Information-Theoretic Generalization Bounds for Batch Reinforcement Learning

Xingtu Liu
Simon Fraser University
rltheory@outlook.com

Abstract

We analyze the generalization properties of batch reinforcement learning (batch RL) with value function approximation from an information-theoretic perspective. We derive generalization bounds for batch RL using (conditional) mutual information. In addition, we demonstrate how to establish a connection between certain structural assumptions on the value function space and conditional mutual information. As a by-product, we derive a *high-probability* generalization bound via conditional mutual information, which was left open in [SZ20] and may be of independent interest.

1 Introduction

Generalization is a fundamental concept in statistical machine learning. It measures how well to learning system performs on unseen data after being trained on a finite dataset. Effective generalization ensures that the learning approach captures the essential patterns in the data. Generalization in supervised learning has been studied for several decades. However, in reinforcement learning (RL), agnostic learning is generally infeasible and realizability is not a sufficient condition for efficient learning. Consequently, the study of generalization in RL poses more challenges.

In this work, we focus on batch reinforcement learning (batch RL), a branch of reinforcement learning where the agent learns a policy from a fixed dataset of previously collected experiences. This setting is favorable when online interaction is expensive, dangerous, or impractical. Batch RL, despite being a special case of supervised learning, still presents distinct challenges due to the complex temporal structures inherent in the data.

Originating from the work of [RZ16, XR17], an information-theoretic framework has been developed to bound the generalization error of learning algorithms using the mutual information between the input dataset and the output hypothesis. This methodology formalizes the intuition that overfitted learning algorithms are less likely to generalize effectively. Unlike traditional approaches such as VC-dimension and Rademacher complexity, this information-theoretic framework offers the significant advantage of capturing all dependencies on the data distribution, hypothesis space, and learning algorithm. Given that reinforcement learning is a learning paradigm in which all the aforementioned aspects differ significantly from those in supervised learning, we believe this novel approach will provide us with more profound insights.

2 Preliminaries

2.1 Batch Reinforcement Learning with Function Approximation

An episodic Markov decision process (MDP) is defined by $\mathcal{M}(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, H)$. We use $\Delta(\mathcal{X})$ to denote the set of the probability distribution over the set $\mathcal{X}$. $\mathcal{M}(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, H)$ is specified by a finite

Mathematics of Modern Machine Learning Workshop at NeurIPS 2024.

state space $\mathcal{S}$, a finite action space $\mathcal{A}$, transition functions $\mathcal{P}_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ at step $h \in [H]$, reward function $r_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ at step h, and H the number of steps in each episode. We assume the reward is bounded, i.e. $r_h(s,a) \in [0,1]$[1], $\forall (s,a,h)$.

Let $\pi = \{\pi_h : \mathcal{S} \to \Delta(\mathcal{A})\}_{h\in[H]}$ where $\pi_h(\cdot \mid s)$ is the action distribution for policy π at state s and step h. Given a policy π, the value function $V_h^\pi : \mathcal{S} \to \mathbb{R}$ at step h is defined as

$$V_h^\pi(s) := \mathbb{E}_\pi \left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \middle| s_h = s \right].$$

The action-value function $Q_h^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ at step h is defined as

$$Q_h^\pi(s,a) := \mathbb{E}_\pi \left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \middle| s_h = s, a_h = a \right].$$

The Bellman operators $\mathcal{T}_h^\pi$ and $\mathcal{T}_h^*$ project functions forward by one step through the dynamics:

$$(\mathcal{T}_h^\pi)(s,a) = r_h(s,a) + \mathbb{E}_{s'\sim\mathcal{P}_h(\cdot|s,a)}[\mathbb{E}_{a'\sim\pi(\cdot|s')}[Q(s',a')]],$$

$$(\mathcal{T}_h^*)(s,a) = r_h(s,a) + \mathbb{E}_{s'\sim\mathcal{P}_h(\cdot|s,a)} \left[\max_{a'} Q(s',a') \right].$$

Now, we denote the dataset $Z = \{(s,a,r,s',h)\}$ where $(s,a) \sim \mu_h$, $r \sim r_h(s,a)$, and $s' \sim \mathcal{P}_h(\cdot|s,a)$ for a fixed h. We also denote $\mathcal{D} = \mathcal{D}_1 \times \cdots \times \mathcal{D}_H$ where $(s,a,r,s',h) \sim \mathcal{D}_h$. We consider batch RL with value function approximation. The learner is given a function class $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_H$ to approximate the optimal Q-value function. Denote $f = (f_1, \cdots, f_H) \in \mathcal{F}$. Since no reward is collected in the $(H+1)^{\text{th}}$ step, we set $f_{H+1} = 0$. For each $f \in \mathcal{F}$, define $\pi_f = \{\pi_{f_h}\}_{h=1}^H$ where $\pi_{f_h}(a|s) = \mathbf{1}\left[a = \arg\max_{a'} f_h(s,a')\right]$. Next, we introduce the Bellman error and its empirical version.

Definition 2.1 (Bellman error)**.** *Under data distribution μ, we define the Bellman error of function $f = (f_1, \cdots, f_H)$ as*

$$\mathcal{E}(f) := \frac{1}{H} \sum_{h=1}^{H} \|f_h - \mathcal{T}_h^\star f_{h+1}\|_{\mu_h}^2. \tag{1}$$

Definition 2.2 (Mean squared empirical Bellman error (MSBE))**.** *Given a dataset $Z \sim \mathcal{D}$, we define the Mean squared empirical Bellman error (MSBE) of function $f = (f_1, \cdots, f_H)$ as*

$$L(f, Z) = \frac{1}{H} \sum_{h=1}^{H} \frac{1}{n} \sum_{(s,a,r,s',h)\in Z_h} (f_h(s,a) - r - V_{f_{h+1}}(s'))^2$$

where $V_{f_{h+1}}(s) := \max_{a\in\mathcal{A}} f_{h+1}(s,a)$.

For convenience, we denote $\ell(f_h, Z_h) = \frac{1}{n} \sum_{(s,a,r,s',h)\in Z_h} (f_h(s,a) - r - V_{f_{h+1}}(s'))^2$.

Bellman error is used in RL as a surrogate loss function to minimize the difference between the estimated value function and the true value function under a policy. The Bellman error serves as a proxy for the optimality gap, which is the difference between the current value function and the optimal value function. Under the concentrability assumption, minimizing the Bellman error is able to reduce the optimality gap.

Lemma 2.3 (Bellman error to value suboptimality [DJL21])**.** *If there exists a constant C such that for any policy π*

$$\sup_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} \frac{\mathrm{d}P_h^\pi}{\mathrm{d}\mu_h}(s,a) \le C$$

then for any $f \in \mathcal{F}$, we have

$$V_1^*(s_1) - V_1^{\pi_f}(s_1) \le 2H\sqrt{C \cdot \mathcal{E}(f)}.$$

[1]For rewards in $[R_{\min}, R_{\max}]$ simply rescale these bounds.

We note that $L(f, Z)$ is a biased estimate of $\mathcal{E}(f)$. A common solution is to use the double sampling method, where for each state and action in the sample, at least two next states are generated [SB99, ASM08, DJL21], and define the unbiased MSBE to be

$$L_{\mathrm{DS}}(f, \tilde{Z}) = \frac{1}{nH} \sum_{(s,a,r,s',\tilde{s}',h)\in\tilde{Z}} \left[\left(f_h(s,a) - r - V_{f_{h+1}}(s')\right)^2 - \frac{1}{2}\left(V_{f_{h+1}}(s') - V_{f_{h+1}}(\tilde{s}')\right)^2 \right].$$

Note that $L(f, Z) \in [0, 4H^2]$, $L_{\mathrm{DS}}(f, \tilde{Z}) \in [-2H^2, 4H^2]$, and double sampling does not increase the sample size, except that it requires an additional generated $\tilde{s}' \sim \mathcal{P}_h(\cdot|s, a)$. Therefore, the results presented in this paper can be easily extended to the double sampling setting.

2.2 Generalization Bounds

Definition 2.4 (Expected generalization bounds)**.** *Given a dataset $Z \sim \mathcal{D}$, and an algorithm $\mathcal{A}$, let $L(\mathcal{A}(Z), Z)$ denotes the training loss and let $L(\mathcal{A}(Z), \mathcal{D})$ denotes the true loss. The expected generalization error is defined as*

$$\left|\mathbb{E}_{Z\sim\mathcal{D}}[L(\mathcal{A}(Z), Z) - L(\mathcal{A}(Z), \mathcal{D})]\right|.$$

Definition 2.5 (High-probability generalization bounds)**.** *Given a dataset $Z \sim \mathcal{D}$, and an algorithm $\mathcal{A}$, let $L(\mathcal{A}(Z), Z)$ denotes the training loss and let $L(\mathcal{A}(Z), \mathcal{D})$ denotes the true loss. Given a failure probability δ and an error tolerance η, the high-probability generalization error is defined as*

$$\mathbb{P}(|L(\mathcal{A}(Z), Z) - L(\mathcal{A}(Z), \mathcal{D})| \geq \eta) \leq \delta.$$

2.3 Mutual Information

We first define the KL-divergence of two distributions.

Definition 2.6 (KL-Divergence)**.** *Let $\mathcal{P}, \mathcal{Q}$ be two distributions over the space Ω and suppose $\mathcal{P}$ is absolutely continuous with respect to $\mathcal{Q}$. The Kullback–Leibler (KL) divergence from $\mathcal{Q}$ to $\mathcal{P}$ is*

$$\mathrm{D}(\mathcal{P}\|\mathcal{Q}) = \mathbb{E}_{X\sim\mathcal{P}_X}\left[\log\frac{\mathcal{P}_X}{\mathcal{Q}_X}\right],$$

where $\mathcal{P}_X$ and $\mathcal{Q}_X$ denote the probability mass/density functions of $\mathcal{P}$ and $\mathcal{Q}$ on X, respectively.

Based on KL-divergence, we can define mutual information and conditional mutual information as follows.

