entropy

MDPI

Article

Information-Theoretic Generalization Bounds for Batch Reinforcement Learning

Xingtu Liu

School of Computing Science, Simon Fraser University, 8888 University Dr W, Burnaby, BC V5A 1S6, Canada; xingtu_liu@sfu.ca

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 and may be of independent interest.

Keywords: reinforcement learning; learning theory; generalization; mutual information

check for updates

Citation: Liu, X. Information-Theoretic Generalization Bounds for Batch Reinforcement Learning. *Entropy* **2024**, *26*, 995. https://doi.org/10.3390/e26110995

Academic Editor: Andrea Murari

Received: 27 September 2024
Revised: 12 November 2024
Accepted: 16 November 2024
Published: 18 November 2024

1. Introduction

Generalization is a fundamental concept in statistical machine learning. It measures how well a 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 [1,2], 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 state space $\mathcal{S}$, a finite action space $\mathcal{A}$, transition functions $\mathcal{P}_h : \mathcal{S}\times\mathcal{A}\rightarrow\Delta(\mathcal{S})$ at step $h\in[H]$, reward function $r_h : \mathcal{S}\times\mathcal{A}\rightarrow\mathbb{R}$ at step h, and H is the number of steps in each episode. We assume the reward is bounded, i.e., $r_h(s,a)\in[0,1]$

(For rewards in $[R_{\min}, R_{\max}]$ simply rescales these bounds.), $\forall(s,a,h)$. See Figure 1 for a graphical illustration.

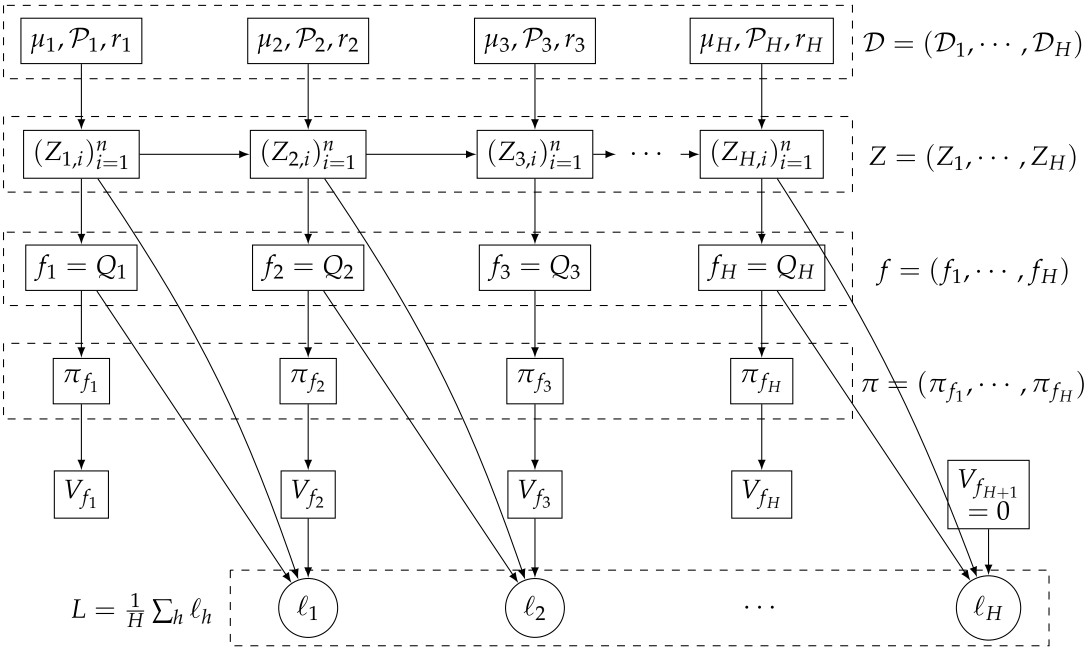

Figure 1. Directed graph representing the training process in Batch RL under episodic MDP.

