On Overfitting and Asymptotic Bias in Batch Reinforcement Learning with Partial Observability (Extended Abstract)*

Vincent François-Lavet [1,3] , **Guillaume Rabusseau** [2,3] , **Joelle Pineau** [1,3] ,
Damien Ernst [4] and **Raphael Fonteneau** [4]
[1] McGill University
[2] Université de Montréal
[3] Mila
[4] Université de Liège

Abstract

When an agent has limited information on its environment, the suboptimality of an RL algorithm can be decomposed into the sum of two terms: a term related to an asymptotic bias (suboptimality with unlimited data) and a term due to overfitting (additional suboptimality due to limited data). In the context of reinforcement learning with partial observability, this paper provides an analysis of the tradeoff between these two sources of error. In particular, our theoretical analysis formally characterizes how a smaller state representation increases the asymptotic bias while decreasing the risk of overfitting.

1 Formalization

When acquisition of new observations is possible (the "online" case), data scarcity is gradually phased out using strategies balancing the exploration / exploitation (E/E) tradeoff. The scientific literature related to this topic is vast; in particular, Bayesian RL techniques [Ross *et al.*, 2011; Ghavamzadeh *et al.*, 2015] offer an elegant way of formalizing the E/E tradeoff.

However, such E/E strategies are not applicable when the acquisition of new observations is not possible anymore. In the pure "batch" setting (also called the "offline" setting), the task is to learn the best possible policy from a fixed set of transition samples [Farahmand, 2011; Lange *et al.*, 2012].

We consider a discrete-time POMDP [Sondik, 1978] model M described by the tuple $(\mathcal{S}, \mathcal{A}, T, R, \Omega, O, \gamma)$ where

- $\mathcal{S}$ is a finite set of states $\{1, \dots, N_{\mathcal{S}}\}$,
- $\mathcal{A}$ is a finite set of actions $\{1, \dots, N_{\mathcal{A}}\}$,
- $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition function,
- $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathcal{R}$ is the reward function, where $\mathcal{R}$ is a continuous set of possible rewards in a range $R_{max} \in \mathbb{R}^+$,
- Ω is a finite set of observations $\{1, \dots, N_{\Omega}\}$,
- $O : \mathcal{S} \times \Omega \to [0, 1]$ is a set of conditional observation probabilities, and
- $\gamma \in [0, 1)$ is the discount factor.

The initial state is drawn from an initial distribution $b(s_0)$. In this paper, the conditional transition probabilities T, the reward function R and the conditional observation probabilities O are unknown. The only information available to the agent is the past experience it gathered while interacting with the POMDP.

1.1 Processing a History of Data

A history of previously observed features can be used to estimate the hidden state dynamics [McCallum, 1996; Littman and Sutton, 2002; Singh *et al.*, 2004]. We denote by $\mathcal{H}_t = \Omega \times (\mathcal{A} \times \mathcal{R} \times \Omega)^t$ the set of histories observed up to time t for $t \in \mathbb{N}_0$, and by $\mathcal{H} = \bigcup_{t=0}^{\infty} \mathcal{H}_t$ the space of all possible observable histories.

In this paper, we consider a mapping $\phi : \mathcal{H} \to \phi(\mathcal{H})$, where $\phi(\mathcal{H}) = \{\phi(H) | H \in \mathcal{H}\}$ is of finite cardinality $|\phi(\mathcal{H})|$. On the one hand, we will show that when ϕ discards information from the whole history, the state representation $\phi(H)$ that the agent uses to make decision might depart from sufficient statistics, which can hurt performance. On the other hand, we will show that it is beneficial to use a mapping ϕ that has a low cardinality $|\phi(\mathcal{H})|$ to avoid overfitting.

Let us first introduce a notion of information on the latent hidden state s through the notion of belief state [Cassandra *et al.*, 1994].

Definition 1.1. *The belief state $b(s|H)$ is defined as the vector of probabilities where the i^{th} component ($i \in \{1, \dots, N_{\mathcal{S}}\}$) is given by $\mathbb{P}(s = i \mid H)$, for any history $H \in \mathcal{H}$.*

Definition 1.2. *The belief state $b_{\phi}(s|\phi(H))$ is defined as the vector of probabilities where the i^{th} component ($i \in \{1, \dots, N_{\mathcal{S}}\}$) is given by $\mathbb{P}(s = i \mid \phi(H))$, for any history $H \in \mathcal{H}$* [1].

