When is Realizability Sufficient for Off-Policy Reinforcement Learning?

Andrea Zanette [1]

Abstract

Understanding when reinforcement learning algorithms can make successful off-policy predictions—and when the may fail to do so– remains an open problem. Typically, model-free algorithms for reinforcement learning are analyzed under a condition called Bellman completeness when they operate off-policy with function approximation, unless additional conditions are met. However, Bellman completeness is a requirement that is much stronger than realizability and that is deemed to be too strong to hold in practice. In this work, we relax this structural assumption and analyze the statistical complexity of off-policy reinforcement learning when only realizability holds for the prescribed function class.

We establish finite-sample guarantees for off-policy reinforcement learning that are free of the approximation error term known as inherent Bellman error, and that depend on the interplay of three factors. The first two are well known: they are the metric entropy of the function class and the concentrability coefficient that represents the cost of learning off-policy. The third factor is new, and it measures the violation of Bellman completeness, namely the mis-alignment between the chosen function class and its image through the Bellman operator. Our analysis directly applies to the solution found by temporal difference algorithms when they converge.

1. Introduction

Markov decision processes (MDP) (Puterman, 1994; Bertsekas, 1995b;a) provide a general framework for reinforcement learning (RL) (Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 2018), which is a general paradigm for prediction and decision making under uncertainty. Modern RL algorithms typically solve sequences of sub-problems that require estimating the value of a policy different from the one that generated the dataset, a task broadly called *off-policy* reinforcement learning. Moreover, function approximations are typically implemented to deal with large state-action spaces.

Various off-policy methods have been proposed, such as importance sampling (Precup, 2000; Thomas & Brunskill, 2016; Jiang & Li, 2016) and weight learning (Uehara et al., 2020; Jiang & Huang, 2020; Zanette & Wainwright, 2022). Nonetheless, methods based on controlling the temporal difference error, such as fitted Q iteration (Ernst et al., 2005; Munos & Szepesvári, 2008), TD (Sutton, 1988), and their variants such as Q-learning (Watkins & Dayan, 1992), remain widely used especially with deep function approximation (Tesauro et al., 1995; Mnih et al., 2013; 2015; 2016; Fujimoto et al., 2018). We collectively refer to these algorithms as *temporal difference (TD) methods*.

Bellman completeness: a fundamental RL notion When the state-action space is large, TD methods are implemented with a function approximation class for the action value function. Their existing analyses (Munos & Szepesvári, 2008; Chen & Jiang, 2019; Duan & Wang, 2020; Fan et al., 2020) rely on a fundamental reinforcement learning notion known as *Bellman completeness*, which must hold for these algorithms to succeed. Completeness requires the chosen approximation space to fully capture each Bellman backup, see Figure 1a. However, such requirement is deemed too strict to hold in practice. What is more, even related algorithms that are theoretically more robust than TD and fitted Q, such as the minimax variant (Antos et al., 2008), also rely on Bellman completeness to properly function without approximation error.

This led researchers to investigate fundamental limits (Chen & Jiang, 2019; Zanette, 2020; Wang et al., 2020; Weisz et al., 2020; Wang et al., 2021; Foster et al., 2021). Recently, (Foster et al., 2021) discovered that completeness is crucial in an information-theoretic sense: even with seemingly benign distribution shifts, exponential lower bounds quickly arise in the absence of Bellman completeness.

Unfortunately, completeness is a very hard condition to meet. For example, when realizability is violated, the

[1] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, United States of America. Correspondence to: Andrea Zanette <zanette@berkeley.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

predictor class can be expanded so as to reduce the approximation error, a balancing act known as bias variance trade-off (Shalev-Shwartz & Ben-David, 2014). On the contrary, when Bellman completeness is not satisfied, enlarging the prescribed function class may make completeness even more violated, because the Bellman backup of this new and bigger function class must now be correctly represented.

In summary, while realizability is also needed to make good predictions in Statistics, Bellman completeness seems like an additional requirement specific to reinforcement learning, one that is intuitively very restrictive and undesirable, and unlikely to hold in practice, but seemingly necessary.

Contribution In this work we analyze the statistical complexity of off-policy reinforcement learning in settings where only realizability is assumed, and bridge the gap between the Bellman complete case and the known exponential lower bounds that arise when Bellman completeness is "extremely" violated. In order to characterize this intermediate regime, we introduce the concept of local inherent Bellman errors to measure the local violation of Bellman completeness. We then establish off-policy error bounds for the solution found by the minimax reinforcement learning formulation (Antos et al., 2008), first with function classes of finite-cardinality and then with more general, nonparametric ones.

Our error bounds depend on three critical factors: 1) the metric entropy of the chosen function class, 2) a certain amplifying factor, called concentrability coefficient, that arises due to the distribution shift, and 3) a new amplifying factor that represents the mis-alignment between the prescribed function class and its image through the Bellman operator. Furthermore, these error bounds apply to the widely used iterative TD methods when and if they do converge.

The main improvement compared to prior analyses is that *the violation of Bellman completeness is expressed as an amplifying factor that affects the sample complexity, instead of as an approximation error term known as inherent Bellman error*. The improvement arises from the application of a localization argument to measure the violation of Bellman completeness. Effectively, this **removes the assumption of Bellman completeness for off-policy evaluation**: instead, the lack of completeness is measured by a certain coefficient —like the metric entropy measures the function capacity and the concentrability measures the distribution shift—that can in principle be computed. We expect the insights of this paper to apply more broadly to other settings such as policy optimization or exploration.

Bellman complete models require all Bellman backups to be contained in the prescribed function class. In contrast, in our work the two are allowed to be only partially aligned. It follows that the decision processes that can be studied with our framework are far richer and more realistic than those that are Bellman complete, because the image of the prescribed function class through the Bellman operator can have a complex, truly high-dimensional structure.

Most literature is discussed in Section 4.3.

2. Preliminaries

Here we recall the basic definitions; some additional background material can be found in Appendix B.

2.1. Notation and Set-up

We focus on infinite-horizon discounted Markov decision processes (Puterman, 1994; Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 2018) with discount factor $\gamma \in [0,1)$, state space $\mathcal{S}$, and an action set $\mathcal{A}$. For each state-action pair (s,a), there is a reward distribution $R(s,a)$ over $[0,1]$ with mean $r(s,a)$, and a transition function $\mathcal{P}(\cdot \mid s,a)$.

A (stationary) target policy π maps states to actions. Its action value function is denoted with $f^\star$. It is defined as the discounted sum of future rewards based on starting from the pair (s,a), and then following the policy π in all future time steps $f^\star(s,a) = r(s,a) + \sum_{h=1}^{\infty} \gamma^h \mathbb{E}[r_h(S_h,A_h) \mid (S_0,A_0) = (s,a)]$, where the expectation is taken over trajectories with $A_h \sim \pi(\cdot \mid S_h)$, and $S_{h+1} \sim \mathcal{P}(\cdot \mid S_h, A_h)$ for $h = 1,2,\ldots$. We also use $f(s,\pi) = \mathbb{E}_{A\sim\pi(\cdot|s)} f(s,A)$ and define the *Bellman evaluation operator* and its empirical counterpart using the observed reward r and successor state s^+ as

$$(\mathcal{T}f)(s,a) = r(s,a) + \gamma \mathbb{E}_{S^+\sim\mathcal{P}(\cdot|s,a)} f(S^+,\pi),$$
$$(\mathrm{T}f)(r,s^+) = r + \gamma f(s^+,\pi).$$

The key property needed in our theorems is that T is a bounded operator. The discounted occupancy measure of the target policy π is given by $d^\pi(s,a) = (1-\gamma)\sum_{h=0}^{\infty} \gamma^h \mathbb{P}_h[(S_h,A_h) = (s,a)]$, where $\mathbb{P}_h$ is the probability of encountering a certain state-action pair when following π from a given initial state.

We are interested in the *prediction error* from a certain initial state s_0, which will be omitted later for brevity

$$\mathcal{E}(f) = (f^\star - f)(s_0,\pi).$$

Throughout the paper we assume that the learner has access to an action value function class $\mathcal{F}$ that contains the correct predictor.

Assumption 1 (Realizability)**.** $f^\star \in \mathcal{F}$.

Learning from a dataset We assume we have access to a dataset $\mathcal{D} = \{(s_i,a_i,r_i,s_i^+)\}_{i=1,\ldots,n}$ that contains n tuples. Each tuple contains a state s, an action a, a reward r and

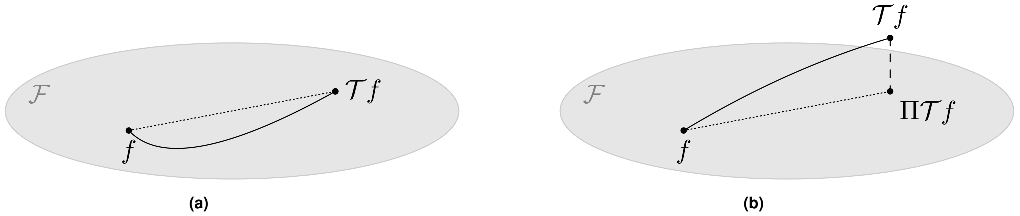


Figure 1: Bellman completeness (Figure 1a) puts strong restrictions on the Bellman operator $\mathcal{T}$, because the Bellman operator $\mathcal{T}$ must map the chosen function class $\mathcal{F}$ onto itself, i.e., $\mathcal{T}\mathcal{F} \subseteq \mathcal{F}$. Without Bellman completeness (Figure 1b), there is no restriction on $\mathcal{T}\mathcal{F}$, although its alignment with $\mathcal{F}$ does influence the statistical complexity of off-policy reinforcement learning.

a successor state s^+. In order to deal with the situation where the dataset is created using different policies, we assume that the states and actions are sampled from an underlying distribution μ. Conditioned on $(s, a) \sim \mu$, the reward and successor state in a certain tuple are sampled from the Markov reward process, i.e., $r \sim R(s, a)$ and $s^+ \sim \mathcal{P}(s, a)$. The associated expectation operator over (s, a, r, s^+) is often denoted with $\mathbb{P}$, while its empirical counterpart over $(s, a, r, s^+) \in \mathcal{D}$ is denoted with $\mathbb{P}_n$.

We commonly measure quantities using the norm induced by the distribution μ and the policy π. Let f be a function defined over the state-action space; they are defined as

$$\|f\|_\mu^2 = \mathbb{E}_{(s,a)\sim\mu}[f(s, a)]^2, \quad \|f\|_\pi^2 = \mathbb{E}_{(s,a)\sim d^\pi}[f(s, a)]^2.$$

Projections The projection operator Π onto $\mathcal{F}$ takes in a function h and finds a function $g \in \mathcal{F}$ closest to h

$$\Pi h = \arg\min_{g\in\mathcal{F}} \|g - h\|_\mu.$$

In most cases we deal with, the function to project is the Bellman backup $h = \mathcal{T}f$, and so it is convenient to denote the projected Bellman backup and the empirically projected backup with specific symbols, defined as

$$g_f = \arg\min_{g\in\mathcal{F}} \|g - \mathcal{T}f\|_\mu^2, \quad \text{and}$$
$$\widehat{g}_f = \arg\min_{g\in\mathcal{F}} \frac{1}{n} \sum_{(s,a,r,s^+)\in\mathcal{D}} \Big(g(s, a) - \mathrm{T}f(r, s^+)\Big)^2. \tag{1}$$

Fitted Q Fitted Q (Ernst et al., 2005) is a classical and well studied (Munos, 2005; Munos & Szepesvári, 2008; Chen & Jiang, 2019; Duan & Wang, 2020; Fan et al., 2020) off-policy prediction and optimization algorithm. In this paper we focus on the policy evaluation version of the algorithm, which starts from an initial iterate $f_0 \in \mathcal{F}$ and updates it iteratively by solving

$$f_{k+1} = \arg\min_{f\in\mathcal{F}} \frac{1}{n} \sum_{(s,a,r,s^+)\in\mathcal{D}} \Big(f(s, a) - r - \gamma f_k(s^+, \pi)\Big)^2.$$

We indicate with $\widehat{f}_{\mathrm{FQ}}$ the fixed point of fitted Q.

2.2. Minimax Formulation and Inherent Bellman Error

The fitted Q algorithm is related to the minimax formulation (Antos et al., 2008) in the sense that when fitted Q converges to a fixed point, such fixed point is a minimizer of the minimax formulation (Chen & Jiang, 2019).

Squared temporal difference cost Consider the following cost function, which is the squared temporal difference error of the tuple (s, a, r, s^+) evaluated using f as next-state value function and g as current function. It is defined as

$$L(g, f) = \Big(g(s, a) - r - \gamma\mathbb{E}_{a^+\sim\pi(s)} f(s^+, a^+)\Big)^2. \tag{2}$$

In order to find a predictor consistent with the dataset $\mathcal{D}$, one can try to minimize the empirical expectation of the above cost function with $g = f$, namely $\widehat{\mathcal{L}}(f, f)$ where[1]

$$\widehat{\mathcal{L}}(g, f) = \frac{1}{|\mathcal{D}|} \sum_{(s,a,r,s^+)\in\mathcal{D}} L(g, f).$$

Unfortunately, due to the double sampling issue (Baird, 1995; Sutton & Barto, 2018), its expectation contains the bias term $\sigma(f)^2$ (made explicit in Lemma 6, but the fact is well known) representing the variance of the backup

$$\mathbb{E}L(g, f) = \|g - \mathcal{T}f\|_\mu^2 + \sigma^2(f). \tag{3}$$

The variance term $\sigma(f)^2$ arises even when $g = f$ in the cost function. This implies that in the limit of infinite data the minimizer of $\mathbb{E}L(f, f)$ must trade-off minimizing the mean-squared Bellman error $\|f - \mathcal{T}f\|_\mu^2$ with minimizing the variance $\sigma(f)^2$ of the backup. The resulting procedure may converge to a solution different from the optimal predictor $f^\star$ even in the realizable setting.

A different cost function To remedy this issue[2], the following cost function was introduced in (Antos et al., 2008):

$$L(f, f) - L(g, f).$$

[1] It is useful to define the cost and its expectation by separating g and f in preparation for the discussion to follow.

