FITTED Q-EVALUATION WITHOUT BELLMAN COMPLETENESS VIA OCCUPANCY WEIGHTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Fitted Q-evaluation (FQE) is a standard regression-based method for off-policy evaluation, but under distribution shift its projected Bellman operator need not be contractive under the offline distribution. Consequently, value-function realizability alone does not ensure convergence, and existing analyses often impose Bellman completeness or related structural conditions. We study occupancy-weighted FQE and show that distribution weighting provides a different route to stability. Weighting each Bellman regression by the density ratio of a suitably discounted target-policy occupancy distribution to the offline distribution aligns the regression norm with the target-policy dynamics, making the population projected Bellman operator contractive. With exact occupancy weights, this removes the need for Bellman completeness. Our main contribution is a finite-sample analysis showing how this population stability extends to estimated weights, nonlinear function classes, and misspecification. The bounds separate finite-iteration, statistical, approximation, and ratio-estimation errors and show that ratio-estimation error enters through its interaction with the Bellman projection residual rather than as a separate additive term. This interaction can be controlled by approximate Bellman completeness or value-function realizability; under exact realizability, the effect of ratio-estimation error is higher order, while under exact Bellman completeness it vanishes. We also analyze an end-to-end procedure combining occupancy-weighted FQE with fitted occupancy-ratio evaluation, obtaining consistency under coverage and joint realizability of the value function and occupancy ratio without completeness assumptions on either class. Controlled experiments illustrate the contraction mechanism and the tradeoff between target alignment and coverage.

1 INTRODUCTION

Fitted Q-evaluation (FQE) is a standard regression-based method for policy evaluation in reinforcement learning (Tsitsiklis & Van Roy, 1996; Gordon, 1995; Lagoudakis & Parr, 2003; Ernst et al., 2005; Munos, 2005; Munos & Szepesvári, 2008; Farahmand et al., 2010; Bertsekas, 2011). FQE approximates Bellman fixed-point iteration by repeatedly forming one-step targets from the current value-function estimate and regressing them onto a function class. It is therefore simple to implement and readily accommodates off-the-shelf supervised learning methods (Voloshin et al., 2019; Fujimoto et al., 2019; Le et al., 2019; Agarwal et al., 2021). Under off-policy distribution shift, however, value-function realizability alone does not ensure convergence, and FQE can fail even with linear function approximation (Baird, 1995; Gordon, 1995; Tsitsiklis & Van Roy, 1996).

This instability arises from a mismatch between the geometry of regression and that of Bellman contraction. At the population level, FQE iterates a projected Bellman operator: each Bellman update is projected onto the fitted function class in the L^2 norm induced by the offline sampling distribution. The Bellman operator, by contrast, is contractive in the supremum norm and in target-policy occupancy norms, but not generally in the offline-sampling norm (Tsitsiklis & Van Roy, 1996; Patterson et al., 2022). The resulting projected Bellman operator therefore need not be contractive, and the classical convergence guarantees for exact value iteration no longer apply.

Contractivity, and hence geometric convergence, is commonly obtained through *Bellman completeness*, which requires the Bellman operator to map the fitted value-function class into itself. Projection is then inactive on Bellman images, so FQE inherits the contraction of exact value iteration. Related

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

analyses replace exact completeness with direct stability assumptions on the projected Bellman operator or with approximate completeness conditions that bound the *inherent Bellman error*, namely, how far Bellman images of functions in the fitted class lie outside that class (Antos et al., 2007; Munos & Szepesvári, 2008; Amortila et al., 2020; Wang et al., 2021b;a; Chang et al., 2022).

Bellman completeness is restrictive under function approximation and behaves differently from an ordinary approximation condition. In particular, it is non-monotone in the fitted class. Enlarging the class can reduce approximation error while worsening completeness because the Bellman images of the newly added functions must also be represented (Zhan et al., 2022; Zanette, 2023). The need for assumptions beyond value-function realizability and coverage is not merely a feature of existing analyses. Chen & Jiang (2019) emphasized the gap between realizability and Bellman completeness, and Foster et al. (2021) showed that standard concentrability and value-function realizability alone are insufficient for sample-efficient offline policy evaluation and learning. Minimax methods provide one alternative by replacing Bellman completeness of the value class with richness conditions on an auxiliary critic class (Uehara et al., 2020; 2021; Zanette & Wainwright, 2022). FQE nevertheless remains widely used for offline policy evaluation because it is simple to implement and readily accommodates flexible supervised-learning methods (Voloshin et al., 2019; Tang & Wiens, 2021).

Another way to stabilize projected Bellman methods under distribution shift is to change the distribution used for projection. Off-policy temporal-difference methods use emphatic or density-ratio weighting to modify the effective sampling distribution (Mahmood et al., 2015; Sutton et al., 2016; Hallak & Mannor, 2017; Gelada & Bellemare, 2019; Sinha et al., 2022; Patterson et al., 2022). In this work, we study this idea for FQE. Specifically, we weight each Bellman regression by the density ratio of a target-policy occupancy distribution to the offline distribution. In a γ-discounted MDP, we use the β-discounted target occupancy, for which any $\beta \in (\gamma^2, 1)$ makes the population projected Bellman operator a $\gamma/\sqrt{\beta}$-contraction in the corresponding L^2 norm. Thus, occupancy weighting changes the projection geometry so that Bellman contraction holds without requiring Bellman completeness at the population level, while preserving the regression structure of FQE. We refer to the resulting procedure as *occupancy-weighted FQE*.

Our main contribution is a finite-sample analysis of occupancy-weighted FQE with estimated weights, nonlinear function approximation, and misspecification. The bounds separate finite-iteration, statistical, approximation, and ratio-estimation errors. With exact occupancy weights, the finite-sample guarantees likewise require no Bellman completeness. With estimated weights, ratio-estimation error enters only through its interaction with the Bellman projection residual rather than as a separate additive error term. Bellman completeness controls this interaction uniformly, whereas value-function realizability controls it near Q^π; under exact realizability, the effect of ratio-estimation error is higher order. Finally, we propose an end-to-end procedure that combines occupancy-weighted FQE with fitted occupancy-ratio evaluation (FORE). Under coverage and joint realizability of the value function and occupancy ratio, the resulting estimator is consistent without completeness assumptions on either class.

1.1 Related Work

Fitted value iteration. Classical analyses of fitted value iteration and FQE establish stability of projected value iteration or control the propagation of one-step approximation and estimation errors (Gordon, 1995; Tsitsiklis & Van Roy, 1996; Munos, 2005; Munos & Szepesvári, 2008; Farahmand et al., 2010; Bertsekas, 2011; Scherrer, 2014). Under function approximation, such guarantees typically require additional structure linking the Bellman operator, fitted class, and sampling distribution, including Bellman completeness, small inherent Bellman error, coverage or concentrability, or structured representations (Antos et al., 2007; Chen & Jiang, 2019; Amortila et al., 2020; Wang et al., 2021a; Chang et al., 2022; Van Roy, 2006; Xie & Jiang, 2021). Even with linear function approximation, projection in an arbitrary off-policy norm need not preserve contraction, and the projected fixed point can be arbitrarily inaccurate despite small approximation error for Q^π (Yu & Bertsekas, 2010; Kolter, 2011). Occupancy weighting instead aligns the regression geometry with a target-occupancy L^2 norm, yielding guarantees in terms of direct approximation of Q^π. Our results are therefore most relevant beyond tabular, linear-MDP, and Bellman-complete settings, where stronger structural assumptions already control fitted iteration (Melo & Ribeiro, 2007; Yang & Wang, 2020; Jin et al., 2023).

Weighted temporal-difference and projected Bellman methods. Our approach builds on a substantial literature that changes the effective sampling or projection distribution to stabilize off-policy value learning. Under linear function approximation, emphatic TD uses follow-on weights to obtain stable off-policy updates (Mahmood et al., 2015; Sutton et al., 2016; Yu, 2015; Yu et al., 2018), while stationary density-ratio methods correct the mismatch between behavior- and target-policy distributions (Hallak & Mannor, 2017; Gelada & Bellemare, 2019). Closest to our setting, Sinha et al. (2022) reweight experience-replay samples toward the current-policy stationary distribution in deep TD learning and motivate this weighting through Bellman contraction. Related work also studies alternative projection norms and generalized projected Bellman objectives (Yu & Bertsekas, 2012; Van Roy, 2006; Patterson et al., 2022). Existing theory is primarily developed for TD-style algorithms, linear function approximation, or population and asymptotic convergence, even when the density ratios themselves are estimated. Our contribution is a finite-sample analysis of occupancy weighting for the regression-based FQE recursion with nonlinear function classes, misspecification, and estimated occupancy weights. We further characterize how ratio-estimation error interacts with value-function realizability and Bellman completeness.

Minimax and adversarial Q-learning. Minimax and adversarial methods replace fitted value iteration with saddle-point objectives that test Bellman residuals against an auxiliary critic class (Dai et al., 2018; Uehara et al., 2020; Feng et al., 2020; Xie & Jiang, 2020; Xie et al., 2021; Uehara et al., 2021; Jin et al., 2021; Uehara et al., 2023; Zanette & Wainwright, 2022). These methods can avoid Bellman completeness of the value-function class, but introduce saddle-point optimization and require the critic class to be sufficiently rich to detect the Bellman errors generated by candidate value functions. In some analyses, this takes the form of conditions such as $\{\mathcal{T}Q - Q : Q \in \mathcal{Q}\} \subseteq c\mathcal{G}$, corresponding to critic-side completeness or related richness assumptions (Feng et al., 2019; Jiang & Huang, 2020; Uehara et al., 2021; Zhan et al., 2022). Occupancy-weighted FQE instead retains the regression-based FQE recursion and uses no auxiliary critic class. Combined with FORE, joint realizability of the Q-function and occupancy ratio is sufficient under coverage for consistent estimation of both functions. Existing minimax results under joint realizability primarily establish consistency of the scalar policy value, while consistent recovery of the nuisance functions generally requires additional critic-side structure (Uehara et al., 2021).

1.2 Preliminaries

We consider a discounted MDP with general state space $\mathcal{S}$, finite action space $\mathcal{A}$, transition kernel $P(\cdot \mid s, a)$, mean reward $r_0(s, a)$, and discount factor $\gamma \in [0, 1)$. A fixed target policy $\pi(a \mid s)$ induces the state-action transition operator $(P_\pi Q)(s, a) := \mathbb{E}[Q(S', A') \mid S = s, A = a]$, $S' \sim P(\cdot \mid s, a)$, $A' \sim \pi(\cdot \mid S')$. The Bellman operator for π and its unique fixed point are $\mathcal{T}Q := r_0 + \gamma P_\pi Q$, $Q^\pi = \mathcal{T}Q^\pi$. The fixed point Q^π is the target-policy Q-function, giving the expected discounted return after taking action a in state s and following π thereafter.

We observe transitions $\mathcal{D}_n = \{(S_i, A_i, R_i, S_i')\}_{i=1}^n$, where $S_i' \sim P(\cdot \mid S_i, A_i)$ and $R_i = r_0(S_i, A_i) + \varepsilon_i$, with $\mathbb{E}[\varepsilon_i \mid S_i, A_i] = 0$. We assume that (S_i, A_i) have common marginal distribution ν_b, while the transition tuples may be dependent. We refer to ν_b as the offline state-action distribution; under stationary trajectory sampling, it is the stationary state-action distribution induced by the behavior policy. For any state-action distribution ν, define $(\pi f)(s) := \sum_{a \in \mathcal{A}} \pi(a \mid s) f(s, a)$, $\|f\|_{2,\nu} := \{\int f(s, a)^2 \, \nu(ds, da)\}^{1/2}$.

1.3 Standard FQE and Norm Mismatch

Classical dynamic programming computes Q^π by iterating the Bellman operator $\mathcal{T}$ (Bellman, 1952; 1966). Because $\mathcal{T}$ is a γ-contraction in the supremum norm, the iterates converge geometrically to Q^π. With function approximation and offline data, FQE replaces exact Bellman iteration with fitted regression (Lagoudakis & Parr, 2003; Ernst et al., 2005; Munos & Szepesvári, 2008). Given a function class $\mathcal{F}$, its empirical update is

$$\widehat{Q}^{(k+1)} \in \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \left\{ R_i + \gamma(\pi \widehat{Q}^{(k)})(S_i') - f(S_i, A_i) \right\}^2 .$$

At the population level, FQE iterates the projected Bellman operator $\Pi_{\mathcal{F}}^{\nu_b} \mathcal{T}$, where $\Pi_{\mathcal{F}}^{\nu_b}$ denotes $L^2(\nu_b)$ projection onto $\mathcal{F}$. Unlike $\mathcal{T}$, this projected operator need not be contractive.

A standard assumption to recover stability is Bellman completeness, $\mathcal{T}\mathcal{F} \subseteq \mathcal{F}$. Then projection is inactive on Bellman images, so $\Pi_{\mathcal{F}}^{\nu_b}\mathcal{T}f = \mathcal{T}f$ for every $f \in \mathcal{F}$, and FQE inherits the contraction of exact Bellman iteration. Bellman completeness is therefore central to many analyses of FQE. Other analyses instead assume stability of the projected Bellman operator directly or control the inherent Bellman error $\sup_{f\in\mathcal{F}} \|\mathcal{T}f - \Pi_{\mathcal{F}}^{\nu_b}\mathcal{T}f\|_{L^2(\nu_b)}$ (Antos et al., 2007; Munos & Szepesvári, 2008; Chen & Jiang, 2019; Amortila et al., 2020; Wang et al., 2021b;a; Chang et al., 2022).

These conditions address the same underlying difficulty. Least-squares fitting and Bellman propagation are controlled in different geometries. Least-squares projection is defined in $L^2(\nu_b)$ and, for closed convex $\mathcal{F}$, is nonexpansive in that norm. By contrast, the Bellman operator is contractive in the supremum norm and in suitable target-policy occupancy norms (Tsitsiklis & Van Roy, 1996). In the off-policy setting, the offline and target-policy occupancy geometries may be misaligned, so $\Pi_{\mathcal{F}}^{\nu_b}\mathcal{T}$ need not inherit the Bellman contraction. We address this mismatch by projecting Bellman updates in a target-policy occupancy norm.

2 Occupancy-Weighted FQE

2.1 Algorithm

We address the norm mismatch by reweighting each FQE regression toward a target-policy occupancy distribution. Let ν_b^S denote the state marginal of ν_b, define $\xi_{\pi,b}(ds, da) := \nu_b^S(ds)\pi(da \mid s)$, and, for $\beta \in (0, 1)$, let $d_\beta := (1-\beta)\sum_{t=0}^{\infty}\beta^t\xi_{\pi,b}P_\pi^t$. When $d_\beta \ll \nu_b$, define the occupancy ratio $w_\beta := \mathrm{d}d_\beta/\mathrm{d}\nu_b$. Given an estimate $\widehat{w}_\beta$ and initialization $\widehat{Q}^{(0)} \in \mathcal{F}$, occupancy-weighted FQE iterates

$$\widehat{Q}^{(k+1)} \in \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \widehat{w}_\beta(S_i, A_i)\left\{R_i + \gamma(\pi\widehat{Q}^{(k)})(S_i') - f(S_i, A_i)\right\}^2.$$

Relative to standard FQE, the Bellman targets are unchanged; only the regression loss is reweighted. The estimated occupancy ratio is computed once and reused across iterations. We return to occupancy-ratio estimation in Section 2.3, where we develop an end-to-end procedure.

The regressions may be fit using any method that supports observation weights. For neural networks, for example, one may take one or more minibatch gradient steps with fixed Bellman targets before refreshing them, as in practical fitted value iteration (Mnih et al., 2013).

2.2 Population Geometry and Contraction

To understand why occupancy weighting stabilizes FQE, consider the corresponding population update with oracle weights. Weighting by w_β makes the regression objective equivalent to projection in $L^2(d_\beta)$. Define $\mathcal{T}_{\mathcal{F},\beta} := \Pi_{\mathcal{F},\beta}\mathcal{T}$, where $\Pi_{\mathcal{F},\beta}g := \arg\min_{f\in\mathcal{F}} \|f - g\|_{2,d_\beta}$. The resulting population recursion is $Q_{\mathcal{F},\beta}^{(k)} = \mathcal{T}_{\mathcal{F},\beta}Q_{\mathcal{F},\beta}^{(k-1)}$. Our first result extends the contraction of the projected Bellman operator under the target-policy stationary distribution (Tsitsiklis & Van Roy, 1996) to discounted target-policy occupancy distributions.

C1 Occupancy discount. The discount satisfies $\beta \in (\gamma^2, 1)$, and $r_0 \in L^2(d_\beta)$.

C2 Closed convex class. The class $\mathcal{F}$ is closed and convex in $L^2(d_\beta)$.

The choice of d_β restores contraction. From the occupancy recursion, $\beta d_\beta P_\pi \le d_\beta$, so Jensen's inequality implies that P_π expands the $L^2(d_\beta)$ norm by at most $\beta^{-1/2}$. Hence the Bellman operator is $\gamma/\sqrt{\beta}$-Lipschitz in this norm. Under Condition C2, projection onto $\mathcal{F}$ is uniquely defined and nonexpansive (Bauschke & Combettes, 2011), so the population recursion is contractive whenever $\beta > \gamma^2$. In particular, $\beta = \gamma$ always satisfies this condition and gives contraction factor $\sqrt{\gamma}$.

Lemma 1 (Occupancy-weighted projected Bellman contraction)**.** *Under Conditions C1–C2, for all* $Q, Q' \in \mathcal{F}$,

$$\|\mathcal{T}_{\mathcal{F},\beta}Q - \mathcal{T}_{\mathcal{F},\beta}Q'\|_{2,d_\beta} \le \bar{\gamma}_\beta\|Q - Q'\|_{2,d_\beta}, \qquad \bar{\gamma}_\beta := \frac{\gamma}{\sqrt{\beta}} < 1.$$

By Banach's fixed-point theorem, the contraction gives a unique fixed point $Q^{\pi}_{\mathcal{F},\beta} \in \mathcal{F}$ and geometric convergence of the population recursion. In particular,

$$\|Q^{(K)}_{\mathcal{F},\beta} - Q^{\pi}_{\mathcal{F},\beta}\|_{2,d_\beta} \le \bar{\gamma}_\beta^K \|Q^{(0)}_{\mathcal{F},\beta} - Q^{\pi}_{\mathcal{F},\beta}\|_{2,d_\beta},$$
$$\|Q^{\pi}_{\mathcal{F},\beta} - Q^{\pi}\|_{2,d_\beta} \le \frac{1}{1-\bar{\gamma}_\beta} \inf_{f\in\mathcal{F}} \|f - Q^{\pi}\|_{2,d_\beta}.$$

Thus, the limiting error is controlled directly by approximation of Q^{π}, rather than approximation of its Bellman images. In particular, if $Q^{\pi} \in \mathcal{F}$, then $Q^{\pi}_{\mathcal{F},\beta} = Q^{\pi}$ without requiring Bellman completeness.

