Mitigating Covariate Shift in Misspecified Regression with Applications to Reinforcement Learning

Philip Amortila PHILIPA4@ILLINOIS.EDU
University of Illinois, Urbana-Champaign

Tongyi Cao TCAO@CS.UMASS.EDU
University of Massachusetts, Amherst

Akshay Krishnamurthy AKSHAYKR@MICROSOFT.COM
Microsoft Research, NYC

Editors: Shipra Agrawal and Aaron Roth

Abstract

A pervasive phenomenon in machine learning applications is *distribution shift*, where training and deployment conditions for a machine learning model differ. As distribution shift typically results in a degradation in performance, much attention has been devoted to algorithmic interventions that mitigate these detrimental effects. This paper studies the effect of distribution shift in the presence of model misspecification, specifically focusing on L_∞-misspecified regression and *adversarial covariate shift*, where the regression target remains fixed while the covariate distribution changes arbitrarily. We show that empirical risk minimization, or standard least squares regression, can result in undesirable *misspecification amplification* where the error due to misspecification is amplified by the density ratio between the training and testing distributions. As our main result, we develop a new algorithm—inspired by robust optimization techniques—that avoids this undesirable behavior, resulting in no misspecification amplification while still obtaining optimal statistical rates. As applications, we use this regression procedure to obtain new guarantees in offline and online reinforcement learning with misspecification and establish new separations between previously studied structural conditions and notions of coverage.

1. Introduction

A majority of machine learning methods are developed and analyzed under the idealized setting where the training conditions accurately reflect those at deployment. Yet, almost all practical applications exhibit *distribution shift*, where these conditions differ significantly. Distribution shift can occur for a plethora of reasons, ranging from quirks in data collection (Recht et al., 2019), to temporal drift (Gama et al., 2014; Besbes et al., 2015), to users adapting to an ML model (Perdomo et al., 2020), and it typically results in a degradation in model performance. Due to the prevalence of this phenomenon and the diversity of applications where it manifests, there is a vast and ever-growing body of literature studying algorithmic interventions to mitigate distribution shift (Quinonero-Candela et al., 2008; Sugiyama and Kawanabe, 2012).

Covariate shift is perhaps the most basic form of distribution shift. Covariate shift is pertinent to supervised learning—where the goal is to predict a label Y from covariates X—and posits a change in the distribution over covariates while keeping the target predictor fixed. This setup, in particular that the target does not change, is natural in applications including neural algorithmic reasoning (Anil et al., 2022; Zhang et al., 2022; Liu et al., 2023), reinforcement learning (Ross et al., 2011; Levine et al., 2020), and computer vision (Koh et al., 2021; Recht et al., 2019; Miller et al., 2021). It is

well known that one can adapt guarantees from statistical learning to the covariate shift setting; specifically, for well-specified regression, a classical density-ratio argument shows that empirical risk minimization (ERM) is consistent under suitably well-behaved covariate shifts.

One stipulation of this consistency guarantee is that the model/hypothesis class be *well-specified* (also referred to as *realizable*). Although statistical learning theory offers a rather complete understanding of misspecification in the absence of covariate shift (via agnostic learning and excess risk bounds), our understanding of how covariate shift can adversely interact with model misspecification remains fairly immature. This interaction is the focus of the present paper.

1.1. Contributions

We study regression under *adversarial covariate shift* where we receive regression samples from a distribution $\mathcal{D}_{\mathrm{train}}$ but are evaluated on an arbitrary distribution $\mathcal{D}_{\mathrm{test}}$ for which no prior knowledge is available; we only assume that the distributions share the same target regression function $f^\star$ and that the worst-case density ratio of the covariate marginals is bounded by $C_\infty \in [1, \infty)$ (formally defined in Section 2). As inductive bias, we have a function class $\mathcal{F}$ of predictors and assume L_∞-*misspecification*: there exists a predictor $\bar{f} \in \mathcal{F}$ that is pointwise close to $f^\star$, i.e., $\|\bar{f} - f^\star\|_\infty \leq \varepsilon_\infty$. This notion is natural for the covariate shift setting because it ensures that $\bar{f}$ has low prediction error on both $\mathcal{D}_{\mathrm{train}}$ and any $\mathcal{D}_{\mathrm{test}}$.

In this setup we obtain the following results:

1. We show that standard empirical risk minimization (ERM) is not robust to covariate shift in the presence of misspecification. Precisely, even in the limit of infinite data, ERM over $\mathcal{F}$ can incur squared prediction error under $\mathcal{D}_{\mathrm{test}}$ scaling as $\Omega(C_\infty \varepsilon_\infty^2)$. Meanwhile the error of the L_∞-misspecified predictor $\bar{f}$ is at most ε_∞^2. We call this phenomenon—where the misspecification error is scaled by the density ratio coefficient (despite there being a predictor avoiding this scaling)—*misspecification amplification*.
2. As our main result, we give a new algorithm, called disagreement-based regression (DBR), that avoids *misspecification amplification* and is therefore robust to adversarial covariate shift under misspecification. DBR has asymptotic prediction error under $\mathcal{D}_{\mathrm{test}}$ scaling as $O(\varepsilon_\infty^2)$, with no dependence on the density ratio coefficient C_∞. At the same time, it has order-optimal finite sample behavior recovering standard "fast rate" guarantees for the well-specified setting, and can be extended to adapt to unknown misspecification level (as shown in Appendix B.4). To our knowledge, this is the first result avoiding misspecification amplification in the adversarial covariate shift setting. Our assumptions—particularly that no information about $\mathcal{D}_{\mathrm{test}}$ is available and that $\mathcal{F}$ is unstructured—rule out prior approaches based on density ratios (Shimodaira, 2000; Duchi and Namkoong, 2021) or sup-norm convergence (Schmidt-Hieber and Zamolodtchikov, 2022); see Appendix A for further discussion.

To demonstrate the utility of disagreement-based regression, we deploy the procedure in value function approximation settings in reinforcement learning (RL), where regression is a standard primitive and mitigating the adverse effects of distribution shift is a central challenge. Here, using DBR as a drop-in replacement for ERM when fitting Bellman backups, we obtain the following results:

1. In the offline RL setting, we instantiate the minimax algorithm of Chen and Jiang (2019) with DBR and show that, under L_∞-misspecification and with coverage measured via the

concentrability coefficient, misspecification amplification can be avoided when learning a near optimal policy. In contrast, prior lower bounds imply that misspecification amplification is unavoidable when coverage is measured via Bellman transfer coefficients (Du et al., 2020; Van Roy and Dong, 2019; Lattimore et al., 2020). Our result therefore establishes a new separation between concentrability and Bellman transfer coefficients.

2. In the online RL setting, we instantiate the GOLF algorithm of Jin et al. (2021) with DBR and obtain analogous results under the structural condition of *coverability* (building on the analysis of Xie et al. (2023)). Taken with the above lower bounds (Du et al., 2020; Van Roy and Dong, 2019; Lattimore et al., 2020), this separates structural conditions involving Bellman errors (e.g., Bellman rank (Jiang et al., 2017), Bellman-eluder dimension (Jin et al., 2021), or sequential extrapolation coefficient (Xie et al., 2023)) from coverability, which does not.

To keep the presentation concise and focused on the interaction between covariate shift and misspecification, we focus on the simplest settings that manifest misspecification amplification. In Section 5, we discuss a number of directions for future work, which include extensions to the core technical and algorithmic results. We defer a detailed discussion of related work to Appendix A.

2. Misspecified regression under distribution shift

We begin by introducing the formal problem setting and our assumptions. Most proofs for results in this section are deferred to Appendix B. There are two joint distributions, called $\mathcal{D}_{\mathrm{train}}$ and $\mathcal{D}_{\mathrm{test}}$, over $\mathcal{X} \times \mathbb{R}$ where $\mathcal{X}$ is a covariate space. We use $\mathbb{P}_{\mathrm{train}}, \mathbb{P}_{\mathrm{test}}$ and $\mathbb{E}_{\mathrm{train}}, \mathbb{E}_{\mathrm{test}}$ to denote the probability law and expectation under these distributions. We hypothesize that $\mathcal{D}_{\mathrm{train}}$ and $\mathcal{D}_{\mathrm{test}}$ share the same *Bayes regression function*, an assumption referred to as covariate shift in the literature (Shimodaira, 2000).

Assumption 2.1 (Covariate shift)**.** *For all $x \in \mathcal{X}$ we have*

$$\mathbb{E}_{\mathrm{train}}[y \mid x] = \mathbb{E}_{\mathrm{test}}[y \mid x].$$

Let $f^\star : x \mapsto \mathbb{E}_{\mathrm{train}}[y \mid x]$ denote the shared Bayes regression function. We posit that the marginal distributions over $\mathcal{X}$ are absolutely continuous with respect to a reference measure and use d_{train} and d_{test} to denote the corresponding marginal densities. We assume these are related via the following density ratio assumption.

Assumption 2.2 (Bounded density ratios)**.** *The density ratio*

$$C_\infty := \sup_{x \in \mathcal{X}} \left| \frac{d_{\mathrm{test}}(x)}{d_{\mathrm{train}}(x)} \right|$$

is bounded, i.e., $C_\infty < \infty$.

Note that $C_\infty \geq 1$ always. Boundedness of density ratios is standard in the covariate shift literature; indeed the coefficient C_∞ appears in the classical covariate shift analyses as well as in many algorithmic interventions (Shimodaira, 2000; Sugiyama et al., 2007). Beyond satisfying these assumption, $\mathcal{D}_{\mathrm{test}}$ can be adaptively and adversarially chosen. In particular, no information about $\mathcal{D}_{\mathrm{test}}$, such as labeled/unlabeled samples or other inductive bias, is available.

We have a dataset $\{(x_i, y_i)\}_{i=1}^n$ of n i.i.d. labeled examples sampled from $\mathcal{D}_{\text{train}}$ and a function class $\mathcal{F} \subset (\mathcal{X} \to \mathbb{R})$ of predictors. We define the (squared) *prediction errors*

$$R_{\text{train}}(f) := \mathbb{E}_{\text{train}}\big[(f(x) - f^\star(x))^2\big], \quad \text{and} \quad R_{\text{test}}(f) := \mathbb{E}_{\text{test}}\big[(f(x) - f^\star(x))^2\big]. \tag{1}$$

We seek to use the dataset to find a predictor $\hat{f}$ for which $R_{\text{test}}(\hat{f})$ is small.

We make two assumptions on $\mathcal{F}$: we assume that $|\mathcal{F}| < \infty$ and that $\mathcal{F}$ is L_∞-misspecified.

Assumption 2.3 (L_∞-misspecification)**.** *For some $\varepsilon_\infty \geq 0$, there exists $\bar{f} \in \mathcal{F}$ with*

$$\big\|\bar{f} - f^\star\big\|_\infty \leq \varepsilon_\infty, \quad \text{where} \quad \|f\|_\infty := \sup_{x \in \mathcal{X}} |f(x)|.$$

Most prior analyses for regression under covariate shift assume that the model class $\mathcal{F}$ is well-specified, i.e., that $\varepsilon_\infty = 0$ so that $f^\star \in \mathcal{F}$. L_∞-misspecification provides a relaxation that is natural for at least two reasons. First, it enables end-to-end learning guarantees via composition with approximation-theoretic results for specific function classes (e.g., neural networks), where it is standard to measure approximation via the L_∞ norm (Telgarsky, 2021). More importantly, L_∞-misspecification is particularly apt in the covariate shift setting because it ensures that $\bar{f}$ has low prediction error on both $\mathcal{D}_{\text{test}}$ and $\mathcal{D}_{\text{train}}$. Thus, there is at least one high-quality predictor whose performance is stable across distributions. In contrast, we have no such guarantee if we, for example, measure misspecification with respect to other norms (which depend on the distribution) or consider the agnostic setting (with no quantified misspecification assumption). Indeed, we will see below that misspecification amplification is unavoidable in such cases.

We also make the following technical assumption.

Assumption 2.4 (Boundedness)**.** $\sup_{f \in \mathcal{F}} \|f\|_\infty \leq 1$ *and* $|y| \leq 1$ *a.s. under* $\mathcal{D}_{\text{train}}$ *and* $\mathcal{D}_{\text{test}}$.

We impose Assumption 2.4 and that $|\mathcal{F}| < \infty$ solely to highlight the novel algorithmic and technical aspects; we expect that relaxing these assumptions is possible.

2.1. Misspecification amplification for empirical risk minimization

When there is no prior knowledge about or data from $\mathcal{D}_{\text{test}}$, perhaps the most natural algorithm for optimizing $R_{\text{test}}(\cdot)$ is empirical risk minimization (ERM) on the data from the training distribution:

$$\hat{f}^{(n)}_{\text{ERM}} := \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2.$$

A standard uniform convergence argument yields the classical covariate shift guarantee for ERM:

Proposition 2.1 (ERM upper bound)**.** *For any $\delta \in (0, 1)$ with probability at least $1 - \delta$,* ERM *satisfies*

$$R_{\text{test}}(\hat{f}^{(n)}_{\text{ERM}}) \leq O\left(C_\infty \varepsilon_\infty^2 + C_\infty \frac{\log(|\mathcal{F}|/\delta)}{n}\right).$$

The second term which scales as $1/n$—the statistical term—is optimal in the generality of our setup (Ma et al., 2023; Ge et al., 2023), the interpretation being that the effective sample size is reduced by a factor of C_∞ due to the mismatch between $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$. The first term—the misspecification

term—represents the asymptotic[1] test error of ERM and demonstrates a phenomenon that we call *misspecification amplification*, whereby the error due to misspecification is amplified by the density ratio coefficient. This phenomenon is simultaneously more concerning and less intuitive than the degradation of the statistical term, because it describes an error which does not decay with larger sample sizes and because $\bar{f} \in \mathcal{F}$ has $R_{\text{test}}(\bar{f}) = \varepsilon_\infty^2$. Since $\mathcal{F}$ contains a predictor that does not incur misspecification amplification, one might hope that misspecification amplification can be avoided.

Our first main result is that misspecification amplification *cannot* be avoided by ERM in the worst case. The result is proved in the asymptotic regime, where ERM is equivalent to the $L_2(\mathcal{D}_{\text{train}})$-projection of $f^\star$ onto the function class $\mathcal{F}$, defined as

$$\hat{f}_{\text{ERM}}^{(\infty)} \in \arg\min_{f \in \mathcal{F}} \|f - f^\star\|_{L_2(\mathcal{D}_{\text{train}})}^2, \qquad \text{with} \qquad \|g\|_{L_2(\mathcal{D}_{\text{train}})}^2 := \mathbb{E}_{\text{train}}\left[g(x)^2\right].$$

The next proposition shows that $\hat{f}_{\text{ERM}}^{(\infty)}$ can incur misspecification amplification.