[2] In practice iterative algorithms are used, but the algorithm studied here is closely related to the iterative TD algorithms.

Compared to the squared TD cost function in Equation (2), which would be minimized with $g = f$, the modified cost function contains the correction term $-L(g,f)$. The expectation of the correction term generates the conditional variance of the backup $\sigma(f)^2$ which then cancels the one present in $\mathbb{E}L(f,f)$. We have $\mathbb{E}[L(f,f) - L(g,f)] =$

$$\begin{aligned} &= \|f - \mathcal{T}f\|_\mu^2 + \sigma(f)^2 - \|g - \mathcal{T}f\|_\mu^2 - \sigma(f)^2 \\ &= \|f - \mathcal{T}f\|_\mu^2 - \|g - \mathcal{T}f\|_\mu^2. \end{aligned} \tag{4}$$

While the modified cost function is successful in cancelling the unwanted term $\sigma(f)^2$, it has introduced a different bias term represented by $\|g - \mathcal{T}f\|_\mu^2$. In order to keep this bias at a minimum, the function g should be selected so as to minimize it, ideally as

$$g_f = \min_{g\in\mathcal{F}} \|g - \mathcal{T}f\|_\mu^2.$$

The population-level loss to minimize is (Antos et al., 2008)

$$\mathcal{M}(f) = \|f - \mathcal{T}f\|_\mu^2 - \min_{g\in\mathcal{F}} \|g - \mathcal{T}f\|_\mu^2. \tag{5}$$

The resulting empirical program to minimize over f is

$$\widehat{\mathcal{M}}(f) = \widehat{\mathcal{L}}(f,f) - \min_{g\in\mathcal{F}} \widehat{\mathcal{L}}(g,f)$$

Its empirical minimizer $\widehat{f}$ is of interest to us:

$$\widehat{f} \in \arg\min_{f\in\mathcal{F}} \widehat{\mathcal{M}}(f).$$

The fact that the fitted Q fixed point minimizes $\widehat{\mathcal{M}}(f)$ (see e.g., (Chen & Jiang, 2019)) motivates the study of the minimax formulation.

Completeness removes the bias Despite the above effort to reduce the bias term, the term $\inf_{g\in\mathcal{F}} \|g - \mathcal{T}f\|_\mu^2$ still affects the estimation quality of the mean-squared Bellman error, and it is unclear whether that is better than $\sigma(f)^2$. A notable case where such correction is desirable is when the bias term $\min_{g\in\mathcal{F}} \|g - \mathcal{T}f\|_\mu^2$ is zero for all $f \in \mathcal{F}$, a condition called Bellman completeness. In this case, the population-level loss $\mathcal{M}(f)$ coincides with the mean-squared Bellman error, i.e., under Bellman completeness we have

$$\mathcal{M}(f) = \|f - \mathcal{T}f\|_\mu^2. \tag{6}$$

Therefore, minimizing $\mathcal{M}$ directly minimizes the mean-squared Bellman error.

Inherent Bellman errors When completeness starts to be violated, only part of $\mathcal{T}f$ is 'captured' by $\mathcal{F}$, and an angle between the two arises, see Figure 1b. Although in this cases the backup $\mathcal{T}f$ is not contained in $\mathcal{F}$, we can still consider its projection onto $\mathcal{F}$ defined in Equation (1). As the projection discards potentially useful informations about the backup $\mathcal{T}f$, we expect an error to arise. Such error is the component of the backup $\mathcal{T}f$ not captured by $\mathcal{F}$:

$$\inf_{g\in\mathcal{F}} \|g - \mathcal{T}f\|_\mu. \tag{7}$$

An algorithm like fitted Q typically considers different functions $f \in \mathcal{F}$ through its execution, and the projection error is propagated through the iterations. Moreover, such error term is present in the definition of the minimax program in Equation (5), and so its presence seems to be unavoidable. Generally, a *worst-case* analysis is adopted, and the worst-case value of the residual over $f \in \mathcal{F}$ is called inherent Bellman error of the function class $\mathcal{F}$

$$\mathcal{I}_\mathcal{F} = \sup_{f\in\mathcal{F}} \inf_{g\in\mathcal{F}} \|g - \mathcal{T}f\|_\mu. \tag{8}$$

(Other definitions based on different norms are possible). The inherent Bellman error is zero for the Bellman complete case in Figure 1a; the less the Bellman backup is aligned with $\mathcal{F}$ the bigger it becomes (cfr. Figure 1b).

3. Local Inherent Bellman Errors

In this section we introduce the core concept of this paper, namely the local inherent Bellman errors and the related notion of β-incompleteness; they are needed to convey the main message of the paper when Bellman completeness is violated. From a technical standpoint, our development is inspired by the localization argument of (Bartlett et al., 2005), which is a now a standard tool in statistics to obtain fast regression rates (Wainwright, 2019). Our use of localization, however, concerns a different quantity—the inherent Bellman error—and brings an even more consequential improvement, i.e., that of removing the approximation error term connected to the lack of Bellman completeness.

Some intuition is provided in Figure 2, while the definitions are motivated as follows. If Bellman completeness was satisfied then minimizing $\mathcal{M}$ would directly minimize the mean-squared Bellman error, see Equation (6). When completeness is violated, our hope is that the mean-squared Bellman error is still minimized by the minimax algorithm. In other words, we hope that $\widehat{f}$ enjoys small mean-squared Bellman error $\|\widehat{f} - \mathcal{T}\widehat{f}\|_\mu^2$. If that is the case, $\widehat{f}$ must belong to the set of predictors $\mathcal{F}(r)$ whose Bellman error is, say, at most r for some positive value r:

$$\mathcal{F}(r) = \{f \in \mathcal{F} \mid \|f - \mathcal{T}f\|_\mu \le r\}.$$

If $\widehat{f}$ is known to belong to the set $\mathcal{F}(r)$, the inherent Bellman error that should arise in a performance bound is one where the predictor f is restricted to the class $\mathcal{F}(r)$. The value of the inherent Bellman error constructed in this way as a function of r is what we call *incompleteness function*.

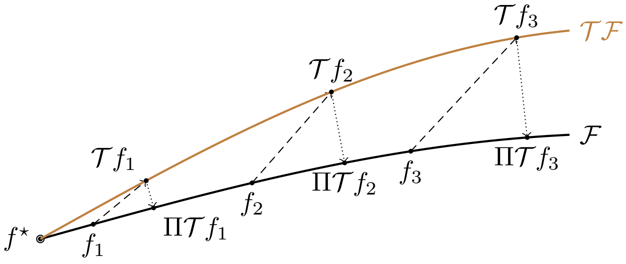


Figure 2: Local inherent Bellman errors. The norm of the un-captured component of the Bellman error $\Pi\mathcal{T}f - \mathcal{T}f$, when maximized over $f \in \mathcal{F}$, is the inherent Bellman error. For every function $f \in \mathcal{F}$, such un-captured component is always a fraction of the Bellman error $f - \mathcal{T}f$. When the Bellman error is reduced, its un-captured component also gets reduced. This means that the 'effective' inherent Bellman error seen by an algorithm decreases as the algorithm approaches the optimal predictor $f^\star$ along $\mathcal{F}$. In order to leverage this observation in the analysis, we *localize* the inherent Bellman error to a subset of functions where the empirical predictor $\widehat{f}$ returned by the minimax algorithm is expected to be. In this way, we can replace the inherent Bellman error in Equation (8), which is defined globally over $\mathcal{F}$, with a more localized version defined over a smaller class $\widetilde{\mathcal{F}} \subset \mathcal{F}$ that contains $\widehat{f}$.

Definition 1 (Incompleteness Function)**.** *The incompleteness function $\mathcal{I}$ (or localized inherent Bellman error) is the function*

$$\mathcal{I}(r) = \sup_{f\in\mathcal{F}(r)} \inf_{g\in\mathcal{F}} \|g - \mathcal{T}f\|_\mu.$$

In other words, the incompleteness function is the inherent Bellman error *localized* to the set of functions of small mean-squared Bellman error $\|f - \mathcal{T}f\|_\mu$. When $r \to \infty$, the localized inherent Bellman error recovers the inherent Bellman error, i.e., $\mathcal{I}(\infty) = \mathcal{I}_{\mathcal{F}}$. Notice that if the model is misspecified ($f^\star \notin \mathcal{F}$) then the set $\mathcal{F}(r)$ may be empty for small values of r, and so the incompleteness function is defined only up to a certain value of r.

To summarize, our expectation is that the empirical solution $\widehat{f}$ belongs to $\mathcal{F}(r)$ for an appropriate value of r. In that case, the inherent Bellman error 'felt' by the minimax algorithm should be $\mathcal{I}(r)$. When r decreases, the function $\mathcal{I}(r)$ should also decrease because it is an error associated to a smaller set. This intuition on the behavior of the local inherent Bellman errors is correct, and it is formalized by the following proposition, which is proved in Appendix C.1.

Proposition 1 (Behavior of Local Inherent Bellman Errors)**.** *The following holds true:*

- *$\mathcal{I}(r)$ is increasing with r;*
- *if realizability holds then $\mathcal{I}(0) = 0$.*

Figures 3a to 3c illustrate possible shapes for the incompleteness function in the realizable case, while Figure 3d shows one where realizability is violated (i.e., when $f^\star \notin \mathcal{F}$).

In the sequel we focus on the realizable case to make the analysis clearer, i.e., on function classes that satisfy Assumption 1. Although in this case the local inherent Bellman error always converges to zero, it might do so at different speeds. The average rate of convergence to zero is denoted with β and it determines the problem complexity.

3.1. β-incomplete MDPs

Let us gain some intuition by considering a linear problem, namely one where the function class $\mathcal{F}$ is linear. It is defined by a feature extractor ϕ that maps state-action pairs to real vectors in $\mathbb{R}^d$, as $\mathcal{F}_{\text{lin}} = \{\phi^\top w \mid w \in \mathbb{R}^d\}$.

When the class is linear and realizability holds, the localized inherent Bellman error $\mathcal{I}(\cdot)$ always increases at a linear rate, a fact that we verify in Appendix F.1.

Proposition 2 (Linearly Incomplete MDPs)**.** *If $\mathcal{F} = \mathcal{F}_{lin}$ then $\mathcal{I}(r) = \beta r$ for all $r \geq 0$.*

In this case, we say that the system is β-incomplete. When $\beta = 0$, the MDP is linear Bellman complete (Zanette et al., 2020; Duan & Wang, 2020) and that corresponds to the situation in Figure 3a. On the contrary, the higher β is, and the farther from $f^\star$ (i.e., the higher the radius r), the more Bellman completeness is violated, a situation in display in Figure 3b.

When $\mathcal{F}$ is non-linear we expect the local inherent Bellman error $\mathcal{I}$ to exhibit a more complex behavior. It must still comply with Proposition 1, namely it must start from zero and increase as the radius increases. In these cases it is a good idea to define a quantity to capture its global behavior. Such quantity should put a bound on the average rate of increase of $\mathcal{I}$, i.e., such that

$$\mathcal{I}(r) \leq \beta r. \tag{9}$$

With this goal in mind, we give the following definition for β, one that applies to the linear and the non-linear setting.

Definition 2 (β-incompleteness)**.** *The incompleteness factor β, or mis-alignment between $\mathcal{F}$ and its image $\mathcal{T}\mathcal{F}$, is the scalar quantity defined as*

$$\sup_{f\in\mathcal{F}} \inf_{g\in\mathcal{F}} \frac{\|g - \mathcal{T}f\|_\mu}{\|f - \mathcal{T}f\|_\mu} = \beta. \tag{10}$$

In other words, β represents the maximum *fraction* of the Bellman error $\|f - \mathcal{T}f\|_\mu$ that is not captured by $\mathcal{F}$. When Bellman completeness holds, $\inf_{g\in\mathcal{F}} \|g - \mathcal{T}f\|_\mu = 0$ for all $f \in \mathcal{F}$, and thus $\beta = 0$. In the worst case, g in the numerator in Definition 2 can at least be chosen equal to f, in which case we have $\beta = 1$. More generally, β is a number between zero and one. The closer it is to zero, the more Bellman complete the MDP is, in the sense that completeness gets violated more slowly when moving away

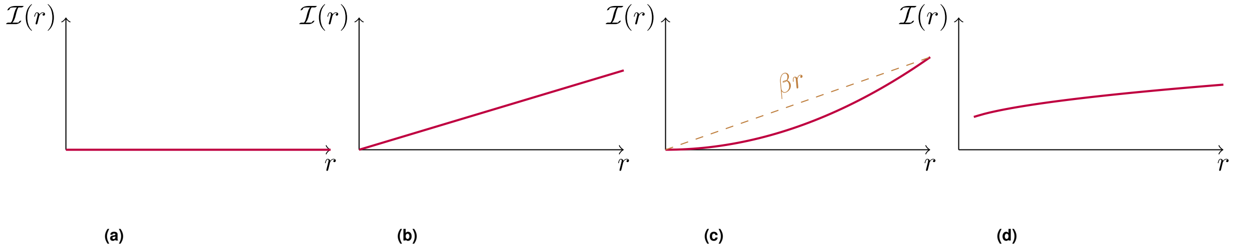


Figure 3: Stylized representations of possible shapes of $\mathcal{I}$

from $f^\star$. See Figure 3c for a visual definition of β. It can be shown that Definition 2 leads to the desired behavior in display in Equation (9), since $\mathcal{I}(r)/r$ can be written as

$$= \sup_{f \in \mathcal{F}(r)} \inf_{g \in \mathcal{F}} \frac{\|g - \mathcal{T}f\|_\mu}{r} \le \sup_{f \in \mathcal{F}(r)} \inf_{g \in \mathcal{F}} \frac{\|g - \mathcal{T}f\|_\mu}{\|f - \mathcal{T}f\|_\mu} \le \beta.$$

How is Definition 2 useful for prediction? Intuitively, the numerator $\inf_{g\in\mathcal{F}} \|g - \mathcal{T}f\|_\mu$ in Definition 2 represents some form of approximation error for the backup $\mathcal{T}f$; the division by the denominator scales such approximation error with respect to the mean-squared Bellman error, which is the quantity that we wish to reduce. When the latter is reduced, the approximation error is also reduced, and the Bellman backup is more faithfully represented. In other words, the approximation error must vanish as we approach $f^\star$.

