FINITE-SAMPLE THEORY FOR FITTED Q-ITERATION WHEN ACTIONS ARE FUNCTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Offline reinforcement learning seeks optimal decision rules from previously collected data. In some applications, a decision can be an entire function, such as a fluence map in radiation therapy or a smooth movement trajectory in robotics. In this paper we study the finite-sample theory for fitted Q-iteration (FQI) with functional actions in a discounted infinite-horizon setting. Three major difficulties arise in this setting: first, the absence of a Lebesgue probability density for functional actions complicates coverage descriptions; second, conventional coverage requirements can be restrictive; and third, the large functional action space makes greedy optimization in FQI challenging. To address these difficulties, we study smoothness-regularized policy search under a critic-relative coverage condition. This condition measures how well logged data distinguish relevant action-value differences without requiring an action density. Our main theorem gives finite-sample guarantees for learned-policy regret relative to the best value within a fixed smooth class of functional-action policies. The results allow trajectory lengths to be either bounded or growing and Q-functions to be fitted by either functional-input kernel ridge regression or adaptive functional neural networks. For a few examples, we can obtain polynomially decaying regret bounds in the number of logged transitions, up to logarithmic factors, with logarithmically many FQI iterations. Numerical experiments show gains of learned functional-action policies over constant-action policies and support our adoption of a critic-relative coverage condition.

1 INTRODUCTION

Reinforcement learning (RL) learns policies that map states to actions to maximize cumulative reward (Sutton & Barto, 2018). When new interactions are costly, one may use offline RL to learn optimal policies from fixed trajectories generated by earlier decision rules, known as behavior policies (Levine et al., 2020; Miao et al., 2022). The action space determines what a decision can express. In the RL literature, the action is typically a categorical, discrete, or continuous variable. Recently in some applications, it is required to choose an entire function as the action at each decision time, such as a radiation-intensity fluence map (Romeijn et al., 2003) or a smooth robot trajectory (Gasparetto et al., 2012). Such decision is referred to as a *functional action*, of which argument is distinct from the sequence of RL decisions. RL work for functional actions is extremely limited. Pan et al. (2018) study function-valued controls for partial differential equations in an online RL setting. For offline RL, Lin et al. (2026) integrate functional actions into fitted Q-iteration (FQI), a popular offline RL method that learns the optimal policy by repeated regression on logged transitions (Ernst et al., 2005), but they do not provide any theoretical guarantee.

In this paper, we fill this theoretical gap and study the finite-sample theory for FQI with functional actions in a discounted infinite-horizon setting. A common challenge in offline FQI theory is to transfer control of critic error from the logged state–action distribution to actions selected during policy improvement. Establishing such a transfer typically requires a coverage condition linking the behavior distribution to policy-selected actions. For example, for finite-dimensional continuous actions, Antos et al. (2007) obtain this link by combining smoothness and regular action geometry with a behavior density bounded away from zero. However, this density-based argument does not extend directly to functional actions since there is no Lebesgue probability density for random functions (Delaigle & Hall, 2010), neighborhood probabilities can decay faster than any power of

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.

their radius (Mas, 2012), and smooth policy actions may lie a positive distance from the behavior support. Consequently, it does not by itself yield the error transfer needed in our setting.

To address this issue, we introduce a coverage alternative for functional-action FQI, motivated by the idea in earlier work that coverage should be measured relative to the error class relevant to the learner. Uehara & Sun (2022) study model-based pessimistic learning and estimate an entire transition model to transfer transition-model errors, whereas our model-free FQI procedure fits critics through repeated Bellman regressions. Xie et al. (2021) formulate function-class-relative coverage through Bellman residuals in finite state and action spaces. By restricting attention to Bellman residuals relevant to the learner, their coefficient is no larger than the corresponding state–action density ratio and can be much smaller. Because its definition does not depend on action geometry, it can also be stated formally for functional actions. However, for functional-action FQI, we also seek to quantify how smoothness of the selected action affects the transfer of critic error. Therefore, we develop coverage conditions that connect this smoothness, controlled by a reproducing kernel Hilbert space (RKHS) penalty on the policy, to critic-specific information in the logged data. These conditions further yield a uniform Hölder bound from logged prediction error to critic error at policy-selected actions, even when the corresponding uniform linear error ratio is infinite.

Building on this transfer, we combine continuous-action policy-search analysis (Antos et al., 2007) with error-propagation arguments (Munos & Szepesvári, 2008) to establish a finite-sample bound on the value gap of the policy learned by functional-action FQI. The bound accounts separately for policy-class approximation error, critic error, policy-improvement error, and finite-iteration error, while allowing reuse of the same data and dependence within trajectories. We further derive convergence rates for the learned policy's value gap when the critics are fitted by either functional-input kernel ridge regression (Wang et al., 2026) or adaptive functional neural networks (Yao et al., 2021), under the respective coverage and complexity conditions, including an extension to continuous state spaces. A running two-state example verifies the conditions throughout the analysis and, with suitable tuning and logarithmically many iterations, yields a value-regret rate of $(NT)^{-2/9}$ up to logarithmic factors for N independent subjects with T transitions each, despite a positive distance between policy-selected actions and the behavior support. Finally, experiments based on the OpenAI Gym Pendulum benchmark (Brockman et al., 2016) compare FQI with functional and constant actions and examine how action proximity relates to critic-relative diagnostics.

2 Problem Setup for Functional-Action Fitted Q-Iteration

Consider a discounted, time-homogeneous Markov decision process (Puterman, 1994) over the infinite horizon $t = 0, 1, \ldots$, with discount factor $0 < \gamma < 1$, Borel state space $\mathcal{S}$, and Borel action space $\mathcal{A} \subset L^2([0,1])$. Each $a = a(\cdot) \in \mathcal{A}$ is an action function, and $a(u)$ is its value at $u \in [0,1]$. We write $P_{\mathrm{env}}(\cdot, \cdot \mid s, a)$ for the conditional distribution of the reward $R_t \in [0,1]$ and next state S_{t+1} given $S_t = s$ and $A_t = a$, and $P_S(\cdot \mid s, a)$ for its next-state marginal. We analyze measurable stationary deterministic policies, represented by maps $\pi : \mathcal{S} \to \mathcal{A}$, so that $\pi(s)$ is the entire action function selected whenever state s is visited. This choice is motivated by discounted optimality theory that a measurable selector attaining the optimal Bellman maxima defines a stationary deterministic optimal policy (Puterman, 1994, Section 6.2).

For initial-state law ρ, define the value of a given π as $J(\pi) = \mathbb{E}^{\pi}_{S_0 \sim \rho} \sum_{t \geq 0} \gamma^t R_t$, which is upper bounded by $V_{\max} = (1-\gamma)^{-1}$ as $R_t \in [0,1]$. Let $\mathcal{H}$ be a separable RKHS of measurable maps from $\mathcal{S}$ to $L^2([0,1])$, with operator-valued kernel $\Gamma : \mathcal{S} \times \mathcal{S} \to \mathcal{B}\{L^2([0,1])\}$ (Micchelli & Pontil, 2005), where $\mathcal{B}(E)$ denotes the bounded linear operators on E. We work with the feasible policy class and value benchmark

$$\Pi = \{\pi \in \mathcal{H} : \pi(s) \in \mathcal{A} \text{ for all } s\}, \qquad J^\star = \sup_{\pi \in \Pi} J(\pi). \tag{1}$$

We observe $\mathcal{D} = \{(S_{i,t}, A_{i,t}, R_{i,t}, S_{i,t+1}) : i \leq N, t \leq T_i\}$ from N independent subjects, each contributing a trajectory of length T_i. For the main analysis, without loss of generality, we assume $T_i = T = T_N$, so the total number of transitions is $n = NT$. See Appendix D.4 for handling unequal trajectory lengths. We aim to develop the finite-sample bound for $J^\star - J(\widehat{\pi}_M)$ where $\widehat{\pi}_M$ is obtained by Algorithm 1.

Algorithm 1 is an extension of standard FQI (Ernst et al., 2005) to functional actions. The key feature of Algorithm 1 is that its policy-improvement step adopts the average-Q objective for continuous

actions (Antos et al., 2007) and additionally regularizes policy smoothness through the RKHS-norm penalty $\lambda_{\pi,n}\|\pi\|_{\mathcal{H}}^2$. In particular, if the zero policy $\pi_0(s) \equiv 0$ is feasible and the approximate maximization has regularized-objective error at most $\bar{\tau}_{\pi,n}$, clipping the critic to $[0, V_{\max}]$ gives $\|\widehat{\pi}_m\|_{\mathcal{H}}^2 \le (V_{\max} + \bar{\tau}_{\pi,n})/\lambda_{\pi,n}$. Under the spectral evaluation condition in Section 3, this norm control also bounds smoothness within each selected action.

The penalty-induced bound above places each returned policy in an RKHS norm ball. Sections 3 and 4 use a fixed such ball for the coverage and value analyses. For $B < \infty$, define

$$\Pi_{\mathcal{H}}(B) = \{\pi \in \Pi : \|\pi\|_{\mathcal{H}} \le B\}.$$

Appendix D.2 gives sufficient conditions for one fixed radius $B = B_{\mathcal{H}} < \infty$ to contain every returned policy throughout the iterations.

Algorithm 1 FQI with functional actions

Require: Data $\mathcal{D}$, action-feasible policy domain Π, critic fitting rule, coefficient $\lambda_{\pi,n} > 0$, discount γ, iterations M.

1: Set $\widehat{Q}_0 = 0$; obtain $\widehat{\pi}_0$ by the update in step 5.
2: **for** $m = 1, \dots, M$ **do**
3: Set $Y_{i,t,m} = R_{i,t} + \gamma \widehat{Q}_{m-1}\{S_{i,t+1}, \widehat{\pi}_{m-1}(S_{i,t+1})\}$.
4: Regress $Y_{i,t,m}$ on $(S_{i,t}, A_{i,t})$ to obtain $\widetilde{Q}_m$; set $\widehat{Q}_m = \mathrm{clip}_{[0,V_{\max}]}(\widetilde{Q}_m)$.
5: $\widehat{\pi}_m \leftarrow \arg\max_{\pi\in\Pi} n^{-1} \sum_{i,t} \widehat{Q}_m\{S_{i,t+1}, \pi(S_{i,t+1})\} - \lambda_{\pi,n}\|\pi\|_{\mathcal{H}}^2$
6: **end for**
7: **return** $\widehat{\pi}_M$.

The best value in the fixed norm ball is $J_{\mathcal{H}}^\star = \sup_{\pi\in\Pi_{\mathcal{H}}(B_{\mathcal{H}})} J(\pi)$. Its approximation gap relative to the best feasible-policy value $J^\star$ is $a_{\mathcal{H}} = J^\star - J_{\mathcal{H}}^\star \ge 0$. Then we can decompose $J^\star - J(\widehat{\pi}_M) = a_{\mathcal{H}} + \{J_{\mathcal{H}}^\star - J(\widehat{\pi}_M)\}$. Based on the coverage condition in Section 3, we develop a bound for $J_{\mathcal{H}}^\star - J(\widehat{\pi}_M)$ in Section 4, with $a_{\mathcal{H}}$ accounting for restriction to the fixed RKHS ball.

3 A COVERAGE ALTERNATIVE FOR FUNCTIONAL ACTIONS

Coverage is a central difficulty for functional-action FQI. Without proper conditions, a small prediction error on logged data does not necessarily imply a small error at the actions selected by a learned policy. However, standard action-density and neighborhood conditions are difficult to apply to functional-action FQI because there exists no Lebesgue probability density for functional actions and actions can easily lie outside the behavior support. In this section, we circumvent this difficulty by adopting critic-relative coverage conditions that combine directional information in logged data with policy smoothness to control evaluation error.

3.1 FROM LOGGED PREDICTION TO POLICY EVALUATION

FQI analyses commonly distinguish the distribution used for critic regression from the state distribution used for policy evaluation (Antos et al., 2007; Munos & Szepesvári, 2008). Accordingly, let ν_X be the average logged state–action law and ν_+ be the average next-state law respectively:

$$\nu_X = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathcal{L}(S_{i,t}, A_{i,t}), \qquad \nu_+ = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathcal{L}(S_{i,t+1}).$$

These pooled laws may depend on $n = NT$. FQI fits critics under ν_X and evaluates them at actions selected by $\Pi_{\mathcal{H}}(B_{\mathcal{H}})$. We therefore compare logged prediction error with a policy-evaluation norm covering these actions.

In Algorithm 1, for a preceding critic Q and policy π, the Bellman label function is $y_{Q,\pi}(s,a,r,s') = r + \gamma Q\{s', \pi(s')\}$. For a fixed label function y, its population regression target is the conditional mean $Q_y(s,a) = \int y(s,a,r,s')P_{\mathrm{env}}(dr, ds' \mid s, a)$. With exact functional inputs, the total error of a critic f fitted to labels y is $f - Q_y$. Let f_y° be a structured approximant of Q_y. Let $\mathcal{V}_n$ be a deterministic family, possibly depending on the sample size n, containing all corresponding pairs of possible fitted critics and target approximants. For $v = (f, f_y^\circ) \in \mathcal{V}_n$, write $h_v = f - f_y^\circ$. We measure this critic difference by its *logged prediction error* e_v and its *policy-evaluation error* υ_v:

$$e_v = \|h_v\|_{L^2(\nu_X)}, \qquad \upsilon_v = \|h_v\|_{\nu_+,\Pi_{\mathcal{H}}(B_{\mathcal{H}}),2} := \Big[\int \sup_{\pi\in\Pi_{\mathcal{H}}(B_{\mathcal{H}})} |h_v\{s,\pi(s)\}|^2 \nu_+(ds)\Big]^{1/2}. \quad (2)$$

Both quantities measure the fitted critic's error relative to its target approximant: e_v uses the logged design, whereas υ_v uses policy-selected actions. The supremum in υ_v is taken separately at each state, so it controls a statewise error envelope. The target-approximation error $f_y^\circ - Q_y$ is controlled separately on the full state–action domain because policy actions may lie outside behavior support.

We seek conditions under which a small e_v implies a small υ_v. A uniform linear inequality $\upsilon_v \le C e_v$ can fail, so we instead allow a Hölder inequality $\upsilon_v \le C_{\mathrm{tr,n}} e_v^{\vartheta}$, $0 < \vartheta \le 1$, which still transfers consistency under suitable control of $C_{\mathrm{tr,n}}$. In particular, when $\vartheta < 1$, this requires controlling the coefficient radius of the analyzed critic differences. The mechanism combines information in the logged actions with smooth evaluation at the actions selected by the localized policy class.

3.2 Directional information and Spectral Transfer

For illustration, we first suppress state dependence and start with an action-linear critic difference $h(a) = \langle d, a\rangle$. Fix an orthonormal basis $\{\phi_j\}$ of $L^2([0,1])$ and let $\mathbf{W}_\alpha \phi_j = j^{-2\alpha}\phi_j$, $\alpha > 0$. Logged-design variation, measured through the second-moment condition $\mathbb{E}h(A)^2 \ge c\|\mathbf{W}_\alpha^{1/2} d\|^2$ for $c > 0$, controls the critic coefficient in an information-weighted norm. For $\zeta > 0$, we quantify action smoothness by $\|a\|_\zeta^2 := \sum_{j\ge 1} j^{2\zeta}|\langle a, \phi_j\rangle|^2$. The ζ-norm emphasizes higher-order spectral variations. For linear features, actions controlled under this norm reduce sensitivity to critic differences along directions weakly observed under the logged design for compatible transfer exponents satisfying $\alpha\vartheta \le \zeta$. See Appendix F for detailed explanations. The policy RKHS supplies action smoothness, and general critic features require the corresponding directional-information and smooth-evaluation conditions stated below.

Assumption 1 (Spectrally regular policy evaluation)**.** There is a measurable $\kappa_\zeta : \mathcal{S} \to [0,\infty)$ such that, on one common full-ν_+-measure set,

$$\|\pi(s)\|_\zeta \le \kappa_\zeta(s)\|\pi\|_{\mathcal{H}} \quad \text{for every } \pi \in \mathcal{H}, \quad \text{and} \quad K_\zeta^2 := \int \kappa_\zeta(s)^2 \nu_+(ds) < \infty. \tag{3}$$

Assumption 1 requires the policy-RKHS norm to control smoothness within each selected action, beyond its L^2 magnitude. It gives the fixed comparison ball the classwise spectral envelope

$$\int \sup_{\pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})} \|\pi(s)\|_\zeta^2 \nu_+(ds) \le K_\zeta^2 B_{\mathcal{H}}^2 =: B_{\Pi,\zeta}^2. \tag{4}$$

The supremum is inside the state integral, so the bound covers every policy action available at each state. The ζ describes action regularity, while $B_{\mathcal{H}}$ controls the size of the policy class. Appendix C.1 gives weighted spectral and Sobolev RKHS constructions satisfying Assumption 1.

For more general critics, the feature need not equal the action. We may associate each v with a coefficient Hilbert space $\mathcal{E}_v$, a coefficient d_v, and a measurable feature map $\psi_v : \mathcal{S}\times\mathcal{A} \to \mathcal{E}_v$ such that $h_v(s,a) = \langle d_v, \psi_v(s,a)\rangle_{\mathcal{E}_v}$. Define the feature subspace $\mathcal{E}_v^{\mathrm{feat}} := \overline{\mathrm{span}}\{\psi_v(s,a) : (s,a) \in \mathcal{S}\times\mathcal{A}\} \subseteq \mathcal{E}_v$, and let $\mathbf{U}_v$ be the orthogonal projection in $\mathcal{E}_v$ onto $\mathcal{E}_v^{\mathrm{feat}}$. The projected coefficient $\mathbf{U}_v d_v$ retains the directions that can affect the critic anywhere on the full domain, including at policy actions outside behavior support. This projection is determined by the full feature map, not the observed sample span. If the features span $\mathcal{E}_v$, then $\mathbf{U}_v$ is the identity. Let $\mathbf{W}_v : \mathcal{E}_v \to \mathcal{E}_v$ be a positive, injective, self-adjoint contraction assigning weights to coefficient directions. The information condition below compares logged variation with these weights. Appendix F.2 defines the inverse powers of $\mathbf{W}_v$.

Assumption 2 (Directional information and smooth evaluation)**.** For constants $c_n > 0$, $0 \le B_{\psi,n} < \infty$ and $0 < \vartheta \le 1$, require the following for every $v \in \mathcal{V}_n$:

(i) *Feature integrability.* $\int \|\psi_v(s,a)\|_{\mathcal{E}_v}^2 \nu_X(ds,da) < \infty$.
(ii) *Directional information.* $e_v^2 \ge c_n \langle \mathbf{U}_v d_v, \mathbf{W}_v \mathbf{U}_v d_v\rangle$.
(iii) *Smooth evaluation.* $\int \sup_{\pi\in\Pi_{\mathcal{H}}(B_{\mathcal{H}})} \|\mathbf{W}_v^{-\vartheta/2}\psi_v\{s,\pi(s)\}\|^2 \nu_+(ds) \le B_{\psi,n}^2$.

In Assumption 2, Part (i) ensures finite logged feature second moments. Part (ii) requires logged variation to distinguish the relevant critic differences, with smaller weights allowing less information in some coefficient directions. Part (iii) limits how strongly policy-selected actions can amplify critic differences along those weakly observed directions. Policy smoothness alone does not verify this

critic-relative condition for an arbitrary feature map: one must also establish how action smoothness controls the weighted feature norm. Appendices H and I give sufficient conditions for this verification in KRR and AdaFNN, respectively.

The restriction to relevant errors follows the model-relative coverage approach of Uehara & Sun (2022), while Xie et al. (2021, Definition 1) use ratios of squared Bellman-residual norms. Here the errors are differences between fitted critics and their target approximants. Coverage is uniform over this declared error family, and evaluation takes a statewise supremum over the comparison policy ball. The resulting Hölder transfer can remain useful when the uniform linear error ratio is infinite. In continuous-action FQI, Antos et al. (2007, Assumptions A2 and A6–A8) combine a positive behavior-density lower bound with regular action geometry and Lipschitz smoothness. Instead, our conditions in Assumption 2 require directional information and smooth evaluation, allowing policy actions outside behavior support. Appendix N gives a detailed comparison.

We now state the main transfer result in Proposition 1, which bounds policy-evaluation error by logged prediction error.

Proposition 1 (Spectral transfer)**.** *Under Assumption 2,* $\upsilon_v \le B_{\psi,n} c_n^{-\vartheta/2} \|d_v\|^{1-\vartheta} e_v^{\vartheta}$.

The proof of Proposition 1 is given in Appendix F. Proposition 1 ensures that a small logged prediction error leads to a small error at smooth policy actions. The exponent ϑ describes the loss of accuracy in this transfer: $\vartheta = 1$ preserves the prediction-error order, while $0 < \vartheta < 1$ gives a bound that decays more slowly. For a fixed weighting operator, stronger logged information (larger c_n) and less amplification by policy-selected features (smaller $B_{\psi,n}$) tighten the bound. When $\vartheta < 1$, the transfer bound also depends on the size of the critic-difference coefficients. Despite this slower decay, the Hölder bound can still support consistent evaluation without nearby logged actions or a finite uniform linear error ratio. Next we provide an example that admits Hölder transfer even though all actions selected by the localized policy class lie outside the support of the logged actions.

Example 1 (Two-state functional-action support gap)**.** Let $\mathcal{S} = \{1, 2\}$ and $\mathcal{A} = \{a : \|a\|_{L^2} \le B_A\}$. Take $\mathcal{G} = H^2(0,1) := \{a \in L^2(0,1) : a', a'' \in L^2(0,1)\}$ with $\|a\|_{\mathcal{G}}^2 = \|a\|_{L^2}^2 + \|a''\|_{L^2}^2$, and let $\mathcal{H} = \mathcal{G} \oplus \mathcal{G}$ with $\|\pi\|_{\mathcal{H}}^2 = \|\pi(1)\|_{\mathcal{G}}^2 + \|\pi(2)\|_{\mathcal{G}}^2$. The feasible domain remains the product action ball $\Pi = \{\pi \in \mathcal{H} : \|\pi(s)\|_{L^2} \le B_A,\ s = 1, 2\}$. For the normalized shifted Legendre basis, the Sobolev embedding in Proposition C.2 in Appendix C verifies (3) with $\zeta = 2$, and every action selected by $\Pi_{\mathcal{H}}(B_{\mathcal{H}})$ has spectral norm at most $CB_{\mathcal{H}}$. At each decision time, logged actions are independently sampled as $A = \varsigma_A \sum_{j \ge 1} j^{-9/4} \xi_j \phi_j$, where ξ_j are independent symmetric signs and ς_A is small enough for feasibility. We also set $\nu_X = \nu_+ \otimes \mathcal{L}(A)$ with ν_+ uniform on the two states and adopt the functional linear critic kernel $K\{(s,a),(s',a')\} = \mathbf{1}\{s = s'\}(1 + \langle a, a' \rangle)$. This model satisfies Assumption 2 with $\vartheta = 8/9$. The logged actions lack order-two spectral smoothness, with an infinite ζ-norm at $\zeta = 2$, and their support lies a positive L^2 distance from all actions selected by $\Pi_{\mathcal{H}}(B_{\mathcal{H}})$. For the bounded critic differences $\Delta_j(s,a) = \kappa_0 \mathbf{1}\{s = 1\}\langle a, \phi_j \rangle$ with fixed sufficiently small $\kappa_0 > 0$, we have $e_j \asymp j^{-9/4}$ and $\upsilon_j \asymp j^{-2} \asymp e_j^{8/9}$. Hence logged-design consistency still implies policy-evaluation consistency through Proposition 1, even though every uniform linear transfer coefficient is infinite. The attained $8/9$ exponent also shows that a uniformly faster transfer rate requires stronger information or smoothness conditions. See Appendix M for covariance, smoothness and support-gap verifications and Section 4 for rewards and transitions for value analysis.

3.3 Transfer at a prescribed accuracy

A large error ratio is not necessarily problematic when policy-evaluation error is already below the desired accuracy. For $\varepsilon > 0$, define amplification

$$C_n\{\varepsilon; \Pi_{\mathcal{H}}(B_{\mathcal{H}})\} = \sup_{v \in \mathcal{V}_n : \upsilon_v > \varepsilon} \upsilon_v^2 / e_v^2, \tag{5}$$

so that $\upsilon_v \le \varepsilon + \sqrt{C_n\{\varepsilon; \Pi_{\mathcal{H}}(B_{\mathcal{H}})\}}\, e_v$. In this definition, a positive number divided by zero is taken as $+\infty$ and an empty supremum is taken as zero.

For a generic logged prediction bound $\delta \ge 0$, define

$$\omega_n(\delta) = \inf_{\varepsilon > 0 :\, C_n\{\varepsilon; \Pi_{\mathcal{H}}(B_{\mathcal{H}})\} < \infty} [\varepsilon + \sqrt{C_n\{\varepsilon; \Pi_{\mathcal{H}}(B_{\mathcal{H}})\}}\, \delta]. \tag{6}$$

Thus $e_v \le \delta$ implies $\upsilon_v \le \omega_n(\delta)$. When $\omega_n(\delta) \lesssim \delta^{\vartheta_*}$ uniformly in n, write $\vartheta_* \in (0,1]$ for the resulting transfer exponent. With uniformly bounded transfer factors in Proposition 1, one may take $\vartheta_* = \vartheta$. Proposition F.1 in Appendix F gives an upper bound for $C_n\{\varepsilon; \Pi_{\mathcal{H}}(B_{\mathcal{H}})\}$ and extends the transfer result to accuracy-dependent directional information.

4 Finite-sample value guarantee

Section 3 provides a transfer control of logged prediction error to critic error at smooth policy actions. In this section we combine this transfer with policy-improvement error and state propagation to bound the regret of Algorithm 1.

4.1 Assumptions and error events

Before observing the data, we specify a clipped critic class $\mathcal{Q}_n$ and a label class $\mathcal{Y}_n$ containing every Bellman label $y_{Q,\pi}$ with $Q \in \mathcal{Q}_n$ and $\pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})$. The deterministic family $\mathcal{V}_n$ of pairs of fitted critics and target approximants is defined in Section 3.1.

Uniform error event. For each $y \in \mathcal{Y}_n$, let $\widehat{f}_y$ be the clipped output of the regression rule in step 4 of Algorithm 1, applied to labels y on $\mathcal{D}$. For the transfer analysis, take F_y to be the raw fit for KRR and the clipped fit evaluated with exact functional scores for AdaFNN. Here clip acts pointwise on critic outputs, with $\mathrm{clip}(z) = \min\{V_{\max}, \max\{0, z\}\}$, so $\widehat{f}_y - \mathrm{clip}\, F_y$ measures numerical input error. For $Q \in \mathcal{Q}_n$, let $\widehat{\pi}_Q$ be its policy update from step 5, and define its empirical and population objectives by

$$\widehat{\Phi}_Q(\pi) = \frac{1}{n}\sum_{i,t} Q\{S_{i,t+1}, \pi(S_{i,t+1})\}, \qquad \Phi_Q(\pi) = \mathbb{E}_{\nu_+} Q\{S, \pi(S)\}.$$

The regularized objective is $\widehat{\Psi}_Q(\pi) = \widehat{\Phi}_Q(\pi) - \lambda_{\pi,n}\|\pi\|_{\mathcal{H}}^2$. Let δ_n, $e_{Q,n}$, u_n, and $\tau_{\pi,n}$ be nonnegative deterministic bounds for prediction relative to the target approximant, numerical input error, objective sampling error, and policy optimization error, respectively. Define the joint event

$$\mathcal{U}_n = \left\{ \begin{array}{ll} (F_y, f_y^\circ) \in \mathcal{V}_n, & \|F_y - f_y^\circ\|_{L^2(\nu_X)} \le \delta_n, \quad \forall y \in \mathcal{Y}_n, \\ \|\widehat{f}_y - \mathrm{clip}\, F_y\|_\infty \le e_{Q,n} & \forall y \in \mathcal{Y}_n, \\ \sup\limits_{Q \in \mathcal{Q}_n,\, \pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})} & |\widehat{\Phi}_Q(\pi) - \Phi_Q(\pi)| \le u_n, \\ \sup\limits_{Q \in \mathcal{Q}_n, \pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})} & \widehat{\Psi}_Q(\pi) - \widehat{\Psi}_Q(\widehat{\pi}_Q) \le \tau_{\pi,n} \end{array} \right\}. \tag{7}$$

Each supremum norm above is taken over the full state–action domain, and $\tau_{\pi,n}$ includes representation and solver errors. For the prescribed, possibly n-dependent iteration count $M = M_n$, define the membership and localization event $\mathcal{R}_n = \bigcap_{m=0}^{M} \{\widehat{Q}_m \in \mathcal{Q}_n,\ \widehat{\pi}_m \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})\}$ and the joint analysis event $\mathcal{E}_n = \mathcal{R}_n \cap \mathcal{U}_n$. On $\mathcal{E}_n$, every realized Bellman label lies in $\mathcal{Y}_n$, and its fitted critic and target approximant satisfy $e_v \le \delta_n$ and hence $\upsilon_v \le \omega_n(\delta_n)$. The uniform bounds permit substitution of all fitted iterates despite data reuse. See the full argument in Appendix E.

For a critic Q and policy π, define the policy Bellman operator by $[\mathbb{B}^\pi Q](s,a) = \mathbb{E}[R + \gamma Q\{S', \pi(S')\} \mid s, a]$. At iteration m, define the statewise greedy defect $g_m(s) = \sup_{\pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})} \widehat{Q}_m\{s, \pi(s)\} - \widehat{Q}_m\{s, \widehat{\pi}_m(s)\}$. It measures how much critic value is lost by using the returned policy at state s. To bound the final value gap, we require uniform control of critic-update and policy-improvement errors across iterations.

Assumption 3 (One-step accuracy)**.** Let $\varepsilon_{Q,n}$ and $\varepsilon_{\pi,n}$ be bounds on critic-update error and policy-improvement error, respectively. On $\mathcal{E}_n$, we assume

$$\max_{1 \le m \le M} \|\widehat{Q}_m - \mathbb{B}^{\widehat{\pi}_{m-1}} \widehat{Q}_{m-1}\|_{\nu_+, \Pi_{\mathcal{H}}(B_{\mathcal{H}}), 2} \le \varepsilon_{Q,n}, \quad \text{and} \quad \max_{0 \le m \le M} \|g_m\|_{L^2(\nu_+)} \le \varepsilon_{\pi,n}.$$

Assumption 3 bounds the critic-update and policy-improvement errors uniformly over FQI iterations. The prediction and numerical-input bounds in (7), together with target approximation and coverage,

supply $\varepsilon_{Q,n}$ while the objective sampling and optimization bounds supply $\varepsilon_{\pi,n}$. Both conversions hold on the same $\mathcal{E}_n$. Appendices F.4 and D give the details and localization conditions.

To turn these errors under ν_+ into a value guarantee, we must also control how they propagate to the states visited by the returned policy. To describe state visitation, let $\mathcal{L}_\rho^\pi(S_t)$ be the distribution of S_t starting from $S_0 \sim \rho$ and following π under the transition kernel P_S. The discounted state occupancy is the probability law $d_\rho^\pi = (1-\gamma)\sum_{t\geq 0}\gamma^t \mathcal{L}_\rho^\pi(S_t)$. Define the pointwise Bellman envelope by $[\mathbb{B}Q](s,a) = \mathbb{E}[R + \gamma \sup_{\pi\in\Pi_\mathcal{H}(B_\mathcal{H})} Q\{S',\pi(S')\} \mid s,a]$, and let $Q^\dagger$ be its bounded fixed point. This envelope supplies an upper value comparator for every policy in the comparison class; it need not be attained by one policy in that coupled class.

Assumption 4 (State transfer)**.** For $Q, Q' \in \mathcal{Q}_n \cup \{Q^\dagger\}$, let $\bar\pi$ be any reference policy in $\Pi_\mathcal{H}(B_\mathcal{H})$. Define the nonnegative state functions $h_{Q,Q'}(s) = \sup_{\pi\in\Pi_\mathcal{H}(B_\mathcal{H})} |Q\{s,\pi(s)\} - Q'\{s,\pi(s)\}|$, $g_{Q,\bar\pi}(s) = \sup_{\pi\in\Pi_\mathcal{H}(B_\mathcal{H})} Q\{s,\pi(s)\} - Q\{s,\bar\pi(s)\}$. For a state function h, write $(P_S h)(s,a) = \int h(s')P_S(ds' \mid s,a)$. There are constants $0 < C_B, C_{\mathrm{occ}}(\pi) < \infty$ such that the following bounds hold for every $h = h_{Q,Q'}$ and every $h = g_{Q,\bar\pi}$, over all the stated choices of $Q, Q', \bar\pi$:

$$\mathbb{E}_{d_\rho^\pi} h \leq C_{\mathrm{occ}}(\pi)\|h\|_{L^2(\nu_+)}, \tag{8}$$

$$\int \sup_{\pi'\in\Pi_\mathcal{H}(B_\mathcal{H})} |P_S h\{s,\pi'(s)\}|^2 \nu_+(ds) \leq C_B^2 \|h\|^2_{L^2(\nu_+)}. \tag{9}$$

The occupancy bound (8) must hold for every possible returned policy π. It controls the average error along that policy's discounted state visits by its $L^2(\nu_+)$ size.

In Assumption 4, the function $h_{Q,Q'}$ measures the largest absolute critic difference among policy-selected actions at each state, and $g_{Q,\bar\pi}$ measures the loss from using the reference policy $\bar\pi$ relative to the statewise supremum. Taking $Q = \widehat{Q}_m$ and $\bar\pi = \widehat{\pi}_m$ recovers g_m. The propagation bound (9) limits how much one state transition can amplify either error, with factor C_B. Eq. (11) in Appendix B gives state-density conditions sufficient for both (8) and (9), and hence for Assumption 4, as in concentrability-based FQI analyses (Munos & Szepesvári, 2008; Farahmand et al., 2010).

4.2 Main value guarantee and examples

We first establish a finite-sample bound on the value gap $J^\star - J(\widehat{\pi}_M)$ and then derive convergence rates for three examples.

Theorem 1 (Finite-sample value guarantee under stable propagation)**.** *Under the data and measurability conditions above as well as Assumptions 3 and 4, suppose that C_{occ} is uniformly bounded over all possible returned policies and $\gamma C_B < 1$ by a fixed positive margin independent of N, T, and M. There exists a constant C, which is independent of N, T, and M, such that after $M \geq 1$ FQI iterations, on $\mathcal{E}_n$, we have*

$$J^\star - J(\widehat{\pi}_M) \leq a_\mathcal{H} + C\Big\{(\gamma C_B)^M + \varepsilon_{Q,n} + \varepsilon_{\pi,n}\Big\}. \tag{10}$$

Theorem 1 separates the bound of $J^\star - J(\widehat{\pi}_M)$ into four components. The approximation gap $a_\mathcal{H}$ defined in Section 2 is the value lost by restricting policies to the fixed RKHS ball. The bounds $\varepsilon_{Q,n}$ and $\varepsilon_{\pi,n}$ defined in Assumption 3 measure the accuracy of critic fitting and policy improvement. The finite-iteration error $(\gamma C_B)^M$ bounds the remaining influence of the initial critic $\widehat{Q}_0 = 0$. It decreases geometrically with the number of FQI updates: γ discounts future errors while C_B from Assumption 4 controls their propagation through one state transition. Keeping γC_B uniformly below one also prevents repeated critic and policy errors from accumulating without bound. Once the finite-iteration error falls below $\varepsilon_{Q,n} + \varepsilon_{\pi,n}$, further iterations have little effect on the guarantee. If $\Pr(\mathcal{E}_n) \to 1$, vanishing one-step errors and increasing M yield convergence in probability to the best value within the fixed smooth class. Theorem G.1 in Appendix G gives the general finite-M bound for every finite γC_B and the full proof.

To derive convergence rates in terms of $n = NT$, we further impose a geometric β-mixing condition on the complete transition records $Z_{i,t} = (S_{i,t}, A_{i,t}, R_{i,t}, S_{i,t+1})$. Let $\beta_{\mathrm{mix}}(k)$ bound their lag-k absolute-regularity coefficients uniformly over subjects and sample sizes, as defined in Appendix L.1.

Assumption 5 (Geometric mixing of transition records)**.** The N subjects are independent, and there are constants $C_{\mathrm{mix}}, c_{\mathrm{mix}} > 0$, such that $\beta_{\mathrm{mix}}(k) \le C_{\mathrm{mix}} e^{-c_{\mathrm{mix}} k}$, $1 \le k < T$.