Proposition 2.2 (ERM lower bound). *For all $\varepsilon_\infty \in (0,1)$ and $C_\infty \in [1,\infty)$ such that $\sqrt{C_\infty} \cdot \varepsilon_\infty \leq 1/2$, and for all $\zeta > 0$ sufficiently small, there exist distributions $\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}$ and a function class $\mathcal{F}$ with $|\mathcal{F}| = 2$ satisfying Assumption 2.1-Assumption 2.4 (with parameters $\varepsilon_\infty, C_\infty$) such that*

$$R_{\text{test}}(\hat{f}_{\text{ERM}}^{(\infty)}) = C_\infty \varepsilon_\infty^2 - \zeta.$$

Combined with the optimality of the statistical term (Ma et al., 2023; Ge et al., 2023), this establishes that Proposition 2.1 characterizes the behavior of ERM under L_∞-misspecification and covariate shift. The construction is based on the following insight, visualized in Figure 1. The fact that $\bar{f}$ is L_∞-close to $f^\star$ guarantees that its prediction errors are "spread out" across the domain $\mathcal{X}$. Since $\bar{f} \in \mathcal{F}$, we know that $\hat{f}_{\text{ERM}}^{(\infty)}$ must satisfy $\|\hat{f}_{\text{ERM}}^{(\infty)} - f^\star\|_{L_2(\mathcal{D}_{\text{train}})}^2 \leq \varepsilon_\infty^2$. Unfortunately, this property does not guarantee that the errors of $\hat{f}_{\text{ERM}}^{(\infty)}$ are "spread out" in a similar manner to $\bar{f}$'s. Indeed, we construct a predictor f_{bad} that concentrates its errors on a region of $\mathcal{X}$ that is amplified by $\mathcal{D}_{\text{test}}$ and makes up for this by having zero error elsewhere. By setting the parameters carefully, we can ensure that this bad predictor is chosen by ERM.

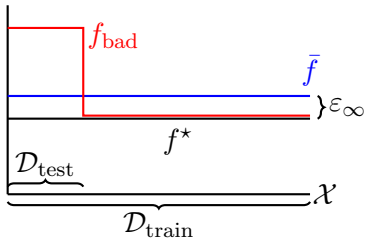

Figure 1: The construction used to prove Proposition 2.2. f_{bad} and $\bar{f}$ have equal risk under $\mathcal{D}_{\text{train}}$ but f_{bad} concentrates errors onto $\mathcal{D}_{\text{test}}$.

We note that essentially the same construction shows that, under the weaker notion of $L_2(\mathcal{D}_{\text{train}})$-misspecification, amplification is unavoidable for any proper learner (which outputs a function in $\mathcal{F}$). Indeed, in Figure 1, the function class $\{f_{\text{bad}}\}$ is $L_2(\mathcal{D}_{\text{train}})$-misspecified but f_{bad} has much higher error on $\mathcal{D}_{\text{test}}$.

Other existing algorithms. Proposition 2.2 only pertains to ERM, and thus, one might ask whether other algorithms can avoid misspecification amplification. Before turning to our positive results in the next section, we briefly note that other standard algorithms (that do not require knowledge of $\mathcal{D}_{\text{test}}$) either incur misspecification amplification to some degree, or have some other failure mode. This pertains to the star algorithm (Audibert, 2007; Liang et al., 2015), other aggregation

1. We consider the asymptotic regime where $n \to \infty$ with all other quantities, like $\log|\mathcal{F}|$ and ε_∞, fixed.

schemes (c.f., Lecué and Rigollet, 2014), and L_∞-regression (Knight, 2017; Yi and Neykov, 2024), as we discuss in Appendix B.2. Several methods for mitigating covariate shift can avoid misspecification amplification, but either require knowledge of $\mathcal{D}_{\text{test}}$ or structural assumptions on $\mathcal{F}$; see Appendix A.

2.2. Main result: Disagreement-based regression

In this section, we provide a new algorithm that avoids misspecification amplification while requiring no knowledge of $\mathcal{D}_{\text{test}}$ and recovering optimal statistical rates. To develop some intuition, observe that in the construction in Figure 1, the only way for the bad predictor (f_{bad}, in red) to be chosen by ERM *and* have large errors on $\mathcal{D}_{\text{test}}$ is for it to have much lower error than $\bar{f}$ on the rest of the domain. Indeed, if we could filter out the points where f_{bad}'s error is less than $\bar{f}$'s, then f_{bad} cannot overcome the large errors on $\mathcal{D}_{\text{test}}$. Stated another way, we can avoid misspecification amplification in this example if we restrict the regression problem to the region where $|f_{\text{bad}}(x) - f^\star(x)| \geq |\bar{f}(x) - f^\star(x)|$.

Generalizing this insight to a larger function class suggests that, when considering a candidate $f \in \mathcal{F}$, we should only measure the square loss for f on the region where $|f(x) - f^\star(x)| \geq |\bar{f}(x) - f^\star(x)|$. Unfortunately, this region depends on $f^\star$ and $\bar{f}$, both of which are unknown. Nevertheless, our approach is based on this intuition, and we avoid the dependence on these unknown functions with two algorithmic ideas.

To eliminate the dependence on $f^\star$, we use the fact that $|\bar{f}(x) - f^\star(x)| \leq \varepsilon_\infty$ and approximate the above region with $I_f := \{x : |f(x) - \bar{f}(x)| \geq c\varepsilon_\infty\}$. Indeed for $c \geq 2$,

$$\{x : |f(x) - \bar{f}(x)| \geq c\varepsilon_\infty\} \subseteq \{x : |f(x) - f^\star(x)| \geq |\bar{f}(x) - f^\star(x)|\}.$$

On the other hand, we know that $|f(x) - f^\star(x)| \leq (c+1)\varepsilon_\infty$ in the complementary region, I_f^C. This is, up to the constant factor, the best pointwise guarantee we can attain, making it safe to ignore the complementary region. This resolves the first issue of dependence on $f^\star$.

To address the dependence on $\bar{f}$, we use that $\bar{f} \in \mathcal{F}$ and formulate a robust optimization objective that implicitly considers all possible pairwise "disagreement regions." Formally, with $W^\tau_{f,g}(x) := \mathbb{1}\{|f(x) - g(x)| \geq \tau\}$ the algorithm is:

$$\hat{f}^{(n)}_{\mathsf{DBR}} \leftarrow \arg\min_{f \in \mathcal{F}} \max_{g \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n W^\tau_{f,g}(x)\big\{(f(x) - y)^2 - (g(x) - y)^2\big\}. \tag{2}$$

We call this algorithm *disagreement-based regression* (DBR) and keep the dependence on τ implicit in the notation for the solution $\hat{f}^{(n)}_{\mathsf{DBR}}$.[2] There are essentially three key ingredients. First, we introduce the "filter" $W^\tau_{f,g}$ to restrict the regression problem to the set of points where the predictions of f and g differ considerably, which we call the *disagreement region*. This formalizes the intuition that we should only measure the square loss for f on points where $|f(x) - \bar{f}(x)| \geq c\varepsilon_\infty$. Second is the robust optimization approach, where for each $f \in \mathcal{F}$, we consider all possible choices $g \in \mathcal{F}$ for filtering, which allows us to take g to be L_∞-close to $f^\star$ in the analysis. Finally, we measure the square loss *regret* in the disagreement region, by subtracting off the square loss of the comparator

2. The name stems from the literature on disagreement-based active learning (Hanneke, 2014), where a similar "range" computation has appeared (Krishnamurthy et al., 2019; Foster et al., 2018, 2021). However our usage is conceptually unrelated: we use disagreement for robustness to covariate shift, while, in active learning, disagreement is used to reduce sample complexity.

function g. Similar to Agarwal and Zhang (2022), this accounts for the fact that each $g \in \mathcal{F}$ yields a different regression problem, with potentially different Bayes error rates.[3]

As our main theorem, we show that disagreement-based regression enjoys the following guarantee.

Theorem 2.1 (Main result for DBR). *Fix $\delta \in (0,1)$. Let $\mathcal{F}$ be a function class with $|\mathcal{F}| < \infty$ satisfying Assumption 2.3 and Assumption 2.4. Then with probability at least $1-\delta$, $\hat{f}^{(n)}_{\mathsf{DBR}}$ with $\tau \geq 3\varepsilon_\infty$ satisfies*

$$\mathbb{E}_{\mathrm{train}}\left[\mathbb{1}\left\{|\hat{f}^{(n)}_{\mathsf{DBR}}(x) - f^\star(x)| \geq \tau + \varepsilon_\infty\right\} \cdot \left\{(\hat{f}^{(n)}_{\mathsf{DBR}}(x) - f^\star(x))^2 - \varepsilon_\infty^2\right\}\right] \leq \frac{160\log(2|\mathcal{F}|/\delta)}{3n}, \tag{3}$$

which directly implies

$$\mathbb{P}_{\mathrm{train}}\left[\left|\hat{f}^{(n)}_{\mathsf{DBR}}(x) - f^\star(x)\right| \geq \tau + \varepsilon_\infty\right] \leq \frac{160\log(2|\mathcal{F}|/\delta)}{3n(\tau^2 + 2\tau\varepsilon_\infty)}. \tag{4}$$

Before turning to a discussion of Theorem 2.1 we state two immediate corollaries. The first addresses the adversarial covariate shift setting, bounding the risk of $\hat{f}^{(n)}_{\mathsf{DBR}}$ under $\mathcal{D}_{\mathrm{test}}$.

Corollary 2.1 (Covariate shift for DBR). *Fix $\delta \in (0,1)$. Under Assumption 2.1–Assumption 2.4, with probability at least $1-\delta$, $\hat{f}^{(n)}_{\mathsf{DBR}}$ with $\tau = 3\varepsilon_\infty$ satisfies*

$$R_{\mathrm{test}}(\hat{f}^{(n)}_{\mathsf{DBR}}) \leq 17\varepsilon_\infty^2 + O\left(C_\infty \frac{\log(|\mathcal{F}|/\delta)}{n}\right). \tag{5}$$

The next result shows that $\hat{f}^{(n)}_{\mathsf{DBR}}$ also recovers the optimal guarantee in the well-specified case, i.e., when $\varepsilon_\infty = 0$.

Corollary 2.2 (Well-specified case). *Fix $\delta \in (0,1)$. Under Assumption 2.1–Assumption 2.4 (with $\varepsilon_\infty = 0$), with probability at least $1-\delta$, $\hat{f}^{(n)}_{\mathsf{DBR}}$ with $\tau \leq O\left(\sqrt{\log(|\mathcal{F}|/\delta)/n}\right)$ satisfies*

$$R_{\mathrm{train}}(\hat{f}^{(n)}_{\mathsf{DBR}}) \leq O\left(\frac{\log(|\mathcal{F}|/\delta)}{n}\right) \quad \text{and} \quad R_{\mathrm{test}}(\hat{f}^{(n)}_{\mathsf{DBR}}) \leq O\left(C_\infty \frac{\log(|\mathcal{F}|/\delta)}{n}\right). \tag{6}$$

We now turn to some remarks regarding Theorem 2.1 and the corollaries.

DBR avoids misspecification amplification Comparing Corollary 2.1 in the $n \to \infty$ limit with Proposition 2.2 highlights the main qualitative difference between DBR and ERM. DBR attains $O(\varepsilon_\infty^2)$ asymptotic test error while the test error for ERM is lower bounded by $\Omega(C_\infty \varepsilon_\infty^2)$. In other words, DBR avoids misspecification amplification while ERM does not. At the same time, the statistical term is identical (up to constants) to that of ERM, enabling us to recover the optimal rate in the well-specified case.

3. More directly, the probability mass of filtered points $\mathbb{P}_{\mathrm{train}}[W^\tau_{f,g}(x)]$ could vary considerably for different $f, g \in \mathcal{F}$.

Quantile guarantee Taking $\tau = O(\varepsilon_\infty)$ in Eq. (3), we have that $\mathbb{P}_{\text{train}}[|\hat{f}^{(n)}_{\mathsf{DBR}}(x) - f^\star(x)| \geq c\varepsilon_\infty] \lesssim 1/n\varepsilon_\infty^2$, which controls the large quantiles of the prediction error. This is reminiscent of what can be achieved by applying Markov's inequality to the guarantee for ERM in the well-specified case. In contrast, ERM only ensures that $R_{\text{train}}(\hat{f}^{(n)}_{\mathsf{ERM}}) = \Omega(\varepsilon_\infty^2)$ under misspecification, which does not imply any meaningful quantile guarantee. One interpretation of our results is that, although such quantile guarantees are not possible for ERM under misspecification, there is no information-theoretic obstruction. We also note that these quantile guarantees are rather different from sup-norm convergence; see Appendix A for further discussion.

Computational efficiency DBR, as described in Eq. (2), does not appear to be computationally tractable, primarily due to the non-smoothness and non-convexity introduced by the filter $W_{f,g}$. A natural direction for future work is to understand the computational challenges involved in avoiding misspecification amplification.

2.2.1. EXTENSIONS

Before closing this section, we mention two extensions that we defer to Appendix B.4.

- *Approximation factor.* The approximation factor of 17 in Corollary 2.1 can be improved to 10 (cf. Proposition B.1); however our approach for doing so degrades the convergence rate of the statistical term. We do not know the optimal approximation factor for this setting or whether there is an inherent trade-off between the statistical term and the approximation/misspecification term.
- *Adapting to unknown misspecification.* Theorem 2.1 requires setting $\tau \geq 3\varepsilon_\infty$ which can always be achieved by setting τ sufficiently large. However, setting $\tau = O(\varepsilon_\infty)$ yields the best guarantee, and so, we would like to choose τ in a data-dependent fashion to adapt to the misspecification level. Proposition B.2 shows that this can be done while recovering essentially the same guarantee as in Theorem 2.1.

3. Proof of Theorem 2.1

This section contains the proof of Theorem 2.1—which we emphasize only requires elementary arguments—and is not essential for understanding the main results of the paper. A reader interested in applications of Theorem 2.1 to reinforcement learning can proceed to Section 4.

The proof of Theorem 2.1 is organized into three steps, each of which is fairly simple. It is helpful to define empirical and population versions of the pairwise objective used by DBR:

$$\text{(Empirical)}: \ \widehat{\mathcal{L}}(f;g) := \frac{1}{n}\sum_{i=1}^{n} W^\tau_{f,g}(x_i)\big\{(f(x_i)-y_i)^2 - (g(x_i)-y_i)^2\big\},$$
$$\text{(Population)}: \ \mathcal{L}(f;g) := \mathbb{E}_{\text{train}}\big[W^\tau_{f,g}(x)\big\{(f(x)-y)^2 - (g(x)-y)^2\big\}\big].$$

First, we establish a certain non-negativity property of the population objective, which is the main structural result. The second step is a uniform convergence argument to show that $\widehat{\mathcal{L}}(\cdot;\cdot)$, which appears in the algorithm, concentrates to the population counterpart $\mathcal{L}(\cdot;\cdot)$. Finally, we study the minimizer $\hat{f}^{(n)}_{\mathsf{DBR}}$ and an L_∞-approximation $\bar{f}$ and relate their objective values to establish the theorem. Details and proofs for the corollaries are deferred to Appendix B.3.

Step 1: Non-negativity. The key lemma for the analysis is the following structural property.