Another possible connection is with the double-sampling issue (Baird, 1995). Although the mean-squared Bellman error cannot be accurately estimated without Bellman completeness (see e.g. (Duan et al., 2021) for a recent lower bound), β-incompleteness ensures that we can estimate it with a certain accuracy relative to its magnitude, and in particular, more accurately for the important functions that are closer to $f^\star$.

4. Error Bounds on Bellman-Incomplete MDPs

In this section we present our main results, which are off-policy error bounds on the prediction error $|\mathcal{E}(\widehat{f})|$ for the minimizer $\widehat{f}$ of the empirical loss $\widehat{\mathcal{M}}$. These error bounds apply to the limit point for fitted Q when it exists (Chen & Jiang, 2019).

Concentrability It is useful to introduce the following concentrability coefficient (Chen & Jiang, 2019; Xie et al., 2021), which represents the increase in the mean-squared Bellman error when moving from the data-generating distribution μ to that induced by the target policy π

$$C = \sup_{f \in \mathcal{F}} \frac{\|f - \mathcal{T}f\|_\pi^2}{\|f - \mathcal{T}f\|_\mu^2}.$$

As the proof shall clarify, the minimax procedure indirectly attempts to minimize the mean-squared Bellman error over μ (even though it cannot estimate it properly), while the prediction error is related to that over d_π. Therefore, the concentrability coefficient[3] translates how minimizing the mean-squared Bellman error over μ affects that over d^π, and hence the prediction error. The higher the value of C, the less effective the minimax algorithm is, because the value of the mean-squared Bellman error over μ is less representative of the prediction error.

4.1. Error bounds with finite classes

For simplicity, let us present the main findings first when the cardinality of $\mathcal{F}$ is finite.

Theorem 1 (Error Bound with Finite Classes)**.** *With probability at least $1-\delta$, the prediction error of the minimizer $\widehat{f}$ satisfies the bound*

$$|\mathcal{E}(\widehat{f})| \le \frac{1}{1-\gamma} \frac{1}{1-\beta} \sqrt{\frac{C \ln(|\mathcal{F}|/\delta)}{n}}. \tag{11}$$

The proof is in Appendix D. The bound above exhibits a typical dependence on several factors: the log failure probability $\ln(1/\delta)$, the square-root of the number of samples n, the effective horizon $\frac{1}{1-\gamma}$, the metric entropy $\ln(|\mathcal{F}|)$ and the concentrability factor C. However, the key novelty is the presence of the pre-factor $\frac{1}{1-\beta}$ that measures the lack of Bellman completeness, and the absence of the inherent Bellman error. Practically speaking, the form of the equation suggests that realizability is sufficient whenever 1) $\beta < 1$, and 2) the TD method converges. When $\beta = 1$, off-policy learning is unviable without additional 'domain knowledge' because the projected Bellman equations—which TD methods aim to solve—may have multiple solutions.

Compared to the state of the art (Chen & Jiang, 2019; Jin et al., 2021; Xie et al., 2021; Duan et al., 2021) analyses of

[3] Some weaker upper bounds, which have the advantage of being independent of $\mathcal{F}$, are the following:

$$C \le \mathbb{E}_{(s,a)\sim\mu}\Big[\frac{d^\pi(s,a)}{\mu(s,a)}\Big]^2 \le \sup_{(s,a)} \frac{d^\pi(s,a)}{\mu(s,a)}.$$

the minimax algorithm, the use of the local inherent Bellman errors has transformed the approximation error term $\mathcal{I}_{\mathcal{F}}$ into the pre-factor $\frac{1}{1-\beta}$ that multiplies the rate of convergence. In other words, Equation (11) establishes that the lack of Bellman completeness does not generate an approximation error—the inherent Bellman error—but instead it affects the rate of convergence.

The factor $\frac{1}{1-\beta}$ could also be interpreted as the cost, in terms of sample complexity, of moving from the double-sampling regime[4] to the single-sampling regime in off-policy reinforcement learning; the work of (Duan et al., 2021) can be used to compare our sample complexity with that of methods based on Bellman residual minimization in the double-sampling regime.

It is instructive to examine in more details the three key components that determine the sample complexity.

- The **metric entropy**, represented by $\ln(|\mathcal{F}|)$, arises already in supervised learning (Wainwright, 2019).
- The **distribution shift**, represented by the concentrability coefficient C, arises (as a simplified expression that does not depend on the Bellman operator) if distribution shift is present in supervised learning.
- The **incompleteness factor**, represented by $\frac{1}{1-\beta}$, measures the adequacy of the chosen function class with respect to the Bellman operator $\mathcal{T}$; this is the key factor that distinguishes the reinforcement learning setting from single-step processes, because it *involves the Bellman operator*. Notice that the notion of β-incompleteness is not an assumption: the value for β can always be computed, and its knowledge is not required by the algorithm. Much like the concentrability coefficient measures the degradation in performance as the target policy π visits different state-action pairs than the dataset distribution μ, the incompleteness factor β represents the loss of efficiency as the chosen function class becomes more and more mis-aligned with the Bellman backups.

Finally, it is worth to highlight the following fact (Chen & Jiang, 2019): if fitted Q converges, its limit point must inherit the bound of Theorem 1, and so our completeness-free result applies to the solution found by fitted Q.

Theorem 1 already contains the key innovation of this paper. However, the result only applies to finite classes, which are statistically simple but also unstructured: they are non-convex and non-differentiable and hence the above result cannot be applied to gradient-based methods such as TD. We deal with more expressive models in Appendix C.3, and make additional considerations in Appendix C.2.

4.2. Comparison with existing guarantees

In reinforcement learning analyses for model free algorithms, an approximation error term is present even if the problem is realizable, i.e., even if the action value function $f^\star$ of the target policy is contained in $\mathcal{F}$. Precisely, the approximation error term is the inherent Bellman error of the function class $\mathcal{F}$. A typical bound[5] (Munos & Szepesvári, 2008; Chen & Jiang, 2019) for the minimax variant reads

$$|V^\pi - \widehat{V}^\pi| \lessapprox \underbrace{\frac{1}{1-\gamma}\sqrt{\frac{C\ln(|\mathcal{F}|/\delta)}{n}}}_{\text{stat error}} + \underbrace{\frac{\sqrt{C}}{1-\gamma}\mathcal{I}_{\mathcal{F}}}_{\text{approx error}}. \tag{12}$$

According to Equation (12), the prediction error can be reduced only up to an error floor represented by the inherent Bellman error $\mathcal{I}_{\mathcal{F}}$ of the function class $\mathcal{F}$.

Figures 4a to 4d display some Bellman errors to help appreciate the results of this paper and the informal definition of β. When Bellman completeness holds such as in Figure 4a, the class $\mathcal{F}$ fully captures the Bellman backup and thus $\beta = 0$ (no component of the Bellman error is left un-captured). In this case, the existing bound in Equation (12) and the new one in Theorem 1 both reduce to $|V^\pi - \widehat{V}^\pi| \lessapprox \frac{1}{1-\gamma}\sqrt{\frac{C\ln(|\mathcal{F}|/\delta)}{n}}$.

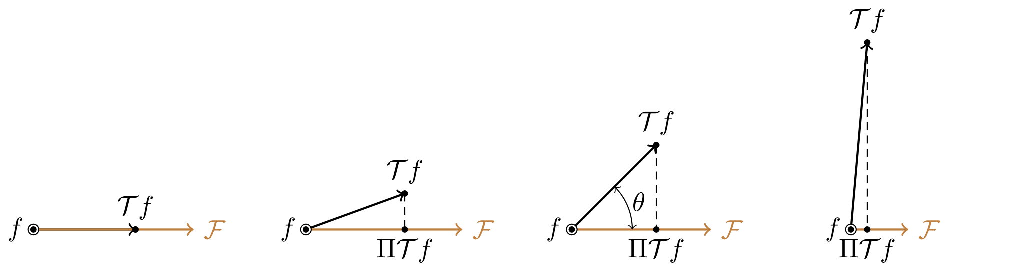
The difference between the new analysis and the existing ones becomes stark when completeness is violated. For example, in Figure 4c, the Bellman backup $\mathcal{T}f$ is mis-aligned with respect to $\mathcal{F}$, and the residual in Equation (7) can be quite large if the Bellman error $f - \mathcal{T}f$ is also large. For the specific example in Figure 4c, the residual in Equation (7) is roughly a fraction $\beta \approx 0.7$ of the full Bellman error, i.e., $\inf_{g\in\mathcal{F}}\|g - \mathcal{T}f\|_\mu \approx \beta\|f - \mathcal{T}f\|_\mu$. If the Bellman error happens to be large, say $\|f - \mathcal{T}f\|_\mu \approx 1$, then the residual $\inf_{g\in\mathcal{F}}\|g - \mathcal{T}f\|_\mu$ will also be large. It follows that the inherent Bellman error will be large as well, and so will the prediction error when estimated via Equation (12):

$$|V^\pi - \widehat{V}^\pi| \gtrapprox \mathcal{I}_{\mathcal{F}} \approx 1. \tag{13}$$

In other words, the bound 12 becomes vacuous. However, if the situation depicted in Figure 4c is representative of the mutual alignment between $\mathcal{T}f$ and $\mathcal{F}$ across various $f \in \mathcal{F}$, then in lieu of a large approximation error, our analysis predicts only a slowdown of a factor of $\frac{1}{1-\beta} \approx 3$ compared

[4] We say that double samples are available when the available dataset contains two independent transitions for each tuple. More precisely, it contains tuples (s, a, r, s^+, s^+_+) such that $s^+_+ \sim \mathcal{P}(s,a)$ and $s^+ \sim \mathcal{P}(s,a)$ are independent successor states, a condition hardly met outside of simulated domains or deterministic MDPs.

[5] Notice that these papers study the case where $\mathcal{T}$ is the Bellman optimality operator, which leads to slightly different expressions.



(a) $\beta = 0$ **(b)** $\beta \approx 0.34$ **(c)** $\beta \approx 0.70$ **(d)** $\beta \approx 0.996$

Figure 4: Local alignments between the Bellman backup $\mathcal{T}f$ and the class $\mathcal{F}$ for various values of β. The setting in Figure 4a is traditionally called 'Bellman complete'. In this simple example $\beta = \sin\theta$.

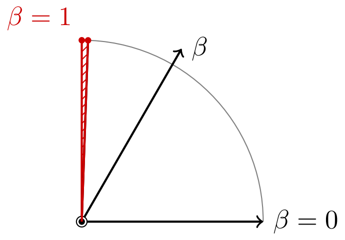


Figure 5: Off-policy reinforcement learning remains viable for values of β in the range $[0, 1)$, while prior analyses expected an unavoidable inherent Bellman error to arise. The red shaded area, which corresponds to $\beta \to 1$, represents problems where the sample complexity becomes unmanageably large, a condition in force in the lower bounds.

to the Bellman complete case:

$$|V^\pi - \widehat{V}^\pi| \lessapprox \underbrace{3}_{\frac{1}{1-\beta}} \times \frac{1}{1-\gamma}\sqrt{\frac{C\ln(|\mathcal{F}|/\delta)}{n}}. \qquad (14)$$

For such problems, the bound in display in Equation (14) is a major improvement compared to the one in Equation (13). While the analyses that lead to Equation (13) suggest that accurate predictions are out of reach due to large inherent Bellman errors, the refined one of this paper expects a minor slowdown in the rate of convergence compared to the Bellman complete case.

It is only when the Bellman backup becomes almost orthogonal to $\mathcal{F}$ that β approaches one and prediction becomes very challenging; such is the situation depicted in Figure 4d and in force in some recent lower bounds (e.g., (Foster et al., 2021)). See Figure 5 for a graphical summary. More precisely, the condition $\beta = 1$ corresponds to the existence of multiple projected fixed points. Any method based on finding projected fixed points to the Bellman equations necessarily fails to converge to the correct predictor on such problems, because the correct predictor is only one of the many possible solutions to the projected Bellman equations.

When β is close to one, the classical bound in Equation (12) can be tighter than the new bound in Equation (11). Of course, one can always select the tighter of the two. Likewise, it is possible to leverage the more general notion of local inherent Bellman error instead of that of β-incompleteness and achieve tighter error guarantees than the ones that we present, but doing so would have only been possible at the expense of the clarity of exposition. Instead, the key contribution of this work is to interpret the inherent Bellman error no longer as an unavoidable approximation error that must be zero for the approximation error to be zero, but as a quantity that naturally decreases when more samples are added. More precisely, if $\beta < 1$, as the number of samples n increases, the bound in Equation (11) eventually becomes tighter than that in Equation (12), establishing convergence to the optimal predictor even when the inherent Bellman error is non-zero. See also Appendix A.

4.3. Further comparison with existing literature

One work close to ours is (Xie & Jiang, 2020b), which operates with stronger concentrability requirements. Another one is the non-linear Bubnov-Galerkin method (Zanette & Wainwright, 2022), for which we may expect similar considerations to apply; however, the violation of completeness is not quantified in an interpretable way in that work.

Our result is due to a refined analysis, as well as to an appropriate definition, and not to a new algorithm. The minimax formulation has been analyzed multiple times, (Antos et al., 2008; Chen & Jiang, 2019; Xie et al., 2021; Jin et al., 2021; Duan et al., 2021; Xie et al., 2022) but to our knowledge all analyses use the inherent Bellman errors. Although our minimax formulation is for policy evaluation, as the proof will clarify, the same argument applies to policy optimization (i.e., when $\mathcal{T}$ is the Bellman optimality operator). Finally, our work removes the binary distinction between Bellman completeness and the lower bound of (Foster et al., 2021).

Additional literature The off-policy prediction task has been widely studied. Earlier methods where based on temporal difference (TD) (Sutton, 1988); they include Q-learning (Watkins & Dayan, 1992) and fitted Q iteration (Ernst et al., 2005; Munos & Szepesvári, 2008). These TD methods are key to the recent successes of RL (Tesauro et al., 1995; Mnih et al., 2013; 2015; 2016; Fujimoto et al., 2018).