Choosing the occupancy discount. A natural default is $\beta = \gamma$, which gives contraction modulus $\sqrt{\gamma}$. This choice also aligns with the γ-discounted occupancies that arise in policy-value estimation, including doubly robust estimators (Kallus & Uehara, 2020). More generally, increasing β improves the contraction modulus $\bar{\gamma}_\beta = \gamma/\sqrt{\beta}$, approaching γ as $\beta \uparrow 1$, but can increase weight variability as the effective horizon grows. When d_β converges to the target-policy stationary distribution as $\beta \uparrow 1$, the classical stationary-distribution contraction with modulus γ is recovered (Tsitsiklis & Van Roy, 1996; Hallak & Mannor, 2017; Van Roy, 2006). We use the offline state marginal as the initial-state distribution because it yields a direct coverage guarantee through action overlap, although the contraction argument applies to any initial-state distribution.

Convexity. Condition C2 covers closed convex subsets of linear, RKHS, and classical smoothness classes (van der Vaart & Wellner, 1996), as well as convex neural-network function classes (Bach, 2017). Standard finite-width neural-network classes are nonconvex and therefore fall outside our projection argument. However, increasingly rich neural networks can approximate functions in convex smoothness classes covered by the theory (Suzuki, 2019). Extending the analysis directly to nonconvex neural-network classes would require additional control of projection and optimization error.

2.3 Occupancy-Ratio Estimation

The weighted FQE update requires an estimate of $w_\beta = \mathrm{d}d_\beta/\mathrm{d}\nu_b$. Occupancy ratios are commonly used to reweight rewards or construct policy-value estimators (Jiang & Li, 2016; Thomas & Brunskill, 2016; Kallus & Uehara, 2020; 2022; van der Laan et al., 2025a); here, the ratio instead determines the regression norm for each Bellman update.

The ratio is characterized by the occupancy-balance equation (Uehara et al., 2021). For every bounded measurable g, $\mathbb{E}_{\nu_b}[w_\beta(S,A)g(S,A)] = (1-\beta)\mathbb{E}_{\xi_{\pi,b}}[g(S,A)] + \beta\mathbb{E}_{\nu_b}[w_\beta(S,A)g(S',A')]$, where $S' \sim P(\cdot \mid S,A)$, $A' \sim \pi(\cdot \mid S')$, and $\mathbb{E}_{\nu_b}[w_\beta(S,A)] = 1$. Existing estimators exploit these restrictions through stationary distribution correction, minimax or primal–dual objectives, and projected balancing methods (Glynn & Henderson, 1998; Liu et al., 2018; Nachum et al., 2019; Zhang et al., 2020; Uehara et al., 2020; 2021; Wang et al., 2023; Amortila et al., 2024).

Our FQE analysis is agnostic to the ratio estimator and requires only control of its estimation error. Minimax and DICE-style estimators typically introduce an auxiliary critic class, with guarantees depending on its richness and completeness relative to the ratio class (Uehara et al., 2021). For an end-to-end procedure, Algorithm 1 uses fitted occupancy-ratio evaluation (FORE) (van der Laan & Kallus, 2026), whose population update contracts in generalized KL under the offline distribution and therefore avoids completeness conditions. Under coverage, its approximation error depends only on direct approximation of the true occupancy ratio by the ratio class and therefore vanishes under ratio realizability, without requiring an auxiliary critic class or critic-side completeness condition.

For theoretical convenience, we use independent samples $\mathcal{D}^w_{n_w} = \{(S_i, A_i, S'_i)\}_{i=1}^{n_w}$ for ratio estimation and $\mathcal{D}^Q_n$ for FQE. For each i, set $X_i = (S_i, A_i)$, draw $A^0_i \sim \pi(\cdot \mid S_i)$ and $A^+_i \sim \pi(\cdot \mid S'_i)$, and define $X^0_i = (S_i, A^0_i)$ and $X^+_i = (S'_i, A^+_i)$. In practice, the ratio may instead be estimated on the same sample or cross-fitted to reuse data.

Algorithm 1 FORE-Weighted Fitted Q-Evaluation

Input: Samples $\mathcal{D}_{n_w}^w, \mathcal{D}_n^Q$, policy π, classes $\mathcal{G}_w, \mathcal{F}$, iterations K_w, K, discounts γ, β

1: Initialize $\widehat{w}^{(0)} \equiv 1$

2: **for** $\ell = 0, \ldots, K_w - 1$ **do**

3: $\quad \widehat{g}^{(\ell+1)} \in \arg\min_{g \in \mathcal{G}_w} \left[\log \left\{ \frac{1}{n_w} \sum_{i=1}^{n_w} e^{g(X_i)} \right\} - (1-\beta) \frac{1}{n_w} \sum_{i=1}^{n_w} g(X_i^0) - \beta \frac{\sum_{i=1}^{n_w} \widehat{w}^{(\ell)}(X_i) g(X_i^+)}{\sum_{i=1}^{n_w} \widehat{w}^{(\ell)}(X_i)} \right]$

4: $\quad \widehat{w}^{(\ell+1)}(x) \leftarrow \frac{e^{\widehat{g}^{(\ell+1)}(x)}}{n_w^{-1} \sum_{i=1}^{n_w} e^{\widehat{g}^{(\ell+1)}(X_i)}}$

5: **end for**

6: Set $\widehat{w}_\beta \leftarrow \widehat{w}^{(K_w)}$, initialize $\widehat{Q}^{(0)} \in \mathcal{F}$, and apply the occupancy-weighted FQE update for $k = 0, \ldots, K-1$ on $\mathcal{D}_n^Q$

Output: $\widehat{w}_\beta, \widehat{Q}^{(K)}$

3 Theoretical Guarantees

We analyze the exact empirical-risk-minimization version of FORE-weighted FQE in Algorithm 1. Each empirical update approximates the population projected Bellman map $\mathcal{T}_{\mathcal{F},\beta}$, whose fixed point is $Q^\pi_{\mathcal{F},\beta}$. The finite-sample analysis therefore bounds the one-step fitted-regression error $\|\widehat{Q}^{(k)} - \mathcal{T}_{\mathcal{F},\beta}\widehat{Q}^{(k-1)}\|_{2,d_\beta}$, including the additional error from estimating the occupancy weights, and propagates this error through the population contraction.

3.1 One-Step Regression Error under Approximate Weighting

We first establish the main technical lemma, which isolates the effect of occupancy-weight estimation error. We require the following conditions.

C3 Boundedness. There exists $M < \infty$ such that $|R| \le M$ and $\sup_{f \in \mathcal{F}} \|f\|_\infty \le M$.

C4 Independent transitions. The transition tuples $\{(S_i, A_i, R_i, S_i')\}_{i=1}^n$ are independent, with $(S_i, A_i) \sim \nu_b$.

C5 Sample splitting. The weight estimator $\widehat{w}_\beta$ is computed from data independent of $\mathcal{D}_n$.

Conditions C3–C5 streamline the finite-sample analysis. Following Chen & Jiang (2019), Condition C4 assumes independent transitions; extensions to mixing trajectories follow by replacing i.i.d. concentration bounds with their mixing counterparts (Jiang et al., 2018). Condition C5 separates weight estimation from Bellman regression, yielding fast ERM rates without controlling the complexity of the weight estimator (Foster & Syrgkanis, 2023). The analysis extends directly to cross-fitting, allowing all observations to contribute to both stages. Without sample splitting, the rate may additionally depend on the complexity or stability of the weight estimator (van der Laan, 2026).

C6 Coverage and weight bounds. There exists $1 \le \kappa_{\mathrm{cov}} < \infty$ such that $\widehat{w}_\beta \, \nu_b \lesssim d_\beta$ and, ν_b-almost surely,

$$w_\beta(S, A) \le \kappa_{\mathrm{cov}}, \qquad 0 \le \widehat{w}_\beta(S, A) \le \kappa_{\mathrm{cov}}.$$

C7 Restricted norm comparability. Let $\lambda_{\mathcal{F}} > 0$ be any constant such that, for every $f, f' \in \mathcal{F}$,

$$\mathbb{E}_{\nu_b}\left[\widehat{w}_\beta (f - f')^2\right] \ge \lambda_{\mathcal{F}} \mathbb{E}_{\nu_b}\left[w_\beta (f - f')^2\right] = \lambda_{\mathcal{F}} \|f - f'\|_{2,d_\beta}^2.$$

Under Condition C6, we set $\widehat{w}_\beta / w_\beta := 1$ on $\{w_\beta = 0\}$. The measure domination ensures that $\widehat{w}_\beta = 0$ ν_b-almost surely on this set. Condition C6 imposes the standard overlap or concentrability bound on the true ratio w_β (Xie et al., 2022; Uehara et al., 2020). The corresponding bound on $\widehat{w}_\beta$ can be enforced by clipping or constrained estimation. Condition C7 prevents the norm induced by the estimated weights from degenerating along directions in $\mathcal{F} - \mathcal{F}$. A simple sufficient condition is to also bound the estimated weights away from zero. In particular, clipping $\widehat{w}_\beta$ to $[\underline{w}, \kappa_{\mathrm{cov}}]$, with $\underline{w} > 0$, enforces both estimated-weight bounds and gives $\lambda_{\mathcal{F}} = \underline{w}/\kappa_{\mathrm{cov}}$. More generally, Appendix B.2 shows that Condition C7 reduces to weighted Gram-matrix comparability for linear classes and covariance-operator comparability for RKHS classes.

We state the ERM rates in terms of metric-entropy critical radii. For $\varepsilon > 0$, let $N(\varepsilon, \mathcal{F}, L^2(\mathsf{Q}))$ denote the ε-covering number of $\mathcal{F}$ under $L^2(\mathsf{Q})$, and define

$$\mathcal{J}(\delta, \mathcal{F}) := \int_0^\delta \sup_{\mathsf{Q}} \sqrt{\log N(\varepsilon, \mathcal{F}, L^2(\mathsf{Q}))}\, d\varepsilon, \quad \delta_n := \inf\left\{\delta > 0 : \mathcal{J}(\delta, \mathcal{F}) \le \sqrt{n}\,\delta^2\right\} \vee \sqrt{\frac{\log(en)}{n}},$$

where the supremum is over finitely supported distributions Q. The entropy component of δ_n yields localized ERM rates for common function classes: order $\sqrt{s\log(p/s)/n}$ for sparse linear classes, $\sqrt{V\log(n/V)/n}$ for VC-subgraph classes, and $n^{-s/(2s+d)}$ for d-variate Hölder or Sobolev classes of smoothness s (van der Vaart & Wellner, 1996; Nickl & Pötscher, 2007; Mendelson & Neeman, 2010; Wainwright, 2019).

We define the weight estimation and statistical regression error as

$$\varepsilon_{\mathrm{Bell},w}(Q) := \left\| \left(\frac{\widehat{w}_\beta}{w_\beta} - 1 \right) (\mathcal{T}Q - \mathcal{T}_{\mathcal{F},\beta} Q) \right\|_{2,d_\beta},$$

$$\varepsilon_{\mathrm{stat}}(m, \tau) := \sqrt{\kappa_{\mathrm{cov}}} \left(\delta_n + \sqrt{\frac{\log(em/\tau)}{n}} \right).$$

Lemma 2 (One-step weighted regression error)**.** *Assume Conditions C1–C7. There exists $C = C(M) < \infty$ such that, for any $\tau \in (0,1)$, with probability at least $1-\tau$,*

$$\left\| \widehat{Q}^{(k+1)} - \mathcal{T}_{\mathcal{F},\beta} \widehat{Q}^{(k)} \right\|_{2,d_\beta} \le C \left\{ \frac{\varepsilon_{\mathrm{stat}}(1,\tau)}{\sqrt{\lambda_{\mathcal{F}}}} + \frac{\varepsilon_{\mathrm{Bell},w}(\widehat{Q}^{(k)})}{\lambda_{\mathcal{F}}} \right\}.$$

The first term is the usual finite-sample regression error. The second captures the effect of estimating the occupancy weights. Let $\varepsilon_w := \|\widehat{w}_\beta / w_\beta - 1\|_{2,d_\beta}$. Hölder's inequality gives, for every $Q \in \mathcal{F}$,

$$\varepsilon_{\mathrm{Bell},w}(Q) \le \varepsilon_w \|\mathcal{T}Q - \mathcal{T}_{\mathcal{F},\beta} Q\|_\infty, \tag{1}$$

where $\mathcal{T}Q - \mathcal{T}_{\mathcal{F},\beta}Q$ is the Bellman projection residual. Hence the contribution of weight-estimation error is bounded by the product of the relative ratio-estimation error and the size of the Bellman projection residual. Exact occupancy weighting sets $\varepsilon_w = 0$, while Bellman completeness sets the projection residual to zero. More generally, approximate completeness or realizability controls the projection residual and therefore limits the effect of ratio-estimation error, as formalized in the next section. The same bound applies to unweighted FQE by setting $\widehat{w}_\beta \equiv 1$, in which case the relative weight error is the fixed distribution mismatch $w_\beta^{-1} - 1$. Occupancy weighting replaces this fixed mismatch with ratio-estimation error, which can vanish with sample size.

3.2 Main Finite-Sample Guarantee

Our main finite-sample result decomposes the error of occupancy-weighted FQE into finite-iteration, approximation, statistical, and occupancy-weight estimation terms. We define the value-function approximation error and inherent Bellman error by $\varepsilon_{\mathrm{app}} := \inf_{f \in \mathcal{F}} \|f - Q^\pi\|_{2,d_\beta}$ and $\varepsilon_{\mathrm{Bell}} := \sup_{Q \in \mathcal{F}} \|\mathcal{T}Q - \mathcal{T}_{\mathcal{F},\beta}Q\|_{2,d_\beta}$.

C8 Interpolation bound for Bellman projection residuals. There exist $\alpha \in [0,1]$ and $C_\infty < \infty$ such that, for every $Q \in \mathcal{F}$,

$$\|\mathcal{T}Q - \mathcal{T}_{\mathcal{F},\beta}Q\|_\infty \le C_\infty \|\mathcal{T}Q - \mathcal{T}_{\mathcal{F},\beta}Q\|_{2,d_\beta}^\alpha.$$

Condition C8 is a regularity condition on the induced class of Bellman projection residuals $\{\mathcal{T}Q - \mathcal{T}_{\mathcal{F},\beta}Q : Q \in \mathcal{F}\}$. It requires residuals that are small in $L^2(d_\beta)$ to also be small uniformly, with α measuring the strength of this interpolation. At $\alpha = 0$, no additional smoothness is required: the condition holds automatically under Condition C3, with the convention $0^0 = 1$ and $C_\infty := \sup_{Q \in \mathcal{F}} \|\mathcal{T}Q - \mathcal{T}_{\mathcal{F},\beta}Q\|_\infty$. Larger exponents follow when the residual class has additional structure or smoothness. Appendix B.3 gives $\alpha = 1$ for finite-dimensional linear classes, $\alpha = 1 - 1/(2r)$ for RKHS classes with eigenvalues $\lambda_j \lesssim j^{-2r}$, $r > 1/2$ (Mendelson & Neeman, 2010), $\alpha = 2s/(2s+d)$ for Hölder classes of smoothness $s > 0$ (Triebel, 2006), and $\alpha = 1 - d/(2s)$ for Sobolev classes of smoothness $s > d/2$ on d-dimensional domains (Adams & Fournier, 2003).

Theorem 1 (Finite-sample error bound)**.** *Assume Conditions C1–C7 and Condition C8 with $\alpha \in [0,1)$. Let $\gamma_{\beta,\star} := (1 + \bar{\gamma}_\beta)/2$. There exists $C = C(M, \alpha, \gamma, \beta) < \infty$ such that, for every $\tau \in (0,1)$*

and $K \in \mathbb{N}$, with probability at least $1-\tau$,

$$\|\widehat{Q}^{(K)} - Q^\pi\|_{2,d_\beta} \le \gamma_{\beta,\star}^K \|\widehat{Q}^{(0)} - Q^\pi_{\mathcal{F},\beta}\|_{2,d_\beta} + \frac{\varepsilon_{\mathrm{app}}}{1-\bar{\gamma}_\beta} + \frac{C}{1-\gamma_{\beta,\star}} \frac{\varepsilon_{\mathrm{stat}}(K,\tau)}{\sqrt{\lambda_{\mathcal{F}}}} + \frac{C}{1-\gamma_{\beta,\star}} \frac{C_\infty \varepsilon_w}{\lambda_{\mathcal{F}}} \min\left\{ \varepsilon_{\mathrm{Bell}}^\alpha, \left(\frac{C_\infty \varepsilon_w}{\lambda_{\mathcal{F}}}\right)^{\alpha/(1-\alpha)} + \left(\frac{\varepsilon_{\mathrm{app}}}{1-\bar{\gamma}_\beta}\right)^\alpha \right\}.$$

The theorem shows how occupancy-weight estimation affects the FQE error bound. Without additional structure, its contribution is first order in ε_w. When $\alpha > 0$, approximate Bellman completeness and approximate realizability attenuate this error in different ways. Exact Bellman completeness eliminates the interaction entirely, whereas exact realizability leaves only the higher-order term $(C_\infty \varepsilon_w/\lambda_{\mathcal{F}})^{1/(1-\alpha)}$. Consequently, consistency does not require Bellman completeness when Q^π is realizable and the occupancy weights are consistently estimated. At $\alpha = 0$, the bound reduces to the generic linear dependence on $C_\infty \varepsilon_w/\lambda_{\mathcal{F}}$ available under Condition C3.

