit from the use of a model-based approach. With our work, we take the first steps towards characterising such problems.

2. A brief note on notation

For much of our results, we will work on a trajectory-level basis; this is a departure from the more traditional methods that involve working with the individual state-action pairs, (s, a). As a result, most of our notation will reflect this change and hence this section can be seen as a gentle introduction to the layout in the remainder of our paper. First, we introduce some general, non-RL specific notation: for any $H \in \mathbb{N}$ we

use $[H]$ to represent the set of integers $\{1, ..., H\}$; $\Delta(\cdot)$ to denote an arbitrary distribution over the argument; $\wedge$ to represent the minimum between a pair of values; and $\lesssim$ to denote an inequality up to constants.

2.1. **Markov decision processes.** We study the offline Reinforcement Learning problem in a finite Horizon setting. We define a finite Horizon MDP as the tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, H, P, R)$ which consists of a (finite) state space, $\mathcal{S}$; a (finite) action space, $\mathcal{A}$; a fixed time horizon, H; a transition function $P := \{P_h\}_{h=0}^H$ where $P_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$; and a reward function $r := \{r_h\}_{h=1}^H$ where $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$. We note that we will sometimes slightly abuse notation and use $P_h(\cdot \mid s_h, a_h)$ to denote the transition probability induced by the P_h formally defined above. We define a (potentially stochastic) policy as the mapping $\pi := \{\pi_h\}_{h=1}^H$ where $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$. When we work with entire trajectories, we will consider the natural extension of the above definitions by defining $P(s_{1:H}, a_{1:H}) := P_0(s_1) \prod_{t=1}^H P_t(s_{t+1} \mid s_t, a_t)$ and $\pi(a_{1:H}|s_{1:H}) := \prod_{t=1}^H \pi_t(a_t \mid s_t)$. For a given transition model and policy, we use $\mathbb{P}^{P,\pi}(s_{1:H}, a_{1:H}) := P_0(s_1) \prod_{t=1}^H P_t(s_{t+1} \mid s_t, a_t)\pi_t(a_t \mid s_t)$ to denote the trajectory probability over the observations $\{s_{1:H}, a_{1:H}\}$; we use $\mathbb{P}^{P,\pi}$ (omitting the argument) to refer to the distribution induced by the above probabilities. When we want to splice the above distributions to a particular range of time-steps, we will use a subscript of the splicing indices to signify this, although this operation will be made much clearer in subsection 2.2.

We use

$$\mathbb{E}^{P,\pi} f(s_{1:H}, a_{1:H}) := \sum_{s_{1:H}, a_{1:H}} f(s_{1:H}, a_{1:H}) \cdot \mathbb{P}^{P,\pi}(s_{1:H}, a_{1:H})$$

to denote the expectation of f over the trajectory distribution induced by P and π, where the summation is over $(s_{1:H}, a_{1:H}) \in (S \times \mathcal{A})^H$. We use $J(P, \pi) := \mathbb{E}_{\tau(P,\pi)} r(s_{1:H}, a_{1:H})$ to denote the expected reward under transition model P and policy π, where $r(s_{1:H}, a_{1:H}) := \sum_{t=1}^H r(s_t, a_t)$. We will specify our results to the case where $r(s_{1:H}, a_{1:H})$ is uniformly bounded above by some constant R: in the literature, this is commonly referred to as the *dense-reward setting*. We note that this setting is often the focus of prior work on Imitation Learning ([5], [6]), and hence considering it here allows for a more complete comparison between our model-based algorithm and IL.

The Q- and Value functions are defined similarly as

$$Q_h^{P,\pi}(s, a) := \mathbb{E}^{P,\pi} \left[\sum_{t=h}^H r(s_t, a_t) \mid s_h = s, a_h = a \right]$$

and $V_h^{P,\pi}(s) := Q_h^{P,\pi}(s, \pi(s))$. It is helpful to note the connection between $J(P, \pi)$ and the more familiar notion of a Value function through the relation $J(P, \pi) = \sum_{s \in \mathcal{S}} P_0(s) V_1^{P,\pi}(s)$; that is, $J(P, \pi)$ is equivalent to $V_1^{P,\pi}(s)$ if s is the fixed starting-state. We say a policy, π, is optimal for a given transition model P if $V_h^{P,\pi}(s) = \max_{\pi'} V_h^{P,\pi'}(s)$ for all $s \in \mathcal{S}$[1] and $h \in [H]$, and use $\pi \in \text{opt}(P)$ to denote that π is within the set of optimal policies for P.

We define a transition function class $\mathcal{P} := \{(f_h)_{h=0}^H \mid f_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})\}$ to be an arbitrary set of transition functions pre-specified by the user. We note that a set of necessary assumptions concerning this choice will be listed in subsection 2.4.

2.2. **Splicing notation.** When we work directly with the induced trajectory distributions, it will sometimes be helpful to truncate the $(s_h, a_h)_{h=1}^H$ pairs under consideration. To this end, for an arbitrary $\mathbb{P}^{P,\pi}$, we define the spliced version as $\mathbb{P}_{0:h}^{P,\pi} := P_0(s_1) \prod_{t=1}^h P_t(s_{t+1} \mid s_t, a_t)\pi_t(a_t \mid s_t)$ for some $h \in [H]$. When we want to sample a state at time-step h, we will sample $s_1 \sim P_0$, $a_1 \sim \pi_1(\cdot \mid s_1)$ and so on, up until $s_h \sim P_{h-1}(\cdot \mid s_{h-1}, a_{h-1})$; we will denote this process by $s_h \sim d_h^{P,\pi}$ and refer to $d_h^{P,\pi}$ as the state-occupancy measure at time-step h induced by P and π.

[1]Note that we only really need to consider the sets where s is possible under P

2.3. **Information theory.** We define the squared Hellinger distance between distributions $\mathbb{P}$ and $\mathbb{Q}$ over the space $\bigotimes_{i=1}^{H}(\mathcal{S}\times\mathcal{A})_i$, as

$$D_H^2(\mathbb{P}||\mathbb{Q}) := \sum_{s_{1:H},a_{1:H}} \left(\sqrt{\mathbb{P}(s_{1:H},a_{1;H})} - \sqrt{\mathbb{Q}(s_{1:H},a_{1:H})}\right)^2$$

We note that the squared Hellinger distance belongs to the class of f-divergences and, although seemingly implied by its name, does *not* in general satisfy the triangle inequality. Interestingly, the Hellinger distance, $D_H(\mathbb{P}||\mathbb{Q}) := \sqrt{D_H^2(\mathbb{P}||\mathbb{Q})}$, *does* define a proper metric on the space of probability distributions, a result which we will prove useful in the sequel.

2.4. **The usual assumptions.** Here, we briefly recall the standard assumptions that persist for all results in our paper. We operate in the finite state and action space setting with fixed time-horizon H, although we allow for $|\mathcal{S}|$ and $|\mathcal{A}|$ to potentially be very large. We assume that each $r_h \in [0,1]$ and $r(s_{1:H},a_{1:H})$ is uniformly bounded above by R for any trajectory; we additionally assume that this function is known and hence each MDP is fully defined by a choice of $P = \{P_h\}_{h=0}^{H}$. We will always use $\hat{\pi}$ to denote the policy returned by Algorithm 1 and π^* to denote the optimal policy under the true MDP, P^*. We additionally assume that both $\hat{\pi}$ and π^* are deterministic, although we note that this is a reasonable assumption in practice since every MDP, P, has at least one deterministic policy $\in \text{opt(P)}$. When we fix a transition class, $\mathcal{P}$, we assume that: 1) $|\mathcal{P}|$ is finite, and; 2) $\mathcal{P}$ *always* contains the true transition function, P^*. We note that the assumptions on $\mathcal{P}$ can be relaxed with the addition of a misspecification-penalty term, although we do not consider this here [4].

3. The algorithm and some results for the Hellinger distance

We now formally define the offline paradigm under which we are operating and introduce the model-based algorithm on which the majority of our results are based. We then provide a collection of information theoretic results concerning the squared-Hellinger distance.

3.1. **The offline setting.** We assume that we have a dataset $\{(s_{1:H},a_{1:H})^i\}_{i=1}^n$ where the trajectories $(s_{1:H},a_{1:H})^i$ are drawn *iid* from $\sim \mathbb{P}^{P^*,\pi^*}$. In more practical terms, this corresponds to collecting n samples of trajectory data under P^* by using π^* as the rollout policy; we note that π^* is assumed to be optimal here, although it's likely that this assumption can be relaxed for many of the results we present. Using this data, we assume that both the model-free and model-based algorithm extract some fitted policy, $\hat{\pi}$, without any further interaction with the environment, and we will study the regret accumulated across a singular episode. We reiterate the fact that our results hold even when $\mathcal{S}$ and $\mathcal{A}$ are not small enough to be feasibly enumerated and stored in memory directly[2], and hence are more compatible with more modern, function-approximation-based approaches to RL. It remains a question of future work to consider the hybrid setting of operating both on- and offline, as is done in [4].