Assumption 5 controls complete transition records uniformly over subjects and sample sizes, while permitting time-varying and subject-specific marginals. In discounted infinite-horizon RL, exponential β-mixing supports finite-sample analyses of fitted Q-iteration (Antos et al., 2007, Assumption A2) and fitted policy iteration (Antos et al., 2008, Assumption 2). Related geometric-ergodicity conditions are used for finite-time temporal-difference learning (Bhandari et al., 2021, Assumption 1) and value-function inference when trajectory length grows (Shi et al., 2022, Assumption A3(ii)).

The following three examples impose Assumption 5 and the premises of Theorem 1, with $\Pr(\mathcal{E}_n) \to 1$ and $M \ge c \log(NT)$ for a sufficiently large fixed constant c. Appendix L.5 gives the probability bounds and logarithmic tuning factors underlying the NT-based rates below. The notation $\widetilde{O}$ suppresses logarithmic factors in $n = NT$.

Example 1 (continued). In the two-state model with functional linear KRR, coverage, fixed-radius policy localization, one-step accuracy, and stable state propagation are jointly verified under the stated tuning in Appendix M. Since the fixed policy ball contains a globally optimal policy, we have $a_{\mathcal{H}} = 0$ and $J^\star - J(\widehat{\pi}_M) = \widetilde{O}_p((NT)^{-2/9})$. Thus the returned policy is value-consistent despite a positive action-support gap and an infinite uniform linear error ratio for the analyzed critic differences. See Appendix M.6 for verifications of the temporal conditions.

The following KRR and AdaFNN examples rely on the stated approximation, information, localization, and optimization conditions. We assume $\omega_n(\delta) \lesssim \delta^{\vartheta_*}$ uniformly over the critic classes, with ϑ_* defined in Section 3. Define the policy-compatibility gap by $\Delta(Q) = \mathbb{E}_{\nu_+} \sup_{\pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})} Q\{S, \pi(S)\} - \sup_{\pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})} \mathbb{E}_{\nu_+} Q\{S, \pi(S)\}$, and let $\Delta_n = \sup_{Q \in \mathcal{Q}_n} \Delta(Q)$. The gap vanishes when each covered critic has a measurable statewise maximizing policy in the comparison ball. Appendix D gives the decomposition.

Example 2 (Functional-input kernel ridge regression)**.** KRR (Wang et al., 2026) uses exact functional inputs and a bounded kernel, with Bellman targets in a fixed bounded critic-RKHS ball. Suppose that the full deterministic Bellman-label and policy-objective log covering numbers are $\widetilde{O}((NT)^{p_K})$ with $0 \le p_K < 1$, at the uniform-norm covering resolutions specified in Appendix H.3. Then we have

$$J^\star - J(\widehat{\pi}_M) \le a_{\mathcal{H}} + \widetilde{O}_p\Big((NT)^{-\vartheta_*(1-p_K)/4} + \sqrt{\Delta_n}\Big).$$

For logarithmic covering complexity $p_K = 0$ and $\Delta_n = 0$, the excess over $a_{\mathcal{H}}$ is $\widetilde{O}_p((NT)^{-\vartheta_*/4})$.

Example 3 (Adaptive functional neural networks)**.** AdaFNN learns d functional scores and an outer neural network (Yao et al., 2021). Here we focus on finitely many states and $\mathcal{H} = \bigoplus_{s \in \mathcal{S}} H^2(0,1)$. Let $s_h \ge 1$ and $s_b > 0$ be the smoothness orders of the outer Bellman-target functions and their score representers. At approximation accuracy $\varepsilon_{\mathrm{app}} > 0$, assume the entire fitted class has action-Lipschitz constants $O(\varepsilon_{\mathrm{app}}^{-p_{\mathrm{Lip}}})$, $p_{\mathrm{Lip}} \ge 0$, and define $D_{\mathrm{Ada}} = \max\{d/s_h,\ 1/s_b,\ 1 + p_{\mathrm{Lip}}/2\}$. Under the conditions in Appendices K and L.5, we have

$$J^\star - J(\widehat{\pi}_M) \le a_{\mathcal{H}} + \widetilde{O}_p\Big((NT)^{-\min\{\vartheta_*, 1/2\}/(2+D_{\mathrm{Ada}})} + \sqrt{\Delta_n}\Big).$$

For the one-score case with first-order smoothness and uniformly bounded action-Lipschitz constants, $d = s_h = s_b = 1$ and $p_{\mathrm{Lip}} = 0$. If $\vartheta_* \ge 1/2$ and $\Delta_n = 0$, the excess over $a_{\mathcal{H}}$ is $\widetilde{O}_p((NT)^{-1/6})$. Appendix L.5 extends the analysis to continuous states, giving an excess value rate over $a_{\mathcal{H}}$ of $\widetilde{O}_p((NT)^{-1/14})$ under the stated conditions.

5 Experiments: Offline Pendulum Control

We modify the Pendulum environment (Brockman et al., 2016) by replacing its scalar torque action with a bounded torque function $a : [0,1] \to [-2,2]$ applied over each decision interval. Then we run experiments in this modified environment to provide empirical motivation for two aspects of the theory: the benefit of selecting functional-action policies over constant-action policies, and assessing logged information through the critic differences that matter for policy improvement. The experiment illustrates these ideas but does not verify all conditions of the value guarantee. Appendix O gives details of implementation and evaluation.

For each sample size $n = NT \in \{2000, 4000, 8000, 16000, 32000\}$, with $T = 20$ and $\gamma = 0.95$, we separately train FQI policies selecting smooth torque functions over each decision interval and state-dependent constant torques on the same 20 independently generated datasets. We estimate each policy value using 1,000 independent Monte Carlo trajectories, separately from those used to select the smoothness penalty. Figure 1(a) shows that as the sample size increases, the mean normalized value increases from 0.587 to 0.702 for functional actions, versus 0.531 to 0.560 for constant actions, illustrating the benefit of retaining within-action shape.

To examine coverage at learned actions, for each dataset we obtain $q = NT = n$ diagnostic pairs $(S_j^{\text{ev}}, A_j^{\text{ev}})$ from an independent cohort of N behavior-policy trajectories with $T = 20$ decisions each. Because states are continuous, we compare each learned action and its paired behavior action with logged actions at the same 32 nearest training states, approximating a same-state comparison. Figure 1(b) shows larger median nearest-action L^2 distances for learned actions than for paired behavior actions, suggesting weaker finite-sample neighborhood coverage at learned actions. Figure S3 in Appendix P.1 reveals the same pattern for other neighborhood sizes.

Our critic-relative coverage condition concerns error transfer from the logged design to policy-selected actions. As a descriptive finite-sample diagnostic, we use the adjacent fitted-critic difference $h = \widehat{Q}_{20} - \widehat{Q}_{19}$ and define the critic-energy ratio $\mathsf{ER}_h = \left[q^{-1}\sum_{j=1}^{q} h\{S_j^{\text{ev}}, \widehat{\pi}_{20}(S_j^{\text{ev}})\}^2\right] / \left[n^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} h(S_{i,t}, A_{i,t})^2\right]$. The numerator and denominator average squared h at learned actions and logged pairs, respectively. In Figure 1(c), the median ratios over the 20 datasets remain near 1 across sample sizes, indicating comparable energy under both designs. This suggests that weaker local action matching does not necessarily imply poor empirical transfer for the observed critic difference.

Together, these findings motivate critic-relative coverage based on error transfer rather than action proximity alone. Figure 1 only focuses on AdaFNN. For KRR (Appendix P), the critic-energy ratio is rising over n despite the same qualitative learned-versus-behavior distance gap, which further illustrates why coverage should be assessed relative to the fitted critic class.

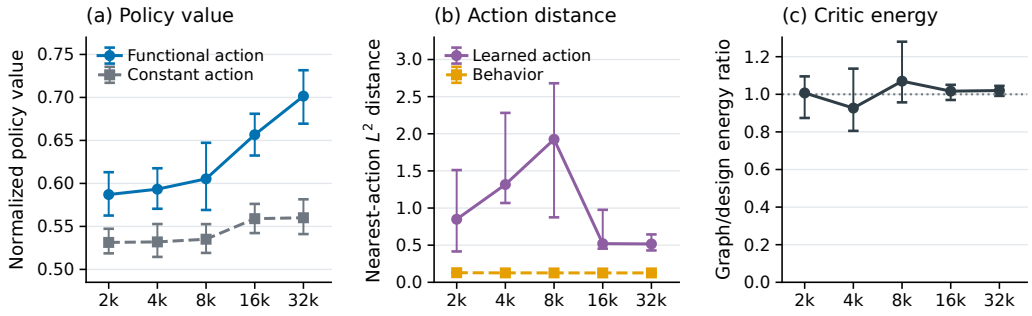

Figure 1: AdaFNN policy value and critic-relative diagnostics versus the number n of offline transitions. (a) Mean normalized policy value for functional- and constant-action policies. (b) Median nearest-action L^2 distance. (c) Median critic-energy ratio ER_h; the dotted line marks equal energy. Bars are pointwise 95% percentile-bootstrap intervals across 20 independent datasets.

6 Discussion

In this paper, we establish finite-sample value guarantees for offline functional-action FQI by combining critic-relative information, policy smoothness, and stable state propagation. Due to the unique nature of functional actions, we propose to use an alternative critic-relative coverage condition, under which we bound learned-policy regret relative to the best value within a fixed smooth policy class. Under suitable conditions for a few special examples, we can obtain a polynomially decaying bound in the number of logged transitions, up to logarithmic factors, with logarithmically many FQI iterations. Numerical experiments show gains of learned functional-action policies over constant-action policies and support our adoption of a critic-relative coverage condition. A future direction is to extend critic-relative coverage to distributional policy evaluation with functional actions.

AI USE STATEMENT

The authors used AI assistance to review mathematical arguments for potential logical gaps, search for counterexamples, check draft derivations, improve the exposition, and polish the language. The authors take full responsibility for all mathematical claims and the final content.

ETHICS STATEMENT

The experiments use a synthetic control environment and do not involve human participants or private personal data.

REPRODUCIBILITY STATEMENT

Proofs and experimental details are provided in the appendices. Code for reproducing the experiments is available at https://anonymous.4open.science/r/Functional-Fitted-Q-Learning-143E/.

REFERENCES

Philip Amortila, Audrey Huang, Akshay Krishnamurthy, and Nan Jiang. A unifying view of coverage in linear off-policy evaluation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 82623–82656, 2026.

András Antos, Rémi Munos, and Csaba Szepesvári. Fitted Q-iteration in continuous action-space MDPs. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 20, pp. 9–16, 2007.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71 (1):89–129, 2008. doi: 10.1007/s10994-007-5038-2.

H. C. P. Berbee. *Random Walks with Stationary Increments and Renewal Theory*, volume 112 of *Mathematical Centre Tracts*. Mathematisch Centrum, Amsterdam, 1979.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. *Operations Research*, 69(3):950–973, 2021. doi: 10.1287/opre.2020.2024.

Jérôme Bolte and Edouard Pauwels. Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Mathematical Programming*, 188(1):19–51, 2021.

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

Richard C. Bradley. Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*, 2:107–144, 2005.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007. doi: 10.1007/s10208-006-0196-8.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1042–1051. PMLR, 2019.

Aurore Delaigle and Peter Hall. Defining probability density for a distribution of random functions. *The Annals of Statistics*, 38(2):1171–1193, 2010.

Qiwei Di, Heyang Zhao, Jiafan He, and Quanquan Gu. Pessimistic nonlinear least-squares value iteration for offline reinforcement learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 4377–4410, 2024.

Herbert Egger and Bernd Hofmann. Tikhonov regularization in hilbert scales under conditional stability assumptions. *Inverse Problems*, 34(11):115015, 2018.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6(18):503–556, 2005.

Amir-massoud Farahmand, Rémi Munos, and Csaba Szepesvári. Error propagation for approximate policy and value iteration. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 23, pp. 568–576, 2010.

Alessandro Gasparetto, Albano Lanzutti, Renato Vidoni, and Valentina Zanotto. Experimental validation and comparative analysis of optimal time-jerk algorithms for trajectory planning. *Robotics and Computer-Integrated Manufacturing*, 28(2):164–181, 2012.

Nikolaus Hansen and Andreas Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001. doi: 10.1162/106365601750190398.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Tailen Hsing and Randall Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley & Sons, Chichester, 2015.

Jianhua Z. Huang and Ya Su. Asymptotic properties of penalized spline estimators in concave extended linear models: Rates of convergence. *The Annals of Statistics*, 49(6):3383–3407, 2021.

Xiang Ji, Minshuo Chen, Mengdi Wang, and Tuo Zhao. Sample complexity of nonparametric off-policy evaluation on low-dimensional manifolds using deep networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

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

Gefei Lin, Rui Miao, Jennifer Sacheck, and Xiaoke Zhang. Precision physical activity prescription via reinforcement learning for functional actions. *arXiv preprint arXiv:2605.19208*, 2026.

André Mas. Lower bound in regression for functional data by representation of small ball probabilities. *Electronic Journal of Statistics*, 6:1745–1778, 2012.

Rui Miao, Zhengling Qi, and Xiaoke Zhang. Off-policy evaluation for episodic partially observable markov decision processes under non-parametric models. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 593–606, 2022.

Charles A. Micchelli and Massimiliano Pontil. On learning vector-valued functions. *Neural Computation*, 17(1):177–204, 2005.

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

Thanh Nguyen-Tang, Sunil Gupta, Hung Tran-The, and Svetha Venkatesh. On sample complexity of offline reinforcement learning with deep ReLU networks in Besov spaces. *Transactions on Machine Learning Research*, 2022.

NIST. Digital Library of Mathematical Functions: Chapter 14, Legendre and related functions, 2026. URL https://dlmf.nist.gov/14. Sections 14.2 and 14.10. Online reference, accessed September 26, 2026.

Yangchen Pan, Amir-massoud Farahmand, Martha White, Saleh Nabi, Piyush Grover, and Daniel Nikovski. Reinforcement learning with function-valued action spaces for partial differential equation control. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3986–3995, 2018.

Juan C. Perdomo, Akshay Krishnamurthy, Peter Bartlett, and Sham Kakade. A complete characterization of linear estimators for offline policy evaluation. *Journal of Machine Learning Research*, 24 (284):1–50, 2023.

Erfan Pirmorad, Faraz Khoshbakhtian, Farnam Mansouri, and Amir-massoud Farahmand. Deep reinforcement learning for online control of stochastic partial differential equations. In *DLDE Workshop at NeurIPS*, 2021.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. John Wiley & Sons, New York, 1994.

James O. Ramsay and Bernard W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer, New York, 2 edition, 2005.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart J. Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 11702–11716, 2021.

H. Edwin Romeijn, Ravindra K. Ahuja, James F. Dempsey, Arvind Kumar, and Jonathan G. Li. A novel linear programming approach to fluence map optimization for intensity modulated radiation therapy treatment planning. *Physics in Medicine and Biology*, 48(21):3521–3542, 2003.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

Chengchun Shi, Sheng Zhang, Wenbin Lu, and Rui Song. Statistical inference of the value function for reinforcement learning in infinite-horizon settings. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):765–793, 2022. doi: 10.1111/rssb.12465.

Dougal J. Sutherland. Fixing an error in Caponnetto and De Vito (2007). *arXiv preprint arXiv:1702.02982*, 2017.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 2 edition, 2018.

Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1–2):1–230, 2015.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. Extended version, arXiv:2107.06226.

Jiayi Wang, Raymond K. W. Wong, Xiaoke Zhang, and Kwun Chuen Gary Chan. Flexible functional treatment effect estimation. *Journal of Machine Learning Research*, 27(16):1–48, 2026.

Ruosong Wang, Dean P. Foster, and Sham M. Kakade. What are the statistical limits of offline RL with linear function approximation? In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Christopher K. I. Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 13, pp. 682–688, 2000.

Tengyang Xie and Nan Jiang. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pp. 550–559. PMLR, 2020.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 6683–6694, 2021.

Junwen Yao, Jonas Mueller, and Jane-Ling Wang. Deep learning for functional data analysis with adaptive basis layers. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11898–11908. PMLR, 2021.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94: 103–114, 2017.

Ming Yin, Mengdi Wang, and Yu-Xiang Wang. Offline reinforcement learning with differentiable function approximation is provably efficient. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1):94–116, 1994.

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

APPENDIX

Appendices A–C give the policy domain, Bellman identities, RKHS embeddings and conforming spline approximation. Appendices D–F establish full-class policy errors, subject concentration, deterministic label covers and spectral transfer. Appendix G gives the general finite-iteration value bound and proves its stable-propagation specialization in the main text. KRR and AdaFNN fitting bounds are in Appendices H and I. Appendices J–L give implementation, quantitative approximation and temporal-dependence refinements. Appendix M verifies the same control model used throughout the main text. Experimental details begin in Appendix O.

CONTENTS

A Technical setting and algorithmic objects **16**

A.1 One RKHS policy domain 16

A.2 Numerical policy representations and the norm penalty 16

B Bellman envelopes and state transfer **17**

C RKHS regularity and conforming spline approximation **18**

C.1 Policy evaluation and functional-action spectral energy 18

C.2 An energy projection with a uniform action error 19

C.3 A compatible continuous-state implementation 20

D Policy improvement and subject concentration **21**

D.1 Policy optimization error and the actual greedy defect 21

D.2 What norm regularization supplies 22

D.3 Equal-length subjects and joint covering inputs 23

D.4 Extension: unequal and random trajectory lengths 23

E Deterministic classes and RKHS policy entropy **26**

F Smooth policy actions and moment-based secant identification **28**

F.1 Actual secants and evaluation at a prescribed accuracy 28

F.2 Directional moments and smooth functional evaluation 29

F.3 A nonlinear information verification 32

F.4 From behavior prediction to policy evaluation 33

G General finite-sample value guarantee and proof **34**

H KRR prediction and moment-based evaluation **35**

H.1 Uniform prediction and a controlled fitted norm 38

H.2 Ridge calibration and explicit prediction rates 39

H.3 An explicit KRR value rate 41

I AdaFNN prediction and exact functional secants **42**

I.1 Exact functional-coordinate secants with all bases learned 46

J Policy-norm implementation and coverage-compatible empirical total curvature **49**

K A quantitative RKHS-compatible approximation regime **51**
K.1 Continuous states with the complete fixed RKHS ball . . . 52
K.2 Retaining growing transfer factors . . . 54
K.3 Unequal-length interpretation . . . 54

L Quantitative beta-mixing rates for the full-sample algorithm **55**
L.1 Dependence and a partition that uses every transition . . . 55
L.2 All three statistical inputs on one event . . . 56
L.3 KRR uses the same dependence event, with its own ridge calibration . . . 58
L.4 Retained-block comparison and omission costs . . . 59
L.5 Dependence scales and value consequences . . . 60

M Verification of Example 1 **61**
M.1 The model, policy class, behavior support and moments . . . 62
M.2 Full linear critics and spectral transfer . . . 62
M.3 State stability and the globally optimal policy . . . 63
M.4 Full-linear ridge fitting for coverage and value . . . 63
M.5 Policy-norm self-localization and the value rate . . . 64
M.6 Temporal refinement for Example 1 . . . 65
M.7 Numerical observations without imposing smooth observed paths . . . 65

N Further related work **66**

O Experimental design and implementation **67**
O.1 Dimensionless empirical total-curvature policy objective . . . 67

P Additional AdaFNN and KRR results **68**
P.1 Empirical locality and critic differences . . . 69

A TECHNICAL SETTING AND ALGORITHMIC OBJECTS

A.1 ONE RKHS POLICY DOMAIN

We use $\mathcal{S}, \mathcal{A}, \mathcal{H}, \Pi, \Pi_{\mathcal{H}}(B_{\mathcal{H}})$, $J^\star$ and $J^\star_{\mathcal{H}}$ from Section 2. The policy RKHS has kernel $\Gamma : \mathcal{S} \times \mathcal{S} \to \mathcal{B}\{L^2([0,1])\}$ and is chosen so that its evaluation maps obey (3). Concrete weighted-spectral and Sobolev constructions are given in Appendix C.1. The scalar critic RKHS $\mathcal{F}$ is a separate space on $\mathcal{S} \times \mathcal{A}$. Every returned policy belongs to the base class Π. The event $\mathcal{R}_n$ from Section 4.1 additionally places the initialization and every returned update in the fixed comparison ball $\Pi_{\mathcal{H}}(B_{\mathcal{H}})$ and their critics in $\mathcal{Q}_n$. The value analysis uses $\mathcal{E}_n = \mathcal{R}_n \cap \mathcal{U}_n$ from Section 4.1, including localization failure in $\Pr(\mathcal{E}_n^c)$. Standalone statistical results below give sufficient events for the four bounds defining $\mathcal{U}_n$ in (7); their failure probabilities are combined with the localization failure probability to bound $\Pr(\mathcal{E}_n^c)$.

Assumption A.1 (Measurability and conditional compatibility)**.** The environment kernel $P_{\mathrm{env}}(dr, ds' \mid s, a)$ is measurable on the standard Borel spaces, has rewards in $[0,1]$, conditional mean reward $\bar{r}(s,a)$ and next-state marginal P_S. Its conditional law agrees with every observed reward–next-state pair given the current pair. Policies and critics are jointly measurable. Policy envelopes of the critics, their absolute differences, and the successive upper-Bellman iterates starting at zero are measurable. These properties extend to their uniform limits.

The common kernel makes the conditional Bellman target the same function at every observed pair, as in the time-homogeneous control model of Antos et al. (2007). Measurable policy envelopes also make the full RKHS suprema legitimate random variables; subject independence and temporal dependence enter separately through the sampling conditions. Here and below $\mathcal{L}(Z)$ denotes a probability law; norms without a subscript use the declared Hilbert or Euclidean space. The main data have fixed common length $T = T_N$, $n = NT$, and independent subjects. Observed marginals can vary with time. Unequal and random lengths are confined to Appendix D.4's extension. In that extension, conditioning on lengths must preserve subject independence, the common conditional model, and the uniform bounds used in the proof.

A useful sufficient measurability construction is a separable policy class in the topology of uniform L^2 evaluation, jointly measurable evaluation maps, and critics continuous in the action at each state. Then a countable dense policy subset gives the same pointwise suprema. It also suffices that these properties hold in an invariant, sup-norm-closed class containing zero and all upper-Bellman iterates. For example, a Feller reward–transition model on compact metric state space with uniformly continuous critic envelopes has this property. More generally, the explicit measurable-envelope assumption above is the premise. We construct the fixed point by the measurable iterates from zero, without claiming that every bounded measurable critic has a measurable supremum over an uncountable policy class.

For $V^\pi(s) = \mathbb{E}_s^\pi \sum_{t\geq 0} \gamma^t R_t$, $Q^\pi = \bar{r} + \gamma P_S V^\pi$, and $J(\pi) = \mathbb{E}_\rho V^\pi$. The smooth-class analysis first bounds $J^\star_{\mathcal{H}} - J(\widehat{\pi}_M)$; adding $a_{\mathcal{H}} = J^\star - J^\star_{\mathcal{H}}$ gives the reported target $J^\star - J(\widehat{\pi}_M)$. No attainment of the supremum defining $J^\star$ is needed.

A.2 NUMERICAL POLICY REPRESENTATIONS AND THE NORM PENALTY

Algorithm 1 optimizes over the full action-feasible domain Π. For its numerical verification, we later use a deterministic feasible subset $\mathcal{C}_n \subset \Pi$ to bound representation and solver errors; $\mathcal{C}_n$ is not part of the theoretical algorithm and need not impose an RKHS-radius constraint. For the B-spline vector $\mathbf{b}_n^{\mathrm{spl}} = (b_{1,n}^{\mathrm{spl}}, \ldots, b_{p_n,n}^{\mathrm{spl}})^\top$ and measurable coefficient maps $c_\theta(s)$, its policies have the form $\pi_\theta(s)(u) = \mathbf{b}_n^{\mathrm{spl}}(u)^\top c_\theta(s)$. Both functional-action regularity and membership in the function-valued policy RKHS must be checked. Appendix C supplies such constructions; membership is not inferred from a finite parameter count.

Write

$$\widehat{\Phi}_Q(\pi) = n^{-1} \sum_{i,t} Q\{S_{i,t+1}, \pi(S_{i,t+1})\}, \qquad \Phi_Q(\pi) = \mathbb{E}_{\nu_+} Q\{S, \pi(S)\},$$

$$\widehat{\Psi}_Q(\pi) = \widehat{\Phi}_Q(\pi) - \lambda_{\pi,n} \|\pi\|^2_{\mathcal{H}}.$$

The RKHS-norm penalty $\lambda_{\pi,n}\|\pi\|_{\mathcal{H}}^2$ is deterministic and is the only regularization used in the theoretical policy update; hence its policy proof requires no empirical-to-population comparison for this penalty. Uniform numerical discrepancies are added to the full-class objective gap in Appendix D. The reported finite-dimensional experiment replaces direct computation of the complete norm by the empirical total curvature documented in Appendix O.1; it does not add that criterion to the RKHS-norm objective. Appendix J proves that this surrogate is coverage-compatible because its localized spline classes satisfy the spectral envelope used by the transfer argument. Algorithm 1 reuses all n transitions and returns the last policy. Clipping and $1 + \gamma V_{\max} = V_{\max}$ keep labels and conditional targets in $[0, V_{\max}]$.

B Bellman envelopes and state transfer

For the fixed smooth policy class $\Pi_{\mathcal{H}}(B_{\mathcal{H}})$, define

$$(\mathbb{B}^\pi Q)(s,a) = \bar{r}(s,a) + \gamma \int Q\{s', \pi(s')\} P_S(ds' \mid s,a), \quad \mathbb{B}Q = \bar{r} + \gamma P_S[\sup_{\pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})} Q(\cdot, \pi(\cdot))].$$

The policy-evaluation norm is (2). It obeys the triangle inequality because the pointwise policy supremum does, followed by Minkowski's inequality. It is generally a seminorm: a zero value need not identify functions outside the state–action pairs induced by $\Pi_{\mathcal{H}}(B_{\mathcal{H}})$. For $\pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})$ write $g_{Q,\pi}(s) = \sup_{\pi' \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})} Q\{s, \pi'(s)\} - Q\{s, \pi(s)\}$.

Lemma B.1 (Upper fixed point and occupancy identity)**.** *There is a bounded measurable fixed point $Q^\dagger \in [0, V_{\max}]$ of $\mathbb{B}$, unique among bounded functions with measurable policy envelopes. It satisfies $Q^\pi \le Q^\dagger$ for every $\pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})$. With $V^\dagger(s) = \sup_{\pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})} Q^\dagger\{s, \pi(s)\}$ and $d_\rho^\pi = (1-\gamma)\sum_{t \ge 0} \gamma^t \mathcal{L}_\rho^\pi(S_t)$,*

$$J_{\mathcal{H}}^\star \le \mathbb{E}_\rho V^\dagger, \qquad \mathbb{E}_\rho V^\dagger - J(\pi) = \frac{1}{1-\gamma} \mathbb{E}_{d_\rho^\pi}[V^\dagger(S) - Q^\dagger\{S, \pi(S)\}].$$

The Bellman contraction and occupancy expansion are the error-propagation tools used in Munos & Szepesvári (2008); Farahmand et al. (2010). Here the envelope may choose different members of a coupled RKHS class at different states, so its fixed point provides an upper value comparator; the actual returned policy is compared to it through its nonnegative greedy defect.

Proof. For admissible measurable envelopes, $|\mathbb{B}Q - \mathbb{B}Q'| \le \gamma P_S \sup_{\pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})} |(Q - Q')(\cdot, \pi(\cdot))| \le \gamma \|Q - Q'\|_\infty$. The iterates of zero stay in $[0, V_{\max}]$ and are uniformly Cauchy. Their measurable limit and the uniform limit of their envelopes give the fixed point; the same inequality proves uniqueness. Monotonicity implies $(\mathbb{B}^\pi)^k Q^\dagger \le Q^\dagger$, and the left side converges uniformly to Q^π. Consequently $J(\pi) \le \mathbb{E}_\rho V^\dagger$ for every $\pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})$, and taking the supremum proves $J_{\mathcal{H}}^\star \le \mathbb{E}_\rho V^\dagger$ directly.

Let $P_S^\pi f(s) = P_S f\{s, \pi(s)\}$ and $g_\pi = V^\dagger - Q^\dagger(\cdot, \pi(\cdot)) \ge 0$. Subtracting the policy Bellman equation gives $V^\dagger - V^\pi = g_\pi + \gamma P_S^\pi (V^\dagger - V^\pi)$. Iterate k times. The remainder has sup norm at most $\gamma^k V_{\max}$ and tends to zero. Integrating the nonnegative sum against ρ and using monotone convergence proves the identity. □

Assumption B.1 (Residual families for state transfer)**.** Let $\mathcal{Q}_n$ be the deterministic fitted class and include $Q^\dagger$ in the following residual families:

$$h(s) = \sup_{\pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})} |Q\{s, \pi(s)\} - Q'\{s, \pi(s)\}|, \quad Q, Q' \in \mathcal{Q}_n \cup \{Q^\dagger\},$$
$$h(s) = g_{Q,\pi_0}(s), \qquad\qquad Q \in \mathcal{Q}_n \cup \{Q^\dagger\}, \quad \pi_0 \in \Pi_{\mathcal{H}}(B_{\mathcal{H}}).$$

For every possible output π, $\mathbb{E}_{d_\rho^\pi} h \le C_{\mathrm{occ}}(\pi) \|h\|_{L^2(\nu_+)}$. For all residuals in these families, $\|P_S h\|_{\nu_+, \Pi_{\mathcal{H}}(B_{\mathcal{H}}), 2} \le C_B \|h\|_{L^2(\nu_+)}$.

The future-state concentrability of Munos & Szepesvári (2008) and the L^2 density norms of Farahmand et al. (2010) motivate the two roles below. We require the inequalities on the displayed state

residuals, separating final-policy visitation from one-step propagation of the uniform policy-evaluation envelope. These are the families used in Assumption 4. Sufficient conditions are

$$\left\| \frac{dd_\rho^\pi}{d\nu_+} \right\|_{L^2(\nu_+)} \le C_{\mathrm{occ}}(\pi), \qquad \sup_{s,\pi} \left\| \frac{dP(\cdot \mid s, \pi(s))}{d\nu_+} \right\|_{L^2(\nu_+)} \le C_B, \tag{11}$$

with absolute continuity of all displayed state laws. The density-ratio bound for d_ρ^π in (11) follows by change of measure and Cauchy–Schwarz. Applying Cauchy–Schwarz at each $(s, \pi(s))$, taking the policy supremum, and integrating over ν_+ proves the transition-density bound in the same display. These state-side sufficient conditions connect with Munos & Szepesvári (2008); Farahmand et al. (2010); Chen & Jiang (2019); Xie & Jiang (2020); Rashidinejad et al. (2021); Zhan et al. (2022). The fixed-point contraction in sup norm uses only $\gamma < 1$. The additional condition $\gamma C_B < 1$ is used for stable propagation in the policy-evaluation norm.

C RKHS REGULARITY AND CONFORMING SPLINE APPROXIMATION

C.1 POLICY EVALUATION AND FUNCTIONAL-ACTION SPECTRAL ENERGY

Condition (3) makes the policy geometry itself compatible with the stronger action norm used in coverage transfer. For every $\pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})$,

$$\|\pi(s)\|_\zeta^2 \le \kappa_\zeta(s)^2 B_{\mathcal{H}}^2.$$

Taking the classwise supremum and integrating proves (4). The following constructions show that this requirement can be built directly into an L^2-valued policy RKHS.

Proposition C.1 (Policy RKHS constructions with spectral evaluation)**.** *Let $\phi_j(u) = \sqrt{2j-1}\,\mathrm{Leg}_{j-1}(2u-1)$ be normalized shifted Legendre polynomials, where Leg_k denotes the degree-k Legendre polynomial. Let $\mathcal{H}_S$ be a scalar RKHS on $\mathcal{S}$ with kernel k_S and suppose $\int k_S(s,s)\nu_+(ds) < \infty$. The L^2-valued space of maps*

$$\pi(s) = \sum_{j\ge1} f_j(s)\phi_j, \qquad \|\pi\|_{\mathcal{H}}^2 = \sum_{j\ge1} j^{2\zeta}\|f_j\|_{\mathcal{H}_S}^2$$

is a separable vector-valued RKHS with operator-valued kernel

$$\Gamma_\zeta(s,t) = k_S(s,t)T_\zeta, \qquad T_\zeta\phi_j = j^{-2\zeta}\phi_j.$$

It satisfies (3) with $\kappa_\zeta(s) = k_S(s,s)^{1/2}$.

Alternatively, let $\mathcal{G} = H^2(0,1)$ with $\langle a,b\rangle_{\mathcal{G}} = \langle a,b\rangle_{L^2} + \langle a'',b''\rangle_{L^2}$. Then $\mathcal{G}$ is an RKHS and $\|a\|_\zeta^2 \le c_{\mathcal{G}}\|a\|_{\mathcal{G}}^2$ for $0 < \zeta \le 2$. Consequently, any $\mathcal{G}$-valued policy RKHS, viewed through the continuous embedding $\mathcal{G} \hookrightarrow L^2$, that satisfies $\|\pi(s)\|_{\mathcal{G}} \le \kappa_{\mathcal{G}}(s)\|\pi\|_{\mathcal{H}}$ and $\kappa_{\mathcal{G}} \in L^2(\nu_+)$ verifies (3) with $\kappa_\zeta = c_{\mathcal{G}}^{1/2}\kappa_{\mathcal{G}}$.

The first construction encodes the coverage norm directly in the operator-valued kernel. The second uses a Sobolev action component, which is convenient for spline computation and for the running example. In both cases the deterministic policy norm, rather than a second population penalty, supplies the spectral radius.

Proof. For the first construction, scalar RKHS evaluation gives

$$\|\pi(s)\|_\zeta^2 = \sum_j j^{2\zeta}|f_j(s)|^2 \le k_S(s,s)\sum_j j^{2\zeta}\|f_j\|_{\mathcal{H}_S}^2.$$

The coefficient-space construction is complete and evaluation into L^2 is bounded because the same display with the weights removed is no larger. Its reproducing operator is Γ_ζ: expanding both arguments in the ϕ_j basis gives the reproducing identity coordinate by coordinate. Integration of the displayed inequality proves the required square-integrability.

For the Sobolev construction, the elementary decomposition $a(u) = a(0) + ua'(0) + \int_0^u (u-t)a''(t)\,dt$ bounds $\|a'\|_{L^2}$ by $C(\|a\|_{L^2} + \|a''\|_{L^2})$: the integral remainder has norm at most $\|a''\|_{L^2}/\sqrt{12}$, and

the norm of a linear polynomial controls its slope. The resulting norm is equivalent to the usual complete H^2 norm. Also $|a(u)| \le \|a\|_{L^2} + \|a'\|_{L^2}$ up to a universal constant, uniformly in u. Evaluation is bounded, so its Riesz representers define a reproducing kernel. The spectral embedding is proved next. Composing it with the assumed $\mathcal{G}$-valued evaluation inequality gives the stated κ_ζ. □

Proposition C.2 (Sobolev control of a common functional spectral norm)**.** *Let* Leg_k *be the degree-*k *Legendre polynomial and* $\phi_j(u) = \sqrt{2j-1}\,\mathrm{Leg}_{j-1}(2u-1)$, $j \ge 1$. *For every* $g \in H^2(0,1)$ *and* $0 < \zeta \le 2$,

$$\sum_{j\ge 1} j^{2\zeta}\langle g, \phi_j\rangle^2 \le C_S\{\|g\|_{L^2}^2 + \|g''\|_{L^2}^2\}. \tag{12}$$

No endpoint condition is imposed on g*: neither* g *nor* g' *is required to vanish at* 0 *or* 1. *Consequently, the spectral bound* (12) *applies to cubic spline actions with the continuity needed for a weak second derivative. The spectral basis is an analytical representation, not a replacement of the policy's B-spline basis.*

The shifted Legendre equation (NIST, 2026, Section 14.2) supplies the spectral identity used below. The estimate verifies that an H^2 component of the policy-RKHS norm controls the weighted feature norm in the transfer assumption, while leaving the endpoint values of the action function unconstrained.