A more robust TD variant which is however harder to optimize numerically is the minimax formulation that we investigate here (Antos et al., 2008); its relation with TD methods has been investigated by (Chen & Jiang, 2019). The minimax formulations has been adopted for provably efficient exploration (Jin et al., 2021) and offline robust optimization (Xie et al., 2021). More recently, the minimax formulation has been used as a proxy to analyze theoretically an empirical algorithm based on TD (Cheng et al., 2022). An analysis based on local Rademacher averages is given in (Duan et al., 2021). All these analyses require Bellman completeness, or otherwise the inherent Bellman error must be suffered.

Many other algorithms for the off-policy prediction problems have been proposed. These include importance sampling methods (Precup, 2000; Thomas & Brunskill, 2016; Jiang & Li, 2016; Liu et al., 2018; Farajtabar et al., 2018), which do not require completeness but can only tolerate small distribution shifts.

More recent literature has proposed weight-learning methods which rely on the knowledge of certain weights, typically the marginalized importance ratios between the distribution that collected the data and the target policy (Liu et al., 2018; Xie & Jiang, 2020a; Zhan et al., 2022; Nachum et al., 2019; Xie et al., 2019; Zhang et al., 2020a;b; Yang et al., 2020; Kallus & Uehara, 2019; Jiang & Huang, 2020; Uehara et al., 2020; Zanette & Wainwright, 2022; Rashidinejad et al., 2022). While these algorithms can avoid Bellman completeness, they rely on additional assumptions, such as realizability of the weight class, and more generally they leverage additional domain knowledge which is implicit in the choice of the weight class. For example, (Uehara et al., 2021) makes completeness assumptions about the weight class, and (Zhan et al., 2022) assume realizability for both the weight and value class. An additional high-level viewpoint is presented in Appendix B.

Two notable exceptions to completeness are (Xie & Jiang, 2020b; Zanette & Wainwright, 2022); however (Xie & Jiang, 2020b) make very strong assumptions on the concentrability factor, while the violation of the completeness condition is not quantified in (Zanette & Wainwright, 2022). The violation of completeness is also examined algebraically and algorithmically for the linear setting by (Perdomo et al., 2022). For off-policy learning with pessimism and linear methods, completeness was removed via a Bubnov-Galerkin approach in (Zanette & Wainwright, 2022) while still ensuring computational tractability; in contrast, here we focus on more general non-linear predictors.

Fundamental limits were investigated in (Zanette, 2020; Wang et al., 2020; Foster et al., 2021). Collectively they show that hard-to-learn structures can arise in absence of Bellman completeness, or with large distribution shift. Our paper describes the intermediate situation between these lower bounds and the Bellman complete setting. Related papers include (Duan & Wang, 2020; Duan et al., 2021; Tang et al., 2019; Nachum & Dai, 2020; Uehara et al., 2021; Chen & Qi, 2022; Chang et al., 2022).

Other papers have implicitly examined settings that are intermediate between realizability and completeness, such as (Wei et al., 2022; Ye et al., 2022). In their setting, if the corruption continues through time then the regret scales linearly. Rather, our setting is corruption free, and we can indeed converge to the optimal solution when $\beta < 1$.

5. Conclusion

In this work we have re-analyzed the statistical complexity of off-policy reinforcement learning on Bellman-incomplete MDPs using temporal-difference-style algorithms. The work establishes that there exists a full spectrum between Bellman completeness and the existing lower bounds where off-policy reinforcement learning remains statistically viable, even without additional domain knowledge, such as weights or test classes, and with no approximation error. The key advancement is due to a localization argument, which removes the approximation error associated to the lack of Bellman completeness.

Even though we presented our findings for the policy evaluation problem, the optimization setting is immediately covered by replacing the Bellman evaluation operator with its optimization counterpart; since our main analysis only relies on the boundedness of the Bellman evaluation operator, this is a straightforward operation. We also expect these insights to extend directly to the setting of exploration and of pessimistic policy learning. More generally, we believe that a local analysis can be a useful tool to analyze new algorithms or existing ones in other settings as well. It can help carefully assess how the violation of a certain assumption affects the performance of an algorithm, so as to relax some structural assumptions in a way that does not introduce an approximation error.

Finally, although our paper exhibits an algorithm to find high-quality solutions in absence of Bellman completeness, there is no guarantee that such points can be found in a computationally efficient way. For example, TD methods do not always converge, although when they do, they inherit such bounds. That raises an interesting question, one that concerns possible statistical-computational trade-offs to be made in reinforcement learning.

Acknowledgments

The author is grateful to the reviewers for their helpful comments which helped improve the quality of the paper. The author is supported by a fellowship from the Foundation of Data Science Institute (FODSI).

References

Antos, A., Szepesvári, C., and Munos, R. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008.

Baird, L. Residual algorithms: Reinforcement learning with function approximation. In *International Conference on Machine Learning (ICML)*. 1995.

Bartlett, P. L., Bousquet, O., and Mendelson, S. Local rademacher complexities. *The Annals of Statistics*, 33(4): 1497–1537, 2005.

Bertsekas, D. *Dynamic programming and stochastic control*, volume 2. Athena Scientific, Belmont, MA, 1995a.

Bertsekas, D. P. *Dynamic programming and stochastic control*, volume 1. Athena Scientific, Belmont, MA, 1995b.

Bertsekas, D. P. and Tsitsiklis, J. N. *Neuro-dynamic programming*. Athena Scientific, 1996.

Chang, J., Wang, K., Kallus, N., and Sun, W. Learning bellman complete representations for offline policy evaluation. In *International Conference on Machine Learning*, pp. 2938–2971. PMLR, 2022.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pp. 1042–1051, 2019.

Chen, X. and Qi, Z. On well-posedness and minimax optimal rates of nonparametric q-function estimation in off-policy evaluation. *arXiv preprint arXiv:2201.06169*, 2022.

Cheng, C.-A., Xie, T., Jiang, N., and Agarwal, A. Adversarially trained actor critic for offline reinforcement learning. *arXiv preprint arXiv:2202.02446*, 2022.

Duan, Y. and Wang, M. Minimax-optimal off-policy evaluation with linear function approximation. *arXiv preprint arXiv:2002.09516*, 2020.

Duan, Y., Jin, C., and Li, Z. Risk bounds and rademacher complexity in batch reinforcement learning. *arXiv preprint arXiv:2103.13883*, 2021.

Ernst, D., Geurts, P., and Wehenkel, L. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6, 2005.

Fan, J., Wang, Z., Xie, Y., and Yang, Z. A theoretical analysis of deep q-learning. In *Learning for Dynamics and Control*, pp. 486–489. PMLR, 2020.

Farajtabar, M., Chow, Y., and Ghavamzadeh, M. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pp. 1447–1456. PMLR, 2018.

Foster, D. J., Krishnamurthy, A., Simchi-Levi, D., and Xu, Y. Offline reinforcement learning: Fundamental barriers for value function approximation, 2021.

Fujimoto, S., Hoof, H., and Meger, D. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pp. 1587–1596. PMLR, 2018.

Jiang, N. and Huang, J. Minimax value interval for off-policy evaluation and policy optimization. *arXiv preprint arXiv:2002.02081*, 2020.

Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 652–661. PMLR, 2016.

Jin, C., Liu, Q., and Miryoosefi, S. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *arXiv preprint arXiv:2102.00815*, 2021.

Kakade, S. M. et al. *On the sample complexity of reinforcement learning*. PhD thesis, University of London London, England, 2003.

Kallus, N. and Uehara, M. Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *arXiv preprint arXiv:1909.05850*, 2019.

Liu, Q., Li, L., Tang, Z., and Zhou, D. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pp. 5356–5366, 2018.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533, 2015.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pp. 1928–1937. PMLR, 2016.

Munos, R. Error bounds for approximate value iteration. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2005.

Munos, R. and Szepesvári, C. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9 (May):815–857, 2008.

Nachum, O. and Dai, B. Reinforcement learning via Fenchel-Rockafellar duality. *arXiv preprint arXiv:2001.01866*, 2020.

Nachum, O., Dai, B., Kostrikov, I., Chow, Y., Li, L., and Schuurmans, D. Algaedice: Policy gradient from arbitrary experience. *arXiv preprint arXiv:1912.02074*, 2019.

Perdomo, J. C., Krishnamurthy, A., Bartlett, P., and Kakade, S. A sharp characterization of linear estimators for offline policy evaluation. *arXiv preprint arXiv:2203.04236*, 2022.

Precup, D. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, pp. 80, 2000.

Puterman, M. L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1994. ISBN 0471619779.

Rashidinejad, P., Zhu, H., Yang, K., Russell, S., and Jiao, J. Optimal conservative offline rl with general function approximation via augmented lagrangian. *arXiv preprint arXiv:2211.00716*, 2022.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Sutton, R. S. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT Press, 2018.

Talagrand, M. A new look at independence. *The Annals of probability*, pp. 1–34, 1996.

Tang, Z., Feng, Y., Li, L., Zhou, D., and Liu, Q. Doubly robust bias reduction in infinite horizon off-policy estimation. *arXiv preprint arXiv:1910.07186*, 2019.

Tesauro, G. et al. Temporal difference learning and td-gammon. *Communications of the ACM*, 38(3):58–68, 1995.

Thomas, P. and Brunskill, E. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 2139–2148, 2016.

Uehara, M., Huang, J., and Jiang, N. Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pp. 9659–9668. PMLR, 2020.

Uehara, M., Imaizumi, M., Jiang, N., Kallus, N., Sun, W., and Xie, T. Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency. *arXiv preprint arXiv:2102.02981*, 2021.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Wang, R., Foster, D. P., and Kakade, S. M. What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*, 2020.

Wang, Y., Wang, R., and Kakade, S. M. An exponential lower bound for linearly-realizable MDPs with constant suboptimality gap. *arXiv preprint arXiv:2103.12690*, 2021.

Watkins, C. J. and Dayan, P. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

Wei, C.-Y., Dann, C., and Zimmert, J. A model selection approach for corruption robust reinforcement learning. In *International Conference on Algorithmic Learning Theory*, pp. 1043–1096. PMLR, 2022.

Weisz, G., Amortila, P., and Szepesvári, C. Exponential lower bounds for planning in MDPs with linearly-realizable optimal action-value functions. *arXiv preprint arXiv:2010.01374*, 2020.

Xie, T. and Jiang, N. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. volume 124 of *Proceedings of Machine Learning Research*, pp. 550–559, Virtual, 03–06 Aug 2020a. PMLR. URL http://proceedings.mlr.press/v124/xie20a.html.

Xie, T. and Jiang, N. Batch value-function approximation with only realizability. *arXiv preprint arXiv:2008.04990*, 2020b.

Xie, T., Ma, Y., and Wang, Y.-X. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Advances in Neural Information Processing Systems*, pp. 9668–9678, 2019.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. *arXiv preprint arXiv:2106.06926*, 2021.

Xie, T., Foster, D. J., Bai, Y., Jiang, N., and Kakade, S. M. The role of coverage in online reinforcement learning. *arXiv preprint arXiv:2210.04157*, 2022.

Yang, M., Nachum, O., Dai, B., Li, L., and Schuurmans, D. Off-policy evaluation via the regularized lagrangian. *arXiv preprint arXiv:2007.03438*, 2020.

Ye, C., Xiong, W., Gu, Q., and Zhang, T. Corruption-robust algorithms with uncertainty weighting for nonlinear contextual bandits and markov decision processes. *arXiv preprint arXiv:2212.05949*, 2022.

Zanette, A. Exponential lower bounds for batch reinforcement learning: Batch rl can be exponentially harder than online RL. *arXiv preprint arXiv:2012.08005*, 2020.

Zanette, A. and Wainwright, M. J. Bellman residual orthogonalization for offline reinforcement learning, 2022. URL https://arxiv.org/abs/2203.12786.

Zanette, A., Lazaric, A., Kochenderfer, M., and Brunskill, E. Learning near optimal policies with low inherent Bellman error. In *International Conference on Machine Learning (ICML)*, 2020.

Zhan, W., Huang, B., Huang, A., Jiang, N., and Lee, J. D. Offline reinforcement learning with realizability and single-policy concentrability. *arXiv preprint arXiv:2202.04634*, 2022.

Zhang, R., Dai, B., Li, L., and Schuurmans, D. Gendice: Generalized offline estimation of stationary values. *arXiv preprint arXiv:2002.09072*, 2020a.

Zhang, S., Liu, B., and Whiteson, S. Gradientdice: Rethinking generalized offline estimation of stationary values. In *International Conference on Machine Learning*, pp. 11194–11203. PMLR, 2020b.

Appendix

The appendix is organized as follows:

- Appendix A presents further comments, particularly related to the weight methods
- Appendix B describes additional notation
- Appendix C describes additional results
- Appendix D presents the main proof of the paper
- Appendix E presents the main technical sub-component of the paper, which is the rate of the minimax program
- Appendix F presents some technical results needed in the prior sections

A. Further Comments on the Relation between TD and Weight Methods

There is a solid high-level connection between TD and weight methods, which we discuss in this section.

If one had access to a generative model, the mean-squared Bellman error can be minimized to find a good predictor. However, without a generative model, it is not possible to directly estimate (and thus minimize) the mean square Bellman error when function approximation is implemented. In this case, the 'standard' approach (e.g., temporal difference learning, fitted Q, but also the minimax formulation that we examine here) is to roughly minimize the projected Bellman error. To be more precise, the Bellman error is projected onto $\mathcal{F}$. Of course, the projection may discard important components of the Bellman error (those orthogonal to $\mathcal{F}$), and so there is a loss in sample efficiency, which our work quantifies with the scalar β. When prior art assumed Bellman completeness, they assumed that there are no orthogonal components.

One might wonder whether it makes sense to 'project' the Bellman error along different spaces (i.e., a space $\mathcal{V}$ different from $\mathcal{F}$). This idea roughly leads to the class of weight methods, although they are normally not presented as methods doing projections; see the paper (Zanette & Wainwright, 2022) for one such viewpoint.

Which one (TD or weight learning) is better? The answer is problem dependent. At a very basic level, if $\mathcal{F}$ is well aligned with the Bellman error, TD-style methods are superior. If one has specific knowledge of a subspace $\mathcal{V}$ that better captures the Bellman error, then a weight learning method can be used. A special case of this is, for instance, when $\mathcal{V}$ contains the density ratio of the target policy with respect to the behavioral policy.