Algorithm 1. *(Our model-based algorithm) Consider the setting where we have a dataset* $\{(s_{1:H},a_{1:H})_i\}_{i=1}^n =: \mathcal{D}$ *with trajectories drawn iid from* $(s_{1:H},a_{1:H})_i \overset{iid}{\sim} \mathbb{P}^{P^*,\pi^*}$ *and a pre-defined transition function class,* $\mathcal{P}$*. Then our algorithm fits* $\hat{P} : (\mathcal{S}\times\mathcal{A})^H \to \Delta(\mathcal{S})^H \subset \mathcal{P}$ *under the objective*

$$\hat{P} := \underset{P\in\mathcal{P}}{\text{argmax}} \sum_{i\in\mathcal{D}} \log\left(P_0(s_1)\prod_{t=1}^{H} P_t(s_{t+1} \mid s_t,a_t)\right)$$

and returns $\hat{\pi}$ *by finding* $\hat{\pi} :\in \textit{opt}(\hat{P})$*.*

We note that the operation of finding $\hat{\pi} \in \text{opt}(\hat{P})$ after fitting $\hat{P}$ is non-trivial in general; in practice, this is a common limitation of model-based approaches and much work has been done on finding fast, approximate ways to compute such a policy ([7], [8]). For the purposes of this paper, we will not concern ourselves with the specifics, although we note that a significant limitation of our results is that they do not currently hold under misspecification of $\hat{\pi}$.

For our next result, it is helpful to note that, under a Markovian assumption, fitting $\hat{P} : (\mathcal{S}\times\mathcal{A})^H \to \Delta(\mathcal{S})^H$

[2]This is known as the tabular setting

is equivalent to fitting $P_t : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ for each $t \in [H]$ by the statement of our optimisation problem. The log-loss is considered, as in [4], since it has been shown to improve Horizon-dependence when utilised in an Imitation Learning framework ([5], [4]).

Theorem 1. *Consider the setting where we have a dataset $\{(s_{1:H}, a_{1:H})^i\}_{i=1}^n =: \mathcal{D}$ with trajectories drawn iid from $(s_{1:H}, a_{1:H})^i \overset{iid}{\sim} \mathbb{P}^{P^*,\pi^*}$. Then the algorithm that fits $\hat{P} : (\mathcal{S} \times \mathcal{A})^H \to \Delta(\mathcal{S})^H \subset \mathcal{P}$ under the objective*

$$\hat{P} := \underset{P \in \mathcal{P}}{\operatorname{argmax}} \sum_{i \in \mathcal{D}} \log \left(P_0(s_1) \prod_{t=1}^{H} P_t(s_{t+1} \mid s_t, a_t) \right)$$

has the same maximiser as

$$\underset{P \in \mathcal{P}}{\operatorname{argmax}} \sum_{i \in \mathcal{D}} \log \left(\mathbb{P}^{P,\pi^*} \left(\{s_{1:H}, a_{1:H}\}^i \right) \right)$$

Hence, fitting $\hat{P}$ as above, we obtain

$$D_H^2(\mathbb{P}^{\hat{P},\pi^*} || \mathbb{P}^{P^*,\pi^*}) \leq \frac{6 \log(2|\mathcal{P}|\delta^{-1})}{n}$$

with probability $\geq 1 - \delta$.

Proof. We first show that both objectives have the same maximiser; this follows immediately from the fact that

$$\begin{aligned} \sum_{i \in \mathcal{D}} \log \left(\mathbb{P}^{P,\pi^*}(\{s_{1:H}, a_{1:H}\}^i) \right) &= \sum_{i \in \mathcal{D}} \left[\log \left(P_0(s_1) \prod_{t=1}^{H} P_t(s_{t+1} \mid s_t, a_t) \right) + \log \left(\prod_{t=1}^{H} \pi_t(a_t|s_t) \right) \right] \\ &\propto \sum_{i \in \mathcal{D}} \log \left(P_0(s_1) \prod_{t=1}^{H} P_t(s_{t+1} \mid s_t, a_t) \right) \end{aligned}$$

To finish our result, we note that the upper bound on $D_H^2(\mathbb{P}^{\hat{P},\pi^*} || \mathbb{P}^{P^*,\pi^*})$ follows from Proposition B.1 in [9] using the density class $\{\mathbb{P}^{P,\pi^*}\}_{P \in \mathcal{P}}$. Since, we have $|\{\mathbb{P}^{P,\pi^*}\}_{P \in \mathcal{P}}| = |\mathcal{P}|$, the result follows. □

At first glance, this result looks very similar to the one presented in [4]. In essence, we have shown that fitting $\hat{P}$ as the MLE within the class P allows us to obtain an upper bound on $D_H^2(\mathbb{P}^{\hat{P},\pi^*} || \mathbb{P}^{P^*,\pi^*})$ as a function of the number of samples, n, and the size of our transition function class. However, we stress the fact that the trajectory-induced distributions used in the squared-Hellinger distance are not consistent across the two papers; in theirs, they obtain a bound on $D_H^2(\mathbb{P}^{P^*,\pi^*} || \mathbb{P}^{P^*,\pi'})$ where π' is a policy fit directly under their algorithm without building $\hat{P}$. This is notably distinct from the bound we obtain since ours does not involve the fitted policy directly; instead, the squared-Hellinger distance we derive is a function of the $\hat{P}$ term which has no direct equivalence in the setting of their paper. Even though Theorem 3.1 in [4] holds for general sets of policies, we have no way to relate $D_H^2(\mathbb{P}^{P^*,\pi^*} || \mathbb{P}^{P^*,\hat{\pi}})$ with $D_H^2(\mathbb{P}^{P^*,\pi^*} || \mathbb{P}^{\hat{P},\pi^*})$ and hence it is difficult to adapt their result directly. Fortunately, we will show that other methods can be used to circumvent this issue, although we will not be able to match the scaling-factors of their bound in general. We note that this is somewhat expected since the distributions they have control over, $\mathbb{P}^{P^*,\pi^*}$ and $\mathbb{P}^{P^*,\pi'}$, are more closely aligned with those used in the statement of regret.

3.2. **Results for the Hellinger distance.** We will now present some standard results for the Hellinger distance which will prove very helpful when we consider stochastic transition classes, $\mathcal{P}$. The first is a simple application of the Data Processing inequality from Information Theory.

Lemma 1. *(Specific case of Data Processing inequality) For any two distributions P_X, $P_{X'}$ defined on a measurable space, $(\mathcal{X}, \mathcal{E})$, we have that*

$$D_H^2\left(Bern(P_X(E)), Bern(P_{X'}(E))\right) \leq D_H^2(P_X, P_{X'})$$

for all $E \subset \mathcal{E}$.

Proof. See Appendix section A □

Our next result is a general, functional inequality which will allow us to relate differences in expectation with the squared-Hellinger distance between the corresponding measures. It is helpful to note that similar results can be derived, as is done in [4], where the RHS does not depend on R; in this case, the expectations under the square root are replaced by the second moment of $h(X)$. For our purposes, it will be sufficient to consider the result as stated, although we note that the dependence on R could potentially be improved if we can show that $\mathbb{E}_{\mathbb{P}}\left[h^2(X)\right]$ and $\mathbb{E}_{\mathbb{Q}}\left[h^2(X)\right]$ are bounded uniformly by R.

Theorem 2 (Change of measure result)**.** *Let $(\mathcal{X}, \mathcal{E})$ be a measurable space, $\mathbb{P}$ and $\mathbb{Q}$ be two probability distributions defined over $(\mathcal{X}, \mathcal{E})$, and $h : \mathcal{X} \to \mathbb{R}$ be a measurable function uniformly bounded in absolute value by R. Then,*

$$|\mathbb{E}_{\mathbb{P}}\left[h(X)\right] - \mathbb{E}_{\mathbb{Q}}[h(X)]| \leq \sqrt{2R \cdot D_H^2(\mathbb{P}||\mathbb{Q})\left(\mathbb{E}_{\mathbb{P}}|h(X)| + \mathbb{E}_{\mathbb{Q}}|h(X)|\right)}$$

Proof. See Appendix section A

□

To conclude this subsection, we present an information-theoretic result which will allow us to control the difference in specific pairs of Value functions as a function of the trajectory-wise squared-Hellinger distance, $D_H^2(\mathbb{P}^{P^*,\pi^*}||\mathbb{P}^{\hat{P},\pi^*})$. This will prove crucial when we consider the stochastic-$\mathcal{P}$ case.

Lemma 2. *Define $P^h := \{P_t^*\}_{t=1}^h \cup \{\hat{P}_t\}_{t=h+1}^H$ for any $h \in [H]$, then*

$$D_H^2(\mathbb{P}^{P^*,\pi^*}||\mathbb{P}^{P^h,\pi^*}) \leq 4D_H^2(\mathbb{P}^{P^*,\pi^*}||\mathbb{P}^{\hat{P},\pi^*})$$

Proof. See Appendix section A □

The below corollary follows almost immediately from Lemma XXX.