Definition 2.7. *Let X, Y, and Z be arbitrary random variables, and let D_{KL} denote the Kullback–Leibler (KL) divergence. The mutual information between X and Y is defined as:*

$$I(X;Y) := D_{\mathrm{KL}}(\mathcal{P}_{X,Y}\|\mathcal{P}_X \otimes \mathcal{P}_Y).$$

The conditional mutual information is defined as:

$$I(X;Y|Z) := \mathbb{E}_Z[D_{\mathrm{KL}}(\mathcal{P}_{X,Y|Z}\|\mathcal{P}_{X|Z} \otimes \mathcal{P}_{Y|Z})].$$

Next, we introduce Rényi's α-Divergence, which is a generalization of KL-divergence. Rényi's α-Divergence has found many applications, such as hypothesis testing, differential privacy, several statistical inference and coding problems [Ver15, VEH14, Csi95, Mir17].

Definition 2.8 (Rényi's α-Divergence)**.** *Let $(\Omega, \mathcal{F}, \mathcal{P})$, $(\Omega, \mathcal{F}, \mathcal{Q})$ be two probability spaces. Let $\alpha > 0$ be a positive real different from 1. Consider a measure μ such that $\mathcal{P} \ll \mu$ and $\mathcal{Q} \ll \mu$ (such a measure always exists, e.g., $\mu = (\mathcal{P} + \mathcal{Q})/2$) and denote with p, q the densities of $\mathcal{P}, \mathcal{Q}$ with respect to μ. The α–Divergence of $\mathcal{P}$ from $\mathcal{Q}$ is defined as follows:*

$$D_\alpha(\mathcal{P}\|\mathcal{Q}) = \frac{1}{\alpha - 1}\log\int p^\alpha q^{1-\alpha}\, d\mu.$$

Note that the above definition is independent of the chosen measure μ. With the definition of Rényi's α-Divergence, we are ready to state the definitions of α-mutual information and α-conditional mutual information.

Definition 2.9 (α-mutual information)**.** *Let X, Y be two random variables jointly distributed according to $\mathcal{P}_{XY}$. Let $\mathcal{Q}_Y$ be any probability measure over $\mathcal{Y}$. For $\alpha > 0$, the α-mutual information between X and Y is defined as:*

$$I_\alpha(X;Y) = \min_{\mathcal{Q}_Y} D_\alpha(\mathcal{P}_{XY} \| \mathcal{P}_X \otimes \mathcal{Q}_Y).$$

Definition 2.10 (Conditional α-mutual information)**.** *Let X, Y, Z be three random variables jointly distributed according to $\mathcal{P}_{XYZ}$. Let $\mathcal{Q}_{Y|Z}$ be any probability measure over $\mathcal{Y}|\mathcal{Z}$. For $\alpha > 0$, a conditional α-mutual information of order α between X and Y given Z is defined as:*

$$I_\alpha^{Y|Z}(X;Y|Z) = \min_{\mathcal{Q}_{Y|Z}} D_\alpha(\mathcal{P}_{XYZ} \| \mathcal{P}_{X|Z} \otimes \mathcal{Q}_{Y|Z} \otimes \mathcal{P}_Z).$$

3 Information-Theoretic Generalization Bounds for Batch RL

We now provide expected and high-probability generalization bounds for batch RL. The generalization bounds are derived from mutual information between the training data and the learned hypothesis. Since mutual information bounds consider the data, algorithm, and hypothesis space comprehensively, they support the design of efficient learning algorithms and fine-grained theoretical analysis.

Theorem 3.1. *Given a dataset $Z \sim \mathcal{D}^n$ consists of nH samples, for any batch RL algorithm $\mathcal{A}$ with output $\mathcal{A}(Z) = f = (f_1, \cdots, f_H) \in \mathcal{F}$, the expected generalization error for the mean squared empirical Bellman error (MSBE) loss is upper bounded by*

$$|\mathbb{E}_{Z\sim\mathcal{D}}[L(\mathcal{A}(Z), Z) - L(\mathcal{A}(Z), \mathcal{D})]| \leq \sqrt{\frac{2H^2 \sum_{h=1}^H I(f_h; Z_h)}{n}}.$$

The above result suggests that reducing the mutual information between the dataset Z_h and the learned function f_h at each step h can improve generalization performance. Note that when the input domain is infinite, mutual information can become unbounded. To address this limitation, an approach based on conditional mutual information was introduced [SZ20]. CMI bounds not only address the issue by normalizing the information content of each data point, but also establish connections with various other generalization concepts, as we will discuss in the next section. We now present a generalization bound using conditional mutual information.

Definition 3.2. *Let $Z \sim \mathcal{D}^{2n}$ consist of $2n$ samples drawn independently from $\mathcal{D}$. Let $U \in \{0,1\}^n$ be uniformly random and independent from Z and the randomness of $\mathcal{A}$. Define $Z_U \in Z$ such that $(Z_U)_i$ is the $(2i - U_i)^{\text{th}}$ sample in Z – that is, Z_U is the subset of Z indexed by U. The conditional mutual information of $\mathcal{A}$ with respect to $\mathcal{D}$ is defined as $I(\mathcal{A}(Z_U); U|Z)$.*

Theorem 3.3. *Let $U \in \{0,1\}^n$ be uniformly random. Given a dataset $Z \sim \mathcal{D}^{2n}$ consists of $2nH$ samples, for any batch RL algorithm $\mathcal{A}$ with output $\mathcal{A}(Z_U) = f = (f_1, \cdots, f_H) \in \mathcal{F}$, the expected generalization error for the mean squared empirical Bellman error (MSBE) loss is upper bounded by*

$$|\mathbb{E}_{Z\sim\mathcal{D}}[L(\mathcal{A}(Z_U), Z_U) - L(\mathcal{A}(Z_U), \mathcal{D})]| \leq \sqrt{\frac{2H^2 \sum_{h=1}^H I(f_h; U|Z_h)}{n}}.$$

Note that our setting is identical to that in [DJL21], i.e. batch RL with value function approximation for episodic MDPs. They established a bound of the order $\tilde{O}\left(H^2\sqrt{\frac{1}{n}} + \sum_{h=1}^H \mathcal{R}(\mathcal{F}_h)\right)$ where $\mathcal{R}(\mathcal{F}_h)$ represents the Rademacher complexity of the function space $\mathcal{F}_h$. In contrast, our result yields an error bound of the order $O\left(H\sqrt{\frac{\sum_{h=1}^H I(f_h;Z_h)}{n}}\right)$. As demonstrated in the subsequent section, under structural assumptions like a finite pseudo-dimension or effective dimension d, this bound can be refined to $\tilde{O}\left(H^2\sqrt{\frac{d}{n}}\right)$.

Next, we proceed to derive the high-probability version of these generalization bounds using α-mutual information.

Theorem 3.4. *Given a dataset $Z \sim \mathcal{D}^n$ consists of nH samples, for any batch RL algorithm $\mathcal{A}$ with output $\mathcal{A}(Z) = f = (f_1, \cdots, f_H) \in \mathcal{F}$, if*

$$n \geq \frac{2H^4}{\varepsilon^2}\left(I_\alpha(\mathcal{A}(Z); Z) + \log 2 + \frac{\alpha}{\alpha - 1}\log\left(\frac{1}{\delta}\right)\right),$$

then the generalization error for the mean squared empirical Bellman error (MSBE) loss is upper bounded by

$$|L(\mathcal{A}(Z), Z) - L(\mathcal{A}(Z), \mathcal{D})| \leq \varepsilon$$

with probability at least $1 - \delta$.

Recall that conditional mutual information is defined as an expectation over the KL divergence. Thus, all prior works using the CMI framework have only provided bounds on the expected generalization error. We wish to establish generalization bounds with high-probability guarantees similar to Theorem 3.4. In the appendix, we prove the desired result using conditional α-mutual information, which directly implies the following theorem.

Theorem 3.5. *Let $U \in \{0,1\}^n$ be uniformly random. Given a dataset $Z \sim \mathcal{D}^{2n}$ consists of $2nH$ samples, for any batch RL algorithm $\mathcal{A}$ with output $\mathcal{A}(Z_U) = f = (f_1, \cdots, f_H) \in \mathcal{F}$, if*

$$n \geq \frac{8H^4}{\varepsilon^2}\left(\max_h I_\alpha^{f_h|Z_h}(f_h; U|Z_h) + \log 2 + \frac{\alpha}{\alpha - 1}\log\left(\frac{H}{\delta}\right)\right).$$

then the generalization error for the mean squared empirical Bellman error (MSBE) loss is upper bounded by

$$|L(\mathcal{A}(Z_U), Z_U) - L(\mathcal{A}(Z_U), \mathcal{D})| \leq \varepsilon$$

with probability at least $1 - \delta$.

4 Value Functions under Structural Assumptions

Due to the challenges stemming from large state-action spaces, long horizons, and the temporal nature of data, there is increasing interest in identifying structural assumptions for RL with value function approximation. These works include but not limited to Bellman rank [JKA+17], Witness rank [SJK+19], and Eluder dimension [WSY20]. These structural conditions aim to develop a unified theory of generalization in RL. In this section, we demonstrate that if a function class satisfies certain structural conditions reflecting a manageable complexity, the mutual information can be effectively upper bounded.

Definition 4.1 (Covering number)**.** *The covering number of a function class $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_H$ under metric $\rho(f, g) = \max_h \|f_h - g_h\|_\infty$, denoted as $\mathcal{N}(\mathcal{F}, \varepsilon)$, is the minimum integer n such that there exists a subset $\mathcal{F}_\varepsilon \subseteq \mathcal{F}$ with $|\mathcal{F}_\varepsilon| = n$, and for any $f \in \mathcal{F}$, there exists $g \in \mathcal{F}_\varepsilon$ such that $\rho(x, y) \leq \varepsilon$.*

Theorem 4.2. *Suppose the function class $\mathcal{F}$ has a covering number of $\mathcal{N}(\mathcal{F}, \varepsilon)$. Let $U \in \{0,1\}^n$ be uniformly random. Given a dataset Z consists of $2nH$ samples, for any batch RL algorithm $\mathcal{A}$ with output $\mathcal{A}(Z_U) = f = (f_1, \cdots, f_H) \in \mathcal{F}$, the expected generalization error for the mean squared empirical Bellman error (MSBE) loss is upper bounded by*

$$|\mathbb{E}_{Z\sim\mathcal{D}}[L(\mathcal{A}(Z_U), Z_U) - L(\mathcal{A}(Z_U), \mathcal{D})]| \leq \sqrt{\frac{2H^3 \log(|\mathcal{N}(\mathcal{F}, \varepsilon)|)}{n}} + 8\varepsilon H + 2\varepsilon^2.$$

Structural assumptions on the function space typically entail a finite covering number. Next, we consider the simplest case: the pseudo-dimension. The pseudo-dimension is a complexity measure of real-valued function classes, analogous to the VC dimension used for binary classification. Although the value function space may be infinite, it remains learnable if it has a finite pseudo-dimension. We also discuss the effective dimension in the appendix.