While weight learning methods are conceptually appealing, it is rare to have such domain knowledge to exploit with a weight learning method, and so TD-style methods (broadly those that we analyze here) remain very popular.

B. Additional Notation

TD and Bellman errors For a given Q-function and policy π, let us define the *temporal difference error* (or TD error) associated to the sample (s, a, r, s^+) and the *Bellman error* at (s, a)

$$(\delta f)(s,a,r,s^+) \overset{def}{=} f(s,a) - r - \gamma f(s^+,\pi), \qquad (\mathcal{B}f)(s,a) \overset{def}{=} f(s,a) - r(s,a) - \gamma \mathbb{E}_{s^+\sim\mathcal{P}(s,a)} f(s^+,\pi). \tag{15}$$

The TD error is a random variable function of (s, a, r, s^+), while the Bellman error is its conditional expectation with respect to the immediate reward and successor state at (s, a).

Function class We deal with a function class $\mathcal{F}$ that contains a set of predictors f defined over the state and action space. They are bounded in supremum norm, i.e., $\sup_{(s,a)}|f(s, a)| \le 1$, a bound that must apply to $f^\star$ as well since we assume realizability.

Some of our results are presented using a statistical complexity notion called Rademacher complexity. The Rademacher complexity of a function class measures the expected worst-case alignment of a predictor $f \in \mathcal{F}$, evaluated in a n-dimensional space over the random covariates $(S_i, A_i) \sim \mu$, with the Rademacher noise ϵ_i, which takes value -1 and $+1$ with equal probability. It is defined for a function class $\mathcal{F}$ as

$$\mathcal{R}_n\big[\mathcal{F}\big] = \mathbb{E}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(S_i, A_i)\Big|.$$

When presenting our results for general function approximation, we use a set that contains functions that are at most $r \geq 0$ away from the optimal one. It is defined as

$$(\mathcal{F} - f^\star)(r) = \{f - f^\star \mid \|f - f^\star\|_\mu \leq r,\ f \in \mathcal{F}\}. \tag{16}$$

C. Additional Results

C.1. Proof of Proposition 1

Proof. Let us focus on the first statement and fix two radii $r \le r'$ where $\mathcal{I}$ exists. The supremum $\sup_f$ for $\mathcal{I}(r)$ is over $\mathcal{F}(r)$ while for $\mathcal{I}(r')$ it is over $\mathcal{F}(r')$; in both cases, the infimum $\inf_g$ is over the original class $\mathcal{F}$. Since $\mathcal{F}(r) \subseteq \mathcal{F}(r')$, taken together these observations imply

$$\mathcal{I}(r) \overset{def}{=} \sup_{f \in \mathcal{F}(r)} \inf_{g \in \mathcal{F}} \|g - \mathcal{T} f\|_\mu \le \sup_{f \in \mathcal{F}(r')} \inf_{g \in \mathcal{F}} \|g - \mathcal{T} f\|_\mu \overset{def}{=} \mathcal{I}(r').$$

Now, for the second statement: when realizability holds, the set $\mathcal{F}(0) = \{f \in \mathcal{F} \mid \|f - \mathcal{T} f\|_\mu \le 0\}$ contains at least $f^\star$, and it is hence non-empty. The fact that $\mathcal{I}(0) = 0$ for a realizable problem then follows from

$$\mathcal{I}(0) \overset{def}{=} \sup_{f \in \mathcal{F}(0)} \inf_{g \in \mathcal{F}} \|g - \mathcal{T} f\|_\mu \le \sup_{f \in \mathcal{F}(0)} \|f - \mathcal{T} f\|_\mu \le 0.$$

□

C.2. Off-policy cost coefficient

The error bound in Equation (11) can be re-written in a more suggestive way:

$$|\mathcal{E}(\widehat{f})| \le \frac{1}{1-\gamma} \sqrt{\frac{C}{1-\beta}} r_\star, \qquad \text{where} \quad r_\star^2 = \frac{\ln(|\mathcal{F}|/\delta)}{(1-\beta)n}.$$

The above regroupment has highlighted the dependence on three key factors. The first is the *rate of convergence* $r_\star$ to zero of the population-level minimax program $\mathcal{M}$ (as the proof will clarify, we have $\mathcal{M}(\widehat{f}) \lesssim r_\star^2$ with high probability). The other two factors are the concentrability coefficient C and the lack of Bellman completeness $\frac{1}{1-\beta}$. They relate how minimizing $\mathcal{M}$—represented by $r_\star$—affects the prediction error $\mathcal{E}(\widehat{f})$.

A natural question to ask is whether it makes sense to have two factors, rather than a single entity, to relate the value of the program $\mathcal{M}(\widehat{f})$ and the prediction error $\mathcal{E}(\widehat{f})$. In fact, it is possible to adopt a more direct approach and directly measure how minimizing $\mathcal{M}(f)$ affects the prediction error $\mathcal{E}(\widehat{f})$, and denote the worst-case ratio by $C^\star$:

$$C^\star \overset{def}{=} \sup_{f \in \mathcal{F}} \frac{\mathcal{E}(f)^2}{\mathcal{M}(f)} \approx \frac{\text{quantity of interest}}{\text{quantity being minimized}}. \tag{17}$$

The off-policy cost coefficient $C^\star$ so defined always leads to tighter bounds: it is always smaller than the product between C and the incompleteness factor $\frac{1}{1-\beta}$ that appears in Theorem 1:

$$C^\star \le \frac{1}{(1-\gamma)^2} \frac{C}{1-\beta}.$$

In fact, the proof of Theorems 1 and 2 computes the performance bound of the minimax algorithm using $C^\star$, only to relax it at the end by using the above display to make the result more interpretable; one can thus directly replace $\frac{1}{(1-\gamma)^2}\frac{C}{1-\beta}$ in Equation (11) and Equation (18) to follow with $C^\star$.

Although $C^\star$ is less interpretable in terms of fundamental reinforcement learning quantities, its use should be preferred for two reasons. The first is that it is smaller, i.e., $C^\star$ can be small even when $\frac{C}{1-\beta}$ is large. The second is that it reflects more truthfully the learning mechanics of the algorithm: $C^\star$ directly bounds the ratio between the quantity of interest—the prediction error $|\mathcal{E}(f)|$—and the one being controlled—the value of the minimax program $\mathcal{M}(f)$—and it is thus the 'correct' way to quantify the cost of off-policy learning with the minimax procedure.

C.3. Error bounds with more general function approximation

In practice, TD methods are implemented as gradient-based algorithms, using differentiable approximators that are far more complex then finite classes and that may operate in a non-parametric regime, such as neural networks. In such cases, we do not expect a $\sqrt{n}$ rate of convergence. In order to provide error bounds that apply to the latter setting, in this section we express the result using Rademacher averages, which are standard ways to quantify the capacity of a function class.

As with the localized inherent Bellman error in Section 3, the relevant sets to determine the statistical complexity—and hence the rate of convergence—are subsets of $\mathcal{F}$ where we expect the predictor $\widehat{f}$ to be. We expect these sets (and their Rademacher complexity) to become smaller as n increases, much like the incompleteness function.

What determines the rate of convergence $r_\star$ then is a certain relation presented in Equation (19). It involves the Rademacher complexity of these localized sets, which is a standard way to express the rates of convergences with generic function classes (Bartlett et al., 2005; Wainwright, 2019). The conditions in Equation (19) must admit a solution $r_\star$ such that the requirement holds for all $r \geq r_\star$; this requirement is met by the bounded classes we consider. We further assume that there are no measurability issues when stating and proving the following theorem; in particular we assume that the prerequisites for using Talagrand are met in order to avoid measurability issues.

Theorem 2 (Error Bounds with General Function Approximation)**.** *With probability at least $1-\delta$, the prediction error of the minimizer $\widehat{f}$ satisfies the bound*

$$|\mathcal{E}(\widehat{f})| \leq \frac{r_\star}{1-\gamma}\sqrt{\frac{C}{1-\beta}} \tag{18}$$

where the rate of convergence $r_\star$ is such that all $r \geq r_\star$ satisfy the inequalities

$$\mathcal{R}_n\Big\{L(f,f) - L(g_f,f) \mid \mathbb{E}[L(f,f) - L(g_f,f)] \leq 2r^2\Big\} \leq c_1 r^2, \tag{19a}$$

$$\mathcal{R}_n\Big[(\mathcal{F} - f^\star)\big(Kr\big)\Big] \leq c_2 r^2, \tag{19b}$$

$$(K+1)\frac{\ln(1/(\delta r))}{n} \leq c_3 r^2. \tag{19c}$$

for three universal constants $c_1, c_2, c_3 > 0$, and $K = 1$ if $\mathcal{F}$ is convex or $K = \frac{1}{1-\beta}$ if $\mathcal{F}$ is non-convex.

The rate of convergence $r_\star$ is that of the minimax procedure, i.e., we have $\mathcal{M}(\widehat{f}) \lesssim r_\star^2$ with high probability. Let us add that convexity always leads to improved bounds. The second and third critical inequalities in Equation (19) are standard, while the first involves the Bellman operator, and can be relaxed only with additional assumptions (Duan et al., 2021).

In all cases, in order to determine the rate of convergence $r_\star$, the first step is to compute the local Rademacher averages in Equation (19) as a function of r, and the second step is to solve for r the resulting relation, finding $r_\star$. It is enough to compute an upper bound to the local Rademacher complexity. Likewise, it is sufficient to identify any value $r_\star$ that solves the resulting relation, but the smaller the $r_\star$, the better the rate of convergence that we can guarantee. Of course, in order to obtain concrete and interpretable bounds, one must consider specific function classes, see the book (Wainwright, 2019) for several parametric as well as non-parametric examples.

Finally, let us mention that the bound that we present here uses the coefficient β which represents the average behavior of $\mathcal{I}$, but intuitively, it is the actual shape of the incompleteness function $\mathcal{I}$ around the origin that determines the problem complexity. It is possible to obtain critical relations involving the incompleteness function, much like those in Equation (19). However, implementing this observation would have made the analysis less clear and the final result less interpretable, and so we leave that for future studies.

D. Main Analysis

In this section we prove Theorems 1 and 2.

Proof techniques Although the minimax formulation has been analyzed previously in a number of works (see e.g., (Chen & Jiang, 2019) for a relatively recent analysis), our proof differs from what is available in the literature from the very set-up, as the concept of local Bellman errors arises quite soon in the proof in Appendix D. In addition, there is substantial technical novelty in the way we bound the minimax program in Appendix E, where the statistical localization, as well as the definition of β, are leveraged explicitly.

Setting up the proof In order to prove the theorems, we need to establish a high probability bound on the estimation error, which is the value function difference at the initial state s_0, i.e., the quantity $|\mathcal{E}(\widehat{f})| = |(f^\star - \widehat{f})(s_0, \pi)|$.

The proof is based on the following key observation: since $\widehat{f}$ minimizes the empirical loss $\widehat{\mathcal{M}}$, we expect that we can bound its population value $\mathcal{M}(\widehat{f})$. Following the suggestion outlined in Appendix C.2, we factorize the squared prediction error as

$$\mathcal{E}(\widehat{f})^2 = \frac{\mathcal{E}(\widehat{f})^2}{\mathcal{M}(\widehat{f})} \times \mathcal{M}(\widehat{f}) \le C^\star \times \mathcal{M}(\widehat{f}).$$

The off-policy cost coefficient $C^\star$ is defined in Equation (17), and connects the prediction error to the population-based value of the minimax program. In order to complete the proof, we need to bound $C^\star$ and $\mathcal{M}(\widehat{f})$.

Bounding $C^\star$ A variation of the simulation lemma (Kakade et al., 2003) allows us to upper bound the numerator in $C^\star$; it is proved in Appendix D.1.

Lemma 1 (Weak Simulation Lemma)**.** *For any $f \in \mathcal{F}$ we have the bound*

$$|\mathcal{E}(f)| \le \frac{1}{1-\gamma}\|f - \mathcal{T}f\|_\pi.$$

In addition, we can lower bound the denominator in $C^\star$ with simple algebra.

Lemma 2 (Effect of β-incompleteness)**.** *For any $f \in \mathcal{F}$ we have the bound*

$$\|f - \mathcal{T}f\|_\mu^2 \le \frac{1}{1-\beta}\mathcal{M}(f).$$

The above lemma is where the definition of β-incompleteness is leveraged; however, β-incompleteness also plays a role in determining the rate of convergence of the minimax program in Proposition 3. After putting together the pieces, we obtain

$$\begin{aligned} C^\star &\le \sup_{f\in\mathcal{F}} \frac{\mathcal{E}(f)^2}{\mathcal{M}(f)} \\ &\le \sup_{f\in\mathcal{F}} \Big(\frac{1}{1-\gamma}\Big)^2 \frac{1}{1-\beta} \frac{\|f-\mathcal{T}f\|_\pi^2}{\|f-\mathcal{T}f\|_\mu^2} \\ &= \Big(\frac{1}{1-\gamma}\Big)^2 \frac{1}{1-\beta} C. \end{aligned}$$

Bounding $\mathcal{M}(\widehat{f})$ In order to conclude, we must establish a high probability rate of convergence for the population loss evaluated at the empirical minimizer $\widehat{f}$. Such rate of convergence depends on the function class $\mathcal{F}$. More precisely, if the function class $\mathcal{F}$ has finite cardinality, the rate of convergence is

$$r_\star^2 \simeq \frac{\ln(|\mathcal{F}|/\delta)}{(1-\beta)n},$$

while for a general function class it must be such that any $r \ge r_\star$ satisfies Equation (19).

Proposition 3 (Rate of Minimax)**.** *With probability at least $1-\delta$*

$$\mathcal{M}(\widehat{f}) \lesssim r_\star^2.$$

The proof of Proposition 3 is in the appendix. Combined with the bound on $C^\star$, the proof of Theorems 1 and 2 is complete.