The key step in the analysis is controlling $\varepsilon_{\mathrm{Bell},w}(\widehat{Q}^{(k)})$ from Lemma 2 along the FQE recursion. Small $\varepsilon_{\mathrm{Bell}}$ controls the Bellman projection residual uniformly over $\mathcal{F}$. Approximate realizability instead provides local control along the recursion. In particular, contraction of $\mathcal{T}$ and nonexpansiveness of projection imply $\|\mathcal{T}\widehat{Q}^{(k)} - \mathcal{T}_{\mathcal{F},\beta}\widehat{Q}^{(k)}\|_{2,d_\beta} \le \bar{\gamma}_\beta \|\widehat{Q}^{(k)} - Q^\pi_{\mathcal{F},\beta}\|_{2,d_\beta} + \varepsilon_{\mathrm{app}}/(1-\bar{\gamma}_\beta)$. Condition C8 converts this $L^2(d_\beta)$ bound into the sup-norm bound required to apply (1). For $\alpha > 0$, the resulting weight error therefore decreases as the iterates approach the projected fixed point, up to the value-function approximation error. Under exact realizability, $Q^\pi_{\mathcal{F},\beta} = Q^\pi$ and this approximation term vanishes; the remaining iterate-dependent term can be absorbed into the contraction, leaving only a higher-order dependence on ratio-estimation error.

For linear function approximation, Corollary S1 in Appendix B.3 treats the endpoint $\alpha = 1$. There, the weight-error term is $(\varepsilon_w/\lambda_{\mathcal{F}}) \min\{\varepsilon_{\mathrm{Bell}}, \varepsilon_{\mathrm{app}}/(1-\bar{\gamma}_\beta)\}$, while weight estimation perturbs the effective contraction factor to $\gamma_w = \bar{\gamma}_\beta(1 + C\varepsilon_w/\lambda_{\mathcal{F}}) < 1$.

End-to-end guarantee with FORE weights. We now specialize the preceding analysis to the FORE estimator in Algorithm 1 (van der Laan & Kallus, 2026). FORE controls occupancy-ratio error in generalized KL divergence, $D^{\mathrm{gen}}_{\nu_b}(f\|g) := \mathbb{E}_{\nu_b}[f\log(f/g) - f + g]$, whereas Theorem 1 depends on relative $L^2(d_\beta)$ error. We therefore assume the target occupancy ratio is bounded above and away from zero.

C9 Target-ratio boundedness. The target occupancy ratio satisfies

$$0 < \underline{\kappa}_\beta := \operatorname*{ess\,inf}_{x\sim\nu_b} w_\beta(x) \le \operatorname*{ess\,sup}_{x\sim\nu_b} w_\beta(x) =: \overline{\kappa}_\beta < \infty.$$

Under the FORE conditions in Appendix G.1–G.2, Theorem S3 gives $\varepsilon_w \le \varepsilon_{\mathrm{FORE}}(K_w, n_w, \tau)$ with probability at least $1-\tau$, where

$$\varepsilon_{\mathrm{FORE}}(K_w, n_w, \tau) \lesssim \frac{\varrho_\beta^{K_w/2}}{\sqrt{\underline{\kappa}_\beta}} \{D^{\mathrm{gen}}_{\nu_b}(1\|w_\beta)\}^{1/2} + \frac{1}{\underline{\kappa}_\beta} \left\{ \frac{\inf_{w\in\mathcal{W}_{\mathrm{F}}} D^{\mathrm{gen}}_{\nu_b}(w\|w_\beta)}{1-\beta} \right\}^{1/2} + \frac{\log(en_w)}{(1-\beta)\sqrt{\underline{\kappa}_\beta}} \left\{ \mathfrak{r}_{n_w,\mathrm{fit}} + \sqrt{\frac{\log(1/\tau)}{n_w}} \right\}, \tag{2}$$

with $\varrho_\beta = (1+\beta)/2$. The three terms are the finite-iteration, ratio-class approximation, and statistical errors. Appendix Corollary S2 formally combines this rate with Theorem 1.

If w_β and Q^π are both realizable, the ratio- and value-class approximation errors vanish. The contribution of FORE to the FQE error is then of order $\{\varepsilon_{\mathrm{FORE}}(K_w, n_w, \tau)/\lambda_{\mathcal{F}}\}^{1/(1-\alpha)}$, which is higher order in the occupancy-weight error whenever $\alpha > 0$. Thus, joint realizability and consistent FORE estimation suffice for consistency without Bellman completeness.

State-value error. The Q-function bound directly yields corresponding state-value bounds. Let d^S_β denote the state marginal of d_β. Since $d_\beta(da \mid s) = \pi(da \mid s)$, Jensen's inequality gives $\|\pi\widehat{Q} - V^\pi\|_{2,d^S_\beta} \le \|\widehat{Q} - Q^\pi\|_{2,d_\beta}$. For the offline state marginal, define $\kappa_{\beta,S} := \inf\{\kappa \ge 1 : \nu_b^S \le$

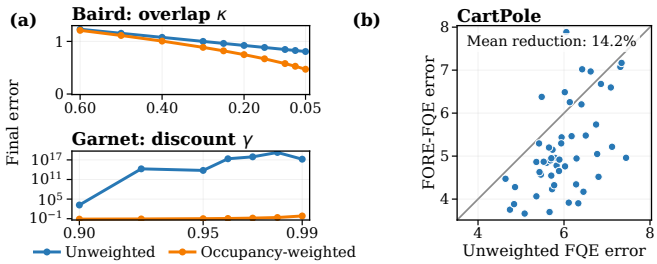

Figure 1: **Occupancy-weighted FQE experiments. (a)** Finite-state experiments. Occupancy weighting reduces final error under severe behavior–target mismatch in Baird's example and prevents divergence in high-discount sparse Garnet MDPs. **(b)** Stochastic continuing CartPole with estimated FORE weights. Each point gives the absolute target-policy value error for one offline dataset; points below the diagonal favor FORE-weighted FQE.

$\kappa d_\beta^S\}$. Then $\|\pi\widehat{Q} - V^\pi\|_{2,\nu_b^S} \le \sqrt{\kappa_{\beta,S}}\|\widehat{Q} - Q^\pi\|_{2,d_\beta}$. The initial occupancy component gives $\kappa_{\beta,S} \le (1-\beta)^{-1}$, so the worst-case transfer factor is $(1-\beta)^{-1/2}$, with a sharper bound when the state-marginal overlap is better. Cauchy–Schwarz gives the same control for the corresponding average value errors.

4 Experimental Investigation

We use two experiments to examine the role of occupancy weighting. The first uses finite-state MDPs, where the projection operator can be computed exactly, to study how mismatch between the offline distribution and the target-policy occupancy affects FQE. The second uses a stochastic continuing-control problem to study FQE when the occupancy ratio must be estimated from data. Appendices C.1 and C.2 give the full experimental details.

Finite-state distribution mismatch. We consider an adaptation of Baird's seven-state example (Baird, 1995) and sparse Garnet MDPs (Bhatnagar et al., 2009). These settings allow us to vary the mismatch between the offline distribution and the target-policy occupancy while computing the relevant projected Bellman operators exactly. Exact occupancy weighting reduces final error under severe mismatch in Baird's example and prevents high-discount divergence in the sparse Garnet MDPs (Figure 1). Smaller κ corresponds to greater behavior–target mismatch, while larger γ weakens Bellman contraction.

Stochastic continuing CartPole with estimated weights. We next consider a stochastic continuing CartPole MDP in which failure resets the system to a random state rather than terminating the process. The behavior and target policies are full-support logistic policies. Each offline dataset contains $25{,}000$ transitions sampled from the stationary behavior distribution after burn-in. Unweighted FQE and FORE-weighted FQE use the same data, function class, and initialization and differ only in their regression weights. We set the occupancy and Bellman discounts to $\beta = \gamma = 0.98$ and use the offline state marginal as the initial-state distribution. We evaluate both estimators on 50 independent offline datasets. For each dataset, absolute target-policy value error is computed relative to an independently simulated Monte Carlo reference value. We report the paired mean reduction in absolute error and its 95% bootstrap confidence interval, along with the effective sample size of the estimated FORE weights.

CartPole results. FORE-weighted FQE reduces mean absolute value error from 6.000 to 5.148, a mean relative reduction of 14.2%. The paired mean error reduction has 95% bootstrap confidence interval $[0.611, 1.087]$, and FORE-weighted FQE has lower absolute error in 44 of the 50 datasets. The median effective sample size (ESS) of the FORE weights, defined as $(\sum_{i=1}^n w_i)^2 / \sum_{i=1}^n w_i^2$ and reported relative to n, is 0.759.

5 Discussion

Occupancy weighting changes the FQE regression geometry so that the projected Bellman operator is contractive, while preserving its supervised-learning structure. Because ratio-estimation error enters through the Bellman projection residual, joint realizability of the value function and occupancy ratio suffices for consistency without completeness assumptions on either class. Weak coverage can make the estimated weights variable; tempering $\widetilde{w}_\alpha \propto \widehat{w}_\beta^\alpha$, $\alpha \in [0,1]$, trades exact occupancy alignment for greater stability whenever the resulting weighted Bellman operator remains contractive (Sinha et al., 2022).

AI USE STATEMENT

Generative AI tools were used to assist with manuscript writing and editing, experiment implementation, and checking mathematical derivations. All research ideas, methodological contributions, experimental design choices, mathematical claims, and interpretation of results were developed by the authors. The authors reviewed all AI-assisted work and take full responsibility for the content of the manuscript.

REFERENCES

Robert A Adams and John JF Fournier. *Sobolev spaces*, volume 140. Elsevier, 2003.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *Advances in neural information processing systems*, 34:29304–29320, 2021.

Philip Amortila, Nan Jiang, and Tengyang Xie. A variant of the wang-foster-kakade lower bound for the discounted setting. *arXiv preprint arXiv:2011.01075*, 2020.

Philip Amortila, Dylan J Foster, Nan Jiang, Ayush Sekhari, and Tengyang Xie. Harnessing density ratios for online reinforcement learning. *arXiv preprint arXiv:2401.09681*, 2024.

András Antos, Csaba Szepesvári, and Rémi Munos. Fitted q-iteration in continuous action-space mdps. *Advances in neural information processing systems*, 20, 2007.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

Leemon C. Baird, III. Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 30–37, 1995.

Heinz H Bauschke and Patrick L Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer, New York, NY, 2011. doi: 10.1007/978-1-4419-9467-7.

Richard Bellman. On the theory of dynamic programming. *Proceedings of the national Academy of Sciences*, 38(8):716–719, 1952.

Richard Bellman. Dynamic programming. *science*, 153(3731):34–37, 1966.

Dimitri P Bertsekas. Approximate policy iteration: A survey and some new methods. *Journal of Control Theory and Applications*, 9(3):310–335, 2011.

Shalabh Bhatnagar, Richard S Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor–critic algorithms. *Automatica*, 45(11):2471–2482, 2009.

Aurelien Bibaut, Maya Petersen, Nikos Vlassis, Maria Dimakopoulou, and Mark van der Laan. Sequential causal inference in a single world of connected units. *arXiv preprint arXiv:2101.07380*, 2021.

Steven J. Bradtke and Andrew G. Barto. Linear least-squares algorithms for temporal difference learning. *Machine Learning*, 22(1–3):33–57, 1996. doi: 10.1007/BF00114723.

Jonathan Chang, Kaiwen Wang, Nathan Kallus, and Wen Sun. Learning bellman complete representations for offline policy evaluation. In *International Conference on Machine Learning*, pp. 2938–2971. PMLR, 2022.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International conference on machine learning*, pp. 1042–1051. PMLR, 2019.

Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. Sbeed: Convergent reinforcement learning with nonlinear function approximation. In *International conference on machine learning*, pp. 1125–1134. PMLR, 2018.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6, 2005.

Amir-massoud Farahmand, Csaba Szepesvári, and Rémi Munos. Error propagation for approximate policy and value iteration. *Advances in neural information processing systems*, 23, 2010.

Yihao Feng, Lihong Li, and Qiang Liu. A kernel loss for solving the bellman equation. *Advances in Neural Information Processing Systems*, 32, 2019.

Yihao Feng, Tongzheng Ren, Ziyang Tang, and Qiang Liu. Accountable off-policy evaluation with kernel bellman statistics. In *International Conference on Machine Learning*, pp. 3102–3111. PMLR, 2020.

Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51(3): 879–908, 2023.

Dylan J Foster, Akshay Krishnamurthy, David Simchi-Levi, and Yunzong Xu. Offline reinforcement learning: Fundamental barriers for value function approximation. *arXiv preprint arXiv:2111.10919*, 2021.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pp. 2052–2062. PMLR, 2019.

Carles Gelada and Marc G. Bellemare. Off-policy deep reinforcement learning by bootstrapping the covariate shift. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 3647–3655. AAAI Press, 2019. doi: 10.1609/AAAI.V33I01.33013647. URL https://doi.org/10.1609/aaai.v33i01.33013647.

Peter W Glynn and Shane G Henderson. Estimation of stationary densities for markov chains. In *1998 Winter Simulation Conference. Proceedings (Cat. No. 98CH36274)*, volume 1, pp. 647–652. IEEE, 1998.

Geoffrey J Gordon. Stable function approximation in dynamic programming. In *Machine learning proceedings 1995*, pp. 261–268. Elsevier, 1995.

Assaf Hallak and Shie Mannor. Consistent on-line off-policy evaluation. In *International Conference on Machine Learning*, pp. 1372–1383. PMLR, 2017.

Bai Jiang, Qiang Sun, and Jianqing Fan. Bernstein's inequalities for general markov chains. *arXiv preprint arXiv:1805.10721*, 2018. URL https://arxiv.org/abs/1805.10721.

Nan Jiang and Jiawei Huang. Minimax value interval for off-policy evaluation and policy optimization. *Advances in Neural Information Processing Systems*, 33:2747–2758, 2020.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International conference on machine learning*, pp. 652–661. PMLR, 2016.

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

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. *Mathematics of Operations Research*, 48(3):1496–1521, 2023.

Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63, 2020.

Nathan Kallus and Masatoshi Uehara. Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *Operations Research*, 70(6):3282–3302, 2022.

J. Zico Kolter. The fixed points of off-policy TD. In *Advances in Neural Information Processing Systems*, volume 24, pp. 2169–2177, 2011.

Michail G Lagoudakis and Ronald Parr. Least-squares policy iteration. *Journal of machine learning research*, 4(Dec):1107–1149, 2003.

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

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in neural information processing systems*, 31, 2018.

Xinyu Liu, Zixuan Xie, and Shangtong Zhang. Linear Q-learning does not diverge in L^2: Convergence rates to a bounded set. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 39871–39897. PMLR, 2025.

A Rupam Mahmood, Huizhen Yu, Martha White, and Richard S Sutton. Emphatic temporal-difference learning. *arXiv preprint arXiv:1507.01569*, 2015.

Prashant Mehta and Sean Meyn. Optimistic training and convergence of Q-learning: Extended version. *arXiv preprint arXiv:2602.06146*, 2026. doi: 10.48550/arXiv.2602.06146.

Francisco S Melo and M Isabel Ribeiro. Convergence of q-learning with linear function approximation. In *2007 European control conference (ECC)*, pp. 2671–2678. IEEE, 2007.

Shahar Mendelson and Joseph Neeman. Regularization in kernel learning. *The Annals of Statistics*, 38(1):526–565, 2010. doi: 10.1214/09-AOS728.

Sean Meyn. The projected bellman equation in reinforcement learning. *IEEE Transactions on Automatic Control*, 69(12):8323–8337, 2024.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Rémi Munos. Error bounds for approximate value iteration. In *Proceedings of the Twentieth National Conference on Artificial Intelligence*, pp. 1006–1011, 2005.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(27):815–857, 2008.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Advances in neural information processing systems*, 32, 2019.

Angelia Nedić and Dimitri P. Bertsekas. Least squares policy evaluation algorithms with linear function approximation. *Discrete Event Dynamic Systems*, 13(1–2):79–110, 2003. doi: 10.1023/A:1022192903948.

Richard Nickl and Benedikt M Pötscher. Bracketing metric entropy rates and empirical central limit theorems for function classes of besov- and sobolev-type. *Journal of Theoretical Probability*, 20 (2):177–199, 2007.

Andrew Patterson, Adam White, and Martha White. A generalized projected bellman error for off-policy value estimation in reinforcement learning. *Journal of Machine Learning Research*, 23 (145):1–61, 2022.

Bruno Scherrer. Approximate policy iteration schemes: A comparison. In *International Conference on Machine Learning*, pp. 1314–1322. PMLR, 2014.

Samarth Sinha, Jiaming Song, Animesh Garg, and Stefano Ermon. Experience replay with likelihood-free importance weights. In *Proceedings of the 4th Annual Learning for Dynamics and Control Conference*, volume 168 of *Proceedings of Machine Learning Research*, pp. 110–123. PMLR, 2022. URL https://proceedings.mlr.press/v168/sinha22a.html.

Richard S Sutton, A Rupam Mahmood, and Martha White. An emphatic approach to the problem of off-policy temporal-difference learning. *Journal of Machine Learning Research*, 17(73):1–29, 2016.

Taiji Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: Optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019.

Shengpu Tang and Jenna Wiens. Model selection for offline reinforcement learning: Practical considerations for healthcare settings. In Ken Jung, Serena Yeung, Mark Sendak, Michael Sjoding, and Rajesh Ranganath (eds.), *Proceedings of the 6th Machine Learning for Healthcare Conference*, volume 149 of *Proceedings of Machine Learning Research*, pp. 2–35. PMLR, 6–7 August 2021. URL https://proceedings.mlr.press/v149/tang21a.html.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International conference on machine learning*, pp. 2139–2148. PMLR, 2016.

Hans Triebel. *Theory of function spaces III*. Springer, 2006.

John Tsitsiklis and Benjamin Van Roy. Analysis of temporal-difference learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 9, pp. 1075–1081, 1996.

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

Masatoshi Uehara, Masaaki Imaizumi, Nan Jiang, Nathan Kallus, Wen Sun, and Tengyang Xie. Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency. *arXiv preprint arXiv:2102.02981*, 2021.

Masatoshi Uehara, Nathan Kallus, Jason D Lee, and Wen Sun. Offline minimax soft-q-learning under realizability and partial coverage. *Advances in Neural Information Processing Systems*, 36: 12797–12809, 2023.

Sara van de Geer. On the uniform convergence of empirical norms and inner products, with application to causal inference. *Electronic Journal of Statistics*, 8:543–574, 2014.

Lars van der Laan. A researcher's guide to empirical risk minimization. *arXiv preprint arXiv:2602.21501*, 2026.