Lemma 3.1 (Non-negativity). *With $\tau \geq 2\varepsilon_\infty$ and for any $\bar{f} \in \mathcal{F}$ such that $\|\bar{f} - f^\star\|_\infty \leq \varepsilon_\infty$, we have*

$$\mathcal{L}(f; \bar{f}) \geq (\tau^2 - 2\tau\varepsilon_\infty) \Pr[W^\tau_{f,\bar{f}}(x)] \geq 0.$$

The proof requires only algebraic manipulations and actually reveals a stronger property: with $\tau \geq 2\varepsilon_\infty$, the random variable $W^\tau_{f,\bar{f}}(x)\big[(f(x) - f^\star(x))^2 - (\bar{f}(x) - f^\star(x))^2\big]$ is non-negative almost surely. By the symmetry $\mathcal{L}(f; g) = -\mathcal{L}(g; f)$, the lemma also shows that any L_∞-misspecified $\bar{f}$ has non-positive population objective.

Step 2: Uniform convergence. Next we establish the following concentration guarantee.

Lemma 3.2 (Concentration). *Fix $\delta \in (0, 1)$ and $\tau \geq 3\varepsilon_\infty$ and define $\varepsilon_{\mathrm{stat}} := \frac{80 \log(|\mathcal{F}|/\delta)}{3n}$. Under Assumption 2.3, for any $\bar{f} \in \mathcal{F}$ such that $\|\bar{f} - f^\star\|_\infty \leq \varepsilon_\infty$, with probability at least $1 - \delta$ we have*

$$\forall f \in \mathcal{F} : \quad \mathcal{L}(f; \bar{f}) \leq 2\widehat{\mathcal{L}}(f; \bar{f}) + \varepsilon_{\mathrm{stat}}, \qquad \text{and equivalently,} \qquad \widehat{\mathcal{L}}(\bar{f}; f) \leq \frac{1}{2}\big(\mathcal{L}(\bar{f}; f) + \varepsilon_{\mathrm{stat}}\big).$$

The proof is based on Bernstein's inequality and importantly exploits a "self-bounding" property of $\widehat{\mathcal{L}}(f; g)$—in particular that $\mathrm{Var}[\widehat{\mathcal{L}}(f; \bar{f})] \leq (12/n)\mathcal{L}(f; \bar{f})$—analogously to the analysis for ERM in the well-specified case.

Step 3: Analysis of $\hat{f}^{(n)}_{\mathsf{DBR}}$. Let $\bar{f} \in \mathcal{F}$ be any function that is L_∞-close to $f^\star$ and condition on the high probability event in Lemma 3.2 holding with the choice $\bar{f}$. The DBR minimizer satisfies

$$\begin{aligned}\mathcal{L}(\hat{f}^{(n)}_{\mathsf{DBR}}; \bar{f}) &\overset{\text{(i)}}{\leq} 2\widehat{\mathcal{L}}(\hat{f}^{(n)}_{\mathsf{DBR}}; \bar{f}) + \varepsilon_{\mathrm{stat}} \overset{\text{(ii)}}{\leq} 2\max_{g \in \mathcal{F}} \widehat{\mathcal{L}}(\hat{f}^{(n)}_{\mathsf{DBR}}; g) + \varepsilon_{\mathrm{stat}} \\ &\overset{\text{(iii)}}{\leq} 2\max_{g \in \mathcal{F}} \widehat{\mathcal{L}}(\bar{f}; g) + \varepsilon_{\mathrm{stat}} \overset{\text{(iv)}}{\leq} \max_{g \in \mathcal{F}} \mathcal{L}(\bar{f}; g) + 2\varepsilon_{\mathrm{stat}} \overset{\text{(v)}}{\leq} 2\varepsilon_{\mathrm{stat}}.\end{aligned}$$

Here inequalities (i) and (iv) are applications of Lemma 3.2, (ii) and (iii) follow from the definition of $\hat{f}^{(n)}_{\mathsf{DBR}}$ since $\bar{f} \in \mathcal{F}$, and (v) is an application of Lemma 3.1 along with the symmetry $\mathcal{L}(f; g) = -\mathcal{L}(g; f)$. Eq. (3) now follows from the fact that $W^\tau_{f,\bar{f}}(x) \geq \mathbb{1}\{|f(x) - f^\star(x)| \geq \tau + \varepsilon_\infty\}$. Eq. (4) follows since under the event $|f(x) - f^\star(x)| \geq \tau + \varepsilon_\infty$ we can lower bound $(f(x) - f^\star(x))^2 - (\bar{f}(x) - f^\star(x))^2 \geq (\tau + \varepsilon_\infty)^2 - \varepsilon_\infty^2$.

4. Applications to online and offline reinforcement learning

In this section, we deploy disagreement-based regression to obtain new results in offline and online RL with function approximation. Algorithmically, this is achieved by using DBR as a drop-in replacement for square loss regression in existing algorithms. We illustrate this by examining and improving the Bellman residual minimization (a.k.a. minimax) algorithm for offline RL (Antos et al., 2008; Chen and Jiang, 2019) (Section 4.1) and the GOLF algorithm (Jin et al., 2021) for online RL (Section 4.2). The analyses also require minimal modifications to those of Xie and Jiang (2021) and Xie et al. (2023), respectively. To emphasize the ease with which DBR can be applied, we adopt the formulations and much of the notation from these works. All proofs for results in this section are deferred to Appendix C.

4.1. Offline reinforcement learning

Setup and notation. We consider a discounted Markov decision process (MDP) $M = (P, R, d_0, \gamma)$ over states $\mathcal{S}$ and actions $\mathcal{A}$, where $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition operator, $R : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function, $d_0 \in \Delta(\mathcal{S})$ is the initial state distribution, and $\gamma \in [0, 1)$ is the discount factor. A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ induces a trajectory $s_0, a_0, r_0, s_1, a_1, r_1, \ldots$ where $s_0 \sim d_0$, and for each $h \in \mathbb{N}$, $a_h \sim \pi(s_h)$, $r_h = R(s_h, a_h)$, and $s_{h+1} \sim P(s_h, a_h)$. We use $\mathbb{P}^\pi[\cdot]$ and $\mathbb{E}^\pi[\cdot]$ to denote probability and expectation under this process. Let $d_h^\pi \in \Delta(\mathcal{S} \times \mathcal{A})$ denote the occupancy measure of π at time-step h, defined as $d_h^\pi(s, a) := \mathbb{P}^\pi[s_h = s, a_h = a]$ and let $d^\pi := (1 - \gamma) \sum_{h=0}^\infty \gamma^h d_h^\pi$.

The value of π is denoted $J(\pi) := \mathbb{E}^\pi\left[\sum_{h=0}^\infty \gamma^h r_h\right]$. Each policy π has value functions $V^\pi : s \mapsto \mathbb{E}^\pi\left[\sum_{h=0}^\infty \gamma^h r_h \mid s_0 = s\right]$ and $Q^\pi : (s, a) \mapsto \mathbb{E}^\pi\left[\sum_{h=0}^\infty \gamma^h r_h \mid s_0 = s, a_0 = a\right]$, and it is known that there exists a policy $\pi^\star$ that maximizes $V^\pi(s)$ simultaneously for all $s \in \mathcal{S}$. This policy also optimizes $J(\cdot)$ and hence is called the optimal policy. It is also known that the value function $Q^\star := Q^{\pi^\star}$ induces the optimal policy via $\pi^\star : s \mapsto \arg\max_a Q^\star(s, a)$ and additionally satisfies *Bellman's optimality equation*: $Q^\star(s, a) := [\mathcal{T}Q^\star](s, a)$ where $\mathcal{T}$ is the Bellman operator, defined via $\mathcal{T}f : (s, a) \mapsto \mathbb{E}[r_0 + \gamma \max_{a'} f(s_1, a') \mid s_0 = s, a_0 = a]$.

In the offline value function approximation setting, we are given a dataset of n tuples $D_n := \{(s_i, a_i, r_i, s_i')\}_{i=1}^n$ generated i.i.d. from the following process: $(s_i, a_i) \sim \mu$ where $\mu \in \Delta(\mathcal{S}, \mathcal{A})$ is the *data collection distribution*, $r_i = R(s_i, a_i)$, and $s_i' \sim P(s_i, a_i)$. We are also given a function class $\mathcal{F} \subset (\mathcal{S} \times \mathcal{A} \to \mathbb{R})$, where each $f \in \mathcal{F}$ induces the policy $\pi_f : s \mapsto \arg\max_a f(s, a)$. Given dataset D_n and function class $\mathcal{F}$, we seek a policy $\hat{\pi}$ that has small suboptimality gap: $J(\pi^\star) - J(\hat{\pi})$. We impose the following assumptions on the function class and on the data collection distribution:

- L_∞**-misspecified realizability/completeness**: There exists $\bar{f} \in \mathcal{F}$ such that $\|\bar{f} - \mathcal{T}\bar{f}\|_\infty \le \varepsilon_\infty$. Additionally, for any $f \in \mathcal{F}$ there exists $g \in \mathcal{F}$ such that $\|g - \mathcal{T}f\|_\infty \le \varepsilon_\infty$.

- **Concentrability**: There exists a constant $C_{\mathrm{conc}} \in [1, \infty)$ such that $\max_{\pi \in \Pi} \left\| \frac{d^\pi}{\mu} \right\|_\infty \le C_{\mathrm{conc}}$. Here $\Pi := \{\pi_f : f \in \mathcal{F}\}$ is the policy class induced by $\mathcal{F}$.

There is a large body of recent work studying various function approximation and coverage assumptions in offline RL (c.f., Xie and Jiang, 2021). Arguably the most standard are concentrability, as we use, and *exact* realizability/completeness, which is stronger than our version with misspecification. Regarding the function approximation assumption, it is not hard to show that misspecification amplification—which in this setting is defined by the suboptimality $J(\pi^\star) - J(\hat{\pi})$ scaling as $\Omega(\varepsilon_\infty \sqrt{C_{\mathrm{conc}}})$—is necessary under weaker notions, such as $L_2(\mu)$-misspecification. Regarding coverage, as we will discuss below, the strength of the coverage assumption determines whether misspecification amplification can be avoided or not.

Algorithm and guarantee. The algorithm we study is a minor modification to the minimax algorithm (Antos et al., 2008; Chen and Jiang, 2019). For each function $\tilde{f} \in \mathcal{F}$ and each tuple (s_i, a_i, r_i, s_i') we can form a regression sample $(s_i, a_i, y_{\tilde{f},i} := r_i + \gamma \max_{a'} \tilde{f}(s_i', a'))$ and define the predictor $\hat{f}$ via the objective:

$$\hat{f} := \arg\min_{f \in \mathcal{F}} \max_{g \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n W_{f,g}^\tau(s_i, a_i)\left\{(f(s_i, a_i) - y_{f,i})^2 - (g(s_i, a_i) - y_{f,i})^2\right\}. \tag{7}$$

Here $W^{\tau}_{f,g}(\cdot)$ is the filter in Eq. (2) with $x = (s, a)$. Given $\hat{f}$, we output $\hat{\pi} := \pi_{\hat{f}}$. Note that the only difference between this algorithm and the original minimax algorithm is the use of the filter $W^{\tau}_{f,g}(\cdot)$ which is essential for obtaining the following guarantee.

Theorem 4.1 (DBR for offline RL). *Fix $\delta \in (0, 1)$, assume that $\mathcal{F}$ is L_∞-misspecified and μ satisfies concentrability (as defined above). Consider the algorithm defined in Eq. (7) with $\tau = 3\varepsilon_\infty$. Then, with probability at least $1 - \delta$ we have*

$$J(\pi^\star) - J(\hat{\pi}) \leq O\left(\frac{\varepsilon_\infty}{1-\gamma} + \frac{1}{1-\gamma}\sqrt{C_{\mathrm{conc}}\frac{\log(|\mathcal{F}|/\delta)}{n}}\right).$$

The theorem is best understood via comparison to the guarantee for the standard minimax algorithm, e.g., Theorem 5 of Xie and Jiang (2020). Under our assumptions (L_∞-misspecification and concentrability), these two bounds differ *only* in the misspecification term: our theorem scales as $\varepsilon_\infty/(1-\gamma)$ while the guarantee for the minimax algorithm scales as $\varepsilon_\infty\sqrt{C_{\mathrm{conc}}}/(1-\gamma)$.[4] Thus, our algorithm inherits the favorable properties of DBR to avoid misspecification amplification in offline RL.

This feature is notable in light of existing lower bounds for misspecified RL (Du et al., 2020; Van Roy and Dong, 2019; Lattimore et al., 2020). Formally, these results consider linear function approximation in various online RL models, but the constructions can be extended to offline RL with general function approximation where coverage is measured via the *Bellman transfer coefficient*. This coefficient is the smallest C_{transfer} such that $\max_{\pi, f\in\mathcal{F}} \frac{\|f-\mathsf{apx}[f]\|^2_{L_2(d^\pi)}}{\|f-\mathsf{apx}[f]\|^2_{L_2(\mu)}} \leq C_{\mathrm{transfer}}$ where $\mathsf{apx}[f] \in \mathcal{F}$ is the L_∞-approximation of $\mathcal{T}f$.[5] The lower bound states that an asymptotic error of $\Omega(\varepsilon_\infty\sqrt{C_{\mathrm{transfer}}})$ is unavoidable.

To contextualize our result with this lower bound, we identify two regimes: the "Bellman transfer regime" where $C_{\mathrm{transfer}} < \infty$ and the "concentrability regime" where $C_{\mathrm{conc}} < \infty$, and note that, since $C_{\mathrm{transfer}} \leq C_{\mathrm{conc}}$, the former is more general. In the Bellman transfer regime, misspecification amplification is unavoidable. In the concentrability regime, Theorem 4.1 avoids misspecification amplification *and* is sample efficient (i.e., has statistical term scaling as $\mathrm{poly}(C_{\mathrm{conc}}, \log(|\mathcal{F}|/\delta), \frac{1}{n}, \frac{1}{1-\gamma})$). This is the first result showing that both of these properties are simultaneously achievable: prior results achieve sample efficiency with misspecification amplification (e.g., Xie and Jiang, 2020), or avoid misspecification amplification with undesirable sample complexity scaling as $\mathrm{poly}(|\mathcal{S}|)$ (the latter is easily achieved under concentrability via a tabular model-based approach). Thus, the regime determines whether misspecification amplification is avoidable or not, and, in the regime where it is avoidable, our algorithm does so in a sample-efficient manner.

4. Xie and Jiang (2020) consider slightly weaker assumptions: they measure both misspecification and concentrability via the $L_2(\mu)$ norm. Our analysis easily accommodates $L_2(\mu)$-concentrability, as can be seen from the proof. On the other hand, as described in Section 2.1, misspecification amplification is necessary under $L_2(\mu)$-misspecification.
5. Many Bellman transfer coefficients exist, but a standard one is the smallest C_{transfer} such that $\max_{\pi, f\in\mathcal{F}} \|f-\mathcal{T}f\|^2_{L_2(d^\pi)}/\|f-\mathcal{T}f\|^2_{L_2(\mu)} \leq C_{\mathrm{transfer}}$. This coincides with ours under exact realizability/completeness, but we believe our definition is more appropriate for the misspecified case because it is equivalent to feature coverage under linear function approximation. Indeed, if $\mathcal{F}$ consists of linear functions in some feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ (but $\mathcal{T}f$ may not be linear due to misspecification) then our definition can be expressed via the features (as $\max_{\pi, \theta\in\mathbb{R}^d} \theta^\top\Sigma_\pi\theta/\theta^\top\Sigma_\mu\theta$ where $\Sigma_d = \mathbb{E}_d\left[\phi(s,a)\phi(s,a)^\top\right]$) but the standard definition cannot.