* This paper is an extended abstract of an article in the Journal of AI Research [François-Lavet *et al.*, 2019].

[1] Note that s and H are random variables and their exact distribution will depend on the context that is considered. For any given probability distribution $\mathcal{D}_H$ over histories: $H \sim \mathcal{D}_H$, the probability $P(s|\phi(H))$ is the expectation of the state when $\phi(H)$ is observed: $b_{\phi}(s \mid \varphi) = \underset{H \sim \mathcal{D}_H, \varphi = \phi(H)}{\mathbb{E}} b(s \mid H)$.

Among all possible mappings ϕ, the notion of sufficient statistics [Kaelbling *et al.*, 1998; Aberdeen *et al.*, 2007] corresponds to the ones that extract enough information from the history to accurately capture the corresponding belief state.

Definition 1.3. *In a POMDP M, a statistic $\phi(H)$ is a sufficient statistic at the condition that $\forall s \in \mathcal{S}$:*

$$\mathbb{P}(s \mid H) = \mathbb{P}(s \mid \phi(H)),$$

for $H \in \mathcal{H}$. A mapping ϕ which provides sufficient statistics for all histories $H \in \mathcal{H}$ is called a sufficient mapping and is denoted as ϕ_0.

One of the key notions on which our analysis relies is the one of "approximately sufficient mappings", i.e. mappings whose corresponding belief state lies in an L_1-ball of radius ϵ centered on $b(\cdot|H)$:

Definition 1.4. *In a POMDP M, a statistic $\phi(H)$ is an ϵ-sufficient statistic at the condition that it meets the following condition with $\epsilon \geq 0$ and with the L_1 norm:*

$$\|b_\phi(\cdot|\phi(H)) - b(\cdot|H)\|_1 \leq \epsilon,$$

for $H \in \mathcal{H}$. A mapping ϕ that provides ϵ-sufficient statistics for all histories $H \in \mathcal{H}$ is called an ϵ-sufficient mapping and is denoted as $\phi_\epsilon(H)$.

1.2 Working with a Limited Dataset

Let $\mathcal{M}(\mathcal{S}, \mathcal{A}, \Omega, \gamma)$ be a set of POMDPs with fixed $\mathcal{S}$, $\mathcal{A}$, Ω, and γ. For any POMDP $M(T, R, O) \in \mathcal{M}$, we denote by D_{M,π_s,N_{tr},N_l} a random dataset generated according to a probability distribution $\mathcal{D}_{M,\pi_s,N_{tr},N_l}$ over the set of N_{tr} trajectories of length N_l. One such trajectory is defined as the observable history $H_{N_l} \in \mathcal{H}_{N_l}$ obtained in M when starting from s_0 and following a stochastic sampling policy π_s that ensures a non-zero probability of taking any action given an observable history $H \in \mathcal{H}$. For simplicity we denote D_{M,π_s,N_{tr},N_l}, simply as $D \sim \mathcal{D}_M$. For the purpose of the analysis, we also introduce the asymptotic dataset $D_\infty = D_{M,\pi_s,N_{tr}\to\infty,N_l\to\infty}$ that would be theoretically obtained in the case where one could generate an infinite number of observations ($N_{tr} \to \infty$ and $N_l \to \infty$).

1.3 Assessing the Performance of a Policy

Let us consider stationary and deterministic control policies $\pi : \phi(\mathcal{H}) \to \mathcal{A}$ with $\pi \in \Pi$. Any particular choice of ϕ induces a particular definition of the policy space Π. We introduce $V_M^\pi(\phi(H))$ with $H \in \mathcal{H}$ as the expected return obtained over an infinite time horizon when the system is controlled using policy π in the POMDP M. For any given distribution $\mathcal{D}_H$ over histories, this is defined as:

$$V_M^\pi(\phi(H)) = \underset{\substack{H' \sim \mathcal{D}_H: \\ \phi(H') = \phi(H)}}{\mathbb{E}} V_M^\pi(H' \mid \phi),$$

with $V_M^\pi(H \mid \phi)$ given by

$$V_M^\pi(H \mid \phi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 \sim b(\cdot|H), \pi\right],$$

where we have $\mathbb{P}\big(\omega_t \mid s_t\big) = O(s_t, \omega_t)$, $a_t = \pi(\phi(H_t))$, $\mathbb{P}\big(s_{t+1}|s_t, a_t\big) = T(s_t, a_t, s_{t+1})$ and $r_t = R\big(s_t, a_t, s_{t+1}\big)$.