Corollary 1. *Define $\pi^h := \hat{\pi}_h \cup \{\pi_t^*\}_{t \neq h}$ and recall that $P^h := \{P_t^*\}_{t=1}^h \cup \{\hat{P}_t\}_{t=h+1}^H$ for any $h \in [H]$. Then, if $D_H^2(\mathbb{P}^{P^*,\pi^*}||\mathbb{P}^{P^*,\pi^h}) \leq \mu D_H^2(\mathbb{P}^{P^*,\pi^*}||\mathbb{P}^{\hat{P},\pi^*})$ for some $\mu > 0$[3], we have*

$$D_H^2(\mathbb{P}^{P^h,\pi^*}||\mathbb{P}^{P^*,\pi^h}) \leq (2+\sqrt{\mu})^2 D_H^2(\mathbb{P}^{P^*,\pi^*}||\mathbb{P}^{\hat{P},\pi^*})$$

Proof. By the triangle inequality for the Hellinger distance, we have

$$\begin{aligned} D_H(\mathbb{P}^{P^h,\pi^*}||\mathbb{P}^{P^*,\pi^h}) &\leq D_H(\mathbb{P}^{P^h,\pi^*}||\mathbb{P}^{P^*,\pi^*}) + D_H(\mathbb{P}^{P^*,\pi^*}||\mathbb{P}^{P^*,\pi^h}) \\ &\leq 2D_H(\mathbb{P}^{P^*,\pi^*}||\mathbb{P}^{\hat{P},\pi^*}) + \sqrt{\mu} D_H(\mathbb{P}^{P^*,\pi^*}||\mathbb{P}^{\hat{P},\pi^*}) \end{aligned}$$

and hence

$$D_H^2(\mathbb{P}^{P^h,\pi^*}||\mathbb{P}^{P^*,\pi^h}) \leq (2+\sqrt{\mu})^2 D_H^2(\mathbb{P}^{P^*,\pi^*}||\mathbb{P}^{\hat{P},\pi^*})$$

□

3.3. **A useful characterisation of regret.** To finish this chapter, we present a result that characterises regret in terms of the action disagreement probabilities. Whilst in the deterministic case it is fairly trivial to show that this term is upper-bounded by a function of $D_H^2(\mathbb{P}^{P^*,\pi^*}||\mathbb{P}^{\hat{P},\pi^*})$, we note that when P^* is stochastic the result becomes a little more involved. We draw attention to the fact that, although the bound holds in general for all choices of $\hat{\pi}$ and π^*, both of our proof techniques will require that π^* and $\hat{\pi}$ are deterministic. Fortunately, we note this is not too impactful an assumption since a deterministic optimal policy always exists for any MDP; however, it still remains an interesting question to consider stochastic policies since they have seen success in practical applications ([10], [11]). We conjecture that a different approach may be needed to extend our results to this setting, an observation in-line with what was done in [4].

Theorem 3. *(Useful characterisation of regret) Assume the usual assumptions from section 2.4 are satisfied; then, the following result holds*

$$J(P^*, \pi^*) - J(\hat{P}, \hat{\pi}) \leq R \cdot \mathbb{P}^{P^*,\pi^*}\left[a_{1:H} \neq \hat{\pi}(s_{1:H})\right]$$

for all choices of P^*, π^*, and $\hat{\pi}$.

[3]This is Definition 1 that we will see in Section 5

Proof. To start, we note that

$$\begin{aligned} J(P^*, \pi^*) &= \mathbb{E}^{P^*, \pi^*}\left[r(s_{1:H}, a_{1:H})\right] \\ &= \mathbb{E}^{P^*, \pi^*}\left[r(s_{1:H}, a_{1:H})\mathbb{1}\left(a_{1:H} = \hat{\pi}(s_{1:H})\right)\right] + \mathbb{E}^{P^*, \pi^*}\left[r(s_{1:H}, a_{1:H})\mathbb{1}\left(a_{1:H} \neq \hat{\pi}(s_{1:H})\right)\right] \end{aligned}$$

We now consider the first term:

$$\begin{aligned} \mathbb{E}^{P^*, \pi^*}\left[r(s_{1:H}, a_{1:H})\mathbb{1}\left(a_{1:H} = \hat{\pi}(s_{1:H})\right)\right] &:= \sum_{s_{1:H}, a_{1:H}} r(s_{1:H}, a_{1:H}) P^*(s_{1:H}, a_{1:H}) \pi^*(a_{1:H}|s_{1:H}) \mathbb{1}(a_{1:H} = \hat{\pi}(s_{1:H})) \\ &= \sum_{s_{1:H}, a_{1:H}} r(s_{1:H}, a_{1:H}) P^*(s_{1:H}, a_{1:H}) \pi^*(a_{1:H}|s_{1:H}) \hat{\pi}(a_{1:H}|s_{1:H}) \\ &\leq \sum_{s_{1:H}, a_{1:H}} r(s_{1:H}, a_{1:H}) P^*(s_{1:H}, a_{1:H}) \hat{\pi}(a_{1:H}|s_{1:H}) \\ &= \mathbb{E}^{P^*, \hat{\pi}} r(s_{1:H}, a_{1:H}) =: J(P^*, \hat{\pi}) \end{aligned}$$

To finish the proof, we rearrange to arrive at

$$\begin{aligned} J(P^*, \pi^*) - J(P^*, \hat{\pi}) &\leq \mathbb{E}^{P^*, \pi^*}\left[r(s_{1:H}, a_{1:H})\mathbb{1}(a_{1:H} \neq \hat{\pi}(s_{1:H})\right] \\ &\leq R \cdot \mathbb{P}^{P^*, \pi^*}\left[a_{1:H} \neq \hat{\pi}(s_{1:H})\right] \end{aligned}$$

as desired. □

4. A regret bound for deterministic $\mathcal{P}$ class

In this section, we assume that our transition functions are deterministic; that is, $P_t(s' \mid s, a) \in \{0, 1\}$ for all $s \in \mathcal{S}$, $a \in \mathcal{A}$, and $t \in [H]$. We recall that the trajectory-level transition function is defined as

$$P(s_{1:H}, a_{1:H}) = P_0(s_1) \prod_{t=1}^{H} P_t(s_{t+1} \mid s_t, a_t) \in \{0, 1\}$$

which we emphasize maps to values within the set $\{0, 1\}$[4] (the [4] is a footnote). Our first result shows that π^*_h and $\hat{\pi}_h$ agree for all $h \in [H]$ when our transition probabilities agree on a trajectory that is either $\hat{\pi}$- or π^*-optimal. This follows almost immediately from the observation that the Value function under a given *deterministic* transition function and *deterministic* policy can only accumulate rewards along a singular trajectory.

Lemma 3. *Let $\hat{P}$ and P^* be deterministic transition functions satisfying*

$$s_{1:H} = \hat{P}(\cdot, a_{1:H}) = P^*(\cdot, a_{1:H})$$

for some $(s_{1:H}, a_{1:H}) \in (\mathcal{S} \times \mathcal{A})^H$ and assume there exists an optimal policy such that $a_{1:H} = \hat{\pi}(s_{1:H})$ for some $\hat{\pi}$ within the set of optimal policies under $\hat{P}$. Then, there exists a (deterministic) optimal policy under P^* such that

$$a_{1:H} = \hat{\pi}(s_{1:H}) = \pi^*(s_{1:H})$$

In other terms, $\pi^*_h(s_h) = \hat{\pi}_h(s_H)$ for all $s_h \in s_{1:H}$

Proof. See Appendix section B □

Our next Lemma provides an upper bound on $\mathbb{P}^{P^*, \pi^*}\left[a_{1:H} \neq \hat{\pi}(s_{1:H})\right]$ in terms of the squared Hellinger distance used in Theorem 1. With this in hand, the regret bound follows trivially from an application of Theorem 3.

Lemma 4. *(Probability bound for deterministic $\mathcal{P}$) Let $\hat{P}$ be the transition function fitted using Algorithm 1 under a deterministic class, $\mathcal{P}$. Then, there exist deterministic policies $\hat{\pi} \in opt(\hat{P})$ and $\pi^* \in opt(P^*)$ such that*

$$\mathbb{P}^{P^*, \pi^*}\left[a_{1:H} \neq \hat{\pi}(s_{1:H})\right] \leq \frac{6 \log(2|\mathcal{P}|\delta^{-1})}{n}$$

with probability $\geq 1 - \delta$

Proof. See Appendix section B □

[4]We will shortly see why this is such a crucial assumption for our proof technique!

4.1. **Statement of the main result.** We are now ready to state the main result for this section, one which follows trivially from a combination of Lemma 4 and Theorem 3. After the statement, we will provide a discussion of the bound and compare our result to the one achieved in [4].