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 following 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 $f = (f_1, \cdots, f_H) \in \mathcal{F}$. As no reward is collected in the $(H+1)$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 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 (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 1 (Bellman error to value suboptimality [3])**.** *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*

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

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 [3–5], and define the unbiased MSBE as:

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

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

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

Definition 4 (High-probability generalization bounds)**.** *Given a dataset* $Z \sim \mathcal{D}$*, and an algorithm* $\mathcal{A}$*, let* $L(\mathcal{A}(Z), Z)$ *denote the training loss and let* $L(\mathcal{A}(Z), \mathcal{D})$ *denote the true loss. Given a failure probability* δ *and an error tolerance* η*, the high-probability generalization error is defined as*

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

2.3. Mutual Information

First, we define the KL-divergence of two distributions.

Definition 5 (KL-Divergence [6])**.** *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 6 ([6])**.** *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 [7–10].

Definition 7 (Rényi's α-Divergence [11])**.** *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 8 (α-mutual information [7])**.** *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 follows:*

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

Definition 9 (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 follows:*

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

3. 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 with traditional methods, particularly in cases where those traditional measures are difficult to compute. See Appendix A for related work.

Theorem 1 ([2])**.** *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 1 ([7])**.** *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 2 ([11])**.** *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

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

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.

Definition 10. *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 ([12])**.** *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., $I(X; Y|Z) := \mathbb{E}_Z[D_{\text{KL}}(\mathcal{P}_{X,Y|Z} \| \mathcal{P}_{X|Z} \otimes \mathcal{P}_{Y|Z})]$, the above theorem does not provide a high-probability bound. Modifying this framework to ensure high-probability guarantees was left as future work in [12]. In the following, we use conditional α-mutual information to address this issue.

Theorem 4. *Let $U \in \{0,1\}^n$ be uniformly random. Given a dataset $Z \sim \mathcal{D}^{2n}$ consists of $2nH$ 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

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

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 $\mathcal{P}_{XYZ} \ll \mathcal{P}_Z \mathcal{Q}_{X|Z} \mathcal{P}_{Y|Z}$. 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 $\alpha \geq 1$,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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, using Hoeffding's inequality, for every X and Z,

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

Therefore, from Equations (2) and (3),

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

Lastly, by setting

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

we obtain the desired conclusion. □

4. 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. As 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 5. *Given that 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*

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

Proof. We first recall the Donsker—Varadhan variational representation ([13]) 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

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

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

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

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,

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

Finally, we observe that

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

□

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 the 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 [12]. 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.

Theorem 6. *Let $U \in \{0,1\}^n$ be uniformly random. Given that 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 $Z = Z_1 \cup \cdots \cup Z_H$ 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 5. We have

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

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

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

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,

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

Finally, we conclude that

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

□

Note that our setting is identical to that in [3], 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 7. *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*

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

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

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

with a probability of at least $1-\delta$.

Proof. Let $Z = Z_1 \cup \cdots \cup Z_H$ 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

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

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

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

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

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

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

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

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

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 7.

Theorem 8. *Let $U \in \{0, 1\}^n$ be uniformly random. Given that 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*

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

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

$$|L(\mathcal{A}(Z_U), Z_U) - L(\mathcal{A}(Z_U), \mathcal{D})| \le \epsilon$$

with probability at least $1 - \delta$.

Proof. By substituting Theorem 2 with Theorem 4 in the proof of Theorem 7, the proof is thereby obtained. □

5. 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 are not limited to, Bellman rank [14], Witness rank [15], and Eluder dimension [16]. 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 11 (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 $\mathcal{N}(\mathcal{F},\epsilon)$, is the minimum integer n, such that there exists a subset $\mathcal{F}_\epsilon \subseteq \mathcal{F}$ with $|\mathcal{F}_\epsilon| = n$, and for any $f \in \mathcal{F}$, there exists $g \in \mathcal{F}_\epsilon$, such that $\rho(x,y) \le \epsilon$.*

Lemma 2. *For discrete random variables X, Y, and Z, we have $I(X;Y|Z) \le \log|X|$.*

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

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

□

Theorem 9. *Suppose the function class $\mathcal{F}$ has a covering number of $\mathcal{N}(\mathcal{F},\epsilon)$. Let $U \in \{0,1\}^n$ be uniformly random. Given that 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*

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

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

$$\rho(f, f^*) \le \epsilon.$$

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

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

Thus,

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

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

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

As $|L(\mathcal{A}(Z_U), Z_U) - L(\mathcal{A}^o(Z_U), Z_U)| \leq \epsilon$ and $|L(\mathcal{A}(Z_U), \mathcal{D}) - L(\mathcal{A}^o(Z_U), \mathcal{D})| \leq \epsilon$, we have

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

By Theorem 6,

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

Therefore,

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

□

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.