4.2. Online reinforcement learning

Setup and notation. We consider a finite horizon episodic MDP (P, R, H, s_1) over state space $\mathcal{S}$ and action space $\mathcal{A}$, where $H \in \mathbb{N}$ is horizon, $P := \{P_h\}_{h=1}^H$ with $P_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the non-stationary transition operator, $R := \{R_h\}_{h=1}^H$ with $R_h : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the non-stationary reward function, and s_1 is a fixed starting state. A (non-stationary) policy $\pi := \{\pi_h\}_{h=1}^H$ is a sequence of mappings $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ which induces a trajectory $(s_1, a_1, r_1, \ldots, s_H, a_H, r_H)$ where $a_h \sim \pi_h(s_h), r_h = R_h(s_h, a_h)$ and $s_{h+1} \sim P_h(s_h, a_h)$ for each time step. We use $\mathbb{P}^\pi[\cdot]$ and $\mathbb{E}^\pi[\cdot]$ to denote probability and expectation under this process, respectively. Let $d_h^\pi \in \Delta(\mathcal{S} \times \mathcal{A})$ denote the occupancy measure of π at time-step h, defined as $d_h^\pi(s,a) := \mathbb{P}^\pi[s_h = s, a_h = a]$.

The value of policy π is denoted $J(\pi) := \mathbb{E}^\pi\left[\sum_{h=1}^H r_h\right]$. Each policy has value functions: $V_h^\pi : s \mapsto \mathbb{E}^\pi\left[\sum_{h'=h}^H r_{h'} \mid s_h = s\right]$ and $Q_h^\pi : (s,a) \mapsto \mathbb{E}^\pi\left[\sum_{h'=h}^H r_{h'} \mid s_h = s, a_h = a\right]$ and there exist an optimal policy $\pi^\star = \{\pi_h^\star\}_{h=1}^H$ that maximizes V_h^π simultaneously for each state $s \in \mathcal{S}$ and hence maximizes $J(\cdot)$. The optimal value function $Q_h^\star := Q_h^{\pi_h^\star}$ induces $\pi^\star$ via $\pi_h^\star : s \mapsto \arg\max_a Q_h^\star(s,a)$ and satisfies Bellman's equation: $Q_h^\star(s,a) = [\mathcal{T}_h Q_{h+1}^\star](s,a)$ where the Bellman operator $\mathcal{T}_h$ is defined via $[\mathcal{T}_h f_{h+1}](s,a) = R_h(s,a) + \mathbb{E}[\max_{a'} f_{h+1}(s_{h+1}, a') \mid s_h = s, a_h = a]$. We assume per-episode rewards satisfy $\sum_{h=1}^H r_h \in [0,1]$.

In online RL, we interact with the MDP for T episodes, where in each episode we select a policy $\pi^{(t)}$ and collect the trajectory $(s_1^{(t)}, a_1^{(t)}, r_1^{(t)}, \ldots, s_H^{(t)}, a_H^{(t)}, r_H^{(t)})$ by taking actions $a_h^{(t)} = \pi_h^{(t)}(s_h^{(t)})$. We measure performance via the cumulative regret, define as $\mathrm{Reg} := \sum_{t=1}^T J(\pi^\star) - J(\pi^{(t)})$. We equip the learner with a value function class $\mathcal{F} := \mathcal{F}_1 \times \ldots \times \mathcal{F}_H$ where each $\mathcal{F}_h \subset \mathcal{S} \times \mathcal{A} \to [0,1]$. Each $f \in \mathcal{F}$ induces a policy π_f which, at time step h takes actions via $\pi_{f,h}(s_h) = \arg\max_a f_h(s_h, a_h)$. We make the following assumptions:

- **L_∞-approximate realizability/completeness.** For each $h \in [H]$ there exists $\bar{f}_h \in \mathcal{F}_h$ such that $\left\|\bar{f}_h - \mathcal{T}_h \bar{f}_{h+1}\right\|_\infty \le \varepsilon_\infty$. Additionally, for each $f_{h+1} \in \mathcal{F}_{h+1}$ there exists $f_h \in \mathcal{F}_h$ such that $\|f_h - \mathcal{T}_h f_{h+1}\|_\infty \le \varepsilon_\infty$.

- **Coverability.** There exists $C_{\mathrm{cov}} \in [1, \infty)$ such that $\inf_{\mu_1,\ldots,\mu_H \in \Delta(\mathcal{S}\times\mathcal{A})} \sup_{\pi \in \Pi, h} \left\|\frac{d_h^\pi}{\mu_h}\right\|_\infty \le C_{\mathrm{cov}}$. Here $\Pi := \{\pi_f : \pi_{f,h}(s) = \arg\max_a f_h(s,a), f \in \mathcal{F}\}$ is the policy class induced by $\mathcal{F}$.

As in offline RL, there is a large body of recent work studying function approximation and structural conditions for sample-efficient online RL (c.f., Agarwal et al., 2019; Foster and Rakhlin, 2023). It is fairly standard to assume exact realizability and completeness, which is stronger than our version with misspecification. Coverability is a recently proposed structural condition (Xie et al., 2023): C_{cov} is known to be small in many MDP models of interest, but weaker conditions that enable sample-efficiency are known. As we will see, the strength of the structural condition determines whether misspecification amplification can be avoided or not.

Algorithm and guarantee. The algorithm is a very minor modification to GOLF (Jin et al., 2021; Xie et al., 2023). To condense the notation, given a sample $(s_h^{(i)}, a_h^{(i)}, r_h^{(i)}, s_{h+1}^{(i)})$ and a function $f' \in \mathcal{F}_{h+1}$, define $x_h^{(i)} := (s_h^{(i)}, a_h^{(i)})$ and $y_{f',h}^{(i)} := r_h^{(i)} + \max_{a'} f'(s_{h+1}^{(i)}, a')$. At the beginning of

episode t, define a version space

$$\mathcal{F}^{(t-1)} := \left\{ f \in \mathcal{F} : \forall h \in [H] : \max_{g \in \mathcal{F}_h} \sum_{i=1}^{t-1} W^{\tau}_{f_h,g}(x_h^{(i)}) \Big\{ (f_h(x_h^{(i)}) - y^{(i)}_{f_{h+1},h})^2 - (g(x_h^{(i)}) - y^{(i)}_{f_{h+1},h})^2 \Big\} \leq \beta \right\},$$

where $\beta > 0$ is a hyperparameter we will set below.

Then, we define the optimistic value function $f^{(t)} := \arg\max_{f \in \mathcal{F}^{(t-1)}} f_1(s_1, \pi_{f,1}(s_1))$ and the induced policy $\pi^{(t)} := \pi_{f^{(t)}}$, collect a trajectory via $\pi^{(t)}$, and proceed to the next episode. Note that the only difference between this algorithm, which we call GOLF.DBR, and the version of GOLF studied by Xie et al. (2023) is that we use the filter $W^{\tau}_{f_h,g}(\cdot)$ in the construction of the version space. GOLF.DBR enjoys the following guarantee.

Theorem 4.2 (DBR for online RL). *Fix $\delta \in (0,1)$, and assume that $\mathcal{F}$ is L_∞-misspecified and coverability (as defined above) holds. Consider* GOLF.DBR *with $\tau = 3\varepsilon_\infty$ and $\beta = c\log(TH|\mathcal{F}|/\delta)$. Then, with probability at least $1-\delta$, we have*

$$\mathrm{Reg} \leq O\Big(\varepsilon_\infty HT + H\sqrt{C_{\mathrm{cov}} T \log(TH|\mathcal{F}|/\delta)\log(T)}\Big).$$

Paralleling the discussion following Theorem 4.1, we emphasize two aspects of the result. The first is that it extends Theorem 1 of Xie et al. (2023) to the misspecified setting, with no degradation of the statistical term and without incurring a dependence on $\varepsilon_\infty\sqrt{C_{\mathrm{cov}}}$. In other words, it avoids misspecification amplification.

The second remark is that, when taken with existing lower bounds (Du et al., 2020; Van Roy and Dong, 2019; Lattimore et al., 2020), Theorem 4.2 establishes a separation between coverability and structural parameters defined in terms of Bellman errors, which include the Bellman-Eluder dimension (Jin et al., 2021), bilinear rank (Du et al., 2021), and Bellman rank (Jiang et al., 2017).[6] This separation is more subtle than in offline RL, because here, as long as the state-action space is finite, one can always use a "tabular" method and eliminate misspecification altogether, at the cost of $\mathrm{poly}(|\mathcal{S}|,|\mathcal{A}|)\cdot\sqrt{T}$ regret. To rule out this algorithm, we restrict to sample-efficient methods: in a setting where a particular structural parameter (e.g., coverability or Bellman rank) is bounded by d we say that an algorithm is sample-efficient if its statistical term scales as $\mathrm{poly}(d, \log(|\mathcal{F}|/\delta), H)\cdot o(T)$. The lower bounds show that, when the structural parameter involves Bellman errors (like the Bellman rank), $\varepsilon_\infty T\sqrt{d}$ misspecification error is necessary for sample efficient algorithms.[7] On the other hand, under coverability, we can achieve misspecification error with no dependence on the structural parameter, in a sample efficient manner.[8] This establishes that whether misspecification amplification can be avoided sample-efficiently depends on the structural properties of the MDP. To our knowledge, this is a novel insight into the interaction between the structural and function approximation assumptions in online RL.

6. As with Bellman transfer coefficients, we believe these definitions should be adjusted to accommodate misspecification. See Definition 10 in Jiang et al. (2017) for an example.
7. Formally, for any $\zeta > 0$ one requires at least $\exp(d^{2\zeta})$ samples to find a $d^{1/2-\zeta}\varepsilon_\infty$ suboptimal policy (Lattimore et al., 2020).
8. We believe that misspecification error $\varepsilon_\infty HT$ is optimal under coverability and that $\varepsilon_\infty HT\sqrt{d}$ is optimal under structural parameters like Bellman rank. However, it remains open to establish the necessity of the horizon factors.

5. Discussion

This paper highlights an intriguing interplay between misspecification and distribution shift, exposing the undesirable *misspecification amplification* property of ERM, and proposing disagreement-based regression as a remedy. We have shown that using disagreement-based regression in online and offline reinforcement learning yields new technical results and reveals new tradeoffs between coverage/structural assumptions and function approximation assumptions.

We close by mentioning several interesting avenues for future work. There are a number of directions that pertain to the core setting of misspecified regression under covariate shift; for example, (a) extending the analysis of DBR to infinite function classes, other loss functions, and other notions of misspecification, (b) deriving a more computationally efficient procedure—perhaps in an oracle model of computation—that avoids misspecification amplification, and (c) determining the optimal achievable approximation factor. Pertaining to reinforcement learning theory, we believe the most pressing direction is to deepen our understanding of the relationship between coverage/structural assumptions (for offline/online RL, respectively) and function approximation assumptions, and we believe misspecification provides a novel lens to study this relationship. It is also worthwhile to consider other applications involving distribution shift where DBR or related procedures may reveal new conceptual insights. Finally, it would also be interesting to study empirical issues, to understand how pervasive and problematic misspecification amplification is, develop practical interventions, and consider applying them to distribution shift and deep reinforcement learning scenarios.

In short, there is much more to understand about the interplay between misspecification and distribution shift, and we look forward to progress in the years to come.

Acknowledgements

We thank Adam Block and Max Simchowitz for helpful feedback on a early version of the manuscript.

References

Alekh Agarwal and Tong Zhang. Minimax regret optimization for robust machine learning under distribution shift. In *Conference on Learning Theory*, 2022.

Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. https://rltheorybook.github.io/, 2019. Version: January 31, 2022.

Anish Agarwal, Devavrat Shah, and Dennis Shen. Synthetic interventions. *arXiv:2006.07691*, 2020.

Anish Agarwal, Munther Dahleh, Devavrat Shah, and Dennis Shen. Causal matrix completion. In *Conference on Learning Theory*, 2023.

Philip Amortila, Nan Jiang, and Csaba Szepesvári. The optimal approximation factors in misspecified off-policy value function estimation. In *International Conference on Machine Learning*, 2023.

Cem Anil, Yuhuai Wu, Anders Andreassen, Aitor Lewkowycz, Vedant Misra, Vinay Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. Exploring length generalization in large language models. *Advances in Neural Information Processing Systems*, 2022.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 2008.

Jean-Yves Audibert. Progressive mixture rules are deviation suboptimal. *Advances in Neural Information Processing Systems*, 2007.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in Neural Information Processing Systems*, 19, 2006.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. *Operations Research*, 2015.

Steven J Bradtke and Andrew G Barto. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 1996.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, 2019.

Corinna Cortes and Mehryar Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 2014.

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. *Advances in Neural Information Processing Systems*, 2010.

Kefan Dong and Tengyu Ma. First steps toward understanding the extrapolation of nonlinear models to unseen domains. In *International Conference on Learning Representations*, 2023a.

Kefan Dong and Tengyu Ma. Toward L_∞-recovery of nonlinear functions: A polynomial sample complexity bound for gaussian random fields. In *Conference on Learning Theory*, 2023b.

Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in RL. In *International Conference on Machine Learning*, 2021.

Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? In *International Conference on Learning Representations*, 2020.

John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 2021.

Dylan Foster, Alekh Agarwal, Miroslav Dudík, Haipeng Luo, and Robert Schapire. Practical contextual bandits with regression oracles. In *International Conference on Machine Learning*, 2018.

Dylan J Foster and Alexander Rakhlin. Foundations of reinforcement learning and interactive decision making. *arXiv:2312.16730*, 2023.

Dylan J Foster, Alexander Rakhlin, David Simchi-Levi, and Yunzong Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. In *Conference on Learning Theory*, 2021.

Dylan J Foster, Akshay Krishnamurthy, David Simchi-Levi, and Yunzong Xu. Offline reinforcement learning: Fundamental barriers for value function approximation. In *Conference on Learning Theory*, 2022.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 2014.

Jiawei Ge, Shange Tang, Jianqing Fan, Cong Ma, and Chi Jin. Maximum likelihood estimation is all you need for well-specified covariate shift. *arXiv:2311.15961*, 2023.

Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. *Dataset Shift in Machine Learning*, 2009.

Steve Hanneke. Theory of disagreement-based active learning. *Foundations and Trends in Machine Learning*, 2014.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, 2018.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. *Advances in Neural Information Processing Systems*, 2006.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In *International Conference on Machine Learning*, 2017.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. *Advances in Neural Information Processing Systems*, 2021.

Keith Knight. On the asymptotic distribution of the L_∞ estimator in linear regression. Technical report, University of Toronto, 2017.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, 2021.

Samory Kpotufe and Guillaume Martinet. Marginal singularity, and the benefits of labels in covariate-shift. In *Conference On Learning Theory*, 2018.