We also define π^* as an optimal policy in M:

$$\pi^* \in \underset{\pi:\phi_0(\mathcal{H})\to\mathcal{A}}{\operatorname{argmax}} V_M^\pi(\phi_0(H_0)),$$

where H_0 is taken out of the distribution of initial observations (compatible with the distribution $b(s_0)$ of initial states through the conditional observation probabilities).

2 Bias-overfitting in RL with Partial Observability

In this section, we study the performance gap between the expected return that can be obtained following the policy built from limited data and the highest possible expected return that we would obtain if the algorithm had access to the POMDP parameters.

To study the importance of the feature space, let us assume that the policies built from limited data are optimal according to frequentist statistics, which allows removing from the analysis how the RL algorithm converges. In order to define the optimal policy according to frequentist statistics, let us first introduce a frequentist-based (augmented) MDP from the dataset D:

Definition 2.1. *With M defined by $(S, A, T, R, \Omega, O, \gamma)$ and the dataset D built from interactions with M, the frequentist-based augmented MDP $\hat{M}_{D,\phi}$, also denoted for simplicity $\hat{M}_D = (\Sigma, \mathrm{A}, \hat{T}, \hat{R}, \Gamma)$, is defined with*

- *the state space: $\Sigma = \phi(\mathcal{H})$,*
- *the action space: $\mathrm{A} = \mathcal{A}$,*
- *the estimated transition function: for $\sigma, \sigma' \in \Sigma$ and $a \in \mathrm{A}$, $\hat{T}(\sigma, a, \sigma')$ is the number of times we observe the transition $(\sigma, a) \to \sigma'$ divided by the number of times we observe (σ, a)* [2]
- *the estimated reward function: for $\sigma, \sigma' \in \Sigma$ and $a \in \mathrm{A}$, $\hat{R}(\sigma, a, \sigma')$ is the mean of the rewards observed for the tuple (σ, a, σ')* [3]
- *the discount factor $\Gamma \leq \gamma$.*

We introduce $\mathcal{V}_{\hat{M}_D}^\pi(\sigma)$ with $\sigma \in \Sigma$ as the expected return obtained over an infinite time horizon when the system is controlled using a policy $\pi : \Sigma \to \mathrm{A}$ in the augmented decision process $\hat{M}_D$:

$$\mathcal{V}_{\hat{M}_D}^\pi(\sigma) = \mathbb{E}\left[\sum_{t=0}^{\infty} \Gamma^k \hat{r}_t | \sigma_0 = \sigma, \pi\right],$$

where $\hat{r}_t$ is the reward s.t. $\hat{r}_t = \hat{R}(\sigma_t, a_t, \sigma_{t+1})$ and the transition is defined by $\mathbb{P}(\sigma_{t+1}|\sigma_t, a_t) = \hat{T}(\sigma_t, a_t, \sigma_{t+1})$.

A policy π is defined to be better than or equal to a policy π' if its expected return is greater than or equal to that of π' for all states. In an MDP, there is always at least one policy

[2] if any (σ, a) has never been encountered in a dataset, we arbitrarily set $\hat{T}(\sigma, a, \sigma') = 1/|\Sigma|, \forall\sigma'$.

[3] if any (σ, a, σ') has never been encountered in a dataset, we arbitrarily set $\hat{R}(\sigma, a, \sigma')$ to the average of rewards observed over the whole dataset D.