Lars van der Laan and Nathan Kallus. Fitted occupancy-ratio evaluation without bellman completeness. *arXiv preprint arXiv:2607.05375*, 2026.

Lars van der Laan, David Hubbard, Allen Tran, Nathan Kallus, and Aurélien Bibaut. Semiparametric double reinforcement learning with applications to long-term causal inference. *arXiv preprint arXiv:2501.06926*, 2025a.

Lars van der Laan, Nathan Kallus, and Aurélien Bibaut. Nonparametric instrumental variable inference with many weak instruments. *arXiv preprint arXiv:2505.07729*, 2025b.

Aad van der Vaart and Jon A Wellner. A local maximal inequality under uniform entropy. *Electronic Journal of Statistics*, 5:192–203, 2011. doi: 10.1214/11-EJS605.

Aad W van der Vaart and Jon A Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York, 1996. doi: 10.1007/978-1-4757-2545-2.

Benjamin Van Roy. Performance loss bounds for approximate value iteration with state aggregation. *Mathematics of Operations Research*, 31(2):234–244, 2006.

Cameron Voloshin, Hoang M Le, Nan Jiang, and Yisong Yue. Empirical study of off-policy policy evaluation for reinforcement learning. *arXiv preprint arXiv:1911.06854*, 2019.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Jiayi Wang, Zhengling Qi, and Raymond KW Wong. Projected state-action balancing weights for offline reinforcement learning. *The Annals of Statistics*, 51(4):1639–1665, 2023.

Ruosong Wang, Dean Foster, and Sham M Kakade. What are the statistical limits of offline rl with linear function approximation? In *International Conference on Learning Representations*, 2021a. URL https://arxiv.org/abs/2010.11895.

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

Tengyang Xie and Nan Jiang. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence*, pp. 550–559. PMLR, 2020.

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

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

Tengyang Xie, Dylan J Foster, Yu Bai, Nan Jiang, and Sham M Kakade. The role of coverage in online reinforcement learning. *arXiv preprint arXiv:2210.04157*, 2022.

Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pp. 10746–10756. PMLR, 2020.

Huizhen Yu. On convergence of emphatic temporal-difference learning. In *Conference on learning theory*, pp. 1724–1751. PMLR, 2015.

Huizhen Yu and Dimitri P. Bertsekas. Error bounds for approximations from projected linear equations. *Mathematics of Operations Research*, 35(2):306–329, 2010. doi: 10.1287/moor.1100.0441.

Huizhen Yu and Dimitri P Bertsekas. Weighted bellman equations and their applications in approximate dynamic programming. *Lab. for Information and Decision Systems Report LIDS-P-2876, MIT*, 2012.

Huizhen Yu, A Rupam Mahmood, and Richard S Sutton. On generalized bellman equations and temporal-difference learning. *Journal of Machine Learning Research*, 19(48):1–49, 2018.

Andrea Zanette. When is realizability sufficient for off-policy reinforcement learning? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 40637–40668. PMLR, 2023.

Andrea Zanette and Martin J Wainwright. Bellman residual orthogonalization for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 35:3137–3151, 2022.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Proceedings of the Thirty-Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 2730–2775. PMLR, 2022.

Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. Gendice: Generalized offline estimation of stationary values. *arXiv preprint arXiv:2002.09072*, 2020.

Appendix roadmap. Appendix A provides additional related work. Appendix B collects the occupancy-weighted algorithm, interpolation bounds for weight–residual interactions, approximate-weight analysis, and the implicit stationary-sampling extension. Appendix C gives details for the finite-state and CartPole experiments, and Appendix D gives the contraction results and proofs for the discounted-occupancy analysis. Appendices E and F contain the technical lemmas and proofs of the main results. Finally, Appendix G gives the weight-error bound for FORE and the resulting end-to-end FQE guarantee.

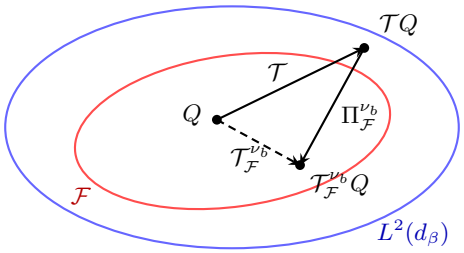

Figure S1: Norm mismatch. The Bellman update $\mathcal{T}Q$ may leave $\mathcal{F}$, so FQE projects it back using the offline-sampling norm $L^2(\nu_b)$. This projection need not preserve the contraction property of policy evaluation in the target-occupancy norm $L^2(d_\beta)$.

A Additional related work

Projected fixed points under linear function approximation.

For fixed-policy evaluation with a linear value-function class, TD and LSTD seek solutions to a projected Bellman equation, whereas exact population least-squares FQE iterates the associated projected Bellman map (Bradtke & Barto, 1996; Tsitsiklis & Van Roy, 1996; Nedić & Bertsekas, 2003). Under an arbitrary off-policy projection norm, this map need not be contractive, and its fixed point can be arbitrarily inaccurate even when Q^π is well approximated by the class (Yu & Bertsekas, 2010; Kolter, 2011). Linearity alone therefore does not resolve the geometric mismatch induced by off-policy projection. Existing convergence guarantees require additional structure, such as Bellman completeness, linear-MDP assumptions, or feature and sampling conditions that restore stability on the chosen representation (Melo & Ribeiro, 2007; Wang et al., 2021a; Chang et al., 2022; Jin et al., 2023).

Related difficulties arise in control, where the policy depends on the current Q-function and the projected Bellman optimality map is nonlinear. Under a specialized adaptive training policy, Meyn (2024) established boundedness of linear Q-learning iterates and existence of a projected Bellman solution, while leaving convergence and uniqueness unresolved. Liu et al. (2025) subsequently obtained an L^2 convergence rate to a bounded set rather than to a particular projected solution. More recently, Mehta & Meyn (2026) constructed realizable examples with multiple projected Bellman solutions, showing that additional structure is needed for uniqueness and convergence. Together, these results demonstrate that linear function approximation alone does not ensure stable projected Bellman iteration. Our work instead concerns fixed-policy evaluation and uses occupancy weighting to make the fitted-regression updates contractive for flexible, possibly nonlinear function classes.

B Additional Details

B.1 Algorithm

Algorithm S2 Occupancy-Weighted Fitted Q-Evaluation

Input: Offline data $\mathcal{D}_n$, target policy π, function class $\mathcal{F}$, iterations K, discounts γ, β

`# Estimate occupancy ratio`

$\widehat{w}_\beta \leftarrow$ estimator of w_β

Initialize $\widehat{Q}^{(0)} \in \mathcal{F}$

for $k = 0, \dots, K-1$ **do**

`# Construct Bellman targets`

$Y_i^{(k)} \leftarrow R_i + \gamma(\pi\widehat{Q}^{(k)})(S_i')$

`# Occupancy-weighted regression`

$\widehat{Q}^{(k+1)} \leftarrow \arg\min_{f\in\mathcal{F}} \sum_{i=1}^n \widehat{w}_\beta(S_i, A_i)\{Y_i^{(k)} - f(S_i, A_i)\}^2$

end for

Output: $\widehat{Q}^{(K)}$

B.2 Verifying Condition C7 for Linear and RKHS Classes

Remark 1 (Linear classes). For linear classes, Condition C7 follows from Gram-matrix conditions weaker than a pointwise lower bound on $\widehat{w}_\beta / w_\beta$. Suppose that

$$\mathcal{F} = \{f_\theta = \phi^\top \theta : \theta \in \Theta\},$$

and define the estimated-weight and target-occupancy Gram matrices

$$\widehat{\Sigma}_\phi := \mathbb{E}_{\nu_b}\left[\widehat{w}_\beta \phi\phi^\top\right], \qquad \Sigma_{\beta,\phi} := \mathbb{E}_{\nu_b}\left[w_\beta \phi\phi^\top\right] = \mathbb{E}_{d_\beta}\left[\phi\phi^\top\right].$$

Then Condition C7 holds whenever the restricted generalized Rayleigh quotient below is positive, and one may take any $\lambda_{\mathcal{F}}$ satisfying

$$0 < \lambda_{\mathcal{F}} \le \inf_{\substack{v \in \Theta - \Theta \\ v^\top \Sigma_{\beta,\phi} v > 0}} \frac{v^\top \widehat{\Sigma}_\phi v}{v^\top \Sigma_{\beta,\phi} v}.$$

The restricted quotient is bounded below by the unrestricted quotient:

$$\inf_{\substack{v \in \Theta - \Theta \\ v^\top \Sigma_{\beta,\phi} v > 0}} \frac{v^\top \widehat{\Sigma}_\phi v}{v^\top \Sigma_{\beta,\phi} v} \ge \inf_{\substack{v \in \mathbb{R}^p \\ v^\top \Sigma_{\beta,\phi} v > 0}} \frac{v^\top \widehat{\Sigma}_\phi v}{v^\top \Sigma_{\beta,\phi} v}.$$

Thus, Condition C7 follows from weighted Gram-matrix comparability. In particular, if

$$\lambda_{\min}(\widehat{\Sigma}_\phi) \ge \underline{\lambda} > 0, \qquad \lambda_{\max}(\Sigma_{\beta,\phi}) \le \overline{\lambda} < \infty,$$

then one may take $\lambda_{\mathcal{F}} = \underline{\lambda}/\overline{\lambda}$. If $\Theta - \Theta$ contains a nonzero multiple of every direction in $\mathbb{R}^p$, as when Θ has nonempty interior, the preceding inequality over $\mathbb{R}^p$ is an equality.

Remark 2 (RKHS classes). Let $\mathcal{H}$ be an RKHS with feature map $K_x := K(x, \cdot)$, and suppose that $\mathcal{F} \subseteq \mathcal{H}$. Define the weighted covariance operators

$$\widehat{\mathcal{C}} := \mathbb{E}_{\nu_b}[\widehat{w}_\beta(X)\, K_X \otimes K_X], \qquad \mathcal{C}_\beta := \mathbb{E}_{d_\beta}[K_X \otimes K_X].$$

For every $h \in \mathcal{H}$,

$$\mathbb{E}_{\nu_b}\left[\widehat{w}_\beta h^2\right] = \langle h, \widehat{\mathcal{C}} h\rangle_{\mathcal{H}}, \qquad \mathbb{E}_{d_\beta}\left[h^2\right] = \langle h, \mathcal{C}_\beta h\rangle_{\mathcal{H}}.$$

Therefore, Condition C7 holds with $\lambda_{\mathcal{F}} = \underline{c}$ whenever

$$\widehat{\mathcal{C}} \succeq \underline{c}\, \mathcal{C}_\beta$$

on $\mathrm{span}(\mathcal{F} - \mathcal{F})$. Thus, for RKHS classes, Condition C7 reduces to covariance-operator comparability along the directions represented by the fitted class.

B.3 Interpolation Bounds for Weight–Residual Interactions

Define the inherent Bellman error in the target-occupancy norm and the relative density-ratio error by

$$\varepsilon_{\mathrm{Bell}} := \sup_{Q \in \mathcal{F}} \|\mathcal{T}Q - \mathcal{T}_{\mathcal{F},\beta} Q\|_{2,d_\beta}, \qquad \varepsilon_w := \left\| \frac{\widehat{w}_\beta}{w_\beta} - 1 \right\|_{2,d_\beta}.$$

Equivalently,

$$\varepsilon_w^2 = \mathbb{E}_{\nu_b}\left[\frac{(\widehat{w}_\beta - w_\beta)^2}{w_\beta}\right].$$

Because $\mathcal{T}_{\mathcal{F},\beta} Q$ is the $L^2(d_\beta)$ projection of $\mathcal{T}Q$ onto $\mathcal{F}$, $\varepsilon_{\mathrm{Bell}}$ measures the inherent Bellman error of $\mathcal{F}$ in the target-occupancy norm. Condition C8 holds with $\alpha > 0$ when the residual class

$$\{\mathcal{T}Q - \mathcal{T}_{\mathcal{F},\beta} Q : Q \in \mathcal{F}\}$$

has suitable finite-dimensional or smoothness structure.

Linear function approximation. Consider a linear-feature class $\mathcal{F} = \{Q_\theta = \phi^\top \theta : \theta \in \Theta\}$, where $\phi = (\phi_1, \ldots, \phi_p)^\top$. Define

$$\mathcal{G}_{\mathrm{res}} := \mathrm{span}\{r_0, \phi_1, \ldots, \phi_p, P_\pi \phi_1, \ldots, P_\pi \phi_p\}.$$

Let $\psi = (\psi_1, \ldots, \psi_q)^\top$ be an $L^2(d_\beta)$-basis of $\mathcal{G}_{\mathrm{res}}$. Then $\Sigma_\psi := \mathbb{E}_{d_\beta}[\psi\psi^\top]$ is positive definite. Set

$$\kappa_{\mathrm{res}} := \operatorname*{ess\,sup}_{x \sim d_\beta} \psi(x)^\top \Sigma_\psi^{-1} \psi(x).$$

Proposition 1 (Linear residual interpolation). *If $\kappa_{\mathrm{res}} < \infty$, then Condition C8 holds with $\alpha = 1$ and $C_\infty = \sqrt{\kappa_{\mathrm{res}}}$.*

The condition $\kappa_{\mathrm{res}} < \infty$ follows from familiar feature boundedness and Gram-matrix conditioning: if $\sup_x \|\psi(x)\|_2 \le B_\psi$ and $\lambda_{\min}(\Sigma_\psi) \ge c_\psi > 0$, then $C_\infty \le B_\psi/\sqrt{c_\psi}$. If each coordinate is bounded by B and c_ψ is dimension-independent, then $C_\infty \le B\sqrt{q/c_\psi} = O(\sqrt{p})$, since $q \le 2p+1$. For an RKHS residual class that is uniformly bounded in RKHS norm, suppose the kernel and the eigenfunctions of its integral operator on $L^2(d_\beta)$ are uniformly bounded. If the eigenvalues satisfy $\lambda_j \lesssim j^{-2r}$ for $r > 1/2$, one may take $\alpha = 1 - 1/(2r)$ (Mendelson & Neeman, 2010, Lemma 5.1). Related interpolation bounds hold for signed convex hulls of suitable bases (van de Geer, 2014, Lemma 2) and for d-variate Hölder or Sobolev residual classes of order $s > d/2$ on regular bounded domains with a design density bounded below, with $\alpha = 1 - d/(2s)$ (Bibaut et al., 2021, Lemma 4).

Under Condition C8 with $\alpha > 0$, the weight-error term obeys

$$\sup_{Q \in \mathcal{F}} \varepsilon_{\mathrm{Bell},w}(Q) \le C_\infty \varepsilon_w \varepsilon_{\mathrm{Bell}}^{\alpha}.$$

Thus, on the completeness branch, density-ratio error and inherent Bellman error enter multiplicatively. The realizability branch of Theorem 1 instead controls the residual locally near $Q^\pi_{\mathcal{F},\beta}$, producing its higher-order and $\varepsilon_{\mathrm{app}}$-dependent terms. In both branches, the small-error guarantee also requires control of ε_w and restricted norm comparability.

Corollary S1 (Linear-class refinement)**.** *Assume Conditions C1–C7 and the conditions of Proposition 1. Define $\varepsilon_{\mathrm{app}}$ and $\varepsilon_{\mathrm{Bell}}$ as in Theorem 1. There exists $C = C(M, \kappa_{\mathrm{res}}) < \infty$ such that, if*

$$\gamma_w := \bar{\gamma}_\beta \left(1 + C\frac{\varepsilon_w}{\lambda_{\mathcal{F}}}\right) < 1,$$

then, for every $\tau \in (0,1)$ and $K \in \mathbb{N}$, with probability at least $1-\tau$,

$$\|\widehat{Q}^{(K)} - Q^\pi\|_{2,d_\beta} \le \gamma_w^K \|\widehat{Q}^{(0)} - Q^\pi_{\mathcal{F},\beta}\|_{2,d_\beta} + \frac{\varepsilon_{\mathrm{app}}}{1-\bar{\gamma}_\beta} + \frac{C}{1-\gamma_w}\left[\frac{\varepsilon_{\mathrm{stat}}(K,\tau)}{\sqrt{\lambda_{\mathcal{F}}}} + \frac{\varepsilon_w}{\lambda_{\mathcal{F}}} \min\left\{\varepsilon_{\mathrm{Bell}}, \frac{\varepsilon_{\mathrm{app}}}{1-\bar{\gamma}_\beta}\right\}\right].$$

Thus, exact Bellman completeness or exact value-function realizability removes the additive weight-error term. Under exact realizability, the direct $\varepsilon_w/\lambda_{\mathcal{F}}$ dependence is absorbed into γ_w, while $\lambda_{\mathcal{F}}$ continues to control norm comparability.

B.4 Approximate Weights and Stationary Limits

Contraction under approximate occupancy weights. The analysis extends beyond exact recovery of w_β. Suppose $\overline{w}_\beta$ is a limiting weight and the corresponding projected Bellman operator is contractive in $L^2(\overline{w}_\beta \nu_b)$ with modulus $\overline{\gamma} < 1$. Then the same inexact Picard argument applies with $\overline{\gamma}$ in place of $\bar{\gamma}_\beta$ and with the projected fixed point defined under the $\overline{w}_\beta\nu_b$-weighted norm. Thus, misspecified weights can still improve stability if they move the regression geometry toward one in which the Bellman operator is contractive.

A simple sufficient condition is multiplicative closeness to the occupancy ratio. Suppose, for some $\varepsilon \in [0,1)$,

$$(1-\varepsilon) w_\beta(s,a) \le \overline{w}_\beta(s,a) \le (1+\varepsilon) w_\beta(s,a).$$

Then $L^2(\overline{w}_\beta\nu_b)$ and $L^2(w_\beta\nu_b) = L^2(d_\beta)$ are equivalent:

$$\|f\|^2_{2,d_\beta} \le \frac{1}{1-\varepsilon}\|f\|^2_{2,\overline{w}_\beta\nu_b}, \qquad \|f\|^2_{2,\overline{w}_\beta\nu_b} \le (1+\varepsilon)\|f\|^2_{2,d_\beta}.$$

Because $\mathcal{T}$ is a $\bar{\gamma}_\beta$-contraction in $L^2(d_\beta)$,

$$\|\mathcal{T}Q - \mathcal{T}Q'\|_{2,\overline{w}_\beta\nu_b} \le \bar{\gamma}_\beta \sqrt{\frac{1+\varepsilon}{1-\varepsilon}}\, \|Q - Q'\|_{2,\overline{w}_\beta\nu_b}.$$

Hence, whenever

$$\bar{\gamma}_\beta \sqrt{\frac{1+\varepsilon}{1-\varepsilon}} < 1,$$

the Bellman operator, and therefore its projection onto a closed convex fitted class, remains contractive in $L^2(\overline{w}_\beta\nu_b)$. Small multiplicative weight errors thus preserve contraction, with an inflated contraction factor.