Definition 12 (VC-Dimension [17])**.** *Given hypothesis class $\mathcal{H} \subseteq \mathcal{X} \to \{0,1\}$, its VC-dimension* 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 13 (Pseudo dimension [18])**.** *Suppose $\mathcal{X}$ is a feature space. Given hypothesis class $\mathcal{H} \subseteq \mathcal{X} \to \mathbb{R}$, its pseudo dimension* Pdim$(\mathcal{H})$ *is defined as* Pdim$(\mathcal{H}) =$ 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 3 (Bounding covering number by pseudo dimension [19])**.** *Given hypothesis class $\mathcal{H} \subseteq \mathcal{X} \to \mathbb{R}$ with* Pdim$(\mathcal{H}) \leq d$*, we have*

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

Corollary 1. *Suppose the function class $\mathcal{F}_h \subset \mathcal{F}$ has a finite pseudo dimension* 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. As Pdim$(\mathcal{F}_h) = d$ and $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_H$, we have $\log \mathcal{N}(\mathcal{F}, \epsilon) \leq O(dH \log(1/\epsilon))$. The claim follows from Theorem 9 by setting $\epsilon = H\sqrt{\frac{d}{n}}$. □

A prior study on finite sample guarantees for minimizing the Bellman error, using pseudo-dimension, demonstrated a sample complexity with a dependence of $\tilde{O}(d^2)$ [5]. In contrast, our sample complexity exhibits a dependence of $\tilde{O}(d)$ on the pseudo-dimension.

Now, we introduce another complexity measure known as the effective dimension [20], 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 14 (ϵ-effective dimension of a set [20]). *The ϵ-effective dimension of a set $\mathcal{X}$ is the minimum integer $d_{\mathit{eff}}(\mathcal{X}, \epsilon) = n$, such that*

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

Definition 15 (ϵ-effective dimension of a function class [20]). *Given a function class $\mathcal{F}$ defined on $\mathcal{X}$, its ϵ-effective dimension $d_{\mathit{eff}}(\mathcal{F}, \epsilon) = n$ 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 $x \in \mathcal{X}$,*
- *$d_{\mathit{eff}}(\phi(\mathcal{X}), \epsilon) = n$, where $\phi(\mathcal{X}) = \{\phi(x) : x \in \mathcal{X}\}$.*

Definition 16 (Kernel MDPs [21]). *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.*
- *$\|\sum_{s \in S} \mathcal{V}(s) \psi_h(s)\|_{\mathcal{H}} \leq 1$ for any function $\mathcal{V} : S \to [0, 1]$.*
- *$\dim_{\mathit{eff}}(X_h, \epsilon) \leq d$ 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 4 (Bounding covering number by effective dimension [21]). *Let $\mathcal{M}$ be a kernel MDP of effective dimension d, then*

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

Corollary 2. *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 $\tilde{O}(H^2\sqrt{d/n})$.*

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.

6. Discussion

In this paper, we analyzed the generalization property of batch reinforcement learning within the framework of information theory. We established 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 [22–28].

As in the supervised learning setting, where various algorithms such as Stochastic Gradient Descent (SGD) [29] and Stochastic Gradient Langevin Dynamics (SGLD) have been studied [30], 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 [12,31]. 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 [32–34]. Therefore, we hope that the information-theoretic approach will provide more insights into understanding the generalization of reinforcement learning.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: No data were created or analyzed in this theoretical study. Data sharing is not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Related Work

Appendix A.1. Batch Reinforcement Learning

A body of literature focuses on finite sample guarantees for batch reinforcement learning with function approximation [35–40]. Common assumptions in batch RL, such as concentrability, realizability, and completeness, have also been examined in more recent studies [41–43]. The most relevant work to ours [3] investigates the generalization performance of batch RL under the same setting using Rademacher complexities.