D.1. Proof of Lemma 1 (*Weak Simulation Lemma*)

For a fixed function $f \in \mathcal{F}$, the simulation lemma (e.g., (Kakade et al., 2003)) ensures

$$
\begin{aligned}
|\mathcal{E}(f)| &= |(f^\star - f)(s_0, \pi)| \\
&= |\frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d^\pi}(f - \mathcal{T}f)(s,a)| \\
&\le \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d^\pi}\sqrt{[(f - \mathcal{T}f)(s,a)]^2}
\end{aligned}
$$

Using the Jensen's inequality we obtain the upper bound

$$
\begin{aligned}
&\le \frac{1}{1-\gamma}\sqrt{\mathbb{E}_{(s,a)\sim d^\pi}[(f - \mathcal{T}f)(s,a)]^2} \\
&= \frac{1}{1-\gamma}\|f - \mathcal{T}f\|_\pi.
\end{aligned}
$$

D.2. Proof of Lemma 2 (*Effect of β-incompleteness*)

We can write

$$
\begin{aligned}
\mathcal{M}(f) &= \|f - \mathcal{T}f\|_\mu^2 - \|g_f - \mathcal{T}f\|_\mu^2 \\
&\ge \|f - \mathcal{T}f\|_\mu^2 - \beta^2\|f - \mathcal{T}f\|_\mu^2 \\
&= (1-\beta^2)\|f - \mathcal{T}f\|_\mu^2 \\
&= (1-\beta)(1+\beta)\|f - \mathcal{T}f\|_\mu^2 \\
&\ge (1-\beta)\|f - \mathcal{T}f\|_\mu^2.
\end{aligned}
$$

The last inequality follows from the fact that $\beta \in [0,1]$.

E. Proof of Proposition 3 (*Rate of Minimax*)

We will show that the population and the empirical loss are related, i.e., that

$$\mathcal{M}(\widehat{f}) \lesssim 2\widehat{\mathcal{M}}(\widehat{f})$$

with high probability. Next, since $\widehat{f}$ minimizes $\widehat{\mathcal{M}}$, and realizability holds, we should have that $\widehat{\mathcal{M}}(\widehat{f})$ is small, or more precisely that with high probability

$$\widehat{\mathcal{M}}(\widehat{f}) \lesssim r_\star^2.$$

Together, they imply the statement. In order to proceed we need to introduce more notation.

E.1. Notation, Empirical Processes and Failure Events

We need to show that the bad event

$$\mathcal{M}(\widehat{f}) \gtrsim 2r_\star^2 \tag{20}$$

occurs with probability at most δ. Since $\widehat{f}$ is random, we establish uniform convergence results, i.e., statements that hold for many (possibly all) functions $f \in \mathcal{F}$. In order to do so, we need to analyze the statistical fluctuations of the empirical process associated to the cost function that defines the loss:

$$X(f) \overset{def}{=} L(f, f) - L(g_f, f).$$

This is a natural quantity to analyze, because its expectation (which is computed with the help of Lemma 6) is precisely the quantity that we wish to control

$$\begin{aligned}
\mathbb{P}X(f) &= \mathbb{E}_{(s,a)\sim\mu}\Big[\mathbb{E}_{r\sim R(s,a),s^+\sim\mathcal{P}(s,a)} X(f)\Big] \\
&= \mathcal{L}(f,f) + \sigma(f)^2 - \mathcal{L}(g_f, f) - \sigma(f)^2 \\
&= \mathcal{L}(f,f) - \mathcal{L}(g_f, f) \\
&= \mathcal{L}(f,f) - \inf_{g\in\mathcal{F}} \mathcal{L}(g, f) \\
&= \mathcal{M}(f),
\end{aligned}$$

while its empirical average is upper bounded by the empirical loss that the agent minimizes

$$\begin{aligned}
\mathbb{P}_n X(f) &= \frac{1}{n} \sum_{(s,a,r,s^+)\in\mathcal{D}} X(f) \\
&= \widehat{\mathcal{L}}(f,f) - \widehat{\mathcal{L}}(g_f, f) \\
&\le \widehat{\mathcal{L}}(f,f) - \inf_{g\in\mathcal{F}} \widehat{\mathcal{L}}(g, f) \\
&= \widehat{\mathcal{M}}(f).
\end{aligned}$$

E.1.1. Setting up the Failure Events

As outlined, we need to establish that it is unlikely that $\widehat{\mathcal{M}}(\widehat{f})$ is large

$$\mathbb{P}(F_1) \le \delta/2 \qquad \text{where } F_1 \;:\; \quad \widehat{\mathcal{M}}(\widehat{f}) > r_\star^2. \tag{21}$$

When the failure event F_1 does not occur, we have $\mathbb{P}_n X(\widehat{f}) \le \widehat{\mathcal{M}}(\widehat{f}) \le r_\star^2$. If we can claim $\mathcal{M}(\widehat{f}) = \mathbb{P}X(\widehat{f}) \le 2\mathbb{P}_n X(\widehat{f})$ then the proof would be complete. Unfortunately, the latter claim is not true in general. However, notice that if $\mathcal{M}(\widehat{f}) = \mathbb{P}X(\widehat{f}) \le r_\star^2$ then we can already jump to the conclusion. Therefore, it is sufficient (and more convenient) to show that it is unlikely that $\mathbb{P}X(\widehat{f})$ is large (i.e., $> r_\star^2$) and at the same time the deviation is large $\mathbb{P}X(\widehat{f}) > 2\mathbb{P}_n X(\widehat{f})$:

$$\mathbb{P}(F_2) \le \delta/2 \qquad \text{where } F_2 \;:\; \quad \mathbb{P}X(\widehat{f}) > 2\mathbb{P}_n X(\widehat{f}) \quad \text{and} \quad \mathbb{P}X(\widehat{f}) > r_\star^2. \tag{22}$$

To recap: when neither F_1 nor F_2 occur either we have

$$\mathcal{M}(\widehat{f}) = \mathbb{P}X(\widehat{f}) \le r_\star^2$$

or otherwise we have

$$\mathcal{M}(\widehat{f}) = \mathbb{P}X(\widehat{f}) \le 2\mathbb{P}_n(\widehat{f}) \le 2\widehat{\mathcal{M}}(\widehat{f}) \le 2r_\star^2,$$

and the proof would be complete. Consequently, the rest of the proof is devoted to showing that F_1 and F_2 are unlikely to occur, namely the claims in Equations (21) and (22).

E.1.2. Relaxing the Failure Events

In this section we define events that are easier to bound and that lead to the stated result in Equations (21) and (22).

Relaxing the claim in Equation (22) The difference $\mathbb{P}X(f) - \mathbb{P}_n X(f)$ is a concentration term. It is convenient to introduce the set of functions under consideration

$$\widetilde{\mathcal{U}}(r_\star) = \{f \in \mathcal{F} \mid \mathbb{P}X(f) > r_\star^2\}.$$

To establish the claim in Equation (22) it is enough to establish that large deviations are unlikely for all functions with large expectation, i.e., that

$$\exists f \in \widetilde{\mathcal{U}}(r_\star) \text{ such that } \quad \mathbb{P}X(f) - \mathbb{P}_n X(f) > \frac{1}{2}\mathbb{P}X(f) \tag{23}$$

can occur with probability at most $\delta/2$.

Relaxing the claim in Equation (21) In order to provide the required bound, we need to leverage the fact that $\widehat{f}$ is minimizing $\widehat{\mathcal{M}}(f)$.

$$\widehat{\mathcal{M}}(\widehat{f}) \le \widehat{\mathcal{M}}(f^\star) = \widehat{\mathcal{L}}(f^\star, f^\star) - \widehat{\mathcal{L}}(\widehat{g}_{f^\star}, f^\star).$$

The term to bound is the empirical (excess) risk of a realizable problem. For convenience, define the empirical process

$$Y(g) \overset{def}{=} L(f^\star, f^\star) - L(g, f^\star).$$

With the above definition we have

$$\begin{aligned} \mathbb{P}_n Y(g) &= \widehat{\mathcal{L}}(f^\star, f^\star) - \widehat{\mathcal{L}}(g, f^\star), \\ \mathbb{P}Y(g) &= \mathcal{L}(f^\star, f^\star) - \mathcal{L}(g, f^\star) \le 0 \end{aligned}$$

To recap: if we can show that with probability $1 - \delta/2$

$$\mathbb{P}_n Y(g) \le \frac{1}{2} r_\star^2 \qquad \text{for all } g \in \mathcal{F} \tag{24}$$

then under the same event we have the desired bound

$$\widehat{\mathcal{M}}(\widehat{f}) \le \mathbb{P}_n Y(\widehat{g}_{f^\star}) \le \frac{1}{2} r_\star^2.$$

E.2. Concentration inequalities for finite classes

In this section we complete the proof for the special case where $\mathcal{F}$ has finite cardinality.

E.2.1. Establishing Equation (23)

To complete the proof, we need to compute the threshold $r_\star$ past which the event in Equation (23) becomes unlikely.

The Bernstein's inequality (see e.g., (Wainwright, 2019) for a reference), coupled with a union bound over each function in $\widetilde{\mathcal{U}}(r_\star) \subseteq \mathcal{F}$ ensures that the following event occurs with probability at most $\delta/2$

$$\exists f \in \widetilde{\mathcal{U}}(r_\star) \text{ such that } \quad \mathbb{P}X(f) - \mathbb{P}_n X(f) \gtrsim \sqrt{\frac{\operatorname{Var} X(f) \ln(|\mathcal{F}|/\delta)}{n}} + \frac{\ln(|\mathcal{F}|/\delta)}{n}.$$

If we make the above right hand side larger then the event becomes even more unlikely. The term involving the variance can be upper bounded by upper bounding the variance

$$\operatorname{Var} X(f) \leq K\mathbb{P}X(f), \qquad \text{where} \qquad K \lesssim \frac{1}{1-\beta},$$

a result stated in Lemma 7. If in addition the fast rate is dominated by the variance term, (we shall see in few lines that this is the case), namely if for all functions in $\widetilde{\mathcal{U}}(r_\star)$

$$\frac{\ln(|\mathcal{F}|/\delta)}{n} \lesssim \sqrt{\frac{K\mathbb{P}X(f) \ln(|\mathcal{F}|/\delta)}{n}}, \tag{25}$$

then we readily obtain the smaller (and more unlikely) event defined below

$$\exists f \in \widetilde{\mathcal{U}}(r_\star) \text{ such that } \quad \mathbb{P}X(f) - \mathbb{P}_n X(f) \gtrsim \sqrt{\frac{K\mathbb{P}X(f) \ln(|\mathcal{F}|/\delta)}{n}}.$$

The fact that Equation (23) holds with probability at most $\delta/2$ then would follow if its right hand side is even bigger than the right hand side in the above display; such situation occurs if for all $f \in \widetilde{\mathcal{U}}(r_\star)$

$$\frac{1}{2}\mathbb{P}X(f) \gtrsim \sqrt{\frac{K\mathbb{P}X(f) \ln(|\mathcal{F}|/\delta)}{n}}. \tag{26}$$

Solving for $\mathbb{P}X(f)$ gives the condition

$$\mathbb{P}X(f) \gtrsim \frac{K \ln(|\mathcal{F}|/\delta)}{n}.$$

Such condition must be satisfied by all functions $f \in \widetilde{\mathcal{U}}(r_\star)$, a fact that holds true by definition of $\widetilde{\mathcal{U}}(r_\star)$ as soon as $r_\star$ satisfies

$$r_\star^2 \gtrsim \frac{K \ln(|\mathcal{F}|/\delta)}{n}. \tag{27}$$

The value for $r_\star$ established by the above inequality ensures that any function $f \in \widetilde{\mathcal{U}}(r_\star)$ satisfies the bound in display in Equation (26) (recall the definition of $\widetilde{\mathcal{U}}(r_\star)$). In addition, it also ensures that Equation (25) is always satisfied, as promised (observe that $K \geq 1$).

To recap: we have computed the critical threshold $r_\star$ past which Equation (23) occurs with vanishing probability, as desired. By doing so, we have also determined the rate of convergence $r_\star$ of the minimax program, up to a constant.

E.2.2. Establishing Equation (24)

In this section we establish Equation (24), or equivalently that the following event has probability at most $\delta/2$:

$$\text{exists } g \in \mathcal{F} \text{ such that } \quad \mathbb{P}_n Y(g) > \frac{1}{2} r_\star^2. \tag{28}$$

We start from the inequality of Bernstein coupled with a union bound over each element of $\mathcal{F}$ to ensure that the following event has probability at most $\delta/2$

$$\exists g \in \mathcal{F} \qquad \text{such that} \qquad \mathbb{P}_n Y(g) - \mathbb{P} Y(g) \gtrsim \sqrt{\frac{\operatorname{Var} Y(g) \ln(|\mathcal{F}|/\delta)}{n}} + \frac{\ln(|\mathcal{F}|/\delta)}{n}.$$

If we make the right hand side in the above display any larger, the event above becomes even more unlikely. We have the following bound on the variance (recall that $\mathbb{P}Y(g) \leq 0$), which we verify in Lemma 8

$$\operatorname{Var} Y(g) \lesssim -\mathbb{P}Y(g).$$

We obtain the following (smaller) event

$$\exists g \in \mathcal{F} \qquad \text{such that} \qquad \mathbb{P}_n Y(g) - \mathbb{P} Y(g) \gtrsim \sqrt{\frac{-\mathbb{P} Y(g) \ln(|\mathcal{F}|/\delta)}{n}} + \frac{\ln(|\mathcal{F}|/\delta)}{n}$$

or equivalently

$$\exists g \in \mathcal{F} \qquad \text{such that} \qquad \mathbb{P}_n Y(g) \geq \mathbb{P} Y(g) + c\sqrt{\frac{-\mathbb{P} Y(g) \ln(|\mathcal{F}|/\delta)}{n}} + \frac{\ln(|\mathcal{F}|/\delta)}{n}$$

for some constant $c > 0$, a bound that holds with probability at most $\delta/2$. We would then be able to conclude that Equation (28) holds with probability at most $\delta/2$ if its right hand side is always larger than the one in the above display, namely when

$$\frac{1}{2} r_\star^2 \geq \mathbb{P} Y(g) + c\sqrt{\frac{-\mathbb{P} Y(g) \ln(|\mathcal{F}|/\delta)}{n}} + \frac{\ln(|\mathcal{F}|/\delta)}{n}$$

The right hand side above is quadratic in $\sqrt{-\mathbb{P}Y(g)}$. Its maximum value[6] is

$$\frac{\ln(|\mathcal{F}|/\delta)}{n} \gtrsim \mathbb{P} Y(g) + c\sqrt{\frac{-\mathbb{P} Y(g) \ln(|\mathcal{F}|/\delta)}{n}} + \frac{\ln(|\mathcal{F}|/\delta)}{n},$$

and therefore it is sufficient that $r_\star$ satisfies the inequality

$$r_\star^2 \gtrsim \frac{\ln(|\mathcal{F}|/\delta)}{n}.$$

In other words, we have determined the minimum value for $r_\star$ past which Equation (28) becomes unlikely; furthermore, this requirement is already satisfied by that presented in Equation (27).