On-policy and ergodic limits. When $\nu_b = \mu_\pi$, stationarity gives $\xi_{\pi,b} = d_\beta = \mu_\pi$ and hence $w_\beta \equiv 1$ for every $\beta \in (0,1)$. Occupancy-weighted FQE then reduces to ordinary FQE, and all weight-related errors vanish. More generally, under ergodic on-policy trajectory sampling, empirical state-action frequencies converge to μ_π, so the population regression norm approaches the target-policy stationary norm. Extending this observation to finite-sample guarantees requires additional mixing and empirical-process control.

B.5 Implicit Stationary Weighting by Sampling

Stationary weighting can also be achieved implicitly when the transition kernel P, and hence μ_π, is known. If we can sample $(\widetilde{S}_i, \widetilde{A}_i, \widetilde{S}'_i) \sim \mu_\pi \otimes P$ directly, then the regression design is already matched to the target stationary distribution. Because these samples do not contain rewards, let $\widehat{r}$ estimate the reward function from the offline data, and run unweighted FQE on

$$\{(\widetilde{S}_i, \widetilde{A}_i, \widehat{r}(\widetilde{S}_i, \widetilde{A}_i), \widetilde{S}'_i)\}_{i=1}^n.$$

This is ordinary FQE using samples from the target-policy stationary distribution, with $w_\pi \equiv 1$ and a perturbed reward.

Define

$$\widehat{\mathcal{T}}Q := \widehat{r} + \gamma P_\pi Q, \qquad \widehat{\mathcal{T}}_{\mathcal{F}} := \Pi_{\mathcal{F}} \widehat{\mathcal{T}},$$

and let $\widehat{Q}^\pi_{\mathcal{F}}$ be the fixed point of $\widehat{\mathcal{T}}_{\mathcal{F}}$. For the K-step iterate $\widehat{Q}^{(K)}$,

$$\|\widehat{Q}^{(K)} - Q^\pi_{\mathcal{F}}\|_{2,\mu_\pi} \le \|\widehat{Q}^{(K)} - \widehat{Q}^\pi_{\mathcal{F}}\|_{2,\mu_\pi} + \|\widehat{Q}^\pi_{\mathcal{F}} - Q^\pi_{\mathcal{F}}\|_{2,\mu_\pi}.$$

The first term is the usual finite-iteration and statistical error under stationary sampling. The second term is the reward-estimation perturbation.

Theorem S1 (Reward perturbation)**.** *Suppose μ_π is stationary for P_π, $r_0, \widehat{r} \in L^2(\mu_\pi)$, and $\mathcal{F}$ is a nonempty closed convex subset of $L^2(\mu_\pi)$. Then*

$$\|\widehat{Q}^\pi_{\mathcal{F}} - Q^\pi_{\mathcal{F}}\|_{2,\mu_\pi} \le \frac{1}{1-\gamma} \|\widehat{r} - r_0\|_{2,\mu_\pi}.$$

If, in addition, $\mathcal{F}$ is a linear subspace of $L^2(\mu_\pi)$, then

$$\|\widehat{Q}^\pi_{\mathcal{F}} - Q^\pi_{\mathcal{F}}\|_{2,\mu_\pi} \le \frac{1}{1-\gamma} \|\Pi_{\mathcal{F}}(\widehat{r} - r_0)\|_{2,\mu_\pi}.$$

C Experiments

C.1 Finite-State Experiments

Baird hub-and-spokes. We use the seven-state hub-and-spokes example of Baird (1995). The solid action transitions deterministically to the hub state, while the dashed action transitions uniformly to one of the six spoke states. The target policy always selects the solid action, and the behavior policy selects it with probability κ, so smaller κ corresponds to greater behavior–target mismatch. We use the standard seven-dimensional linear feature map, zero rewards, ridge-regularized FQE, and exact target-policy occupancy ratios, which equal κ^{-2} on hub–solid pairs and zero otherwise. Since $Q^\pi \equiv 0$, any nonzero error is due to the behavior of the projected FQE iteration rather than reward approximation. We vary κ from 0.05 to 0.60, using a finer grid under severe mismatch, and average over 100 independent datasets.

Sparse Garnet MDPs. We also consider Garnet MDPs with 100 states, four actions, and branching factor five. For each state–action pair, transition probabilities over five successor states are drawn from a Dirichlet distribution, and rewards are Gaussian. The target policy is randomly generated, and the behavior policy is ε-greedy with $\varepsilon = 0.1$. Offline transitions are generated using a reset mechanism that induces a sampling distribution different from the target-policy stationary distribution. The five-dimensional linear feature class contains Q^π but is not Bellman complete. We compare unweighted FQE with FQE weighted by the exact target-to-offline occupancy ratio.

C.2 Stochastic Continuing CartPole

Environment and policies. We use a stochastic continuing version of CartPole with state $S = (x, \dot{x}, \theta, \dot{\theta})$ and action $A \in \{0, 1\}$. Transitions follow the standard CartPole dynamics with Euler step 0.02, followed by independent Gaussian noise with standard deviation 0.001 in each state

coordinate. If the resulting successor state satisfies $|x| > 2.4$ or $|\theta| > 12$ degrees, it is replaced by an independent draw from $\mathrm{Unif}([-0.05, 0.05]^4)$. This reset is part of the transition kernel, so the process is continuing.

The reward is

$$r(s) = \left(2 - \frac{|\theta|}{12 \text{ degrees}}\right)\left(2 - \frac{|x|}{2.4}\right) - 1.$$

Both behavior and target policies are logistic. With $c = (0.8, 1.3, 12, 2)^\top$,

$$\Pr(A = 1 \mid S = s) = \sigma(c^\top s/T),$$

where $T = 1$ for the behavior policy and $T = 1.5$ for the target policy. We set the Bellman and occupancy discounts to

$$\gamma = \beta = 0.98.$$

Offline data. Each offline dataset contains $n = 25{,}000$ transitions sampled from the stationary behavior distribution after burn-in. Unweighted FQE and FORE-FQE use the same rewards and transitions.

FORE and FQE. FORE uses 128 joint random Fourier features with bandwidth multiplier one and ridge penalty 10^{-2}. Estimated ratios are truncated to $[10^{-4}, 50]$ and normalized to have empirical mean one.

The Q-class uses 64 random Fourier features with bandwidth 1.4. Both unweighted FQE and FORE-FQE use ridge penalty 10^{-5}, analytic expectations over the target action distribution, and the empirical projected fixed point. The two estimators differ only in the regression weights: unweighted FQE assigns unit weight to each transition, whereas FORE-FQE uses the normalized estimated occupancy ratio.

Evaluation. We compare the estimators across 50 independent offline datasets. Policy value is the discounted return under the target policy from the reset distribution. A large independent Monte Carlo sample of target-policy trajectories provides the reference value.

For dataset j, define the paired improvement in absolute value error by

$$G_j = \left|\widehat{V}_{j,\mathrm{unweighted}} - V^\pi\right| - \left|\widehat{V}_{j,\mathrm{FORE}} - V^\pi\right|.$$

We report the mean absolute error of each estimator, the mean paired improvement $\overline{G}$, its 95% paired-bootstrap confidence interval, and the proportion of datasets for which FORE-FQE has smaller absolute error.

D Contraction Results and Proofs for Discounted-Occupancy FQE

D.1 Stationary-Limit Contraction

Lemma S1 (Bellman contraction for Q-functions under a stationary measure)**.** *If μ_π is stationary for P_π and $r_0 \in L^2(\mu_\pi)$, then, for all $Q_1, Q_2 \in L^2(\mu_\pi)$,*

$$\|P_\pi(Q_1 - Q_2)\|_{2,\mu_\pi} \le \|Q_1 - Q_2\|_{2,\mu_\pi}.$$

Consequently, the Bellman operator $\mathcal{T}$ is a γ-contraction:

$$\|\mathcal{T}Q_1 - \mathcal{T}Q_2\|_{2,\mu_\pi} \le \gamma\|Q_1 - Q_2\|_{2,\mu_\pi}.$$

Proof. Let $h := Q_1 - Q_2$. Conditional Jensen's inequality gives

$$\{P_\pi h(s,a)\}^2 \le P_\pi(h^2)(s,a).$$

Integrating with respect to μ_π and using $\mu_\pi P_\pi = \mu_\pi$ yields

$$\|P_\pi h\|_{2,\mu_\pi}^2 \le \int P_\pi(h^2)\, d\mu_\pi = \int h^2\, d\mu_\pi = \|h\|_{2,\mu_\pi}^2.$$

Since $\mathcal{T}Q_1 - \mathcal{T}Q_2 = \gamma P_\pi h$, it follows that

$$\|\mathcal{T}Q_1 - \mathcal{T}Q_2\|_{2,\mu_\pi} \le \gamma\|Q_1 - Q_2\|_{2,\mu_\pi},$$

which proves both claims. □

Lemma S2 (Contraction of the projected Bellman operator)**.** *Assume that μ_π is stationary for P_π, $r_0 \in L^2(\mu_\pi)$, and $\mathcal{F}$ is a nonempty closed convex subset of $L^2(\mu_\pi)$. Then $\mathcal{T}_\mathcal{F} = \Pi_\mathcal{F}\mathcal{T}$ is a γ-contraction on $L^2(\mu_\pi)$; that is, for all $Q_1, Q_2 \in L^2(\mu_\pi)$,*

$$\|\mathcal{T}_\mathcal{F}Q_1 - \mathcal{T}_\mathcal{F}Q_2\|_{2,\mu_\pi} \le \gamma \|Q_1 - Q_2\|_{2,\mu_\pi}.$$

Proof. The metric projection onto a nonempty closed convex subset of a Hilbert space is nonexpansive. Therefore, Lemma S1 gives

$$\begin{aligned}\|\mathcal{T}_\mathcal{F}Q_1 - \mathcal{T}_\mathcal{F}Q_2\|_{2,\mu_\pi} &\le \|\mathcal{T}Q_1 - \mathcal{T}Q_2\|_{2,\mu_\pi} \\ &\le \gamma\|Q_1 - Q_2\|_{2,\mu_\pi},\end{aligned}$$

as claimed. □

Proof of Lemma 1. Let $h := Q - Q'$. Conditional Jensen's inequality and the occupancy recursion $\beta d_\beta P_\pi \le d_\beta$ give

$$\|P_\pi h\|^2_{2,d_\beta} \le \int h^2 \, d(d_\beta P_\pi) \le \beta^{-1}\|h\|^2_{2,d_\beta}.$$

Hence $\|\mathcal{T}Q - \mathcal{T}Q'\|_{2,d_\beta} \le \bar{\gamma}_\beta\|Q - Q'\|_{2,d_\beta}$. Condition C2 makes $\Pi_{\mathcal{F},\beta}$ nonexpansive in $L^2(d_\beta)$, so composing with this projection proves the claim. □

E Technical Lemmas

E.1 Local Maximal Inequality

Let $O_1, \dots, O_n \in \mathcal{O}$ be independent random variables. For any function $f : \mathcal{O} \to \mathbb{R}$, define

$$\|f\| := \sqrt{\frac{1}{n}\sum_{i=1}^n \mathbb{E}[f(O_i)^2]}. \tag{S1}$$

For a star-shaped class of functions $\mathcal{F}$ and a radius $\delta \in (0, \infty)$, define the localized Rademacher complexity

$$\mathcal{R}_n(\mathcal{F}, \delta) := \mathbb{E}\left[\sup_{\substack{f\in\mathcal{F} \\ \|f\|\le\delta}} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(O_i)\right],$$

where ϵ_i are i.i.d. Rademacher random variables.

The following lemma restates a local maximal inequality from Lemma 11 of Foster & Syrgkanis (2023); see also Lemma 11 of van der Laan et al. (2025b).

Lemma S3 (Local maximal inequality)**.** *Let $\mathcal{F}$ be a star-shaped class of functions satisfying $\sup_{f\in\mathcal{F}}\|f\|_\infty \le M$. Let $\delta = \delta_n \in (0,1)$ satisfy the critical radius condition $\mathcal{R}_n(\mathcal{F}, \delta) \le \delta^2$, and suppose that, as $n \to \infty$,*

$$\frac{1}{\sqrt{n}}\sqrt{\log\log(1/\delta_n)} = o(\delta_n).$$

Then there exists a constant $C > 0$ such that, for all $\eta \in (0,1)$, with probability at least $1 - \eta$, every $f \in \mathcal{F}$ satisfies

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^n \big(f(O_i) - \mathbb{E}[f(O_i)]\big)\right| &\le C\Big(\delta^2 + \delta\,\|f\|\Big) \\ &+ C\Big(\frac{\sqrt{\log(e/\eta)}\,\|f\|}{\sqrt{n}} + \frac{M\,\log(e/\eta)}{n}\Big).\end{aligned}$$

Proof. Apply the cited one-sided inequality to the symmetric star-shaped class $\mathcal{F}_\pm := \mathcal{F} \cup (-\mathcal{F})$. This class has the same envelope as $\mathcal{F}$, and its localized Rademacher complexity is at most twice that of $\mathcal{F}$. Thus a universal multiple of δ_n is a critical radius for $\mathcal{F}_\pm$; we absorb this multiple into the constant below. Lemma 11 in van der Laan et al. (2025b) gives a constant $C > 0$ such that, for all $u \ge 1$, with probability at least $1 - e^{-u^2}$, every $f \in \mathcal{F}_\pm$ satisfies

$$\frac{1}{n}\sum_{i=1}^n \big(f(O_i) - \mathbb{E}[f(O_i)]\big) \le C\Big(\delta_n^2 + \delta_n\,\|f\| + \frac{u\,\|f\|}{\sqrt{n}} + \frac{M\,u^2}{n}\Big).$$

Set $u := \sqrt{\log(e/\eta)}$, for which $e^{-u^2} = \eta/e \le \eta$. On the resulting event, apply the inequality to both f and $-f$. Then, with probability at least $1-\eta$, every $f \in \mathcal{F}$ satisfies

$$\left|\frac{1}{n}\sum_{i=1}^{n}\big(f(O_i) - \mathbb{E}[f(O_i)]\big)\right| \le C\Big(\delta_n^2 + \delta_n \|f\|\Big) + C\Big(\frac{\sqrt{\log(e/\eta)}\,\|f\|}{\sqrt{n}} + \frac{M\,\log(e/\eta)}{n}\Big).$$

This is the claimed two-sided maximal inequality. □

The next lemma bounds the localized Rademacher complexity by the uniform entropy integral in the form obtained from Theorem 2.1 of van der Vaart & Wellner (2011).

For any distribution Q and any uniformly bounded function class $\mathcal{F}$, let $N(\varepsilon, \mathcal{F}, L^2(\mathsf{Q}))$ denote the ε-covering number of $\mathcal{F}$ under the $L^2(\mathsf{Q})$ norm (van der Vaart & Wellner, 1996). Define the uniform entropy integral of $\mathcal{F}$ by

$$\mathcal{J}(\delta, \mathcal{F}) := \int_0^\delta \sup_{\mathsf{Q}} \sqrt{\log N(\epsilon, \mathcal{F}, L^2(\mathsf{Q}))}\, d\epsilon,$$

where the supremum is over all discrete probability distributions Q.

Lemma S4. *Let $\mathcal{F}$ be a star-shaped class of functions such that $\sup_{f\in\mathcal{F}} \|f\|_\infty \le M$. Then, for every $\delta > 0$,*

$$\mathcal{R}_n(\mathcal{F}, \delta) \lesssim \frac{1}{\sqrt{n}}\,\mathcal{J}(\delta, \mathcal{F})\left(1 + \frac{\mathcal{J}(\delta, \mathcal{F})}{\delta^2\sqrt{n}}\right),$$

where the implicit constant depends only on M.

Proof of Lemma S4. Rademacher symmetrization followed by Theorem 2.1 of van der Vaart & Wellner (2011), applied to the class localized by $\|f\| \le \delta$ and rescaled to have unit envelope, gives

$$\mathcal{R}_n(\mathcal{F}, \delta) \le \frac{C_M}{\sqrt{n}}\,\mathcal{J}(\delta, \mathcal{F})\left\{1 + \frac{\mathcal{J}(\delta, \mathcal{F})}{\delta^2\sqrt{n}}\right\},$$

where C_M depends only on M. This is the claimed bound. □

E.2 One-Step Error for an Inexact FQE Update

Recall the occupancy density ratio $w_\beta = \mathrm{d}d_\beta/\mathrm{d}\nu_b$, and let $\widehat{w}_\beta$ denote an estimate of this ratio trained independently of $\mathcal{D}_n$, as required by Condition C5.