that is better than or equal to all other policies and this is an optimal policy [Sutton and Barto, 1998]. In the augmented MDP $\hat{M}_D$, we denote the optimal policy as $\pi_{D,\phi}$ and we also call it the frequentist-based policy. Let us now decompose the error of using a frequentist-based policy $\pi_{D,\phi}$ in the actual POMDP:

$$\begin{aligned}\underset{D\sim\mathcal{D}_M}{\mathbb{E}}\left[V_M^{\pi^*}(\phi_0(H)) - V_M^{\pi_{D,\phi}}(\phi(H))\right] = \\ \underbrace{\left(V_M^{\pi^*}(\phi_0(H)) - V_M^{\pi_{D_\infty,\phi}}(\phi(H))\right)}_{\substack{\text{bias function of dataset } D_\infty \text{ (function of } \pi_s) \\ \text{and frequentist-based policy } \pi_{D_\infty,\phi} \text{ (function of } \phi \text{ and } \Gamma)}} \\ + \underbrace{\underset{D\sim\mathcal{D}_M}{\mathbb{E}}\left[V_M^{\pi_{D_\infty,\phi}}(\phi(H)) - V_M^{\pi_{D,\phi}}(\phi(H))\right]}_{\substack{\text{overfitting due to finite dataset D (function of } \pi_s,\, N_l,\, N_{tr}) \\ \text{in the context of frequentist-based policy } \pi_{D,\phi} \\ \text{(function of } \phi \text{ and } \Gamma)}} .\end{aligned} \tag{1}$$

The term *bias* actually refers to an asymptotic bias when the size of the dataset tends to infinity, while the term *overfitting* refers to the expected suboptimality due to the finite size of the dataset (and thus due to the variance in the estimated transition function and reward function).

We start by providing a bound on the bias, which is an original result based on the belief states via the ϵ-sufficient statistic.

Theorem 1. ***"Bound on the bias":*** *Let M be a POMDP described by the 7-tuple $(\mathcal{S}, \mathcal{A}, T, R, \Omega, O, \gamma)$. Let $\hat{M}_{D_\infty}$ be an augmented MDP $(\Sigma, \mathrm{A}, \hat{T}, \hat{R}, \Gamma = \gamma)$ estimated, according to Definition 2.1, from a dataset D_∞. Then, for any ϵ-sufficient mapping $\phi = \phi_\epsilon$, the asymptotic bias can be bounded as follows:*

$$\max_{H\in\mathcal{H}}\left(V_M^{\pi^*}(\phi_0(H)) - V_M^{\pi_{D_\infty,\phi}}(\phi(H))\right) \leq \frac{2\epsilon R_{max}}{(1-\gamma)^3}. \tag{2}$$

Proof. We consider the frequentist-based MDP $\hat{M}_{D_\infty,\phi_0}(\Sigma_0, \mathrm{A}, \hat{T}, \hat{R}, \Gamma = \gamma)$, for $H \in \mathcal{H}$ and $a \in \mathcal{A}$, let us define

$$\begin{aligned}\mathcal{Q}_{\hat{M}_{D_\infty,\phi_0}}^{\pi_{D_\infty,\phi_0}}(\phi_0(H), a) = \hat{R}'(\phi_0(H), a) + \\ \gamma \sum_{\varphi\in\phi_0(\mathcal{H})} \hat{T}(\phi_0(H), a, \varphi)\mathcal{V}_{\hat{M}_{D_\infty,\phi_0}}^{\pi_{D_\infty,\phi_0}}(\varphi),\end{aligned}$$

where the reward

$$\hat{R}'(\phi_0(H), a) = \sum_{\varphi\in\phi_0(\mathcal{H})} \hat{T}(\phi_0(H), a, \varphi)\hat{R}(\phi_0(H), a, \varphi).$$

Then the main part of the proof is to demonstrate Proposition 2 below. From there, by applying Lemma 1 by Abel *et al.* 2016, we have:

$$\left\|\mathcal{V}_{M_{D_\infty,\phi_0}}^{\pi_{D_\infty,\phi_0}} - \mathcal{V}_{M_{D_\infty,\phi_0}}^{\pi_{D_\infty,\phi_\epsilon}}\right\|_\infty \leq \frac{2\frac{\epsilon R_{max}}{1-\gamma}}{(1-\gamma)^2} = \frac{2\epsilon R_{max}}{(1-\gamma)^3}.$$

By further noticing that, when starting in s_0, $\hat{M}_{D_\infty,\phi_0}$ and M provide an identical value function for a given policy $\pi_{D,\phi}$ and that $\pi_{D_\infty,\phi_0} \sim \pi^*$, i.e. $V_M^{\pi^*} = V_M^{\pi_{D_\infty,\phi_0}}$, the theorem follows. □

Remark. As compared to Hutter 2014 and Abel *et al.* 2016, this bound relates directly to the capacity of the mapping $\phi(H)$ to retrieve sufficient information on the latent hidden state. As compared to PBVI [Pineau *et al.*, 2003] and similar approaches, we do not make the assumption that T, R and O are known and, as such, they need to be estimated from data.