Theorem 4. *(A regret bound for deterministic $\mathcal{P}$) Consider the setting where we have the dataset $\{(s_{1:H}, a_{1:H})^i\}_{i=1}^n$ with trajectories drawn iid from $(s_{1:H}, a_{1:H})^i \overset{iid}{\sim} \mathbb{P}^{P^*,\pi^*}$ and $\hat{P}$ fit as in Algorithm 1 under a deterministic function class, $\mathcal{P}$. Then, under all the usual conditions, the associated optimal policy, $\hat{\pi}$, has regret bounded by*

$$\begin{aligned} J(P^*, \pi^*) - J(P^*, \hat{\pi}) &\leq R\frac{12\log(2|\mathcal{P}|\delta^{-1})}{n} \\ &\lesssim R\frac{\log(2|\mathcal{P}|\delta^{-1})}{n} \end{aligned}$$

with probability $\geq 1 - \delta$.

Proof. The result follows immediately from Theorem 3 and Lemma 4. □

We note that this bound decays at a fast rate, indicated by presence of the $\mathcal{O}(1/n)$ term, and maintains a dependence on R which matches the state-of-the-art for Imitation Learning under a similar loss [4]. The caveat here is that, whereas the analogous result in [4] held for potentially-stochastic instances of P^*, we require that $\mathcal{P}$ is limited to *only* deterministic transition functions and hence P^* must be deterministic [5]. Although this limits the applicability of Theorem 4, we note that the dependence on $|\mathcal{P}|$ is of interest in its own right since modelling the transition function is often thought to be of lower *complexity* than modelling the policy directly ([1], [12]). We will explore this thought further in Section 6, although we highlight that this observation offers an explanation as to why we might expect our algorithm to be more sample efficient than those from IL.

5. An unoptimised bound for stochastic $\mathcal{P}$

In this section, we generalise the results obtained in Section 4 and present a bound for the setting where $\mathcal{P}$ contains potentially-stochastic transition functions. To do so, we will first need to introduce some additional machinery in the form of a recoverability condition.

Definition 1. *(Recoverability condition) Recall that we have defined $\pi^h := \hat{\pi}_h \cup \{\pi^*_t\}_{t \neq h}$; then, we say that our problem satisfies a recoverability condition if*

$$D^2_H(\mathbb{P}^{P^*,\pi^*} || \mathbb{P}^{P^*,\pi^h}) \leq \mu \cdot D^2_H(\mathbb{P}^{P^*,\pi^*} || \mathbb{P}^{\hat{P},\pi^*})$$

for all $h \in [H]$. We call $\mu > 0$ the recoverability constant and note that it has the loose interpretation of implying that $\hat{\pi}_h$ cannot be too dissimilar to π^*_h if both $\hat{P}$ and P^* are 'close'.

Although such an assumption will prove crucial in allowing us to obtain a regret bound for the stochastic case, we note that more work needs to be done on understanding its implications. We unfortunately did not have enough time to give this consideration; however, it would certainly be an interesting question for future work to investigate whether such a μ can hold uniformly over all MDPs. Failing this, it would alternatively be useful to characterise certain classes of P^* for which such a constant can be obtained.

With this definition in hand, we now present a result analogous to Lemma 4 for the stochastic case. The crux of the proof is an action-gap argument (utilised in [13], [14]), which is then combined with some information-theoretic results to obtain an upper bound in terms of the squared-Hellinger distance. The unfortunate consequence of such a technique is that it introduces an unavoidable dependence on the Horizon through the use of a union bound across Q_h-functions; it remains a question of future work to see whether a more careful analysis can eliminate this constant.

[5] Recall that our realisability assumption enforces this

Theorem 5. *(Probability bound for stochastic $\mathcal{P}$) Consider the setting where we have the dataset $\{(s_{1:H}, a_{1:H})^i\}_{i=1}^n$ with trajectories drawn iid from $(s_{1:H}, a_{1:H})^i \overset{iid}{\sim} \mathbb{P}^{P^*,\pi^*}$ and fit $\hat{P}$ as in Algorithm 1, where $\mathcal{P}$ is now a class of (potentially stochastic) transition functions. Then, under all the usual assumptions,*

$$\begin{aligned}\mathbb{P}^{P^*,\pi^*}\left[a_{1:H} \neq \hat{\pi}(s_{1:H})\right] &\leq \sum_{h=1}^{H} \frac{12(2+\sqrt{\mu})\left[R \wedge (H-h)\right]}{\Omega_{min}(h)} \sqrt{D_H^2(\mathbb{P}^{P^*,\pi^*}||\mathbb{P}^{\hat{P},\pi^*})} \\ &\lesssim HR\sqrt{D_H^2(\mathbb{P}^{P^*,\pi^*}||\mathbb{P}^{\hat{P},\pi^*})}\end{aligned}$$

where μ is the recoverability constant defined above, π^* and $\hat{\pi}$ are the deterministic[6] optimal policies under P^* and $\hat{P}$, and $\Omega_{\min}(h)$ is a (potentially very small) scaling factor.

Proof. See Appendix section C □

5.1. **Statement of the main result for stochastic $\mathcal{P}$.** We now state a regret bound for the model-based algorithm that holds under general classes of $\mathcal{P}$. Importantly, this extends the result presented in Section 4 since it allows for the consideration of stochastic transitions; the caveat in this respect, is that we trade off a Horizon-free dependence for this increased generality. We will provide a discussion of this fact, alongside a comparison with the results presented in [9], after the statement of the bound.

Theorem 6. *(A regret bound for stochastic $\mathcal{P}$) Consider the setting where we have the dataset $\{(s_{1:H}, a_{1:H})^i\}_{i=1}^n$ with trajectories drawn iid from $(s_{1:H}, a_{1:H})^i \overset{iid}{\sim} \mathbb{P}^{P^*,\pi^*}$ and fit $\hat{P}$ as in Algorithm 1 under a general function class, $\mathcal{P}$. Then, under all the usual conditions, the associated optimal policy, $\hat{\pi}$, has regret bounded by*

$$\begin{aligned}J(P^*,\pi^*) - J(P^*,\hat{\pi}) &\leq 12\sqrt{6}(2+\sqrt{\mu}) \cdot R \sum_{h=1}^{H} \frac{R \wedge (H-h)}{\Omega_{min}(h)} \sqrt{\frac{\log(2|\mathcal{P}|\delta^{-1})}{n}} \\ &\lesssim HR^2\sqrt{\frac{\log(2|\mathcal{P}|\delta^{-1})}{n}}\end{aligned}$$

with probability $\geq 1-\delta$

Proof. The result follows immediately from an application of Theorem 1, Theorem 3, and Theorem 5. The order bound follows from noting that

$$\begin{aligned}12\sqrt{6}(2+\sqrt{\mu}) \cdot R \sum_{h=1}^{H} \frac{R \wedge (H-h)}{\Omega_{\min}(h)} \sqrt{\frac{\log(2|\mathcal{P}|\delta^{-1})}{n}} &\leq 12\sqrt{6}(2+\sqrt{\mu}) \cdot R^2 \sqrt{\frac{\log(2|\mathcal{P}|\delta^{-1})}{n}} \sum_{h=1}^{H} \frac{1}{\Omega_{\min}(h)} \\ &\lesssim HR^2\sqrt{\frac{\log(2|\mathcal{P}|\delta^{-1})}{n}}\end{aligned}$$

□

We note that Theorem 6 is weaker than Theorem 4 in 2 respects: 1) we incur a HR scaling cost over the deterministic result; and 2) the regret now scales at the slower rate of $\mathcal{O}(1/\sqrt{n})$. Whereas the first issue may potentially be mitigated through a more careful analysis, we conjecture that issue 2) may be fundamentally unavoidable without adapting the fitting procedure outlined in Algorithm 1. One of the reasons for this, stems from the observation seen for stochastic policy-classes in [4]; when fitting with respect to a stochastic class, they show that a lower bound which scales at a rate of $\mathcal{O}(1/\sqrt{n})$ is unavoidable in general for the Imitation Learning setting[7]. It remains a question of future work to see whether such an observation extends to the model-based case, although we note that, at the very least, a stronger result than Theorem 2 would be needed in order to disprove the above observation.

Despite the incurred HR scaling cost, we reiterate that the main advantage of our approach lies in the fact that the transition class, $\mathcal{P}$, often exhibits a lower *complexity* than modelling the policies directly ([1], [12]). In practice, the additional HR scaling factor may be dwarfed by the difference between $|\mathcal{P}|$ and $|\Pi|$ for some general policy-class Π; in this respect, the number of trajectories, n, needed to obtain a suitable control on

[6]This is crucial for the argument we present, although it may be possible to relax this assumption with further work

[7][15] show that a faster rate can be obtained in the tabular case

regret could *potentially* be much lower under Theorem 6 than it otherwise would be under the corresponding results in [4].