Definition 4.3 (VC-Dimension)**.** *Given hypothesis class $\mathcal{H} \subseteq \mathcal{X} \to \{0,1\}$, its VC-dimension $\mathrm{VCdim}(\mathcal{H})$ is defined as the maximal cardinality of a set $X = \{x_1, \ldots, x_{|X|}\} \subseteq \mathcal{X}$ that satisfies $|\mathcal{H}_X| = 2^{|X|}$ (or X is shattered by $\mathcal{H}$), where $\mathcal{H}_X$ is the restriction of $\mathcal{H}$ to X, namely $\{(h(x_1), \ldots, h(x_{|X|})) : h \in \mathcal{H}\}$.*

Definition 4.4 (Pseudo dimension [Hau92]). *Suppose $\mathcal{X}$ is a feature space. Given hypothesis class $\mathcal{H} \subseteq \mathcal{X} \to \mathbb{R}$, its pseudo dimension $\mathrm{Pdim}(\mathcal{H})$ is defined as $\mathrm{Pdim}(\mathcal{H}) = \mathrm{VCdim}(\mathcal{H}^+)$, where $\mathcal{H}^+ = \{(x, \xi) \mapsto 1[h(x) > \xi] : h \in \mathcal{H}\} \subseteq \mathcal{X} \times \mathbb{R} \to \{0, 1\}\}$.*

Lemma 4.5 (Bounding covering number by pseudo dimension [Hau95]). *Given hypothesis class $\mathcal{H} \subseteq \mathcal{X} \to \mathbb{R}$ with $\mathrm{Pdim}(\mathcal{H}) \leq d$, we have*

$$\log \mathcal{N}(\mathcal{H}, \varepsilon) \leq O(d \log(1/\varepsilon)).$$

Corollary 4.6. *Suppose the function class $\mathcal{F}_h \subset \mathcal{F}$ has a finite pseudo dimension $\mathrm{Pdim}(\mathcal{F}_h) = d$. For any batch RL algorithm with n training samples, the expected generalization error for the mean squared empirical Bellman error (MSBE) loss is upper bounded by $\tilde{O}(H^2\sqrt{d/n})$.*

Proof. Since $\mathrm{Pdim}(\mathcal{F}_h) = d$ and $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_H$, we have $\log \mathcal{N}(\mathcal{F}, \varepsilon) \leq O(dH \log(1/\varepsilon))$. The claim follows from Theorem 4.2 by setting $\varepsilon = H\sqrt{\frac{d}{n}}$. □

A prior study on finite sample guarantees for minimizing the Bellman error, using pseudo-dimension, demonstrated a sample complexity with a dependence of $\tilde{O}(d^2)$ [ASM08]. In contrast, our sample complexity exhibits a dependence of $\tilde{O}(d)$ on the pseudo-dimension.

We showed that when a function class contains infinitely many elements, a finite covering number can be used to upper bound the generalization error. Just as the VC-dimension imposes a finite cardinality, various concepts in real-valued function classes, such as pseudo-dimension and effective dimension, result in a finite covering number, thereby ensuring efficient learning.

5 Discussion

In this paper, we analyzed the generalization property of batch reinforcement learning within the framework of information theory. We derived generalization bounds using both conditional and unconditional mutual information. Besides, we demonstrated how to leverage the structure of the function space to guarantee generalization. Due to the merits of the information-theoretic approach, there are several appealing future research directions.

The first interesting avenue is to extend the results to the online setting. It is noteworthy that in on-policy learning, the inputs (e.g. the reward and the next state), are influenced by the output (e.g. the policy or the model), which highlights a significant disparity compared to off-policy and supervised learning. In supervised learning, a small mutual information between the input and the output indicates that the model is not overfitting. In on-policy learning, analyzing the mutual information between the input and the output can be more complicated and insightful. For example, in model-based reinforcement learning, where the model is a part of the output, a small mutual information might indicate that the learned model focuses more on the goal of maximizing the cumulative reward rather than solely capturing the transition dynamics. How to learn an effective model beyond merely fitting the transition is the central theme in decision-aware model-based reinforcement learning [WLM$^+$23, FBN17, Far18, Aba20, JFZL19, JLS$^+$22, WZS$^+$23].

As in the supervised learning setting, where various algorithms such as Stochastic Gradient Descent (SGD) [NDHR21] and Stochastic Gradient Langevin Dynamics (SGLD) have been studied [NHD$^+$19], a promising future direction is to analyze information-theoretic generalization bounds for specific reinforcement learning algorithms such as stochastic policy gradient methods.

In addition, the information-theoretic approach has the potential to unify various concepts related to generalization, such as differential privacy and stability [SZ20, HDMR21]. It would be interesting to explore how these notions in reinforcement learning can be leveraged to guarantee generalization.

Analyzing generalization for reinforcement learning is inherently more challenging than in supervised learning [DKWY19, WAS21, WWK21]. Therefore, we hope that the information-theoretic approach will provide more insights into understanding the generalization of reinforcement learning.

References

[AAV18] Amir Asadi, Emmanuel Abbe, and Sergio Verdú. Chaining mutual information and tightening generalization bounds. *Advances in Neural Information Processing Systems*, 31, 2018.

[Aba20] Romina Abachi. *Policy-aware model learning for policy gradient methods*. University of Toronto (Canada), 2020.

[Ala20] Ibrahim Alabdulmohsin. Towards a unified theory of learning and information. *Entropy*, 22(4):438, 2020.

[AMTR24] Gholamali Aminian, Saeed Masiha, Laura Toni, and Miguel RD Rodrigues. Learning algorithm generalization error bounds via auxiliary distributions. *IEEE Journal on Selected Areas in Information Theory*, 2024.

[ASM08] András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71:89–129, 2008.

[ATR21] Gholamali Aminian, Laura Toni, and Miguel RD Rodrigues. Information-theoretic bounds on the moments of the generalization error of learning algorithms. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 682–687. IEEE, 2021.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 02 2013.

[CJ19] Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR, 2019.

[CR23] Yifeng Chu and Maxim Raginsky. A unified framework for information-theoretic generalization bounds. *Advances in Neural Information Processing Systems*, 36:79260–79278, 2023.

[Csi95] Imre Csiszár. Generalized cutoff rates and rényi's information measures. *IEEE Transactions on information theory*, 41(1):26–34, 1995.

[CSM21] Qi Chen, Changjian Shui, and Mario Marchand. Generalization bounds for meta-learning: An information-theoretic analysis. *Advances in Neural Information Processing Systems*, 34:25878–25890, 2021.

[DJL21] Yaqi Duan, Chi Jin, and Zhiyuan Li. Risk bounds and rademacher complexity in batch reinforcement learning. In *International Conference on Machine Learning*, pages 2892–2902. PMLR, 2021.

[DKL+21] Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pages 2826–2836. PMLR, 2021.

[DKWY19] Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019.

[EGI21] Amedeo Roberto Esposito, Michael Gastpar, and Ibrahim Issa. Generalization error bounds via rényi-, f-divergences and maximal leakage. *IEEE Transactions on Information Theory*, 67(8):4986–5004, 2021.

[Far18] Amir-massoud Farahmand. Iterative value-aware model learning. *Advances in Neural Information Processing Systems*, 31, 2018.

[FBN17] Amir-massoud Farahmand, Andre Barreto, and Daniel Nikovski. Value-aware loss function for model-based reinforcement learning. In *Artificial Intelligence and Statistics*, pages 1486–1494. PMLR, 2017.

[FGMS08] Amir Farahmand, Mohammad Ghavamzadeh, Shie Mannor, and Csaba Szepesvári. Regularized policy iteration. *Advances in Neural Information Processing Systems*, 21, 2008.

[FGSM16] Amir-massoud Farahm, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized policy iteration with nonparametric function spaces. *Journal of Machine Learning Research*, 17(139):1–66, 2016.

[FKQR21] Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

[Hau92] David Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and computation*, 100(1):78–150, 1992.

[Hau95] David Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995.

[HDMR21] Mahdi Haghifam, Gintare Karolina Dziugaite, Shay Moran, and Dan Roy. Towards a unified information-theoretic framework for generalization. *Advances in Neural Information Processing Systems*, 34:26370–26381, 2021.

[HHT21] Haiyun He, YAN Hanshu, and Vincent Tan. Information-theoretic generalization bounds for iterative semi-supervised learning. 2021.

[HKGKS20] Hassan Hafez-Kolahi, Zeinab Golgooni, Shohreh Kasaei, and Mahdieh Soleymani. Conditioning and processing: Techniques to improve information-theoretic generalization bounds. *Advances in Neural Information Processing Systems*, 33:16457–16467, 2020.

[HNK+20] Mahdi Haghifam, Jeffrey Negrea, Ashish Khisti, Daniel M Roy, and Gintare Karolina Dziugaite. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. *Advances in Neural Information Processing Systems*, 33:9925–9935, 2020.

[HRVSG21] Hrayr Harutyunyan, Maxim Raginsky, Greg Ver Steeg, and Aram Galstyan. Information-theoretic generalization bounds for black-box learning algorithms. In *Advances in Neural Information Processing Systems*, volume 34, pages 24670–24682, 2021.

[HYG24] Haiyun He, Christina Lee Yu, and Ziv Goldfeld. Information-theoretic generalization bounds for deep neural networks. *arXiv preprint arXiv:2404.03176*, 2024.

[JFZL19] Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. *Advances in neural information processing systems*, 32, 2019.

[JKA+17] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

[JLM21] Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021.

[JLS+22] Tianying Ji, Yu Luo, Fuchun Sun, Mingxuan Jing, Fengxiang He, and Wenbing Huang. When to update your model: Constrained model-based reinforcement learning. *Advances in Neural Information Processing Systems*, 35:23150–23163, 2022.

[JS21] Sharu Theresa Jose and Osvaldo Simeone. Information-theoretic generalization bounds for meta-learning and applications. *Entropy*, 23(1):126, 2021.