Let $\mathcal{F}$ be a convex function class and let $\widehat{Q}^{\mathrm{init}} \in \mathcal{F}$ be an initial, possibly data-dependent, estimate. Define the weighted population risk

$$\widehat{R}_0(Q) = \mathbb{E}_{\nu_b}\Big[w_\beta(S,A)\left\{R + \gamma(\pi\widehat{Q}^{\mathrm{init}})(S') - Q(S,A)\right\}^2\Big].$$

Following Algorithm 1, define the empirical minimizer

$$\widehat{Q}_n := \arg\min_{Q\in\mathcal{F}} \widehat{R}_n(Q),$$

where

$$\widehat{R}_n(Q) = \sum_{i=1}^{n} \widehat{w}_\beta(S_i, A_i)\left\{R_i + \gamma(\pi\widehat{Q}^{\mathrm{init}})(S_i') - Q(S_i, A_i)\right\}^2.$$

Lemma S5 (Projected Bellman update as an occupancy-weighted population-risk minimizer)**.** *Assume Conditions C1, C2, C3, and C6. Then*

$$\mathcal{T}_{\mathcal{F},\beta}(\widehat{Q}^{\mathrm{init}}) = \arg\min_{Q\in\mathcal{F}} \widehat{R}_0(Q).$$

Proof. Because $w_\beta = \mathrm{d}d_\beta/\mathrm{d}\nu_b$, the measures $w_\beta\nu_b$ and d_β coincide. By definition of the occupancy projection, $\mathcal{T}_{\mathcal{F},\beta}(\widehat{Q}^{\mathrm{init}})$ minimizes the $L^2(w_\beta\nu_b)$ distance to $\mathcal{T}(\widehat{Q}^{\mathrm{init}})$ over $\mathcal{F}$. Moreover,

$$E\{R + \gamma(\pi\widehat{Q}^{\mathrm{init}})(S') \mid S, A\} = \mathcal{T}(\widehat{Q}^{\mathrm{init}})(S, A).$$

Expanding the square conditional on (S, A) shows that replacing the conditional Bellman target by the noisy target adds only $E_{\nu_b}[w_\beta \operatorname{Var}\{R + \gamma(\pi\widehat{Q}^{\mathrm{init}})(S') \mid S, A\}]$, which does not depend on Q. Hence $\mathcal{T}_{\mathcal{F},\beta}(\widehat{Q}^{\mathrm{init}})$ also minimizes $\widehat{R}_0$ over $\mathcal{F}$. □

Recall the critical radius

$$\delta_n := \inf\Big\{\delta > 0 : \mathcal{J}(\delta, \mathcal{F}) \le \sqrt{n}\,\delta^2\Big\} \vee \sqrt{\frac{\log(en)}{n}}.$$

Lemma S6. *Let w be a state-action function with $\|w\|_\infty \le B$, where $B \ge 1$. Assume that R and $\mathcal{F}$ are uniformly bounded by M, and that the critical radius δ_n above is finite and satisfies $\delta_n < 1$. Then there exists a constant $C = C(M) > 0$ such that, for all $\tau \in (0,1)$, with probability at least $1-\tau$, simultaneously for all $Q_1, Q_2 \in \mathcal{F}$ and $V_1 \in \mathcal{V}_\mathcal{F} := \{\pi Q : Q \in \mathcal{F}\}$,*

$$\begin{aligned}
&\big|(P_n - P_0)\big[w\,(Q_1 - Q_2)\,\{R + \gamma V_1 - Q_2\}\big]\big| \\
&\le C\Big(B(\delta_n)^2 + \delta_n\,\|w(Q_1 - Q_2)\|\Big) \\
&\qquad + C\Big(\frac{\sqrt{\log(e/\tau)}\,\|w(Q_1 - Q_2)\|}{\sqrt{n}} + \frac{B\log(e/\tau)}{n}\Big),
\end{aligned}$$

where $\|\cdot\|$ denotes the offline-sampling norm $L^2(P_0)$.

Proof. Fix a bounded weight function w with $\|w\|_\infty \le B$. If w is estimated on an independent sample, condition on that sample throughout. Let

$$\mathcal{V}_\mathcal{F} := \{\,\pi Q : Q \in \mathcal{F}\,\}, \qquad (\pi Q)(s) := \sum_{a \in \mathcal{A}} \pi(a \mid s) Q(s,a).$$

Define the function class

$$\begin{aligned}
\mathcal{G}_w := \Big\{&(s,a,r,s') \mapsto w(s,a)\,\big(Q_1(s,a) - Q_2(s,a)\big) \\
&\times \big(r + \gamma V_1(s') - Q_2(s,a)\big) :\ Q_1, Q_2 \in \mathcal{F},\ V_1 \in \mathcal{V}_\mathcal{F}\Big\}.
\end{aligned}$$

By boundedness, each $g \in \mathcal{G}_w$ satisfies

$$|g(s,a,r,s')| \le C(M)\,\big|w(s,a)\{Q_1(s,a) - Q_2(s,a)\}\big|,$$

and hence

$$\|g\|_{L^2(P_0)} \le C(M)\,\|w(Q_1 - Q_2)\|_{L^2(P_0)}. \tag{S2}$$

Entropy control. For any discrete distribution Q_S on $\mathcal{S}$, let $\bar{\mathsf{Q}}$ be its target-policy lift to the state–action space:

$$\bar{\mathsf{Q}}(s,a) := \mathsf{Q}_S(s)\pi(a \mid s).$$

For any $Q, \widetilde{Q} \in \mathcal{F}$, conditional Jensen's inequality gives

$$\big|\pi(Q - \widetilde{Q})(s)\big|^2 \le \sum_{a \in \mathcal{A}} \pi(a \mid s)\{Q(s,a) - \widetilde{Q}(s,a)\}^2.$$

Therefore

$$\|\pi Q - \pi\widetilde{Q}\|_{L^2(\mathsf{Q}_S)} \le \|Q - \widetilde{Q}\|_{L^2(\bar{\mathsf{Q}})}.$$

It follows that, for every $\varepsilon > 0$,

$$N(\varepsilon, \mathcal{V}_\mathcal{F}, L^2(\mathsf{Q}_S)) \le N(\varepsilon, \mathcal{F}, L^2(\bar{\mathsf{Q}})).$$

Since the entropy supremum for $\mathcal{F}$ ranges over all discrete state–action distributions, the uniform entropy of $\mathcal{V}_\mathcal{F}$ is bounded by that of $\mathcal{F}$.

Let $\mathcal{G}_{w/B} := \{g/B : g \in \mathcal{G}_w\}$. The map $(Q_1, Q_2, V_1) \mapsto (w/B)(Q_1 - Q_2)(r + \gamma V_1 - Q_2)$ is a Lipschitz pointwise transformation of uniformly bounded functions with multiplier envelope one. Standard permanence properties of entropy under Lipschitz maps and bounded multipliers (van der Vaart & Wellner, 1996) imply that there is a constant $C_1 = C_1(M)$ such that, for all $\delta > 0$,

$$\mathcal{J}(\delta, \mathcal{G}_{w/B}) \le C_1\,\mathcal{J}(C_1\delta, \mathcal{F}).$$

Thus $\mathcal{G}_{w/B}$ has the same critical radius as $\mathcal{F}$, up to constants absorbed into C.

Let

$$\mathcal{H}_{w/B} := \{\,tg : g \in \mathcal{G}_{w/B},\ t \in [-1,1]\,\}$$

be the symmetric star-shaped hull, and let $M_G = C(M)$ be an envelope for $\mathcal{G}_{w/B}$. The covering bound for star-shaped hulls gives, for all $\delta > 0$,

$$\mathcal{J}(\delta, \mathcal{H}_{w/B}) \ \lesssim\ \mathcal{J}(C_1\delta, \mathcal{G}_{w/B}) + \delta\sqrt{\log\Big(1 + \frac{M_G}{\delta}\Big)}.$$

The additional term is at most a constant multiple of $\sqrt{n}\,\delta^2$ whenever $\delta \gtrsim \sqrt{\log(eM_G n)/n}$. Since M_G is structural and the definition of δ_n includes the parametric floor, δ_n is a critical radius for $\mathcal{H}_{w/B}$, up to constants depending only on M. The same floor also makes the side condition in Lemma S3 automatic: for all sufficiently large n with $\delta_n < 1$,

$$\frac{n^{-1/2}\sqrt{\log\log(1/\delta_n)}}{\delta_n} \le \sqrt{\frac{\log\log\{\sqrt{n/\log(en)}\}}{\log(en)}} = o(1).$$

For the finitely many remaining sample sizes, boundedness gives the same conclusion after enlarging the constant. Lemmas S4 and S3 applied to $\mathcal{H}_{w/B}$ give that, with probability at least $1-\tau$, every $h \in \mathcal{G}_{w/B}$ satisfies

$$\begin{aligned}|(P_n - P_0)h| &\le C\Big((\delta_n)^2 + \delta_n \|h\|_{L^2(P_0)}\Big) \\ &\quad + C\Big(\frac{\sqrt{\log(e/\tau)}\,\|h\|_{L^2(P_0)}}{\sqrt{n}} + \frac{\log(e/\tau)}{n}\Big).\end{aligned}$$

Multiplying this display by B gives that every $g \in \mathcal{G}_w$ satisfies

$$\begin{aligned}|(P_n - P_0)g| &\le C\Big(B(\delta_n)^2 + \delta_n \|g\|_{L^2(P_0)}\Big) \\ &\quad + C\Big(\frac{\sqrt{\log(e/\tau)}\,\|g\|_{L^2(P_0)}}{\sqrt{n}} + \frac{B\,\log(e/\tau)}{n}\Big).\end{aligned}$$

Using (S2) yields the claimed uniform bound. Because the event is simultaneous over $Q_1, Q_2 \in \mathcal{F}$ and $V_1 \in \mathcal{V}_{\mathcal{F}}$, it can later be specialized to fitted, data-dependent choices such as $Q_1 = \mathcal{T}_{\mathcal{F},\beta}(\widehat{Q}^{\mathrm{init}})$, $Q_2 = \widehat{Q}_n$, and $V_1 = \pi\widehat{Q}^{\mathrm{init}}$. □

Proof of Lemma 2. Fix an iteration k, and write $\widehat{Q}^{\mathrm{init}} = \widehat{Q}^{(k)}$ and $\widehat{Q}_n = \widehat{Q}^{(k+1)}$. Under occupancy weighting, $w_\beta = \mathrm{d}d_\beta/\mathrm{d}\nu_b$, so $w_\beta \nu_b = d_\beta$. For brevity, define

$$\|f\|^2_{2,\widehat{w}_\beta} := P_0\big[\widehat{w}_\beta f^2\big].$$

We first use the empirical first-order condition, then lower bound its population counterpart and control the remaining empirical-process term.

By first-order optimality of $\widehat{Q}_n = \arg\min_{Q\in\mathcal{F}} \widehat{R}_n(Q)$ and convexity of $\mathcal{F}$, we have, for all $Q \in \mathcal{F}$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \widehat{w}_\beta(S_i,A_i)\,\{Q(S_i,A_i) - \widehat{Q}_n(S_i,A_i)\} \\ \times \big\{R_i + \gamma\,(\pi\widehat{Q}^{\mathrm{init}})(S_i') - \widehat{Q}_n(S_i,A_i)\big\} \le 0.\end{aligned}$$

Since $\mathcal{T}_{\mathcal{F},\beta}(\widehat{Q}^{\mathrm{init}}) \in \mathcal{F}$, taking $Q = \mathcal{T}_{\mathcal{F},\beta}(\widehat{Q}^{\mathrm{init}})$ gives

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \widehat{w}_\beta(S_i,A_i)\,\{\mathcal{T}_{\mathcal{F},\beta}(\widehat{Q}^{\mathrm{init}})(S_i,A_i) - \widehat{Q}_n(S_i,A_i)\} \\ \times \big\{R_i + \gamma\,(\pi\widehat{Q}^{\mathrm{init}})(S_i') - \widehat{Q}_n(S_i,A_i)\big\} \le 0.\end{aligned}$$

Adding and subtracting the population expectation and rearranging,

$$\begin{aligned}&P_0\big[\widehat{w}_\beta\,(\mathcal{T}_{\mathcal{F},\beta}(\widehat{Q}^{\mathrm{init}}) - \widehat{Q}_n)\,\{R + \gamma(\pi\widehat{Q}^{\mathrm{init}}) - \widehat{Q}_n\}\big] \\ &\le (P_0 - P_n)\big[\widehat{w}_\beta\,(\mathcal{T}_{\mathcal{F},\beta}(\widehat{Q}^{\mathrm{init}}) - \widehat{Q}_n)\,\{R + \gamma(\pi\widehat{Q}^{\mathrm{init}}) - \widehat{Q}_n\}\big].\end{aligned} \tag{S3}$$

In the displays below, R, $\widehat{Q}_n$, and $\pi\widehat{Q}^{\mathrm{init}}$ denote their coordinate evaluations at (S, A, R, S'); in particular, the Bellman target uses $(\pi\widehat{Q}^{\mathrm{init}})(S')$.

Set

$$\Delta := \mathcal{T}_{\mathcal{F},\beta}(\widehat{Q}^{\mathrm{init}}) - \widehat{Q}_n, \qquad b := \mathcal{T}(\widehat{Q}^{\mathrm{init}}) - \mathcal{T}_{\mathcal{F},\beta}(\widehat{Q}^{\mathrm{init}}).$$

Taking conditional expectations given (S, A) gives

$$P_0\Big[\widehat{w}_\beta\Delta\{R + \gamma(\pi\widehat{Q}^{\mathrm{init}}) - \widehat{Q}_n\}\Big] = P_0[\widehat{w}_\beta\Delta(\Delta + b)].$$

Because $\Delta \in \mathcal{H}_{\mathcal{F}}$, Condition C7 gives

$$\|\Delta\|_{2,d_\beta} \le \lambda_{\mathcal{F}}^{-1/2}\|\Delta\|_{2,\widehat{w}_\beta}.$$

Moreover, $w_\beta\nu_b = d_\beta$, and the variational inequality for projection onto a convex set gives $P_0[w_\beta\Delta b] \ge 0$: the projection point is $\mathcal{T}_{\mathcal{F},\beta}(\widehat{Q}^{\mathrm{init}})$, the comparison point is $\widehat{Q}_n \in \mathcal{F}$, and b is

the residual from the unprojected Bellman target. By Condition C6 and the ratio convention above, $\widehat{w}_\beta - w_\beta = w_\beta(\widehat{w}_\beta/w_\beta - 1)$ ν_b-almost surely. Therefore, by Cauchy–Schwarz and the definition of $\varepsilon_{\mathrm{Bell},w}(\widehat{Q}^{\mathrm{init}})$,

$$P_0[\widehat{w}_\beta \Delta b] = P_0[w_\beta \Delta b] + P_0\left[w_\beta\left(\frac{\widehat{w}_\beta}{w_\beta} - 1\right)\Delta b\right] \geq -\frac{\varepsilon_{\mathrm{Bell},w}(\widehat{Q}^{\mathrm{init}})}{\sqrt{\lambda_{\mathcal{F}}}}\|\Delta\|_{2,\widehat{w}_\beta}.$$

Combining these inequalities yields

$$P_0\left[\widehat{w}_\beta \Delta\{R + \gamma(\pi\widehat{Q}^{\mathrm{init}}) - \widehat{Q}_n\}\right] \geq \|\Delta\|^2_{2,\widehat{w}_\beta} - \frac{\varepsilon_{\mathrm{Bell},w}(\widehat{Q}^{\mathrm{init}})}{\sqrt{\lambda_{\mathcal{F}}}}\|\Delta\|_{2,\widehat{w}_\beta}.$$

Applying this lower bound to (S3) and rearranging,

$$\begin{aligned}
&\|\Delta\|^2_{2,\widehat{w}_\beta} \\
&\leq (P_0 - P_n)\left[\widehat{w}_\beta\,\Delta\,\{R + \gamma(\pi\widehat{Q}^{\mathrm{init}}) - \widehat{Q}_n\}\right] \\
&\quad + \frac{\varepsilon_{\mathrm{Bell},w}(\widehat{Q}^{\mathrm{init}})}{\sqrt{\lambda_{\mathcal{F}}}}\|\Delta\|_{2,\widehat{w}_\beta}.
\end{aligned} \tag{S4}$$

We now invoke Condition C5. Conditional on the independent sample used to estimate the weights, $\widehat{w}_\beta$ is a fixed bounded multiplier. Lemma S6 applies uniformly over the regression and Bellman-target classes, so it may be specialized to the data-dependent fitted functions $\mathcal{T}_{\mathcal{F},\beta}(\widehat{Q}^{\mathrm{init}})$, $\widehat{Q}_n$, and $\pi\widehat{Q}^{\mathrm{init}}$ even though the FQE regressions reuse the same Bellman-regression sample. Condition C6 gives

$$\|\widehat{w}_\beta\Delta\|^2_{L^2(P_0)} \leq \kappa_{\mathrm{cov}}\|\Delta\|^2_{2,\widehat{w}_\beta}.$$

Applying Lemma S6 with $w = \widehat{w}_\beta$, and then integrating over the independent weight-estimation sample, gives, for $C = C(M) > 0$, with probability at least $1 - \tau$,

$$\begin{aligned}
&\left|(P_0 - P_n)\left[\widehat{w}_\beta\,\Delta\,\{R + \gamma(\pi\widehat{Q}^{\mathrm{init}}) - \widehat{Q}_n\}\right]\right| \\
&\leq C\kappa_{\mathrm{cov}}\left((\delta_n)^2 + \frac{\log(e/\tau)}{n}\right) \\
&\quad + C\sqrt{\kappa_{\mathrm{cov}}}\left(\delta_n + \sqrt{\frac{\log(e/\tau)}{n}}\right)\|\Delta\|_{2,\widehat{w}_\beta}.
\end{aligned}$$

Combining this high-probability bound with (S4), and writing

$$y := \|\Delta\|_{2,\widehat{w}_\beta}, \qquad a := \delta_n + \sqrt{\log(e/\tau)/n}, \qquad \omega := \varepsilon_{\mathrm{Bell},w}(\widehat{Q}^{\mathrm{init}}),$$

gives

$$y^2 \leq C\kappa_{\mathrm{cov}}a^2 + \left(C\sqrt{\kappa_{\mathrm{cov}}}\,a + \frac{\omega}{\sqrt{\lambda_{\mathcal{F}}}}\right)y.$$

Solving this scalar inequality yields

$$y \leq C\left\{\sqrt{\kappa_{\mathrm{cov}}}\,a + \frac{\omega}{\sqrt{\lambda_{\mathcal{F}}}}\right\}.$$

Condition C7 and $\sqrt{\kappa_{\mathrm{cov}}}a = \varepsilon_{\mathrm{stat}}(1,\tau)$ therefore give

$$\|\mathcal{T}_{\mathcal{F},\beta}(\widehat{Q}^{\mathrm{init}}) - \widehat{Q}_n\|_{2,d_\beta} \leq C\left\{\frac{\varepsilon_{\mathrm{stat}}(1,\tau)}{\sqrt{\lambda_{\mathcal{F}}}} + \frac{\varepsilon_{\mathrm{Bell},w}(\widehat{Q}^{\mathrm{init}})}{\lambda_{\mathcal{F}}}\right\},$$

which is the bound in Lemma 2. □

F Proofs of Main Results

Theorem S2 (Approximation error of the occupancy-projected fixed point)**.** *Under Conditions C1–C2,*

$$\|Q^\pi_{\mathcal{F},\beta} - Q^\pi\|_{2,d_\beta} \leq \frac{1}{1 - \bar{\gamma}_\beta}\inf_{f\in\mathcal{F}}\|f - Q^\pi\|_{2,d_\beta}.$$