6. Extension to unknown rewards and further work

In this section we briefly describe a range of potential extensions to our work that we didn't have time to fully consider. The first, and perhaps most insightful extension, would be to consider specific classes of $\mathcal{P}$ and compare how their complexity scales with the policy classes presented in [4]. As mentioned before, we conjecture that this this could potentially lead to less trajectory samples, n, being needed compared to the Imitation Learning setting; hence, a concrete characterisation of which $\mathcal{P}$-classes benefit under our results would provide significant progress towards a more complete understanding of when model-based methods are appropriate.

The second extension would be to consider how an unknown reward function would impact our current results. However, it is important to note that one of the main advantages of offline Imitation Learning is that it does not require access to the rewards generated by rolling out π^*; in the unknown-reward setting, such a dependence would ultimately be unavoidable for our algorithm since $\hat{\pi}$ is fit with respect to a learnt model of the environment. To enable a more suitable comparison in the unknown-reward setting, it would make more sense to compare Algorithm 1 to hybrid (both offline and online) approaches to Imitation Learning [4].

Appendix A. Results for the Hellinger distance

In this section, we provide the proofs omitted in the main text, categorised by their respective chapters. We will restate each Theorem for the sake of clarity.

Lemma 1. *(Specific case of Data Processing inequality) For any two distributions P_X, $P_{X'}$ defined on a measurable space, $(\mathcal{X}, \mathcal{E})$, we have that*

$$D_H^2\left(Bern(P_X(E)), Bern(P_{X'}(E))\right) \leq D_H^2(P_X, P_{X'})$$

for all $E \subset \mathcal{E}$.

Proof of lemma 1. Let P_X and $P_{X'}$ be distributions over some measurable space $(\mathcal{X}, \mathcal{E})$, and define $X : \Omega \to \mathcal{E}$ and $X' : \Omega \to \mathcal{E}$ such that P_X is the distribution of X over $\mathcal{E}$ and $P_{X'}$ is the distribution of X'. Next, define $Y = g(X)$ and $Y' = g(X')$ for some measurable $g : (\mathcal{X}, \mathcal{E}) \to (\mathbb{R}, \mathcal{B})$, $Y' := g(X')$, where P_Y and $P_{Y'}$ denote the respective distributions. To finish our setup, we define the transition kernel, $K : \mathcal{E} \times \mathcal{B} \to [0, 1]$, as

$$K(\cdot \mid x) := \mathbb{1}(g(x) \in \cdot)$$

and set $g(x) := \mathbb{1}(x \in E)$ for an arbitrary $E \in \mathcal{E}$. Then,

$$\begin{aligned} P_Y(A) &= \int_{\mathcal{E}} K(A \mid x) dP_X(x) \\ &= \int_{\mathcal{E}} \mathbb{1}(g(x) \in A) dP_X(x) \\ &= P_X(g^{-1}(A)) \end{aligned}$$

for all $A \subset \mathcal{B}$. Hence $P_Y(1) = P_X(g^{-1}(1)) = P_X(E)$ and $P_Y(0) = P_X(g^{-1}(0)) = P(E^c)$; this implies that $Y \sim Bern(P_X(E))$, and similarly $Y' \sim Bern(P_{X'}(E))$. Following immediately from an application of the Data processing inequality [16], we therefore have that,

$$\begin{aligned} & D_H^2(P_Y, P_{Y'}) \leq D_H^2(P_X, P_{X'}) \\ \implies & D_H^2\left(Bern(P_X(E)), Bern(P_{X'}(E))\right) \leq D_H^2(P_X, P_{X'}) \end{aligned}$$

□

Theorem 2 (Change of measure result)**.** *Let $(\mathcal{X}, \mathcal{E})$ be a measurable space, $\mathbb{P}$ and $\mathbb{Q}$ be two probability distributions defined over $(\mathcal{X}, \mathcal{E})$, and $h : \mathcal{X} \to \mathbb{R}$ be a measurable function uniformly bounded in absolute value by R. Then,*

$$|\mathbb{E}_{\mathbb{P}}[h(X)] - \mathbb{E}_{\mathbb{Q}}[h(X)]| \leq \sqrt{2R \cdot D_H^2(\mathbb{P}||\mathbb{Q}) \left(\mathbb{E}_{\mathbb{P}}|h(X)| + \mathbb{E}_{\mathbb{Q}}|h(X)|\right)}$$

Proof of theorem 2. This result is very similar to the one in [9], although we trivially extend it to non-negative $h(\cdot)$'s. Fix a measurable space $(\mathcal{X}, \mathcal{E})$ and let $\mathbb{P}$ and $\mathbb{Q}$ be two probability measures over $(\mathcal{X}, \mathcal{E})$. We first note that for all $p, q \geq 0$, we have the result through simple algebra

$$\frac{(p-q)^2}{2(p+q)} \leq (\sqrt{p} - \sqrt{q})^2$$

since $(p-q) = (\sqrt{p} - \sqrt{q})(\sqrt{p} + \sqrt{q})$. We now fix a set $A \subset \mathcal{E}$ and set $p = \mathbb{P}(A)$ and $q = \mathbb{Q}(A)$; by Lemma 1, we then have

$$\left(\sqrt{\mathbb{P}(A)} - \sqrt{\mathbb{Q}(A)}\right)^2 \leq D_H^2\left(Bern(\mathbb{P}(A)), Bern(\mathbb{Q}(A))\right) \leq D_H^2(P_X, P_{X'})$$

and hence for all $A \subset \mathcal{E}$,

$$|\mathbb{P}(A) - \mathbb{Q}(A)| \leq \sqrt{2[\mathbb{P}(A) + \mathbb{Q}(A)] \cdot D_H^2(\mathbb{P}||\mathbb{Q})}$$

Next, we have

$$\begin{aligned}
|\mathbb{E}_{\mathbb{P}}\left[h(X)\right] - \mathbb{E}_{\mathbb{Q}}[h(X)]| &= \left|\mathbb{E}_{\mathbb{P}}\left[h(X)^+\right] - \mathbb{E}_{\mathbb{Q}}[h(X)^+] + \mathbb{E}_{\mathbb{Q}}\left[h(X)^-\right] - \mathbb{E}_{\mathbb{P}}[h(X)^-]\right| \\
&\leq \left|\mathbb{E}_{\mathbb{P}}\left[h(X)^+\right] - \mathbb{E}_{\mathbb{Q}}[h(X)^+]\right| + \left|\mathbb{E}_{\mathbb{P}}\left[h(X)^-\right] - \mathbb{E}_{\mathbb{Q}}[h(X)^-]\right|
\end{aligned}$$

For simplicity, we will focus on bounding $|\mathbb{E}_{\mathbb{P}}\left[h(X)^+\right] - \mathbb{E}_{\mathbb{Q}}[h(X)^+]|$, since the bound on the negative component of $h(X)$ follows similarly. As $h(X)^+$ is non-negative, we have that

$$\begin{aligned}
\left|\mathbb{E}_{\mathbb{P}}\left[h(X)^+\right] - \mathbb{E}_{\mathbb{Q}}[h(X)^+]\right| &= \left|\int_0^H \mathbb{P}(h(X)^+ \geq t) - \mathbb{Q}(h(X)^+ \geq t)dt\right| \\
&\leq \int_0^H |\mathbb{P}(h(X)^+ \geq t) - \mathbb{Q}(h(X)^+ \geq t)|dt
\end{aligned}$$

By setting $A(t) := \{x \in \mathcal{E} : h(x)^+ \geq t\}$, we get

$$\begin{aligned}
&\leq \int_0^H \sqrt{2[\mathbb{P}(A(t)) + \mathbb{Q}(A(t))] \cdot D_H^2(\mathbb{P}||\mathbb{Q})}dt \\
&\leq \sqrt{2H \cdot D_H^2(\mathbb{P}||\mathbb{Q})}\left(\int_0^H \mathbb{P}(A(t)) + \mathbb{Q}(A(t))dt\right)^{1/2} \\
&= \sqrt{2H\left(\mathbb{E}_{\mathbb{P}}\left[h(X)^+\right] + \mathbb{E}_{\mathbb{Q}}[h(X)^+]\right) \cdot D_H^2(\mathbb{P}||\mathbb{Q})}
\end{aligned}$$

where the penultimate line follows by the measure-theoretic version of Jensen's inequality with μ taken as the uniform distribution on $[0, H]$. By an equivalent argument for the negative component of $h(X)$, we therefore have

$$\begin{aligned}
|\mathbb{E}_{\mathbb{P}}\left[h(X)\right] - \mathbb{E}_{\mathbb{Q}}[h(X)]| &\leq \sqrt{2H \cdot D_H^2(\mathbb{P}||\mathbb{Q}) \sum_{*\in\{+,-\}} \left(\mathbb{E}_{\mathbb{P}}\left[h(X)^*\right] + \mathbb{E}_{\mathbb{Q}}[h(X)^*]\right)} \\
&= \sqrt{2H \cdot D_H^2(\mathbb{P}||\mathbb{Q})\left(\mathbb{E}_{\mathbb{P}}|h(X)| + \mathbb{E}_{\mathbb{Q}}|h(X)|\right)}
\end{aligned}$$

which completes the argument □

Lemma 2. *Define* $P^h := \{P_t^*\}_{t=1}^h \cup \{\hat{P}_t\}_{t=h+1}^H$ *for any* $h \in [H]$*, then*

$$D_H^2(\mathbb{P}^{P^*,\pi^*}||\mathbb{P}^{P^h,\pi^*}) \leq 4D_H^2(\mathbb{P}^{P^*,\pi^*}||\mathbb{P}^{\hat{P},\pi^*})$$