[LGM12] Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Finite-sample analysis of least-squares policy iteration. *Journal of Machine Learning Research*, 13:3041–3074, 2012.

[LJ18] Adrian Tovar Lopez and Varun Jog. Generalization error bounds using wasserstein distances. In *2018 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2018.

[LVY19] Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, pages 3703–3712. PMLR, 2019.

[Mir17] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE, 2017.

[MS08] Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

[NDHR21] Gergely Neu, Gintare Karolina Dziugaite, Mahdi Haghifam, and Daniel M Roy. Information-theoretic generalization bounds for stochastic gradient descent. In *Conference on Learning Theory*, pages 3526–3545. PMLR, 2021.

[NHD+19] Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M Roy. Information-theoretic generalization bounds for sgld via data-dependent estimates. *Advances in Neural Information Processing Systems*, 32, 2019.

[PJL18] Ankit Pensia, Varun Jog, and Po-Ling Loh. Generalization error bounds for noisy, iterative algorithms. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 546–550. IEEE, 2018.

[RZ16] Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics*, pages 1232–1240. PMLR, 2016.

[SB99] Richard S Sutton and Andrew G Barto. Reinforcement learning: An introduction. *Robotica*, 17(2):229–235, 1999.

[SJK+19] 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. In *Conference on learning theory*, pages 2898–2933. PMLR, 2019.

[SZ20] Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. In *Conference on Learning Theory*, pages 3437–3452. PMLR, 2020.

[VEH14] Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[Ver15] Sergio Verdú. α-mutual information. In *2015 Information Theory and Applications Workshop (ITA)*, pages 1–6. IEEE, 2015.

[WAS21] Gellért Weisz, Philip Amortila, and Csaba Szepesvári. Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory*, pages 1237–1264. PMLR, 2021.

[WDSFC19] Hao Wang, Mario Diaz, José Cândido S Santos Filho, and Flavio P Calmon. An information-theoretic view of generalization via wasserstein distance. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 577–581. IEEE, 2019.

[WFK20] Ruosong Wang, Dean P Foster, and Sham M Kakade. What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*, 2020.

[WGC23] Hao Wang, Rui Gao, and Flavio P Calmon. Generalization bounds for noisy iterative algorithms using properties of additive noise channels. *Journal of machine learning research*, 24(26):1–43, 2023.

[WLM+23] Ran Wei, Nathan Lambert, Anthony McDonald, Alfredo Garcia, and Roberto Calandra. A unified view on solving objective mismatch in model-based reinforcement learning. *arXiv preprint arXiv:2310.06253*, 2023.

[WSY20] Ruosong Wang, Russ R Salakhutdinov, and Lin Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 33:6123–6135, 2020.

[WWK21] Yuanhao Wang, Ruosong Wang, and Sham Kakade. An exponential lower bound for linearly realizable mdp with constant suboptimality gap. *Advances in Neural Information Processing Systems*, 34:9521–9533, 2021.

[WZS+23] Xiyao Wang, Ruijie Zheng, Yanchao Sun, Ruonan Jia, Wichayaporn Wongkamjan, Huazhe Xu, and Furong Huang. Coplanner: Plan to roll out conservatively but to explore optimistically for model-based rl. *arXiv preprint arXiv:2310.07220*, 2023.

[XJ21] Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, pages 11404–11413. PMLR, 2021.

[XR17] Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in neural information processing systems*, 30, 2017.

[ZLKB20] Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR, 2020.

A Paradigm of Batch RL under the Episodic MDP Setting

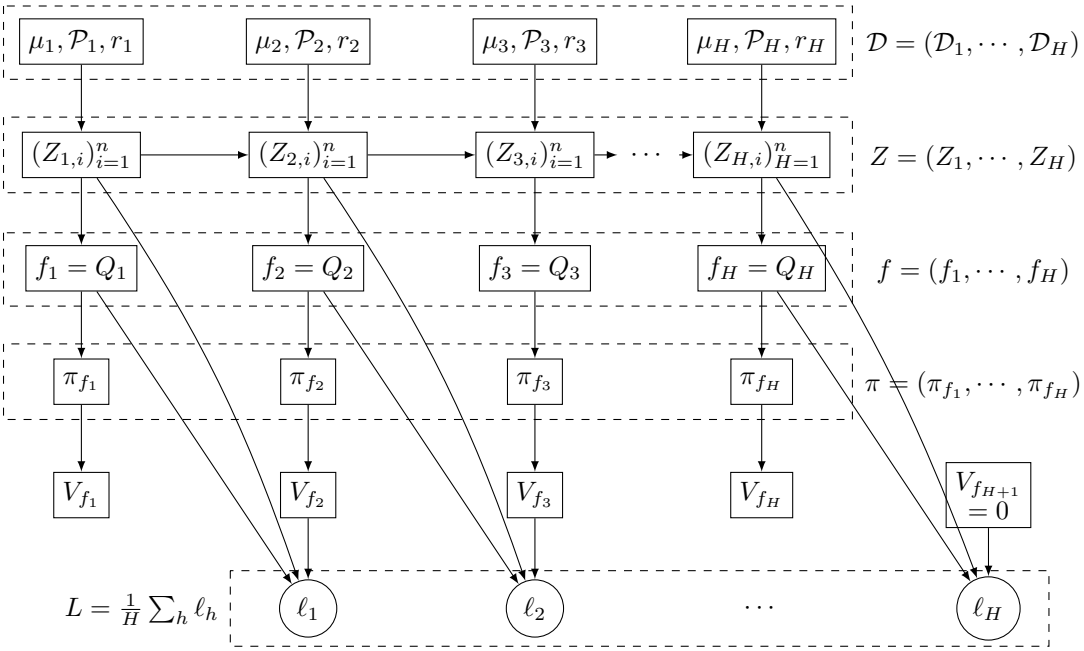

Figure 1: Directed graph representing the training process in Batch RL under episodic MDP.

B Related Work

Batch reinforcement learning A body of literature focuses on finite sample guarantees for batch reinforcement learning with function approximation [MS08, FGMS08, ZLKB20, LGM12, FGSM16, LVY19]. Common assumptions in batch RL, such as concentrability, realizability, and completeness, have also been examined in more recent studies [CJ19, WFK20, XJ21]. The most relevant work to

ours [DJL21] investigates the generalization performance of batch RL under the same setting using Rademacher complexities.

Structural conditions for efficient RL Analogous to complexity measures in supervised learning, several structural conditions have been studied to enable efficient reinforcement learning, including Bellman rank [JKA$^+$17], Witness rank [SJK$^+$19], Eluder dimension [WSY20], Bellman Eluder dimension [JLM21], and more [ZLKB20, DKL$^+$21, FKQR21]. Identifying structural conditions and classifying RL problems clarifies the limits of what can be learned and guides the design of efficient algorithms.

Information-theoretic study of generalization The information-theoretic approach was initially introduced by [RZ16, XR17] and subsequently refined to derive tighter bounds [AAV18, HKGKS20]. Besides, various other information-theoretic bounds have been proposed, leveraging concepts such as conditional mutual information [SZ20], f-divergence [EGI21], the Wasserstein distance [LJ18, WDSFC19], and more [ATR21, AMTR24]. Some studies have focused on analyzing specific algorithms [PJL18, NHD$^+$19, HNK$^+$20, HRVSG21, NDHR21, WGC23] while others have examined particular settings such as deep learning [HYG24], iterative semi-supervised learning [HHT21], and meta-learning [JS21, CSM21]. There are also works attempting to provide a unified framework for generalization from an information-theoretic perspective [HDMR21, CR23, Ala20].

C Generalization Bounds via Mutual Information

Mutual information bounds provide a direct link between the generalization error and the amount of information shared between the training data and the learned hypothesis. This offers a clear information-theoretic understanding of how overfitting can be controlled by reducing the dependency on the training data. Mutual information bounds are applicable to a wide range of learning algorithms and settings, including those with unbounded loss functions and complex hypothesis spaces. Moreover, the use of mutual information can simplify the analysis of generalization compared to traditional methods, particularly in cases where those traditional measures are difficult to compute.

Theorem C.1 ([XR17])**.** *Let $\mathcal{D}$ be a distribution on Z. Let $\mathcal{A} : Z \to \mathcal{W}$ be a randomized algorithm. Let $\ell : \mathcal{W} \times Z \to \mathbb{R}$ be a loss function which is σ-subgaussian with respect to Z. Let $L : \mathcal{W} \times Z \to \mathbb{R}$ be the empirical risk. Then*

$$|\mathbb{E}_{Z\sim\mathcal{D}}[L(\mathcal{A}(Z), Z) - L(\mathcal{A}(Z), \mathcal{D})]| \leq \sqrt{\frac{2\sigma^2}{n} I(\mathcal{A}(Z); Z)}.$$

The above theorem provides a bound on the expected generalization error. High-probability generalization bounds can be obtained using the α-mutual information. Note that the α-mutual information shares many properties with standard mutual information.

Proposition C.2 ([Ver15])**.** *For discrete random variables X and Y, the following holds:*

(i) Data Processing Inequality: given $\alpha > 0$, $I_\alpha(X, Z) \leq \min\{I_\alpha(X, Y), I_\alpha(Y, Z)\}$ if the markov chain $X - Y - Z$ holds.

(ii) $I_\alpha(X; Y)$ is non-decreasing in α.

(iii) $I_\alpha(X, Y) \leq \min\{\log |X|, \log |Y|\}$.

(iv) $I_\alpha(X, Y) \geq 0$ with equality iff X and Y are independent.

Theorem C.3 ([EGI21])**.** *Let $\mathcal{D}$ be a distribution on Z. Let $\mathcal{A} : Z \to \mathcal{W}$ be a randomized algorithm. Let $\ell : \mathcal{W} \times Z \to \mathbb{R}$ be a loss function which is σ-subgaussian with respect to Z. Let $L : \mathcal{W} \times Z \to \mathbb{R}$ be the empirical risk. Given $\eta, \delta \in (0, 1)$ and fix $\alpha \geq 1$, if the number of samples n satisfies*

$$n \geq \frac{2\sigma^2}{\eta^2} \left(I_\alpha(\mathcal{A}(Z); Z) + \log 2 + \frac{\alpha}{\alpha - 1} \log \left(\frac{1}{\delta} \right) \right).$$

then we have

$$\mathbb{P}\left(|L(\mathcal{A}(Z), Z) - L(\mathcal{A}(Z), \mathcal{D})| \leq \eta\right) \geq 1 - \delta.$$

The mutual information bound can be infinite in some cases and thus be vacuous. To address this, the conditional mutual information (CMI) approach was introduced. CMI bounds normalize the information content for each data point, preventing the problem of infinite information content, particularly in continuous data distributions. This makes CMI a more robust and applicable method in scenarios where mutual information would otherwise be unbounded.