Appendix A.2. 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 [14], Witness rank [15], Eluder dimension [16], Bellman Eluder dimension [21], and more [20,37,44]. Identifying structural conditions and classifying RL problems clarifies the limits of what can be learned and guides the design of efficient algorithms.

Appendix A.3. Information-Theoretic Study of Generalization

The information-theoretic approach was initially introduced by [1,2] and subsequently refined to derive tighter bounds [45–47]. Besides, various other information-theoretic bounds have been proposed, leveraging concepts such as conditional mutual information [12], f-divergence [11], the Wasserstein distance [48,49], and more [50,51]. Some studies have focused on analyzing specific algorithms [29,30,52–55] while others have examined particular settings such as deep learning [56], iterative semi-supervised learning [57], transfer learning [58], and meta-learning [59,60]. There are also works attempting to provide a unified framework for generalization from an information-theoretic perspective [31,61,62].

References

1. Russo, D.; Zou, J. Controlling bias in adaptive data analysis using information theory. In Proceedings of the Artificial Intelligence and Statistics, Cadiz, Spain, 9–11 May 2016; pp. 1232–1240.
2. Xu, A.; Raginsky, M. Information-theoretic analysis of generalization capability of learning algorithms. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 2521–2530.
3. Duan, Y.; Jin, C.; Li, Z. Risk bounds and rademacher complexity in batch reinforcement learning. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 2892–2902.
4. Sutton, R.S.; Barto, A.G. Reinforcement learning: An introduction. *Robotica* **1999**, *17*, 229–235. [CrossRef]
5. Antos, A.; Szepesvári, C.; Munos, R. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Mach. Learn.* **2008**, *71*, 89–129. [CrossRef]
6. Cover, T.M. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 1999.
7. Verdú, S. α-mutual information. In Proceedings of the 2015 Information Theory and Applications Workshop (ITA), San Diego, CA, USA, 1–6 February 2015; pp. 1–6.
8. Van Erven, T.; Harremos, P. Rényi divergence and Kullback-Leibler divergence. *IEEE Trans. Inf. Theory* **2014**, *60*, 3797–3820. [CrossRef]
9. Csiszár, I. Generalized cutoff rates and Rényi's information measures. *IEEE Trans. Inf. Theory* **1995**, *41*, 26–34. [CrossRef]
10. Mironov, I. Rényi differential privacy. In Proceedings of the 2017 IEEE 30th Computer Security Foundations Symposium (CSF), Santa Barbara, CA, USA, 21–25 August 2017; pp. 263–275.
11. Esposito, A.R.; Gastpar, M.; Issa, I. Generalization error bounds via Rényi-, f-divergences and maximal leakage. *IEEE Trans. Inf. Theory* **2021**, *67*, 4986–5004. [CrossRef]
12. Steinke, T.; Zakynthinou, L. Reasoning about generalization via conditional mutual information. In Proceedings of the Conference on Learning Theory, Graz, Austria, 9–12 July 2020; pp. 3437–3452.
13. Boucheron, S.; Lugosi, G.; Massart, P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*; Oxford University Press: Oxford, UK, 2013.
14. Jiang, N.; Krishnamurthy, A.; Agarwal, A.; Langford, J.; Schapire, R.E. Contextual decision processes with low bellman rank are pac-learnable. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 1704–1713.
15. Sun, W.; Jiang, N.; Krishnamurthy, A.; Agarwal, A.; Langford, J. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In Proceedings of the Conference on Learning Theory, Phoenix, AZ, USA, 25–28 June 2019; pp. 2898–2933.
16. Wang, R.; Salakhutdinov, R.R.; Yang, L. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6123–6135.
17. Shalev-Shwartz, S.; Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*; Cambridge University Press: Cambridge, UK, 2014.
18. Haussler, D. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Inf. Comput.* **1992**, *100*, 78–150. [CrossRef]
19. Haussler, D. Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *J. Comb. Theory Ser. A* **1995**, *69*, 217–232. [CrossRef]
20. Du, S.; Kakade, S.; Lee, J.; Lovett, S.; Mahajan, G.; Sun, W.; Wang, R. Bilinear classes: A structural framework for provable generalization in rl. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 2826–2836.
21. Jin, C.; Liu, Q.; Miryoosefi, S. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 13406–13418.
22. Wei, R.; Lambert, N.; McDonald, A.; Garcia, A.; Calandra, R. A Unified View on Solving Objective Mismatch in Model-Based Reinforcement Learning. *arXiv* **2023**, arXiv:2310.06253.
23. Farahmand, A.M.; Barreto, A.; Nikovski, D. Value-aware loss function for model-based reinforcement learning. In Proceedings of the Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 20–22 April 2017; pp. 1486–1494.
24. Farahmand, A.M. Iterative value-aware model learning. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 9090–9101.
25. Abachi, R. *Policy-Aware Model Learning for Policy Gradient Methods*; University of Toronto (Canada): Toronto, ON, Canada, 2020.
26. Janner, M.; Fu, J.; Zhang, M.; Levine, S. When to trust your model: Model-based policy optimization. *Adv. Neural Inf. Process. Syst.* **2018**, *32*, 12519–12530.
27. Ji, T.; Luo, Y.; Sun, F.; Jing, M.; He, F.; Huang, W. When to update your model: Constrained model-based reinforcement learning. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 23150–23163.
28. Wang, X.; Zheng, R.; Sun, Y.; Jia, R.; Wongkamjan, W.; Xu, H.; Huang, F. Coplanner: Plan to roll out conservatively but to explore optimistically for model-based RL. *arXiv* **2023**, arXiv:2310.07220.