Akshay Krishnamurthy, Alekh Agarwal, Tzu-Kuo Huang, Hal Daumé III, and John Langford. Active learning for cost-sensitive classification. *Journal of Machine Learning Research*, 2019.

Tor Lattimore, Csaba Szepesvari, and Gellert Weisz. Learning with good feature representations in bandits and in RL with a generative model. In *International Conference on Machine Learning*, 2020.

Guillaume Lecué and Philippe Rigollet. Optimal learning with Q-aggregation. *The Annals of Statistics*, 2014.

Qi Lei, Wei Hu, and Jason Lee. Near-optimal linear regression under distribution shift. In *International Conference on Machine Learning*, 2021.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv:2005.01643*, 2020.

Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with square loss: Localization through offset rademacher complexity. In *Conference on Learning Theory*, 2015.

Bingbin Liu, Jordan T Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Exposing attention glitches with flip-flop language modeling. *Advances in Neural Information Processing Systems*, 2023.

Cong Ma, Reese Pathak, and Martin J Wainwright. Optimally tackling covariate shift in RKHS-based nonparametric regression. *The Annals of Statistics*, 2023.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv:0902.3430*, 2009.

John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, 2021.

Wenlong Mou, Ashwin Pananjady, and Martin J Wainwright. Optimal oracle inequalities for solving projected fixed-point equations. *Mathematics of Operations Research*, 2022.

Rémi Munos. Error bounds for approximate policy iteration. In *International Conference on Machine Learning*, 2003.

Rémi Munos. Performance bounds in L_p-norm for approximate value iteration. *SIAM Journal on Control and Optimization*, 2007.

Reese Pathak, Cong Ma, and Martin Wainwright. A new similarity measure for covariate shift with applications to nonparametric regression. In *International Conference on Machine Learning*, 2022.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, 2020.

Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset Shift in Machine Learning*. Mit Press, 2008.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, 2019.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics*, 2011.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

Johannes Schmidt-Hieber and Petr Zamolodtchikov. Local convergence rates of the least squares estimator with applications to transfer learning. *arXiv:2204.05003*, 2022.

Devavrat Shah, Dogyoon Song, Zhi Xu, and Yuzhe Yang. Sample efficient reinforcement learning via low-rank matrix estimation. *Advances in Neural Information Processing Systems*, 2020.

Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv:2108.13624*, 2021.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 2000.

Max Simchowitz, Abhishek Gupta, and Kaiqing Zhang. Tackling combinatorial distribution shift: A matrix completion perspective. In *Conference on Learning Theory*, 2023.

Masashi Sugiyama and Motoaki Kawanabe. *Machine Learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press, 2012.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in Neural Information Processing Systems*, 2007.

Matus Telgarsky. Deep learning theory lecture notes. https://mjt.cs.illinois.edu/dlt/, 2021. Version: 2021-10-27 v0.0-e7150f2d (alpha).

John Tsitsiklis and Benjamin Van Roy. Analysis of temporal-difference learning with function approximation. *Advances in Neural Information Processing Systems*, 1996.

Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer, 2008.

Benjamin Van Roy and Shi Dong. Comments on the Du-Kakade-Wang-Yang lower bounds. *arXiv:1911.07910*, 2019.

Tengyang Xie and Nan Jiang. $Q^\star$ approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence*, 2020.

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, 2021.

Tengyang Xie, Dylan J Foster, Yu Bai, Nan Jiang, and Sham M Kakade. The role of coverage in online reinforcement learning. In *International Conference on Learning Representations*, 2023.

Yufei Yi and Matey Neykov. Non-asymptotic bounds for the L_∞ estimator in linear regression with uniform noise. *Bernoulli*, 2024.

Huizhen Yu and Dimitri P Bertsekas. Error bounds for approximations from projected linear equations. *Mathematics of Operations Research*, 2010.

Yaoliang Yu and Csaba Szepesvári. Analysis of kernel mean matching under covariate shift. In *International Conference on Machine Learning*, 2012.

Yi Zhang, Arturs Backurs, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, and Tal Wagner. Unveiling transformers with lego: a synthetic reasoning task. *arXiv:2206.04301*, 2022.

Appendix A. Additional related work

There is a vast body of work studying distribution shift broadly and covariate shift in particular. We focus on the most closely related techniques for the covariate shift setting and refer the reader to Quinonero-Candela et al. (2008); Sugiyama and Kawanabe (2012); Shen et al. (2021) for a more comprehensive treatment.

Reweighting and robust optimization. Perhaps the most common way to correct for covariate shift is by reweighting each example (x, y) in the objective function by the density ratio $w(x) := d_{\text{test}}(x)/d_{\text{train}}(x)$. This method has been studied in a long series of works (Shimodaira, 2000; Cortes et al., 2010; Cortes and Mohri, 2014). In its simplest form it requires knowledge of $\mathcal{D}_{\text{test}}$ via the density ratios, so it is not directly applicable to our adversarial covariate shift setting. Extensions include approaches that estimate density ratios using unlabeled samples from $\mathcal{D}_{\text{test}}$ (Huang et al., 2006; Sugiyama et al., 2007; Gretton et al., 2009; Yu and Szepesvári, 2012) and robust optimization approaches that employ an auxiliary hypothesis class of distributions $\mathcal{P}$ containing $\mathcal{D}_{\text{test}}$ (Hashimoto et al., 2018; Sagawa et al., 2020; Duchi and Namkoong, 2021; Agarwal and Zhang, 2022). However, these still require prior knowledge about $\mathcal{D}_{\text{test}}$, in particular it is known that the sample complexity of robust optimization scales with the statistical complexity of the auxiliary class $\mathcal{P}$ (Duchi and Namkoong, 2021), leading to vacuous bounds in the absence of inductive bias.

Ge et al. (2023) study statistical inference under covariate shift in well- and misspecified settings. They show that maximum likelihood estimation on $\mathcal{D}_{\text{train}}$ is inconsistent with misspecification, a result which is conceptually similar to our lower bound for ERM. However, their construction is not L_∞-misspecified so it is not directly comparable. Algorithmically, they use reweighting for the misspecified case, which, as mentioned, cannot be implemented in our setting.

Sup-norm convergence and function class-specific results. Another line of work provides specialized analyses for specific function classes of interest, such as linear (Lei et al., 2021), bilinear (Simchowitz et al., 2023), nonparametric (Kpotufe and Martinet, 2018; Pathak et al., 2022; Ma et al., 2023), and some neural network (Dong and Ma, 2023a) classes. The overarching technical approach in these works is to measure distance between distributions in a manner that captures the structure of the function class, analogously to learning-theoretic results for domain adaptation (Ben-David et al., 2006; Mansour et al., 2009). Many of these works operate closer to the well-specified regime than we do (e.g., in the nonparametric setting. An exception is the work of Simchowitz et al. (2023) who studied misspecification amplification that arises from low rank approximation under bilinear structure, a setting that is rather different from our own.

A complementary approach is based on sup-norm convergence which seeks to control $\|\hat{f} - f^\star\|_\infty$ for a predictor $\hat{f}$ and is naturally robust to covariate shift. Sup-norm convergence has been studied for various function classes (c.f., Shah et al., 2020; Agarwal et al., 2020, 2023; Schmidt-Hieber and Zamolodtchikov, 2022; Dong and Ma, 2023b), but unfortunately is not possible in the general statistical learning setup (Dong and Ma, 2023b). We mention sup-norm convergence primarily to contrast with our quantile guarantee in Eq. (4), which controls the probability over x of large errors rather than the magnitude of the errors themselves and which is attainable for any function class, even with misspecification.

Related work in reinforcement learning. Our results for offline and online RL build directly on the analyses in Xie and Jiang (2020) and Xie et al. (2023) respectively. The former contributes to a long line of work on offline RL (Munos, 2003, 2007; Antos et al., 2008; Chen and Jiang, 2019) while the latter is part of a series of works establishing structural conditions under which online

reinforcement learning is statistically tractable (c.f., Agarwal et al., 2019; Foster and Rakhlin, 2023). Many of these works do account for misspecification, but the question of whether misspecification amplification can be avoided is not considered.

Results that do focus on misspecification primarily consider linear function approximation. In the simpler offline policy evaluation setting, several works study least squares temporal difference learning (LSTD) (Bradtke and Barto, 1996) with misspecification (Tsitsiklis and Van Roy, 1996; Yu and Bertsekas, 2010; Mou et al., 2022). Recently, Amortila et al. (2023) precisely characterized the optimal misspecification amplification (i.e., approximation factors) achievable across a range of settings, showing that LSTD is essentially optimal in most regimes. The exception is when the offline data distribution is supported on the entire state space, one can employ a "tabular" model-based algorithm to incur no approximation error whatsoever, but the sample complexity scales polynomially with $|\mathcal{S}|$. Our offline RL results are conceptually similar because under concentrability (which essentially implies full support), the standard minimax algorithm does not achieve the optimal approximation factor. A crucial difference is that our disagreement-based variant achieves an improved approximation factor *without* incurring any sample complexity overhead.

For the more challenging offline policy optimization and online RL, Du et al. (2020); Van Roy and Dong (2019); Lattimore et al. (2020) establish conditions under which misspecification amplification is necessary. As discussed above, combining our results with these lower bounds and their variations, reveals new tradeoffs between coverage/structural and function approximation conditions, distinct from tradeoffs established by prior work (Xie and Jiang, 2021; Foster et al., 2022).

Appendix B. Proofs for Section 2

B.1. Analysis for ERM

Proposition 2.1 (ERM upper bound)**.** *For any $\delta \in (0,1)$ with probability at least $1-\delta$,* ERM *satisfies*

$$R_{\mathrm{test}}(\hat{f}^{(n)}_{\mathsf{ERM}}) \leq O\left(C_\infty \varepsilon_\infty^2 + C_\infty \frac{\log(|\mathcal{F}|/\delta)}{n}\right).$$

Proof of Proposition 2.1. The proof of Proposition 2.1 is fairly standard, particularly in the well-specified case when $\varepsilon_\infty = 0$. Our analysis that handles misspecification is adapted from the proof of Lemma 16 in Chen and Jiang (2019). For the majority of the proof we only consider $\mathcal{D}_{\mathrm{train}}$, and we consequently omit the subscript when indexing expectations, variances, and the risk functional. Define

$$R(f) := \mathbb{E}[(f(x) - f^\star(x))^2] \quad \text{and} \quad \widehat{R}(f) := \frac{1}{n}\sum_{i=1}^n (f(x_i) - y_i)^2,$$

so that $\hat{f}^{(n)}_{\mathsf{ERM}} := \arg\min_{f\in\mathcal{F}} \widehat{R}(f)$. We establish concentration on the "excess risk" functional $\widehat{R}(f) - \widehat{R}(\bar{f})$. For any $f \in \mathcal{F}$, we establish the following facts:

$$\mathbb{E}[(f(x)-y)^2 - (\bar{f}(x)-y)^2] = \mathbb{E}[(f(x)-f^\star(x))^2 - (\bar{f}(x)-f^\star(x))^2] \tag{8}$$

$$\mathrm{Var}[(f(x)-y)^2 - (\bar{f}(x)-y)^2] \leq 8\mathbb{E}[(f(x)-y)^2 - (\bar{f}(x)-y)^2] + 16\varepsilon_\infty^2. \tag{9}$$

Eq. (8) implies that $\mathbb{E}[\widehat{R}(f) - \widehat{R}(\bar{f})] = R(f) - R(\bar{f})$ as desired. Eq. (9) will enable us to achieve a fast convergence rate. The former is derived as follows. Observe that conditional on any x we have

$$\begin{aligned}
&\mathbb{E}[(f(x) - y)^2 - (\bar{f}(x) - y)^2 \mid x] \\
&\quad = \mathbb{E}[(f(x) - y)^2 - (\bar{f}(x) - f^\star(x) + f^\star(x) - y)^2 \mid x] \\
&\quad = \mathbb{E}[(f(x) - y)^2 - (\bar{f}(x) - f^\star(x))^2 - 2(\bar{f}(x) - f^\star(x))(f^\star(x) - y) - (f^\star(x) - y)^2 \mid x] \\
&\quad = \mathbb{E}[(f(x) - y)^2 - (\bar{f}(x) - f^\star(x))^2 - (f^\star(x) - y)^2 \mid x] \\
&\quad = f(x)^2 - f^\star(x)^2 - 2\mathbb{E}_{\text{train}}[y \mid x](f(x) - f^\star(x)) - (\bar{f}(x) - f^\star(x))^2 \\
&\quad = (f(x) - f^\star(x))^2 - (\bar{f}(x) - f^\star(x))^2.
\end{aligned}$$

Eq. (9) is derived as follows.

$$\begin{aligned}
\mathrm{Var}[(f(x) - y)^2 - (\bar{f}(x) - y)^2] &\le \mathbb{E}[((f(x) - y)^2 - (\bar{f}(x) - y)^2)^2] \\
&= \mathbb{E}[(f(x) - \bar{f}(x))^2 (f(x) + \bar{f}(x) - 2y)^2] \\
&\le 4\mathbb{E}[(f(x) - \bar{f}(x))^2] \\
&\le 8\mathbb{E}[(f(x) - f^\star(x))^2 + (\bar{f}(x) - f^\star(x))^2] \\
&= 8\mathbb{E}[(f(x) - f^\star(x))^2 - (\bar{f}(x) - f^\star(x))^2 + 2(\bar{f}(x) - f^\star(x)^2] \\
&\le 8\mathbb{E}[(f(x) - f^\star(x))^2 - (\bar{f}(x) - f^\star(x))^2] + 16\varepsilon_\infty^2.
\end{aligned}$$

Finally, we apply Eq. (8).