[6] Notice that $\mathbb{P}Y(g)$ is negative while the square root term is positive; in particular, the right hand side is maximized when $\mathbb{P}Y(g) \simeq \frac{\ln(|\mathcal{F}|/\delta)}{n}$.

E.3. Concentration inequalities for general functions

In this section we establish Equations (23) and (24) for general function classes. It is useful to define the following factor (up to a constant).

$$K \simeq \begin{cases} \frac{1}{1-\beta}, & \text{if } \mathcal{F} \text{ is non-convex} \\ 1, & \text{if } \mathcal{F} \text{ is convex.} \end{cases}$$

E.3.1. Establishing the claim in Equation (23)

In order to provide a bound to Equation (23), we need a suitable concentration inequality that can ensure fast rates by leveraging the variance of the process. However, the analysis to follow deals with the worst-case variance represented by the quantity $\sup_{f \in \widetilde{\mathcal{U}}(r_\star)} \mathbb{P}X(f)$ which can be[7] of order one. In order to tightly connect the worst-case maximum variance to the actual value of $\mathbb{P}X(f)$ of the function responsible for violating the inequality in Equation (23), it is best to partition the set $\widetilde{\mathcal{U}}(r_\star)$

$$\widetilde{\mathcal{U}}(r_\star) = \cup_{m \in [M]} \widetilde{\mathcal{U}}_m$$

according to the value of $\mathbb{P}X(f)$, i.e., using intervals that tightly bracket the possible values of $\mathbb{P}X(f)$, as follows:

$$\widetilde{\mathcal{U}}_m = \Big\{ f \in \widetilde{\mathcal{U}}(r_\star) \;\mid\; r^2 < \mathbb{P}X(f) \le 2r^2 \Big\}, \qquad \text{where } r^2 = 2^{m-1} r_\star^2.$$

The partition starts at $m = 1$ where $r = r_\star$ and since (see footnote) $\mathbb{P}X(f) \lesssim 1$, the partition can stop at $M \simeq \log_2(1/r_\star)$. When $f \in \widetilde{\mathcal{U}}_m$ we have $\mathbb{P}X(f) \ge r^2$ and therefore we can create a larger event which is easier to bound

$$\Big\{ \exists f \in \widetilde{\mathcal{U}}_m \;\mid \mathbb{P}X(f) - \mathbb{P}_n X(f) > \frac{1}{2}\mathbb{P}X(f) \Big\} \subseteq \Big\{ \exists f \in \widetilde{\mathcal{U}}_m \;\mid \mathbb{P}X(f) - \mathbb{P}_n X(f) > \frac{1}{2} r^2 \Big\} \overset{def}{=} E_m.$$

Let E be the event in Equation (23); using the above inclusion, we can claim

$$E \subseteq \cup_{m \in [M]} E_m.$$

At this point we can apply Lemma 3; rescaling δ coupled with the union bound now gives a bound on the original event

$$\mathbb{P}(E) \le \sum_{m \in [M]} \mathbb{P}(E_m) \le \delta/2.$$

In order to apply Lemma 3, several conditions must be met. The bound on the variance is ensured by Lemma 7; in addition, r must satisfy the following two critical relations for appropriate constants and for all $m \in [M]$

$$\mathbb{E} \sup_{f \in \widetilde{\mathcal{U}}_m} \Big\{ \mathbb{P}X(f) - \mathbb{P}_n X(f) \Big\} \lesssim r^2, \qquad \text{and} \qquad (K+1)\frac{\ln(1/(\delta r_\star))}{n} \lesssim r^2. \tag{29}$$

The condition on the left involves the Bellman operator $\mathcal{T}$ through $X(f)$. The requirement is relaxed in Lemma 4; we obtain that it is sufficient that r satisfies an inequality that involves the following local Rademacher averages:

$$\mathcal{R}_n \Big\{ L(f,f) - L(g_f, f) \mid \mathbb{P}X(f) \le 2r^2 \Big\} \lesssim r^2 \qquad \text{and} \qquad (K+1)\frac{\log_2(1/(\delta r_\star))}{n} \lesssim r^2. \tag{30}$$

If both conditions admit a smallest positive solution $r_\star$ such that Equation (30) holds for all $r \ge r_\star$ then we can cover all cases $m \in [M]$ with the condition $r \ge r_\star$ where $r_\star$ satisfies

$$\mathcal{R}_n \Big\{ L(f,f) - L(g_f, f) \mid \mathbb{P}X(f) \le 2r_\star^2 \Big\} \lesssim r_\star^2 \qquad \text{and} \qquad (K+1)\frac{\log_2(1/(\delta r_\star))}{n} \lesssim r_\star^2. \tag{31}$$

Since $r \ge r_\star$, when the inequalities in Equation (31) are satisfied, Equation (30) is automatically satisfied as well.

[7] We have $\mathbb{P}X(f) \le \|f - \mathcal{T}f\|_\mu^2 \le \|f - \mathcal{T}f\|_\infty^2 \lesssim 1$.

E.3.2. Establishing the claim in Equation (24)

Since $\mathbb{P}Y(g) \le 0$, it is sufficient to claim that we are unlikely to witness large deviations such as the one below[8]:

$$\exists g \in \mathcal{F} \text{ such that } \mathbb{P}_n Y(g) - \mathbb{P}Y(g) > -\frac{1}{2}\mathbb{P}Y(g) + \frac{1}{2}r_\star^2. \tag{32}$$

Let E be the above event; we show that E can occur with probability at most $\delta/2$. In the complement event, Equation (24) must hold.

We construct a family of sets $\{E_m\}$ such that

$$E \subseteq \cup_{m\in\{0,1,2,\dots,M\}} E_m$$

where each event E_m is described in the analysis to follow.

Small variance event Let us consider the set of functions with small variance

$$\mathcal{F}_0 = \{g \in \mathcal{F} \mid 0 \le -\mathbb{P}Y(g) \le \frac{1}{2}r_\star^2\}.$$

The associated event is

$$\begin{aligned}
&\left\{\exists g \in \mathcal{F}_0 \text{ such that } \mathbb{P}_n Y(g) - \mathbb{P}Y(g) > -\frac{1}{2}\mathbb{P}Y(g) + \frac{1}{2}r_\star^2\right\} \\
\subseteq &\left\{\exists g \in \mathcal{F}_0 \text{ such that } \mathbb{P}_n Y(g) - \mathbb{P}Y(g) > \frac{1}{2}r_\star^2\right\} \\
\overset{def}{=} &E_0.
\end{aligned}$$

Large variance events Consider the following partitioning to control the variance of the empirical process

$$\mathcal{F}_m = \{r^2 < -\mathbb{P}Y(g) \le 2r^2\}, \qquad \text{where } r^2 \overset{def}{=} 2^{m-2}r_\star^2 \ge \frac{1}{2}r_\star^2, \quad \text{for } m = 1,2,\dots,M.$$

The partition stops at $M \simeq \ln(1/r_\star)$ as $-\mathbb{P}Y(g) \lesssim 1$. The associated events are

$$\begin{aligned}
&\left\{\exists g \in \mathcal{F}_m \text{ such that } \mathbb{P}_n Y(g) - \mathbb{P}Y(g) \ge -\frac{1}{2}\mathbb{P}Y(g) + \frac{1}{2}r_\star^2\right\} \\
= &\left\{\exists g \in \mathcal{F}_m \text{ such that } \mathbb{P}_n Y(g) - \mathbb{P}Y(g) \ge -\frac{1}{2}\mathbb{P}Y(g)\right\} \\
\subseteq &\left\{\exists g \in \mathcal{F}_m \text{ such that } \mathbb{P}_n Y(g) - \mathbb{P}Y(g) \ge \frac{1}{2}r^2\right\} \\
\overset{def}{=} &E_m.
\end{aligned}$$

Putting together the pieces After rescaling δ to become $\delta/(2(M+1))$ and using the union bound we can finally apply Lemma 3 to bound the event in Equation (32)

$$\mathbb{P}(E) \le \sum_m \mathbb{P}(E_m) \le \delta/2,$$

In order to apply Lemma 3, we need to verify the assumptions in the statement of the lemma.

[8] Notice that Equation (32) is a bit weaker than Equation (23) due to the additional $r_\star$ term on the right hand side; this is due to the fact that we must also consider the small variance regime $-\mathbb{P}Y(g) \le r_\star^2$ in this section.

We have the following variance calculation reported in Lemma 8

$$\mathrm{Var}[Y(g)] \lesssim -\mathbb{P}Y(g), \qquad \text{for all } g \in \mathcal{F}. \tag{33}$$

By the symmetry of $\mathcal{F}$, every time Lemma 3 is invoked, for every $m = 1, 2, \dots, M$ the associated value for r must satisfy

$$\mathbb{E} \sup_{g \in \mathcal{F}_m} \Big\{ \mathbb{P}Y(g) - \mathbb{P}_n Y(g) \Big\} \lesssim r^2, \qquad \text{and} \qquad \frac{\ln(1/(\delta r_\star))}{n} \lesssim r^2.$$

The condition on the right is already in the final form; the one on the left involves the Bellman operator $\mathcal{T}$ through $Y(f)$. In order to obtain a bound that only depends on the class $\mathcal{F}$, the requirement is relaxed in Lemma 5. After the relaxation, we obtain that it is sufficient that r satisfies the inequalities

$$\mathcal{R}_n\big[\big(\mathcal{F} - f^\star\big)(r)\big] \lesssim r^2 \qquad \text{and} \qquad \frac{\ln(1/(\delta r_\star))}{n} \lesssim r^2.$$

E.3.3. Talagrand's Bound

In this section we assume that the prerequisites for using Talagrand are met in order to avoid measurability issues, and bound the supremum of an empirical process. Let W be a random variable on a certain probability space. For a given function class $\mathcal{H}$, define the supremum of the empirical process

$$Z = \sup_{h \in \mathcal{H}} \Big\{ \mathbb{P}h(W) - \mathbb{P}_n h(W) \Big\}.$$

We use Talagrand's bound (Talagrand, 1996) to derive a tail bound to Z when the variance of the process is tightly bracketed. (Here $\nu \geq 1$).

Lemma 3 (Talagrand's Bound with Bracketed Variance)**.** *The event*

$$T_1 \;:\quad Z > \frac{1}{2}r^2$$

occurs with probability at most δ if the following conditions are satisfied for appropriate universal constants

$$\mathrm{Var}[h(W)] \leq \nu \mathbb{P}h(W) \leq 2\nu r^2, \qquad \textit{and} \qquad \mathbb{E}Z \lesssim r^2, \qquad \textit{and} \qquad (\nu+1)\frac{\log(1/\delta)}{n} \lesssim r^2. \tag{34}$$

The strategy is to create a more 'natural' tail event T_2 associated to a variance-based concentration inequality. The concentration inequality will ensure that $\mathbb{P}(T_2) \leq \delta$. Then we show that the event T_1 is contained in T_2, ensuring $\mathbb{P}(T_1) \leq \mathbb{P}(T_2) \leq \delta$.

Talagrand's bound (see Thm 3.27 and Eq. 3.85 in the book (Wainwright, 2019)) applies to the tail event

$$T_2 \;:\quad Z \gtrsim \mathbb{E}Z + \sqrt{\frac{\sigma^2 \log(1/\delta)}{n}} + \frac{\log(1/\delta)}{n}. \tag{35}$$

where the variance-proxy of the process is

$$\sigma^2 \overset{def}{=} \sup_{h \in \mathcal{H}} \mathbb{P}\Big\{ h(W) - \mathbb{P}h(W) \Big\}^2 + 2\mathbb{P}Z.$$

It ensures that such large deviations have small probability of occurring

$$\mathbb{P}(T_2) \leq \delta. \tag{36}$$

We now proceed to showing that

$$T_1 = \Big\{ Z \geq \frac{1}{2}r^2 \Big\} \subseteq \Big\{ Z \geq \mathbb{E}Z + \sqrt{\frac{\sigma^2 \log(1/\delta)}{n}} + \frac{\log(1/\delta)}{n} \Big\} = T_2,$$

which allows us to conclude. In order to show the inclusion, we need to ensure that

$$\frac{1}{2}r^2 \geq \mathbb{E}Z + \sqrt{\frac{\sigma^2 \log(1/\delta)}{n}} + \frac{\log(1/\delta)}{n} \overset{def}{=} RHS.$$

We start from the above rhs and upper bound it until we obtain r^2. By combining the bound on the variance with the one on the expectation we obtain

$$\begin{aligned}
\sigma^2 &= \sup_{h \in \mathcal{H}} \mathbb{P}\Big\{ h(W) - \mathbb{P}h(W) \Big\}^2 + 2\mathbb{P}Z \\
&= \sup_{h \in \mathcal{H}} \mathrm{Var}[h(W)] + 2\mathbb{P}Z \\
&\leq \sup_{h \in \mathcal{H}} \nu \mathbb{P}h(W) + 2\mathbb{P}Z \\
&\leq 2\nu r^2 + 2\mathbb{P}Z \\
&\lesssim (\nu+1) r^2,
\end{aligned}$$

where the last step used Equation (34). This implies the upper bound

$$\begin{aligned}RHS &\le \mathbb{E}Z + r\sqrt{(\nu+1)\frac{\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{n} \\ &\le \frac{1}{2}r^2,\end{aligned}$$

where the last step used again Equation (34) for appropriately tuned numerical constants. Therefore, we have shown the inclusion $T_1 \subseteq T_2$; combined with the tail bound Equation (36) we obtain

$$\mathbb{P}(T_1) \le \mathbb{P}(T_2) \le \delta,$$

as claimed.