Proof. Let $w(u) = u(1-u)$ and $\mathrm{LegOp}\, g = -(wg')'$. The shifted Legendre equation gives $\mathrm{LegOp}\,\phi_j = \varpi_j\phi_j$, with $\varpi_j = (j-1)j$ (NIST, 2026, Section 14.2). For smooth g, integration by parts twice gives

$$\langle \mathrm{LegOp}\, g, \phi_j\rangle = [-wg'\phi_j + wg\phi_j']_0^1 - \int_0^1 g(w\phi_j')'\,du = \varpi_j\langle g, \phi_j\rangle.$$

Both boundary terms vanish because $w(0) = w(1) = 0$, not because of a boundary condition on g. Since $j^{2\zeta} \le j^4 \le 4(1+\varpi_j^2)$, Parseval's identity yields

$$\sum_j j^{2\zeta}\langle g, \phi_j\rangle^2 \le 4\|g\|_{L^2}^2 + 4\|\,\mathrm{LegOp}\, g\|_{L^2}^2.$$

To bound $\mathrm{LegOp}\, g$, write $g(u) = g(0) + ug'(0) + v(u)$, where $v(u) = \int_0^u (u-t)g''(t)\,dt$. Cauchy–Schwarz gives $\|v\|_{L^2} \le \|g''\|_{L^2}/\sqrt{12}$. Also,

$$\begin{aligned}
\|a + bu\|_{L^2}^2 &= (a + b/2)^2 + b^2/12,\\
|g'(0)| &\le \sqrt{12}\|g - v\|_{L^2} \le \sqrt{12}\|g\|_{L^2} + \|g''\|_{L^2},\\
\left\|\int_0^{\cdot} g''(t)\,dt\right\|_{L^2} &\le \|g''\|_{L^2}/\sqrt{2}.
\end{aligned}$$

Consequently $\|g'\|_{L^2} \le \sqrt{12}\|g\|_{L^2} + (1 + 1/\sqrt{2})\|g''\|_{L^2}$. Using $\|w\|_\infty = 1/4$ and $\|w'\|_\infty = 1$ gives

$$\begin{aligned}
\|\,\mathrm{LegOp}\, g\|_{L^2} &\le \sqrt{12}\|g\|_{L^2} + 2\|g''\|_{L^2},\\
\sum_j j^{2\zeta}\langle g, \phi_j\rangle^2 &\le 100\|g\|_{L^2}^2 + 32\|g''\|_{L^2}^2 \le 200(\|g\|_{L^2}^2 + \|g''\|_{L^2}^2).
\end{aligned}$$

For general $g \in H^2(0,1)$, choose smooth $g_k \to g$ in H^2. Apply the bound to each finite coordinate sum, pass to $k \to \infty$ by continuity of the coordinates, and then let the number of coordinates increase by monotone convergence. This proves the same inequality for g. □

C.2 An energy projection with a uniform action error

Let a quasi-uniform partition have d_u intervals and maximum width $h \asymp d_u^{-1}$. Use the space of piecewise cubic, globally C^1 splines. This is an H^2-conforming B-spline space with double interior knots. Write $\mathbf{U}_{d_u}$ for its $\mathcal{G}$-orthogonal projection. The use of an energy projection is part of this verified numerical construction.

Proposition C.3 (Feasible functional-action and objective approximation). *For every $a \in H^2(0,1)$,*

$$\|\mathbf{U}_{d_u} a\|_{\mathcal{G}} \le \|a\|_{\mathcal{G}}, \qquad \|a - \mathbf{U}_{d_u} a\|_{L^2} \le C d_u^{-2} \|a\|_{\mathcal{G}}.$$

For $B_A > 0$ and $\|a\|_{L^2} \le B_A$, define $\widetilde{a} = B_A \mathbf{U}_{d_u} a / \max\{B_A, \|\mathbf{U}_{d_u} a\|_{L^2}\}$. Then $\|\widetilde{a}\|_{L^2} \le B_A$, $\|\widetilde{a}\|_{\mathcal{G}} \le \|a\|_{\mathcal{G}}$ and

$$\|a - \widetilde{a}\|_{L^2} \le 2C d_u^{-2} \|a\|_{\mathcal{G}}.$$

If $Q(s,\cdot)$ is Lip_Q-Lipschitz, replacing a by $\widetilde{a}$ in $Q(s,a) - \lambda\|a\|_{\mathcal{G}}^2$ incurs loss at most $\mathrm{Lip}_Q \|a - \widetilde{a}\|_{L^2}$.

Spline approximation with derivative-penalty control is developed, for example, by Huang & Su (2021, Proposition 2.1). The energy projection here additionally preserves the fixed action-RKHS budget, and the radial adjustment preserves action feasibility; the final inequality converts these properties into a policy optimization error.

Proof. Orthogonal projection is a contraction. Put $e = a - \mathbf{U}_{d_u} a$. Riesz representation gives $z \in H^2$ satisfying $\langle z, w\rangle_{\mathcal{G}} = \langle e, w\rangle_{L^2}$ for every $w \in H^2$. Testing with z gives $\|z\|_{\mathcal{G}} \le \|e\|_{L^2}$. For compactly supported tests the weak equation is $z'''' + z = e$. A fourfold primitive of $e - z \in L^2$ is in H^4; its difference from z has fourth distributional derivative zero and is a polynomial of degree at most three. The coefficients of that polynomial are controlled by its L^2 norm. It follows that $\|z\|_{H^4} \le C(\|z\|_{L^2} + \|e - z\|_{L^2}) \le C\|e\|_{L^2}$. Integration by parts in the weak identity gives the natural conditions $z''(0) = z''(1) = z'''(0) = z'''(1) = 0$. These conditions apply to the dual solution, not to the action a.

For clarity, the required spline approximation can be obtained by cubic Hermite interpolation of z and z' at the mesh nodes. On the unit reference interval subtract its third-order Taylor polynomial. The remainder and its first three derivatives are integrals of $z^{(4)}$, so their sup norms are bounded by a constant times $\|z^{(4)}\|_{L^2}$. The Hermite interpolant of this remainder has coefficients bounded by the same endpoint trace bounds. Write $\mathbf{J}_h z$ for the resulting piecewise Hermite interpolant. Scaling to an interval I of width h_I gives, for $j = 0, 1, 2$,

$$\|(z - \mathbf{J}_h z)^{(j)}\|_{L^2(I)} \le C h_I^{4-j} \|z^{(4)}\|_{L^2(I)}.$$

The pieces share value and first derivative at nodes and thus belong to the conforming cubic space. Summing squared bounds gives $\|z - \mathbf{J}_h z\|_{\mathcal{G}} \le C h^2 \|z\|_{H^4}$. Projection orthogonality now yields

$$\|e\|_{L^2}^2 = \langle e, z - \mathbf{U}_{d_u} z\rangle_{\mathcal{G}} \le C h^2 \|e\|_{\mathcal{G}} \|e\|_{L^2} \le C h^2 \|a\|_{\mathcal{G}} \|e\|_{L^2}.$$

Division when $e \neq 0$ proves the error estimate; the zero case is immediate.

Radial scaling is a contraction in $\mathcal{G}$ and changes the projected action function by $(\|\mathbf{U}_{d_u} a\|_{L^2} - B_A)_+ \le \|\mathbf{U}_{d_u} a - a\|_{L^2}$. The triangle inequality proves the factor two. The RKHS-norm penalty cannot increase because $\|\widetilde{a}\|_{\mathcal{G}} \le \|a\|_{\mathcal{G}}$; action Lipschitz continuity therefore proves the last assertion. □

For the finite-state product class of Example 1, apply this construction at each state. The resulting policy belongs to $\mathcal{C}_n \subset \Pi$. If $T_{\mathcal{G}} : L^2 \to \mathcal{G} \subset L^2$ is defined by $\langle T_{\mathcal{G}} a, b\rangle_{\mathcal{G}} = \langle a, b\rangle_{L^2}$, its L^2-valued operator kernel is $\Gamma(s,t) = \mathbf{1}\{s = t\} T_{\mathcal{G}}$. Uniform averaging of the pointwise objective bound controls the full-class-to-spline empirical objective gap, with no value-level approximation term.

C.3 A COMPATIBLE CONTINUOUS-STATE IMPLEMENTATION

Here is an independent implementation verification on $\mathcal{S} = [0,1]$, not another control model. Let $\mathcal{H} = H^1([0,1]; \mathcal{G})$ with norm

$$\|\pi\|_{\mathcal{H}}^2 = \int_0^1 \{\|\pi(s)\|_{\mathcal{G}}^2 + \|\partial_s \pi(s)\|_{\mathcal{G}}^2\}\, ds.$$

Bounded Hilbert-valued evaluation makes this a $\mathcal{G}$-valued RKHS. For the fixed joint ball with $\|\pi(s)\|_{L^2} \le B_A$, take the tensor product of an H^1 energy projection onto continuous piecewise linear state functions and the functional-action projection $\mathbf{U}_{d_u}$. Both projections are contractions in $\mathcal{H}$.

If the state mesh width is h_s, its uniform evaluation error is at most $Ch_s^{1/2}\|\pi\|_{\mathcal{H}}$ in $\mathcal{G}$. To verify this, the dual equation for the state energy projection is $z - z'' = e$ with natural Neumann boundary conditions; a twofold-primitive argument gives $\|z\|_{H^2(\mathcal{G})} \le C\|e\|_{L^2(\mathcal{G})}$. Piecewise linear interpolation and energy orthogonality give $\|e\|_{L^2(\mathcal{G})} \le Ch_s\|\pi\|_{\mathcal{H}}$. The one-dimensional interpolation inequality $\|e\|^2_{\infty,\mathcal{G}} \le C\|e\|_{L^2(\mathcal{G})}\|e\|_{H^1(\mathcal{G})}$ then gives the claim. Indeed, choose $t_0 \in [0,1]$ with $\|e(t_0)\|_{\mathcal{G}} \le \|e\|_{L^2(\mathcal{G})}$. Absolute continuity of $\|e(\cdot)\|^2_{\mathcal{G}}$ and Cauchy–Schwarz give

$$\|e\|^2_{\infty,\mathcal{G}} \le \|e\|^2_{L^2(\mathcal{G})} + 2\|e\|_{L^2(\mathcal{G})}\|e'\|_{L^2(\mathcal{G})} \le 3\|e\|_{L^2(\mathcal{G})}\|e\|_{H^1(\mathcal{G})}.$$

Applying Proposition C.3 after the state projection yields uniform L^2 action error $CB_{\mathcal{H}}(h_s^{1/2} + d_u^{-2})$ on $\Pi_{\mathcal{H}}(B_{\mathcal{H}})$.

A single scalar shrinkage by $B_A/\max\{B_A, \sup_s \|\pi_{\mathrm{proj}}(s)\|_{L^2}\}$ preserves the tensor spline space and the $\mathcal{H}$ budget and at most doubles this error. Because both projections and the final shrinkage contract the $\mathcal{H}$ norm, the deterministic RKHS-norm penalty cannot increase. Thus the full regularized empirical representation loss is bounded by

$$C\,\mathrm{Lip}_Q B_{\mathcal{H}}(h_s^{1/2} + d_u^{-2}).$$

This verifies membership, action feasibility and an objective bound for a continuous-state implementation. The coupled class still retains $\Delta(Q)$. Discontinuous state-cell policies are not used as approximants in this $\mathcal{H}$.

D Policy improvement and subject concentration

The approximation gap $a_{\mathcal{H}}$ concerns restriction to the fixed policy ball, whereas Δ_n concerns compatibility of statewise maximization with a single policy in that ball. Policy localization separately requires all returned policies to remain in the ball on $\mathcal{E}_n$. Example 1 establishes localization with $a_{\mathcal{H}} = 0$ and controls the greedy defect directly, without requiring $\Delta_n = 0$. A positive penalty coefficient $\lambda_{\pi,n}$ alone does not establish a sample-size-independent radius. Appendix D.2 gives a sufficient low-norm comparator calibration.

D.1 Policy optimization error and the actual greedy defect

For $Q : \mathcal{S} \times \mathcal{A} \to [0, V_{\max}]$, define

$$\Delta(Q) = \mathbb{E}_{\nu_+} \sup_{\pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})} Q\{S, \pi(S)\} - \sup_{\pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})} \Phi_Q(\pi),$$
$$g_Q^{\mathrm{sel}}(s) = \sup_{\pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})} Q\{s, \pi(s)\} - Q\{s, \widehat{\pi}_Q(s)\}.$$

A bound on the policy optimization error is $\tau_{\pi,n} \ge \sup_Q[\sup_{\pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})} \widehat{\Psi}_Q(\pi) - \widehat{\Psi}_Q(\widehat{\pi}_Q)]$. The theory uses this policy optimization error directly. For a later numerical verification based on $\widehat{\pi}_Q \in \mathcal{C}_n \subset \Pi$, it can be bounded by

$$\sup_{\pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})} \widehat{\Psi}_Q(\pi) - \widehat{\Psi}_Q(\widehat{\pi}_Q) \le \left[\sup_{\pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})} \widehat{\Psi}_Q(\pi) - \sup_{\pi \in \mathcal{C}_n} \widehat{\Psi}_Q(\pi)\right]_+ + \{\sup_{\pi \in \mathcal{C}_n} \widehat{\Psi}_Q(\pi) - \widehat{\Psi}_Q(\widehat{\pi}_Q)\}. \tag{13}$$

The brackets are numerical representation and solver errors. If the computed objective differs uniformly from $\widehat{\Psi}_Q$ by at most ϵ_{num}, add $2\epsilon_{\mathrm{num}}$ to its solver bound.

Lemma D.1 (Objective-to-defect conversion)**.** *Suppose every returned policy is in $\Pi_{\mathcal{H}}(B_{\mathcal{H}})$, $\sup_{Q,\pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})} |\widehat{\Phi}_Q(\pi) - \Phi_Q(\pi)| \le u_n$, and the policy optimization error is at most $\tau_{\pi,n}$. Then, for every covered Q,*

$$\mathbb{E}_{\nu_+} g_Q^{\mathrm{sel}} \le \Delta(Q) + 2u_n + \tau_{\pi,n} + \lambda_{\pi,n} B_{\mathcal{H}}^2, \quad \|g_Q^{\mathrm{sel}}\|^2_{L^2(\nu_+)} \le V_{\max}\{\Delta(Q) + 2u_n + \tau_{\pi,n} + \lambda_{\pi,n} B_{\mathcal{H}}^2\}.$$

For the deterministic fitted-critic class $\mathcal{Q}_n$, let u_n bound uniform sample-to-population objective deviation and let $\tau_{\pi,n}$ bound the policy optimization error, including the representation and numerical errors described above. A uniform policy-improvement error bound is

$$\varepsilon_{\pi,n}^2 = V_{\max}\Big\{ \underbrace{\sup_{Q\in\mathcal{Q}_n} \Delta(Q)}_{\text{policy-compatibility gap}} + \underbrace{2u_n + \tau_{\pi,n}}_{\text{sampling and computation}} + \underbrace{\lambda_{\pi,n} B_{\mathcal{H}}^2}_{\text{RKHS-norm penalty}} \Big\}. \tag{14}$$

The algorithm optimizes one average-objective policy, whereas the greedy defect compares actions separately at each state. The policy-compatibility gap measures this difference. The square root converts the average-objective bound to the required $L^2(\nu_+)$ greedy-defect bound. The policy-compatibility gap plays the role of the approximate-greedy error in Antos et al. (2007, Section 3.2.3). Here $\Delta(Q)$ records the compatibility of statewise choices with one RKHS policy, while the remaining terms quantify estimation, numerical representation, optimization and norm regularization within that same policy class.

Proof. For every $\pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})$, empirical near-optimality gives

$$\Phi_Q(\pi) - \Phi_Q(\widehat{\pi}_Q) \le 2u_n + \tau_{\pi,n} + \lambda_{\pi,n}\{\|\pi\|_{\mathcal{H}}^2 - \|\widehat{\pi}_Q\|_{\mathcal{H}}^2\} \le 2u_n + \tau_{\pi,n} + \lambda_{\pi,n} B_{\mathcal{H}}^2.$$

Take the supremum over π and add $\Delta(Q)$. The displayed $L^2(\nu_+)$ bound follows from $0 \le g_Q^{\mathrm{sel}} \le V_{\max}$. This proves (14), including initialization, whenever the stated bounds and policy localization hold. In the value analysis, these conditions are verified on $\mathcal{E}_n$. □

If the class is a product of compact action sets over finitely many states, a continuous critic admits independent statewise maximizers and $\Delta(Q) = 0$. In a coupled RKHS ball this equality requires verification. More generally, for a measurable statewise selector $a_Q^\star$ and policies $\pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})$, action Lipschitz continuity gives

$$\Delta(Q) \le \mathrm{Lip}_Q \inf_{\pi\in\Pi_{\mathcal{H}}(B_{\mathcal{H}})} \{\mathbb{E}_{\nu_+}\|a_Q^\star(S) - \pi(S)\|_{L^2}^2\}^{1/2}.$$

If the pointwise loss instead has a verified quadratic upper bound $Q(s, a_Q^\star(s)) - Q(s,\pi(s)) \le \mathrm{Lip}_{Q,2}\|a_Q^\star(s) - \pi(s)\|_{L^2}^2/2$, the corresponding squared approximation bound applies. A constrained boundary maximizer by itself does not make the first derivative vanish.

D.2 What norm regularization supplies

The positive deterministic penalty first gives a sample-size-dependent radius without any concentration argument. Suppose the zero policy belongs to Π and the computed update has regularized-objective gap at most $\bar{\tau}_{\pi,n}$ relative to zero. Since $0 \le Q \le V_{\max}$,

$$\|\widehat{\pi}_Q\|_{\mathcal{H}}^2 \le B_{n,\mathrm{pre}}^2 := \frac{V_{\max} + \bar{\tau}_{\pi,n}}{\lambda_{\pi,n}}. \tag{15}$$

This preliminary ball is deterministic once the tuning and numerical tolerance are fixed, but it grows if $\lambda_{\pi,n} \downarrow 0$.

A fixed-radius refinement follows from low-norm near-optimal comparators. Suppose that, for every covered Q, there is a policy $\pi_Q^0 \in \Pi$ such that

$$\|\pi_Q^0\|_{\mathcal{H}} \le B_0, \qquad \sup_{\pi\in\Pi} \Phi_Q(\pi) - \Phi_Q(\pi_Q^0) \le z_n,$$

and the numerical representation and solver guarantees imply

$$\widehat{\Psi}_Q(\widehat{\pi}_Q) \ge \widehat{\Psi}_Q(\pi_Q^0) - \bar{\tau}_{\pi,n}.$$

Assume $B_0 \le B_{n,\mathrm{pre}}$ and

$$\sup_{Q,\pi\in\Pi_{\mathcal{H}}(B_{n,\mathrm{pre}})} |\widehat{\Phi}_Q(\pi) - \Phi_Q(\pi)| \le \bar{u}_n.$$

Then

$$\|\widehat{\pi}_Q\|_{\mathcal{H}}^2 \le B_0^2 + \frac{z_n + 2\bar{u}_n + \bar{\tau}_{\pi,n}}{\lambda_{\pi,n}}. \tag{16}$$

Indeed, regularized near-optimality bounds the returned squared norm by B_0^2 plus the empirical objective advantage of $\widehat{\pi}_Q$ over π_Q^0, divided by $\lambda_{\pi,n}$, and the latter advantage is at most $z_n + 2\bar{u}_n$. If the right side is at most $B_{\mathcal{H}}^2$, every returned policy belongs to the fixed ball $\Pi_{\mathcal{H}}(B_{\mathcal{H}})$. Thus errors of order $\lambda_{\pi,n}$ and a uniformly bounded comparator norm are sufficient for the localization event used by the coverage and value proofs. All classes in this argument, including the preliminary ball, are declared before fitting.

D.3 Equal-length subjects and joint covering inputs

Assumption D.1 (Independent subjects)**.** The subject-level data have a common trajectory length and satisfy the data conditions in Section 2; arbitrary dependence within a subject is permitted.

Each subject is one independent sampling unit for the concentration argument, so the unconditional statistical scale is N. This sampling model complements the stationary mixing trajectory of Antos et al. (2007, Assumption A2): temporal information is used here through the separate mixing refinement, while the subject-level bounds allow arbitrary within-trajectory dependence. For a deterministic function class $\mathcal{O}$, write $H_{\mathcal{O}}(\epsilon) = \log \mathrm{Cov}(\epsilon, \mathcal{O}, \|\cdot\|_\infty)$. A covering number is the minimum number of radius-ϵ balls needed to cover the class. Unpenalized policy objectives on the localization event form $\mathcal{O}_n = \{s \mapsto Q(s, \pi(s)) : Q \in \mathcal{Q}_n, \pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})\}$. If critics have action Lipschitz bound Lip_n, their product cover gives

$$H_{\mathcal{O}_n}(\epsilon_Q + \mathrm{Lip}_n \epsilon_\pi) \le H_{\mathcal{Q}_n}(\epsilon_Q) + H_{\Pi_{\mathcal{H}}(B_{\mathcal{H}})}(\epsilon_\pi), \qquad d_\Pi(\pi, \pi') = \sup_s \|\pi(s) - \pi'(s)\|_{L^2}.$$

Bellman label covers use these unpenalized objectives. The deterministic RKHS penalty has no deviation term. Optional empirical derivative diagnostics use their own density class or matrix event.

For $0 \le f \le B$, a deterministic net and independent subject averages give

$$\sup_{f \in \mathcal{O}} |(\widehat{\mathbb{P}} - \mathbb{P})f| \le 2\epsilon + B\sqrt{\frac{H_{\mathcal{O}}(\epsilon) + \log(2/\delta)}{2N}} \tag{17}$$

with probability at least $1-\delta$. In particular, this supplies u_n for $\mathcal{O}_n$. For clipped fits and bounded labels with conditional targets Q_y, the same subject-level argument with relative Bernstein concentration yields

$$\sup_y \|\widehat{f}_y - Q_y\|_{L^2(\nu_X)}^2 \le C\left\{\eta_Q^2 + \tau_Q + V_{\max}(\epsilon + \eta) + \frac{V_{\max}^2[H_F(\epsilon) + H_Y(\eta) + \log(2/\delta)]}{N}\right\}. \tag{18}$$

The full proof, including the different observed marginals within subjects, is the weighted calculation below with weights $1/N$. There, all three subject scales are exactly N. For a derivative state representation with rank r_{st} and standardized feature envelope Cr_{st}, its relative Gram event has failure at most $2r_{\mathrm{st}} \exp(-cN/r_{\mathrm{st}})$. These events can be intersected with failure at most their sum; they need not be independent.

D.4 Extension: unequal and random trajectory lengths

The main presentation has equal weights $1/N$. This extension retains the original unequal-length concentration scales. No within-trajectory independence, marginal stationarity, mixing, subsampling or change to the empirical objective is introduced.

Length arrays and the pooled design. For N subjects, let $T_{i,N} \ge 1$ be the number of logged transitions from subject i. Put

$$n = \sum_{i=1}^N T_{i,N}, \qquad \overline{T}_N = n/N, \qquad T_{\max,N} = \max_i T_{i,N}, \qquad w_{i,N} = T_{i,N}/n.$$

The imbalance ratio remains $C_{T,n} = T_{\max,N}/\overline{T}_N \geq 1$. Lengths are deterministic, or probability statements are conditional on their vector under the preceding convention: conditional on that vector, subjects remain independent, each observed reward–transition conditional law remains compatible with the common kernel, and the claimed constants are uniform. Conditioning on arbitrary informative stopping times is not asserted to preserve these properties. For readability omit the subscript N on T_i, w_i within the proof.

For $Z_{it} = (S_{it}, A_{it}, R_{it}, S_{i,t+1})$, write

$$\widehat{\mathbb{P}}f = \frac{1}{n}\sum_{i,t} f(Z_{it}) = \sum_i w_i \overline{f}_i, \qquad \overline{f}_i = \frac{1}{T_i}\sum_{t=1}^{T_i} f(Z_{it}), \qquad \mathbb{P}f = \mathbb{E}\widehat{\mathbb{P}}f.$$

Thus neither fitting nor policy improvement becomes an equally weighted mean of subject means. The pooled laws remain

$$\nu_{X,N} = \sum_i w_i \frac{1}{T_i}\sum_t \mathcal{L}(S_{it}, A_{it}), \qquad \nu_N^+ = \sum_i w_i \frac{1}{T_i}\sum_t \mathcal{L}(S_{i,t+1}).$$

They may depend on the length array and on N. All approximation, membership and localization, information, state-transfer and state-leverage conditions used below refer to these laws and to the same classes fixed before fitting. A rate for the lengths alone does not establish any of those other conditions.

Define two deterministic cluster scales

$$\mathcal{N}_\infty = \frac{1}{\max_i w_i} = \frac{n}{T_{\max}} = \frac{N}{C_{T,n}}, \qquad \mathcal{N}_2 = \frac{1}{\sum_i w_i^2} = \frac{n^2}{\sum_i T_i^2}.$$

Because $\sum_i w_i = 1$, $\sum_i w_i^2 \leq \max_i w_i$ and $\sum_i w_i^2 \geq 1/N$,

$$N_\star = \frac{N}{C_{T,n}^2} \leq \mathcal{N}_\infty \leq \mathcal{N}_2 \leq N.$$

Unlike $N_\star$, $\mathcal{N}_\infty$ is always at least one. It diverges exactly when no single subject has a nonvanishing fraction of the transition weight. The next proof uses $\mathcal{N}_2$ for bounded objective deviations but only $\mathcal{N}_\infty$ as a common scale for prediction and relative penalty control. The larger $\mathcal{N}_2$ is not substituted throughout without a separate argument.

Three simultaneous statistical inputs.

Lemma D.2 (Independent subjects with unequal trajectory lengths)**.** *Assume independent subjects and the common observed conditional model. All classes and net centers below are deterministic, conditional on the lengths when needed.*

For any class $\mathcal{O}$ with $0 \leq f \leq B$ and logarithmic sup-norm covering number $H_{\mathcal{O}}(\epsilon)$, with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{O}} |(\widehat{\mathbb{P}} - \mathbb{P})f| \leq 2\epsilon + B\sqrt{\frac{H_{\mathcal{O}}(\epsilon) + \log(2/\delta)}{2\mathcal{N}_2}} \leq 2\epsilon + CB\sqrt{\frac{H_{\mathcal{O}}(\epsilon) + \log(2/\delta)}{\mathcal{N}_\infty}}. \tag{19}$$

For full-label clipped least-squares regression, suppose $f, Y_y, Q_y \in [0, V_{\max}]$, $Q_y(X_{it}) = \mathbb{E}(Y_y(Z_{it}) \mid X_{it})$ at every observed pair, and the uniform full-domain approximation error is at most η_Q. Let $H(\epsilon,\eta)$ be the sum of the log covers of the fit and label classes at sup-norm accuracies ϵ, η. If the returned fit is τ_Q-optimal for the full empirical squared loss, then, on one full-label event of probability at least $1-\delta$,

$$\sup_y \|\widehat{f}_y - Q_y\|^2_{L^2(\nu_{X,N})} \leq C\left\{\eta_Q^2 + \tau_Q + V_{\max}(\epsilon+\eta) + \frac{V_{\max}^2\{H(\epsilon,\eta) + \log(2/\delta)\}}{\mathcal{N}_\infty}\right\}. \tag{20}$$

For any deterministic state-feature map w_{st}, let $\Sigma_{\mathrm{st}} = \mathbb{E}_{\nu_N^+} w_{\mathrm{st}} w_{\mathrm{st}}^\top$ and $\widehat{\Sigma}_{\mathrm{st}} = n^{-1}\sum_{i,t} w_{\mathrm{st}}(S_{i,t+1}) w_{\mathrm{st}}(S_{i,t+1})^\top$, with rank $r_{\mathrm{st}} \leq d_S$. If $\operatorname{ess\,sup}_{\nu_N^+} w_{\mathrm{st}}^\top \Sigma_{\mathrm{st}}^\dagger w_{\mathrm{st}} \leq C d_S$, then

$$\Pr\left\{\tfrac{1}{2}\Sigma_{\mathrm{st}} \not\preccurlyeq \widehat{\Sigma}_{\mathrm{st}} \text{ or } \widehat{\Sigma}_{\mathrm{st}} \not\preccurlyeq \tfrac{3}{2}\Sigma_{\mathrm{st}}\right\} \leq 2r_{\mathrm{st}} \exp\{-c\mathcal{N}_\infty/d_S\}. \tag{21}$$

The zero-rank case has equality almost surely and failure probability zero. In particular, $\mathcal{N}_\infty \geq Cd_S \log(2d_S/\delta)$ *gives the displayed relative Gram comparison. Each bound uses whole subjects as independent units.*

Weighted subject averages let the unequal-range inequality of Hoeffding (1963, Theorem 2) retain $\mathcal{N}_2$ for bounded objectives, whereas the relative-risk and Gram bounds use the largest subject weight through $\mathcal{N}_\infty$. The matrix step applies Tropp (2015, Theorem 5.1.1) to whole-subject feature sums, preserving all logged transitions in the fitted objective.

Proof. For a fixed f, $w_i\overline{f}_i$ are independent and lie in intervals of lengths w_iB. The unequal-range Hoeffding inequality gives

$$\Pr\{|(\widehat{\mathbb{P}} - \mathbb{P})f| > x\} \leq 2\exp\left\{-\frac{2x^2}{B^2\sum_i w_i^2}\right\}.$$

This is Hoeffding (1963, Theorem 2) applied to weighted subject averages. A union bound over a deterministic ϵ-net and an off-net cost 2ϵ prove (19). It applies to unpenalized policy objectives by using $f(s') = Q\{s', \pi(s')\}$ and their joint cover.

For a fixed fit/label pair, put

$$\begin{aligned} d_{f,y}(Z) &= (f(X) - Y_y(Z))^2 - (Q_y(X) - Y_y(Z))^2, \\ D_i &= T_i^{-1}\sum_t d_{f,y}(Z_{it}), \\ \mu_i &= \mathbb{E}D_i = T_i^{-1}\sum_t \mathbb{E}(f - Q_y)^2(X_{it}) \geq 0, \\ \epsilon_{f,y}^2 &= \sum_i w_i\mu_i = \|f - Q_y\|_{L^2(\nu_{X,N})}^2. \end{aligned}$$

Boundedness and conditional compatibility imply $|d_{f,y}| \leq V_{\max}^2$ and $d_{f,y}^2 \leq 4V_{\max}^2(f - Q_y)^2$. Jensen's inequality within the subject, without independence, gives

$$\begin{aligned} &|w_i(D_i - \mu_i)| \leq 2w_iV_{\max}^2, \\ &\sum_i \mathrm{Var}(w_iD_i) \leq \sum_i w_i^2\mathbb{E}D_i^2 \leq 4V_{\max}^2\sum_i w_i^2\mu_i \leq \frac{4V_{\max}^2}{\mathcal{N}_\infty}\epsilon_{f,y}^2. \end{aligned}$$

The scalar Bernstein inequality for the independent centered terms $w_i(D_i - \mu_i)$, followed by Young's inequality, yields, for $x > 0$, with failure at most $2e^{-x}$,

$$|(\widehat{\mathbb{P}} - \mathbb{P})d_{f,y}| \leq \sqrt{\frac{8V_{\max}^2\epsilon_{f,y}^2x}{\mathcal{N}_\infty}} + \frac{4V_{\max}^2x}{3\mathcal{N}_\infty} \leq \tfrac{1}{4}\epsilon_{f,y}^2 + \frac{CV_{\max}^2x}{\mathcal{N}_\infty}.$$

One may use the one-dimensional specialization of Tropp (2015, Theorem 6.1.1). The linear remainder $4V_{\max}^2x/(3\mathcal{N}_\infty)$ in the Bernstein bound is harmless. Take a product fit/label net and $x = H(\epsilon, \eta) + \log(2/\delta)$. The common conditional kernel gives $\|Q_y - Q_{y'}\|_\infty \leq \|y - y'\|_\infty$. For representatives (f', y') with $\|f - f'\|_\infty \leq \epsilon$ and $\|y - y'\|_\infty \leq \eta$, difference of squares gives

$$\|d_{f,y} - d_{f',y'}\|_\infty \leq 2V_{\max}\epsilon + 6V_{\max}\eta =: a_1, \qquad |\epsilon_{f,y}^2 - \epsilon_{f',y'}^2| \leq 2V_{\max}(\epsilon + \eta) =: a_2.$$

The centered empirical difference costs at most $2a_1$, and replacing the risk in the relative term costs $a_2/4$. Thus the same event extends to all pairs with a remainder $C\{V_{\max}(\epsilon + \eta) + V_{\max}^2x/\mathcal{N}_\infty\}$. Net centers may be chosen in the deterministic classes; a factor-two change of radius for a restricted subclass is absorbed into C.

Let f_y° approach the deterministic approximation oracle. On that event the approximate empirical minimization inequality gives

$$\tfrac{3}{4}\epsilon_{\widehat{f}_y,y}^2 \leq \tfrac{5}{4}\epsilon_{f_y^\circ,y}^2 + \tau_Q + C\left\{V_{\max}(\epsilon + \eta) + V_{\max}^2x/\mathcal{N}_\infty\right\}.$$

Let the near-minimizer tolerance decrease to zero. This proves (20). No common subject marginal distribution was required: each μ_i may differ. Consequently the step $\sum_i w_i^2\mu_i \leq (\sum_i w_i^2)\epsilon_{f,y}^2$ is not valid in general, which is why (20) does not claim the scale $\mathcal{N}_2$.

For (21), every observed feature vector belongs to $\mathrm{Range}(\Sigma_{\mathrm{st}})$ almost surely. Indeed, each marginal is a positive-weight component of ν_N^+; apply the zero second moment to a basis of $\ker \Sigma_{\mathrm{st}}$. On the nonzero eigenspace let $\widetilde{w} = \Sigma_{\mathrm{st}}^{\dagger/2} w_{\mathrm{st}}$ and define

$$\Xi_i = \frac{1}{n} \sum_{t=1}^{T_i} \widetilde{w}(S_{i,t+1}) \widetilde{w}(S_{i,t+1})^\top .$$

These matrices are independent across subjects, positive semidefinite, and satisfy

$$0 \preceq \Xi_i \preceq C d_S w_i \,\mathrm{Id} \preceq \frac{C d_S}{\mathcal{N}_\infty} \,\mathrm{Id}, \qquad \sum_i \mathbb{E}\Xi_i = \mathrm{Id}\,.$$

The matrix Chernoff inequality (Tropp, 2015, Theorem 5.1.1) at relative error $1/2$ proves the two eigenvalue tails in (21). Transform back on $\mathrm{Range}(\Sigma_{\mathrm{st}})$ and retain the common null space. If an empirical finite-feature curvature diagnostic is used, integrating the same quadratic-form inequalities against its derivative coefficients gives the corresponding comparison; this optional consequence is not needed for the deterministic RKHS penalty. □

The three events can be intersected with failure at most the sum of their stated probabilities. In particular, this weighted refinement does not rely on independence between the policy update, fitted critics and Gram matrix.

E Deterministic classes and RKHS policy entropy

For a critic Q and policy π, write

$$y_{Q,\pi}(s,a,r,s') = r + \gamma Q\{s', \pi(s')\}.$$

Step 3 uses the realized function $y_m(\mathcal{D}) = y_{\widehat{Q}_{m-1}, \widehat{\pi}_{m-1}}$ to form $Y_{i,t,m} = y_m(\mathcal{D})(S_{i,t}, A_{i,t}, R_{i,t}, S_{i,t+1})$. The following deterministic construction covers these data-dependent labels and their fitted-critic/target pairs.

Assumption E.1 (Classes fixed before fitting)**.** Fix a deterministic clipped critic class $\mathcal{Q}_n$ and a deterministic label class

$$\mathcal{Y}_n \supset \{(s,a,r,s') \mapsto r + \gamma Q\{s', \pi(s')\} : Q \in \mathcal{Q}_n,\ \pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})\}.$$

The membership and localization event is $\mathcal{R}_n$ defined in Section 4.1: it requires $\widehat{Q}_m \in \mathcal{Q}_n$ and $\widehat{\pi}_m \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})$ for $0 \le m \le M$. The event $\mathcal{U}_n$ is defined by the four uniform bounds in (7), and $\mathcal{E}_n = \mathcal{R}_n \cap \mathcal{U}_n$. Sufficient concentration and computational events are established on these deterministic classes before fitted functions are substituted.