Now, Bernstein's inequality and a union bound over $f \in \mathcal{F}$ gives that with probability at least $1-\delta$

$$\begin{aligned}
&\forall f \in \mathcal{F} : R(f) - R(\bar{f}) - (\widehat{R}(f) - \widehat{R}(\bar{f})) \\
&\qquad \le \sqrt{\frac{(16(R(f) - R(\bar{f})) + 32\varepsilon_\infty^2)\log(|\mathcal{F}|/\delta)}{n}} + \frac{4\log(|\mathcal{F}|/\delta)}{3n}.
\end{aligned}$$

Since $\hat{f}^{(n)}_{\mathsf{ERM}}$ minimizes $\widehat{R}(f)$ we have that $\widehat{R}(\hat{f}^{(n)}_{\mathsf{ERM}}) - \widehat{R}(\bar{f}) \le 0$, we can deduce that

$$R(\hat{f}^{(n)}_{\mathsf{ERM}}) - R(\bar{f}) \le \sqrt{\frac{(16(R(\hat{f}^{(n)}_{\mathsf{ERM}}) - R(\bar{f})) + 32\varepsilon_\infty^2)\log(|\mathcal{F}|/\delta)}{n}} + \frac{4\log(|\mathcal{F}|/\delta)}{3n}.$$

Using the AM-GM inequality ($\sqrt{ab} \le a/2 + b/2$), the right hand side can be simplified to yield

$$R(\hat{f}^{(n)}_{\mathsf{ERM}}) - R(\bar{f}) \le \frac{1}{2}(R(\hat{f}^{(n)}_{\mathsf{ERM}}) - R(\bar{f})) + \varepsilon_\infty^2 + \frac{28\log(|\mathcal{F}|/\delta)}{3n}.$$

Re-arranging and using that $R_{\text{train}}(\bar{f}) \le \varepsilon_\infty^2$ we obtain

$$R_{\text{train}}(\hat{f}^{(n)}_{\mathsf{ERM}}) \le 3\varepsilon_\infty^2 + \frac{56\log(|\mathcal{F}|/\delta)}{3n}.$$

Finally we bound the risk under $\mathcal{D}_{\text{test}}$ via a standard importance weighting argument:

$$R_{\text{test}}(\hat{f}^{(n)}_{\mathsf{ERM}}) = \mathbb{E}_{\text{train}}\left[\frac{d_{\text{test}}(x)}{d_{\text{train}}(x)}(\hat{f}^{(n)}_{\mathsf{ERM}}(x) - f^\star(x))^2\right] \le \sup_{x\in\mathcal{X}}\left|\frac{d_{\text{test}}(x)}{d_{\text{train}}(x)}\right| \cdot \left(3\varepsilon_\infty^2 + \frac{56\log(|\mathcal{F}|/\delta)}{3n}\right).$$

Note that we crucially use that $(\hat{f}^{(n)}_{\mathsf{ERM}}(x) - f^\star(x))^2$ is non-negative here. This proves the proposition. □

Proposition 2.2 (ERM lower bound). *For all $\varepsilon_\infty \in (0,1)$ and $C_\infty \in [1,\infty)$ such that $\sqrt{C_\infty} \cdot \varepsilon_\infty \le 1/2$, and for all $\zeta > 0$ sufficiently small, there exist distributions $\mathcal{D}_{\mathrm{train}}, \mathcal{D}_{\mathrm{test}}$ and a function class $\mathcal{F}$ with $|\mathcal{F}| = 2$ satisfying Assumption 2.1-Assumption 2.4 (with parameters $\varepsilon_\infty, C_\infty$) such that*

$$R_{\mathrm{test}}(\hat{f}^{(\infty)}_{\mathsf{ERM}}) = C_\infty \varepsilon_\infty^2 - \zeta.$$

Proof of Proposition 2.2. Fix $\varepsilon_\infty \in (0,1)$ and $C_\infty \ge 1$ such that $\sqrt{C_\infty} \cdot \varepsilon_\infty \le 1/2$. Let $0 < \zeta < \sqrt{C_\infty} \cdot \varepsilon_\infty$. Let $\mathcal{X} = [0,1]$ and let $\mathcal{D}_{\mathrm{train}}$ be the distribution over (x,y) where $x \sim \mathrm{Uniform}(\mathcal{X})$ and $y \sim \mathrm{Ber}(1/2)$. Let $\widetilde{\mathcal{X}} := [0, 1/C_\infty] \subset \mathcal{X}$ and let $\mathcal{D}_{\mathrm{test}}$ be the distribution over (x,y) where $x \sim \mathrm{Uniform}(\widetilde{\mathcal{X}})$ and $y \sim \mathrm{Ber}(1/2)$. These choices yield $f^\star(x) = 1/2$ for all $x \in \mathcal{X}$, satisfy Assumption 2.1, and ensure that $\sup_{x \in \mathcal{X}} \left| \frac{d_{\mathrm{test}}(x)}{d_{\mathrm{train}}(x)} \right| = C_\infty$.

Let $\mathcal{F} = \{\bar{f}, f_{\mathrm{bad}}\}$ where $\bar{f}(x) = 1/2 + \varepsilon_\infty$ for all $x \in \mathcal{X}$ (satisfying Assumption 2.3) and f_{bad} is defined as

$$f_{\mathrm{bad}}(x) = \begin{cases} 1/2 & \text{if } x \notin \widetilde{\mathcal{X}} \\ 1/2 + \zeta & \text{if } x \in \widetilde{\mathcal{X}} \end{cases}.$$

By definition, observe that $\hat{f}^{(\infty)}_{\mathsf{ERM}} = f_{\mathrm{bad}}$ as long as $\|f_{\mathrm{bad}} - f^\star\|^2_{L_2(\mathcal{D}_{\mathrm{train}})} < \|\bar{f} - f^\star\|^2_{L_2(\mathcal{D}_{\mathrm{train}})}$. A direct calculation shows that this inequality is satisfied for any $\zeta < \sqrt{C_\infty} \cdot \varepsilon_\infty$. However, f_{bad} has large population risk under $\mathcal{D}_{\mathrm{test}}$, in particular

$$R_{\mathrm{test}}(f_{\mathrm{bad}}) = \mathbb{E}_{\mathrm{test}}[(f_{\mathrm{bad}}(x) - f^\star(x))^2] = \zeta^2,$$

which we can make arbitrarily close to $C_\infty \varepsilon_\infty^2$. □

B.2. Discussion of other algorithms

Star algorithm. Audibert's star algorithm (Audibert, 2007; Liang et al., 2015) is a two-stage regression procedure that achieves the fast convergence rate for non-convex classes in misspecified or agnostic regression. Given that the construction used to prove Proposition 2.2 has a finite (and hence non-convex) function class, one might ask whether the star algorithm can avoid misspecification amplification. We briefly sketch here why this is not the case. In the context of the construction, where $\mathcal{F} = \{f_{\mathrm{bad}}, \bar{f}\}$, the asymptotic version of the star algorithm is to compute

$$\hat{f}_{\mathrm{star}} := \operatorname*{arg\,min}_{f_\alpha : \alpha \in [0,1]} \mathbb{E}_{\mathrm{train}}[(f_\alpha(x) - f^\star(x))^2] \quad \text{where} \quad f_\alpha(x) = (1-\alpha) f_{\mathrm{bad}}(x) + \alpha \bar{f}(x).$$

We claim that when $\zeta = \sqrt{C_\infty} \cdot \varepsilon_\infty$, the optimal choice for α is exactly $1/2$. The prediction error under $\mathcal{D}_{\mathrm{test}}$ for this choice is, unfortunately, exactly $1/4(\sqrt{C_\infty}+1)^2\varepsilon_\infty^2$, which still manifests misspecification amplification. Note that, due to the simplicity of our construction, the same argument applies to other improper learning schemes based on convexification (c.f., Lecué and Rigollet, 2014).

To see that the minimum is achieved at $\alpha = 1/2$, we write the optimization problem over α as

$$\begin{aligned} &\operatorname*{arg\,min}_{\alpha \in [0,1]} \frac{1}{C_\infty} \cdot \left((1-\alpha)\sqrt{C_\infty}\varepsilon_\infty + \alpha\varepsilon_\infty \right)^2 + \left(1 - \frac{1}{C_\infty}\right) \cdot (\alpha\varepsilon_\infty)^2 \\ &\quad = \operatorname*{arg\,min}_{\alpha \in [0,1]} \alpha^2 + (1-\alpha)^2 + \frac{2\alpha(1-\alpha)}{\sqrt{C_\infty}}. \end{aligned}$$

The derivative, w.r.t. α, of the latter is

$$\frac{d\left(\alpha^2 + (1-\alpha)^2 + \frac{2\alpha(1-\alpha)}{\sqrt{C_\infty}}\right)}{d\alpha} = 2\alpha - 2(1-\alpha) + \frac{2}{\sqrt{C_\infty}} - \frac{4\alpha}{\sqrt{C_\infty}} = \left(2 - \frac{2}{\sqrt{C_\infty}}\right)(2\alpha - 1).$$

Since $C_\infty > 1$, the second derivative is non-negative, so the optimization problem is convex. Moreover, the derivative is zero at $\alpha = 1/2$, showing that this is a minimizer of the optimization problem.

L_∞ regression. Given that we assume L_∞-misspecification, and in light of the construction for Proposition 2.2, it is tempting to optimize the maximal absolute deviation instead of the square loss:

$$\hat{f}_\infty^{(n)} \leftarrow \arg\min_{f \in \mathcal{F}} \max_i |f(x_i) - y_i|.$$

This procedure is known as L_∞ regression or the Chebyshev estimator and has been studied in the statistics community (Knight, 2017; Yi and Neykov, 2024). These analyses primarily consider the well-specified setting with noise that is uniformly distributed, i.e., $y_i = f^\star(x_i) + \epsilon_i$ where $\epsilon_i \sim \mathrm{Unif}([-a, a])$ for some $a \geq 0$. We believe such analyses can extend to the L_∞-misspecified setting to show that the procedure avoids misspecification amplification. However, strong assumptions on the noise are crucial, as L_∞ regression can be inconsistent under more general conditions.

We illustrate with a simple example. Let $\mathcal{X} = \{x\}$ be a singleton, $y = \mathrm{Ber}(1/4)$ and $\mathcal{F} = \{f^\star : x \mapsto 1/4, f : x \mapsto 1/2\}$ be a class with two functions. For all n sufficiently large, the dataset will contain the sample $(x, 1)$ at which point $f^\star$ will have L_∞ error $3/4$, while f will have error $1/2$. Thus the method will be inconsistent.

B.3. Analysis for DBR

We begin with the proofs of Lemma 3.1 and Lemma 3.2, thus completing steps one and two of the proof. Then we turn to proving the corollaries.

Lemma 3.1 (Non-negativity). *With $\tau \geq 2\varepsilon_\infty$ and for any $\bar{f} \in \mathcal{F}$ such that $\|\bar{f} - f^\star\|_\infty \leq \varepsilon_\infty$, we have*

$$\mathcal{L}(f; \bar{f}) \geq (\tau^2 - 2\tau\varepsilon_\infty) \Pr[W^\tau_{f,\bar{f}}(x)] \geq 0.$$

Proof of Lemma 3.1. Following the calculation used to derive Eq. (8) we have that, conditional on any x:

$$\mathbb{E}_{\text{train}}[(f(x) - y)^2 - (\bar{f}(x) - y)^2 \mid x] = (f(x) - f^\star(x))^2 - (\bar{f}(x) - f^\star(x))^2$$

Under the event $x \in W^\tau_{f,\bar{f}}$ with $\tau \geq 2\varepsilon_\infty$ we claim that this must be non-negative. In particular

$$|f(x) - f^\star(x)| \geq |f(x) - \bar{f}(x)| - |\bar{f}(x) - f^\star(x)| \geq \tau - \varepsilon_\infty \geq \varepsilon_\infty \geq 0$$

Therefore,

$$(f(x) - f^\star(x))^2 - (\bar{f}(x) - f^\star(x))^2 \geq (\tau - \varepsilon_\infty)^2 - \varepsilon_\infty^2 \geq \tau^2 - 2\tau\varepsilon_\infty.$$

The right hand side is non-negative whenever $\tau \geq 2\varepsilon$. □

Lemma 3.2 (Concentration). *Fix $\delta \in (0,1)$ and $\tau \geq 3\varepsilon_\infty$ and define $\varepsilon_{\text{stat}} := \frac{80\log(|\mathcal{F}|/\delta)}{3n}$. Under Assumption 2.3, for any $\bar{f} \in \mathcal{F}$ such that $\|\bar{f} - f^\star\|_\infty \leq \varepsilon_\infty$, with probability at least $1-\delta$ we have*

$$\forall f \in \mathcal{F}: \quad \mathcal{L}(f;\bar{f}) \leq 2\widehat{\mathcal{L}}(f;\bar{f}) + \varepsilon_{\text{stat}}, \qquad \text{and equivalently,} \qquad \widehat{\mathcal{L}}(\bar{f};f) \leq \frac{1}{2}\big(\mathcal{L}(\bar{f};f) + \varepsilon_{\text{stat}}\big).$$

Proof of Lemma 3.2. The concentration inequality is similar to the one used in the proof of Proposition 2.1. We apply Bernstein's inequality and a union bound to the empirical disagreement-based loss $\widehat{\mathcal{L}}(f;\bar{f})$ for each $f \in \mathcal{F}$. To do so, we must calculate the mean, variance, and range of $\widehat{\mathcal{L}}(f;\bar{f})$. Note that by the same calculation as in the proof of Proposition 2.1, we have that $\mathbb{E}[\widehat{\mathcal{L}}(f;\bar{f})] = \mathcal{L}(f;\bar{f})$ and that the range of each random variable in the empirical average is 1. The variance calculation however is slightly different:

$$\begin{aligned}
\operatorname{Var}[W^\tau_{f,\bar{f}}(x)\cdot\{(f(x)-y)^2 - (\bar{f}(x)-y)^2\}] &\leq \mathbb{E}[W^\tau_{f,\bar{f}}(x)\cdot\{(f(x)-y)^2 - (\bar{f}(x)-y)^2\}^2] \\
&\leq \mathbb{E}[W^\tau_{f,\bar{f}}(x)(f(x)-\bar{f}(x))^2(f(x)+\bar{f}(x)-2y)^2] \\
&\leq 4\mathbb{E}[W^\tau_{f,\bar{f}}(x)(f(x)-\bar{f}(x))^2].
\end{aligned}$$

Next, we consider a fixed x and define $a := (f(x) - f^\star(x))$ and $b := (f^\star(x) - \bar{f}(x))$, so that we can write $(f(x)-\bar{f}(x))^2 = (f(x)-f^\star(x)+f^\star(x)-\bar{f}(x))^2 = (a+b)^2$. Now, when $\tau \geq 3\varepsilon_\infty$ we have:

$$W^\tau_{f,\bar{f}}(x) = 1 \Rightarrow |a| = |f(x)-f^\star(x)| \geq |f(x)-\bar{f}(x)| - \varepsilon_\infty \geq 2\varepsilon_\infty.$$

Along with the fact that $|b| = |\bar{f}(x) - f^\star(x)| \leq \varepsilon_\infty$, this implies that $|b| \leq |a|/2$ or equivalently that $b^2 \leq a^2/4$. Using this, we can deduce that

$$(a+b)^2 \leq \frac{9a^2}{4} \leq \frac{9a^2}{4} - 3b^2 + \frac{3a^2}{2} = 3(a^2-b^2).$$

Re-introducing the definitions for a and b we have that

$$\operatorname{Var}[W^\tau_{f,\bar{f}}(x)\cdot\{(f(x)-y)^2 - (\bar{f}(x)-y)^2\}] \leq 12\mathcal{L}(f;\bar{f})$$

Now, applying Bernstein's inequality and a union bound over all $f \in \mathcal{F}$ yields that with probability $1-\delta$:

$$\begin{aligned}
\forall f \in \mathcal{F}: \mathcal{L}(f;\bar{f}) - \widehat{\mathcal{L}}(f;\bar{f}) &\leq \sqrt{\frac{24\mathcal{L}(f;\bar{f})\log(|\mathcal{F}|/\delta)}{n}} + \frac{4\log(|\mathcal{F}|/\delta)}{3n} \\
&\leq \frac{1}{2}\mathcal{L}(f;\bar{f}) + \frac{40\log(|\mathcal{F}|/\delta)}{3n}.
\end{aligned}$$

Re-arranging proves the first statement, and the second statement follows from the symmetries $\widehat{\mathcal{L}}(f;g) = -\widehat{\mathcal{L}}(g;f)$ and $\mathcal{L}(f;g) = -\mathcal{L}(g;f)$. □