29. Neu, G.; Dziugaite, G.K.; Haghifam, M.; Roy, D.M. Information-theoretic generalization bounds for stochastic gradient descent. In Proceedings of the Conference on Learning Theory, Boulder, CO, USA, 15–19 August 2021; pp. 3526–3545.
30. Negrea, J.; Haghifam, M.; Dziugaite, G.K.; Khisti, A.; Roy, D.M. Information-theoretic generalization bounds for SGLD via data-dependent estimates. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 11015–11025.
31. Haghifam, M.; Dziugaite, G.K.; Moran, S.; Roy, D. Towards a unified information-theoretic framework for generalization. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 26370–26381.
32. Du, S.S.; Kakade, S.M.; Wang, R.; Yang, L.F. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv* **2019**, arXiv:1910.03016.
33. Weisz, G.; Amortila, P.; Szepesvári, C. Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. In Proceedings of the Algorithmic Learning Theory, Virtual, 16–19 March 2021; pp. 1237–1264.
34. Wang, Y.; Wang, R.; Kakade, S. An exponential lower bound for linearly realizable mdp with constant suboptimality gap. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 9521–9533.
35. Munos, R.; Szepesvári, C. Finite-Time Bounds for Fitted Value Iteration. *J. Mach. Learn. Res.* **2008**, *9* , 815–857.
36. Farahmand, A.; Ghavamzadeh, M.; Mannor, S.; Szepesvári, C. Regularized policy iteration. *Adv. Neural Inf. Process. Syst.* **2008**, *21*, 441–448.
37. Zanette, A.; Lazaric, A.; Kochenderfer, M.; Brunskill, E. Learning near optimal policies with low inherent bellman error. In Proceedings of the International Conference on Machine Learning, Virtual, 13–18 July 2020; pp. 10978–10989.
38. Lazaric, A.; Ghavamzadeh, M.; Munos, R. Finite-sample analysis of least-squares policy iteration. *J. Mach. Learn. Res.* **2012**, *13*, 3041–3074.
39. Farahm, A.M.; Ghavamzadeh, M.; Szepesvári, C.; Mannor, S. Regularized policy iteration with nonparametric function spaces. *J. Mach. Learn. Res.* **2016**, *17*, 1–66.
40. Le, H.; Voloshin, C.; Yue, Y. Batch policy learning under constraints. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 3703–3712.
41. Chen, J.; Jiang, N. Information-theoretic considerations in batch reinforcement learning. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 1042–1051.
42. Wang, R.; Foster, D.P.; Kakade, S.M. What are the statistical limits of offline RL with linear function approximation? *arXiv* **2020**, arXiv:2010.11895.
43. Xie, T.; Jiang, N. Batch value-function approximation with only realizability. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 11404–11413.
44. Foster, D.J.; Kakade, S.M.; Qian, J.; Rakhlin, A. The statistical complexity of interactive decision making. *arXiv* **2021**, arXiv:2112.13487.
45. Asadi, A.; Abbe, E.; Verdú, S. Chaining mutual information and tightening generalization bounds. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 7245–7254.
46. Hafez-Kolahi, H.; Golgooni, Z.; Kasaei, S.; Soleymani, M. Conditioning and processing: Techniques to improve information-theoretic generalization bounds. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 16457–16467.