Theorem C.4 ([SZ20])**.** *Let $\mathcal{D}$ be a distribution on Z. Let $\mathcal{A} : Z \to \mathcal{W}$ be a randomized algorithm. Let $L : \mathcal{W} \times Z \to \mathbb{R}$ be a function such that $|L(w, z_1) - L(w, z_2)| \leq \Delta(z_1, z_2)$ for all $z_1, z_2 \in \mathcal{Z}$ and $w \in \mathcal{W}$ given $\Delta : Z^2 \to \mathbb{R}$. Let $U \in \{0,1\}^n$ be uniformly random. Then*

$$|\mathbb{E}_{Z\sim\mathcal{D}}[L(\mathcal{A}(Z_U), Z_U) - L(\mathcal{A}(Z_U), \mathcal{D})]| \leq \sqrt{\frac{2\mathbb{E}_{z_1,z_2}[\Delta(z_1,z_2)^2]}{n} I(\mathcal{A}(Z_U); U|Z)}.$$

Another advantage of the CMI bounds is that they can be derived from various concepts such as VC-dimension, compression schemes, stability, and differential privacy, offering a unified framework for generalization analysis. However, because CMI is defined as an expectation, i.e. $I(X;Y|Z) := \mathbb{E}_Z[D_{\mathrm{KL}}(\mathcal{P}_{X,Y|Z} \| \mathcal{P}_{X|Z} \otimes \mathcal{P}_{Y|Z})]$, the above theorem does not provide a high-probability bound. Modifying this framework to ensure high-probability guarantees was left as future work in [SZ20]. In the following, we use conditional α-mutual information to address this issue.

Theorem C.5. *Let $U \in \{0,1\}^n$ be uniformly random. Given a dataset $Z \sim \mathcal{D}^{2n}$ consists of $2n$ samples. Let $\mathcal{A} : Z_U \to \mathcal{W}$ be a randomized algorithm. Let $\ell : \mathcal{W} \times Z \to \mathbb{R}$ be a loss function which is σ-subgaussian with respect to Z. Let $L : \mathcal{W} \times Z_U \to \mathbb{R}$ be the empirical risk. Given $\eta, \delta \in (0,1)$ and fix $\alpha \geq 1$, if the number of samples n satisfies*

$$n \geq \frac{2\sigma^2}{\eta^2}\left(I_\alpha^{\mathcal{A}(Z_U)|Z}(\mathcal{A}(Z_U); U|Z) + \log 2 + \frac{\alpha}{\alpha - 1}\log\left(\frac{1}{\delta}\right)\right)$$

then we have

$$\mathbb{P}\left(|L(\mathcal{A}(Z_U), Z_U) - L(\mathcal{A}(Z_U), \mathcal{D})| \leq \eta\right) \geq 1 - \delta.$$

Proof. See Appendix E. □

D Effective Dimension

In this section, we introduce another complexity measure known as the effective dimension [DKL+21], which has a similar covering number to the pseudo-dimension. The effective dimension quantifies how the function class responds to data, indicating the minimum number of samples required to learn effectively.

Definition D.1 (ε-effective dimension of a set [DKL+21])**.** *The ε-effective dimension of a set $\mathcal{X}$ is the minimum integer $d_{\mathit{eff}}(\mathcal{X}, \varepsilon) = n$ such that*

$$\sup_{x_1,\ldots,x_n \in \mathcal{X}} \frac{1}{n} \log\det\left(I + \frac{1}{\varepsilon^2}\sum_{i=1}^n x_i x_i^\top\right) \leq e^{-1}.$$

Definition D.2 (ε-effective dimension of a function class [DKL+21])**.** *Given a function class $\mathcal{F}$ defined on $\mathcal{X}$, its ε-effective dimension $d_{\mathit{eff}}(\mathcal{F}, \varepsilon) = n$ is the minimum integer n such that there exists a separable Hilbert space $\mathcal{H}$ and a mapping $\phi : \mathcal{X} \to \mathcal{H}$ so that*

- *for every $f \in \mathcal{F}$ there exists $\theta_f \in B_{\mathcal{H}}(1)$ satisfying $f(x) = \langle \theta_f, \phi(x)\rangle_{\mathcal{H}}$ for all $x \in \mathcal{X}$,*
- *$d_{\mathit{eff}}(\phi(\mathcal{X}), \varepsilon) = n$ where $\phi(\mathcal{X}) = \{\phi(x) : x \in \mathcal{X}\}$.*

Definition D.3 (Kernel MDPs [JLM21])**.** *In a kernel MDP of effective dimension d, for each step $h \in [H]$, there exist feature mappings $\phi_h : S \times A \to \mathcal{H}$ and $\psi_h : S \to \mathcal{H}$ where $\mathcal{H}$ is a separable Hilbert space, so that the transition measure can be represented as the inner product of features, i.e.,*

$$\mathbb{P}_h(s' \mid s, a) = \langle \phi_h(s,a), \psi_h(s')\rangle_{\mathcal{H}}.$$

Besides, the reward function is linear in ϕ, i.e.,

$$r_h(s,a) = \langle \phi_h(s,a), \theta_h^r\rangle_{\mathcal{H}}$$

for some $\theta_h^r \in \mathcal{H}$. Here, ϕ is known to the learner while ψ and θ_h^r are unknown. Moreover, a kernel MDP satisfies the following regularization conditions: for all h

- $\|\theta_h^r\|_{\mathcal{H}} \leq 1$ *and* $\|\phi_h(s,a)\|_{\mathcal{H}} \leq 1$ *for all* s, a.
- $\left\|\sum_{s\in S} \mathcal{V}(s)\psi_h(s)\right\|_{\mathcal{H}} \leq 1$ *for any function* $\mathcal{V} : S \to [0,1]$.
- $dim_{\mathit{eff}}(X_h, \varepsilon) \leq d$ *for all* h, *where* $X_h = \{\phi_h(s,a) : (s,a) \in S \times A\}$.

Kernel MDPs are extensions of the traditional MDPs where the transition dynamics and rewards are represented in a Reproducing Kernel Hilbert Space (RKHS). In this setup, the value functions or Q-functions are approximated using kernel methods, allowing the model to capture more complex dependencies in the data compared to linear models. To learn kernel MDPs, it is necessary to construct a function class $\mathcal{F}$.

Lemma D.4 (Bounding covering number by effective dimension [JLM21])**.** *Let* $\mathcal{M}$ *be a kernel MDP of effective dimension* d, *then*

$$\log \mathcal{N}(\mathcal{F}, \varepsilon) \leq O(Hd \log(1 + dH/\varepsilon)).$$

Corollary D.5. *Suppose the function class* $\mathcal{F}$ *has a finite effective dimension* d. *For any batch RL algorithm with* n *training samples, the expected generalization error for the mean squared empirical Bellman error (MSBE) loss is upper bounded by* $\tilde{O}(H^2\sqrt{d/n})$.

E Proof of Theorem C.5

Theorem E.1 (Theorem C.5 restated)**.** *Let* $U \in \{0,1\}^n$ *be uniformly random. Given a dataset* $Z \sim \mathcal{D}^{2n}$ *consists of* $2n$ *samples. Let* $\mathcal{A} : Z_U \to \mathcal{W}$ *be a randomized algorithm. Let* $\ell : \mathcal{W} \times Z \to \mathbb{R}$ *be a loss function which is* σ*-subgaussian with respect to* Z. *Let* $L : \mathcal{W} \times Z_U \to \mathbb{R}$ *be the empirical risk. Given* $\eta, \delta \in (0,1)$ *and fix* $\alpha \geq 1$, *if the number of samples* n *satisfies*

$$n \geq \frac{2\sigma^2}{\eta^2}\left(I_\alpha^{\mathcal{A}(Z_U)|Z}(\mathcal{A}(Z_U); U|Z) + \log 2 + \frac{\alpha}{\alpha - 1}\log\left(\frac{1}{\delta}\right)\right)$$

then we have

$$\mathbb{P}\left(|L(\mathcal{A}(Z_U), Z_U) - L(\mathcal{A}(Z_U), \mathcal{D})| \leq \eta\right) \geq 1 - \delta.$$

Proof. Let $(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}, \mathcal{F}, \mathcal{P}_{XYZ})$ be a probability space, and let $\mathcal{Q}(\mathcal{X}|\mathcal{Z})$ be the set of conditional probability measures $\mathcal{Q}_{X|Z}$ such that $\mathcal{P}_{XYZ} \ll \mathcal{P}_Z Q_{X|Z} \mathcal{P}_{Y|Z}$. Given $E \in \mathcal{F}$ and $z \in \mathcal{Z}, x \in \mathcal{X}$, let $E_{z,x} = \{y \in \mathcal{Y} : (x,y,z) \in E\}$. We first prove that for a fixed $\alpha \geq 1$,

$$\mathcal{P}_{XYZ}(E) \leq \mathbb{E}_Z\left[\operatorname{ess\,sup}_{Q_{X|Z} \in \mathcal{Q}(\mathcal{X}|\mathcal{Z})} \mathcal{P}_{Y|Z}(E_{Z,X})\right]^{\frac{\alpha-1}{\alpha}} \exp\left(\frac{\alpha - 1}{\alpha} I_\alpha^{X|Z}(X; Y|Z)\right). \quad (2)$$

Using the Radon-Nikodym derivative of $\mathcal{P}_{XYZ}$ with respect to the product measure $\mathcal{P}_Z \mathcal{Q}_{X|Z} \mathcal{P}_{Y|Z}$, we have

$$\mathcal{P}_{XYZ}(E) = \mathbb{E}_{\mathcal{P}_Z Q_{X|Z} \mathcal{P}_{Y|Z}}\left[\frac{d\mathcal{P}_{XYZ}}{d\mathcal{P}_Z \mathcal{Q}_{X|Z} \mathcal{P}_{Y|Z}} \mathcal{I}_E\right]$$

where $\mathcal{I}_E$ is the indicator function of the event E. Next, we introduce three sets of exponents $\alpha'', \alpha', \alpha$ and $\gamma'', \gamma', \gamma$, such that

$$\frac{1}{\alpha''} + \frac{1}{\gamma''} = \frac{1}{\alpha'} + \frac{1}{\gamma'} = \frac{1}{\alpha} + \frac{1}{\gamma} = 1.$$