Proof of Theorem S2. Let $f^\star := \Pi_{\mathcal{F},\beta} Q^\pi$. The fixed-point identities, $\mathcal{T} Q^\pi = Q^\pi$, and Lemma 1 give

$$\begin{aligned}\|Q^\pi_{\mathcal{F},\beta} - Q^\pi\|_{2,d_\beta} &\le \|Q^\pi_{\mathcal{F},\beta} - f^\star\|_{2,d_\beta} + \|f^\star - Q^\pi\|_{2,d_\beta} \\ &= \|\mathcal{T}_{\mathcal{F},\beta} Q^\pi_{\mathcal{F},\beta} - \mathcal{T}_{\mathcal{F},\beta} Q^\pi\|_{2,d_\beta} + \|f^\star - Q^\pi\|_{2,d_\beta} \\ &\le \bar{\gamma}_\beta \|Q^\pi_{\mathcal{F},\beta} - Q^\pi\|_{2,d_\beta} + \inf_{f\in\mathcal{F}} \|f - Q^\pi\|_{2,d_\beta}.\end{aligned}$$

Rearranging proves the stated bound. □

F.1 Proofs of the Completeness and Realizability Refinements

Proof of Proposition 1. For $Q_\theta = \phi^\top \theta$, write $\mathcal{T}_{\mathcal{F},\beta} Q_\theta = \phi^\top \overline{\theta}$ for some $\overline{\theta} \in \Theta$. Then

$$\mathcal{T} Q_\theta - \mathcal{T}_{\mathcal{F},\beta} Q_\theta = r_0 + \gamma (P_\pi \phi)^\top \theta - \phi^\top \overline{\theta} \in \mathcal{G}_{\mathrm{res}}.$$

For any $g = a^\top \psi \in \mathcal{G}_{\mathrm{res}}$, Cauchy–Schwarz gives, for d_β-almost every x,

$$|g(x)| \le \{a^\top \Sigma_\psi a\}^{1/2} \{\psi(x)^\top \Sigma_\psi^{-1} \psi(x)\}^{1/2} \le \sqrt{\kappa_{\mathrm{res}}}\, \|g\|_{2,d_\beta}.$$

Applying this inequality to each Bellman projection residual proves Condition C8 with $\alpha = 1$ and $C_\infty = \sqrt{\kappa_{\mathrm{res}}}$. □

Proof of Theorem 1. Let

$$e_k := \|\widehat{Q}^{(k)} - Q^\pi_{\mathcal{F},\beta}\|_{2,d_\beta}, \qquad B(Q) := \mathcal{T} Q - \mathcal{T}_{\mathcal{F},\beta} Q.$$

Throughout the proof, $C = C(M, \alpha, \beta, \gamma)$ may change from line to line and is independent of C_∞, ε_w, and $\lambda_{\mathcal{F}}$. We first bound the projection residual using value-function approximation. Let $Q_0 := \Pi_{\mathcal{F},\beta} Q^\pi$, so $\|Q_0 - Q^\pi\|_{2,d_\beta} = \varepsilon_{\mathrm{app}}$. Since $Q^\pi_{\mathcal{F},\beta}$ is the fixed point of $\mathcal{T}_{\mathcal{F},\beta}$, the definition of metric projection gives

$$\begin{aligned}\|B(Q^\pi_{\mathcal{F},\beta})\|_{2,d_\beta} &= \mathrm{dist}_{2,d_\beta}(\mathcal{T} Q^\pi_{\mathcal{F},\beta}, \mathcal{F}) \\ &\le \|\mathcal{T} Q^\pi_{\mathcal{F},\beta} - Q_0\|_{2,d_\beta} \\ &\le \bar{\gamma}_\beta \|Q^\pi_{\mathcal{F},\beta} - Q^\pi\|_{2,d_\beta} + \varepsilon_{\mathrm{app}} \le \frac{\varepsilon_{\mathrm{app}}}{1-\bar{\gamma}_\beta},\end{aligned}$$

where the last inequality uses Theorem S2. Because distance to a closed set is nonexpansive and $\mathcal{T}$ is a $\bar{\gamma}_\beta$-contraction in $L^2(d_\beta)$, it follows that

$$\|B(\widehat{Q}^{(k)})\|_{2,d_\beta} \le \min\left\{\varepsilon_{\mathrm{Bell}}, \bar{\gamma}_\beta e_k + \frac{\varepsilon_{\mathrm{app}}}{1-\bar{\gamma}_\beta}\right\}. \tag{S5}$$

Condition C8, Cauchy–Schwarz, and (S5) imply

$$\begin{aligned}\varepsilon_{\mathrm{Bell},w}(\widehat{Q}^{(k)}) &\le \varepsilon_w \|B(\widehat{Q}^{(k)})\|_\infty \\ &\le C_\infty \varepsilon_w \min\left\{\varepsilon^\alpha_{\mathrm{Bell}}, \left(\bar{\gamma}_\beta e_k + \frac{\varepsilon_{\mathrm{app}}}{1-\bar{\gamma}_\beta}\right)^\alpha\right\}.\end{aligned} \tag{S6}$$

Apply Lemma 2 at confidence level τ/K, take a union bound over the K regressions, and combine the resulting pointwise one-step bounds with Lemma 1. On an event of probability at least $1-\tau$, for every $0 \le k < K$,

$$e_{k+1} \le \bar{\gamma}_\beta e_k + C\frac{\varepsilon_{\mathrm{stat}}(K,\tau)}{\sqrt{\lambda_{\mathcal{F}}}} + C\frac{C_\infty \varepsilon_w}{\lambda_{\mathcal{F}}} \min\left\{\varepsilon^\alpha_{\mathrm{Bell}}, \left(\bar{\gamma}_\beta e_k + \frac{\varepsilon_{\mathrm{app}}}{1-\bar{\gamma}_\beta}\right)^\alpha\right\}. \tag{S7}$$

Using the completeness branch in (S7) and $\bar{\gamma}_\beta \le \gamma_{\beta,\star}$ gives

$$e_{k+1} \le \gamma_{\beta,\star} e_k + C\frac{\varepsilon_{\mathrm{stat}}(K,\tau)}{\sqrt{\lambda_{\mathcal{F}}}} + C\frac{C_\infty \varepsilon_w}{\lambda_{\mathcal{F}}} \varepsilon^\alpha_{\mathrm{Bell}}.$$

For the realizability branch, suppose first that $\alpha \in (0,1)$. By subadditivity of $t \mapsto t^\alpha$ and Young's inequality with coefficient $\gamma_{\beta,\star} - \bar{\gamma}_\beta > 0$,

$$C\frac{C_\infty \varepsilon_w}{\lambda_{\mathcal{F}}}\left(\bar{\gamma}_\beta e_k + \frac{\varepsilon_{\mathrm{app}}}{1-\bar{\gamma}_\beta}\right)^\alpha \le (\gamma_{\beta,\star} - \bar{\gamma}_\beta) e_k + C\left\{\left(\frac{C_\infty \varepsilon_w}{\lambda_{\mathcal{F}}}\right)^{1/(1-\alpha)} + \frac{C_\infty \varepsilon_w}{\lambda_{\mathcal{F}}}\left(\frac{\varepsilon_{\mathrm{app}}}{1-\bar{\gamma}_\beta}\right)^\alpha\right\}.$$

When $\alpha = 0$, the same bound follows directly from the convention $x^0 = 1$, after enlarging C. Therefore both branches imply

$$e_{k+1} \le \gamma_{\beta,\star} e_k + C\frac{\varepsilon_{\mathrm{stat}}(K,\tau)}{\sqrt{\lambda_{\mathcal{F}}}} + C\min\left\{\frac{C_\infty \varepsilon_w}{\lambda_{\mathcal{F}}}\varepsilon^\alpha_{\mathrm{Bell}}, \left(\frac{C_\infty \varepsilon_w}{\lambda_{\mathcal{F}}}\right)^{1/(1-\alpha)} + \frac{C_\infty \varepsilon_w}{\lambda_{\mathcal{F}}}\left(\frac{\varepsilon_{\mathrm{app}}}{1-\bar{\gamma}_\beta}\right)^\alpha\right\}.$$

Unrolling this recursion and using $\sum_{j=0}^{K-1} \gamma_{\beta,\star}^j \le (1-\gamma_{\beta,\star})^{-1}$ bounds e_K. Finally, Theorem S2 gives

$$\|\widehat{Q}^{(K)} - Q^\pi\|_{2,d_\beta} \le e_K + \frac{\varepsilon_{\text{app}}}{1-\bar{\gamma}_\beta},$$

which proves Theorem 1. □

Proof of Corollary S1. Let

$$e_k := \|\widehat{Q}^{(k)} - Q^\pi_{\mathcal{F},\beta}\|_{2,d_\beta}.$$

Equation (S5), Proposition 1, and Cauchy–Schwarz give

$$\varepsilon_{\text{Bell},w}(\widehat{Q}^{(k)}) \le \sqrt{\kappa_{\text{res}}}\,\varepsilon_w \min\left\{\varepsilon_{\text{Bell}}, \bar{\gamma}_\beta e_k + \frac{\varepsilon_{\text{app}}}{1-\bar{\gamma}_\beta}\right\}.$$

Applying Lemma 2 at confidence level τ/K, taking a union bound, and using Lemma 1 yield, simultaneously for $0 \le k < K$,

$$e_{k+1} \le \bar{\gamma}_\beta e_k + C\frac{\varepsilon_{\text{stat}}(K,\tau)}{\sqrt{\lambda_{\mathcal{F}}}} + C\frac{\varepsilon_w}{\lambda_{\mathcal{F}}} \min\left\{\varepsilon_{\text{Bell}}, \bar{\gamma}_\beta e_k + \frac{\varepsilon_{\text{app}}}{1-\bar{\gamma}_\beta}\right\}. \tag{S8}$$

Using the completeness branch of (S8) and $\bar{\gamma}_\beta \le \gamma_w$ gives

$$e_{k+1} \le \gamma_w e_k + C\frac{\varepsilon_{\text{stat}}(K,\tau)}{\sqrt{\lambda_{\mathcal{F}}}} + C\frac{\varepsilon_w}{\lambda_{\mathcal{F}}}\varepsilon_{\text{Bell}}.$$

Using the realizability branch and the definition of γ_w gives

$$e_{k+1} \le \gamma_w e_k + C\frac{\varepsilon_{\text{stat}}(K,\tau)}{\sqrt{\lambda_{\mathcal{F}}}} + C\frac{\varepsilon_w}{\lambda_{\mathcal{F}}}\frac{\varepsilon_{\text{app}}}{1-\bar{\gamma}_\beta}.$$

Taking the smaller bound and unrolling the resulting recursion gives

$$e_K \le \gamma_w^K e_0 + \frac{C}{1-\gamma_w}\left[\frac{\varepsilon_{\text{stat}}(K,\tau)}{\sqrt{\lambda_{\mathcal{F}}}} + \frac{\varepsilon_w}{\lambda_{\mathcal{F}}}\min\left\{\varepsilon_{\text{Bell}}, \frac{\varepsilon_{\text{app}}}{1-\bar{\gamma}_\beta}\right\}\right].$$

Finally, $\|\widehat{Q}^{(K)} - Q^\pi\|_{2,d_\beta} \le e_K + \varepsilon_{\text{app}}/(1-\bar{\gamma}_\beta)$ by Theorem S2, proving Corollary S1. □

Proof of Theorem S1. By Lemma S2, both projected Bellman operators are γ-contractions on $\mathcal{F}$; changing the reward does not affect differences of Bellman updates. Their fixed-point identities and the triangle inequality therefore give

$$(1-\gamma)\|\widehat{Q}^\pi_{\mathcal{F}} - Q^\pi_{\mathcal{F}}\|_{2,\mu_\pi} \le \|\widehat{\mathcal{T}}_{\mathcal{F}}(\widehat{Q}^\pi_{\mathcal{F}}) - \mathcal{T}_{\mathcal{F}}(\widehat{Q}^\pi_{\mathcal{F}})\|_{2,\mu_\pi}. \tag{S9}$$

Next, for any Q,

$$\begin{aligned}\|\widehat{\mathcal{T}}_{\mathcal{F}}(Q) - \mathcal{T}_{\mathcal{F}}(Q)\|_{2,\mu_\pi} &= \|\Pi_{\mathcal{F}}\widehat{\mathcal{T}}(Q) - \Pi_{\mathcal{F}}\mathcal{T}(Q)\|_{2,\mu_\pi} \\ &\le \|\widehat{\mathcal{T}}(Q) - \mathcal{T}(Q)\|_{2,\mu_\pi} \\ &= \|\widehat{r} - r_0\|_{2,\mu_\pi},\end{aligned}$$

because $\widehat{\mathcal{T}}(Q) - \mathcal{T}(Q) = \widehat{r} - r_0$ and $\Pi_{\mathcal{F}}$ is nonexpansive in $L^2(\mu_\pi)$. Substituting into (S9) yields

$$\|\widehat{Q}^\pi_{\mathcal{F}} - Q^\pi_{\mathcal{F}}\|_{2,\mu_\pi} \le \frac{1}{1-\gamma}\|\widehat{r} - r_0\|_{2,\mu_\pi}.$$

Linear subspaces. If $\mathcal{F}$ is a linear subspace of $L^2(\mu_\pi)$, then $\Pi_{\mathcal{F}}$ is a linear orthogonal projection. Thus, for every $Q \in L^2(\mu_\pi)$,

$$\widehat{\mathcal{T}}_{\mathcal{F}}(Q) - \mathcal{T}_{\mathcal{F}}(Q) = \Pi_{\mathcal{F}}(\widehat{r} - r_0).$$

Substituting this into (S9) yields

$$\|\widehat{Q}^\pi_{\mathcal{F}} - Q^\pi_{\mathcal{F}}\|_{2,\mu_\pi} \le \frac{1}{1-\gamma}\|\Pi_{\mathcal{F}}(\widehat{r} - r_0)\|_{2,\mu_\pi}.$$

This proves the second bound. □

G FORE Guarantees for Occupancy-Ratio Estimation and FQE

This appendix gives a finite-sample error bound for FORE and verifies that the estimated weights satisfy Conditions C6 and C7 required by the FQE analysis. Combining these results with Theorem 1 proves Corollary S2.

G.1 FORE Estimator

Write $X = (S, A)$. From each observed state–action pair, draw $A^0 \sim \pi(\cdot \mid S)$ and set $X^0 = (S, A^0)$. Independently, draw

$$S' \sim P(\cdot \mid X), \qquad A' \sim \pi(\cdot \mid S'),$$

and set $X^+ = (S', A')$. Let $P_\pi(\cdot \mid x)$ denote the resulting target-policy transition kernel on $\mathcal{S} \times \mathcal{A}$. For a normalized ratio w, define

$$\mathsf{B}^\pi_\beta w := (1-\beta)\omega_{\pi/b} + \beta \frac{\mathrm{d}\{(w\nu_b)P_\pi\}}{\mathrm{d}\nu_b}.$$

The target occupancy ratio is the fixed point $\mathsf{B}^\pi_\beta w_\beta = w_\beta$.

For nonnegative f and $g > 0$, define

$$D^{\mathrm{gen}}_{\nu_b}(f\|g) := \mathbb{E}_{\nu_b}\left[f(X)\log\frac{f(X)}{g(X)} - f(X) + g(X)\right].$$

Let $\mathcal{G}_w$ be a class of log-ratio functions and define the normalized exponential class

$$\mathcal{W}_{\mathrm{F}} := \left\{ w_g(x) = \frac{\exp\{g(x)\}}{\mathbb{E}_{\nu_b}[\exp\{g(X)\}]} : g \in \mathcal{G}_w \right\}.$$

When both arguments integrate to one under ν_b, $D^{\mathrm{gen}}_{\nu_b}$ is the ordinary KL divergence between the corresponding distributions. In particular,

$$D^{\mathrm{gen}}_{\nu_b}(w\|w_\beta) = D_{\mathrm{KL}}(w\nu_b\|d_\beta).$$

Define the population approximation error

$$\varepsilon_{\mathrm{KL}} := \inf_{w\in\mathcal{W}_{\mathrm{F}}} D^{\mathrm{gen}}_{\nu_b}(w\|w_\beta).$$

Hence $\varepsilon_{\mathrm{KL}} = 0$ whenever $\log w_\beta$ belongs to $\mathcal{G}_w$ up to an additive constant.

Let

$$\mathcal{D}^w_{n_w} = \{(X_i, X_i^0, X_i^+)\}_{i=1}^{n_w}$$

denote the sample used to fit the weights, and let $\mathbb{P}^w_{n_w}$ denote its empirical average. FORE (van der Laan & Kallus, 2026) initializes $\widehat{w}^{(0)} \equiv 1$ and, for $k = 0, \ldots, K_w - 1$, computes

$$\widehat{g}_{k+1} \in \arg\min_{g\in\mathcal{G}_w}\left[\log \mathbb{P}^w_{n_w}\left\{e^{g(X)}\right\} - (1-\beta)\mathbb{P}^w_{n_w}\{g(X^0)\} - \beta\frac{\mathbb{P}^w_{n_w}\left\{\widehat{w}^{(k)}(X)g(X^+)\right\}}{\mathbb{P}^w_{n_w}\left\{\widehat{w}^{(k)}(X)\right\}}\right],$$

$$\widehat{w}^{(k+1)}(x) = \frac{\exp\{\widehat{g}_{k+1}(x)\}}{\mathbb{P}^w_{n_w}[\exp\{\widehat{g}_{k+1}(X)\}]}.$$

The X^0 term represents the source distribution $\xi_{\pi,b}$, while the weighted X^+ term represents one-step propagation of the current weighted distribution under P_π. Its denominator normalizes the current empirical weights. This gives the discounted FORE recursion with source distribution $\xi_{\pi,b}$.

G.2 Conditions and Localized Complexity

In addition to Conditions C1 and C9, FORE requires source coverage, $\xi_{\pi,b} \ll \nu_b$ and $\underline{\omega}_{\pi/b} > 0$. The following conditions specify the FORE sample and function class. Condition C5 is needed only when the estimated weights are subsequently used in FQE.

C10 FORE sample. The triples in $\mathcal{D}^w_{n_w}$ are independent, with $X_i = (S_i, A_i) \sim \nu_b$. Conditionally on X_i, $X_i^0 = (S_i, A_i^0)$, where $A_i^0 \sim \pi(\cdot \mid S_i)$, and $X_i^+ \sim P_\pi(\cdot \mid X_i)$, with the two draws independent.