This is the fit-and-target uniformity used to justify data reuse in Nguyen-Tang et al. (2022). The label family ranges over the deterministic norm ball containing every realized policy on the membership and localization event, so a fitted policy and critic can be substituted together after all iterations have reused the data. For every deterministic $y \in \mathcal{Y}_n$, define its conditional target by

$$Q_y(s,a) = \int y(s,a,r,s') P_{\mathrm{env}}(dr, ds' \mid s,a), \qquad (s,a) \in \mathcal{S} \times \mathcal{A}.$$

The environment kernel defines this function on the full admissible domain, including actions outside behavior support. Choose deterministic target approximants f_y° satisfying

$$\sup_{y \in \mathcal{Y}_n} \|f_y^\circ - Q_y\|_\infty \le \eta_{Q,n},$$

where the supremum norm is over that full domain. Approximation only on behavior support would not control the approximation error at policy-selected actions. For exact functional inputs, let $\mathcal{F}_n(y)$ be a deterministic class containing every critic that step 4 may return for label y. The pair family in Section 3 is

$$\mathcal{V}_n = \{(f, f_y^\circ) : y \in \mathcal{Y}_n,\ f \in \mathcal{F}_n(y)\}.$$

With numerical functional scores, use the corresponding exact-score fit family in this definition and control the discrepancy between implemented and exact-score critics separately by $e_{Q,n}$, as in Corollary F.1. For KRR, Appendix H applies the secant argument to raw fits before clipping. On $\mathcal{R}_n$, $y_m(\mathcal{D}) \in \mathcal{Y}_n$. Thus, on $\mathcal{E}_n$,

$$\max_m E_{\mathcal{D}}(y_m(\mathcal{D})) \le \sup_{y \in \mathcal{Y}_n} E_{\mathcal{D}}(y).$$

Here $E_{\mathcal{D}}(y)$ denotes any error functional controlled uniformly in y on $\mathcal{U}_n$. This is a pathwise substitution, and uses no new independence of fitted labels; the fitted critics are paired with the approximants of their realized labels in the same fixed $\mathcal{Y}_n$. If $\mathcal{Q}_n$ is enlarged to include idealized counterparts that evaluate functional integrals exactly, or other numerical approximants, the corresponding label cover is enlarged before fitting as well.

Proposition E.1 (Policy entropy from RKHS evaluation tails)**.** *Suppose $\{e_\ell\}$ is an orthonormal basis of the policy RKHS, uniform L^2 evaluation has norm at most $\kappa_{\mathcal{H}}$, and for some $p_{\mathrm{eig}} > 0$,*

$$\sup_s \sum_{\ell > m} \|e_\ell(s)\|_{L^2}^2 \le C m^{-2p_{\mathrm{eig}}}.$$

For every subset Π of its radius-B_Π ball,

$$\log \mathrm{Cov}(\epsilon, \Pi, d_\Pi) \le C\epsilon^{-1/p_{\mathrm{eig}}} \log(1 + C/\epsilon), \qquad 0 < \epsilon \le 1,$$

with constants depending on the radius and the displayed evaluation bounds.

The reproducing evaluation bound of Micchelli & Pontil (2005, Proposition 2.1) controls the finite coefficient block, and the assumed uniform tail controls the remaining directions. Their combination gives a cover of the entire RKHS ball; the truncation is a covering argument and leaves the policy domain unchanged.

Proof. Write $\pi = \sum_\ell \theta_\ell e_\ell$, $\sum_\ell \theta_\ell^2 \le B_\Pi^2$. Cauchy–Schwarz gives uniform tail error at most $C^{1/2} B_\Pi m^{-p_{\mathrm{eig}}}$. Choose m so this is at most $\epsilon/4$. On the first m coefficients, evaluation is $\kappa_{\mathcal{H}}$-Lipschitz from Euclidean norm to d_Π. A radius-$\epsilon/(4\kappa_{\mathcal{H}})$ Euclidean net in the radius-B_Π ball has at most $(1 + 8\kappa_{\mathcal{H}} B_\Pi/\epsilon)^m$ points. The truncation and cover therefore give error at most $\epsilon/2$. For a restricted feasible class, replace each ball intersecting Π by a representative from that intersection, at cost at most a factor two in radius. Taking logarithms proves the bound. □

Bounded kernel diagonal supplies the evaluation bound, whereas the tail condition supplies uniform approximation of the whole RKHS ball. These are separate properties. A Mercer eigenvalue bound alone does not supply the uniform eigenfunction tail in the proposition.

For action-Lipschitz critics, the label map satisfies

$$\|y_{Q,\pi} - y_{Q',\pi'}\|_\infty \le \gamma\{\|Q - Q'\|_\infty + \mathrm{Lip}_n d_\Pi(\pi, \pi')\}.$$

Combining this with the preceding cover gives full label and objective entropy over Π. A finite parameter count may be used only for a verified numerical subclass; it is not the entropy of the full infinite-dimensional policy domain.

For the two-state fixed comparison ball, which is contained in a product of two fixed H^2 balls, Proposition C.3 also gives $\log \mathrm{Cov}\{\epsilon, \Pi_{\mathcal{H}}(B_{\mathcal{H}}), d_\Pi\} \le C\epsilon^{-1/2} \log(1 + C/\epsilon)$. Indeed use $d_u \asymp \epsilon^{-1/2}$, cover the projected functional actions in an L^2-orthonormal basis of their $O(d_u)$-dimensional spline space, and take a product of the two covers. The radius of that Euclidean ball is bounded by the continuous embedding of $\mathcal{G}$ into L^2; moving centers into the feasible class has the same factor-two cost. For its actual Bellman labels an even smaller continuation-cube cover is available, as proved in Appendix M.

Proposition E.2 (Entropy of a full continuous-state policy RKHS ball)**.** *Let $\mathcal{S} = [0,1]$, $\mathcal{G} = H^2(0,1)$ with the Sobolev norm of Proposition C.1, and $\mathcal{H} = H^1([0,1]; \mathcal{G})$. Fix*

$$\Pi_{\mathcal{H}}(B_{\mathcal{H}}) = \{\pi \in \mathcal{H} : \|\pi\|_{\mathcal{H}} \le B_{\mathcal{H}}, \sup_s \|\pi(s)\|_{L^2} \le B_A\}.$$

For this entire fixed ball and $d_\Pi(\pi, \pi') = \sup_s \|\pi(s) - \pi'(s)\|_{L^2}$,

$$\log \mathrm{Cov}\{\epsilon, \Pi_{\mathcal{H}}(B_{\mathcal{H}}), d_\Pi\} \le C\epsilon^{-5/2} \log(1 + C/\epsilon), \qquad 0 < \epsilon \le 1. \tag{22}$$

The constant depends on the fixed radii and the continuous embeddings, not on a computational dimension.

The Hilbert-valued RKHS evaluation property of Micchelli & Pontil (2005, Proposition 2.1) supplies the dimension-independent evaluation bound. Here the approximation in Appendix C.3 supplies a quantitative cover of the whole ball. The finite-rank spaces below are a device for proving that cover; policies, Bellman labels and the value benchmark still range over $\Pi_{\mathcal{H}}(B_{\mathcal{H}})$.

Proof. Let $\mathbf{U}_{h_s,d_u}$ be the tensor energy projection from Appendix C.3. The state and functional-action projections commute and are contractions in $\mathcal{H}$, so

$$\|\mathbf{U}_{h_s,d_u}\pi\|_{\mathcal{H}} \le B_{\mathcal{H}}, \qquad d_\Pi(\pi, \mathbf{U}_{h_s,d_u}\pi) \le CB_{\mathcal{H}}(h_s^{1/2} + d_u^{-2}).$$

Choose $h_s \asymp \epsilon^2$ and $d_u \asymp \epsilon^{-1/2}$, with constants giving approximation error at most $\epsilon/8$. The image space has dimension $D \le Ch_s^{-1}d_u \le C\epsilon^{-5/2}$. An $\mathcal{H}$-orthonormal coordinate system identifies its radius-$B_{\mathcal{H}}$ ball with a Euclidean ball. Uniform state evaluation gives $d_\Pi(f,g) \le \kappa_{\mathcal{H}}\|f-g\|_{\mathcal{H}}$, with fixed $\kappa_{\mathcal{H}}$. A Euclidean net of radius $\epsilon/(8\kappa_{\mathcal{H}})$ has at most $(1+16\kappa_{\mathcal{H}}B_{\mathcal{H}}/\epsilon)^D$ centers. Combining the projection error and this net covers $\Pi_{\mathcal{H}}(B_{\mathcal{H}})$ by ambient balls of radius $\epsilon/4$. For every ball meeting the feasible class, choose a center in that intersection; the radius at most doubles. Taking logarithms proves the claim, including the action-feasibility restriction. □

F SMOOTH POLICY ACTIONS AND MOMENT-BASED SECANT IDENTIFICATION

Norm localization and (3) supply the spectral radius in (4), while observed actions in $L^2[0,1]$ satisfy the regression moment or envelope conditions. Identification is relative to critic differences and their second moments, with no local action probability condition. State transfer in Section B remains a separate requirement.

Use the deterministic smooth class $\Pi_{\mathcal{H}}(B_{\mathcal{H}})$ defined in Section 2. All fitted critics and Bellman labels are covered on their full deterministic classes. Write

$$\|h\|^2_{\nu_+,\Pi_{\mathcal{H}}(B_{\mathcal{H}}),2} = \int \sup_{\pi\in\Pi_{\mathcal{H}}(B_{\mathcal{H}})} |h\{s,\pi(s)\}|^2 \nu_+(ds).$$

For a fixed orthonormal basis $\{\phi_j\}$ of $L^2([0,1])$, define

$$\|a\|_\zeta^2 = \sum_{j\ge1} j^{2\zeta}|\langle a,\phi_j\rangle|^2, \qquad \mathbf{W}_\alpha\phi_j = j^{-2\alpha}\phi_j, \quad \alpha>0.$$

The spectral weight operator $\mathbf{W}_\alpha$ sets the scale assigned to functional coordinates in the transfer argument. It does not assert that the observed action covariance is diagonal or has these eigenvalues. The directional comparison (26) must be verified separately, and eigenvalue decay alone does not imply it. If an infinite-dimensional covariance dominates $\mathbf{W}_\alpha$ on all of $L^2([0,1])$, finite second moments force $\alpha > 1/2$.

For the state-free linear illustration $h(a) = \langle d, a\rangle$ in Section 3, the feature is the action itself. For $0 < \vartheta \le 1$ with $\alpha\vartheta \le \zeta$,

$$\|\mathbf{W}_\alpha^{-\vartheta/2}a\|^2 = \sum_{j\ge1} j^{2\alpha\vartheta}|\langle a,\phi_j\rangle|^2 \le \|a\|_\zeta^2.$$

Thus actions with $\|a\|_\zeta \le B$ have smooth-evaluation factor at most B. If the logged second-moment condition is $e^2 := \mathbb{E}h(A)^2 \ge c\|\mathbf{W}_\alpha^{1/2}d\|^2$ for $c>0$, weighted Cauchy–Schwarz and spectral interpolation give

$$|h(a)| \le B\|\mathbf{W}_\alpha^{\vartheta/2}d\| \le B\|d\|^{1-\vartheta}\|\mathbf{W}_\alpha^{1/2}d\|^\vartheta \le Bc^{-\vartheta/2}\|d\|^{1-\vartheta}e^\vartheta.$$

This is the action-linear specialization of Proposition F.1; smoothness controls evaluation, while the logged second-moment condition supplies information.

F.1 ACTUAL SECANTS AND EVALUATION AT A PRESCRIBED ACCURACY

All critic pairs in this section belong to a family $\mathcal{V}_n$ fixed before fitting. It contains every possible fitted critic and its population target approximant for each Bellman label under analysis. When

numerical functional scores are used, the corresponding exact-score critic is the idealized version obtained by evaluating those integrals exactly rather than by quadrature; the numerical discrepancy is added separately. Neither observed prediction errors nor the fitted path are used to discard pairs. For $v = (f, f^\circ) \in \mathcal{V}_n$, put

$$\upsilon_v = \|f - f^\circ\|_{\nu_+, \Pi_{\mathcal{H}}(B_{\mathcal{H}}), 2}, \qquad e_v = \|f - f^\circ\|_{L^2(\nu_X)}.$$

For an evaluation accuracy $\varepsilon > 0$, define

$$C_n\{\varepsilon; \Pi_{\mathcal{H}}(B_{\mathcal{H}})\} = \sup_{v \in \mathcal{V}_n : \upsilon_v > \varepsilon} \frac{\upsilon_v^2}{e_v^2}. \tag{23}$$

A positive numerator divided by zero is $+\infty$, and the supremum of an empty set is zero. This amplification concerns actual critic differences, rather than every vector in a feature space. Accuracy ε is an analysis parameter, not a modification of either algorithmic penalty. Uniform control as $\varepsilon \downarrow 0$ is not required.

For finite $C_\varepsilon = C_n\{\varepsilon; \Pi_{\mathcal{H}}(B_{\mathcal{H}})\}$, the definition separates two cases:

$$\begin{aligned} \upsilon_v \le \varepsilon \quad &\Longrightarrow \quad \upsilon_v \le \varepsilon + \sqrt{C_\varepsilon}\, e_v, \\ \upsilon_v > \varepsilon \quad &\Longrightarrow \quad e_v > 0, \quad \upsilon_v \le \sqrt{C_\varepsilon}\, e_v. \end{aligned}$$

Consequently

$$\upsilon_v \le \varepsilon + \sqrt{C_n\{\varepsilon; \Pi_{\mathcal{H}}(B_{\mathcal{H}})\}}\, e_v \qquad (v \in \mathcal{V}_n), \tag{24}$$

and a deterministic bound $e_v \le \delta$ gives

$$\upsilon_v \le \omega_n(\delta) := \inf_{\varepsilon > 0 :\, C_n\{\varepsilon; \Pi_{\mathcal{H}}(B_{\mathcal{H}})\} < \infty} \left[\varepsilon + \sqrt{C_n\{\varepsilon; \Pi_{\mathcal{H}}(B_{\mathcal{H}})\}}\, \delta\right]. \tag{25}$$

An empty set of pairs with $\upsilon_v > \varepsilon$ gives $C_\varepsilon = 0$ and $\upsilon_v \le \varepsilon$ for every pair. An empty admissible-accuracy set gives $\omega_n = +\infty$. Along a sequence of fits on the fixed policy class, the prediction-error sequence $\delta_{Q,n}$ must satisfy $\omega_n(\delta_{Q,n}) \to 0$ for consistency; finiteness at each fixed accuracy alone is insufficient.

The coefficient in (23) is not assumed to be small without a verification. The following directional moment and smooth-feature calculations supply quantitative sufficient conditions for it.

F.2 Directional moments and smooth functional evaluation

For each $v = (f, f^\circ) \in \mathcal{V}_n$, put $h_v = f - f^\circ$. Let a separable Hilbert space $\mathcal{E}_v$, a coefficient $d_v \in \mathcal{E}_v$, and a measurable feature map give the exact identity

$$h_v(s,a) = \langle d_v, \psi_v(s,a)\rangle, \qquad \Sigma_v = \mathbb{E}_{\nu_X}[\psi_v(X) \otimes \psi_v(X)], \qquad e_v^2 = \langle d_v, \Sigma_v d_v\rangle.$$

Here $(u \otimes w)z = u\langle w, z\rangle$ is a rank-one operator; thus $\Sigma_v z = \mathbb{E}_{\nu_X}[\psi_v(X)\langle \psi_v(X), z\rangle]$. No feature mean is subtracted. Assume $\mathbb{E}\|\psi_v(X)\|^2 < \infty$. Let $\mathbf{U}_v$ be the orthogonal projection onto the closed span of feature values on the entire domain relevant to fitting and evaluation, not onto the span of the observed sample. Thus $\mathbf{U}_v^* = \mathbf{U}_v$, $\mathbf{U}_v^2 = \mathbf{U}_v$, $\|\mathbf{U}_v z\| \le \|z\|$, and $\psi_v(s,a) = \mathbf{U}_v \psi_v(s,a)$, including at policy-selected actions outside behavior support.

Assumption F.1 (Information along actual secants and smooth features)**.** For every pair, let $\mathbf{W}_v$ be the fixed smoothness-weight operator on $\mathcal{E}_v$, specified by the functional coordinates and critic representation. It is positive, injective and self-adjoint with norm at most one, so every nonzero direction receives a positive weight no larger than one. Uniformly on $\mathcal{V}_n$, require

$$\langle d_v, \Sigma_v d_v\rangle \ge c_n \langle \mathbf{U}_v d_v, \mathbf{W}_v \mathbf{U}_v d_v\rangle, \qquad c_n > 0. \tag{26}$$

For some $0 < \vartheta \le 1$, the candidate policies satisfy the smooth-feature bound

$$\int \sup_{\pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})} \|\mathbf{W}_v^{-\vartheta/2} \psi_v\{s, \pi(s)\}\|^2 \nu_+(ds) \le B_{\psi,n}^2. \tag{27}$$

The inverse is understood on its domain. All families, constants and suprema are fixed before fitting. If they are controlled on an event, the bound is uniform over a fixed class containing every possible fitted realization.

As in the Bellman-residual transfer coefficient of Xie et al. (2021, Definition 1), information is required on the error class used by the analysis. Here that class consists of actual critic secants, and the smooth-feature condition identifies which evaluations at functional policy actions can be controlled from their behavior second moments.

Spectral powers. Let $\mathbf{E}_{\mathbf{W}}$ be the spectral resolution of the positive injective contraction $\mathbf{W} = \mathbf{W}_v$, and $\mu_z(B) = \langle z, \mathbf{E}_{\mathbf{W}}(B)z\rangle$. No bounded inverse is assumed. Spectral calculus gives

$$\mathbf{W}^p = \int_{(0,1]} t^p \mathbf{E}_{\mathbf{W}}(dt) \quad (p \geq 0),$$

$$\mathrm{Dom}(\mathbf{W}^{-\vartheta/2}) = \{z : \int_{(0,1]} t^{-\vartheta} d\mu_z(t) < \infty\}, \qquad \|\mathbf{W}^{-\vartheta/2} z\|^2 = \int_{(0,1]} t^{-\vartheta} d\mu_z(t).$$

The last equality holds on that domain. In an eigenbasis it becomes $\sum_j \varpi_j^{-\vartheta} |\langle z, e_j\rangle|^2$, so the feature condition requires finite energy after upweighting directions with small smoothness weights.

Condition (26) concerns the actual coefficient d_v of each critic difference, rather than unrelated parameter directions. It is a joint state–action information condition, separate from policy smoothness and the eigenvalues of $\mathbb{E}[A \otimes A]$. Propositions H.1 and I.2 verify the feature bound from smoothness of policy-selected actions. These actions may lie outside behavior support.

A distribution-family-free sufficient verification of the directional moment condition is

$$\Pr_{\nu_X}\Big\{|h_v(X)| \geq b_n^{\mathrm{dir}} \|\mathbf{W}_v^{1/2} \mathbf{U}_v d_v\|\Big\} \geq p_n^{\mathrm{dir}} > 0 \tag{28}$$

for pairs with nonzero weighted coefficient energy. It gives $c_n = p_n^{\mathrm{dir}} (b_n^{\mathrm{dir}})^2$. Indeed, put $a_v = \|\mathbf{W}_v^{1/2} \mathbf{U}_v d_v\|$ and let $\mathsf{Obs}_v^{\mathrm{dir}}$ be the event inside (28). Nonnegativity of h_v^2 gives

$$e_v^2 = \mathbb{E} h_v(X)^2 \geq \mathbb{E}[h_v(X)^2 \mathbf{1}_{\mathsf{Obs}_v^{\mathrm{dir}}}] \geq (b_n^{\mathrm{dir}})^2 a_v^2 \Pr(\mathsf{Obs}_v^{\mathrm{dir}}) \geq p_n^{\mathrm{dir}} (b_n^{\mathrm{dir}})^2 a_v^2.$$

When $a_v = 0$ the required lower bound is immediate. This event measures whether the data distinguish two critics; the distinguishing observations need not be close to any policy-selected action. The quadratic-form comparison can instead be checked directly. Normality, independent action coordinates, and coordinatewise densities are not required.

Proposition F.1 (Secant transfer under uniform or accuracy-dependent information)**.** (i) Uniform information. *Under Assumption F.1,*

$$\upsilon_v \leq B_{\psi,n} c_n^{-\vartheta/2} \|d_v\|^{1-\vartheta} e_v^{\vartheta}.$$

If $\|d_v\| \leq B_{Q,n} < \infty$*, then, for* $\varepsilon > 0$*,*

$$C_n\{\varepsilon; \Pi_{\mathcal{H}}(B_{\mathcal{H}})\} \leq B_{\psi,n}^{2/\vartheta} c_n^{-1} B_{Q,n}^{2(1-\vartheta)/\vartheta} \varepsilon^{-2(1-\vartheta)/\vartheta}. \tag{29}$$

At $\vartheta = 1$ *no uniform coefficient radius is required, and the exact evaluation-to-prediction ratio is at most* $B_{\psi,n}^2/c_n$*. If all* $d_v = 0$*, all secants vanish; expressions with exponent zero are then unnecessary.*

(ii) Accuracy-dependent information. *Retain the smooth-feature condition and* $\|d_v\| \leq B_{Q,n}$*, but replace the uniform moment condition by the following weaker, accuracy-dependent requirement. For a fixed* $\varepsilon_0 > 0$ *and* $0 < \varepsilon \leq \varepsilon_0$*, only pairs with* $\upsilon_v > \varepsilon$ *must satisfy*

$$e_v^2 \geq c_n(\varepsilon) \langle \mathbf{U}_v d_v, \mathbf{W}_v \mathbf{U}_v d_v\rangle, \qquad c_n(\varepsilon) > 0. \tag{30}$$

Then (29) holds with $c_n(\varepsilon)$ *in place of* c_n*. Also, for every pair with* $e_v \leq \delta$*,*

$$\upsilon_v \leq \inf_{0<\varepsilon\leq\varepsilon_0} \Big\{\varepsilon + B_{\psi,n} c_n(\varepsilon)^{-\vartheta/2} B_{Q,n}^{1-\vartheta} \delta^{\vartheta}\Big\}. \tag{31}$$

If $c_n(\varepsilon) \geq c_* \varepsilon^{\beta_{\mathrm{info}}}$ *on a fixed accuracy interval, where* $\beta_{\mathrm{info}} \geq 0$ *is the information-decay exponent, and* c_*^{-1}*,* $B_{\psi,n}$ *and* $B_{Q,n}$ *are bounded, this is* $O(\delta^{\vartheta_*})$ *as* $\delta \downarrow 0$*, where*

$$\vartheta_* := \frac{2\vartheta}{2 + \beta_{\mathrm{info}}\vartheta}.$$

The directional probability verification (28) need only hold for the same pairs, and its constants may depend on ε*. This gives* $c_n(\varepsilon) = p_n^{\mathrm{dir}}(\varepsilon)\{b_n^{\mathrm{dir}}(\varepsilon)\}^2$ *without a local-positivity argument.*

The exponent follows the Hilbert-scale interpolation principle in Egger & Hofmann (2018, equation (1.4)), applied here to the coefficient and smooth policy-evaluation feature of each critic difference. Accuracy-dependent information permits weaker identification of differences whose policy-evaluation effect is already small, and the balance in part (ii) quantifies the resulting change in the transfer exponent.

Proof. Fix v, and write $\mathbf{W} = \mathbf{W}_v$, $z = \mathbf{U}_v d_v$ and $\psi = \psi_v(s,a)$. The full-domain projection gives $h_v(s,a) = \langle z, \psi \rangle$ and $\|z\| \le \|d_v\|$.

First justify the weighted inner product. For the spectral projections $\mathbf{E}_k = \mathbf{E}_{\mathbf{W}}([1/k, 1])$, the inverse power is bounded on $\mathrm{Range}(\mathbf{E}_k)$, and

$$\langle \mathbf{E}_k z, \mathbf{E}_k \psi \rangle = \langle \mathbf{W}^{\vartheta/2} \mathbf{E}_k z, \mathbf{W}^{-\vartheta/2} \mathbf{E}_k \psi \rangle.$$

Injectivity gives $\mathbf{E}_k z \to z$, and the assumed inverse-power domain gives $\mathbf{W}^{-\vartheta/2} \mathbf{E}_k \psi \to \mathbf{W}^{-\vartheta/2} \psi$. Passing to the limit and applying Cauchy–Schwarz yields

$$|h_v(s,a)|^2 \le \langle z, \mathbf{W}^{\vartheta} z \rangle \|\mathbf{W}^{-\vartheta/2} \psi_v(s,a)\|^2.$$

For $0 < \vartheta < 1$, Hölder's inequality with exponents $1/\vartheta$ and $1/(1-\vartheta)$ in the spectral measure μ_z gives

$$\begin{aligned} \langle z, \mathbf{W}^{\vartheta} z \rangle &= \int t^{\vartheta} \, d\mu_z(t) \le \left(\int t \, d\mu_z \right)^{\vartheta} \left(\int 1 \, d\mu_z \right)^{1-\vartheta} \\ &= \langle z, \mathbf{W} z \rangle^{\vartheta} \|z\|^{2(1-\vartheta)} \le c_n^{-\vartheta} e_v^{2\vartheta} \|d_v\|^{2(1-\vartheta)}. \end{aligned}$$

Taking the policy supremum before integrating the pointwise bound now gives

$$\begin{aligned} \upsilon_v^2 &\le \langle z, \mathbf{W}^{\vartheta} z \rangle \int \sup_{\pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})} \|\mathbf{W}^{-\vartheta/2} \psi_v(s, \pi(s))\|^2 \nu_+(ds) \\ &\le B_{\psi,n}^2 c_n^{-\vartheta} e_v^{2\vartheta} \|d_v\|^{2(1-\vartheta)}. \end{aligned}$$

This proves the transfer inequality. At $\vartheta = 1$, use $\langle z, \mathbf{W} z \rangle \le c_n^{-1} e_v^2$ directly, without a coefficient radius. No commutation of $\mathbf{U}_v$ and $\mathbf{W}_v$ has been used.

For $C_{\mathrm{tr}} = B_{\psi,n} c_n^{-\vartheta/2} B_{Q,n}^{1-\vartheta} > 0$ and $\upsilon_v > \varepsilon$, the transfer bound implies

$$e_v^2 \ge (\upsilon_v / C_{\mathrm{tr}})^{2/\vartheta}, \qquad \frac{\upsilon_v^2}{e_v^2} \le C_{\mathrm{tr}}^{2/\vartheta} \upsilon_v^{-2(1-\vartheta)/\vartheta} \le C_{\mathrm{tr}}^{2/\vartheta} \varepsilon^{-2(1-\vartheta)/\vartheta}.$$

Taking the supremum proves the amplification bound. Zero coefficients or $C_{\mathrm{tr}} = 0$ give vanishing secants, so there is no positive-numerator case to consider.

Accuracy-dependent information. Set $\Lambda_\varepsilon^{\mathrm{tr}} = B_{\psi,n} c_n(\varepsilon)^{-\vartheta/2} B_{Q,n}^{1-\vartheta}$. On the restricted family $\upsilon_v > \varepsilon$, the preceding spectral calculation applies with $c_n(\varepsilon)$ and gives

$$\upsilon_v \le \Lambda_\varepsilon^{\mathrm{tr}} e_v^{\vartheta}, \qquad \upsilon_v^2 / e_v^2 \le (\Lambda_\varepsilon^{\mathrm{tr}})^{2/\vartheta} \varepsilon^{-2(1-\vartheta)/\vartheta}.$$

For $e_v \le \delta$, separating $\upsilon_v \le \varepsilon$ from $\upsilon_v > \varepsilon$ gives $\upsilon_v \le \varepsilon + \Lambda_\varepsilon^{\mathrm{tr}} \delta^{\vartheta}$. Taking the infimum proves the transfer-envelope bound.

If $c_n(\varepsilon) \ge c_* \varepsilon^{\beta_{\mathrm{info}}}$, put $q_{\mathrm{bal}} = \beta_{\mathrm{info}} \vartheta / 2$ and absorb the bounded factors into C_*. When $q_{\mathrm{bal}} > 0$, choose $\varepsilon = \delta^{\vartheta/(1+q_{\mathrm{bal}})} \le \varepsilon_0$ for sufficiently small δ. Then

$$\upsilon_v \le \delta^{\vartheta/(1+q_{\mathrm{bal}})} + C_* \delta^{\vartheta} \delta^{-q_{\mathrm{bal}}\vartheta/(1+q_{\mathrm{bal}})} = (1 + C_*) \delta^{2\vartheta/(2+\beta_{\mathrm{info}}\vartheta)}.$$

When $q_{\mathrm{bal}} = 0$, let $r \downarrow 0$ to obtain $\upsilon_v \le C_* \delta^{\vartheta}$; when $\delta = 0$, the same limit gives $\upsilon_v = 0$.

Finally, on the same accuracy-restricted family, the directional probability event with threshold $b_{\mathrm{dir}} \|\mathbf{W}_v^{1/2} \mathbf{U}_v d_v\|$ and probability p gives

$$e_v^2 \ge \mathbb{E}\Big[h_v(X)^2 \mathbf{1}\{ |h_v(X)| \ge b_{\mathrm{dir}} \|\mathbf{W}_v^{1/2} \mathbf{U}_v d_v\| \} \Big] \ge p b_{\mathrm{dir}}^2 \langle \mathbf{U}_v d_v, \mathbf{W}_v \mathbf{U}_v d_v \rangle.$$

Thus its constants may depend on r without requiring observations near the policy-selected action. □

Remark F.1 (Actual secants versus full feature leverage)**.** For $\lambda_{\mathrm{lev}} > 0$, Proposition F.1 and $x^{\vartheta} y^{1-\vartheta} \le \lambda_{\mathrm{lev}}^{\vartheta-1}(x + \lambda_{\mathrm{lev}} y)$ give

$$\Lambda_{\Pi_{\mathcal{H}}(B_{\mathcal{H}})}(\lambda_{\mathrm{lev}}) := \sup_{v : d_v \ne 0} \frac{\upsilon_v^2}{e_v^2 + \lambda_{\mathrm{lev}} \|d_v\|^2} \le B_{\psi,n}^2 c_n^{-\vartheta} \lambda_{\mathrm{lev}}^{\vartheta - 1}. \tag{32}$$

To verify the scalar inequality used here, set $u = x/(x + \lambda_{\mathrm{lev}} y) \in [0, 1]$ when $x + \lambda_{\mathrm{lev}} y > 0$. Then

$$\frac{x^{\vartheta} y^{1-\vartheta}}{x + \lambda_{\mathrm{lev}} y} = \lambda_{\mathrm{lev}}^{\vartheta-1} u^{\vartheta} (1-u)^{1-\vartheta} \le \lambda_{\mathrm{lev}}^{\vartheta-1};$$

zero denominators are treated by continuity. Applying this to $x = e_v^2$ and $y = \|d_v\|^2$ proves (32). The full-feature coefficient based on the inverse second-moment operator $\Lambda_v\{\lambda_{\mathrm{lev}}; \Pi_{\mathcal{H}}(B_{\mathcal{H}})\} = \int \sup_{\pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})} \langle \psi_v\{s, \pi(s)\}, (\Sigma_v + \lambda_{\mathrm{lev}} \mathrm{Id})^{-1} \psi_v\{s, \pi(s)\} \rangle \nu_+(ds)$ instead controls all feature directions. Here Id is the identity, and $\Sigma_{v,\mathrm{lev}} = \Sigma_v + \lambda_{\mathrm{lev}} \mathrm{Id} \succeq \lambda_{\mathrm{lev}} \mathrm{Id}$ means $\langle z, \Sigma_{v,\mathrm{lev}} z \rangle \ge \lambda_{\mathrm{lev}} \|z\|^2$ for every z, so $\Sigma_{v,\mathrm{lev}}^{-1}$ is bounded. Weighted Cauchy–Schwarz and its equality case give the variational identity

$$|\langle d_v, \psi \rangle|^2 \le \langle d_v, \Sigma_{v,\mathrm{lev}} d_v \rangle \langle \psi, \Sigma_{v,\mathrm{lev}}^{-1} \psi \rangle, \qquad \langle \psi, \Sigma_{v,\mathrm{lev}}^{-1} \psi \rangle = \sup_{u \ne 0} \frac{|\langle u, \psi \rangle|^2}{\langle u, \Sigma_{v,\mathrm{lev}} u \rangle}.$$

Indeed, write the inner product as $\langle \Sigma_{v,\mathrm{lev}}^{1/2} u, \Sigma_{v,\mathrm{lev}}^{-1/2} \psi \rangle$. Equality holds for $u = \Sigma_{v,\mathrm{lev}}^{-1} \psi$ when $\psi \ne 0$. Taking the policy supremum in the weighted Cauchy–Schwarz bound and integrating gives $v_v^2 \le \Lambda_v\{\lambda_{\mathrm{lev}}; \Pi_{\mathcal{H}}(B_{\mathcal{H}})\}(e_v^2 + \lambda_{\mathrm{lev}} \|d_v\|^2)$. The variational supremum ranges over every u, not only d_v. Under the stronger comparison $\Sigma_v \succeq c_n \mathbf{U}_v \mathbf{W}_v \mathbf{U}_v$, the spectral argument in the proof of Proposition F.1 applies to every $\mathbf{U}_v u$. Combining it with $x^{\vartheta} y^{1-\vartheta} \le \lambda_{\mathrm{lev}}^{\vartheta-1}(x + \lambda_{\mathrm{lev}} y)$ yields pointwise

$$\langle \psi, \Sigma_{v,\mathrm{lev}}^{-1} \psi \rangle \le c_n^{-\vartheta} \lambda_{\mathrm{lev}}^{\vartheta-1} \|\mathbf{W}_v^{-\vartheta/2} \psi\|^2.$$

The feature-envelope bound then proves the same numerical bound for Λ_v. At $\vartheta = 1$, the spectral integrands $(t + \lambda_{\mathrm{lev}})^{-1}$ increase to t^{-1} as $\lambda_{\mathrm{lev}} \downarrow 0$; monotone convergence gives the zero-ridge bound $B_{\psi,n}^2 / c_n$ on the finite-energy domain $\mathrm{Range}(\Sigma_v^{1/2}) = \{\Sigma_v^{1/2} z : z \in \mathcal{E}_v\}$, the range of the square-root operator. None of these full-feature conditions is required by the value theorem. The second-moment operator is that of the integrated feature; averaging Jacobian second-moment operators or comparing sorted eigenvalues does not preserve this directional information.

Restricting a transfer ratio to a prescribed error family also appears for Bellman residuals in Xie et al. (2021) and for transition-model errors in Uehara & Sun (2022). The distinction here is between the coefficient direction associated with each exact critic secant and the variational supremum over the entire feature space, which explains the two information conditions in the display above.

F.3 A NONLINEAR INFORMATION VERIFICATION

Proposition F.2 (Accuracy-dependent information in a nonlinear critic family)**.** *Consider one state, an L^2 action ball, and a fixed smooth RKHS policy ball containing the zero action. Let the behavior action satisfy $|\langle A, \phi_1 \rangle| = \varsigma_A > 0$, as in the signed-series design of Appendix M. For a fixed $\kappa > 1$, set*

$$z(a) = \min\{1, |\langle a, \phi_1 \rangle| / \varsigma_A\}, \qquad F_\ell(a) = \ell\{1 - z(a)\} + \ell^\kappa z(a), \quad 0 \le \ell \le 1.$$

For exact pair features $\psi(a) = (1 - z(a), z(a))$, coefficients $d_{\ell,u} = (\ell - u, \ell^\kappa - u^\kappa)$ and smoothness-weight operator $\mathbf{W} = \mathrm{Id}_2$, the smooth-feature constant is one. There is no positive uniform directional-information constant over all pairs, but pairs with evaluation error $\upsilon_{\ell,u} > r$ satisfy the moment inequality with

$$c(r) = \frac{r^{2\kappa-2}}{(1+\kappa^2)\kappa^{2\kappa-2}}.$$

Every pair obeys $\upsilon_{\ell,u} \le \kappa \epsilon_{\ell,u}^{1/\kappa}$. Full-feature zero-ridge leverage at the zero action is infinite.

Function-class-relative coefficients in Uehara & Sun (2022); Xie et al. (2021) motivate restricting which errors require information. This construction additionally makes the information threshold depend on the desired evaluation accuracy: actual coefficient pairs obey a nonlinear relation even though their ambient feature second-moment operator is singular. It verifies the $\beta_{\mathrm{info}} > 0$ branch of Proposition F.1, complementing the infinite-dimensional spectral mechanism in Example 1.