We now provide Proposition 2, which is the key result required in the proof of Theorem 1.

Proposition 2. *Let ϕ_ϵ be an ϵ-sufficient mapping, and let ϕ_0 be a sufficient mapping. Then, for any $H^{(1)}, H^{(2)}$ such that $\phi_\epsilon(H^{(1)}) = \phi_\epsilon(H^{(2)})$, we have*

$$\begin{aligned}\max_a \left|\mathcal{Q}_{\hat{M}_{D_\infty,\phi_0}}^{\pi_{D_\infty,\phi_0}}(\phi_0(H^{(1)}), a) - \mathcal{Q}_{\hat{M}_{D_\infty,\phi_0}}^{\pi_{D_\infty,\phi_0}}(\phi_0(H^{(2)}), a)\right| \\ \leq \epsilon\frac{R_{max}}{(1-\gamma)}.\end{aligned}$$

Proof. For this proposition, we rely on the fact that since $\phi_\epsilon(H^{(1)}) = \phi_\epsilon(H^{(2)})$, we are able to bound the L1 error terms of the associated belief states of $H^{(1)}$ and $H^{(2)}$. From that bound, we then present two different ways of independent interest to prove Proposition 2. The first proof makes use of a tree of possible future observations, rewards and corresponding actions given a policy π, and we show that when starting from $H^{(1)}, H^{(2)}$ such that $\phi_\epsilon(H^{(1)}) = \phi_\epsilon(H^{(2)})$, the bound holds. We also provide an alternative proof that makes use of the formalism of the bisimulation metric [Ferns *et al.*, 2004] along with the data processing inequality. The details of the proofs are given in the full paper. □

We now provide a bound on the overfitting error that monotonically grows with $|\phi(\mathcal{H})|$. Theorem 3 shows that using a large set of features potentially leads to a stronger drop in performance when the available dataset D is limited (the bound decreases proportionally to $\frac{1}{\sqrt{n}}$). A theoretical analysis in the context of MDPs with a finite dataset was performed by Jiang *et al.* 2015.

Theorem 3. ***"Bound on the overfitting":*** *Let M be a POMDP described by the 7-tuple $(\mathcal{S}, \mathcal{A}, T, R, \Omega, O, \gamma)$. Let $\hat{M}_D$ be an augmented MDP $(\Sigma, \mathrm{A}, \hat{T}, \hat{R}, \Gamma = \gamma)$ estimated, according to Definition 2.1, from a dataset D with the assumption that D has n transitions from any possible pair $(\phi(H), a) \in \phi(\mathcal{H}) \times \mathrm{A}$ (sampled i.i.d according to $\mathcal{D}_M$). Then the overfitting due to using the frequentist-based policy $\pi_{D,\phi}$ instead of $\pi_{D_\infty,\phi}$ in the POMDP M can be bounded as follows:*

$$\begin{aligned}\max_{H\in\mathcal{H}}\left(V_M^{\pi_{D_\infty,\phi}}(\phi(H)) - V_M^{\pi_{D,\phi}}(\phi(H))\right) \leq \\ \frac{2R_{max}}{(1-\gamma)^2}\sqrt{\frac{1}{2n} ln\left(\frac{2|\phi(\mathcal{H})||\mathrm{A}|^{1+|\phi(\mathcal{H})|}}{\delta}\right)},\end{aligned} \tag{3}$$

with probability at least $1-\delta$.

Proof. The proof of Theorem 3 is deferred to the full paper. □

Overall, Theorems 1 and 3 can help choose a good state representation for POMDPs as they provide bounds on the two terms that appear in the bias-overfitting decomposition of Equation 1. An additional feature in the mapping ϕ has an overall positive effect only if the increase of information on the belief state prevails over the additional risk of overfitting.

3 Experiments

This section provides empirical illustrations of the theoretical results on a distribution of synthetic POMDPs. Experiments on a POMDP with real-world data are available in the full paper.