Proof of lemma 2. First, fix an arbitrary $h \in [H]$ and define P^h as above. To complete our new set of notation, we will use $\mathbb{P}_{0:h}^{P,\pi}(s_{1:h}, a_{1:h}) := P_0(s_1)\prod_{t=1}^h P_t(s_{t+1}|s_t, a_t)\pi_t(a_t|s_t)$ to denote the trajectory distribution up to time h. Then, by the triangle inequality for Hellinger distances[8], we have

$$D_H(\mathbb{P}^{P^*,\pi^*}||\mathbb{P}^{P^h,\pi^*}) \leq D_H(\mathbb{P}^{P^*,\pi^*}||\mathbb{P}^{\hat{P},\pi^*}) + D_H(\mathbb{P}^{\hat{P},\pi^*}||\mathbb{P}^{P^h,\pi^*})$$

[8]For those wondering why we don't work with the squared Hellinger directly, we note that this property does not hold in general for the squared Hellinger

Since $\mathbb{P}^{P,\pi^*}_{h+1:H} \overset{d}{=} \mathbb{P}^{\hat{P},\pi^*}_{h+1:H}$ conditional on $(s_t, a_t)_{t=1}^h$, using Proposition 7.1 in [17] and the fact that the squared Hellinger distance is an f-divergence, we have that[9]

$$D_H^2(\mathbb{P}^{\hat{P},\pi^*}||\mathbb{P}^{P^h,\pi^*}) = D_H^2(\mathbb{P}^{\hat{P},\pi^*}_{0:h}||\mathbb{P}^{P^h,\pi^*}_{0:h}) = D_H^2(\mathbb{P}^{\hat{P},\pi^*}_{0:h}||\mathbb{P}^{P^*,\pi^*}_{0:h})$$

To upper bound the term on the right, we use the monotonicity property of f-divergences from Theorem 2.2 in [16]; with this, we have

$$D_H^2(\mathbb{P}^{\hat{P},\pi^*}_{0:h}||\mathbb{P}^{P^*,\pi^*}_{0:h}) \leq D_H^2(\mathbb{P}^{\hat{P},\pi^*}||\mathbb{P}^{P^*,\pi^*})$$

Hence, to finish our proof we note that $(x+y)^2 \leq 2(x^2+y^2)$ for all $x, y \in \mathbb{R}$, and therefore

$$\begin{aligned} D_H^2(\mathbb{P}^{P^*,\pi^*}||\mathbb{P}^{P^h,\pi^*}) &\leq 2\left(D_H^2(\mathbb{P}^{P^*,\pi^*}||\mathbb{P}^{\hat{P},\pi^*}) + D_H^2(\mathbb{P}^{\hat{P},\pi^*}||\mathbb{P}^{P^h,\pi^*})\right) \\ &\leq 4D_H^2(\mathbb{P}^{P^*,\pi^*}||\mathbb{P}^{\hat{P},\pi^*}) \end{aligned}$$

which completes our result for the specified h. Since this choice was arbitrary, we note that our bound holds for all $h \in [H]$ and hence the result follows. □

Appendix B. Results for deterministic $\mathcal{P}$

Lemma 3. *Let $\hat{P}$ and P^* be deterministic transition functions satisfying*

$$s_{1:H} = \hat{P}(\cdot, a_{1:H}) = P^*(\cdot, a_{1:H})$$

for some $(s_{1:H}, a_{1:H}) \in (\mathcal{S} \times \mathcal{A})^H$ and assume there exists an optimal policy such that $a_{1:H} = \hat{\pi}(s_{1:H})$ for some $\hat{\pi}$ within the set of optimal policies under $\hat{P}$. Then, there exists a (deterministic) optimal policy under P^* such that

$$a_{1:H} = \hat{\pi}(s_{1:H}) = \pi^*(s_{1:H})$$

In other terms, $\pi_h^*(s_h) = \hat{\pi}_h(s_H)$ for all $s_h \in s_{1:H}$

Proof of lemma 3. By the Bellman equation,

$$V_1^{P^*,\pi^*}(s_1) = \sum_{a\in\mathcal{A}} \pi^*(a \mid s_1)\left[r(s_1,a) + \sum_{s\in\mathcal{S}} P^*(s \mid s_1, a) V_2^{P^*,\pi^*}(s_2)\right] \tag{1}$$

$$= r(s_1, \pi^*(s_1)) + V_2^{P^*,\pi^*}(s_2) \tag{2}$$

$$= \sum_{t=1}^{H} r(s_t, \pi^*(s_t)) \tag{3}$$

$$= V_1^{\hat{P},\pi^*}(s_1) \tag{4}$$

where we have used recursion in line (3), and (4) follows since $\hat{P}$ and P^* agree on the unique trajectory which contributes to the Value function. By definition of an optimal deterministic policy, we therefore have that π^* is optimal for $\hat{P}$ when starting in s_1, and so $\pi^*(s_i) = \hat{\pi}(s_i)$ for all $s_i \in s_{1:H}$ as desired. We note that determinism of the transitions is needed for the result to hold in the case when $\hat{P}$ and P^* only agree on a singular trajectory; without this, we would require the transition probabilities to agree for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ to argue in a similar manner.

□

Lemma 4. *(Probability bound for deterministic $\mathcal{P}$) Let $\hat{P}$ be the transition function fitted using Algorithm 1 under a deterministic class, $\mathcal{P}$. Then, there exist deterministic policies $\hat{\pi} \in opt(\hat{P})$ and $\pi^* \in opt(P^*)$ such that*

$$\mathbb{P}^{P^*,\pi^*}\left[a_{1:H} \neq \hat{\pi}(s_{1:H})\right] \leq \frac{6\log(2|\mathcal{P}|\delta^{-1})}{n}$$

with probability $\geq 1-\delta$

[9] We note that the last equality follows by definition of P_h

Proof of lemma 4. To start, we note that $\{a_{1:H} \neq \hat{\pi}(s_{1:H})\}$ is decomposable as the union of the two sets $\{s_{1:H} = \hat{P}(\cdot, a_{1:H}) = P^*(\cdot, a_{1:H}), a_{1:H} \neq \pi^*(s_{1:H})\}$ and $\{s_{1:H} \neq \hat{P}(\cdot, a_{1:H}) \cup s_{1:H} \neq P^*(\cdot, a_{1:H}), a_{1:H} \neq \hat{\pi}(s_{1:H})\}$, where we have used Lemma 3 to substitute $\hat{\pi}$ with π^*. We will deal with each term separately then employ a union bound: first, we have

$$\mathbb{P}^{P^*,\pi^*}\left[s_{1:H} = \hat{P}(\cdot, a_{1:H}) = P^*(\cdot, a_{1:H}), a_{1:H} \neq \pi^*(s_{1:H})\right] \leq \sum_{s_{1:H}, a_{1:H}} P^*(s_{1:H}, a_{1:H})\pi^*(a_{1:H}|s_{1:H})\mathbb{1}\left(a_{1:H} \neq \pi^*(s_{1:H})\right)$$

which is trivially 0. Now for the other term:

$$\begin{aligned}
\mathbb{P}^{P^*,\pi^*}\left[s_{1:H} \neq \hat{P}(\cdot, a_{1:H}) \cup s_{1:H} \neq P^*(\cdot, a_{1:H}), a_{1:H} \neq \hat{\pi}(s_{1:H})\right] &\leq \mathbb{P}^{P^*,\pi^*}\left[s_{1:H} \neq \hat{P}(\cdot, a_{1:H}) \cup s_{1:H} \neq P^*(\cdot, a_{1:H})\right] \\
&\leq \mathbb{P}^{P^*,\pi^*}\left[s_{1:H} \neq \hat{P}(\cdot, a_{1:H})\right] + \mathbb{P}^{P^*,\pi^*}\left[s_{1:H} \neq P^*(\cdot, a_{1:H})\right] \\
&= \mathbb{P}^{P^*,\pi^*}\left[s_{1:H} \neq \hat{P}(\cdot, a_{1:H})\right] + 0
\end{aligned}$$