Proof. Since $z(A) = 1$ and $z(0) = 0$,

$$\Sigma = \mathrm{diag}(0, 1), \qquad \psi(0) = (1, 0), \qquad \sup_a \|\psi(a)\|^2 \le 1.$$

For $\ell > u$, put $\Delta = \ell - u$ and $\Delta_\kappa = \ell^\kappa - u^\kappa$. Integration gives $\Delta^\kappa \le \Delta_\kappa \le \kappa\Delta$. Hence

$$\epsilon_{\ell,u} = \Delta_\kappa, \qquad \Delta \le \upsilon_{\ell,u} \le \kappa\Delta, \qquad \|d_{\ell,u}\|^2 \le (1+\kappa^2)\Delta^2.$$

If $\upsilon_{\ell,u} > r$, then $\Delta > r/\kappa$, so

$$\frac{\epsilon_{\ell,u}^2}{\|d_{\ell,u}\|^2} \ge \frac{\Delta^{2\kappa-2}}{1+\kappa^2} \ge c(r).$$

Swapping endpoints gives all other pairs. Also $\upsilon_{\ell,u} \le \kappa\Delta \le \kappa\Delta_\kappa^{1/\kappa}$. For (F_ℓ, F_0) the errors are exactly $\upsilon = \ell$ and $\epsilon = \ell^\kappa$, so $\epsilon^2/\|d\|^2 = \ell^{2\kappa-2}/(1+\ell^{2\kappa-2}) \to 0$. The ridge feature leverage at zero is $\langle \psi(0), (\Sigma + \lambda_{\text{lev}}\,\mathrm{Id})^{-1}\psi(0)\rangle = \lambda_{\text{lev}}^{-1}$. Finally $\vartheta = 1$ and $\beta_{\text{info}} = 2\kappa - 2$ yield $\vartheta_* = 1/\kappa$, agreeing with the direct bound. The family is compact in the uniform norm and uniformly action-Lipschitz, since z is ς_A^{-1}-Lipschitz and $\ell \in [0,1]$. □

F.4 From behavior prediction to policy evaluation

Regression bounds hold uniformly over deterministic classes containing every possible Bellman label and corresponding fit, allowing the realized same-data labels to be substituted (Nguyen-Tang et al., 2022). Appendix E constructs these classes. Target approximation must hold on the full state–action domain: approximation only on behavior support cannot control evaluation at policy actions outside that support. This subsection justifies the error conversion (35) used in the main examples.

For a deterministic Bellman label y, let Q_y be its conditional mean defined by the common reward–transition kernel on the whole admissible domain, including policy-selected actions outside observed support. For each label choose a deterministic exact-score approximant f_y° with $\sup_y \|f_y^\circ - Q_y\|_\infty \le \eta_{Q,n}$.

Corollary F.1 (A common regression-to-policy error bound)**.** *Suppose one event gives simultaneously for the full deterministic label class*

$$\|\widehat{f}_y - Q_y\|_{L^2(\nu_X)} \le \delta_{Q,n}, \qquad \|\widehat{f}_y - F_y\|_\infty \le e_{Q,n},$$

where F_y is the exact-score version of the fit. Let a deterministic family $\mathcal{V}_n$ contain every possible pair (F_y, f_y°) substituted on this event. Then

$$\sup_y \|\widehat{f}_y - Q_y\|_{\nu_+,\Pi_{\mathcal{H}}(B_{\mathcal{H}}),2} \le \eta_{Q,n} + e_{Q,n} + \omega_n(\delta_{Q,n} + \eta_{Q,n} + e_{Q,n}). \tag{33}$$

Every quantitative upper bound for the secant transfer envelope in (24) may replace the last term. In particular, under Assumption F.1 and $\|d_v\| \le B_{Q,n}$, the sharper bound is

$$\sup_y \|\widehat{f}_y - Q_y\|_{\nu_+,\Pi_{\mathcal{H}}(B_{\mathcal{H}}),2} \le E_n(\delta_{Q,n}, \eta_{Q,n}, e_{Q,n}, B_{Q,n}),$$
$$E_n(\delta,\eta,e,B_d) := \eta + e + B_{\psi,n} c_n^{-\vartheta/2} B_d^{1-\vartheta}(\delta+\eta+e)^\vartheta. \tag{34}$$

Under the scale-dependent condition of Proposition F.1(ii)*, another valid bound is*

$$E_n^{\text{sc}}(\delta,\eta,e,B_d) := \eta + e + \inf_{0<\varepsilon\le\varepsilon_0} \left\{\varepsilon + B_{\psi,n} c_n(\varepsilon)^{-\vartheta/2} B_d^{1-\vartheta}(\delta+\eta+e)^\vartheta\right\}.$$

At $\vartheta = 1$ omit the coefficient factor; if $B_d = 0$ use $\eta + e$ directly.

The bound combines the logged prediction error $\delta_{Q,n}$, full-domain target-approximation error $\eta_{Q,n}$, and numerical input error $e_{Q,n}$. On $\mathcal{E}_n$, it supplies the critic-update error in Assumption 3 by taking

$$\varepsilon_{Q,n} = \eta_{Q,n} + e_{Q,n} + \omega_n(\delta_{Q,n} + \eta_{Q,n} + e_{Q,n}). \tag{35}$$

In the notation of (7), these component bounds give $\delta_n = \delta_{Q,n} + \eta_{Q,n} + e_{Q,n}$ and hence $\varepsilon_{Q,n} = \eta_{Q,n} + e_{Q,n} + \omega_n(\delta_n)$. The KRR specialization instead uses its raw fit with $f_y^\circ = Q_y$, $\delta_n = \delta_{K,n}$, and exact inputs; Proposition H.3 verifies both pair membership and this prediction bound before clipping. This separates statistical prediction from its conditional-stability conversion, following the stability-versus-noise distinction in Egger & Hofmann (2018). Full-domain approximation and numerical errors enter on both sides of the secant comparison, while the label-uniform event permits their use at every same-data fitted-Q step (Nguyen-Tang et al., 2022).

Proof. Work on the simultaneous event and take $v = (F_y, f_y^\circ)$. The behavior norm of this exact-score secant is bounded by the triangle inequality:

$$\begin{aligned} e_v &\le \|F_y - \widehat{f}_y\|_{L^2(\nu_X)} + \|\widehat{f}_y - Q_y\|_{L^2(\nu_X)} + \|Q_y - f_y^\circ\|_{L^2(\nu_X)} \\ &\le e_{Q,n} + \delta_{Q,n} + \eta_{Q,n}. \end{aligned}$$

The full-domain bounds also dominate the policy-evaluation norm, so

$$\begin{aligned} \|\widehat{f}_y - Q_y\|_{\nu_+,\Pi_{\mathcal{H}}(B_{\mathcal{H}}),2} &\le \|\widehat{f}_y - F_y\|_{\nu_+,\Pi_{\mathcal{H}}(B_{\mathcal{H}}),2} \\ &\quad + \|F_y - f_y^\circ\|_{\nu_+,\Pi_{\mathcal{H}}(B_{\mathcal{H}}),2} \\ &\quad + \|f_y^\circ - Q_y\|_{\nu_+,\Pi_{\mathcal{H}}(B_{\mathcal{H}}),2} \\ &\le e_{Q,n} + \upsilon_v + \eta_{Q,n}. \end{aligned}$$

Apply the finite-accuracy bound (24) to the displayed behavior-norm bound for e_v and substitute it into the displayed policy-evaluation bound. Proposition F.1(i)–(ii) give the two quantitative alternatives with the same inputs. If $B_{Q,n} = 0$, the secant vanishes; at $\vartheta = 1$, its coefficient factor is absent. The events and bounds cover the full deterministic label class, so taking $\sup_y$ is valid. □

Prediction-rate substitution. If $\delta_{Q,n} + \eta_{Q,n} + e_{Q,n} = \widetilde{O}(N^{-b_Q})$, $b_Q > 0$, and $B_{\psi,n}$, c_n^{-1} and (when $\vartheta < 1$) $B_{Q,n}$ are bounded, (34) gives

$$\sup_y \|\widehat{f}_y - Q_y\|_{\nu_+,\Pi_{\mathcal{H}}(B_{\mathcal{H}}),2} = \widetilde{O}(N^{-\vartheta b_Q}).$$

Under (30) with $c_n(\varepsilon) \ge c_* \varepsilon^{\beta_{\mathrm{info}}}$ and bounded c_*^{-1}, replace ϑ by $\vartheta_* = 2\vartheta/(2 + \beta_{\mathrm{info}}\vartheta)$. More generally, for $\kappa_{\mathrm{tr}} \ge 0$, $C_n\{\varepsilon; \Pi_{\mathcal{H}}(B_{\mathcal{H}})\} \le C\varepsilon^{-\kappa_{\mathrm{tr}}}$ with bounded C gives exponent $2b_Q/(2 + \kappa_{\mathrm{tr}})$. These follow by substitution into (31), or by choosing ε of order $(\delta_{Q,n} + \eta_{Q,n} + e_{Q,n})^{2/(2+\kappa_{\mathrm{tr}})}$ in (25); for $\kappa_{\mathrm{tr}} = 0$ let $\varepsilon \downarrow 0$. All conditions are evaluated along the chosen iteration sequence $M = M_n$, including its full critic and label classes.

If some actual difference satisfies $e_v = 0$ but $\upsilon_v > 0$, then $C_n\{\varepsilon; \Pi_{\mathcal{H}}(B_{\mathcal{H}})\} = \infty$ for $0 < \varepsilon < \upsilon_v$. Such a fixed identification failure is retained by the theorem, not removed by a smoothing penalty or by taking a proof-only ridge parameter positive. Appendix M verifies the information condition for the full linear critic kernel of Example 1.

Under the deterministic-class convention of Appendix E, each realized label is substituted into the same uniform bound on the fixed smooth class $\Pi_{\mathcal{H}}(B_{\mathcal{H}})$.

G General finite-sample value guarantee and proof

Theorem 1 is the stable-propagation specialization of the following finite-iteration statement.

Theorem G.1 (General finite-iteration value bound)**.** *Under the data and measurability conditions of Section 2 and Assumption A.1, suppose the uniform critic and policy error bounds in Assumption 3 hold on $\mathcal{E}_n$ from Section 4.1. This event also places all fitted critics in $\mathcal{Q}_n$ and all returned policies in $\Pi_{\mathcal{H}}(B_{\mathcal{H}})$. Under Assumption 4, put*

$$\chi = \gamma C_B, \qquad G_M(\chi) = \sum_{j=0}^{M-1} \chi^j.$$

On $\mathcal{E}_n$, for the prescribed finite $M \ge 1$,

$$J^\star - J(\widehat{\pi}_M) \le a_{\mathcal{H}} + \frac{C_{\mathrm{occ}}(\widehat{\pi}_M)}{1-\gamma}\{2V_{\max}\chi^M + 2G_M(\chi)\varepsilon_{Q,n} + (1 + 2\chi G_M(\chi))\varepsilon_{\pi,n}\}. \tag{36}$$

The critic bound $\varepsilon_{Q,n}$ can be supplied by (35) under the verified secant conditions. The bound holds for every finite χ, with probability at least $\Pr(\mathcal{E}_n)$.

Beyond $a_{\mathcal{H}}$, the terms in (36) are finite-iteration error, accumulated critic-update error and policy-improvement error. Stable propagation bounds their accumulation uniformly in M.

Proof. We first prove the value bound relative to $\Pi_{\mathcal{H}}(B_{\mathcal{H}})$ and then add the approximation gap $a_{\mathcal{H}} = J^\star - J^\star_{\mathcal{H}}$. On $\mathcal{E}_n$, put

$$E^Q_m = \|Q^\dagger - \widehat{Q}_m\|_{\nu_+,\Pi_{\mathcal{H}}(B_{\mathcal{H}}),2}, \qquad h_m(s) = \sup_{\pi\in\Pi_{\mathcal{H}}(B_{\mathcal{H}})} |(Q^\dagger - \widehat{Q}_m)\{s,\pi(s)\}|.$$

Allow nonuniform error bounds $\varepsilon_{Q,m}$ and $\varepsilon_{\pi,m}$ in the argument. The Bellman decomposition is

$$\begin{aligned} Q^\dagger - \widehat{Q}_m = \{\mathbb{B}Q^\dagger - \mathbb{B}\widehat{Q}_{m-1}\} + \{\mathbb{B}\widehat{Q}_{m-1} - \mathbb{B}^{\widehat{\pi}_{m-1}}\widehat{Q}_{m-1}\} \\ + \{\mathbb{B}^{\widehat{\pi}_{m-1}}\widehat{Q}_{m-1} - \widehat{Q}_m\}. \end{aligned}$$

The Bellman-optimality difference is bounded by $\gamma P h_{m-1}$, and the greedy-policy difference is bounded by $\gamma P g_{m-1}$. Minkowski and the one-step state condition give, with $\chi = \gamma C_B$,

$$E^Q_m \le \chi E^Q_{m-1} + \varepsilon_{Q,m} + \chi\varepsilon_{\pi,m-1}.$$

Induction yields

$$E^Q_M \le \chi^M E^Q_0 + \sum_{m=1}^{M} \chi^{M-m}\varepsilon_{Q,m} + \chi \sum_{m=0}^{M-1} \chi^{M-1-m}\varepsilon_{\pi,m}.$$

Clipping and probability normalization give $E^Q_0 \le V_{\max}$.

For any clipped Q and $\pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})$, adding and subtracting Q gives

$$g_{Q^\dagger,\pi}(s) \le g_{Q,\pi}(s) + 2 \sup_{\pi'\in\Pi_{\mathcal{H}}(B_{\mathcal{H}})} |(Q^\dagger - Q)\{s,\pi'(s)\}|.$$

Apply Lemma B.1, which gives $J^\star_{\mathcal{H}} \le \mathbb{E}_\rho V^\dagger$ by a supremum over the localized class. The occupancy condition for the two displayed residuals yields

$$J^\star_{\mathcal{H}} - J(\pi) \le \frac{C_{\mathrm{occ}}(\pi)}{1-\gamma}\{\|g_{Q,\pi}\|_{L^2(\nu_+)} + 2\|Q^\dagger - Q\|_{\nu_+,\Pi_{\mathcal{H}}(B_{\mathcal{H}}),2}\}.$$

Set $(Q,\pi) = (\widehat{Q}_M, \widehat{\pi}_M)$ and substitute the recursion:

$$\begin{aligned} J^\star_{\mathcal{H}} - J(\widehat{\pi}_M) \le \frac{C_{\mathrm{occ}}(\widehat{\pi}_M)}{1-\gamma}\Bigg\{\varepsilon_{\pi,M} + 2\chi^M E^Q_0 + 2\sum_{m=1}^{M}\chi^{M-m}\varepsilon_{Q,m} \\ + 2\chi\sum_{m=0}^{M-1}\chi^{M-1-m}\varepsilon_{\pi,m}\Bigg\}. \end{aligned}$$

Uniform bounds and the identity $J^\star - J(\widehat{\pi}_M) = a_{\mathcal{H}} + J^\star_{\mathcal{H}} - J(\widehat{\pi}_M)$ give precisely (36). No policy attaining $J^\star_{\mathcal{H}}$ has been selected in the proof, and the only comparison with the base class Π is the explicit approximation gap $a_{\mathcal{H}}$. All events are intersected once; their dependence is unrestricted and their failure probabilities can be added. The proof is valid for every finite M. If $\chi = \gamma C_B$ is bounded above by a fixed $\chi_0 < 1$, independent of N and M, then $G_M(\chi) \le (1-\chi_0)^{-1}$ and the initialization term is bounded by $V_{\max}\chi^M = V_{\max}(\gamma C_B)^M$. With fixed γ and uniformly bounded occupancy constants, this gives (10) and proves Theorem 1. □

Direct policy errors. Any bound on the actual $\Pi_{\mathcal{H}}(B_{\mathcal{H}})$ defect may supply $\varepsilon_{\pi,m}$. In particular, a product policy class with separated statewise optimization can use Appendix M's direct bound. For a coupled policy class, the policy-compatibility gap in Appendix D remains in the objective route.

H KRR PREDICTION AND MOMENT-BASED EVALUATION

We first record kernel feature calculations, then establish uniform prediction, fitted-norm and conditional evaluation bounds. Kernel input regularity controls covering and numerical errors; evaluation additionally requires the directional information and smooth-feature conditions. All statements use

the full deterministic label and prediction classes. Let $\mathcal{F}$ be a separable RKHS on $\mathcal{S} \times \mathcal{A}$, with measurable kernel, $\sup_x K(x,x) \le \kappa_K^2$. The kernel section $K_x = K(\cdot, x) \in \mathcal{F}$ represents evaluation through $f(x) = \langle f, K_x \rangle_{\mathcal{F}}$ (the reproducing property). Assume

$$\|K_{(s,a)} - K_{(s,a')}\|_{\mathcal{F}} \le \mathrm{Lip}_K \|a - a'\|_{L^2}. \tag{37}$$

The second-moment operator is $\Sigma_K = \int K_x \otimes K_x \, \nu_X(dx)$ and its empirical version is $\widehat{\Sigma}_K = n^{-1} \sum_{i,t} K_{X_{i,t}} \otimes K_{X_{i,t}}$. For each deterministic label y, let $\widetilde{f}_{\varrho,y}$ be an approximate minimizer of

$$\widehat{\Psi}_{\varrho,y}(f) = n^{-1} \sum_{i,t} \{Y_{i,t}(y) - f(X_{i,t})\}^2 + \varrho \|f\|_{\mathcal{F}}^2,$$

with $\varrho > 0$ and objective gap at most $\tau_K \ge 0$. FQI returns $\widehat{f}_y = \mathrm{clip}_{[0, V_{\max}]} \widetilde{f}_{\varrho,y}$.

Proposition H.1 (Functional kernel secants and a smoothness verification)**.** *For an RKHS $\mathcal{F}$, the un-clipped difference $f - g$ has $\mathcal{E} = \mathcal{F}$, $d = f - g$, and $\psi(s,a) = K_{(s,a)}$. Thus the same moment theorem applies with a deterministic spectral weight operator $\mathbf{W}_K$ whenever its quadratic-form and smooth-feature conditions are verified.*

A direct infinite-dimensional example uses finitely many states and a kernel length scale $\ell_K > 0$:

$$K\{(s,a),(s',a')\} = \mathbf{1}\{s = s'\}(1 + \ell_K^{-2} \langle a, a' \rangle_{L^2})^m, \qquad m \in \mathbb{N}.$$

For each state use the ambient direct sum of symmetric tensor powers $\{L^2([0,1])\}^{\odot k}$, $0 \le k \le m$. Here $\{L^2([0,1])\}^{\odot k}$ is the subspace of the k-fold Hilbert tensor product unchanged by permutations of its factors, and $a^{\otimes k} = a \otimes \cdots \otimes a$. The conventions $\{L^2([0,1])\}^{\odot 0} = \mathbb{R}$, $a^{\otimes 0} = 1$ and $\mathbf{W}_\alpha^{\otimes 0} = 1$ give the intercept. Direct sums, denoted by $\bigoplus$, stack orthogonal blocks, so their squared norms add. RKHS coefficients are isometrically identified with the closed feature span, while the spectral weight operator acts on the ambient space; invariance of that span is not required. The features are

$$\psi_m(a) = \bigoplus_{k=0}^{m} \binom{m}{k}^{1/2} \ell_K^{-k} a^{\otimes k}, \qquad \mathbf{W}_K = \bigoplus_s \bigoplus_{k=0}^{m} \mathbf{W}_\alpha^{\otimes k}.$$

Let $\mathbf{U}_K$ denote the projection onto that closed feature span. For $\vartheta = \min\{1, \zeta/\alpha\}$ and $\|a\|_\zeta \le B_{\Pi,\zeta}$,

$$\|\mathbf{W}_K^{-\vartheta/2} K_{(s,a)}\|^2 \le (1 + \ell_K^{-2} B_{\Pi,\zeta}^2)^m.$$

For a uniform policy radius $B_{\Pi,\zeta}$, this gives $B_{\psi,n}^2 = (1 + \ell_K^{-2} B_{\Pi,\zeta}^2)^m$. For only the integrated radius (4), the full functional linear kernel ($m = 1$) still has $B_{\psi,n}^2 = 1 + \ell_K^{-2} B_{\Pi,\zeta}^2$. For $m > 1$, also retain a deterministic uniform bound $\sup_{s,\pi} \|\pi(s)\|_\zeta \le \overline{B}_{\Pi,\zeta}$; then a valid integrated factor is

$$B_{\psi,n}^2 = (1 + \ell_K^{-2} \overline{B}_{\Pi,\zeta}^2)^{m-1} (1 + \ell_K^{-2} B_{\Pi,\zeta}^2).$$

Alternatively, directly bound the integrated $2m$-th policy-smoothness moment. These are not finite-dimensional action discretizations: infinitely many action coordinates enter the kernel.

Functional-input kernels such as the linear and Gaussian examples in Wang et al. (2026, Remark 2) encode functions directly through Hilbert-space geometry. The tensor calculation here identifies how functional-policy regularity controls the feature energy required by secant transfer, with infinitely many action coordinates retained in the kernel.

Proof. For $d = f - g \in \mathcal{F}$, the reproducing property gives

$$(f - g)(s,a) = \langle d, K_{(s,a)} \rangle_{\mathcal{F}},$$
$$\|f - g\|_{L^2(\nu_X)}^2 = \mathbb{E}_{\nu_X} \langle d, K_X \rangle^2 = \langle d, \mathbb{E}_{\nu_X}[K_X \otimes K_X] d \rangle = \langle d, \Sigma_K d \rangle.$$

For polynomial features, the tensor inner-product identity and binomial theorem yield

$$\langle a^{\otimes k}, (a')^{\otimes k} \rangle = \prod_{q=1}^{k} \langle a, a' \rangle = \langle a, a' \rangle^k,$$

$$\langle \psi_m(a), \psi_m(a') \rangle = \sum_{k=0}^{m} \binom{m}{k} \ell_K^{-2k} \langle a, a' \rangle^k = (1 + \ell_K^{-2} \langle a, a' \rangle)^m,$$

$$\langle e_s \otimes \psi_m(a), e_{s'} \otimes \psi_m(a') \rangle = \mathbf{1}\{s = s'\}(1 + \ell_K^{-2} \langle a, a' \rangle)^m.$$

The coefficient lies in the closed feature span. Writing $a_j = \langle a, \phi_j \rangle$, spectral calculus gives

$$\begin{aligned}
\|(\mathbf{W}_\alpha^{\otimes k})^{-\vartheta/2} a^{\otimes k}\|^2 &= \sum_{j_1,\dots,j_k \geq 1} \prod_{q=1}^{k} j_q^{2\alpha\vartheta} |a_{j_q}|^2 \\
&= \left(\sum_{j \geq 1} j^{2\alpha\vartheta} |a_j|^2 \right)^k = \|a\|_{\alpha\vartheta}^{2k}, \\
\|\mathbf{W}_K^{-\vartheta/2} \psi_m(a)\|^2 &= \sum_{k=0}^{m} \binom{m}{k} \ell_K^{-2k} \|a\|_{\alpha\vartheta}^{2k} \\
&= (1 + \ell_K^{-2} \|a\|_{\alpha\vartheta}^2)^m \leq (1 + \ell_K^{-2} \|a\|_\zeta^2)^m \qquad (\alpha\vartheta \leq \zeta).
\end{aligned}$$

The infinite sums follow from finite coordinate truncations by monotone convergence. Thus $\|a\|_\zeta \leq B_{\Pi,\zeta}$ gives the pointwise bound $(1 + \ell_K^{-2} B_{\Pi,\zeta}^2)^m$.

For an integrated radius, set $x(s) = \ell_K^{-2} \sup_{\pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})} \|\pi(s)\|_\zeta^2$. Then

$$\begin{aligned}
\int x \, d\nu_+ &\leq \ell_K^{-2} B_{\Pi,\zeta}^2, \\
\int (1 + x) \, d\nu_+ &\leq 1 + \ell_K^{-2} B_{\Pi,\zeta}^2 && (m = 1), \\
\int (1 + x)^m \, d\nu_+ &\leq (1 + \bar{x})^{m-1} \int (1 + x) \, d\nu_+ && (m > 1,\ x \leq \bar{x}), \\
&\leq (1 + \ell_K^{-2} \overline{B}_{\Pi,\zeta}^2)^{m-1} (1 + \ell_K^{-2} B_{\Pi,\zeta}^2) && (\bar{x} = \ell_K^{-2} \overline{B}_{\Pi,\zeta}^2).
\end{aligned}$$

The feature second-moment comparison remains a separate condition; no invariance of the feature span under $\mathbf{W}_K$ is used. □

For the full functional-distance Gaussian kernel, the following Lipschitz calculation controls entropy and numerical errors, not the spectral feature condition. Its explicit secant rate still requires a compatible $\mathbf{W}_K$ and a feature second-moment comparison; otherwise (33) retains the amplification at the prescribed accuracy. A Gaussian kernel imposes no Gaussian distribution on observed actions.

Proposition H.2 (Kernel input regularity for entropy and numerical errors)**.** *For $f, g \in \mathcal{F}$, condition* (37) *implies*

$$\mathrm{Lip}_a(f - g) \leq \mathrm{Lip}_K \|f - g\|_{\mathcal{F}}, \qquad \mathrm{Lip}_a(\mathrm{clip}\, f) \leq \mathrm{Lip}_K \|f\|_{\mathcal{F}}.$$

The full functional-distance Gaussian kernel

$$K\{(s,a),(s',a')\} = k_S(s,s') \exp\{-\|a - a'\|_{L^2}^2 / (2\ell_K^2)\}, \qquad \sup_s k_S(s,s) \leq \kappa_S^2,$$

satisfies (37) *with* $\mathrm{Lip}_K = \kappa_S / \ell_K$. *Adding an action-independent intercept leaves this constant unchanged; composing with a bounded linear score map multiplies it by that map's norm.*

Kernel-section regularity also controls functional-input reconstruction errors in Wang et al. (2026, Assumption 6). Here the same calculation supplies the action-Lipschitz constants needed for joint critic–policy covers and numerical objective comparisons, while the preceding proposition addresses the different spectral energy used for evaluation.

Proof. The reproducing property and Cauchy–Schwarz give

$$\begin{aligned}
|(f - g)(s,a) - (f - g)(s,a')| &= |\langle f - g, K_{(s,a)} - K_{(s,a')} \rangle| \\
&\leq \|f - g\|_{\mathcal{F}} \|K_{(s,a)} - K_{(s,a')}\|_{\mathcal{F}} \\
&\leq \mathrm{Lip}_K \|f - g\|_{\mathcal{F}} \|a - a'\|_{L^2}, \\
|\,\mathrm{clip}\, f(s,a) - \mathrm{clip}\, f(s,a')| &\leq |f(s,a) - f(s,a')| \qquad \text{(projection)} \\
&\leq \mathrm{Lip}_K \|f\|_{\mathcal{F}} \|a - a'\|_{L^2}.
\end{aligned}$$

For the Gaussian kernel, put $x = \|a - a'\|_{L^2}^2/(2\ell_K^2)$. Then

$$\begin{aligned}\|K_{(s,a)} - K_{(s,a')}\|_{\mathcal{F}}^2 &= K((s,a),(s,a)) + K((s,a'),(s,a')) \\ &\quad - 2K((s,a),(s,a')) \\ &= 2k_S(s,s)(1-e^{-x}) = 2k_S(s,s)\int_0^x e^{-t}\,dt \\ &\le 2\kappa_S^2 x = \kappa_S^2\ell_K^{-2}\|a-a'\|_{L^2}^2, \\ \|K_{(s,a)} - K_{(s,a')}\|_{\mathcal{F}} &\le (\kappa_S/\ell_K)\|a-a'\|_{L^2}.\end{aligned}$$

An action-independent intercept contributes $c(s) + c(s) - 2c(s) = 0$. For a bounded linear score map Score, the same calculation gives

$$\|K_{(s,a)} - K_{(s,a')}\|_{\mathcal{F}} \le (\kappa_S/\ell_K)\|\operatorname{Score}(a-a')\| \le (\kappa_S/\ell_K)\|\operatorname{Score}\|_{\mathrm{op}}\|a-a'\|_{L^2}.$$

□

H.1 Uniform prediction and a controlled fitted norm

The cluster bound controls the score and second-moment operator uniformly over the label class. Suppose

$$Q_y \in \mathcal{F}, \quad \sup_y \|Q_y\|_{\mathcal{F}} \le B_F, \quad 0 \le Y(y), Q_y \le V_{\max}. \tag{38}$$

This is a regression approximation condition, not an identification condition of the general FQI theorem. Let $H_Y(\eta) = \log \operatorname{Cov}(\eta, \mathcal{Y}_n, \|\cdot\|_\infty)$, with net centers in the label class, and use the independent-subject scale N. For $0 < \delta < 1$ define

$$s_{K,n} = 2\kappa_K\eta + C\kappa_K V_{\max}\sqrt{\{H_Y(\eta) + \log(8/\delta) + 1\}/N}, \tag{39}$$

$$v_{K,n} = C\kappa_K^2\sqrt{\{\log(8/\delta) + 1\}/N}, \tag{40}$$

$$B_{K,n} = B_F + s_{K,n}/\varrho + \sqrt{\tau_K/\varrho}, \tag{41}$$

$$\delta_{K,n}^2 = \varrho B_F^2 + 2(B_{K,n} + B_F)s_{K,n} + \tau_K + v_{K,n}(B_{K,n} + B_F)^2. \tag{42}$$

The universal constant C is chosen large enough for the following simultaneous event.

Proposition H.3 (Uniform KRR prediction, fitted norm and conditional evaluation)**.** (i) Prediction and fitted norm. *Under Assumptions A.1 and D.1, (38), finite $H_Y(\eta)$, and minimization of $\widehat{\Psi}_{\varrho,y}$ to objective gap τ_K, with probability at least $1-\delta$,*

$$\sup_y \|\widetilde{f}_{\varrho,y}\|_{\mathcal{F}} \le B_{K,n}, \qquad \sup_y \left\|\widetilde{f}_{\varrho,y} - Q_y\right\|_{L^2(\nu_X)} \le \delta_{K,n}, \qquad \sup_y \left\|\widehat{f}_y - Q_y\right\|_{L^2(\nu_X)} \le \delta_{K,n}.$$

The probability refers to the full deterministic label class and hence permits all same-data FQI iterations.

(ii) Evaluation at policy-selected actions. *Fix before fitting a deterministic family containing all raw fitted-critic/Bellman-target pairs covered by the event in part (i). Their clipped error at policy-selected actions is at most $\omega_n(\delta_{K,n})$. If the actual raw kernel secants satisfy Assumption F.1, then*

$$\sup_y \|\widehat{f}_y - Q_y\|_{\nu_+,\Pi_{\mathcal{H}}(B_{\mathcal{H}}),2} \le B_{\psi,n}c_n^{-\vartheta/2}(B_{K,n} + B_F)^{1-\vartheta}\delta_{K,n}^{\vartheta}.$$

Under Proposition F.1(ii), replace the right-hand side by the bound (31) at $\delta = \delta_{K,n}$ and $B_{Q,n} = B_{K,n} + B_F$. The raw prediction radius is proved before clipping; neither evaluation accuracy nor a proof-only ridge changes the KRR tuning parameter.

The fit-and-target uniformity of Nguyen-Tang et al. (2022) is combined here with RKHS ridge fitting: a single subject-level score and operator event controls prediction and the raw fitted norm for every Bellman label. The norm bound supplies the coefficient radius needed for the subsequent regression-to-policy-evaluation transfer, linking the regression calculation to the functional-action geometry.

Proof. First control the score uniformly over labels. Put $\epsilon_{i,t,y} = Y_{i,t,y} - Q_y(X_{i,t})$ and $\mathbf{z}_{i,y}^{\mathrm{sub}} = (N/n)\sum_t \epsilon_{i,t,y} K_{X_{i,t}}$, so that $\mathbf{z}_y = N^{-1}\sum_i \mathbf{z}_{i,y}^{\mathrm{sub}}$. Conditional compatibility gives $\mathbb{E}\mathbf{z}_{i,y}^{\mathrm{sub}} = 0$, and $\|\mathbf{z}_{i,y}^{\mathrm{sub}}\| \le B_{\mathrm{sc}} := \kappa_K V_{\max}$. Independence across subjects, followed by the bounded-differences inequality (Boucheron et al., 2013), yields

$$\mathbb{E}\|\mathbf{z}_y\|^2 \le B_{\mathrm{sc}}^2/N,$$
$$\Pr\Big\{\|\mathbf{z}_y\| > B_{\mathrm{sc}}/\sqrt{N} + B_{\mathrm{sc}}\sqrt{2x/N}\Big\} \le e^{-x}.$$

Indeed, replacing one subject changes the norm by at most $2B_{\mathrm{sc}}/N$. Use $x = H_Y(\eta) + \log(8/\delta)$ and a union bound on a deterministic η-label net. The common kernel implies $\|Q_y - Q_{y'}\|_\infty \le \|y - y'\|_\infty$, whence $\|\mathbf{z}_y - \mathbf{z}_{y'}\| \le 2\kappa_K\eta$ for a net representative. Thus $\sup_y \|\mathbf{z}_y\| \le s_{K,n}$ on the stated score event.

Next write $\widehat{\Sigma}_K = N^{-1}\sum_i \mathbf{C}_i$, where $\mathbf{C}_i = (N/n)\sum_t K_{X_{i,t}} \otimes K_{X_{i,t}}$. The identity $\|K_x \otimes K_x\|_{\mathrm{HS}} = \|K_x\|^2$ gives $\|\mathbf{C}_i\|_{\mathrm{HS}} \le \kappa_{\mathrm{cov}} := \kappa_K^2$. The independent centered Hilbert–Schmidt sums satisfy $\mathbb{E}\|\widehat{\Sigma}_K - \Sigma_K\|_{\mathrm{HS}}^2 \le \kappa_{\mathrm{cov}}^2/N$, and replacement of one subject changes this norm by at most $2\kappa_{\mathrm{cov}}/N$. The preceding bounded-differences calculation with $x = \log(8/\delta)$ gives

$$\|\widehat{\Sigma}_K - \Sigma_K\|_{\mathrm{op}} \le \|\widehat{\Sigma}_K - \Sigma_K\|_{\mathrm{HS}} \le v_{K,n}.$$

Intersect the score and second-moment events. Their total failure is at most δ.

For the exact ridge minimizer, the normal equation and spectral calculus give

$$\widehat{f}_{\varrho,y} = (\widehat{\Sigma}_K + \varrho\,\mathrm{Id})^{-1}\widehat{\Sigma}_K Q_y + (\widehat{\Sigma}_K + \varrho\,\mathrm{Id})^{-1}\mathbf{z}_y,$$
$$\|\widehat{f}_{\varrho,y}\| \le B_F + s_{K,n}/\varrho, \qquad \|\widetilde{f}_{\varrho,y} - \widehat{f}_{\varrho,y}\|^2 \le \tau_K/\varrho.$$

Here $(\widehat{\Sigma}_K + \varrho\,\mathrm{Id})^{-1}\widehat{\Sigma}_K$ has norm at most one; the bound for $\|\widetilde{f}_{\varrho,y} - \widehat{f}_{\varrho,y}\|^2$ follows by completing the square in the ridge objective, whose excess is at least $\varrho\|\widetilde{f}_{\varrho,y} - \widehat{f}_{\varrho,y}\|^2$. Consequently $\|\widetilde{f}_{\varrho,y}\| \le B_{K,n}$ uniformly in y.

Finally, set $e_y = \widetilde{f}_{\varrho,y} - Q_y$. Comparing the approximate ridge fit with Q_y and discarding the nonnegative fitted penalty gives

$$\begin{aligned}\langle e_y, \widehat{\Sigma}_K e_y\rangle &\le 2\langle e_y, \mathbf{z}_y\rangle + \varrho B_F^2 + \tau_K,\\ \|e_y\|_{L^2(\nu_X)}^2 &= \langle e_y, \widehat{\Sigma}_K e_y\rangle + \langle e_y, (\Sigma_K - \widehat{\Sigma}_K)e_y\rangle\\ &\le 2(B_{K,n} + B_F)s_{K,n} + \varrho B_F^2 + \tau_K + v_{K,n}(B_{K,n} + B_F)^2 = \delta_{K,n}^2.\end{aligned}$$

We used $\|e_y\| \le B_{K,n} + B_F$. Since $Q_y \in [0, V_{\max}]$, $|\operatorname{clip}(\widetilde{f}_{\varrho,y}) - Q_y| \le |e_y|$ pointwise. This proves the clipped prediction bound on the same full-label event.

For evaluation, the same raw difference $e_y = \widetilde{f}_{\varrho,y} - Q_y$ satisfies $\|e_y\|_{L^2(\nu_X)} \le \delta_{K,n}$ and $\|e_y\|_{\mathcal{F}} \le B_{K,n} + B_F$. The pointwise clipping inequality already proved gives

$$\sup_y \|\widehat{f}_y - Q_y\|_{\nu_+,\Pi_{\mathcal{H}}(B_{\mathcal{H}}),2} \le \sup_y \|e_y\|_{\nu_+,\Pi_{\mathcal{H}}(B_{\mathcal{H}}),2} \le \omega_n(\delta_{K,n}).$$

The last inequality is (25) on the deterministic raw-secant family. Substitution of these prediction and coefficient radii into the uniform or accuracy-dependent part of Proposition F.1 proves the two quantitative alternatives. No clipped critic is treated as an RKHS element, and no new probability event is introduced.