3.1 Protocol

We randomly sample N_P POMDPs such that $N_{\mathcal{S}} = 5$, $N_{\mathcal{A}} = 2$ and $N_\Omega = 5$ (except when stated otherwise) from a distribution $\mathcal{P}$ that we refer to as Random POMDP. Random transition functions $T(\cdot,\cdot,\cdot)$ are drawn by assigning, for each entry (s, a, s'), a zero value with probability 3/4, and, with probability 1/4, a non-zero entry with a value drawn uniformly in $[0, 1]$. For all (s, a), if all $T(s, a, s')$ are zeros, we enforce one non-zero value for a random $s' \in \mathcal{S}$. Values are normalized. Random reward functions are generated by associating to all possible (s, a, s') a reward sampled uniformly and independently from $[-1, 1]$. Random conditional observation probabilities $O(\cdot,\cdot)$ are generated in the following way: the probability of observing $o^{(i)}$ when in state $s^{(i)}$ is equal to 0.5, while all other values are chosen uniformly randomly so that it is normalized for any s. For all POMDPs, we have $\gamma = 1$ and $\Gamma = 0.95$ if not stated otherwise and we truncate the trajectories to a length of $N_l = 100$ time steps.

For each generated POMDP $P \sim \mathcal{P}$, we generate 20 datasets $D \in \mathcal{D}_P$ where $\mathcal{D}_P$ is a probability distribution over all possible sets of n trajectories ($n \in [2, 5000]$); where each trajectory is made up of a history H_{100} of 100 time steps, when starting from an initial state $s_0 \in \mathcal{S}$ and taking uniform random decisions. Each dataset D induces a policy $\pi_{D,\phi}$, and we want to evaluate the expected return of this policy while discarding the variance related to the stochasticity of the transitions, observations and rewards. To do so, policies are tested with 1000 rollouts of the policy. For each POMDP P, we are then able to get an estimate of the average score μ_P which is defined as:

$$\mu_P = \mathop{\mathbb{E}}_{D\sim\mathcal{D}_P} \mathbb{E}\left[\sum_{t=0}^{N_l} \gamma^t r_t | s_0, \pi_{D,\phi}\right].$$

We are also able to get an estimate of a parametric variance σ_P^2 defined as:

$$\sigma_P^2 = \mathop{\mathrm{Var}}_{D\sim\mathcal{D}_P} \mathbb{E}\left[\sum_{t=0}^{N_l} \gamma^t r_t | s_0, \pi_{D,\phi}\right].$$

3.2 History Processing

We show experimentally that any additional feature from the history H_t is likely to reduce the asymptotic bias, but may also increase the overfitting. For any history length h, we consider the mapping ϕ_h that extracts the current observation and the last $h-1$ (observation, action) tuples. In the experiments, we compare the policies π_{D,ϕ_h} for $h = 1, 2, 3$.

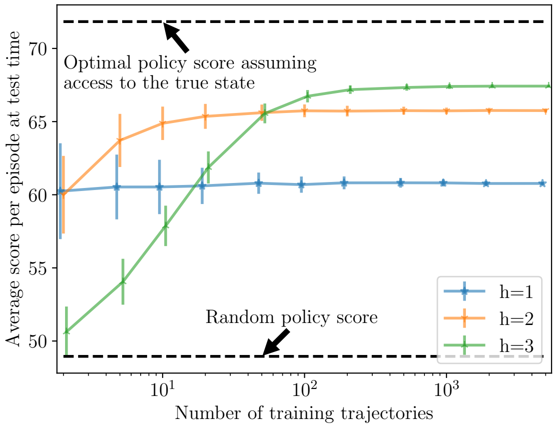

Figure 1: Evolution (as a function of the size of the dataset) of estimated values of $\mathop{\mathbb{E}}_{P\sim\mathcal{P}} \mu_P \pm \mathop{\mathbb{E}}_{P\sim\mathcal{P}} \sigma_P$ computed from a sample of $N_P = 50$ POMDPs drawn from $\mathcal{P}$. The bars are used to represent the variance observed when dealing with different datasets drawn from the distribution.

The values $\mathop{\mathbb{E}}_{P\sim\mathcal{P}} \mu_P$ and $\mathop{\mathbb{E}}_{P\sim\mathcal{P}} \sigma_P$ are displayed in Figure 1. One can observe that a small set of features (small history) appears to be a better choice when the dataset is small (only a few trajectories). With an increasing number of trajectories, the optimal number of features increases.