E.3.4. Simplifying the Rademacher Complexities

Lemma 4 (Rademacher Complexities for the X process)**.** *We have the relation*

$$\mathbb{E} \sup_{\mathbb{P}X(f) \le 2r^2} \Big\{ \mathbb{P}X(f) - \mathbb{P}_n X(f) \Big\} \lesssim \mathbb{E} \sup_{\mathbb{P}X(f) \le 2r^2} \frac{1}{n} \sum_i \epsilon_i [L(f,f) - L(g_f, f)].$$

First notice that we have

$$\|f - g_f\|_\mu \le \|f - \mathcal{T}f\|_\mu + \|g_f - \mathcal{T}f\|_\mu \le 2\|f - \mathcal{T}f\|_\mu,$$

and so using Lemma 2

$$\|f - g_f\|_\mu^2 \lesssim \frac{1}{1-\beta} \mathbb{P}X(f).$$

If instead $\mathcal{F}$ is convex, the Pythagoras' theorem ensures

$$\|f - g_f\|_\mu^2 \le \mathbb{P}X(f), \tag{37}$$

We handle both cases with

$$\|f - g_f\|_\mu^2 \lesssim K\mathbb{P}X(f). \tag{38}$$

It is useful to rewrite the left hand side in the statement of the lemma in a better form first by using a symmetrization argument, for which it is temporarily useful to emphasize the dependence on the random variable $G = (s, a, r, s^+)$

$$\begin{aligned} &= \mathbb{E} \sup_{\mathbb{P}X(f) \le 2r^2} \Big\{ \mathbb{P}X(f)(G) - \mathbb{P}_n X(f)(G_i) \Big\} \\ &= \mathbb{E} \sup_{\mathbb{P}X(f) \le 2r^2} \frac{1}{n} \sum_i \Big\{ \mathbb{P}X(f)(G) - X(f)(G_i) \Big\} \end{aligned}$$

Upon defining i.i.d. random variables $\widetilde{G}_i$ we can write

$$= \mathbb{E} \sup_{\mathbb{P}X(f) \le 2r^2} \frac{1}{n} \sum_i \Big\{ \mathbb{P}X(f)(\widetilde{G}_i) - X(f)(G_i) \Big\}.$$

Using Jensen's inequality we obtain

$$\le \mathbb{E} \sup_{\mathbb{P}X(f) \le 2r^2} \frac{1}{n} \sum_i \Big\{ X(f)(\widetilde{G}_i) - X(f)(G_i) \Big\}.$$

Now introduce the Rademacher random variables ϵ_i

$$\begin{aligned} &= \mathbb{E} \sup_{\mathbb{P}X(f) \le 2r^2} \frac{1}{n} \sum_i \epsilon_i \Big\{ X(f)(\widetilde{G}_i) - X(f)(G_i) \Big\} \\ &\le 2\mathbb{E} \sup_{\mathbb{P}X(f) \le 2r^2} \frac{1}{n} \sum_i \epsilon_i \Big\{ X(f)(G_i) \Big\} \\ &= 2\mathbb{E} \sup_{\mathbb{P}X(f) \le 2r^2} \frac{1}{n} \sum_i \epsilon_i [L(f,f) - L(g_f, f)]. \end{aligned}$$

The above argument is standard (see the textbook (Wainwright, 2019), chapter 4).

Lemma 5 (Rademacher Complexities for the Y process)**.** *We have the relation*

$$\mathbb{E} \sup_{-\mathbb{P}Y(g) \le 2r^2} \Big\{ \mathbb{P}Y(g) - \mathbb{P}_n Y(g) \Big\} \le \mathcal{R}_n\{\Delta \in \mathcal{F} - \mathcal{F}, \mid \|\Delta\|_\mu^2 \le 2r^2\}.$$

Notice that we have

$$\|g - f^\star\|_\mu^2 = -\mathbb{P}Y(g). \tag{39}$$

by definition of Y. It is useful to rewrite the left hand side in the statement of the lemma in a better form first by using a symmetrization argument; this step is analogous to that in Lemma 4, and thus here we report only the final bound:

$$\mathbb{E} \sup_{-\mathbb{P}Y(g)\le 2r^2} \Big\{\mathbb{P}Y(g) - \mathbb{P}_n Y(g)\Big\} \le 2\mathbb{E} \sup_{-\mathbb{P}Y(g)\le 2r^2} \frac{1}{n}\sum_i \epsilon_i [L(f^\star, f^\star) - L(g, f^\star)].$$

Using Equation (39) we obtain

$$= 2\mathbb{E} \sup_{\|g-f^\star\|_\mu^2\le 2r^2} \frac{1}{n}\sum_i \epsilon_i [L(f^\star, f^\star) - L(g, f^\star)].$$

Since

$$|L(f^\star, f^\star) - L(g, f^\star)| = |(f^\star - \mathrm{T}f^\star)^2 - (g - \mathrm{T}f^\star)^2| = |(f^\star - \mathrm{T}f^\star - g + \mathrm{T}f^\star)(f^\star - \mathrm{T}f^\star + g - \mathrm{T}f^\star)| \lesssim |f^\star - g|,$$

the Talagrand's contraction principle (see (Talagrand, 1996) or Thm A.6 in (Bartlett et al., 2005)) ensures

$$\le 2\mathbb{E} \sup_{\|g-f^\star\|_\mu^2\le 2r^2} \Big| \frac{1}{n}\sum_i \epsilon_i (g - f^\star)\Big|,$$

Thus we can re-write the above as

$$2\mathbb{E} \sup_{\Delta\in\mathcal{F}-f^\star,\ \|\Delta\|_\mu^2\le 2r^2} \Big| \frac{1}{n}\sum_i \epsilon_i \Delta\Big|,$$

which is the Rademacher complexity of the set

$$(\mathcal{F} - f^\star)(2r) \overset{def}{=} \Big\{\Delta \in \mathcal{F} - f^\star,\ \|\Delta\|_\mu^2 \le 2r^2\Big\}.$$

F. Technical Results

F.1. Proof of Proposition 2 (*Linearly Incomplete MDPs*)

Since $\mathcal{F}$ is linear, the projector Π onto $\mathcal{F}$ is a linear map. We can write

$$\begin{aligned} f - \mathcal{T}f &= f - \mathcal{T}f - (f^\star - \mathcal{T}f^\star) \\ &= (f - f^\star) - \gamma\mathcal{P}(f - f^\star) \\ &= (\mathcal{I} - \gamma\mathcal{P})(f - f^\star) \end{aligned}$$

and

$$\begin{aligned} g_f - \mathcal{T}f &= \Pi\mathcal{T}f - \mathcal{T}f \\ &= \Pi\mathcal{T}f - \mathcal{T}f - (\underbrace{\Pi\mathcal{T}f^\star}_{f^\star} - \mathcal{T}f^\star) \\ &= \gamma\Pi\mathcal{P}(f - f^\star) - \gamma\mathcal{P}(f - f^\star) \\ &= \gamma(\Pi - \mathcal{I})\mathcal{P}(f - f^\star). \end{aligned}$$

Notice that $A = \gamma(\Pi - \mathcal{I})\mathcal{P}$ and $B = (\mathcal{I} - \gamma\mathcal{P})$ are both linear operators. If we denote with $\Delta = f - f^\star$ the increments, we have

$$\mathcal{I}(r) = \sup_{\Delta \in \mathcal{F}_{\text{lin}},\ \|B\Delta\|_\mu \le r} \|A\Delta\|_\mu. \tag{40}$$

Fix $r > 0$ and let β satisfy $\mathcal{I}(r) = \beta r$ for that specific value of r. Now, consider any other radius $r' > r$; it must be representable as $r' = cr$ for some constant $c > 1$. Then the function $\Delta' = c\Delta \in \mathcal{F}_{\text{lin}}$ is feasible for the program below if Δ is feasible for the one in Equation (40)

$$\mathcal{I}(cr) = \sup_{\Delta \in \mathcal{F}_{\text{lin}},\ \|B\Delta\|_\mu \le cr} \|A\Delta\|_\mu. \tag{41}$$

This implies $\mathcal{I}(cr) \ge \beta cr$. Now assume that the inequality is strict to derive a contradiction. That is, assume $\mathcal{I}(cr) > \beta cr$ and let Δ' be a maximizer of Equation (41). Then the function $\Delta = \Delta'/c$ is feasible for Equation (40) and it gives $\mathcal{I}(r) > \beta r$, contradiction, because we assumed $\mathcal{I}(r) = \beta r$. Therefore we must have $\mathcal{I}(r') = \mathcal{I}(cr) = \beta cr = \beta r'$ for any $r' > r$. Since r is arbitrary, and $\mathcal{I}(0) = 0$ follows from Proposition 1, the proof is complete.

Lemma 6 (Expectation of the Single Cost)**.**

$$\mathbb{E}L(g, f) = \mathcal{L}(g, f) + \mathbb{E}_{(s,a)\sim\mu} \operatorname*{Var}_{r\sim R(s,a),\ s^+\sim\mathcal{P}(s,a)} \Big[(\mathrm{T}f)(r, s^+)\Big].$$

Proof. Recall the definition of Bellman backup T. By some algebra steps we have

$$\begin{aligned}
\mathbb{E}[L(g,f)] &= \mathbb{E}\Big(g(s,a) - r - \gamma f(s^+,\pi)\Big)^2 \\
&= \mathbb{E}\Big(g(s,a) - (\mathcal{T}f)(s,a) + (\mathcal{T}f)(s,a) - \underbrace{[r + \gamma f(s^+,\pi)]}_{=(\mathrm{T}f)(r,s^+)}\Big)^2 \\
&= \mathbb{E}\Bigg\{\Big(g(s,a) - (\mathcal{T}f)(s,a)\Big)^2 \\
&+ 2\Big(g(s,a) - (\mathcal{T}f)(s,a)\Big)\Big((\mathcal{T}f)(s,a) - \mathrm{T}f(r,s^+)\Big) \\
&+ \Big((\mathcal{T}f)(s,a) - \mathrm{T}f(r,s^+)\Big)^2\Bigg\} \\
&= \mathcal{L}(g,f) \\
&+ 2\mathbb{E}_{(s,a)\sim\mu}\Big[\Big(g(s,a) - (\mathcal{T}f)(s,a)\Big)\underbrace{\mathbb{E}_{r\sim R(s,a),\, s^+\sim\mathcal{P}_\pi(s,a)}\Big((\mathcal{T}f)(s,a) - \mathrm{T}f(r,s^+)\Big)}_{=0}\Big] \\
&+ \mathbb{E}_{(s,a)\sim\mu}\mathbb{E}_{r\sim R(s,a),\, s^+\sim\mathcal{P}(s,a)}\Big[(\mathcal{T}f)(s,a) - \mathrm{T}f(r,s^+)\Big]^2 \\
&= \mathcal{L}(g,f) + \mathbb{E}_{(s,a)\sim\mu}\operatorname*{Var}_{r\sim R(s,a),\, s^+\sim\mathcal{P}(s,a)}\Big[\mathrm{T}f(r,s^+)\Big]
\end{aligned}$$

□

F.2. Variance Bounds

Lemma 7 (Variance of the X-process)**.** *We have the following bound on the variance*

$$\mathrm{Var}[X(f)] \lesssim \frac{1}{1-\beta}\mathbb{P}X(f).$$

In addition, when $\mathcal{F}$ is convex then we have the tighter inequality

$$\mathrm{Var}[X(f)] \lesssim \mathbb{P}X(f).$$

Proof.

$$\begin{aligned}
\mathrm{Var}[X(f)] &= \mathrm{Var}\left[L(f,f) - L(g_f,f)\right] \\
&\le \mathbb{E}\left[L(f,f) - L(g_f,f)\right]^2 \\
&= \mathbb{E}\Big[(f-g_f)^2(f - \mathrm{T}f + g_f - \mathrm{T}f)^2\Big] \\
&\lesssim \mathbb{E}(f-g_f)^2 \\
&= \|f-g_f\|_\mu^2.
\end{aligned}$$

When $\mathcal{F}$ is convex Pythagoras' theorem ensures

$$\begin{aligned}
\|f-g_f\|_\mu^2 &\le \|f-\mathcal{T}f\|_\mu^2 - \|g_f-\mathcal{T}f\|_\mu^2 \\
&= \mathbb{P}X(f).
\end{aligned}$$

Otherwise, for arbitrary $\mathcal{F}$ we have the bound

$$\begin{aligned}
\|f-g_f\|_\mu &= \|f-\mathcal{T}f+\mathcal{T}f-g_f\|_\mu \\
&\le \|f-\mathcal{T}f\|_\mu + \|g_f-\mathcal{T}f\|_\mu \\
&\le 2\|f-\mathcal{T}f\|_\mu.
\end{aligned}$$

Coupled with Lemma 2, we obtain the bound

$$\|f - g_f\|_\mu^2 \lesssim \|f - \mathcal{T} f\|_\mu^2 \lesssim \frac{1}{1-\beta} \mathbb{P} X(f).$$

□

Lemma 8 (Variance of the Y-process)**.** *For any $g \in \mathcal{F}$ we have the bound*

$$\mathrm{Var}[Y(g)] \leq -\mathbb{P} Y(g).$$

Proof.

$$\begin{aligned}
\mathrm{Var}[Y(g)] &= \mathrm{Var}\left[L(f^\star, f^\star) - L(g, f^\star)\right] \\
&\leq \mathbb{E}\big[L(f^\star, f^\star) - L(g, f^\star)\big]^2 \\
&= \mathbb{E}\Big[(f^\star - g)^2 (f^\star - \mathrm{T} f^\star + g - \mathrm{T} f^\star)^2\Big] \\
&\lesssim \mathbb{E}(f^\star - g)^2 \\
&= \|f^\star - g\|_\mu^2 \\
&= \|\mathcal{T} f^\star - g\|_\mu^2 \\
&= \mathcal{L}(g, f^\star) \\
&= \mathcal{L}(g, f^\star) - \underbrace{\mathcal{L}(f^\star, f^\star)}_{=0} \\
&= -\mathbb{P} Y(g).
\end{aligned}$$

□