Corollary 2.1 (Covariate shift for DBR). *Fix $\delta \in (0,1)$. Under Assumption 2.1–Assumption 2.4, with probability at least $1-\delta$, $\hat{f}^{(n)}_{\text{DBR}}$ with $\tau = 3\varepsilon_\infty$ satisfies*

$$R_{\text{test}}(\hat{f}^{(n)}_{\text{DBR}}) \leq 17\varepsilon_\infty^2 + O\left(C_\infty \frac{\log(|\mathcal{F}|/\delta)}{n}\right). \tag{5}$$

Proof of Corollary 2.1. Beginning the with risk under $\mathcal{D}_{\text{test}}$ and assuming that $\tau = 3\varepsilon_\infty$ we can write

$$\begin{aligned}
R_{\text{test}}(\hat{f}^{(n)}_{\mathsf{DBR}}) &= \mathbb{E}_{\text{test}}[(\hat{f}^{(n)}_{\mathsf{DBR}}(x) - f^\star(x))^2] \\
&= \mathbb{E}_{\text{test}}[\mathbb{1}\{|\hat{f}^{(n)}_{\mathsf{DBR}}(x) - f^\star(x)| < 4\varepsilon_\infty\} \cdot (\hat{f}^{(n)}_{\mathsf{DBR}}(x) - f^\star(x))^2] \\
&\quad + \mathbb{E}_{\text{test}}[\mathbb{1}\{|\hat{f}^{(n)}_{\mathsf{DBR}}(x) - f^\star(x)| \geq 4\varepsilon_\infty\} \cdot (\hat{f}^{(n)}_{\mathsf{DBR}}(x) - f^\star(x))^2] \\
&\leq 16\varepsilon_\infty^2 + \mathbb{E}_{\text{test}}[\mathbb{1}\{|\hat{f}^{(n)}_{\mathsf{DBR}}(x) - f^\star(x)| \geq 4\varepsilon_\infty\} \cdot (\hat{f}^{(n)}_{\mathsf{DBR}}(x) - f^\star(x))^2] \\
&\leq 17\varepsilon_\infty^2 + \mathbb{E}_{\text{test}}[\mathbb{1}\{|\hat{f}^{(n)}_{\mathsf{DBR}}(x) - f^\star(x)| \geq 4\varepsilon_\infty\} \cdot \{(\hat{f}^{(n)}_{\mathsf{DBR}}(x) - f^\star(x))^2 - \varepsilon_\infty^2\}].
\end{aligned}$$

Note that, due to the indicator, the quantity inside the expectation is non-negative. Therefore, via exactly the same importance weighting argument as we used in the proof of Proposition 2.1, the latter is at most C_∞ times the quantity bounded in Eq. (3). □

Corollary 2.2 (Well-specified case)**.** *Fix $\delta \in (0,1)$. Under Assumption 2.1–Assumption 2.4 (with $\varepsilon_\infty = 0$), with probability at least $1-\delta$, $\hat{f}^{(n)}_{\mathsf{DBR}}$ with $\tau \leq O\left(\sqrt{\log(|\mathcal{F}|/\delta)/n}\right)$ satisfies*

$$R_{\text{train}}(\hat{f}^{(n)}_{\mathsf{DBR}}) \leq O\left(\frac{\log(|\mathcal{F}|/\delta)}{n}\right) \quad \text{and} \quad R_{\text{test}}(\hat{f}^{(n)}_{\mathsf{DBR}}) \leq O\left(C_\infty \frac{\log(|\mathcal{F}|/\delta)}{n}\right). \tag{6}$$

Proof of Corollary 2.2. Let Δ denote the right hand side of Eq. (3). Note that in the well-specified case where $\varepsilon_\infty = 0$, Theorem 2.1 ensures that

$$\mathbb{E}_{\text{train}}[\mathbb{1}\{|\hat{f}^{(n)}_{\mathsf{DBR}}(x) - f^\star(x)| \geq \tau\} \cdot (\hat{f}^{(n)}_{\mathsf{DBR}}(x) - f^\star(x))^2] \leq \Delta.$$

Then, if we take $\tau \leq \sqrt{\Delta}$, we have

$$\begin{aligned}
R_{\text{train}}(\hat{f}^{(n)}_{\mathsf{DBR}}) &= \mathbb{E}_{\text{train}}[\mathbb{1}\{|\hat{f}^{(n)}_{\mathsf{DBR}}(x) - f^\star(x)| < \tau\} \cdot (\hat{f}^{(n)}_{\mathsf{DBR}}(x) - f^\star(x))^2] \\
&\quad + \mathbb{E}_{\text{train}}[\mathbb{1}\{|\hat{f}^{(n)}_{\mathsf{DBR}}(x) - f^\star(x)| \geq \tau\} \cdot (\hat{f}^{(n)}_{\mathsf{DBR}}(x) - f^\star(x))^2] \\
&\leq \tau^2 + \Delta \leq 2\Delta.
\end{aligned}$$

This proves the corollary. □

B.4. Extensions

In this section, we provide two results mentioned in Section 2. First we improve the approximation factor in Corollary 2.1 from 17 to 10 albeit at the cost of a worse statistical term. Second we show how to choose τ in a data-driven fashion to adapt to unknown misspecification level ε_∞.

Proposition B.1 (Improved approximation factor)**.** *Under Assumption 2.1–Assumption 2.4, with $\tau = 2\varepsilon_\infty$ and for $\delta \in (0,1)$, we have that, with probability at least $1-\delta$:*

$$R_{\text{test}}(\hat{f}^{(n)}_{\mathsf{DBR}}) \leq 10\varepsilon_\infty^2 + C_\infty \cdot O\left(\sqrt{\frac{\log(|\mathcal{F}|/\delta)}{n}}\right). \tag{10}$$

Proof sketch. The proof is essentially identical to that of Theorem 2.1, except that we replace the concentration statement of Lemma 3.2 with a simpler one that relies on Hoeffding's inequality. The new concentration statement is that for any $\tau \geq 0$ and $\delta \in (0,1)$ with probability $1-\delta$ we have

$$\forall f \in \mathcal{F}: \quad \mathcal{L}(f;\bar{f}) \leq \widehat{\mathcal{L}}(f;\bar{f}) + \varepsilon_{\mathrm{slow}},$$

where $\varepsilon_{\mathrm{slow}} := c\sqrt{\frac{\log(|\mathcal{F}|/\delta)}{n}}$ for some universal constant $c > 0$. This follows by a standard application of Hoeffding's inequality and a union bound, but importantly does not impose the restriction that $\tau \geq 3\varepsilon_\infty$.

Now the analysis to prove Theorem 2.1 yields that for any $\tau \geq 2\varepsilon_\infty$:

$$\mathbb{E}_{\mathrm{train}}\left[\mathbb{1}\{|\hat{f}^{(n)}_{\mathsf{DBR}}(x) - f^\star(x)| \geq \tau + \varepsilon_\infty\} \cdot \left\{(\hat{f}^{(n)}_{\mathsf{DBR}}(x) - f^\star(x))^2 - (\bar{f}(x) - f^\star(x))^2\right\}\right] \leq c\varepsilon_{\mathrm{slow}}.$$

Taking $\tau = 2\varepsilon_\infty$ and following the derivation used to prove Corollary 2.1, we get

$$R_{\mathrm{test}}(\hat{f}^{(n)}_{\mathsf{DBR}}) \leq 10\varepsilon_\infty^2 + c\varepsilon_{\mathrm{slow}}$$

(Note that this requires the non-negativity property provided by Lemma 3.1, which we still have.) □

The next result considers adapting to an unknown misspecification level.

Proposition B.2 (Adapting to ε_∞)**.** *Let $\delta \in (0,1)$ and define $S := \{2^i : \tau_{\min} \leq 2^i \leq \tau_{\max}\}$ where $\tau_{\min} := \sqrt{\frac{160\log(|\mathcal{F}||S|/\delta)}{3n}}$ and $\tau_{\max} := 1$. Let $\tau^\star := \min\{\tau \in S : \tau \geq 3\varepsilon_\infty\}$. Then there is an algorithm that, without knowledge of ε_∞ and with probability at least $1-\delta$, computes $\hat{f}$ satisfying*

$$\mathbb{E}_{\mathrm{train}}\left[\mathbb{1}\{|\hat{f}(x) - f^\star(x)| \geq \tau^\star + \varepsilon_\infty\} \cdot \left\{(\hat{f}(x) - f^\star(x))^2 - \varepsilon_\infty^2\right\}\right] \leq \frac{160\log(2|\mathcal{F}||S|/\delta)}{3n}.$$

Note that when $\varepsilon_\infty \ll \tau_{\min}$, we are essentially in the realizable regime. Thus, via the proof of Corollary 2.2 the above guarantee with $\tau^\star := \tau_{\min}$ suffices. On the other hand if $\varepsilon_\infty \geq 1/3$ then $\tau^\star$ is undefined, but due to Assumption 2.4 the guarantee in Theorem 2.1 is vacuous. Thus, the above theorem recovers essentially the same result as Theorem 2.1, but without knowledge of ε_∞.

Proof sketch. The algorithm is as follows. We run a slight variation of disagreement based regression for each $\tau \in S$: Instead of computing the minimizer of the objective in Eq. (2) we form the version space of near-minimizers. Specifically, define

$$\forall \tau \in S: \quad F_\tau := \left\{f \in \mathcal{F} : \max_{g \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n W^\tau_{f,g}(x_i)\left\{(f(x_i)-y_i)^2 - (g(x_i)-y_i)^2\right\} \leq \varepsilon_{\mathrm{stat}}/2\right\},$$

where we define $\varepsilon_{\mathrm{stat}} = \frac{80\log(|\mathcal{F}||S|/\delta)}{3n}$. Note this is slightly inflated from the definition in the statement of Lemma 3.2, which accounts for a union bound over all $|S|$ runs of the algorithm. Next, we define

$$\hat{\tau} := \arg\min\left\{\tau \in S : \bigcap_{\tau' \in S : \tau' \geq \tau} F_{\tau'} \neq \emptyset\right\},$$

and return any function in this intersection, i.e., let $\hat{f}$ be any function in $\bigcap_{\tau' \in S : \tau' \geq \hat{\tau}} F_{\tau'}$.

For the analysis, via the analysis of Theorem 2.1 and a union bound over the $|S|$ choices for τ, we have

$$\forall \tau \geq \tau^\star : \quad \bar{f} \in F_\tau \quad \text{and} \quad f \in F_\tau \Rightarrow \mathcal{L}^\tau(f; \bar{f}) \leq \varepsilon_{\text{stat}},$$

where $\mathcal{L}^\tau(f; g)$ is the population objective with parameter τ. The first statement directly implies that $\hat{\tau} \leq \tau^\star$. This in turn implies that $\hat{f} \in F_{\tau^\star}$ and so $\hat{f}$ achieves the same statistical guarantee as if we ran DBR with parameter $\tau^\star$ (up to the additional union bound). □

Appendix C. Proofs for Section 4

C.1. Offline RL

Theorem 4.1 (DBR for offline RL). *Fix $\delta \in (0, 1)$, assume that $\mathcal{F}$ is L_∞-misspecified and μ satisfies concentrability (as defined above). Consider the algorithm defined in Eq. (7) with $\tau = 3\varepsilon_\infty$. Then, with probability at least $1 - \delta$ we have*

$$J(\pi^\star) - J(\hat{\pi}) \leq O\left(\frac{\varepsilon_\infty}{1-\gamma} + \frac{1}{1-\gamma} \sqrt{C_{\text{conc}} \frac{\log(|\mathcal{F}|/\delta)}{n}} \right).$$

Proof of Theorem 4.1. For each "target" function $f_{\text{trg}} \in \mathcal{F}$ such that $f_{\text{trg}} \neq \bar{f}$, let us define $\mathsf{apx}[f_{\text{trg}}] \in \mathcal{F}$ to be *any* approximation to the Bellman backup $\mathcal{T} f_{\text{trg}}$ s.t. $\|\mathsf{apx}[f_{\text{trg}}] - \mathcal{T} f_{\text{trg}}\|_\infty \leq \varepsilon_\infty$. Define $\mathsf{apx}[\bar{f}] = \bar{f}$, which also satisfies $\|\mathsf{apx}[\bar{f}] - \mathcal{T}\bar{f}\|_\infty \leq \varepsilon_\infty$ by assumption. Let us define the empirical and population losses for the disagreement-based regression problem with regression targets derived from f_{trg}.

$$\text{(Empirical)} \quad \widehat{\mathcal{L}}_{f_{\text{trg}}}(f; g) := \frac{1}{n} \sum_{i=1}^n W^\tau_{f,g}(s_i, a_i) \big\{ (f(s_i, a_i) - y_{f_{\text{trg}},i})^2 - (g(s_i, a_i) - y_{f_{\text{trg}},i})^2 \big\},$$

$$\text{(Population)} \quad \mathcal{L}_{f_{\text{trg}}}(f; g) := \mathbb{E}_\mu \big[W^\tau_{f,g}(s, a) \big\{ (f(s, a) - y_{f_{\text{trg}}})^2 - (g(s, a) - y_{f_{\text{trg}}})^2 \big\} \big].$$

Here recall that $y_{f_{\text{trg}}} := r + \max_{a'} f_{\text{trg}}(s', a')$ is derived from the sample (s, a, r, s'). Also note that we use $\mathbb{E}_\mu[\cdot]$ to denote expectation with respect to the data collection policy.

First, we apply Lemma 3.1 and Lemma 3.2 to each of the $|\mathcal{F}|$ regression problems. By approximate completeness and the definition of $\mathsf{apx}[f_{\text{trg}}]$ this yields

$$\forall f_{\text{trg}}, f \in \mathcal{F} : \quad 0 \leq \mathcal{L}_{f_{\text{trg}}}(f; \mathsf{apx}[f_{\text{trg}}]) \leq 2\widehat{\mathcal{L}}_{f_{\text{trg}}}(f; \mathsf{apx}[f_{\text{trg}}]) + \varepsilon_{\text{stat}}, \tag{11}$$

where $\varepsilon_{\text{stat}} := \frac{160 \log(|\mathcal{F}|/\delta)}{3n}$. The above uniform bound holds with probability $1 - \delta$. Note that this $\varepsilon_{\text{stat}}$ is twice as large as the one in the proof of Theorem 2.1, which accounts for the additional union bound over all $|\mathcal{F}|$ regression problems.

The main statistical guarantee for $\hat{f}$ is derived as follows

$$\mathcal{L}_{\hat{f}}(\hat{f}; \mathsf{apx}[\hat{f}]) \overset{\text{(i)}}{\leq} 2\widehat{\mathcal{L}}_{\hat{f}}(\hat{f}; \mathsf{apx}[\hat{f}]) + \varepsilon_{\text{stat}} \overset{\text{(ii)}}{\leq} 2 \max_{g \in \mathcal{F}} \widehat{\mathcal{L}}_{\hat{f}}(\hat{f}; g) + \varepsilon_{\text{stat}}$$
$$\overset{\text{(iii)}}{\leq} 2 \max_{g \in \mathcal{F}} \widehat{\mathcal{L}}_{\bar{f}}(\bar{f}; g) + \varepsilon_{\text{stat}} \overset{\text{(iv)}}{\leq} 2\varepsilon_{\text{stat}}.$$

Here (i) is the second inequality in Eq. (11), (ii) follows since $\mathsf{apx}[\hat{f}] \in \mathcal{F}$, (iii) uses the optimality property of $\hat{f}$, and (iv) uses Eq. (11) again, noting the symmetry of $\mathcal{L}_{\bar{f}}(\cdot;\cdot)$ and using $\mathsf{apx}[\bar{f}] = \bar{f}$.