In addition, one can also observe that the expected variance of the score decreases as the number of samples increases. As the variance decreases, the risk of overfitting also decreases, and it becomes possible to target a larger policy class (using a larger feature set).

4 Conclusion and Future Works

In the context of POMDPs, this paper discusses the bias-overfitting tradeoff of RL algorithms with limited available data (batch RL). We show that it may be preferable to concede an asymptotic bias in order to reduce overfitting. The main theoretical results of this extended abstract relate to the state representation, for which we introduce the notion of the ϵ-sufficient statistics and showed that it enables the formalization of the bias-overfitting trade-off.

Acknowledgments

We thank the anonymous reviewers for many helpful comments. The authors thank the Walloon Region (Belgium) that funded this research in the context of the BATWAL project. We also acknowledge financial support by the Natural Science and Engineering Research Council Canada (NSERC) and Samsung Electronics Co Ltd. Guillaume Rabusseau acknowledges support of an IVADO postdoctoral fellowship.

References

[Abel *et al.*, 2016] David Abel, David Hershkowitz, and Michael Littman. Near optimal behavior via approximate state abstraction. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 2915–2923, 2016.

[Aberdeen *et al.*, 2007] Douglas Aberdeen, Olivier Buffet, and Owen Thomas. Policy-gradients for PSRs and POMDPs. In *Artificial Intelligence and Statistics*, pages 3–10, 2007.

[Cassandra *et al.*, 1994] Anthony R Cassandra, Leslie Pack Kaelbling, and Michael L Littman. Acting optimally in partially observable stochastic domains. In *Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence*, volume 94, pages 1023–1028, 1994.

[Farahmand, 2011] Amir-massoud Farahmand. *Regularization in reinforcement learning*. PhD thesis, University of Alberta, 2011.

[Ferns *et al.*, 2004] Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite Markov decision processes. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 162–169. AUAI Press, 2004.

[François-Lavet *et al.*, 2019] Vincent François-Lavet, Guillaume Rabusseau, Joelle Pineau, Damien Ernst, and Raphael Fonteneau. On overfitting and asymptotic bias in batch reinforcement learning with partial observability. *Journal of Artificial Intelligence Research*, 65:1–30, 2019.

[Ghavamzadeh *et al.*, 2015] Mohammad Ghavamzadeh, Shie Mannor, Joelle Pineau, and Aviv Tamar. Bayesian reinforcement learning: a survey. *Foundations and Trends® in Machine Learning*, 8(5-6):359–483, 2015.

[Hutter, 2014] Marcus Hutter. Extreme state aggregation beyond mdps. In *International Conference on Algorithmic Learning Theory*, pages 185–199. Springer, 2014.

[Jiang *et al.*, 2015] Nan Jiang, Alex Kulesza, and Satinder Singh. Abstraction selection in model-based reinforcement learning. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 179–188, 2015.

[Kaelbling *et al.*, 1998] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1):99–134, 1998.

[Lange *et al.*, 2012] Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch reinforcement learning. In *Reinforcement learning*, pages 45–73. Springer, 2012.

[Littman and Sutton, 2002] Michael L Littman and Richard S Sutton. Predictive representations of state. In *Advances in neural information processing systems*, pages 1555–1561, 2002.

[McCallum, 1996] Andrew Kachites McCallum. *Reinforcement learning with selective perception and hidden state*. PhD thesis, University of Rochester, 1996.

[Pineau *et al.*, 2003] Joelle Pineau, Geoff Gordon, and Sebastian Thrun. Point-based value iteration: An anytime algorithm for POMDPs. In *Proceedings of the 18th international joint conference on Artificial intelligence*, volume 3, pages 1025–1032, 2003.

[Ross *et al.*, 2011] Stéphane Ross, Joelle Pineau, Brahim Chaib-draa, and Pierre Kreitmann. A Bayesian approach for learning and planning in partially observable Markov decision processes. *The Journal of Machine Learning Research*, 12:1729–1770, 2011.

[Singh *et al.*, 2004] Satinder Singh, Michael R James, and Matthew R Rudary. Predictive state representations: A new theory for modeling dynamical systems. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 512–519. AUAI Press, 2004.

[Sondik, 1978] Edward J Sondik. The optimal control of partially observable Markov processes over the infinite horizon: Discounted costs. *Operations research*, 26(2):282–304, 1978.

[Sutton and Barto, 1998] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.