□

H.2 Ridge calibration and explicit prediction rates

Corollary H.1 (Boundary and interior ridge calibration)**.** *Retain the full-domain target ball and the simultaneous score and operator event of Proposition H.3, with bounded target norm and kernel envelope. Use the radii* (41)–(42)*.*

(i) Bounded ratios, including boundary tuning. *If* $\varrho_n \to 0$, $s_{K,n} + v_{K,n} = O(\varrho_n)$ *and* $\tau_{K,n} = O(\varrho_n)$, *then* $B_{K,n} = O(1)$ *and* $\delta_{K,n} = O(\varrho_n^{1/2})$. *Under bounded raw-secant information and smooth-feature constants, evaluation is* $O(\varrho_n^{\bar{\vartheta}/2})$, *where* $\bar{\vartheta} = \vartheta$ *for uniform information and* $\bar{\vartheta} = 2\vartheta/(2 + \beta_{\mathrm{info}}\vartheta)$ *when* $c_n(\varepsilon) \geq c_* \varepsilon^{\beta_{\mathrm{info}}}$ *on a fixed interval, with fixed* $c_* > 0$.

(ii) Vanishing ratios under interior tuning. *Suppose in addition that* Lip_K *is bounded,* $N \to \infty$, *and the full label entropy at a polynomially small* η_n *satisfies* $H_Y(\eta_n) = \widetilde{O}(N^{p_Y^{\mathrm{ent}}})$ *for* $0 \leq p_Y^{\mathrm{ent}} < 1$. *For polynomially small confidence levels, take*

$$\varrho_n = N^{-p_\varrho}, \qquad 0 < p_\varrho < (1 - p_Y^{\mathrm{ent}})/2, \qquad \eta_n = o(\varrho_n), \quad \tau_{K,n} = o(\varrho_n).$$

Then $B_{K,n} = B_F + o(1)$, $\delta_{K,n} = \widetilde{O}(\varrho_n^{1/2})$, *and under bounded uniform-information transfer constants the evaluation error is* $\widetilde{O}_p(\varrho_n^{\vartheta/2})$.

Classical spectral KRR rates track effective dimension as ridge decreases (Caponnetto & De Vito, 2007), with the effective-dimension bound corrected by Sutherland (2017). The present sufficient calibration instead tracks the full-label score and operator bounds, making explicit the ridge scale that keeps all fitted secants inside one bounded coefficient ball.

Proof. The radius expressions can be written as

$$B_{K,n} = B_F + \frac{s_{K,n}}{\varrho_n} + \sqrt{\frac{\tau_{K,n}}{\varrho_n}},$$
$$\frac{\delta_{K,n}^2}{\varrho_n} = B_F^2 + 2(B_{K,n} + B_F)\frac{s_{K,n}}{\varrho_n} + \frac{\tau_{K,n}}{\varrho_n} + (B_{K,n} + B_F)^2 \frac{v_{K,n}}{\varrho_n}. \tag{43}$$

For part (i), every displayed ratio is bounded. Equation (43) directly bounds both $B_{K,n}$ and $\delta_{K,n}^2/\varrho_n$. The conditional evaluation part of Proposition H.3 then gives $O(\varrho_n^{\bar{\vartheta}/2})$; it needs boundedness of $B_{K,n}$, not convergence to B_F.

For part (ii), put $p_\varrho^+ = (1 - p_Y^{\mathrm{ent}})/2 > p_\varrho$. The score and second-moment bounds give

$$s_{K,n} = O(\eta_n) + \widetilde{O}(N^{-p_\varrho^+}), \qquad v_{K,n} = \widetilde{O}(N^{-1/2}).$$

Because $p_\varrho < p_\varrho^+ \leq 1/2$, these powers dominate the logarithms. Thus $s_{K,n}/\varrho_n, v_{K,n}/\varrho_n, \tau_{K,n}/\varrho_n \to 0$. The radius identity in (43) gives $B_{K,n} = B_F + o(1)$, and its prediction identity gives $\delta_{K,n}^2/\varrho_n = B_F^2 + o(1)$. Apply the same raw-secant and clipping bound on events of probability $1 - o(1)$. □

For example, if the entropy of the full label class fixed before fitting at the chosen polynomial net resolution is $\widetilde{O}(N^{p_Y^{\mathrm{ent}}})$, $0 \leq p_Y^{\mathrm{ent}} < 1$, choose a deterministic logarithmic majorant $\kappa_{\log,n}$ and

$$\varrho_n = \kappa_{\log,n} N^{-(1-p_Y^{\mathrm{ent}})/2}, \qquad s_{K,n} + v_{K,n} \leq C\varrho_n, \qquad \tau_{K,n} \leq C\varrho_n.$$

This gives $\delta_{K,n} = \widetilde{O}(N^{-(1-p_Y^{\mathrm{ent}})/4})$. The majorant must dominate the actual entropy and confidence factors. The label cover is evaluated at the unconditional pre-fit radius $\{(V_{\max}^2 + \tau_{K,n})/\varrho_n\}^{1/2}$, not at the sharper fitted radius subsequently proved on the event. Thus this example is a compatible deterministic tuning condition, not an implicit equation solved with a realized entropy or prediction error.

The deterministic class used to define the labels cannot be chosen retrospectively as the observed small-norm fitted path. Comparing the empirical objective with zero gives the unconditional radius $\|\widetilde{f}_{\varrho,y}\|_{\mathcal{F}} \leq \{(V_{\max}^2 + \tau_K)/\varrho\}^{1/2}$. Its clipped ball contains all iterates. The corresponding policy-indexed label class defines H_Y in Section H.1. The sharper radius $B_{K,n}$ is then derived on the simultaneous event of Proposition H.3.

For general policy classes, the full label entropy is supplied by the product cover in Appendix E; it includes the actual RKHS policy cover. A finite-dimensional parameter count applies only when that is the declared implementation class. A full functional-distance Gaussian kernel needs its own

full-domain entropy and information verification. Example 1 instead provides a two-dimensional continuation-label cover for the full functional linear kernel.

The full-sample mixing extension is Proposition L.2; the retained-block alternative retains its omission cost in Appendix L.4.

H.3 An explicit KRR value rate

Retain the bounded target RKHS ball, uniformly bounded kernel, and exact functional inputs of Corollary H.1. Let $p_K \in [0,1)$ be an exponent such that the full deterministic Bellman-label and policy-objective log covering numbers are $\widetilde{O}(N^{p_K})$ at uniform-norm resolutions $O(\varrho_n)$. These covers must contain all possible fits and labels before the data are observed, using the unconditional fitted radius described above. For a sufficiently large deterministic logarithmic factor $\kappa_{\log,n}$, choose

$$\varrho_n = \kappa_{\log,n} N^{-(1-p_K)/2}.$$

The factor must dominate the actual covering and confidence logarithms so that $s_{K,n} + v_{K,n} = O(\varrho_n)$ on an event with probability tending to one. Suppose the KRR regularized empirical squared-loss objective gap satisfies $\tau_{K,n} = O(\varrho_n)$, the policy optimization error satisfies $\tau_{\pi,n} = O(\varrho_n)$, and $\lambda_{\pi,n} = O(\varrho_n)$. Require fixed-radius policy localization separately, as in (16).

Let $\Delta_n = \sup_{Q \in \mathcal{Q}_n} \Delta(Q)$, using the clipped deterministic fitted class. Assume the raw fit/target differences satisfy the uniform or accuracy-dependent transfer conditions of Proposition H.3, with bounded factors and exponent ϑ_* from Section 3. Under the bounded occupancy and stable state-propagation conditions of Theorem 1, for $M \geq c \log N$ with sufficiently large fixed c,

$$J^\star - J(\widehat{\pi}_M) \leq a_{\mathcal{H}} + \widetilde{O}_p\left(N^{-\vartheta_*(1-p_K)/4} + \sqrt{\Delta_n}\right).$$

This is the independent-subject version of the rate in Example 2; its mixing version follows below.

Proof. The boundary calibration gives $B_{K,n} = O(1)$ and $\delta_{K,n} = O(\varrho_n^{1/2})$. The raw-secant transfer bound and clipping give $\varepsilon_{Q,n} = O(\varrho_n^{\vartheta_*/2})$. Applying (17) to the deterministic policy-objective cover yields $u_n = O(\varrho_n)$, since both its net resolution and its concentration remainder have this order. Equation (14) then gives

$$\varepsilon_{\pi,n} = O(\varrho_n^{1/2} + \sqrt{\Delta_n}).$$

As $0 < \vartheta_* \leq 1$, the critic term dominates $\varrho_n^{1/2}$. The iteration term in Theorem 1 is no larger for a sufficiently large constant c, proving the result. These uniform events, the stated computational guarantees, and $\mathcal{R}_n$ imply $\mathcal{E}_n$ before substituting the same-data iterates. □

Mixing version. Under Assumption 5, use the block length $b = \min\{T, \lceil c_b \log(eNT) \rceil\}$ for sufficiently large fixed $c_b > 0$ and effective sample size $\mathcal{N}_b = NT/b$, as justified in Appendix L.5. Proposition L.2 supplies the same score and second-moment bounds at scale $\mathcal{N}_b$, while (54) supplies the policy-objective bound. Replace N by $\mathcal{N}_b$ in the preceding deterministic covers, ridge coefficient, and computational tolerances, and take $M \geq c \log(NT)$ with the covers valid for the actual iteration sequence. In particular,

$$\varrho_n = \kappa_{\log,n} \mathcal{N}_b^{-(1-p_K)/2} = \kappa_{\log,n} b^{(1-p_K)/2} (NT)^{-(1-p_K)/2},$$

where $\kappa_{\log,n}$ dominates the actual covering and confidence logarithms. Thus the ridge calibration for Example 2 also includes the block factor $b^{(1-p_K)/2}$. The prediction and critic-update bounds are

$$\delta_{K,n} = \widetilde{O}((NT)^{-(1-p_K)/4}), \qquad \varepsilon_{Q,n} = \widetilde{O}((NT)^{-\vartheta_*(1-p_K)/4}).$$

Combining these with the policy-error bound gives the value rate with $\mathcal{N}_b^{-\vartheta_*(1-p_K)/4}$. Since $b = O(\log(eNT))$, it gives

$$J^\star - J(\widehat{\pi}_M) \leq a_{\mathcal{H}} + \widetilde{O}_p\left((NT)^{-\vartheta_*(1-p_K)/4} + \sqrt{\Delta_n}\right),$$

with additional dependence-comparison failure at most $(NT)^{-2}$ (zero when $b = T$). The target, raw-secant, policy-localization, and state-transfer conditions remain uniform over the recalibrated classes and pooled laws. This proves the stronger rate in Example 2.

I AdaFNN prediction and exact functional secants

We use the learned-basis architecture of Yao et al. (2021). The finite-rank Bellman condition supplies uniform target approximation, while the prediction bounds cover all critic and policy parameters. Evaluation uses exact secants in outer-parameter and functional-basis coordinates, with a separate directional information condition.

Assumption I.1 (Adaptive-basis clipped ReLU fitted-Q regression)**.** The functional-action set is bounded in L^2:

$$B_A := \sup_{a\in\mathcal{A}} \|a\|_{L^2} < \infty.$$

Fix a deterministic state feature map $\varphi_{S,n} : \mathcal{S} \to \mathbb{R}^{d_{S,n}}$ with compact range $\mathcal{X}_{S,n}$. When the state is already finite-dimensional, as in the Pendulum experiment, $\varphi_{S,n}$ is the identity. Assumption I.2 uses this state representation in its finite-rank factorization.

Fix an integer $d_{\mathrm{Ada}} \geq 1$, independent of the sample size, as the number of adaptive basis nodes. For $j = 1, \ldots, d_{\mathrm{Ada}}$, let

$$\beta_{\theta_{\beta,j}}(u) = \mathrm{NN}_{\theta_{\beta,j}}(u), \qquad u \in [0,1],$$

where the micro-network class $\mathcal{B}_n$ has deterministic depth, width, sparsity, activation, and weight bounds, and satisfies $\sup_{\beta\in\mathcal{B}_n} \|\beta\|_{L^2} \leq B_{\beta,n}$. Its exact basis score is

$$c_{\theta_{\beta,j}}(a) := \langle a, \beta_{\theta_{\beta,j}}\rangle_{L^2} = \int_0^1 a(u)\beta_{\theta_{\beta,j}}(u)\,du.$$

The raw adaptive-basis ReLU critic is

$$f_{\theta,\boldsymbol{\theta}_\beta}(s,a) = \mathrm{NN}_\theta\left[\varphi_{S,n}(s), c_{\theta_{\beta,1}}(a), \ldots, c_{\theta_{\beta,d_{\mathrm{Ada}}}}(a)\right],$$

where NN_θ ranges over a deterministic bounded-weight ReLU class $\mathcal{M}_{\mathrm{out},n}$ on

$$\mathcal{X}_{\mathrm{Ada},n} := \mathcal{X}_{S,n} \times [-B_A B_{\beta,n}, B_A B_{\beta,n}]^{d_{\mathrm{Ada}}}.$$

Write $\mathcal{Q}_{\mathrm{Ada},n}$ for this composite class and $\overline{\mathcal{Q}}_{\mathrm{Ada},n} := \{\mathrm{clip}_{[0,V_{\max}]} f : f \in \mathcal{Q}_{\mathrm{Ada},n}\}$. The class is fixed before the data are observed even though the fitted basis functions are learned from the data. The micro-network class contains the zero function, so unused basis nodes in a lower-rank target can be zero-padded, by setting their output weights to zero.

The fitted-Q routine returns a $\tau_{\mathrm{Ada},n}$-approximate global minimizer of the clipped empirical squared loss: for every deterministic label $y \in \mathcal{Y}_n$,

$$\begin{aligned}&\frac{1}{n}\sum_{i,t}\{\mathrm{clip}_{[0,V_{\max}]}\,\widehat{f}_y^{\mathrm{Ada}}(X_{i,t}) - Y_{i,t}(y)\}^2 \\ &\quad\leq \inf_{f\in\mathcal{Q}_{\mathrm{Ada},n}} \frac{1}{n}\sum_{i,t}\{\mathrm{clip}_{[0,V_{\max}]}\,f(X_{i,t}) - Y_{i,t}(y)\}^2 + \tau_{\mathrm{Ada},n},\end{aligned}$$

where $X_{i,t} = (S_{i,t}, A_{i,t})$ and $Y_{i,t}(y) = y(S_{i,t}, A_{i,t}, R_{i,t}, S_{i,t+1})$. The fitted-$Q$ update uses $\mathrm{clip}_{[0,V_{\max}]}\,\widehat{f}_y^{\mathrm{Ada}}$. Since rewards lie in $[0,1]$ and every preceding fitted iterate is clipped to $[0, V_{\max}]$, the generated Bellman labels satisfy

$$0 \leq Y(y) \leq 1 + \gamma V_{\max} = V_{\max}, \qquad 0 \leq Q_y \leq V_{\max}.$$

The same uniform bound is imposed on any enlarged deterministic label class.

For numerical implementation, let $\widetilde{c}_{\beta,J_{\mathrm{quad},n}}$ denote the score computed from the supplied observation record. Assume it is measurable and that its declared representation supplies the bound

$$A_{\mathrm{quad},n} = \sup_{a,\beta} |\widetilde{c}_{\beta,J_{\mathrm{quad},n}}(a) - \langle a,\beta\rangle_{L^2}|.$$

This discrepancy is retained, and no vanishing rate is asserted for arbitrary raw point samples of an L^2 function. Exact inner products give zero discrepancy. Section M.7 verifies numerical scores from cell averages using regularity of the learned basis, not of the observed action. Proposition I.3 transfers

the resulting score discrepancy to critic outputs. Point evaluation on an arbitrary L^2 equivalence class is not treated as a bounded functional.

Prediction and covering bounds use the implemented scores; secant identities use the exact-score counterpart, with discrepancy bounded in Proposition I.3. Throughout the approximation and prediction statements, $\mathcal{Q}_{\mathrm{Ada},n}$ denotes the implemented class, and it coincides with the exact-score class when $A_{\mathrm{quad},n} = 0$. With numerical scores, the common state–score box is enlarged to

$$\mathcal{X}_{S,n} \times [-B_A B_{\beta,n} - A_{\mathrm{quad},n}, B_A B_{\beta,n} + A_{\mathrm{quad},n}]^{d_{\mathrm{Ada}}},$$

and the stated outer-network and entropy conditions are imposed there.

For $\varepsilon, \eta > 0$, define

$$H^{\mathrm{Ada}}_{F,n}(\varepsilon) := \log \mathrm{Cov}(\varepsilon, \overline{\mathcal{Q}}_{\mathrm{Ada},n}, \|\cdot\|_\infty), \qquad H^{\mathrm{Ada}}_{Y,M,n}(\eta) := \log \mathrm{Cov}(\eta, \mathcal{Y}_n, \|\cdot\|_\infty),$$

$$H_{\mathrm{Ada},M,n}(\varepsilon, \eta, \delta) := H^{\mathrm{Ada}}_{F,n}(\varepsilon) + H^{\mathrm{Ada}}_{Y,M,n}(\eta) + \log(8M/\delta).$$

The reachable-label entropy is obtained from the sup-norm Lipschitz Bellman maps, together with a policy cover when labels are indexed jointly by (Q, π).

The primary sampling conditions are independent subject clusters (Assumption D.1) or the full-sample two-group construction under Assumption L.1. Their effective sample sizes are N and $\mathcal{N}_b$, respectively; only the full-sample two-group construction carries the comparison failure δ_b^{mix}. The optional one-group route uses $N_{\mathrm{eff}}(b, q)$, omission $\mathsf{o}_{b,q}$ and failure $\delta_{\mathrm{coup}}(b, q)$ from the coupling bound (61). It carries the additional prediction radius $\Delta^{\mathrm{Ada}}_{\mathrm{dep},Q} = V_{\max}\sqrt{\mathsf{o}_{b,q}}$, as proved in Appendix L.4; equality of retained and full design laws is not assumed. The conversion to the policy-evaluation norm is supplied by Proposition F.1 after verifying the exact secant representation, smooth-feature bound, and directional moment comparison.

The learned score layer is the AdaFNN construction of Yao et al. (2021, Section 3), with both basis micro-networks and the outer network included in the fitted class. Declaring that entire class in advance allows the same-data uniformity argument of Nguyen-Tang et al. (2022) to cover the learned representation and its numerical scores together.

Assumption I.2 (Uniform finite-rank Bellman-target factorization)**.** For every $y \in \mathcal{Y}_n$, there exist an integer $d_y^\circ \le d_{\mathrm{Ada}}$, a continuous linear map

$$\mathrm{Score}_y : L^2([0,1]) \to \mathbb{R}^{d_y^\circ}, \qquad \|\mathrm{Score}_y\|_{\mathrm{op}} \le B_{\mathrm{score}},$$

and a continuous outer map

$$h_y : \mathcal{X}_{S,n} \times [-B_A B_{\mathrm{score}}, B_A B_{\mathrm{score}}]^{d_y^\circ} \to [0, V_{\max}]$$

such that

$$Q_y(s,a) = h_y\{\varphi_{S,n}(s), \mathrm{Score}_y\, a\}, \qquad (s,a) \in \mathcal{S} \times \mathcal{A}.$$

Assume $B_{\beta,n} \ge B_{\mathrm{score}}$. After zero-padding $\mathrm{Score}_y\, a$ to dimension d_{Ada}, assume that h_y admits a $[0, V_{\max}]$-valued extension to the common score domain $\mathcal{X}_{\mathrm{Ada},n}$, including its numerical enlargement when used. All stated regularity and approximation bounds apply to this extension; padding alone does not extend its domain. The outer maps share a deterministic modulus of continuity $\omega_{h,M,n}$ on their compact domains:

$$|h_y(z,v) - h_y(z,v')| \le \omega_{h,M,n}(\|v - v'\|_2), \qquad \omega_{h,M,n}(r) \downarrow 0 \text{ as } r \downarrow 0.$$

The learned action directions are global; state dependence and interaction enter through h_y. The fixed d_{Ada} bounds the functional rank of every reachable target. We take the common modulus $\omega_{h,M,n}$ to be continuous and nondecreasing, allowing passage to unattained approximation infima in Lemma I.1.

This uses the finite-rank composition behind Yao et al. (2021, Theorem 1), now with score maps explicitly continuous on L^2 and a common approximation modulus over the Bellman-label family. The directions may vary with the target label, so the assumption matches adaptive bases rather than a fixed action projection.

By the Hilbert-space Riesz representation theorem, each coordinate of Score_y has a representer $b_{y,j} \in L^2([0,1])$ satisfying

$$\mathrm{Score}_y\, a = \big(\langle a, b_{y,1}\rangle, \dots, \langle a, b_{y,d_y^\circ}\rangle\big), \qquad \max_j \|b_{y,j}\|_{L^2} \le B_{\mathrm{score}}.$$

Choose such representers and define the uniform micro-network and outer-network approximation errors

$$A_{\beta,M,n} := \sup_y \max_{1\le j\le d_y^\circ} \inf_{\beta\in\mathcal{B}_n} \|b_{y,j} - \beta\|_{L^2},$$

$$A_{h,M,n} := \sup_y \inf_{\mathrm{NN}\in\mathcal{M}_{\mathrm{out},n}} \sup_{(z,v)\in\mathcal{X}_{\mathrm{Ada},n}} |\,\mathrm{NN}(z,v) - h_y(z,v)|,$$

where $\mathcal{X}_{\mathrm{Ada},n}$ is the corresponding zero-padded compact state-score domain.

Lemma I.1 (Uniform adaptive-basis approximation for reachable Bellman targets)**.** *Under Assumptions I.1 and I.2,*

$$A^{\infty}_{\mathrm{Ada},M,n} := \sup_{y\in\mathcal{Y}_n} \inf_{f\in\mathcal{Q}_{\mathrm{Ada},n}} \Big\|\mathrm{clip}_{[0,V_{\max}]}\, f - Q_y\Big\|_{\infty}$$
$$\le A_{h,M,n} + \omega_{h,M,n}\Big[\sqrt{d_{\mathrm{Ada}}}\{B_A A_{\beta,M,n} + A_{\mathrm{quad},n}\}\Big].$$

In particular, the behavior-design approximation error

$$A_{\mathrm{Ada},M,n} := \sup_y \inf_{f\in\mathcal{Q}_{\mathrm{Ada},n}} \Big\|\mathrm{clip}_{[0,V_{\max}]}\, f - Q_y\Big\|_{L^2(\nu_X)}$$

and the corresponding approximation error in the policy-evaluation norm are both bounded by $A^{\infty}_{\mathrm{Ada},M,n}$*. If the outer maps are uniformly* $\mathrm{Lip}_{h,M,n}$*-Lipschitz in their score arguments, the final term is bounded by*

$$\mathrm{Lip}_{h,M,n}\sqrt{d_{\mathrm{Ada}}}\{B_A A_{\beta,M,n} + A_{\mathrm{quad},n}\}.$$

The composition argument follows the basis-plus-outer approximation in Yao et al. (2021); the displayed bound makes each contribution uniform over reachable Bellman targets and retains numerical score error. Smoothness assumptions can then be converted into explicit ReLU approximation rates using Schmidt-Hieber (2020, Theorem 5).

Proof. By Riesz representation, $(\mathrm{Score}_y\, a)_j = \langle a, b_{y,j}\rangle$ with $\|b_{y,j}\|_{L^2} \le B_{\mathrm{score}}$. Fix $\tau > 0$ and choose basis approximants with $\|\beta_{y,j} - b_{y,j}\|_{L^2} \le A_{\beta,M,n} + \tau$; use zero in unused coordinates. If $z_y(a)$ is the zero-padded target score, Cauchy–Schwarz and the numerical score bound give

$$|\widetilde{c}_j(a) - (z_y(a))_j| \le A_{\mathrm{quad},n} + B_A(A_{\beta,M,n} + \tau),$$
$$\|\widetilde{c}(a) - z_y(a)\|_2 \le \sqrt{d_{\mathrm{Ada}}}\{A_{\mathrm{quad},n} + B_A(A_{\beta,M,n} + \tau)\} =: b_\tau.$$

Both vectors lie in the declared enlarged score box. Choose an outer approximant NN_y^τ with error at most $A_{h,M,n} + \tau$ there. For $f_y^\tau = \mathrm{clip}\,\mathrm{NN}_y^\tau(\varphi_{S,n}, \widetilde{c})$, nonexpansive clipping and the modulus of the extended target imply

$$|f_y^\tau(s,a) - Q_y(s,a)| \le |\,\mathrm{NN}_y^\tau(\varphi_{S,n},\widetilde{c}) - h_y(\varphi_{S,n},\widetilde{c})| + |h_y(\varphi_{S,n},\widetilde{c}) - h_y(\varphi_{S,n},z_y)|$$
$$\le A_{h,M,n} + \tau + \omega_{h,M,n}(b_\tau).$$

Take the approximation infimum and then $\sup_y$. Letting $\tau \downarrow 0$ uses continuity of the common modulus and gives the stated full-domain bound. The Lipschitz version follows by $\omega_{h,M,n}(u) \le \mathrm{Lip}_{h,M,n}u$. Both claimed prediction and policy-evaluation norms are bounded by the same supremum norm because their underlying measures are probability laws. □

Fix a deterministic sequence $t_n \downarrow 0$. For every $y \in \mathcal{Y}_n$, choose a deterministic parameter Θ_y° in the adaptive-basis class such that, with $F_\Theta := \mathrm{clip}_{[0,V_{\max}]}\, f_\Theta$,

$$\Big\|F_{\Theta_y^\circ} - Q_y\Big\|_{\infty} \le A^{\infty}_{\mathrm{Ada},M,n} + t_n.$$

These are deterministic near-oracle parameters for the fixed reachable-label class; they are not fitted from the observed sample.

Lemma I.2 (Composite entropy of the adaptive-basis critic)**.** *Assume every* $\mathrm{NN} \in \mathcal{M}_{\mathrm{out},n}$ *is* $\mathrm{Lip}_{\mathrm{out},n}$*-Lipschitz in its score coordinates. Let*

$$H_{\mathrm{out},n}(u) := \log \mathrm{Cov}(u, \mathcal{M}_{\mathrm{out},n}, \|\cdot\|_{\infty,\mathcal{X}_{\mathrm{Ada},n}}), \qquad H_{\beta,n}(u) := \log \mathrm{Cov}(u, \mathcal{B}_n, \|\cdot\|_{L^2}).$$

For exact L^2 *scores,*

$$H^{\mathrm{Ada}}_{F,n}(\varepsilon) \le H_{\mathrm{out},n}(\varepsilon/2) + d_{\mathrm{Ada}}\, H_{\beta,n}\left(\frac{\varepsilon}{2\mathrm{Lip}_{\mathrm{out},n} B_A \sqrt{d_{\mathrm{Ada}}}}\right).$$

For a deterministic quadrature implementation, the same statement holds with $B_A \left\|\beta - \widetilde{\beta}\right\|_{L^2}$ *replaced by a deterministic uniform score-Lipschitz bound for that rule. Depth and weight factors remain in the network covers, or are absorbed when they are logarithmic under the chosen architecture.*

The sparse-network cover of Schmidt-Hieber (2020, Lemma 5) applies separately to the outer network and every basis micro-network. Composing their covers through the score Lipschitz bound accounts for the entire learned critic, and the Bellman-label cover subsequently adds the entropy of the full RKHS policy class.

Proof. Set $d = d_{\mathrm{Ada}}$. Choose an $\varepsilon/2$-net of the outer class and a u-net of the basis class, with representatives NN_k and $\widetilde{\beta}_j$. Cauchy–Schwarz and the outer Lipschitz bound give

$$\|c_\beta(a) - c_{\widetilde{\beta}}(a)\|_2 \le B_A \sqrt{d}\, u,$$
$$|\,\mathrm{NN}(\varphi_{S,n}, c_\beta) - \mathrm{NN}_k(\varphi_{S,n}, c_{\widetilde{\beta}})| \le \varepsilon/2 + \mathrm{Lip}_{\mathrm{out},n} B_A \sqrt{d}\, u.$$

Take $u = \varepsilon/(2\mathrm{Lip}_{\mathrm{out},n} B_A \sqrt{d})$ when the denominator is positive. Clipping does not increase the error. The product net consequently has at most

$$\mathrm{Cov}(\varepsilon/2, \mathcal{M}_{\mathrm{out},n}, \|\cdot\|_\infty)\, \mathrm{Cov}(u, \mathcal{B}_n, \|\cdot\|_{L^2})^d$$

elements. Taking logarithms proves the bound. If the denominator vanishes, one basis choice suffices. For numerical scores use their declared score-Lipschitz bound in the score-perturbation inequality, with the outer cover and Lipschitz bound on the enlarged box. □

Proposition I.1 (Uniform adaptive-basis prediction under clustered and mixing observations)**.** *Under Assumption I.1, fix* $\varepsilon, \eta > 0$*. For independent subjects set* $(\mathcal{N}_{\mathrm{stat}}, \mathsf{o}, \delta_{\mathrm{dep}}) = (N, 0, 0)$*; for the full-sample two-group partition set* $(\mathcal{N}_{\mathrm{stat}}, \mathsf{o}, \delta_{\mathrm{dep}}) = (\mathcal{N}_b, 0, \delta_b^{\mathrm{mix}})$*. With probability at least* $1 - \delta - \delta_{\mathrm{dep}}$*, uniformly over the full deterministic label class,*

$$\|\,\mathrm{clip}_{[0,V_{\max}]}\, \widehat{f}^{\mathrm{Ada}}_y - Q_y\|^2_{L^2(\nu_X)} \le C\Bigg[A^2_{\mathrm{Ada},M,n} + V_{\max}(\varepsilon+\eta) + \tau_{\mathrm{Ada},n} + \frac{V^2_{\max} H_{\mathrm{Ada},M,n}(\varepsilon,\eta,\delta)}{\mathcal{N}_{\mathrm{stat}}} + V^2_{\max}\mathsf{o}\Bigg].$$

Define $\delta_{\mathrm{Ada},n}$ *by the square root, replacing the approximation radius by its full-domain upper bound* $A^\infty_{\mathrm{Ada},M,n}$*. The secondary retained-block version uses* $(\mathcal{N}_{\mathrm{stat}}, \mathsf{o}, \delta_{\mathrm{dep}}) = (N_{\mathrm{eff}}(b,q), \mathsf{o}_{b,q}, \delta_{\mathrm{coup}}(b,q))$ *and requires* $N_{\mathrm{eff}}(b,q) \ge 1$.

The regression step uses the double-uniform principle of Nguyen-Tang et al. (2022), with the entropy of both learned network components and the policy-indexed labels. Whole subjects or the stated mixing blocks determine the effective sample size, so the same architecture-specific bound connects the independent-subject and temporally informative sampling regimes.

Proof. Use the deterministic implemented fit and label classes in Lemma D.2 or Proposition L.1. Their product nets have the cost $H_{\mathrm{Ada},M,n}$, and labels and clipped fits are bounded by $V_{\max}$. Those proofs establish the full-class relative excess-loss event before substituting an empirical minimizer.

In their final basic inequality, the approximation term can be left as the population infimum rather than replaced immediately by a sup-norm bound:

$$\tfrac{3}{4}\|\widehat{f}_y - Q_y\|^2_{L^2(\nu_X)} \le \tfrac{5}{4} \inf_{f\in\overline{\mathcal{Q}}_{\mathrm{Ada},n}} \|f - Q_y\|^2_{L^2(\nu_X)} + \tau_{\mathrm{Ada},n} + C\{V_{\max}(\varepsilon+\eta) + V^2_{\max}H/\mathcal{N}_{\mathrm{stat}}\}.$$

This proves the assertion with $A_{\mathrm{Ada},M,n}$; its full-domain upper bound then gives the declared radius. The numerical discrepancy is already part of the implemented-class approximation and is not discarded. For one-group blocks use (63), whose proof adds $CV^2_{\max}\mathsf{o}$ without equating the design laws. Every event covers the full class before fitting, so no separate conditioning on the fitted labels or independence across FQI iterations is needed. □

I.1 Exact functional-coordinate secants with all bases learned

Let

$$F_{\theta,\boldsymbol{\beta}}(s,a) = \mathrm{clip}_{[0,V_{\max}]}\,\mathrm{NN}_\theta\{\varphi_{S,n}(s), \langle a,\beta_1\rangle, \dots, \langle a,\beta_d\rangle\}.$$

Each endpoint basis β_j is still the output of its learned micro-network. For the proof only, use the functional basis elements themselves as coordinates in $\mathcal{E} = \mathbb{R}^{p_\theta} \oplus \{L^2([0,1])\}^d$. A segment between two basis elements need not be another micro-network output; it is an analytical path between the two endpoint critics. Thus no basis is frozen and no different fitting procedure is introduced.

Proposition I.2 (Functional secant features of AdaFNN)**.** *For two endpoint critics let $d_v = (\theta - \theta', \beta_1 - \beta'_1, \dots, \beta_d - \beta'_d)$. Take the straight segment $(\theta_t, \boldsymbol{\beta}_t)$ in these coordinates. Suppose the clipped outer map along its deterministic parameter hull is Lipschitz in its score vector with constant $\mathrm{Lip}_{\mathrm{out},n}$, and has a pathwise parameter-derivative bound $B_{\partial,n}$ on the bounded input domain. Then there are measurable integrated derivatives $u_v(s,a) \in \mathbb{R}^{p_\theta}$ and $w_v(s,a) \in \mathbb{R}^d$ such that*

$$\psi_v(s,a) = (u_v(s,a), w_{v,1}(s,a)a, \dots, w_{v,d}(s,a)a),$$
$$F_{\theta,\boldsymbol{\beta}} - F_{\theta',\boldsymbol{\beta}'} = \langle d_v, \psi_v\rangle, \qquad \|u_v(s,a)\| \le B_{\partial,n}, \quad \|w_v(s,a)\| \le \mathrm{Lip}_{\mathrm{out},n}.$$

For

$$\mathbf{W}_v = \mathrm{Id}_{p_\theta} \oplus \mathbf{W}_\alpha \oplus \cdots \oplus \mathbf{W}_\alpha, \qquad \vartheta = \min\{1, \zeta/\alpha\},$$

the integrated policy radius (4) verifies (27) with

$$B^2_{\psi,n} = B^2_{\partial,n} + \mathrm{Lip}^2_{\mathrm{out},n} B^2_{\Pi,\zeta}. \tag{44}$$

Observed actions require only their $L^2([0,1])$ norm for feature integrability. If the endpoint parameter and basis bounds are $\|\theta\| \le B_{\theta,n}$ and $\|\beta_j\| \le B_{\beta,n}$, then $\|d_v\| \le B_{Q,n} := 2(B^2_{\theta,n} + dB^2_{\beta,n})^{1/2}$.

The pathwise chain rule of Bolte & Pauwels (2021, Lemma 2 and Theorem 8) gives an exact difference representation even with ReLU and clipping. Keeping the learned basis functions as functional coordinates exposes the factors $w_{v,j}a$, through which the functional-action norm directly controls the policy-evaluation feature energy.

Proof. Put $\delta\theta = \theta - \theta'$ and $\delta\beta_j = \beta_j - \beta'_j$. Along the analytical path $\theta_t = \theta' + t\delta\theta$, $\beta_{t,j} = \beta'_j + t\delta\beta_j$, the scores satisfy $\dot{z}_{t,j}(a) = \langle a, \delta\beta_j\rangle$. The intermediate bases need not be micro-network outputs.