47. Bu, Y.; Zou, S.; Veeravalli, V.V. Tightening mutual information-based bounds on generalization error. *IEEE J. Sel. Areas Inf. Theory* **2020**, *1*, 121–130. [CrossRef]
48. Lopez, A.T.; Jog, V. Generalization error bounds using Wasserstein distances. In Proceedings of the 2018 IEEE Information Theory Workshop (ITW), Guangzhou, China, 25–29 November 2018; pp. 1–5.
49. Wang, H.; Diaz, M.; Santos Filho, J.C.S.; Calmon, F.P. An information-theoretic view of generalization via Wasserstein distance. In Proceedings of the 2019 IEEE International Symposium on Information Theory (ISIT), Paris, France, 7–12 July 2019; pp. 577–581.
50. Aminian, G.; Toni, L.; Rodrigues, M.R. Information-theoretic bounds on the moments of the generalization error of learning algorithms. In Proceedings of the 2021 IEEE International Symposium on Information Theory (ISIT), Virtual, 12–20 July 2021; pp. 682–687.
51. Aminian, G.; Masiha, S.; Toni, L.; Rodrigues, M.R. Learning algorithm generalization error bounds via auxiliary distributions. *IEEE J. Sel. Areas Inf. Theory* **2024**, *5*, 273–284. [CrossRef]
52. Pensia, A.; Jog, V.; Loh, P.L. Generalization error bounds for noisy, iterative algorithms. In Proceedings of the 2018 IEEE International Symposium on Information Theory (ISIT), Vail, CO, USA, 17–22 June 2018; pp. 546–550.
53. Haghifam, M.; Negrea, J.; Khisti, A.; Roy, D.M.; Dziugaite, G.K. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9925–9935.
54. Harutyunyan, H.; Raginsky, M.; Ver Steeg, G.; Galstyan, A. Information-theoretic generalization bounds for black-box learning algorithms. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 24670–24682.
55. Wang, H.; Gao, R.; Calmon, F.P. Generalization bounds for noisy iterative algorithms using properties of additive noise channels. *J. Mach. Learn. Res.* **2023**, *24*, 1–43.
56. He, H.; Yu, C.L.; Goldfeld, Z. Information-Theoretic Generalization Bounds for Deep Neural Networks. *arXiv* **2024**, arXiv:2404.03176.
57. He, H.; Hanshu, Y.; Tan, V. Information-theoretic generalization bounds for iterative semi-supervised learning. In Proceedings of the The Tenth International Conference on Learning Representations, Virtual, 25 April 2022.
58. Wu, X.; Manton, J.H.; Aickelin, U.; Zhu, J. On the generalization for transfer learning: An information-theoretic analysis. *IEEE Trans. Inf. Theory* **2024**, *70*, 7089–7124. [CrossRef]

59. Jose, S.T.; Simeone, O. Information-theoretic generalization bounds for meta-learning and applications. *Entropy* **2021**, *23*, 126. [CrossRef]
60. Chen, Q.; Shui, C.; Marchand, M. Generalization bounds for meta-learning: An information-theoretic analysis. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 25878–25890.
61. Chu, Y.; Raginsky, M. A unified framework for information-theoretic generalization bounds. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 79260–79278.
62. Alabdulmohsin, I. Towards a unified theory of learning and information. *Entropy* **2020**, *22*, 438. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.