C11 Bounded convex FORE class. The class $\mathcal{G}_w$ is convex, compact in the supremum norm, contains 0, and, for some $R_w < \infty$, satisfies

$$\sup_{g\in\mathcal{G}_w} |g(x) - \mathbb{E}_{\nu_b}\{g(X)\}| \le R_w \qquad \text{for } \nu_b\text{-almost every } x.$$

Convexity makes each log-ratio update a convex optimization problem, while compactness ensures that the empirical minimum is attained. The centered bound implies

$$e^{-2R_w} \le w(x) \le e^{2R_w}$$

for every $w \in \mathcal{W}_{\mathrm{F}}$ and every empirical FORE iterate. Together with the target-ratio bounds and occupancy balance equation, it also gives

$$\frac{d\{(w\nu_b)P_\pi\}}{d\nu_b} \le e^{2R_w}\frac{\overline{\kappa}_\beta}{\beta\underline{\kappa}_\beta} \qquad \nu_b\text{-almost surely.} \tag{S10}$$

These bounds verify the ratio and one-step coverage conditions used in Theorem S3.

Localized critical radius. The FORE bound involves empirical processes for log-ratio differences and weighted one-step terms. Define the centered and difference classes

$$\mathcal{G}_w^\circ := \{g - \mathbb{E}_{\nu_b} g : g \in \mathcal{G}_w\}, \qquad \mathcal{G}_w^\Delta := \{g_1 - g_2 : g_1, g_2 \in \mathcal{G}_w^\circ\},$$

and the multiplier class

$$\mathcal{M}_w := \left\{(x, x^+) \mapsto w(x) g_\Delta(x^+) : w \in \mathcal{W}_{\mathrm{F}},\ g_\Delta \in \mathcal{G}_w^\Delta\right\}.$$

Let $Q_{b,\Delta}$ denote the law of (X, X) for $X \sim \nu_b$, and let $Q_{b,\pi}$ denote the law of (X, X^+) under

$$X \sim \nu_b, \qquad X^+ \mid X \sim P_\pi(\cdot \mid X).$$

For a square-integrable class $\mathcal{A}$ under a distribution Q, define the localized Rademacher complexity

$$\mathcal{R}_{n_w}(\mathcal{A}, r; Q) := \mathbb{E}\left[\sup_{\substack{a\in\mathcal{A}:\\ \|a\|_{L^2(Q)}\le r}} \left|\frac{1}{n_w}\sum_{i=1}^{n_w}\sigma_i a(Z_i)\right|\right],$$

where $Z_i \overset{\mathrm{iid}}{\sim} Q$, and the σ_i's are independent Rademacher variables. Set

$$\mathfrak{C}_{n_w}(r) := \max\Big\{\mathcal{R}_{n_w}(\mathcal{G}_w^\Delta, r; \nu_b), \mathcal{R}_{n_w}(\mathcal{G}_w^\Delta, r; \xi_{\pi,b}), \mathcal{R}_{n_w}(\mathcal{M}_w, r; Q_{b,\Delta}), \mathcal{R}_{n_w}(\mathcal{M}_w, r; Q_{b,\pi})\Big\},$$
$$\mathfrak{r}_{n_w,\mathrm{fit}} := n_w^{-1/2} \vee \inf\left\{r > 0 : \mathfrak{C}_{n_w}(r) \le r^2\right\}.$$

Thus, $\mathfrak{r}_{n_w,\mathrm{fit}}$ controls the log-ratio difference and multiplier classes that enter the FORE finite-sample bound.

G.3 Weight-Error Bound for FORE

Theorem S3 (Weight error of the FORE estimator)**.** *Assume Conditions C1, C9, and C10–C11. Also suppose that $\xi_{\pi,b} \ll \nu_b$ and $\underline{\omega}_{\pi/b} > 0$. Let*

$$\widehat{w}_\beta := \widehat{w}^{(K_w)}, \qquad \varrho_\beta := \frac{1+\beta}{2},$$

and define

$$\varepsilon_w := \left\|\frac{\widehat{w}_\beta}{w_\beta} - 1\right\|_{2,d_\beta} = \left\{\mathbb{E}_{\nu_b}\left[\frac{(\widehat{w}_\beta - w_\beta)^2}{w_\beta}\right]\right\}^{1/2}.$$

Set

$$L_\beta := 1 + \log\left(\frac{e^{2R_w}}{\underline{\kappa}_\beta}\right).$$

There exists a finite constant $C^\circ_{\mathrm{FORE},\beta} = C(R_w, \beta, \text{coverage constants})$ such that, for every $\tau \in (0,1)$, with probability at least $1-\tau$,

$$D^{\mathrm{gen}}_{\nu_b}(\widehat{w}_\beta \| w_\beta) \le C^\circ_{\mathrm{FORE},\beta} L_\beta \left[\varrho_\beta^{K_w} D^{\mathrm{gen}}_{\nu_b}(1\|w_\beta) + \frac{\varepsilon_{\mathrm{KL}}}{(1-\beta)\underline{\kappa}_\beta} + \frac{\log^2(en_w)}{(1-\beta)^2}\left\{\mathfrak{r}^2_{n_w,\mathrm{fit}} + \frac{\log(1/\tau)}{n_w}\right\}\right].$$

Consequently, after increasing this constant if necessary,

$$\varepsilon_w \le C^\circ_{\mathrm{FORE},\beta}\left[\frac{\varrho_\beta^{K_w/2}}{\sqrt{\underline{\kappa}_\beta}}\left\{D^{\mathrm{gen}}_{\nu_b}(1\|w_\beta)\right\}^{1/2} + \frac{1}{\underline{\kappa}_\beta}\sqrt{\frac{\varepsilon_{\mathrm{KL}}}{1-\beta}} + \frac{\log(en_w)}{(1-\beta)\sqrt{\underline{\kappa}_\beta}}\left\{\mathfrak{r}_{n_w,\mathrm{fit}} + \sqrt{\frac{\log(1/\tau)}{n_w}}\right\}\right].$$

The generalized-KL inequality is the native FORE guarantee. The second inequality converts it to the relative $L^2(d_\beta)$ error used by the FQE analysis. Its three terms are, respectively, the finite-iteration, ratio-class approximation, and statistical errors.

G.4 Formal End-to-End Guarantee for FORE-Weighted FQE

The following corollary formally combines the FORE rate in (2) with Theorem 1.

Corollary S2 (End-to-end guarantee for FQE with FORE weights)**.** *Assume Conditions C1–C5, Condition C8 with $\alpha \in [0,1)$, Condition C9, and Conditions C10–C11. Let $(\widehat{w}_\beta, \widehat{Q}^{(K)})$ be the output of Algorithm 1. Then Conditions C6 and C7 hold with*

$$\kappa_{\mathrm{cov}} = \overline{\kappa}_{\mathrm{F}} := \max\{\overline{\kappa}_\beta, e^{2R_w}\}, \qquad \lambda_{\mathcal{F}} \geq \frac{e^{-2R_w}}{\overline{\kappa}_\beta}.$$

Moreover, for every $\tau \in (0,1)$, with probability at least $1-\tau$, the conclusion of Theorem 1 holds with ε_w replaced by $\varepsilon_{\mathrm{FORE}}(K_w, n_w, \tau/2)$ and $\varepsilon_{\mathrm{stat}}(K,\tau)$ replaced by $\varepsilon_{\mathrm{stat}}(K,\tau/2)$, for a constant $C = C(M, C_\infty, \alpha, \beta, \gamma, R_w, \underline{\kappa}_\beta^{-1}, \overline{\kappa}_\beta)$.

A uniform lower bound on the target-to-behavior action ratio is sufficient for the lower bound in Condition C9. Let $\pi_b(\cdot \mid s)$ denote the conditional action law under ν_b, and define $\underline{\omega}_{\pi/b} := \operatorname{ess\,inf}_{(s,a)\sim\nu_b} \pi(a \mid s)/\pi_b(a \mid s)$. If $\underline{\omega}_{\pi/b} > 0$, the occupancy recursion gives $\underline{\kappa}_\beta \geq (1-\beta)\underline{\omega}_{\pi/b}$, and therefore $\underline{\kappa}_\beta^{-1} \leq \{(1-\beta)\underline{\omega}_{\pi/b}\}^{-1}$. Thus, under uniform action overlap, the inverse lower bound grows at most linearly with the effective horizon $(1-\beta)^{-1}$.

G.5 Proof of the End-to-End Guarantee for FQE with FORE Weights

Proof of Corollary S2. Set

$$\overline{\kappa}_{\mathrm{F}} := \max\{\overline{\kappa}_\beta, e^{2R_w}\}, \qquad \lambda_{\mathcal{F}} := \frac{e^{-2R_w}}{\overline{\kappa}_\beta}.$$

All pointwise inequalities below hold ν_b-almost surely.

Step 1: Uniform weight bounds. For any log-ratio g fitted by FORE, define $g^\circ := g - \mathbb{E}_{\nu_b} g$. Condition C11 gives

$$-R_w \leq g^\circ(x) \leq R_w.$$

Because empirical normalization is invariant to additive constants, every fitted iterate can be written as

$$\widehat{w}^{(k)}(x) = \frac{e^{g^\circ(x)}}{\mathbb{P}_{n_w}^w e^{g^\circ(X)}}.$$

Both the numerator and denominator lie in $[e^{-R_w}, e^{R_w}]$, so

$$e^{-2R_w} \leq \widehat{w}^{(k)}(x) \leq e^{2R_w}.$$

Condition C9 gives

$$0 < \underline{\kappa}_\beta \leq w_\beta(x) \leq \overline{\kappa}_\beta.$$

Hence $w_\beta \leq \overline{\kappa}_{\mathrm{F}}$, $\widehat{w}_\beta \leq \overline{\kappa}_{\mathrm{F}}$, and

$$\widehat{w}_\beta \, \nu_b \leq \frac{e^{2R_w}}{\underline{\kappa}_\beta} \, d_\beta.$$

Thus Condition C6 holds.

Step 2: Restricted norm comparability. For every $h \in \mathcal{F} - \mathcal{F}$ with $\|h\|_{2,d_\beta} > 0$,

$$\frac{\mathbb{E}_{\nu_b}[\widehat{w}_\beta h^2]}{\mathbb{E}_{\nu_b}[w_\beta h^2]} \geq \frac{e^{-2R_w}\mathbb{E}_{\nu_b}[h^2]}{\overline{\kappa}_\beta \mathbb{E}_{\nu_b}[h^2]} = \lambda_{\mathcal{F}}.$$

Therefore Condition C7 holds with the stated $\lambda_{\mathcal{F}}$.

Step 3: Combine the FORE and FQE bounds. Apply Theorem S3 with failure probability $\tau/2$. On the resulting event,

$$\varepsilon_w \leq \varepsilon_{\mathrm{FORE}}(K_w, n_w, \tau/2).$$

Conditional on the FORE sample, apply Theorem 1 with failure probability $\tau/2$, taking $\kappa_{\mathrm{cov}} = \overline{\kappa}_{\mathrm{F}}$ and $\lambda_{\mathcal{F}} = e^{-2R_w}/\overline{\kappa}_\beta$. Both branches of the weight-error term are nondecreasing in ε_w, so substituting $\varepsilon_{\mathrm{FORE}}(K_w, n_w, \tau/2)$ for ε_w gives the bound in Corollary S2. Condition C5 makes the FORE and FQE samples independent, and a union bound gives probability at least $1-\tau$. □

G.6 Proof of the Weight-Error Bound for FORE

Proof. Step 1: Generalized-KL error. The adjoint Bellman equation gives $\mathsf{B}^\pi_\beta w_\beta = w_\beta$. For normalized density ratios w and v, the images $\mathsf{B}^\pi_\beta w$ and $\mathsf{B}^\pi_\beta v$ have the same initial component. Joint convexity of KL and data processing under P_π give

$$D^{\mathrm{gen}}_{\nu_b}\big(\mathsf{B}^\pi_\beta w \| \mathsf{B}^\pi_\beta v\big) \le \beta D_{\mathrm{KL}}\big((w\nu_b)P_\pi \| (v\nu_b)P_\pi\big) \le \beta D^{\mathrm{gen}}_{\nu_b}(w\|v).$$

Taking $v = w_\beta$ yields

$$D^{\mathrm{gen}}_{\nu_b}\big(\mathsf{B}^\pi_\beta w \| w_\beta\big) \le \beta D^{\mathrm{gen}}_{\nu_b}(w\|w_\beta).$$

The FORE theorem measures approximation on the log-ratio scale, while our statement uses generalized KL. Fix $w \in \mathcal{W}_{\mathrm{F}}$ and set $t = w/w_\beta$. Conditions C11 and C9 imply

$$\frac{e^{-2R_w}}{\overline{\kappa}_\beta} \le t \le \frac{e^{2R_w}}{\underline{\kappa}_\beta}.$$

Let $h(t) := t\log t - t + 1$, with $h(t)/(\log t)^2 := 1/2$ at $t = 1$. Since $h(t)/(\log t)^2 \to \infty$ as $t \to \infty$,

$$c_{\mathrm{KL}} := \inf_{t \ge e^{-2R_w}/\overline{\kappa}_\beta} \frac{h(t)}{(\log t)^2} > 0.$$

The constant c_{KL} depends only on R_w and $\overline{\kappa}_\beta$, not on $\underline{\kappa}_\beta$, and hence

$$\|\log w - \log w_\beta\|^2_{2,\nu_b} \le \frac{1}{\underline{\kappa}_\beta c_{\mathrm{KL}}} D^{\mathrm{gen}}_{\nu_b}(w\|w_\beta).$$

Taking the infimum over $w \in \mathcal{W}_{\mathrm{F}}$, the squared log-ratio approximation error in the FORE theorem is at most $\varepsilon_{\mathrm{KL}}/(\underline{\kappa}_\beta c_{\mathrm{KL}})$.

Condition C10 gives independence of the transition triples. Under Condition C11, the empirical objective is continuous in the supremum norm, compactness guarantees existence of exact minimizers and $L^2(\nu_b)$ total boundedness, and $w \le e^{2R_w}$ uniformly over $\mathcal{W}_{\mathrm{F}}$.

The target-ratio bounds and occupancy balance equation also give the one-step envelope in (S10). As inequalities between measures,

$$\nu_b \le \underline{\kappa}_\beta^{-1} d_\beta, \qquad d_\beta P_\pi \le \beta^{-1} d_\beta \le \beta^{-1}\overline{\kappa}_\beta \nu_b.$$

Therefore, for $w \in \mathcal{W}_{\mathrm{F}}$,

$$(w\nu_b)P_\pi \le e^{2R_w}\frac{\overline{\kappa}_\beta}{\beta\underline{\kappa}_\beta}\nu_b,$$

which proves (S10). Together with source coverage, these bounds verify the bounded-log-ratio, positivity, lower-tail, and one-step empirical-process conditions of Theorem 4.2 in van der Laan & Kallus (2026). Applying that theorem with source distribution $\xi_{\pi,b}$ and discount β gives

$$D^{\mathrm{gen}}_{\nu_b}(\widehat{w}_\beta\|w_\beta) \le C^\circ_{\mathrm{FORE},\beta} L_\beta \Bigg[\varrho_\beta^{K_w} D^{\mathrm{gen}}_{\nu_b}(1\|w_\beta) + \frac{\varepsilon_{\mathrm{KL}}}{(1-\beta)\underline{\kappa}_\beta} + \frac{\log^2(en_w)}{(1-\beta)^2}\left\{\mathfrak{r}^2_{n_w,\mathrm{fit}} + \frac{\log(1/\tau)}{n_w}\right\}\Bigg],$$

where

$$\varrho_\beta = \frac{1+\beta}{2}, \qquad L_\beta = 1 + \log\left(\frac{e^{2R_w}}{\underline{\kappa}_\beta}\right),$$

after absorbing c_{KL}^{-1} into $C^\circ_{\mathrm{FORE},\beta}$. The same sample can be reused across all K_w iterations because the empirical-process bound is uniform over the log-ratio and induced multiplier classes.

Step 2: Relative $L^2(d_\beta)$ error. Conditions C11 and C9 imply

$$\widehat{w}_\beta \le e^{2R_w}, \qquad w_\beta \ge \underline{\kappa}_\beta.$$

The one-sided generalized-KL-to-chi-square inequality of van der Laan & Kallus (2026), applied with $a = \widehat{w}_\beta$, $b = w_\beta$, $M = e^{2R_w}$, and $m = \underline{\kappa}_\beta$, gives

$$\varepsilon_w \le C_\chi \left\{ D^{\mathrm{gen}}_{\nu_b}(\widehat{w}_\beta \| w_\beta) \right\}^{1/2},$$

where $C_\chi < \infty$ depends only on R_w and $\underline{\kappa}_\beta$.

Write $U = e^{2R_w}/\underline{\kappa}_\beta$ and $\phi(t) = t \log t - t + 1$. The proof of the cited inequality gives

$$C_\chi = \left\{ \inf_{0 \le t \le U} \frac{\phi(t)}{(t-1)^2} \right\}^{-1/2},$$

where the ratio is defined by continuity at $t = 1$. Since this ratio is nonincreasing, $C_\chi = (U - 1)/\sqrt{\phi(U)}$. Moreover,

$$\sup_{U \ge 1} \frac{(U-1)\sqrt{1 + \log U}}{\sqrt{U \phi(U)}} < \infty,$$

and therefore

$$C_\chi \sqrt{L_\beta} \le \frac{C(R_w)}{\sqrt{\underline{\kappa}_\beta}}.$$

Taking square roots of the generalized-KL bound and using subadditivity gives

$$\varepsilon_w \le C^\circ_{\mathrm{FORE},\beta} \left[\frac{\varrho_\beta^{K_w/2}}{\sqrt{\underline{\kappa}_\beta}} \left\{ D^{\mathrm{gen}}_{\nu_b}(1 \| w_\beta) \right\}^{1/2} + \frac{1}{\underline{\kappa}_\beta} \sqrt{\frac{\varepsilon_{\mathrm{KL}}}{1-\beta}} + \frac{\log(e n_w)}{(1-\beta)\sqrt{\underline{\kappa}_\beta}} \left\{ \mathfrak{r}_{n_w,\mathrm{fit}} + \sqrt{\frac{\log(1/\tau)}{n_w}} \right\} \right],$$

after increasing $C^\circ_{\mathrm{FORE},\beta}$ if necessary. This and the generalized-KL bound prove Theorem S3. □