By applying Hölder's inequality three times to separate the different components of the expectation, we derive

$$\begin{aligned}
&\mathbb{E}_{\mathcal{P}_Z Q_{X|Z} \mathcal{P}_{Y|Z}}\left[\frac{d\mathcal{P}_{XYZ}}{d\mathcal{P}_Z \mathcal{Q}_{X|Z} \mathcal{P}_{Y|Z}} \mathcal{I}_E\right] \\
&\leq \mathbb{E}_{\mathcal{P}_Z}^{\frac{1}{\alpha''}}\left[\mathbb{E}_{\mathcal{Q}_{X|Z}}^{\frac{\alpha''}{\alpha'}}\left[\mathbb{E}_{\mathcal{Q}_{Y|Z}}^{\frac{\alpha}{\alpha'}}\left[\left(\frac{d\mathcal{P}_{XYZ}}{d\mathcal{P}_Z \mathcal{Q}_{X|Z} \mathcal{P}_{Y|Z}}\right)^{\alpha}\right]\right]\right] \mathbb{E}_{\mathcal{P}_Z}^{\frac{1}{\gamma''}}\left[\mathbb{E}_{\mathcal{Q}_{X|Z}}^{\frac{\gamma''}{\gamma'}}\left[\mathbb{E}_{\mathcal{P}_{Y|Z}}^{\frac{\gamma'}{\gamma}}\left[\mathcal{I}_E^{\gamma}\right]\right]\right].
\end{aligned}$$

By setting $\alpha'' = \alpha$ and $\alpha' = 1$,

$$\mathbb{E}_{\mathcal{P}_Z}^{\frac{1}{\alpha''}}\left[\mathbb{E}_{\mathcal{Q}_{X|Z}}^{\frac{\alpha''}{\alpha'}}\left[\mathbb{E}_{\mathcal{Q}_{Y|Z}}^{\frac{\alpha}{\alpha'}}\left[\left(\frac{d\mathcal{P}_{XYZ}}{d\mathcal{P}_Z \mathcal{Q}_{X|Z} \mathcal{P}_{Y|Z}}\right)^{\alpha}\right]\right]\right] \leq \exp\left(\frac{\alpha - 1}{\alpha} I_\alpha^{X|Z}(X; Y|Z)\right).$$

Since $\alpha' = 1$ and $\frac{1}{\alpha'} + \frac{1}{\gamma'} = 1$, we have $\gamma' \to \infty$. As $\gamma' \to \infty$, $\mathbb{E}_{\mathcal{Q}_{X|Z}}^{\frac{\gamma''}{\gamma'}}\left[\mathbb{E}_{\mathcal{P}_{Y|Z}}^{\frac{\gamma'}{\gamma}}\left[\mathcal{I}_E^{\gamma}\right]\right]$ tends to the essential supremum

$$\operatorname{ess\,sup}_{Q_{X|Z}\in\mathcal{Q}(\mathcal{X}|\mathcal{Z})} \mathcal{P}_{Y|Z}(E_{Z,X}).$$

As $\frac{1}{\gamma''} = \frac{\alpha-1}{\alpha}$, we have

$$\mathbb{E}_{\mathcal{P}_Z}^{\frac{1}{\gamma''}}\left[\mathbb{E}_{\mathcal{Q}_{X|Z}}^{\frac{\gamma''}{\gamma'}}\left[\mathbb{E}_{\mathcal{P}_{Y|Z}}^{\frac{\gamma'}{\gamma}}\left[\mathcal{I}_E^{\gamma}\right]\right]\right] \le \mathbb{E}_Z\left[\operatorname{ess\,sup}_{Q_{X|Z}\in\mathcal{Q}(\mathcal{X}|\mathcal{Z})} \mathcal{P}_{Y|Z}(E_{Z,X})\right]^{\frac{\alpha-1}{\alpha}}.$$

Thus, eq. (2) holds by combining all the inequalities.

Now, let $X = \mathcal{A}(Z_U)$ and $Y = U$. Consider the event

$$E = \{(X, Y, Z) : |L(X, Z_Y) - \mathbb{E}_Y[L(X, \mathcal{D})]| \ge \eta\},$$

where $L(X, Z_Y)$ denotes the empirical risk defined as the average of n loss functions, and each loss function is σ-subgaussian. We can express $E_{Z,X}$, the fibers of E with respect to Z and X, as

$$E_{Z,X} = \{Y : |L(X, Z_Y) - \mathbb{E}_Y[L(X, \mathcal{D})]| \ge \eta\}.$$

For any fixed Z and X, the random variable Y remains independent of Z and X under any $\mathcal{Q}_{X|Z} \in \mathcal{Q}(\mathcal{X}|\mathcal{Z})$. Now, by Hoeffding's inequality, for every X and Z,

$$\mathcal{P}_Y(E_{Z,X}) \le 2\exp\left(\frac{-n\eta^2}{2\sigma^2}\right). \tag{3}$$

Therefore, from eq. (2) and eq. (3),

$$\begin{aligned}\mathbb{P}(E) &\le 2\exp\left(\frac{\alpha-1}{\alpha}\cdot\frac{-n\eta^2}{2\sigma^2}\right)\exp\left(\frac{\alpha-1}{\alpha}I_\alpha^{\mathcal{A}(Z_U)|Z}(\mathcal{A}(Z_U); U|Z)\right)\\ &= 2\exp\left(\frac{\alpha-1}{\alpha}\left(I_\alpha^{\mathcal{A}(Z_U)|Z}(\mathcal{A}(Z_U); U|Z) - \frac{-n\eta^2}{2\sigma^2}\right)\right).\end{aligned}$$

Lastly, by setting

$$n \ge \frac{2\sigma^2}{\eta^2}\left(I_\alpha^{\mathcal{A}(Z_U)|Z}(\mathcal{A}(Z_U); U|Z) + \log 2 + \frac{\alpha}{\alpha-1}\log\left(\frac{1}{\delta}\right)\right)$$

we obtain the desired conclusion. □

F Proof of Theorem 3.1

Theorem F.1 (Theorem 3.1 restated)**.** *Given a dataset $Z \sim \mathcal{D}^n$ consists of nH samples, for any batch RL algorithm $\mathcal{A}$ with output $\mathcal{A}(Z) = f = (f_1, \cdots, f_H) \in \mathcal{F}$, the expected generalization error for the mean squared empirical Bellman error (MSBE) loss is upper bounded by*

$$|\mathbb{E}_{Z\sim\mathcal{D}}[L(\mathcal{A}(Z), Z) - L(\mathcal{A}(Z), \mathcal{D})]| \le \sqrt{\frac{2H^2\sum_{h=1}^H I(f_h; Z_h)}{n}}.$$

Proof. We first recall the Donsker–Varadhan variational representation ([BLM13]) of the KL-divergence between any two probability measures π and ρ on a common measurable space $(\Omega, \mathcal{F})$

$$D_{\mathrm{KL}}(\pi\|\rho) = \sup_F\left\{\int_\Omega F\, d\pi - \log\int_\Omega e^F\, d\rho\right\}$$

where the supremum is over all measurable functions $F : \Omega \to \mathbb{R}$ such that $e^F \in L^1(\rho)$.

Let be $Z = Z_1 \cup \cdots \cup Z_H$ be a dataset where $Z_h = \{(s, a, r, s', h)\} \sim \mathcal{D}_h$. Let $\mathcal{A}(Z) = f = (f_1, \cdots, f_H) \in \mathcal{F}$ be the output of some batch RL algorithm $\mathcal{A}$. Let $\tilde{f}_h$ and $\tilde{Z}_h$ be the independent

copies of f_h and Z_h. Let

$$\begin{aligned}L(f, Z) &= \frac{1}{H}\sum_{h=1}^{H} \ell(f_h, Z_h) \\ &= \frac{1}{H}\sum_{h=1}^{H} \frac{1}{n} \sum_{(s,a,r,s',h)\in Z_h} (f_h(s,a) - r - V_{f_{h+1}}(s'))^2.\end{aligned}$$

Now, we have

$$\begin{aligned}I(f_h; Z_h) &= D_{\mathrm{KL}}(P_{f_h,Z_h} \| P_{f_h} \otimes P_{Z_h}) \\ &= \sup_g \left\{ \mathbb{E}_{f_h,Z_h}[g(f_h, Z_h)] - \log \mathbb{E}_{\tilde{f}_h,\tilde{Z}_h}[e^{g(\tilde{f}_h,\tilde{Z}_h)}] \right\} \\ &\qquad \text{(Donsker–Varadhan variational representation)} \\ &\geq \lambda \mathbb{E}_{f_h,Z_h}[\ell(f_h, Z_h)] - \log \mathbb{E}_{\tilde{f}_h,\tilde{Z}_h}[e^{\lambda \ell(\tilde{f}_h,\tilde{Z}_h)}]. \qquad (\forall \lambda \in \mathbb{R})\end{aligned}$$

Since $\ell(f_h, Z_h) = \frac{1}{n}\sum_{(s,a,r,s',h)\in Z_h} (f_h(s,a) - r - V_{f_{h+1}}(s'))^2$ and $(f_h(s,a) - r - V_{f_{h+1}}(s'))^2 \in [0, 4H^2]$ for any h, it follows that

$$\log \mathbb{E}_{\tilde{f}_h,\tilde{Z}_h}[e^{\lambda(\ell(\tilde{f}_h,\tilde{Z}_h) - \mathbb{E}_{\tilde{f}_h,\tilde{Z}_h}[\ell(\tilde{f}_h,\tilde{Z}_h)])}] \leq \frac{2\lambda^2 H^4}{n}.$$