Since the Bayes regression function defined by targets $y_{\hat{f}}$ is $\mathcal{T}\hat{f}$, this yields

$$\mathbb{E}_\mu\Big[\mathbb{1}\Big\{|\hat{f}(s,a) - \mathsf{apx}[\hat{f}](s,a)| \geq 3\varepsilon_\infty\Big\} \cdot \Big\{(\hat{f}(s,a) - [\mathcal{T}\hat{f}](s,a))^2 - (\mathsf{apx}[\hat{f}](s,a) - [\mathcal{T}\hat{f}](s,a))^2\Big\}\Big] \leq 2\varepsilon_{\mathrm{stat}}. \tag{12}$$

We translate this to the squared Bellman error on any other distribution $\nu \in \Delta(\mathcal{X} \times \mathcal{A})$ via a slightly stronger argument than the one used to prove Corollary 2.1.

$$\begin{aligned}
\mathbb{E}_\nu\Big[\Big|\hat{f}(s,a) - [\mathcal{T}\hat{f}](s,a)\Big|\Big] &\leq \varepsilon_\infty + \mathbb{E}_\nu\Big[\Big|\hat{f}(s,a) - \mathsf{apx}[\hat{f}](s,a)\Big|\Big] \\
&\leq 4\varepsilon_\infty + \mathbb{E}_\nu\Big[\mathbb{1}\Big\{|\hat{f}(s,a) - \mathsf{apx}[\hat{f}](s,a)| \geq 3\varepsilon_\infty\Big\} \cdot \Big|\hat{f}(s,a) - \mathsf{apx}[\hat{f}](s,a)\Big|\Big] \\
&\leq 4\varepsilon_\infty + \sqrt{\mathbb{E}_\mu\left[\left(\frac{\nu(s,a)}{\mu(s,a)}\right)^2\right]} \cdot \sqrt{\mathbb{E}_\mu\left[\mathbb{1}\Big\{|\hat{f}(s,a) - \mathsf{apx}[\hat{f}](s,a)| \geq 3\varepsilon_\infty\Big\} \cdot \Big(\hat{f}(s,a) - \mathsf{apx}[\hat{f}](s,a)\Big)^2\right]} \\
&= 4\varepsilon_\infty + \|\nu/\mu\|_{L_2(\mu)} \cdot \sqrt{\mathbb{E}_\mu\left[\mathbb{1}\Big\{|\hat{f}(s,a) - \mathsf{apx}[\hat{f}](s,a)| \geq 3\varepsilon_\infty\Big\} \cdot \Big(\hat{f}(s,a) - \mathsf{apx}[\hat{f}](s,a)\Big)^2\right]} \\
&\leq 4\varepsilon_\infty + \|\nu/\mu\|_{L_2(\mu)} \cdot \sqrt{6\varepsilon_{\mathrm{stat}}}.
\end{aligned}$$

The last inequality is based on the "self-bounding" argument we used to control the variance in the proof of Lemma 3.2, which showed that under the event $|\hat{f}(s,a) - \mathsf{apx}[\hat{f}](s,a)| \geq 3\varepsilon_\infty$:

$$\Big(\hat{f}(s,a) - \mathsf{apx}[\hat{f}](s,a)\Big)^2 \leq 3 \cdot \Big\{\Big(\hat{f}(s,a) - [\mathcal{T}\hat{f}](s,a)\Big)^2 - \Big(\mathsf{apx}[\hat{f}](s,a) - [\mathcal{T}\hat{f}](s,a)\Big)^2\Big\}.$$

Note that $\|\nu/\mu\|^2_{L_2(\mu)} \leq \|\nu/\mu\|_\infty$ since $\mathbb{E}_\mu[\nu(s,a)/\mu(s,a)] = \mathbb{E}_\nu[1] = 1$.

Finally, we appeal to the telescoping performance difference lemma (c.f., Xie and Jiang, 2020, Theorem 2), which states that for an action-value function f,

$$J(\pi^\star) - J(\pi_f) \leq \frac{\mathbb{E}_{d^{\pi^\star}}[[\mathcal{T}f](s,a) - f(s,a)]}{1-\gamma} + \frac{\mathbb{E}_{d^{\pi_f}}[f(s,a) - [\mathcal{T}f](s,a)]}{1-\gamma},$$

where $d^\pi := (1-\gamma)\sum_{h=0}^\infty \gamma^h d^\pi_h$. Both terms are controlled by the distribution shift argument above and the concentrability coefficient, yielding the theorem. □

C.2. Online RL

Theorem 4.2 (DBR for online RL). *Fix $\delta \in (0,1)$, and assume that $\mathcal{F}$ is L_∞-misspecified and coverability (as defined above) holds. Consider* GOLF.DBR *with $\tau = 3\varepsilon_\infty$ and $\beta = c\log(TH|\mathcal{F}|/\delta)$. Then, with probability at least $1-\delta$, we have*

$$\mathrm{Reg} \leq O\Big(\varepsilon_\infty HT + H\sqrt{C_{\mathrm{cov}} T \log(TH|\mathcal{F}|/\delta)\log(T)}\Big).$$

Proof of Theorem 4.2. The proof makes essentially two modifications to the proof of Theorem 1 of Xie et al. (2023). The first step is a concentration argument, which is essentially a martingale version of Theorem 2.1. The second is the distribution shift argument, which is very similar to the one we used to prove Theorem 4.1. To keep the presentation concise, we focus on these arguments, and explain how they fit into the analysis of Xie et al. (2023), but we do not provide a self-contained proof.

Notation. We adopt the following notation. Recall that $\mathcal{F}^{(t-1)}$ is the version space used in episode t and that $f^{(t)} \in \mathcal{F}^{(t-1)}$ induces the policy $\pi^{(t)}$ deployed in the episode. As before, let $\mathsf{apx}[f_{h+1}] \in \mathcal{F}_h$ denote the L_∞-approximation to $\mathcal{T}_h f_{h+1}$. For each episode t let

$$\delta_h^{(t)}(\cdot) := f_h^{(t)}(\cdot) - [\mathcal{T}_h f_{h+1}^{(t)}](\cdot) \qquad \text{and}$$
$$\mathrm{err}^{(t)}(\cdot) := \mathbb{1}\left\{\left|f_h^{(t)}(\cdot) - \mathsf{apx}[f_{h+1}^{(t)}](\cdot)\right| \geq 3\varepsilon_\infty\right\} \cdot \left\{\left(f_h^{(t)}(\cdot) - [\mathcal{T}_h f_{h+1}^{(t)}](\cdot)\right)^2 - \left(\mathsf{apx}[f_{h+1}^{(t)}](\cdot) - [\mathcal{T}_h f_{h+1}^{(t)}](\cdot)\right)^2\right\}.$$

Let $d_h^{(t)} = d_h^{\pi^{(t)}}$ and define $\widetilde{d}_h^{(t)}(x,a) = \sum_{i=1}^{t-1} d_h^{(i)}(x,a)$ and $\mu_h^\star$ to be the distribution that achieves the value C_{cov} for layer h.

Concentration. By a martingale version of Theorem 2.1, we can show that with probability at least $1-\delta$, for all $t \in [T]$:

$$(i)\ \ \bar{f} \in \mathcal{F}^{(t)}, \quad \text{and} \quad (ii)\ \ \forall h \in [H]: \ \sum_{s,a} \widetilde{d}_h^{(t)}(s,a)\mathrm{err}_h^{(t)}(s,a) \leq O(\beta), \tag{13}$$

where $\beta = c\log(TH|\mathcal{F}|/\delta)$. We do not provide a complete proof of this statement, noting that it is essentially the same guarantee as in Eq. (12), except that (a) it is a non-stationary version with a union bound over each time step h and episode t and (b) it uses martingale concentration (i.e., Freedman's inequality instead of Bernstein's inequality). It is also worth comparing with the concentration guarantee of (Xie et al., 2023) under exact realizability/completeness, which is that $Q^\star \in \mathcal{F}^{(t)}$ and that $\sum_{s,a} \widetilde{d}_h^{(t)}(s,a)(\delta_h^{(t)}(s,a))^2 \leq O(\beta)$.

Distribution shift. To bound the regret, first note that by a simple inductive argument, we have that $\|\bar{f}_1 - Q_1^\star\|_\infty \leq H\varepsilon_\infty$. Thus, under the event in Eq. (13) which ensures that $\bar{f} \in \mathcal{F}^{(t)}$ for each t, we have that $f^{(t)}$ is approximately optimistic in the sense that

$$J(\pi^\star) := \mathbb{E}[Q_1^\star(s_1, \pi_1^\star(s_1))] \leq \mathbb{E}[\bar{f}_1(s_1, \pi_{\bar{f},1}(s_1))] + H\varepsilon_\infty \leq \mathbb{E}[f_1^{(t)}(s_1, \pi_{f^{(t)},1}(s_1))] + H\varepsilon_\infty.$$

Thus, we can bound the regret by

$$\mathrm{Reg} \leq TH\varepsilon_\infty + \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{(s,a)\sim d_h^{(t)}}\left[\delta_h^{(t)}(s,a)\right].$$

For distribution shift, we must translate the above on-policy Bellman errors to the "DBR" errors on the historical data $\widetilde{d}_h^{(t)}$, which is controlled by Eq. (13). Following (Xie et al., 2023) we consider *burn-in* and *stable* phases. Let

$$\gamma_h(s,a) := \min\left\{t : \widetilde{d}_h^{(t)}(s,a) \geq C_{\mathrm{cov}} \cdot \mu_h^\star(s,a)\right\},$$

and decompose

$$\sum_{t=1}^{T} \mathbb{E}_{(s,a)\sim d_h^{(t)}}\left[\delta_h^{(t)}(s,a)\right] = \sum_{t=1}^{T} \mathbb{E}_{(s,a)\sim d_h^{(t)}}\left[\delta_h^{(t)}(s,a)\mathbb{1}\{t < \gamma_h(s,a)\}\right] + \mathbb{E}_{(s,a)\sim d_h^{(t)}}\left[\delta_h^{(t)}(s,a)\mathbb{1}\{t \geq \gamma_h(s,a)\}\right].$$

The first term is the regret incurred during the burn-in phase, which is bounded by $2C_{\mathrm{cov}}$ following exactly the argument of Xie et al. (2023). This contributes a total regret of $2HC_{\mathrm{cov}}$.

The second term is the regret incurred during the stable phase, for which we must perform a distribution shift argument. To condense the notation, define

$$\bar{\delta}_h^{(t)}(\cdot) := \mathsf{apx}[f_{h+1}^{(t)}](\cdot) - [\mathcal{T}_h f_{h+1}^{(t)}](\cdot), \quad \text{and} \quad \tilde{\delta}_h^{(t)}(\cdot) := f_h^{(t)}(\cdot) - \mathsf{apx}[f_{h+1}^{(t)}](\cdot).$$

Note that, by assumption, $\left|\bar{\delta}_h^{(t)}(s,a)\right| \leq \varepsilon_\infty$. Then,

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}_{d_h^{(t)}}\left[\delta_h^{(t)}(s,a)\mathbb{1}\{t > \gamma_h(s,a)\}\right] &= \sum_{t=1}^{T} \mathbb{E}_{d_h^{(t)}}\left[\left(\tilde{\delta}_h^{(t)}(s,a) + \bar{\delta}_h^{(t)}(s,a)\right)\mathbb{1}\{t > \gamma_h(s,a)\}\right] \\
&\leq \sum_{t=1}^{T} \mathbb{E}_{d_h^{(t)}}\left[\tilde{\delta}_h^{(t)}(s,a)\mathbb{1}\{t > \gamma_h(s,a)\}\right] + T\varepsilon_\infty \\
&\leq \sum_{t=1}^{T} \mathbb{E}_{d_h^{(t)}}\left[\mathbb{1}\left\{|\tilde{\delta}_h^{(t)}(s,a)| \geq 3\varepsilon_\infty, t > \gamma_h(s,a)\right\}\tilde{\delta}_h^{(t)}(s,a)\right] + 4T\varepsilon_\infty.
\end{aligned}$$

We proceed by applying the Cauchy-Schwarz inequality to the first term:

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}_{d_h^{(t)}}\left[\mathbb{1}\left\{|\tilde{\delta}_h^{(t)}(s,a)| \geq 3\varepsilon_\infty\right\}\tilde{\delta}_h^{(t)}(s,a)\mathbb{1}\{t > \gamma_h(s,a)\}\right] \\
&\leq \sqrt{\sum_{t=1}^{T}\sum_{s,a} \frac{\left(\mathbb{1}\{t > \gamma_h(s,a)\}d_h^{(t)}(s,a)\right)^2}{\widetilde{d}_h^{(t)}(s,a)}} \cdot \sqrt{\sum_{t=1}^{T}\sum_{x,a} \widetilde{d}_h^{(t)}(x,a)\mathbb{1}\left\{|\tilde{\delta}_h^{(t)}(s,a)| \geq 3\varepsilon_\infty\right\}(\tilde{\delta}_h^{(t)}(s,a))^2} \\
&\leq \sqrt{\sum_{t=1}^{T}\sum_{s,a} \frac{\left(\mathbb{1}\{t > \gamma_h(s,a)\}d_h^{(t)}(s,a)\right)^2}{\widetilde{d}_h^{(t)}(s,a)}} \cdot \sqrt{3\sum_{t=1}^{T}\sum_{x,a} \widetilde{d}_h^{(t)}(x,a)\mathrm{err}_h^{(t)}(s,a)}.
\end{aligned}$$

The final inequality follows from the self-bounding property that we used in the proof of Lemma 3.2 and Theorem 4.1. In particular under the event that $\left|\tilde{\delta}_h^{(t)}(s,a)\right| \geq 3\varepsilon_\infty$, we can bound $(\tilde{\delta}_h^{(t)}(s,a))^2 \leq 3\left((\delta_h^{(t)}(s,a))^2 - (\bar{\delta}_h^{(t)}(s,a))^2\right)$. Thus we have converted from the on-policy Bellman error to the historical "DBR" errors, i.e., we bound the regret in the stable phase by

$$\leq \sqrt{\sum_{t=1}^{T}\sum_{s,a} \frac{\left(\mathbb{1}\{t > \gamma_h(s,a)\}d_h^{(t)}(s,a)\right)^2}{\widetilde{d}_h^{(t)}(s,a)}} \cdot O\left(\sqrt{\beta T}\right) + 4T\varepsilon_\infty.$$

Meanwhile the density ratio term is bounded by $O(\sqrt{C_{\mathrm{cov}}\log(T)})$ via the analysis of Xie et al. (2023). Repeating this analysis for each time step h proves the theorem. □