Write $F_t(s,a) = \overline{\mathrm{NN}}_{\theta_t}(s, z_t(a))$, including clipping. On the bounded path this function is Lipschitz, hence absolutely continuous. ReLU preactivations are piecewise polynomial in t; each nonzero piece has finitely many roots, and an identically zero piece has zero path derivative. The chain-rule selections in the declared derivative envelopes (Bolte & Pauwels, 2021, Lemma 2 and Theorem 8) therefore give

$$\dot{F}_t = \langle u_t, \delta\theta\rangle + \sum_{j=1}^d w_{t,j}\langle a, \delta\beta_j\rangle \quad \text{for almost every } t,$$

$$u_v = \int_0^1 u_t\,dt, \qquad w_v = \int_0^1 w_t\,dt, \qquad \|u_v\| \le B_{\partial,n}, \quad \|w_v\| \le \mathrm{Lip}_{\mathrm{out},n}.$$

The fundamental theorem of calculus now yields the exact identity

$$F_1 - F_0 = \langle u_v, \delta\theta\rangle + \sum_{j=1}^{d} \langle w_{v,j} a, \delta\beta_j\rangle_{L^2} = \langle d_v, \psi_v\rangle_{\mathcal{E}},$$
$$\psi_v = (u_v, w_{v,1}a, \dots, w_{v,d}a).$$

Thus all outer and basis differences are retained. For $\mathbf{W}_v = \mathrm{Id} \oplus \mathbf{W}_\alpha \oplus \cdots \oplus \mathbf{W}_\alpha$ and $\alpha\vartheta \le \zeta$,

$$\|\mathbf{W}_v^{-\vartheta/2}\psi_v(s,a)\|^2 = \|u_v\|^2 + \|w_v\|^2 \sum_{k\ge 1} k^{2\alpha\vartheta} |\langle a, \phi_k\rangle|^2$$
$$\le B_{\partial,n}^2 + \mathrm{Lip}_{\mathrm{out},n}^2 \|a\|_\zeta^2.$$

Taking the policy supremum before integrating proves the feature bound $B_{\partial,n}^2 + \mathrm{Lip}_{\mathrm{out},n}^2 B_{\Pi,\zeta}^2$. Behavior integrability and the coefficient radius use only ordinary $L^2([0,1])$ norms:

$$\mathbb{E}\|\psi_v(S,A)\|^2 \le B_{\partial,n}^2 + \mathrm{Lip}_{\mathrm{out},n}^2 \mathbb{E}\|A\|_{L^2}^2 < \infty,$$
$$\|d_v\|^2 = \|\delta\theta\|^2 + \sum_j \|\delta\beta_j\|_{L^2}^2 \le 4B_{\theta,n}^2 + 4dB_{\beta,n}^2.$$

The directional lower-moment comparison is the existing actual-direction hypothesis on this full block feature second-moment operator. □

The associated second-moment operator is the full block operator

$$\Sigma_v = \mathbb{E}_{\nu_X} \begin{bmatrix} u_v \\ w_{v,1}A \\ \vdots \\ w_{v,d}A \end{bmatrix} \begin{bmatrix} u_v \\ w_{v,1}A \\ \vdots \\ w_{v,d}A \end{bmatrix}^*.$$

Its cross blocks and its dependence on both endpoint critics must be retained. Neither a covariance of A alone nor the last hidden layer verifies its lower bound. The representation is an exact secant in outer-parameter and functional-basis coordinates, not a claim about a frozen basis or about the native micro-network parameter Jacobian.

For any fixed finite bounded-weight architecture and bounded state and score inputs, $B_{\partial,n}$ is finite on the parameter hull. Growing width, depth, or layer norms can enlarge it. Bounded individual weights do not prove a uniformly bounded product of layer operator norms, $B_{\partial,n}$, or the moment coefficient c_n^{-1}.

Proposition I.3 (Numerical score accuracy and action regularity)**.** *Every endpoint critic is* $\mathrm{Lip}_{\mathrm{Ada},n} = \mathrm{Lip}_{\mathrm{out},n}\sqrt{d}B_{\beta,n}$*-Lipschitz in its action. If each numerical integral score has uniform error at most* $A_{\mathrm{quad},n}$*, then*

$$e_{\mathrm{Ada},n} = \mathrm{Lip}_{\mathrm{out},n}\sqrt{d}\, A_{\mathrm{quad},n}$$

bounds the numerical input error.

Yao et al. (2021, Algorithm 1) evaluates each learned score by numerical integration. This bound quantifies how the score accuracy propagates through the outer network and supplies a common error budget for fitting, policy objectives and the exact-score secant comparison.

Proof. For a fixed endpoint basis, Cauchy–Schwarz and the outer Lipschitz bound give

$$\|c_{\boldsymbol{\beta}}(a) - c_{\boldsymbol{\beta}}(a')\|_2^2 = \sum_{j=1}^{d} |\langle a - a', \beta_j\rangle|^2$$
$$\le \|a - a'\|_{L^2}^2 \sum_{j=1}^{d} \|\beta_j\|_{L^2}^2 \le dB_{\beta,n}^2 \|a - a'\|_{L^2}^2,$$
$$|F(s,a) - F(s,a')| \le \mathrm{Lip}_{\mathrm{out},n}\|c_{\boldsymbol{\beta}}(a) - c_{\boldsymbol{\beta}}(a')\|_2$$
$$\le \mathrm{Lip}_{\mathrm{out},n}\sqrt{d}\, B_{\beta,n}\|a - a'\|_{L^2}.$$

For numerical scores, both vectors lie in the stipulated enlarged box, so

$$\|\widetilde{c}(a) - c(a)\|_2^2 = \sum_{j=1}^{d} |\widetilde{c}_j(a) - c_j(a)|^2 \le d A_{\mathrm{quad},n}^2,$$

$$\begin{aligned} |\widehat{F}^{\mathrm{num}}(s,a) - F^{\mathrm{exact}}(s,a)| &\le \mathrm{Lip}_{\mathrm{out},n} \|\widetilde{c}(a) - c(a)\|_2 \\ &\le \mathrm{Lip}_{\mathrm{out},n} \sqrt{d}\, A_{\mathrm{quad},n}. \end{aligned}$$

Taking the endpoint and full-domain suprema gives the stated uniform bounds. □

Corollary I.1 (AdaFNN evaluation and its joint-complexity rate)**.** (i) Conditional evaluation. *Use the deterministic prediction and label classes of Proposition I.1. For the deterministic exact-score approximants, take*

$$\eta_{Q,n} = A_{h,M,n} + \omega_{h,M,n}(\sqrt{d} B_A A_{\beta,M,n}) + t_n.$$

If the exact functional-coordinate family satisfies the actual-direction comparison (26) with the spectral weight operator in Proposition I.2, then

$$\sup_y \|\widehat{f}_y - Q_y\|_{\nu_+, \Pi_{\mathcal{H}}(B_{\mathcal{H}}),2} \le E_n(\delta_{\mathrm{Ada},n}, \eta_{Q,n}, e_{\mathrm{Ada},n}, B_{Q,n}),$$

where $B_{\psi,n}$ is given by (44) and $\vartheta = \min\{1, \zeta/\alpha\}$. For the accuracy-dependent directional condition, use E_n^{sc} instead. Without either verification, the general bound is (33). These are information requirements on the declared family, not automatic properties of every AdaFNN architecture.

(ii) Joint-complexity specialization. *Use $\mathcal{N}_{\mathrm{stat}} = N$ for equal-length subject clusters or $\mathcal{N}_{\mathrm{stat}} = \mathcal{N}_b$ for full-sample two-group mixing. Along $M = M_n$, retain the approximation conditions and bounded $V_{\max}$, and let D_n be a deterministic bound on the full joint covering cost below, including the actual policy RKHS entropy. At polynomial net resolutions satisfying $\varepsilon_n + \eta_n = O(D_n/\mathcal{N}_{\mathrm{stat}})$, assume explicitly*

$$H_{F,n}^{\mathrm{Ada}}(\varepsilon_n) + H_{Y,M_n,n}^{\mathrm{Ada}}(\eta_n) + \log(8M_n/\delta_n) = \widetilde{O}(D_n).$$

The product cover verifies this absorption only when the additional depth, width, weight and action-Lipschitz factors in its logarithmic entropy bound are themselves absorbed; otherwise retain those factors in the displayed prediction bound. Under either primary regime,

$$\delta_{\mathrm{Ada},n} = \widetilde{O}\Big(\eta_{Q,n} + e_{\mathrm{Ada},n} + \sqrt{D_n/\mathcal{N}_{\mathrm{stat}}} + \sqrt{\tau_{\mathrm{Ada},n}}\Big).$$

For the secondary retained-block route use $\mathcal{N}_{\mathrm{stat}} = N_{\mathrm{eff}}(b,q)$ and retain $V_{\max}\sqrt{\mathsf{o}_{b,q}}$ and the coupling probability. When the full prediction-plus-approximation expression is $\widetilde{O}(\mathcal{N}_{\mathrm{stat}}^{-b_Q})$ and the constants in part (i) are bounded, policy evaluation error is $\widetilde{O}(\mathcal{N}_{\mathrm{stat}}^{-\vartheta b_Q})$. Under the scale-dependent condition replace ϑ by ϑ_*; if the feature or moment constants grow, retain them explicitly.

The prediction rate combines the target approximation and joint covering inputs used in neural fitted-Q analysis (Nguyen-Tang et al., 2022). Evaluation at functional policy actions then contributes the separate exponent ϑ (or ϑ_*) through secant stability, making the roles of critic capacity and functional-action regularity visible in the final rate.

Proof. Choose deterministic exact-score approximants f_y° from Lemma I.1, including their near-minimizer tolerance in t_n. For the exact-score fit F_y, the supplied bounds are

$$\|f_y^\circ - Q_y\|_\infty \le \eta_{Q,n}, \qquad \|\widehat{f}_y - Q_y\|_{L^2(\nu_X)} \le \delta_{\mathrm{Ada},n}, \qquad \|\widehat{f}_y - F_y\|_\infty \le e_{\mathrm{Ada},n}.$$

Thus the two triangle inequalities in Corollary F.1 apply with prediction argument $r_n^\circ = \delta_{\mathrm{Ada},n} + \eta_{Q,n} + e_{\mathrm{Ada},n}$ and additive error $\eta_{Q,n} + e_{\mathrm{Ada},n}$. Proposition I.2 supplies $B_{\psi,n} = \{B_{\partial,n}^2 + \mathrm{Lip}_{\mathrm{out},n}^2 B_{\Pi,\zeta}^2\}^{1/2}$ and radius $B_{Q,n}$. Substitution gives

$$\sup_y \|\widehat{f}_y - Q_y\|_{\nu_+, \Pi_{\mathcal{H}}(B_{\mathcal{H}}),2} \le \eta_{Q,n} + e_{\mathrm{Ada},n} + B_{\psi,n} c_n^{-\vartheta/2} B_{Q,n}^{1-\vartheta} (r_n^\circ)^\vartheta.$$

The finite-accuracy and scale-dependent versions use respectively $\omega_n(r_n^\circ)$ and the infimum in Proposition F.1(ii) for the last term. All substitutions occur on the same full-label event.

Joint-complexity specialization. For exact scores, action Lipschitz continuity gives the joint label comparison

$$\|y_{Q,\pi} - y_{Q',\pi'}\|_\infty \le \gamma\{\|Q - Q'\|_\infty + \mathrm{Lip}_{\mathrm{Ada},n}\|\pi - \pi'\|_{\infty,L^2}\}.$$

Choose critic and policy net radii $\eta/(2\gamma)$ and $\eta/(2\gamma\mathrm{Lip}_{\mathrm{Ada},n})$; when $\mathrm{Lip}_{\mathrm{Ada},n} = 0$ only one policy representative is needed. Lemma I.2 and the assumed logarithmic absorption give $H_F(\varepsilon_n) + H_Y(\eta_n) + \log(8M_n/\delta_n) = \widetilde{O}(D_n)$. For numerical scores, use a verified action-Lipschitz bound for the implemented critic in this product-cover argument, or impose the full implemented-class entropy bound in part (ii) directly; a uniform score discrepancy alone does not imply action Lipschitz continuity. The numerical version of an exact approximant has full-domain error at most $\eta_{Q,n}+e_{\mathrm{Ada},n}$. The prediction theorem therefore yields

$$\delta^2_{\mathrm{Ada},n} \le \widetilde{O}\{(\eta_{Q,n} + e_{\mathrm{Ada},n})^2 + D_n/\mathcal{N}_{\mathrm{stat}} + \tau_{\mathrm{Ada},n} + V^2_{\max}\mathsf{o}\},$$
$$\delta_{\mathrm{Ada},n} \le \widetilde{O}\{\eta_{Q,n} + e_{\mathrm{Ada},n} + \sqrt{D_n/\mathcal{N}_{\mathrm{stat}}} + \sqrt{\tau_{\mathrm{Ada},n}} + V_{\max}\sqrt{\mathsf{o}}\}.$$

The displayed bound for $\delta_{\mathrm{Ada},n}$ uses $\sqrt{\sum_j x_j} \le \sum_j \sqrt{x_j}$. Both primary regimes have $\mathsf{o} = 0$; the one-group alternative retains its omission with $\mathcal{N}_{\mathrm{stat}} = N_{\mathrm{eff}}(b,q)$. Add the relevant dependence failure to δ_n. Finally, if $r_n = \delta_{\mathrm{Ada},n} + \eta_{Q,n} + e_{\mathrm{Ada},n} = \widetilde{O}(\mathcal{N}_{\mathrm{stat}}^{-b_Q})$, substitute this bound into part (i). Bounded transfer factors give $r_n + Cr_n^\vartheta = \widetilde{O}(\mathcal{N}_{\mathrm{stat}}^{-\vartheta b_Q})$. The accuracy-dependent information condition replaces ϑ by ϑ_*. □

J Policy-norm implementation and coverage-compatible empirical total curvature

The theoretical update uses only $\|\pi\|^2_{\mathcal{H}}$. For the Sobolev construction in Proposition C.1, this norm already contains the within-action terms that imply the spectral bound in Proposition C.2; no second smoothness penalty is added to Algorithm 1. The reported spline computation uses empirical total curvature in place of direct evaluation of this complete norm. The two objectives are not asserted to be identical. The role verified below is the one needed for coverage: empirical total curvature is a computable surrogate whose localized policies obey the same classwise spectral envelope. Curvature alone is a seminorm because it vanishes on affine action functions; the action-feasibility bound controls this null component through the L^2 term in Proposition C.2, while the declared coefficient-map class controls state-side complexity.

For a fixed spline basis vector $\mathbf{b}^{\mathrm{spl}}$ in the action argument, set

$$G_0 = \int_0^1 \mathbf{b}^{\mathrm{spl}}(u)\mathbf{b}^{\mathrm{spl}}(u)^\top du, \qquad G_2 = \int_0^1 (\mathbf{b}^{\mathrm{spl}})''(u)\{(\mathbf{b}^{\mathrm{spl}})''(u)\}^\top du, \qquad G_{\mathcal{G}} = G_0 + G_2.$$

In the two-state running example, writing $\pi(s)(u) = \mathbf{b}^{\mathrm{spl}}(u)^\top c_s$ gives the exact computable penalty

$$\|\pi\|^2_{\mathcal{H}} = \sum_{s=1}^{2} c_s^\top G_{\mathcal{G}} c_s.$$

For the continuous-state construction $\mathcal{H} = H^1([0,1];\mathcal{G})$ and a coefficient map $c \in H^1([0,1];\mathbb{R}^{p_n})$,

$$\|\pi\|^2_{\mathcal{H}} = \int_0^1 \{c(s)^\top G_{\mathcal{G}} c(s) + c'(s)^\top G_{\mathcal{G}} c'(s)\}\, ds.$$

Thus the same quadratic penalty is implementable from spline mass and second-derivative matrices, with an additional state-derivative term when the policy class couples continuous states. For bounded differentiable features $z(s)$ and $c(s) = a_{\max}\tanh(\Theta_{\mathrm{pol}}^\top z(s))$, the chain rule and $|\tanh'| \le 1$ give $\|c'(s)\| \le a_{\max}\|\Theta_{\mathrm{pol}}\|_{\mathrm{op}}\|z'(s)\|$.

For a deterministic feasible spline sieve $\mathcal{C}_n$, define

$$\widehat{\Omega}_{\mathcal{D}}(\pi) = \frac{1}{n}\sum_{i,t} \|\partial_u^2\pi(S_{i,t+1})\|^2_{L^2}, \qquad \Omega(\pi) = \int \|\partial_u^2\pi(s)\|^2_{L^2}\nu_+(ds). \tag{45}$$

The first quantity is the implemented surrogate and the second is its population counterpart. They measure the within-action component of the Sobolev geometry; state-side complexity remains controlled by the declared spline coefficient-map class.

Proposition J.1 (Curvature sublevels supply a classwise spectral envelope)**.** *Suppose* $\sup_{\pi\in\mathcal{C}_n,s}\|\pi(s)\|_{L^2}\le R_A$. *Let*

$$\mathcal{G}_n^{(2)}=\{s\mapsto\partial_u^2\pi(s):\pi\in\mathcal{C}_n\},\qquad U_n=\sup_{g\in\mathcal{G}_n^{(2)},s}\|g(s)\|_{L^2}^2<\infty,$$

and write

$$\mathcal{N}_n(v)=\mathrm{Cov}\{v,\mathcal{G}_n^{(2)},\|\cdot\|_{\infty,L^2}\}.$$

For $B\ge 0$, *put* $\mathcal{C}_n(B)=\{\pi\in\mathcal{C}_n:\Omega(\pi)\le B\}$ *and*

$$W_n(B)=\min\left[U_n,\inf_{v>0}\left\{\sqrt{\mathcal{N}_n(v)B}+2v\right\}^2\right].\tag{46}$$

If the displayed covering numbers are finite, then, for the shifted Legendre basis in Proposition C.2 and $0<\zeta\le 2$,

$$\int\sup_{\pi\in\mathcal{C}_n(B)}\|\pi(s)\|_\zeta^2\nu_+(ds)\le C_S\{R_A^2+W_n(B)\}.\tag{47}$$

The bound $C_S(R_A^2+U_n)$ *remains valid without the covering-number refinement.*

Proof. Choose a v-cover of $\mathcal{G}_n^{(2)}$. From every cover ball meeting the population sublevel class, choose one member of that intersection. At most $\mathcal{N}_n(v)$ representatives g_j are needed, and they form a $2v$-cover of the derivative maps in $\mathcal{C}_n(B)$. Because each representative belongs to the sublevel class,

$$\left\|\sup_{\pi\in\mathcal{C}_n(B)}\|\partial_u^2\pi(\cdot)\|_{L^2}\right\|_{L^2(\nu_+)}\le\left\|\max_j\|g_j(\cdot)\|_{L^2}\right\|_{L^2(\nu_+)}+2v\le\sqrt{\mathcal{N}_n(v)B}+2v.$$

The pointwise envelope gives the alternative bound $U_n^{1/2}$. Apply Proposition C.2 to every selected action, take the policy supremum at each state and then integrate. The uniform L^2 action bound supplies R_A^2, and optimizing over v proves (47). □

Proposition J.2 (Empirical total curvature localizes the population envelope)**.** *Suppose that, on an event* $\mathcal{E}_{\Omega,n}$, *uniformly over* $\mathcal{C}_n$,

$$\Omega(\pi)\le c_{\Omega,n}\widehat{\Omega}_{\mathcal{D}}(\pi)+v_{\Omega,n}.\tag{48}$$

Then the empirical sublevel set $\widehat{\mathcal{C}}_n(b)=\{\pi\in\mathcal{C}_n:\widehat{\Omega}_{\mathcal{D}}(\pi)\le b\}$ *is contained in* $\mathcal{C}_n(c_{\Omega,n}b+v_{\Omega,n})$ *on that event. Consequently, its classwise spectral radius is bounded by the right side of* (47) *with* $B=c_{\Omega,n}b+v_{\Omega,n}$.

One sufficient additive comparison follows directly from the independent-subject empirical process. Let $r_\pi(s)=\|\partial_u^2\pi(s)\|_{L^2}^2\in[0,U_n]$ *and* $H_{\Omega,n}(\epsilon)=\log\mathrm{Cov}\{\epsilon,\{r_\pi:\pi\in\mathcal{C}_n\},\|\cdot\|_\infty\}$. *For every* $\epsilon>0$, *with probability at least* $1-\delta$,

$$\sup_{\pi\in\mathcal{C}_n}|\widehat{\Omega}_{\mathcal{D}}(\pi)-\Omega(\pi)|\le 2\epsilon+U_n\sqrt{\frac{H_{\Omega,n}(\epsilon)+\log(2/\delta)}{2N}}.\tag{49}$$

Thus (48) *holds with* $c_{\Omega,n}=1$ *and the displayed right side as* $v_{\Omega,n}$.

Proof. The set inclusion is immediate from (48), after which Proposition J.1 applies. For (49), regard the average of $r_\pi(S_{i,t+1})$ over t as one $[0,U_n]$-valued observation from subject i. These N subject averages are independent and their mean expectation is $\Omega(\pi)$. Apply Hoeffding's inequality on a fixed ϵ-net, take a union bound, and pay 2ϵ to pass from the net to the full class. No within-subject independence is used. □

The next relative comparison can sharpen the additive bound when the derivative maps share a fixed finite state representation.

Proposition J.3 (Relative comparison for an empirical derivative criterion)**.** *Suppose* $\partial_u^q\pi(s)(u)=w(s)^\top v_\pi(u)$ *in a declared numerical family, and put* $\Sigma_w=\mathbb{E}_{\nu_+}w(S)w(S)^\top$ *and* $\widehat{\Sigma}_w=n^{-1}\sum_{i,t}w(S_{i,t+1})w(S_{i,t+1})^\top$. *If* $c_-\Sigma_w\preceq\widehat{\Sigma}_w\preceq c_+\Sigma_w$, *including null directions, then the empirical and population averages of* $\|\partial_u^q\pi(S)\|_{L^2}^2$ *lie within the same factors, uniformly over that family.*

Proof. Tonelli writes the population quantity as $\int v_\pi(u)^\top \Sigma_w v_\pi(u)\, du$ and the empirical quantity with $\widehat{\Sigma}_w$. Integrating the assumed quadratic-form comparison proves the claim. The Gram concentration calculation in Appendix D.4 gives $c_- = 1/2, c_+ = 3/2$ under its stated leverage and sample-size conditions. □

In particular, the lower comparison gives (48) with $c_{\Omega,n} = c_-^{-1}$ and $v_{\Omega,n} = 0$. A numerical derivative matrix must likewise be compared with G_2 on the coefficient space. The ordering $G_2 \preceq c_{\mathrm{num}}\widehat{G}_2$ on the coefficient space gives the one-sided comparison required above, including the null directions; alternatively, a uniform numerical discrepancy can be absorbed into $v_{\Omega,n}$.

Corollary J.1 (The Pendulum empirical total curvature is coverage-compatible)**.** *For the experimental cubic B-spline policies in Appendix O.1, suppose the B-spline basis is nonnegative and forms a partition of unity and $c_C(s) \in [-2,2]^{p_u}$, as in the implemented bounded output map. Then, for every state and policy parameter,*

$$\|\pi_C(s)\|_{L^2}^2 \le 4, \qquad \|\partial_u^2 \pi_C(s)\|_{L^2}^2 = c_C(s)^\top W_{p_u} c_C(s) \le 4p_u \lambda_{\max}(W_{p_u}) = \Omega_{\mathrm{scale}}(p_u).$$

Consequently, for $0 < \zeta \le 2$,

$$\int \sup_C \|\pi_C(s)\|_\zeta^2 \nu_+(ds) \le C_S\{4 + \Omega_{\mathrm{scale}}(p_u)\} < \infty. \tag{50}$$

Empirical total-curvature sublevels admit the sharper radius in Propositions J.1–J.2 whenever their stated comparison and covering conditions are used. Hence the implemented criterion verifies the policy-smoothness input to Assumption 2(iii)*. Together with the directional-information comparison already imposed on the relevant critic secants in Assumption 2*(ii)*, it yields the spectral transfer bound of Proposition 1. Example 1 proves this comparison from its positive logged coordinate weights; the KRR and AdaFNN results use the corresponding moment comparisons stated in their specializations.*

Proof. Partition of unity and the coefficient box give $|\pi_C(s)(u)| \le 2$. The quadratic-form bound follows from $\|c_C(s)\|_2^2 \le 4p_u$. Proposition C.2 then gives (50) pointwise before integration. □

The reported experiment fixes $p_u = 12$, so (50) is a deterministic finite radius. If the spline dimension grows with sample size, the displayed p_u- and matrix-dependent radius must be retained in the transfer factor; this corollary does not assert a dimension-uniform bound.

K A QUANTITATIVE RKHS-COMPATIBLE APPROXIMATION REGIME

This section records how explicit approximation and entropy conditions enter the value theorem. The RKHS policy class remains fixed. The full-kernel KRR rate for Example 1 is proved separately in Appendix M.

For a finite-state model whose policy RKHS is a fixed Sobolev product space, let the label-uniform Bellman targets have at most d functional scores as in Assumption I.2. Suppose their score representers belong to a fixed one-dimensional C^{s_b} ball and their outer functions, separately for each state, belong to a fixed C^{s_h} ball on a compact d-dimensional score box. Require the extended outer functions to be uniformly Lipschitz in scores. Sparse ReLU approximations (Yarotsky, 2017; Schmidt-Hieber, 2020) at accuracy radius $\varepsilon_{\mathrm{app}}$ can be chosen with outer and basis parameter counts

$$p_h(\varepsilon_{\mathrm{app}}) = \widetilde{O}(\varepsilon_{\mathrm{app}}^{-d/s_h}), \qquad p_\beta(\varepsilon_{\mathrm{app}}) = \widetilde{O}(\varepsilon_{\mathrm{app}}^{-1/s_b}),$$

under the stated bounded-domain smoothness conditions. The full-domain approximation bound in Lemma I.1 is then $O(\varepsilon_{\mathrm{app}})$. Take logarithmic depths and polynomial weight bounds so their covering factors at polynomial resolutions are logarithmic. For the declared whole critic class, retain a verified action-Lipschitz bound $\mathrm{Lip}(\varepsilon_{\mathrm{app}}) \le C\varepsilon_{\mathrm{app}}^{-p_{\mathrm{Lip}}}$ with $p_{\mathrm{Lip}} \ge 0$; smooth targets alone do not bound every fitted network's Lipschitz constant.

The entropy of the localized policy class from Appendix E, at resolution $\varepsilon_{\mathrm{app}}^2/\mathrm{Lip}(\varepsilon_{\mathrm{app}})$, is $\widetilde{O}(\varepsilon_{\mathrm{app}}^{-1-p_{\mathrm{Lip}}/2})$. The composite critic and label covers consequently have order $\widetilde{O}(\varepsilon_{\mathrm{app}}^{-d_{\mathrm{app}}})$, where

$$D_{\mathrm{Ada}} = d_{\mathrm{app}} = \max\{d/s_h,\ 1/s_b,\ 1 + p_{\mathrm{Lip}}/2\}.$$

This exponent includes full RKHS policy complexity. It is not the parameter count of the numerical spline optimizer.

Corollary K.1 (A conditional structural value rate)**.** *Under the preceding smoothness, architecture, entropy and numerical conditions, suppose the state constants are bounded with $\chi \le \chi_0 < 1$. Require the membership and localization event and the secant conditions to hold with probability tending to one, with bounded feature/coefficient/information factors. Write ϑ_* for the verified transfer exponent, allowing accuracy-dependent information. Assume the fixed-radius localization condition in (16) holds for $B_{\mathcal{H}}$, and let $\Delta_{\varepsilon_{\mathrm{app}}}$ bound the policy-compatibility gap $\sup_{Q\in\mathcal{Q}_{\varepsilon_{\mathrm{app}}}} \Delta(Q)$. Choose a deterministic logarithmic majorant $\kappa_{\log,N}$ and*

$$\varepsilon_{\mathrm{app},N} = (\kappa_{\log,N}/N)^{1/(2+d_{\mathrm{app}})}, \quad \lambda_{\pi,n} = O(\varepsilon_{\mathrm{app},N}), \quad \tau_{\mathrm{Ada,n}} = O(\varepsilon^2_{\mathrm{app},N}).$$

Use exactly evaluated scores or a discrepancy of order $O(\varepsilon_{\mathrm{app},N})$ with the same implemented-class cover. For conforming product splines take $d_u \gtrsim \varepsilon_{\mathrm{app},N}^{-(1+p_{\mathrm{Lip}})/2}$ and a policy optimization error of $O(\varepsilon_{\mathrm{app},N})$. Then, with $M \ge c \log N$ for sufficiently large c,

$$J^\star - J(\widehat{\pi}_M) = a_{\mathcal{H}} + \widetilde{O}_p\Big\{N^{-\min\{\vartheta_*,1/2\}/(2+d_{\mathrm{app}})} + \sqrt{\Delta_{\varepsilon_{\mathrm{app},N}}}\Big\}.$$

The approximation and network covers use Schmidt-Hieber (2020, Theorem 5 and Lemma 5), while the policy cover uses the full Sobolev product envelope containing the localized class. The minimum in the exponent separates spectral transfer from the objective-to-greedy conversion: the objective-to-greedy conversion contributes $\varepsilon_{\mathrm{app},N}^{1/2}$ even when critic evaluation is faster.

Proof. Use fit and label resolutions of order $\varepsilon^2_{\mathrm{app},N}$. The entropy bound and choice of $\kappa_{\log,N}$ give $\{H_F + H_Y + \log(1/\delta_N)\}/N = O(\varepsilon^2_{\mathrm{app},N})$ for polynomially small δ_N. The clipped squared-loss proposition therefore gives prediction radius $O(\varepsilon_{\mathrm{app},N})$. The same joint unpenalized objective cover yields $u_n = O(\varepsilon_{\mathrm{app},N})$. The fixed ball bounds the deterministic penalty term by $\lambda_{\pi,n}B^2_{\mathcal{H}} = O(\varepsilon_{\mathrm{app},N})$. Proposition C.3 gives representation gap $O(\varepsilon_{\mathrm{app},N}^{-p_{\mathrm{Lip}}} d_u^{-2}) = O(\varepsilon_{\mathrm{app},N})$. Together with the stipulated solver and numerical objective tolerances, Lemma D.1 gives $\varepsilon_{\pi,n} = O\{\varepsilon^{1/2}_{\mathrm{app},N} + \sqrt{\Delta_{\varepsilon_{\mathrm{app},N}}}\}$. The verified transfer proposition gives $\varepsilon_{Q,n} = O(\varepsilon^{\vartheta_*}_{\mathrm{app},N})$. Theorem 1 and $M \asymp \log N$ finish the proof. □

The information and derivative envelopes remain explicit conditions on the whole growing critic family. If their product grows, retain that product in (35) before propagating errors; the displayed exponent requires the bounded-factor premise. For a coupled continuous-state RKHS class, the verified implementation in Appendix C.3 supplies representation and membership bounds, and the same value theorem additionally retains $\sup_Q \Delta(Q)$.

K.1 Continuous states with the complete fixed RKHS ball

Use the complete policy class of Proposition E.2 as an entropy envelope, without restricting it to its numerical projections. The value and coverage arguments use the same fixed comparison ball $\Pi_{\mathcal{H}}(B_{\mathcal{H}})$. At accuracy radius $\varepsilon_{\mathrm{app}}$, let $\mathcal{Q}_{\varepsilon_{\mathrm{app}}}$ be a clipped fitting class fixed before fitting. Suppose its full-domain Bellman targets, indexed by every $Q \in \mathcal{Q}_{\varepsilon_{\mathrm{app}}}$ and $\pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})$, have approximants in the fitting class with uniform error $O(\varepsilon_{\mathrm{app}})$. Assume

$$\log \mathrm{Cov}(\varepsilon^2_{\mathrm{app}}, \mathcal{Q}_{\varepsilon_{\mathrm{app}}}, \|\cdot\|_\infty) \le C\varepsilon_{\mathrm{app}}^{-p^{\mathrm{ent}}_Q} \log^c(e/\varepsilon_{\mathrm{app}}), \qquad \mathrm{Lip}(\varepsilon_{\mathrm{app}}) \le C\varepsilon_{\mathrm{app}}^{-p_{\mathrm{Lip}}}, \quad p_{\mathrm{Lip}} \ge 0, \tag{51}$$

where $\mathrm{Lip}(\varepsilon_{\mathrm{app}})$ is an action-Lipschitz bound on the entire class and c is fixed. Define

$$D_{\mathrm{ct}} = \max\{p^{\mathrm{ent}}_Q, 5 + 5p_{\mathrm{Lip}}/2\}, \qquad \Delta_{\varepsilon_{\mathrm{app}}} = \sup_{Q\in\mathcal{Q}_{\varepsilon_{\mathrm{app}}}} \Delta(Q).$$

The policy part of this exponent follows by evaluating (22) at resolution $\varepsilon^2_{\mathrm{app}}/\{1 + \mathrm{Lip}(\varepsilon_{\mathrm{app}})\}$. Thus the joint fit, objective and Bellman-label log-covering cost is $\widetilde{O}(\varepsilon_{\mathrm{app}}^{-D_{\mathrm{ct}}})$.

Corollary K.2 (Continuous-state value rate under a full-ball entropy envelope)**.** *Under (51), use independent subjects, clipped least-squares fitting with objective tolerance $O(\varepsilon^2_{\mathrm{app},N})$, and exact scores. Suppose the uniform membership and localization, secant and state-transfer conditions hold*

with probability tending to one, with bounded transfer factors, bounded occupancy constants and $\chi \le \chi_0 < 1$. *Let*

$$\varepsilon_{\mathrm{app},N} = (\kappa_{\log,N}/N)^{1/(2+D_{\mathrm{ct}})}, \qquad \lambda_{\pi,n} = O(\varepsilon_{\mathrm{app},N}), \qquad M \ge c_M \log N,$$

where $\kappa_{\log,N}$ *is a sufficiently large deterministic logarithmic majorant and* c_M *is sufficiently large. Require the localized-class empirical policy-objective gap* $\tau_{\pi,n} = O(\varepsilon_{\mathrm{app},N})$ *and the fixed-radius norm-localization condition* (16)*. Then*

$$J^\star - J(\widehat{\pi}_M) = a_{\mathcal{H}} + \widetilde{O}_p\left\{N^{-\min\{\vartheta_*,1/2\}/(2+D_{\mathrm{ct}})} + \sqrt{\Delta_{\varepsilon_{\mathrm{app},N}}}\right\}, \qquad \vartheta_* = \frac{2\vartheta}{2+\beta_{\mathrm{info}}\vartheta}.$$

Here $\beta_{\mathrm{info}} = 0$ *for uniform directional information; otherwise require the accuracy-dependent condition with fixed prefactor and accuracy interval. If* $\Delta_{\varepsilon_{\mathrm{app},N}} = O(\varepsilon_{\mathrm{app},N})$*, this is* $a_{\mathcal{H}}$ *plus the displayed polynomial rate; the approximation term vanishes when a globally optimal policy belongs to* $\Pi_{\mathcal{H}}(B_{\mathcal{H}})$.

This extends the restricted-policy improvement issue in Antos et al. (2007) to a complete continuous-state, function-valued RKHS entropy envelope. The localized class inherits the entropy bound of that ball, while the residual propagation uses the same state-side roles as Munos & Szepesvári (2008); Farahmand et al. (2010). The additional $\Delta_{\varepsilon_{\mathrm{app}}}$ term bounds the policy-compatibility gap for the single coupled policy class, and $a_{\mathcal{H}}$ records approximation by the fixed comparison ball.

Proof. Use policy resolution $\varepsilon^2_{\mathrm{app},N}/\{1+\mathrm{Lip}(\varepsilon_{\mathrm{app},N})\}$ and critic resolution $\varepsilon^2_{\mathrm{app},N}$ in the product label cover. Equations (17)–(18), with a polynomially small failure probability, give

$$u_n = \widetilde{O}(\varepsilon_{\mathrm{app},N}), \qquad \delta_{Q,n} = \widetilde{O}(\varepsilon_{\mathrm{app},N}), \qquad \frac{\widetilde{O}(\varepsilon^{-D_{\mathrm{ct}}}_{\mathrm{app},N})}{N} = O(\varepsilon^2_{\mathrm{app},N}).$$

The target approximation and numerical input bounds are included in the prediction radius. The deterministic penalty contribution is at most $\lambda_{\pi,n}B^2_{\mathcal{H}} = O(\varepsilon_{\mathrm{app},N})$. Lemma D.1 therefore gives $\varepsilon_{\pi,n} = \widetilde{O}(\sqrt{\Delta_{\varepsilon_{\mathrm{app},N}}} + \varepsilon^{1/2}_{\mathrm{app},N})$. Proposition F.1 and Corollary F.1 give $\varepsilon_{Q,n} = \widetilde{O}(\varepsilon^{\vartheta_*}_{\mathrm{app},N})$. Substitute these bounds into Theorem 1; geometric decay controls the iteration remainder.

For a concrete computation, the energy projections and global shrinkage of Appendix C.3 preserve this same full ball. Choose $h_s \lesssim \varepsilon^{2(1+p_{\mathrm{Lip}})}_{\mathrm{app},N}$ and $d_u \gtrsim \varepsilon^{-(1+p_{\mathrm{Lip}})/2}_{\mathrm{app},N}$. Their empirical representation loss is

$$CL(\varepsilon_{\mathrm{app},N})B_{\mathcal{H}}(h_s^{1/2} + d_u^{-2}) = O(\varepsilon_{\mathrm{app},N}).$$

Solver and numerical objective tolerances $O(\varepsilon_{\mathrm{app},N})$ then supply the stated localized-class $\tau_{\pi,n}$ through (13). These projections change the implementation, not the localized policy class in any statistical supremum. □

The selector condition can be checked directly. If each $Q \in \mathcal{Q}_{\varepsilon_{\mathrm{app}}}$ has a measurable pointwise maximizing selector in the same $\Pi_{\mathcal{H}}(B_{\mathcal{H}})$, then $\Delta_{\varepsilon_{\mathrm{app}}} = 0$. For example, fix $v_\circ \in \mathcal{G}$ with $\|v_\circ\|_{L^2} = 1$ and take

$$Q(s,a) = \mathrm{clip}_{[0,V_{\max}]}\{q(s) - \eta(s)(\langle v_\circ, a\rangle - g(s))^2\},$$

where $0 \le q \le V_{\max}$, $\eta \ge 0$, $\|g\|_{H^1}\|v_\circ\|_{\mathcal{G}} \le B_{\mathcal{H}}$ and $\|g\|_\infty \le B_A$. The policy $\pi_Q(s) = g(s)v_\circ$ lies in $\Pi_{\mathcal{H}}(B_{\mathcal{H}})$ and attains $q(s)$ at every state. This verifies the selector requirement for this critic family; the policy domain itself remains infinite dimensional.