Thus, we obtain

$$\begin{aligned}I(f_h; Z_h) &\geq \lambda \left(\mathbb{E}_{f_h,Z_h}[\ell(f_h, Z_h)] - \mathbb{E}_{\tilde{f}_h,\tilde{Z}_h}[\ell(\tilde{f}_h, \tilde{Z}_h)]) \right] - \frac{2\lambda^2 H^4}{n} \\ \Rightarrow \frac{I(f_h; Z_h)}{\lambda} + \frac{2\lambda^2 H^4}{n} &\geq \mathbb{E}_{f_h,Z_h}[\ell(f_h, Z_h)] - \mathbb{E}_{\tilde{f}_h,\tilde{Z}_h}[\ell(\tilde{f}_h, \tilde{Z}_h)])].\end{aligned}$$

By optimizing the above inequality over $\lambda > 0$ and $\lambda < 0$, respectively, we derive

$$-H^2\sqrt{\frac{2I(f_h; Z_h)}{n}} \leq \mathbb{E}_{f_h,Z_h}[\ell(f_h, Z_h)] - \mathbb{E}_{\tilde{f}_h,\tilde{Z}_h}[\ell(\tilde{f}_h, \tilde{Z}_h)])] \leq H^2\sqrt{\frac{2I(f_h; Z_h)}{n}},$$

and thus,

$$\left| \mathbb{E}_{f_h,Z_h}[\ell(f_h, Z_h)] - \mathbb{E}_{\tilde{f}_h,\tilde{Z}_h}[\ell(\tilde{f}_h, \tilde{Z}_h)])] \right| \leq H^2\sqrt{\frac{2I(f_h; Z_h)}{n}}. \tag{4}$$

Finally, we observe that

$$\begin{aligned}|\mathbb{E}_{Z\sim\mathcal{D}}[L(\mathcal{A}(Z), Z) - L(\mathcal{A}(Z), \mathcal{D})]| &= \left| \mathbb{E}_{Z\sim\mathcal{D}} \left[\frac{1}{H}\sum_{h=1}^{H} \ell(f_h, Z_h) - \mathbb{E}_{Z\sim\mathcal{D}} \left[\frac{1}{H}\sum_{h=1}^{H} \ell(f_h, Z_h) \right] \right] \right| \\ &= \left| \frac{1}{H}\sum_{h=1}^{H} \mathbb{E}_{Z_h\sim\mathcal{D}_h} \left[\ell(f_h, Z_h) - \mathbb{E}_{Z_h\sim\mathcal{D}_h} \left[\ell(f_h, Z_h) \right] \right] \right| \\ &= \frac{1}{H}\sum_{h=1}^{H} \left| \mathbb{E}_{f_h,Z_h}[\ell(f_h, Z_h)] - \mathbb{E}_{\tilde{f}_h,\tilde{Z}_h}[\ell(\tilde{f}_h, \tilde{Z}_h)]) \right| \\ &\leq \frac{1}{H}\sum_{h=1}^{H} H^2\sqrt{\frac{2I(f_h; Z_h)}{n}} \qquad \text{(eq. (4))} \\ &= \sqrt{\frac{2H^2\sum_{h=1}^{H} I(f_h; Z_h)}{n}}.\end{aligned}$$

□

G Proof of Theorem 3.3

Theorem G.1 (Theorem 3.3 restated)**.** *Let $U \in \{0,1\}^n$ be uniformly random. Given a dataset $Z \sim \mathcal{D}^{2n}$ consists of $2nH$ samples, for any batch RL algorithm $\mathcal{A}$ with output $\mathcal{A}(Z_U) = f = (f_1, \cdots, f_H) \in \mathcal{F}$, the expected generalization error for the mean squared empirical Bellman error (MSBE) loss is upper bounded by*

$$|\mathbb{E}_{Z\sim\mathcal{D}}[L(\mathcal{A}(Z_U), Z_U) - L(\mathcal{A}(Z_U), \mathcal{D})]| \leq \sqrt{\frac{2H^2 \sum_{h=1}^{H} I(f_h; U|Z_h)}{n}}.$$

Proof. Let $U \in \{0,1\}^n$ be uniformly random. Let be $Z = Z_1 \cup \cdots \cup Z_H$ be a dataset where each $Z_h = \{(s, a, r, s', h)\} \sim \mathcal{D}_h$ consists of $2n$ samples. Define $Z_U = (Z_1)_U \cup \cdots \cup (Z_H)_U$. Let $\mathcal{A}(Z_U) = f = (f_1, \cdots, f_H) \in \mathcal{F}$ be the output of some batch RL algorithm $\mathcal{A}$. Let $\bar{f}_h = \mathcal{A}(Z_{\bar{U}})_h$, $\tilde{Z}_h = (Z_h)_U$ and $\bar{Z}_h = (Z_h)_{\bar{U}}$. Note that $Z_h = \tilde{Z}_h \cup \bar{Z}_h$. We define the disintegrated mutual information

$$I^Z(X;Y) := D_{KL}(P_{X,Y|Z} \| P_X P_Y | Z).$$

Note that $I(X;Y|Z) = \mathbb{E}_Z[I^Z(X;Y)]$. The rest of the proof is analogous to Theorem 3.1. We have

$$\begin{aligned}
I^{Z_h}(f_h; \tilde{Z}_h|Z_h) &= D_{\mathrm{KL}}(P_{f_h,\tilde{Z}_h|Z_h} \| P_{f_h|Z_h} \otimes P_{\tilde{Z}_h|Z_h}) \\
&= \sup_g \left\{ \mathbb{E}_{f_h,\tilde{Z}_h|Z_h}[g(f_h, \tilde{Z}_h)] - \log \mathbb{E}_{\bar{f}_h,\bar{Z}_h|Z_h}[e^{g(\bar{f}_h,\bar{Z}_h)}] \right\} \\
&\qquad \text{(Donsker–Varadhan variational representation)} \\
&\geq \lambda \mathbb{E}_{f_h,\tilde{Z}_h|Z_h}[\ell(f_h, \tilde{Z}_h)] - \log \mathbb{E}_{\bar{f}_h,\bar{Z}_h|Z_h}[e^{\lambda \ell(\bar{f}_h,\bar{Z}_h)}]. \qquad (\forall \lambda \in \mathbb{R})
\end{aligned}$$

Since $\ell(f_h, Z_h) = \frac{1}{n} \sum_{(s,a,r,s',h)\in Z_h} (f_h(s,a) - r - V_{f_{h+1}}(s'))^2$ and $(f_h(s,a) - r - V_{f_{h+1}}(s'))^2 \in [0, 4H^2]$ for any h, it follows that

$$\log \mathbb{E}_{\bar{f}_h,\bar{Z}_h|Z_h}[e^{\lambda(\ell(\bar{f}_h,\bar{Z}_h) - \mathbb{E}_{\bar{f}_h,\bar{Z}_h|Z_h}[\ell(\bar{f}_h,\bar{Z}_h)])}] \leq \frac{2\lambda^2 H^4}{n}.$$

Thus, we obtain

$$\begin{aligned}
I^{Z_h}(f_h; \tilde{Z}_h|Z_h) &\geq \lambda \left(\mathbb{E}_{f_h,\tilde{Z}_h|Z_h}[\ell(f_h, \tilde{Z}_h)] - \mathbb{E}_{\bar{f}_h,\bar{Z}_h|Z_h}[\ell(\bar{f}_h, \bar{Z}_h)]) \right] \right) - \frac{2\lambda^2 H^4}{n} \\
\Rightarrow \frac{I^{Z_h}(f_h; \tilde{Z}_h|Z_h)}{\lambda} + \frac{2\lambda^2 H^4}{n} &\geq \mathbb{E}_{f_h,\tilde{Z}_h|Z_h}[\ell(f_h, \tilde{Z}_h)] - \mathbb{E}_{\bar{f}_h,\bar{Z}_h|Z_h}[\ell(\bar{f}_h, \bar{Z}_h)])].
\end{aligned}$$

By optimizing the above inequality over $\lambda > 0$ and $\lambda < 0$, respectively, we derive

$$-H^2 \sqrt{\frac{2I^{Z_h}(f_h; \tilde{Z}_h|Z_h)}{n}} \leq \mathbb{E}_{f_h,\tilde{Z}_h|Z_h}[\ell(f_h, \tilde{Z}_h)] - \mathbb{E}_{\bar{f}_h,\bar{Z}_h|Z_h}[\ell(\bar{f}_h, \bar{Z}_h)])] \leq H^2 \sqrt{\frac{2I^{Z_h}(f_h; \tilde{Z}_h|Z_h)}{n}},$$

and thus,

$$\left| \mathbb{E}_{f_h,\tilde{Z}_h|Z_h}[\ell(f_h, \tilde{Z}_h)] - \mathbb{E}_{\bar{f}_h,\bar{Z}_h|Z_h}[\ell(\bar{f}_h, \bar{Z}_h)]) \right] \right| \leq H^2 \sqrt{\frac{2I^{Z_h}(f_h; \tilde{Z}_h|Z_h)}{n}}. \tag{5}$$

Finally, we conclude that

$$
\begin{aligned}
|\mathbb{E}_{Z\sim\mathcal{D}}[L(\mathcal{A}(Z_U), Z_U) - L(\mathcal{A}(Z_U), \mathcal{D})]| &= \left|\mathbb{E}_{Z\sim\mathcal{D}}\left[\frac{1}{H}\sum_{h=1}^{H}\ell(f_h, \tilde{Z}_h) - \mathbb{E}_{Z\sim\mathcal{D}}\left[\frac{1}{H}\sum_{h=1}^{H}\ell(f_h, \tilde{Z}_h)\right]\right]\right| \\
&= \left|\frac{1}{H}\sum_{h=1}^{H}\mathbb{E}_{Z_h\sim\mathcal{D}_h}\left[\ell(f_h, \tilde{Z}) - \mathbb{E}_{Z_h\sim\mathcal{D}_h}\left[\ell(f_h, \tilde{Z})\right]\right]\right| \\
&\le \frac{1}{H}\sum_{h=1}^{H}\left|\mathbb{E}_{Z_h\sim\mathcal{D}_h}\left[\mathbb{E}_{\bar{f}_h, \bar{Z}_h|Z_h}[\ell(\bar{f}_h, \bar{Z}_h)])] - \mathbb{E}_{f_h, \tilde{Z}_h|Z_h}[\ell(f_h, \tilde{Z}_h)])\right]\right| \\
&\le \frac{1}{H}\sum_{h=1}^{H}H^2\mathbb{E}_{Z_h\sim\mathcal{D}_h}\left[\sqrt{\frac{2I^{Z_h}(f_h; \tilde{Z}_h|Z_h)}{n}}\right] \quad \text{(eq. (5))} \\
&\le \frac{1}{H}\sum_{h=1}^{H}H^2\sqrt{\frac{2\mathbb{E}_{Z_h\sim\mathcal{D}_h}[I^{Z_h}(f_h; \tilde{Z}_h|Z_h)]}{n}} \\
&= \sqrt{\frac{2H^2\sum_{h=1}^{H}I(f_h; \tilde{Z}_h|Z_h)}{n}} \\
&= \sqrt{\frac{2H^2\sum_{h=1}^{H}I(f_h; U|Z_h)}{n}}.
\end{aligned}
$$

□

H Proof of Theorem 3.4

Theorem H.1 (Theorem 3.4 restated)**.** *Given a dataset $Z \sim \mathcal{D}^n$ consists of nH samples, for any batch RL algorithm $\mathcal{A}$ with output $\mathcal{A}(Z) = f = (f_1, \cdots, f_H) \in \mathcal{F}$, if*

$$
n \ge \frac{8H^4}{\varepsilon^2}\left(\max_h I_\alpha(f_h; Z_h) + \log 2 + \frac{\alpha}{\alpha - 1}\log\left(\frac{H}{\delta}\right)\right).
$$

then the generalization error for the mean squared empirical Bellman error (MSBE) loss is upper bounded by

$$
|L(\mathcal{A}(Z), Z) - L(\mathcal{A}(Z), \mathcal{D})| \le \varepsilon
$$

with probability at least $1 - \delta$.