Hence, we have $\mathbb{P}^{P^*,\pi^*}\left[a_{1:H} \neq \hat{\pi}(s_{1:H})\right] \leq \mathbb{P}^{P^*,\pi^*}\left[s_{1:H} \neq \hat{P}(\cdot, a_{1:H})\right]$ and so it suffices to work with this upper bound. To finish the proof, it remains to show that this quantity is bounded above by $D_H^2(\mathbb{P}^{P^*,\pi^*}||\mathbb{P}^{\hat{P},\pi^*})$; to do this, we observe that

$$\begin{aligned}
D_H^2(\mathbb{P}^{\hat{P},\pi^*}||\mathbb{P}^{P^*,\pi^*}) &:= \sum_{s_{1:H}, a_{1:H}} \left(\sqrt{\pi^*(a_{1:H}|s_{1:H})\hat{P}(s_{1:H}, a_{1:H})} - \sqrt{\pi^*(a_{1:H}|s_{1:H})P^*(s_{1:H}, a_{1:H})}\right)^2 \\
&= \sum_{s_{1:H}, a_{1:H}} \pi^*(a_{1:H}|s_{1:H}) \left(\sqrt{\hat{P}(s_{1:H}, a_{1:H})} - \sqrt{P^*(s_{1:H}, a_{1:H})}\right)^2 \\
&\geq \sum_{s_{1:H}, a_{1:H}} \pi^*(a_{1:H}|s_{1:H}) \left(\sqrt{0} - \sqrt{P^*(s_{1:H}, a_{1:H})}\right)^2 \mathbb{1}(s_{1:H} \neq \hat{P}(\cdot, a_{1:H})) \\
&= \mathbb{P}^{P^*,\pi^*}\left[s_{1:H} \neq \hat{P}(\cdot, a_{1:H})\right]
\end{aligned}$$

With this, we employ Theorem 1 and arrive at the final result of:

$$\begin{aligned}
\mathbb{P}^{P^*,\pi^*}\left[a_{1:H} \neq \hat{\pi}(s_{1:H})\right] &\leq D_H^2(\mathbb{P}^{\hat{P},\pi^*}||\mathbb{P}^{P^*,\pi^*}) \\
&\leq \frac{6\log(2|\mathcal{P}|\delta^{-1})}{n}
\end{aligned}$$

with probability $\geq 1 - \delta$

We note that although the lower bound on $D_H^2(\mathbb{P}^{\hat{P},\pi^*}||\mathbb{P}^{P^*,\pi^*})$ holds independent of a deterministic $\mathcal{P}$, it becomes infeasible to work with such a quantity in the stochastic case; hence, our method of proof will need to be adapted in the sequel. □

Appendix C. Results for general $\mathcal{P}$

Theorem 5. *(Probability bound for stochastic $\mathcal{P}$) Consider the setting where we have the dataset $\{(s_{1:H}, a_{1:H})^i\}_{i=1}^n$ with trajectories drawn iid from $(s_{1:H}, a_{1:H})^i \overset{iid}{\sim} \mathbb{P}^{P^*,\pi^*}$ and fit $\hat{P}$ as in Algorithm 1, where $\mathcal{P}$ is now a class of (potentially stochastic) transition functions. Then, under all the usual assumptions,*

$$\begin{aligned}
\mathbb{P}^{P^*,\pi^*}\left[a_{1:H} \neq \hat{\pi}(s_{1:H})\right] &\leq \sum_{h=1}^{H} \frac{12(2+\sqrt{\mu})\left[R \wedge (H-h)\right]}{\Omega_{min}(h)} \sqrt{D_H^2(\mathbb{P}^{P^*,\pi^*}||\mathbb{P}^{\hat{P},\pi^*})} \\
&\lesssim HR\sqrt{D_H^2(\mathbb{P}^{P^*,\pi^*}||\mathbb{P}^{\hat{P},\pi^*})}
\end{aligned}$$

where μ is the recoverability constant defined above, π^* and $\hat{\pi}$ are the deterministic[10] optimal policies under P^* and $\hat{P}$, and $\Omega_{\min}(h)$ is a (potentially very small) scaling factor.

[10]This is crucial for the argument we present, although it may be possible to relax this assumption with further work

Proof of theorem 5. We define the action gap for a given starting state and time-step as

$$\Omega(s,h) := Q_h^*(s,\pi^*(s)) - \operatorname*{argmax}_{a\neq\pi^*(s)} Q_h^{P^*,\pi^*}(s,a) < R$$

and note that, by the definition of an optimal policy, if $Q_h^*(s,\pi^*(s)) - Q_h^*(s,\hat{\pi}(s)) \leq \Omega(s,h)$, then $\hat{\pi} := \hat{\pi}_h \cup \{\pi_t^*\}_{t=h+1}^H$ is optimal under P^* and hence $\hat{\pi}_h(s) = \pi_h^*(s)$. We note that we have assumed uniqueness of π^* here to simplify the argument, although our results can be extended (as they were in the deterministic case) by showing that there exists a π^* such that the above implication holds. Additionally, we recall that $d_h^{P^*,\pi^*}$ is defined as the state-occupancy measure under roll-outs induced by P^* and π^* up to time-step h; for this section only, we will abbreviate this distribution as d_h^*.

To start our proof: since $Q_h^*(s,\pi^*(s)) - Q_h^*(s,\hat{\pi}(s)) \leq \Omega(s,h) \implies \hat{\pi}_h(s) = \pi_h^*(s)$, we therefore have

$$\mathbb{P}_{s\sim d_h^*}\left[Q_h^*(s,\pi^*(s)) - Q_h^*(s,\hat{\pi}(s)) \leq \Omega(s,h)\right] \leq \mathbb{P}_{s\sim d_h^*}\left[\hat{\pi}_h(s) = \pi_h^*(s)\right]$$

and hence

$$\mathbb{P}_{s\sim d_h^*}\left[\hat{\pi}_h(s) \neq \pi_h^*(s)\right] \leq \mathbb{P}_{s\sim d_h^*}\left[Q_h^*(s,\pi^*(s)) - Q_h^*(s,\hat{\pi}(s)) > \Omega(s,h)\right]$$

To proceed, we now decompose the difference in Q-functions as

$$Q_h^*(s,\pi^*(s)) - Q_h^*(s,\hat{\pi}(s)) = \underbrace{Q_h^*(s,\pi^*(s)) - Q_h^{\hat{P},\pi^*}(s,\pi^*(s)) + Q_h^{\hat{P},\pi^*}(s,\pi^*(s)) - Q_h^*(s,\hat{\pi}(s))}_{\xi(h,s)}$$

which implies that

$$\mathbb{P}_{s\sim d_h^*}\left[Q^*(s,\pi^*(s)) - Q^*(s,\hat{\pi}(s)) \leq \Omega(h,s)\right] \geq \mathbb{P}_{s\sim d_h^*}\left[\xi(h,s) \leq \Omega(h,s)\right]$$

and therefore, by taking complements,

$$\mathbb{P}_{s\sim d_h^*}\left[Q_h^*(s,\pi^*(s)) - Q_h^*(s,\hat{\pi}(s)) > \Omega(h,s)\right] \leq \mathbb{P}_{s\sim d_h^*}\left[\xi(h,s) > \Omega(h,s)\right]$$

Now, consider the set $\{\xi(h,s) > \Omega(h,s)\}$; by a standard argument, we note that $\xi(h,s) > \Omega(h,s)$ only if each component of $\xi(h,s)$ is larger than $\Omega(h,s)/2$. Hence,

$$\mathbb{P}_{s\sim d_h^*}\left[\xi(h,s) > \Omega(h,s)\right] \leq$$

$$\mathbb{P}_{s\sim d_h^*}\left[Q_h^*(s,\pi^*(s)) - Q_h^{\hat{P},\pi^*}(s,\pi^*(s)) > \frac{\Omega(h,s)}{2} \cup Q_h^{\hat{P},\pi^*}(s,\hat{\pi}(s)) - Q_h^*(s,\hat{\pi}(s)) > \frac{\Omega(h,s)}{2}\right]$$

and therefore, through the use of a union bound, we can work with each event separately. First, we consider $\{Q_h^*(s,\pi^*(s)) - Q_h^{\hat{P},\pi^*}(s,\pi^*(s)) > \frac{\Omega(h,s)}{2}\} \equiv \{V_h^*(s) - V_h^{\hat{P},\pi^*}(s) > \frac{\Omega(h,s)}{2}\}$; to remove $\Omega(h,s)$'s dependence on the state, we define $\Omega_{min}(h) := \min_s \Omega(h,s)$ and note that $\{V_h^*(s) - V_h^{\hat{P},\pi^*}(s) > \frac{\Omega(h,s)}{2}\} \subset \{V_h^*(s) - V_h^{\hat{P},\pi^*}(s) > \frac{\Omega_{\min}(h)}{2}\}$. Therefore, by Markov's inequality, we have

$$\mathbb{P}_{s\sim d_h^*}\left[V_h^*(s) - V_h^{\hat{P},\pi^*}(s) > \frac{\Omega(h,s)}{2}\right] \leq \frac{2}{\Omega_{\min}(h)}\mathbb{E}_{s\sim d_h^*}\left[V_h^*(s) - V_h^{\hat{P},\pi^*}(s)\right]$$