Proof. Let be $Z = Z_1 \cup \cdots \cup Z_H$ be a dataset where $Z_h = \{(s, a, r, s', h)\} \sim \mathcal{D}_h$. Let $\mathcal{A}(Z) = f = (f_1, \cdots, f_H) \in \mathcal{F}$ be the output of some batch RL algorithm $\mathcal{A}$. Let

$$
\begin{aligned}
L(f, Z) &= \frac{1}{H}\sum_{h=1}^{H}\ell(f_h, Z_h) \\
&= \frac{1}{H}\sum_{h=1}^{H}\frac{1}{n}\sum_{(s,a,r,s',h)\in Z_h}(f_h(s, a) - r - V_{f_{h+1}}(s'))^2.
\end{aligned}
$$

Since $\ell(f, Z) \in [0, 4H^2]$ for every f, it is $2H^2$-sub-Gaussian. By Theorem C.3, we have

$$
|\ell(f_h, Z_h) - \mathbb{E}_{Z_h\sim\mathcal{D}_h}[\ell(f_h, Z_h)]| \le \varepsilon
$$

with probability at least $1 - \delta'$ for

$$
n \ge \frac{8H^4}{\varepsilon^2}\left(I_\alpha(f_h; Z_h) + \log 2 + \frac{\alpha}{\alpha - 1}\log\left(\frac{1}{\delta'}\right)\right).
$$

Since we have n samples at each $h \in [H]$, we require

$$n \geq \frac{8H^4}{\varepsilon^2}\left(\max_h I_\alpha(f_h; Z_h) + \log 2 + \frac{\alpha}{\alpha-1}\log\left(\frac{1}{\delta'}\right)\right).$$

The claim now follows by the union bound by setting $\delta' = \delta/H$.

□

I Proof of Theorem 3.5

Theorem I.1 (Theorem 3.5 restated)**.** *Let $U \in \{0,1\}^n$ be uniformly random. Given a dataset $Z \sim \mathcal{D}^{2n}$ consists of $2nH$ samples, for any batch RL algorithm $\mathcal{A}$ with output $\mathcal{A}(Z_U) = f = (f_1, \cdots, f_H) \in \mathcal{F}$, if*

$$n \geq \frac{8H^4}{\varepsilon^2}\left(\max_h I_\alpha^{f_h|Z_h}(f_h; U|Z_h) + \log 2 + \frac{\alpha}{\alpha-1}\log\left(\frac{H}{\delta}\right)\right).$$

then the generalization error for the mean squared empirical Bellman error (MSBE) loss is upper bounded by

$$|L(\mathcal{A}(Z_U), Z_U) - L(\mathcal{A}(Z_U), \mathcal{D})| \leq \varepsilon$$

with probability at least $1-\delta$.

Proof. By substituting Theorem C.3 with Theorem C.5 in the proof of Theorem 3.4, the proof is thereby obtained. □

J Proof of Theorem 4.2

Lemma J.1. *For discrete random variables X, Y and Z, we have $I(X;Y|Z) \leq \log|X|$.*

Proof. Denote $H(X \mid Z)$ the conditional entropy of X given Z.

$$\begin{aligned} I(X;Y|Z) &= H(X|Z) - H(X|Y,Z) \\ &\leq H(X|Z) && (H(X|Y,Z) \geq 0) \\ &= \mathbb{E}_z[H(X|Z=z)] \\ &\leq \mathbb{E}_z[\log|X|] \\ &= \log|X|. \end{aligned}$$

□

Theorem J.2 (Theorem 4.2 restated)**.** *Suppose the function class $\mathcal{F}$ has a covering number of $\mathcal{N}(\mathcal{F}, \varepsilon)$. Let $U \in \{0,1\}^n$ be uniformly random. Given a dataset Z consists of $2nH$ samples, for any batch RL algorithm $\mathcal{A}$ with output $\mathcal{A}(Z_U) = f = (f_1, \cdots, f_H) \in \mathcal{F}$, the expected generalization error for the mean squared empirical Bellman error (MSBE) loss is upper bounded by*

$$|\mathbb{E}_{Z\sim\mathcal{D}}[L(\mathcal{A}(Z_U), Z_U) - L(\mathcal{A}(Z_U), \mathcal{D})]| \leq \sqrt{\frac{2H^3\log(|\mathcal{N}(\mathcal{F},\varepsilon)|)}{n}} + 8\varepsilon H + 2\varepsilon^2.$$

Proof. Let $\tilde{Z}_h = (Z_h)_U$. We first define an oracle algorithm $\mathcal{A}^o$ capable of outputting a function $\mathcal{A}^o(Z_U) = f^* = (f_1^*, \ldots, f_H^*)$ such that

$$\rho(f, f^*) \leq \varepsilon.$$

Note that $\mathcal{A}^o$ is only used for theoretical analysis. Observe that

$$\begin{aligned}
L(\mathcal{A}(Z_U), Z_U) &= \frac{1}{H}\sum_{h=1}^{H}\frac{1}{n}\sum_{(s,a,r,s',h)\in\tilde{Z}_h}(f_h(s,a)-r-V_{f_{h+1}}(s'))^2 \\
&= \frac{1}{H}\sum_{h=1}^{H}\frac{1}{n}\sum_{(s,a,r,s',h)\in\tilde{Z}_h}(f_h(s,a)-f_h^*(s,a)+f_h^*(s,a)-r-V_{f_{h+1}}(s'))^2 \\
&= \varepsilon^2 + \frac{1}{H}\sum_{h=1}^{H}\frac{1}{n}\sum_{(s,a,r,s',h)\in\tilde{Z}_h}(f_h^*(s,a)-r-V_{f_{h+1}}(s'))^2 \\
&\quad + 2\varepsilon\frac{1}{H}\sum_{h=1}^{H}\frac{1}{n}\sum_{(s,a,r,s',h)\in\tilde{Z}_h}(f_h^*(s,a)-r-V_{f_{h+1}}(s')) \\
&\le \varepsilon^2 + L(\mathcal{A}^o(Z_U), Z_U) + 4\varepsilon H.
\end{aligned}$$

Thus,

$$L(\mathcal{A}(Z_U), Z_U) - L(\mathcal{A}^o(Z_U), Z_U) \le 4\varepsilon H + \varepsilon^2.$$

Bounding $|L(\mathcal{A}(Z_U), \mathcal{D}) - L(\mathcal{A}^o(Z_U), \mathcal{D})|$ is similar. Now we have

$$\begin{aligned}
L(\mathcal{A}(Z_U), Z_U) - L(\mathcal{A}(Z_U), \mathcal{D}) = {} & L(\mathcal{A}(Z_U), Z_U) - L(\mathcal{A}^o(Z_U), Z_U) + L(\mathcal{A}^o(Z_U), Z_U) \\
& - L(\mathcal{A}^o(Z_U), \mathcal{D}) + L(\mathcal{A}^o(Z_U), \mathcal{D}) - L(\mathcal{A}(Z_U), \mathcal{D}).
\end{aligned}$$

Since $|L(\mathcal{A}(Z_U), Z_U) - L(\mathcal{A}^o(Z_U), Z_U)| \le \varepsilon$ and $|L(\mathcal{A}(Z_U), \mathcal{D}) - L(\mathcal{A}^o(Z_U), \mathcal{D})| \le \varepsilon$, we have

$$L(\mathcal{A}(Z_U), Z_U) - L(\mathcal{A}(Z_U), \mathcal{D}) \le L(\mathcal{A}^o(Z_U), Z_U) - L(\mathcal{A}^o(Z_U), \mathcal{D}) + 8\varepsilon H + 2\varepsilon^2.$$

By Theorem 3.3,

$$\begin{aligned}
|\mathbb{E}_{Z\sim\mathcal{D}}[L(\mathcal{A}^o(Z_U), Z_U) - L(\mathcal{A}^o(Z_U), \mathcal{D})]| &\le \sqrt{\frac{2H^2\sum_{h=1}^{H} I(f_h^*; U|Z_h)}{n}} \\
&\le \sqrt{\frac{2H^2\sum_{h=1}^{H}\log(|\mathcal{F}_\varepsilon|)}{n}} \qquad \text{(Lemma J.1)} \\
&= \sqrt{\frac{2H^3\log(|\mathcal{F}_\varepsilon|)}{n}} \\
&= \sqrt{\frac{2H^3\log(|\mathcal{N}(\mathcal{F},\varepsilon)|)}{n}}.
\end{aligned}$$

Therefore,

$$|\mathbb{E}_{Z\sim\mathcal{D}}[L(\mathcal{A}(Z_U), Z_U) - L(\mathcal{A}(Z_U), \mathcal{D})]| \le \sqrt{\frac{2H^3\log(|\mathcal{N}(\mathcal{F},\varepsilon)|)}{n}} + 8\varepsilon H + 2\varepsilon^2.$$

□