Using Lemma 2 and Theorem 2, we then obtain

$$\begin{aligned}
\mathbb{E}_{s\sim d_h^*}\left[V_h^*(s) - V_h^{\hat{P},\pi^*}(s)\right] &\leq |\mathbb{E}_{s\sim d_h^*} V_h^*(s) - \mathbb{E}_{s\sim d_h^*} V_h^{\hat{P},\pi^*}(s)| \\
&\leq 2[R \wedge (H-h)]\sqrt{4D_H^2(\mathbb{P}^{P^*,\pi^*}||\mathbb{P}^{\hat{P},\pi^*})} \\
&= 4[R \wedge (H-h)]\sqrt{D_H^2(\mathbb{P}^{P^*,\pi^*}||\mathbb{P}^{\hat{P},\pi^*})}
\end{aligned}$$

and hence

$$\mathbb{P}_{s\sim d_h^*}\left[V_h^*(s) - V_h^{\hat{P},\pi^*}(s) > \frac{\Omega(h,s)}{2}\right] \leq \frac{8[R \wedge (H-h)]}{\Omega_{\min}(h)}\sqrt{D_H^2(\mathbb{P}^{P^*,\pi^*}||\mathbb{P}^{\hat{P},\pi^*})}$$

for all $h \in [H]$. To deal with the remaining term, we first note that

$$Q_h^{\hat{P},\pi^*}(s,\pi^*(s)) - Q_h^*(s,\hat{\pi}(s)) = \sum_{s'\in\mathcal{S}} \hat{P}_h(s' \mid s,\pi^*(s))V_{h+1}^{\hat{P},\pi^*}(s') - P_h^*(s' \mid s,\hat{\pi}(s))V_{h+1}^*(s')$$

since we have assumed that rewards are known. Then, by Assumption 1 and Corollary 1, we have that

$$\begin{aligned}
\mathbb{E}_{s\sim d_h^*}\left[Q_h^{\hat{P},\pi^*}(s,\pi^*(s)) - Q_h^*(s,\hat{\pi}(s))\right] &= \mathbb{E}_{s\sim d_h^*}\left[\mathbb{E}_{s'\sim\hat{P}(\cdot|s,\pi^*(s))}V_{h+1}^{\hat{P},\pi^*}(s')\right] - \mathbb{E}_{s\sim d_h^*}\left[\mathbb{E}_{s'\sim P^*(\cdot|s,\hat{\pi}(s))}V_{h+1}^{P^*,\pi^*}(s')\right] \\
&\leq \left|\mathbb{E}_{s\sim d_h^*}\left[\mathbb{E}_{s'\sim\hat{P}(\cdot|s,\pi^*(s))}V_{h+1}^{\hat{P},\pi^*}(s')\right] - \mathbb{E}_{s\sim d_h^*}\left[\mathbb{E}_{s'\sim P^*(\cdot|s,\hat{\pi}(s))}V_{h+1}^{P^*,\pi^*}(s')\right]\right| \\
&= \left|\mathbb{E}^{P^h,\pi^*}r(s_{1:H},a_{1:H}) - \mathbb{E}^{P^*,\pi^h}r(s_{1:H},a_{1:H})\right| \\
&\leq 2[R\wedge(H-h)]\sqrt{D_H^2(\mathbb{P}^{P^h,\pi^*}||\mathbb{P}^{P^*,\pi^h})} \\
&\leq 2[R\wedge(H-h)]\sqrt{(2+\sqrt{\mu})^2\cdot D_H^2(\mathbb{P}^{P^*,\pi^*}||\mathbb{P}^{\hat{P},\pi^*})} \\
&= 2(2+\sqrt{\mu})[R\wedge(H-h)]\sqrt{D_H^2(\mathbb{P}^{\hat{P},\pi^*}||\mathbb{P}^{P^*,\pi^*})}
\end{aligned}$$

and therefore

$$\begin{aligned}
\mathbb{P}_{s\sim d_h^*}\left[Q_h^{\hat{P},\pi^*}(s,\pi^*(s)) - Q_h^*(s,\hat{\pi}(s)) > \frac{\Omega(h,s)}{2}\right] &\leq \frac{2}{\Omega_{\min}(s)}\mathbb{E}_{s\sim d_h^*}\left[Q_h^{\hat{P},\pi^*}(s,\pi^*(s)) - Q_h^*(s,\hat{\pi}(s))\right] \\
&\leq \frac{4(2+\sqrt{\mu})[R\wedge(H-h)]}{\Omega_{\min}(s)}\sqrt{D_H^2(\mathbb{P}^{\hat{P},\pi^*}||\mathbb{P}^{P^*,\pi^*})}
\end{aligned}$$

By combining the above results with a union bound, we obtain

$$\begin{aligned}
\mathbb{P}_{s\sim d_h^*}\left[\hat{\pi}_h(s)\neq\pi^*(s)\right] &\leq \frac{8[R\wedge(H-h)]}{\Omega_{\min}(h)}\sqrt{D_H^2(\mathbb{P}^{P^*,\pi^*}||\mathbb{P}^{\hat{P},\pi^*})} + \frac{4(2+\sqrt{\mu})[R\wedge(H-h)]}{\Omega_{\min}(h)}\sqrt{D_H^2(\mathbb{P}^{\hat{P},\pi^*}||\mathbb{P}^{P^*,\pi^*})} \\
&\leq \frac{12(2+\sqrt{\mu})[R\wedge(H-h)]}{\Omega_{\min}(h)}\sqrt{D_H^2(\mathbb{P}^{\hat{P},\pi^*}||\mathbb{P}^{P^*,\pi^*})}
\end{aligned}$$

for all $h\in[H]$. Using the union bound once again, we have

$$\begin{aligned}
\mathbb{P}^{P^*,\pi^*}[a_{1:H}\neq\hat{\pi}(s_{1:H})] &= \mathbb{P}^{P^*,\pi^*}[\exists h : a_h\neq\hat{\pi}_h(s_h)] \\
&= \mathbb{P}^{P^*,\pi^*}\left[\bigcup_{h=1}^H a_h\neq\hat{\pi}_h(s_h)\right] \\
&\leq \sum_{h=1}^H \mathbb{P}_{s_h\sim d_h^*}[\pi^*(s_h)\neq\hat{\pi}_h(s_h)] \\
&\leq \sum_{h=1}^H \frac{12(2+\sqrt{\mu})[R\wedge(H-h)]}{\Omega_{\min}(h)}\sqrt{D_H^2(\mathbb{P}^{\hat{P},\pi^*}||\mathbb{P}^{P^*,\pi^*})}
\end{aligned}$$

which completes the result. □

References

[1] Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based RL in Contextual Decision Processes: PAC bounds and Exponential Improvements over Model-free Approaches, May 2019. arXiv:1811.08540 [cs].

[2] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual Decision Processes with Low Bellman Rank are PAC-Learnable, December 2016. arXiv:1610.09512 [cs].

[3] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm, December 2017. arXiv:1712.01815 [cs].

[4] Dylan J Foster, Adam Block, and Dipendra Misra. Is Behavior Cloning All You Need? Understanding Horizon in Imitation Learning. *Neurips*, 2024.

[5] Stephane Ross and Drew Bagnell. 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, March 2010.

[6] Nived Rajaraman, Yanjun Han, Lin F. Yang, Kannan Ramchandran, and Jiantao Jiao. Provably Breaking the Quadratic Error Compounding Barrier in Imitation Learning, Optimally, February 2021. arXiv:2102.12948 [cs].

[7] Andrei A. Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy Distillation, January 2016. arXiv:1511.06295 [cs].

[8] Sergey Levine and Vladlen Koltun. Guided Policy Search. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1–9. PMLR, May 2013.

[9] Dylan J. Foster, Alexander Rakhlin, Ayush Sekhari, and Karthik Sridharan. On the Complexity of Adversarial Decision Making, June 2022. arXiv:2206.13063 [cs].
[10] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion, March 2024. arXiv:2303.04137 [cs].
[11] Adam Block, Ali Jadbabaie, Daniel Pfrommer, Max Simchowitz, and Russ Tedrake. Provable Guarantees for Generative Behavior Cloning: Bridging Low-Level Stability and High-Level Behavior, October 2023. arXiv:2307.14619 [cs].
[12] Aditya Modi, Jinglin Chen, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. Model-free Representation Learning and Exploration in Low-rank MDPs, June 2022. arXiv:2102.07035 [cs].
[13] Jingfeng Wu, Vladimir Braverman, and Lin F Yang. Gap-Dependent Unsupervised Exploration for Reinforcement Learning.
[14] Amir-massoud Farahmand. Action-Gap Phenomenon in Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.
[15] Nived Rajaraman, Lin F Yang, Jiantao Jiao, and Kannan Ramachandran. Toward the Fundamental Limits of Imitation Learning. *Neurips*, 2020.
[16] Yury Polyanskiy. Mit lecture notes on information theory. *MIT*, 2020.
[17] Yury Polyanskiy. Mit lecture notes on f-divergences. *MIT*, 2020.