Numerical scores give the same conclusion when implemented critics differ from their exact-score counterparts by $O(\varepsilon_{\mathrm{app},N})$ uniformly on the fitting and policy-evaluation domains and the implemented critics and labels, together with those counterparts, admit the same deterministic covering bounds. The proof then uses these implemented-class covers in concentration and charges the uniform discrepancy in Corollary F.1.

For learned-basis critics with d scores, outer C^{s_h} targets ($s_h \ge 1$) on a fixed compact continuous state–score domain, uniformly Lipschitz in the scores, and C^{s_b} score representers give the approximation budgets of Schmidt-Hieber (2020, Theorem 5 and Lemma 5) with

$$p^{\mathrm{ent}}_Q = \max\{(1+d)/s_h, 1/s_b\}, \qquad D_{\mathrm{ct}} = \max\{(1+d)/s_h, 1/s_b, 5+5p_{\mathrm{Lip}}/2\}.$$

The budgets use the finite-rank factorization of Assumption I.2, uniformly over the full policy-indexed labels. Any restriction enforcing the declared Lipschitz or selector bounds must also contain the target approximants required in (51); this containment is a premise of the specialization. The same argument permits other critic covers by retaining their p_Q^{ent}. These are sufficient rates from an explicit full-ball cover.

K.2 Retaining growing transfer factors

Corollary K.3 (Value rates with growing critic transfer factors)**.** *Use either structural setting above, with $\varepsilon_{\mathrm{app},N} = \widetilde{O}(N^{-1/(2+D)})$, prediction, approximation and numerical errors $\widetilde{O}(\varepsilon_{\mathrm{app},N})$, and policy error $\widetilde{O}(\varepsilon_{\mathrm{app},N}^{1/2} + \sqrt{\Delta_{\varepsilon_{\mathrm{app},N}}})$. Retain bounded state constants, $\chi \le \chi_0 < 1$ and $M \ge c_M \log N$ for sufficiently large c_M. Replace bounded transfer factors by the accuracy-dependent information bound $c_n(\varepsilon) \ge \underline{c}_n \varepsilon^{\beta_{\mathrm{info}}}$ on a fixed accuracy interval, where $\beta_{\mathrm{info}} \ge 0$ is the information-decay exponent, and assume*

$$B_{\psi,n}\underline{c}_n^{-\vartheta/2} B_{Q,n}^{1-\vartheta} = \widetilde{O}(\varepsilon_{\mathrm{app},N}^{-p_{\mathrm{grow}}}), \qquad 0 \le p_{\mathrm{grow}} < \vartheta.$$

At $\vartheta = 1$ the coefficient-radius factor is omitted. Then

$$J^\star - J(\widehat{\pi}_M) = a_{\mathcal{H}} + \widetilde{O}_p\left\{ N^{-\frac{1}{2+D}\min\{\frac{2(\vartheta - p_{\mathrm{grow}})}{2+\beta_{\mathrm{info}}\vartheta}, 1/2\}} + \sqrt{\Delta_{\varepsilon_{\mathrm{app},N}}} \right\}.$$

For a finite-state setting with a product localized class, or with the direct statewise defect bound used in Example 1, take $\Delta_{\varepsilon_{\mathrm{app},N}} = 0$.

Hilbert-scale interpolation relates weak and strong errors through smoothness (Egger & Hofmann, 2018, equation (1.4)). This corollary additionally retains the sample-dependent multiplier of the critic secants before balancing accuracy. The fixed comparison-ball radius keeps the policy-action evaluation radius bounded; only verified growth of critic features, coefficients or inverse information is charged here.

Proof. The finite-accuracy version of Proposition F.1 bounds the critic evaluation error by an additive $\widetilde{O}(\varepsilon_{\mathrm{app},N})$ plus

$$\inf_{0<\varepsilon\le\varepsilon_0} \{\varepsilon + \widetilde{O}(\varepsilon_{\mathrm{app},N}^{\vartheta - p_{\mathrm{grow}}})\varepsilon^{-\beta_{\mathrm{info}}\vartheta/2}\}.$$

For $\beta_{\mathrm{info}} > 0$, balancing gives exponent $b_{\mathrm{tr}} = 2(\vartheta - p_{\mathrm{grow}})/(2 + \beta_{\mathrm{info}}\vartheta) > 0$; for $\beta_{\mathrm{info}} = 0$, let ε decrease to zero and take $b_{\mathrm{tr}} = \vartheta - p_{\mathrm{grow}}$. Since $b_{\mathrm{tr}} \le 1$, the additive error is no slower. The main value theorem with the stated policy error yields $a_{\mathcal{H}} + \widetilde{O}_p(\varepsilon_{\mathrm{app},N}^{\min\{b_{\mathrm{tr}},1/2\}} + \sqrt{\Delta_{\varepsilon_{\mathrm{app},N}}})$, which is the displayed bound. Setting $p_{\mathrm{grow}} = 0$ recovers the bounded-factor rate. □

K.3 Unequal-length interpretation

For the weighted extension of Appendix D.4, replace the structural proof's concentration scale by $\mathcal{N}_\infty = n/T_{\max}$, retaining its actual objective scale $\mathcal{N}_2 = n^2/\sum_i T_i^2$ where useful. The conservative KRR score/operator argument instead uses $N_\star = N/C_{T,n}^2$. If $\overline{T} \asymp N^{p_{\mathrm{avg}}}$ and $C_{T,n} \asymp N^{p_C}$, these scales obey

$$n \asymp N^{1+p_{\mathrm{avg}}}, \quad \mathcal{N}_\infty \asymp N^{1-p_C}, \quad N_\star \asymp N^{1-2p_C}, \qquad \mathcal{N}_\infty \le \mathcal{N}_2 \le N.$$

Thus a verified neural structural power p has order $\widetilde{O}_p(N^{-(1-p_C)p})$ for $p_C < 1$, while the conservative KRR bound must use its own scale and constants. For $N - N^{p_{\mathrm{long}}}$ subjects with $N^{p_{\mathrm{short}}}$ transitions each and $N^{p_{\mathrm{long}}}$ subjects with $N^{p_{\mathrm{longlen}}}$ transitions each, $p_{\mathrm{longlen}} \ge p_{\mathrm{short}}$, $0 < p_{\mathrm{long}} < 1$, put $p_T = \max\{1 + p_{\mathrm{short}}, p_{\mathrm{long}} + p_{\mathrm{longlen}}\}$. Direct summation gives

$$n \asymp N^{p_T}, \qquad \mathcal{N}_\infty \asymp N^{p_T - p_{\mathrm{longlen}}}, \qquad \mathcal{N}_2 \asymp N^{2p_T - \max\{1+2p_{\mathrm{short}}, p_{\mathrm{long}}+2p_{\mathrm{longlen}}\}}.$$

These are length-array calculations. They do not verify policy entropy or information for the changing pooled design. For conditional random-length statements the constants must hold uniformly on the stated length event, and its failure probability is added.

L QUANTITATIVE BETA-MIXING RATES FOR THE FULL-SAMPLE ALGORITHM

We retain the approximation, information and Bellman arguments, and prove one full-sample event for prediction and policy improvement, together with an optional finite-feature Gram comparison. Two alternating block groups include all terminal fragments, so their sum is the unchanged transition-weighted objective and has no omission error. Section L.4 treats the secondary one-group construction, where omission and coupling costs remain explicit.

L.1 DEPENDENCE AND A PARTITION THAT USES EVERY TRANSITION

Use the common-length data with $n = NT$ and independent subjects. The same proof extends to the length array of Appendix D.4 by retaining each block's actual length. Write

$$Z_{i,t} = (S_{i,t}, A_{i,t}, R_{i,t}, S_{i,t+1}), \qquad \widehat{\mathbb{P}} f = n^{-1} \sum_{i,t} f(Z_{i,t}), \qquad \mathbb{P} f = \mathbb{E}\widehat{\mathbb{P}} f.$$

The dependence condition is on these complete transition records, not only on the state sequence. This matters, for example, when a persistent subject-level action component remains after state dependence has weakened.

Assumption L.1 (Absolute regularity of complete records)**.** For $Z_{i,t} = (S_{i,t}, A_{i,t}, R_{i,t}, S_{i,t+1})$, require

$$\sup_{i,N} \sup_{1 \le t \le T-k} \beta_{\mathrm{mix}}\{\sigma(Z_{i,1}, \ldots, Z_{i,t}), \sigma(Z_{i,t+k}, \ldots, Z_{i,T})\} \le \beta_{\mathrm{mix}}(k) \longrightarrow 0.$$

Here $\beta_{\mathrm{mix}}(\mathcal{U}, \mathcal{V})$ is one half the supremum of $\sum_{j,\ell} |\Pr(U_j \cap V_\ell) - \Pr(U_j)\Pr(V_\ell)|$ over finite measurable partitions from the two sigma fields. Its nonincreasing envelope and constants are uniform in subjects and sample size.

We use this uniform array condition. Its bound concerns the complete records and is uniform in subjects and sample size; terminal fragments and nonstationary marginals remain allowed. This uses the absolute-regularity convention of Bradley (2005, equations (1.5), (2.2)). Marginal stationarity, a common subject marginal distribution, and independence between the two groups below are not assumed. The common conditional model still requires $Q_y(X_{i,t}) = \mathbb{E}\{Y_y(Z_{i,t}) \mid X_{i,t}\}$ at every observed pair, for each deterministic covered label.

For an integer $1 \le b \le T$, partition each subject into consecutive blocks

$$m_i = \lceil T/b \rceil, \qquad I_{i,j} = \{(j-1)b + 1, \ldots, \min(jb, T)\}, \qquad \ell_{i,j} = |I_{i,j}|,$$
$$\mathcal{I}_c = \{(i,j) : j \equiv c \pmod 2\}, \quad c = 0, 1, \qquad \widehat{\mathbb{P}}_c f = \frac{1}{n} \sum_{(i,j) \in \mathcal{I}_c} \sum_{t \in I_{i,j}} f(Z_{i,t}).$$

Thus $\widehat{\mathbb{P}} = \widehat{\mathbb{P}}_0 + \widehat{\mathbb{P}}_1$, $\mathbb{P} = \mathbb{P}_0 + \mathbb{P}_1$ for $\mathbb{P}_c = \mathbb{E}\widehat{\mathbb{P}}_c$, and $\sum_{i,j} \ell_{i,j} = n$. The normalization in each group is n, not that group's sample size. A last block may be short; it is never deleted. Define

$$L_b^{\mathrm{link}} = \sum_i (m_i - 2)_+, \qquad \delta_b^{\mathrm{mix}} = \min\{1, L_b^{\mathrm{link}} \beta_{\mathrm{mix}}(b+1)\}, \qquad \mathcal{N}_b = \frac{n}{b}.$$

Consecutive blocks within the same group are separated by one complete length-b block. There are exactly L_b^{link} within-subject links across the two groups, and

$$L_b^{\mathrm{link}} \le n/b, \qquad \delta_b^{\mathrm{mix}} \le \min\{1, (n/b)\beta_{\mathrm{mix}}(b)\}, \qquad \mathcal{N}_b \ge N.$$

At $b = T$ there is one block per subject and $\delta_b^{\mathrm{mix}} = 0$, whether or not the mixing envelope is small. The effective sample size $\mathcal{N}_b$ is justified by the concentration bounds, not a claim that all transitions are independent.

Lemma L.1 (A shared independent-block comparison)**.** *For each c, compare the joint law of the blocks in $\mathcal{I}_c$ with the product of their actual blockwise marginal laws. Their total-variation distance is at most the sum of the within-subject link coefficients in that group. Consequently, if events $\mathsf{Good}_c^{\mathrm{mix}}$, each measurable with respect to its own block group, have failure probabilities at most δ_c under their respective product laws, then*

$$\Pr(\mathsf{Good}_0^{\mathrm{mix}} \cap \mathsf{Good}_1^{\mathrm{mix}}) \ge 1 - \delta_0 - \delta_1 - \delta_b^{\mathrm{mix}}. \tag{52}$$

An event $\mathsf{Good}_c^{\mathrm{mix}}$ *may simultaneously cover every deterministic fit, Bellman label, policy objective, second-moment operator and optional derivative matrix considered below. The comparison cost is not multiplied by a covering number or by the number of FQI iterations.*

The independent-block comparison of Yu (1994, Corollary 2.7 and Lemma 4.1) transfers one uniform event at a time. Keeping the actual nonstationary block marginals and subject boundaries gives the displayed array bound; the two groups are combined by a probability union bound rather than an independence assertion.

Proof. For two standard-Borel random elements, the partition definition of β_{mix} equals the total-variation distance between their joint law and the product of their marginals; this follows first for finite partitions and then by measurable approximation. For ordered same-group block records $\mathbf{Z}_1, \dots, \mathbf{Z}_m$ of one subject, telescoping the joint law against the product law gives

$$\left\| \mathcal{L}(\mathbf{Z}_1, \dots, \mathbf{Z}_m) - \bigotimes_{j=1}^{m} \mathcal{L}(\mathbf{Z}_j) \right\|_{\mathrm{TV}} \le \sum_{j=2}^{m} \beta_{\mathrm{mix}}\{\sigma(\mathbf{Z}_1, \dots, \mathbf{Z}_{j-1}), \sigma(\mathbf{Z}_j)\} \le (m-1)\beta_{\mathrm{mix}}(b+1). \tag{53}$$

The product-law telescoping bound in (53) follows inductively by separating the last block and using that tensoring by a fixed probability law does not increase total variation. Its $\beta_{\mathrm{mix}}(b+1)$ bound uses monotonicity under restriction of sigma fields and the actual time separation. Independence across subjects permits another product-law telescoping, adding their bounds. The sum of the two group link counts is $\sum_i (m_i - 2)_+$, including cases $m_i = 1, 2$. Transfer each group's entire bad event once from its product law and apply a union bound. If the uncapped comparison sum exceeds one, (52) with $\delta_b^{\mathrm{mix}} = 1$ is vacuous but valid. The two product laws are used separately; their groups are never asserted to be independent of one another. This is the independent-block comparison principle used by Yu (1994, Lemma 4.1), written here for nonstationary subject arrays and unequal terminal blocks. □

L.2 All three statistical inputs on one event

The next result replaces the cluster concentration steps, not just the sample-size notation. The Gram part is included for optional finite-feature diagnostics and is not needed by the deterministic policy-norm penalty.

Proposition L.1 (Full-sample objective, prediction and Gram control)**.** *Under Assumption L.1, fix b and deterministic classes before fitting. Let $0 < \delta < 1$. For a class $\mathcal{O}$ with $0 \le f \le B$ and sup-norm log covering number $H_{\mathcal{O}}(\epsilon)$,*

$$\sup_{f \in \mathcal{O}} |(\widehat{\mathbb{P}} - \mathbb{P})f| \le 2\epsilon + CB\sqrt{\frac{H_{\mathcal{O}}(\epsilon) + \log(16/\delta)}{\mathcal{N}_b}}. \tag{54}$$

For clipped least squares, let fits, labels and their targets take values in $[0, V_{\max}]$. Suppose the full-domain approximation error is at most η_Q and the fitted objective gap is at most τ_Q. Let $H(\epsilon, \eta)$ be the sum of the fit and label log covers. Then

$$\sup_y \|\widehat{f}_y - Q_y\|^2_{L^2(\nu_X)} \le C\left\{\eta_Q^2 + \tau_Q + V_{\max}(\epsilon + \eta) + \frac{V_{\max}^2[H(\epsilon, \eta) + \log(16/\delta)]}{\mathcal{N}_b}\right\}.$$

For the linear state-feature derivative representation, put

$$\Sigma_{\mathrm{st}} = \mathbb{E}_{\nu_+} w_{\mathrm{st}} w_{\mathrm{st}}^\top, \qquad \widehat{\Sigma}_{\mathrm{st}} = n^{-1} \sum_{i,t} w_{\mathrm{st}}(S_{i,t+1}) w_{\mathrm{st}}(S_{i,t+1})^\top,$$

and suppose $\mathrm{rank}(\Sigma_{\mathrm{st}}) = r_{\mathrm{st}} \le d_S$ *and* $\operatorname{ess\,sup}_{\nu_+} w_{\mathrm{st}}^\top \Sigma_{\mathrm{st}}^\dagger w_{\mathrm{st}} \le C_{\mathrm{lev}} d_S$. *If* $\mathcal{N}_b \ge C d_S \log(16(d_S \vee 1)/\delta)$, *then*

$$\tfrac{1}{2}\Sigma_{\mathrm{st}} \preceq \widehat{\Sigma}_{\mathrm{st}} \preceq \tfrac{3}{2}\Sigma_{\mathrm{st}}. \tag{55}$$

The three conclusions hold jointly with probability at least $1 - \delta - \delta_b^{\mathrm{mix}}$, after choosing the universal constants consistently. The zero-rank case of (55) holds exactly. If the optional empirical derivative criterion of Appendix J is used, the state Gram event gives its relative comparison with $c_- = 1/2$ and $c_+ = 3/2$.

The block principle of Yu (1994) and unequal-range concentration of Hoeffding (1963, Theorem 2) are applied to all transitions through the two groups. This simultaneously changes the objective, regression and optional finite-feature curvature inputs of FQI, so the temporal gain is established for the fitted full-sample procedure.

Proof. All concentration arguments in this proof first take place under a group's product-block law; its block marginals are unchanged. For a fixed bounded f, a block contribution to $\widehat{\mathbb{P}}_c f$ has range length at most $B\ell_{i,j}/n$. Unequal-range Hoeffding concentration (Hoeffding, 1963, Theorem 2) and

$$\sum_{i,j} \frac{\ell_{i,j}^2}{n^2} \le \frac{b}{n} = \mathcal{N}_b^{-1}$$

give the two group deviations. Their sum is at most the right side of (54) on a deterministic net, by Cauchy–Schwarz over the two groups. The centered off-net extension costs 2ϵ for the full mean. This applies to unpenalized policy objectives with $f(Z) = Q\{S', \pi(S')\}$ and the full joint cover.

For prediction, fix f, y and define

$$d_{f,y}(Z) = (f(X) - Y_y(Z))^2 - (Q_y(X) - Y_y(Z))^2, \qquad \epsilon_{f,y}^2 = \mathbb{P}d_{f,y} = \|f - Q_y\|_{L^2(\nu_X)}^2,$$
$$\Delta_{i,j}^{f,y} = n^{-1} \sum_{t \in I_{i,j}} d_{f,y}(Z_{i,t}), \qquad \mu_{i,j} = \mathbb{E}\Delta_{i,j}^{f,y} = n^{-1} \sum_{t \in I_{i,j}} \mathbb{E}(f - Q_y)^2(X_{i,t}) \ge 0.$$

Compatibility and boundedness give $|d_{f,y}| \le V_{\max}^2$ and $d_{f,y}^2 \le 4V_{\max}^2 (f - Q_y)^2$. Jensen's inequality within each block therefore gives

$$|\Delta_{i,j}^{f,y} - \mu_{i,j}| \le 2V_{\max}^2 b/n, \qquad \mathbb{E}(\Delta_{i,j}^{f,y})^2 \le 4V_{\max}^2 (\ell_{i,j}/n)\mu_{i,j}.$$

Within group c the centered contributions are independent, and their variance sum is at most $4V_{\max}^2 \mathcal{N}_b^{-1} \epsilon_c^2$, where $\epsilon_c^2 = \sum_{\mathcal{I}_c} \mu_{i,j} \ge 0$ and $\epsilon_0^2 + \epsilon_1^2 = \epsilon_{f,y}^2$. Scalar Bernstein followed by Young's inequality yields

$$|(\widehat{\mathbb{P}}_c - \mathbb{P}_c) d_{f,y}| \le \tfrac{1}{4}\epsilon_c^2 + C V_{\max}^2 x / \mathcal{N}_b$$

with failure at most $2e^{-x}$. Sum the two inequalities and take product fit/label nets with $x = H(\epsilon, \eta) + \log(16/\delta)$. The common full-domain kernel gives $\|Q_y - Q_{y'}\|_\infty \le \|y - y'\|_\infty$; the loss and risk extensions cost $C V_{\max}(\epsilon + \eta)$. Thus, on a full-class event,

$$|(\widehat{\mathbb{P}} - \mathbb{P}) d_{f,y}| \le \tfrac{1}{4}\epsilon_{f,y}^2 + C\{V_{\max}(\epsilon + \eta) + V_{\max}^2 x / \mathcal{N}_b\}. \tag{56}$$

Apply the full-sample approximate minimization inequality to the fit and a deterministic approximation oracle:

$$\tfrac{3}{4}\epsilon_{\widehat{f}_y,y}^2 \le \tfrac{5}{4}\epsilon_{f_y^\circ,y}^2 + \tau_Q + C\{V_{\max}(\epsilon + \eta) + V_{\max}^2 x / \mathcal{N}_b\}. \tag{57}$$

Let the oracle tolerance decrease to zero. No equality of group design laws was used, and no empirical loss was fitted to only one group.

For the Gram bound, work on $\mathrm{Range}(\Sigma_{\mathrm{st}})$ and whiten by $\widetilde{w} = \Sigma_{\mathrm{st}}^{\dagger/2} w_{\mathrm{st}}$. The pooled law gives $\mathbb{E}_{\nu_+} \widetilde{w}\widetilde{w}^\top = \mathrm{Id}$ and $\|\widetilde{w}\|^2 \le \kappa_{\mathrm{st}} := C_{\mathrm{lev}} d_S$ almost surely. Every individual observed marginal shares these properties of the support and envelope because it is a positive-weight component of ν_+. For each block set

$$\Xi_{i,j} = \frac{1}{n} \sum_{t \in I_{i,j}} \widetilde{w}(S_{i,t+1}) \widetilde{w}(S_{i,t+1})^\top, \qquad \overline{\Xi}_c = \sum_{\mathcal{I}_c} \mathbb{E}\Xi_{i,j}. \tag{58}$$

Then $0 \preceq \Xi_{i,j} \preceq \kappa_{\mathrm{st}}/\mathcal{N}_b \,\mathrm{Id}$, $\overline{\Xi}_0 + \overline{\Xi}_1 = \mathrm{Id}$, and $0 \preceq \overline{\Xi}_c \preceq \mathrm{Id}$. A group mean can be singular and need not be a scalar multiple of Id. With $\Xi_{i,j}^\circ = \Xi_{i,j} - \mathbb{E}\Xi_{i,j}$, the group variance obeys

$$\|\Xi_{i,j}^\circ\| \le 2\kappa_{\mathrm{st}}/\mathcal{N}_b, \qquad \sum_{\mathcal{I}_c} \mathbb{E}(\Xi_{i,j}^\circ)^2 \preceq \sum_{\mathcal{I}_c} \mathbb{E}\Xi_{i,j}^2 \preceq (\kappa_{\mathrm{st}}/\mathcal{N}_b)\overline{\Xi}_c \preceq (\kappa_{\mathrm{st}}/\mathcal{N}_b)\,\mathrm{Id}\,.$$

Matrix Bernstein (Tropp, 2015, Theorem 6.1.1), at group deviation $1/4$, gives failure at most $2r_{\mathrm{st}} \exp(-c\mathcal{N}_b/d_S)$. Both group deviations imply a full deviation at most $1/2$. Transform back on

the nonzero eigenspace and retain the common null space. Integrating the same action-derivative quadratic form against the coefficient functions gives the optional empirical derivative comparison. The numerical derivative matrix, if used, still needs its own stated two-sided comparison.

Allocate the independent-product failure budget δ among these groupwise events, take their unions separately within each group, and apply Lemma L.1 once. This proves the joint probability and avoids paying the dependence comparison separately for each net element, statistic or fitted iteration. □

L.3 KRR USES THE SAME DEPENDENCE EVENT, WITH ITS OWN RIDGE CALIBRATION

The score and operator bounds have their own ridge calibration.

Proposition L.2 (Uniform KRR prediction under the full-sample block comparison)**.** *Let the measurable kernel have separable RKHS $\mathcal{F}$, diagonal bound κ_K^2, and full-domain targets satisfying $\sup_y \|Q_y\|_{\mathcal{F}} \le B_F$. Assume $Y_y, Q_y \in [0, V_{\max}]$, common conditional compatibility, and a full deterministic label log cover $H_Y(\eta)$. Write $\Sigma_K, \widehat{\Sigma}_K$ for the population and empirical RKHS second-moment operators and*

$$\mathbf{z}_y = \frac{1}{n}\sum_{i,t}\{Y_y(Z_{i,t}) - Q_y(X_{i,t})\}K_{X_{i,t}}.$$

Under Assumption L.1, one event of probability at least $1 - \delta - \delta_b^{\mathrm{mix}}$ gives

$$\sup_y \|\mathbf{z}_y\|_{\mathcal{F}} \le s_{K,n}, \qquad \|\widehat{\Sigma}_K - \Sigma_K\|_{\mathrm{op}} \le v_{K,n},$$

$$s_{K,n} = 2\kappa_K\eta + C\kappa_K V_{\max}\sqrt{\frac{H_Y(\eta) + \log(16/\delta) + 1}{\mathcal{N}_b}},$$

$$v_{K,n} = C\kappa_K^2\sqrt{\frac{\log(16/\delta) + 1}{\mathcal{N}_b}}.$$

The full-sample approximate ridge fit with objective gap τ_K has the simultaneous raw norm and prediction bounds

$$B_{K,n} = B_F + s_{K,n}/\varrho_n + \sqrt{\tau_K/\varrho_n}, \tag{59}$$

$$\delta_{K,n}^2 = \varrho_n B_F^2 + 2(B_{K,n} + B_F)s_{K,n} + \tau_K + v_{K,n}(B_{K,n} + B_F)^2. \tag{60}$$

Clipping does not increase error toward the bounded target. If $\varrho_n \to 0$, $s_{K,n} + v_{K,n} = O(\varrho_n)$ and $\tau_K = O(\varrho_n)$, then $B_{K,n} = O(1)$ and $\delta_{K,n} = O(\varrho_n^{1/2})$. The raw-kernel secant conditions, when verified with controlled constants, give evaluation error $\widetilde{O}(\varrho_n^{\vartheta_*/2})$.

This uses the same block comparison as Yu (1994), now on the ridge score and second-moment terms. The fit-and-target uniformity emphasized by Nguyen-Tang et al. (2022) is retained, and the fitted-norm bound supplies the coefficient radius required by the functional secant calculation.

Proof. Fix a group c and first work under its product-block law. Write

$$\mathbf{z}_{i,j,y}^{\mathrm{blk}} = \frac{1}{n}\sum_{t\in I_{i,j}}\{Y_y(Z_{i,t}) - Q_y(X_{i,t})\}K_{X_{i,t}}, \qquad \mathbf{z}_{c,y}^{\mathrm{blk}} = \sum_{\mathcal{I}_c}\mathbf{z}_{i,j,y}^{\mathrm{blk}}, \qquad \ell_c^{(2)} = \sum_{\mathcal{I}_c}\ell_{i,j}^2/n^2.$$

Conditional compatibility gives $\mathbb{E}\mathbf{z}_{i,j,y}^{\mathrm{blk}} = 0$, without independence within a block. Thus

$$\|\mathbf{z}_{i,j,y}^{\mathrm{blk}}\| \le \kappa_K V_{\max}\ell_{i,j}/n, \quad \mathbb{E}\|\mathbf{z}_{c,y}^{\mathrm{blk}}\|^2 = \sum_{\mathcal{I}_c}\mathbb{E}\|\mathbf{z}_{i,j,y}^{\mathrm{blk}}\|^2 \le \kappa_K^2 V_{\max}^2\ell_c^{(2)}, \quad \ell_0^{(2)} + \ell_1^{(2)} \le \mathcal{N}_b^{-1}.$$

The cross terms vanish by product-block independence. Replacing one block changes $\|\mathbf{z}_{c,y}^{\mathrm{blk}}\|$ by at most $2\kappa_K V_{\max}\ell_{i,j}/n$. Bounded differences (Boucheron et al., 2013) and Jensen therefore imply

$$\Pr_{\otimes,c}\left\{\|\mathbf{z}_{c,y}^{\mathrm{blk}}\| > \kappa_K V_{\max}\sqrt{\ell_c^{(2)}}(1 + \sqrt{2x})\right\} \le e^{-x}.$$

An empty group contributes zero. Use $x = H_Y(\eta) + \log(16/\delta)$ and a deterministic label net in each group. Since $\|Q_y - Q_{y'}\|_\infty \le \|y - y'\|_\infty$, the two groups' off-net costs sum to $2\kappa_K\eta$, not a term proportional to the number of blocks. Cauchy–Schwarz gives $\sqrt{\ell_0^{(2)}} + \sqrt{\ell_1^{(2)}} \le \sqrt{2/\mathcal{N}_b}$, proving the score radius.

For the second-moment operator, let $\mathbf{K}_{i,j}^{\mathrm{blk}} = n^{-1}\sum_{t\in I_{i,j}} K_{X_{i,t}} \otimes K_{X_{i,t}}$ and $\Delta_c^\Sigma = \sum_{\mathcal{I}_c}(\mathbf{K}_{i,j}^{\mathrm{blk}} - \mathbb{E}\mathbf{K}_{i,j}^{\mathrm{blk}})$. The Hilbert–Schmidt identity $\|K_X \otimes K_X\|_{\mathrm{HS}} = \|K_X\|^2 \le \kappa_K^2$ gives

$$\mathbb{E}\|\Delta_c^\Sigma\|_{\mathrm{HS}}^2 \le \sum_{\mathcal{I}_c} \mathbb{E}\|\mathbf{K}_{i,j}^{\mathrm{blk}}\|_{\mathrm{HS}}^2 \le \kappa_K^4 \ell_c^{(2)}.$$

The replacement bound is $2\kappa_K^2 \ell_{i,j}/n$. The same norm argument, now with $x = \log(16/\delta)$ and no label net, controls $\|\Delta_c^\Sigma\|_{\mathrm{HS}}$. Summing groups and using operator norm at most Hilbert–Schmidt norm yields $v_{K,n}$. Within each group combine the score and second-moment bad events, then apply Lemma L.1 once. These score and operator bounds are the only sampling-specific inputs to the ridge argument. Substitute them in the normal-equation and approximate-minimizer bounds of Proposition H.3 to obtain (59)–(60); its algebra uses the original full-sample objective, not a fitted block subsample. Corollary H.1 then supplies the boundary calibration. The label cover remains fixed before fitting, and the same raw-secant and clipping argument gives evaluation. □

The common label entropy is fixed before fitting. Combining these product-law events with the objective events before the group comparison pays the same dependence failure only once.

L.4 Retained-block comparison and omission costs

Under Assumption L.1, consider the following alternative to the full-sample two-group partition. Retain the first b_{ret} transitions in each complete cycle of length $b_{\mathrm{ret}} + q_{\mathrm{ret}}$, with $b_{\mathrm{ret}}, q_{\mathrm{ret}} \ge 1$:

$$m_{i,\mathrm{ret}} = \lfloor T/(b_{\mathrm{ret}} + q_{\mathrm{ret}})\rfloor, \quad I_{i,j}^{\mathrm{ret}} = \{(j-1)(b_{\mathrm{ret}} + q_{\mathrm{ret}}) + 1, \dots, (j-1)(b_{\mathrm{ret}} + q_{\mathrm{ret}}) + b_{\mathrm{ret}}\},$$
$$\mathcal{N}_{\mathrm{ret}} = \sum_i m_{i,\mathrm{ret}}, \qquad \mathsf{o}_{\mathrm{ret}} = 1 - b_{\mathrm{ret}}\mathcal{N}_{\mathrm{ret}}/n.$$

Equivalently, $\mathcal{N}_{\mathrm{ret}} = N_{\mathrm{eff}}(b_{\mathrm{ret}}, q_{\mathrm{ret}})$ and $\mathsf{o}_{\mathrm{ret}} = \mathsf{o}_{b_{\mathrm{ret}},q_{\mathrm{ret}}}$; all incomplete cycles are included in $\mathsf{o}_{\mathrm{ret}}$.

Consecutive retained blocks are separated by $q_{\mathrm{ret}} + 1$ indices. Apply the product-law argument in Lemma L.1 with this separation and use monotonicity of the mixing envelope. Successive maximal coupling on standard Borel spaces (Berbee, 1979, Corollary 4.2.5) then gives independent copies with the same block marginals and joint failure at most

$$\delta_{\mathrm{coup}}(b_{\mathrm{ret}}, q_{\mathrm{ret}}) \le \sum_i (m_{i,\mathrm{ret}} - 1)_+ \beta_{\mathrm{mix}}(q_{\mathrm{ret}}). \tag{61}$$

On this one event all deterministic blockwise function evaluations agree. For $|f| \le B$, retained and omitted averages satisfy

$$\widehat{\mathbb{P}} = (1 - \mathsf{o}_{\mathrm{ret}})\widehat{\mathbb{P}}_{\mathrm{ret}} + \mathsf{o}_{\mathrm{ret}}\widehat{\mathbb{P}}_{\mathrm{omit}},$$
$$\sup_f |(\widehat{\mathbb{P}} - \mathbb{P})f| \le 4B\mathsf{o}_{\mathrm{ret}} + \sup_f |(\widehat{\mathbb{P}}_{\mathrm{ret}} - \mathbb{P}_{\mathrm{ret}})f|. \tag{62}$$

The copied block means have $\mathcal{N}_{\mathrm{ret}}$ independent units. Hoeffding concentration and a deterministic cover therefore give the objective bound with scale $\mathcal{N}_{\mathrm{ret}}$ and the additional $4B\mathsf{o}_{\mathrm{ret}}$, at failure $\delta + \delta_{\mathrm{coup}}(b_{\mathrm{ret}}, q_{\mathrm{ret}})$. The fitted objective continues to use all transitions and $\mathbb{P}_{\mathrm{ret}}$ is not identified with $\mathbb{P}$.

With $\mathcal{N}_{\mathrm{ret}} \ge 1$, the full-sample prediction bound is

$$\delta_{Q,n}^2 \le C\left\{\eta_Q^2 + \tau_Q + V_{\max}(\epsilon + \eta) + \frac{V_{\max}^2[H(\epsilon,\eta) + \log(C/\delta)]}{\mathcal{N}_{\mathrm{ret}}} + V_{\max}^2\mathsf{o}_{\mathrm{ret}}\right\}. \tag{63}$$

The objective comparison (62) carries an $O(\mathsf{o}_{\mathrm{ret}})$ term. The full-sample relative Gram event additionally requires

$$\mathcal{N}_{\mathrm{ret}} \ge Cd_S\log(C(d_S \vee 1)/\delta), \qquad Cd_S\mathsf{o}_{\mathrm{ret}} \le \tfrac{1}{8}. \tag{64}$$

All three events have joint failure at most $\delta + \delta_{\mathrm{coup}}(b_{\mathrm{ret}}, q_{\mathrm{ret}})$.

For completeness, normalize retained and omitted sums by the full n, writing $\widehat{\mathbb{P}} = \widehat{\mathbb{P}}^0_{\mathrm{ret}} + \widehat{\mathbb{P}}^0_{\mathrm{omit}}$ and likewise for their expectations. For the excess loss $d = d_{f,y}$, compatibility gives $\mathbb{P}^0_{\mathrm{ret}} d, \mathbb{P}^0_{\mathrm{omit}} d \geq 0$. The same blockwise Bernstein and product-net argument as (56) yields

$$|(\widehat{\mathbb{P}}^0_{\mathrm{ret}} - \mathbb{P}^0_{\mathrm{ret}})d| \leq \tfrac{1}{4}\mathbb{P}^0_{\mathrm{ret}} d + C\{V_{\max}(\epsilon + \eta) + V^2_{\max} x/\mathcal{N}_{\mathrm{ret}}\}.$$

Since $|d| \leq V^2_{\max}$, the omitted centered term is at most $2V^2_{\max}\mathsf{o}_{\mathrm{ret}}$. Use $\mathbb{P}^0_{\mathrm{ret}} d \leq \mathbb{P}d$ and the full empirical minimization inequality (57) to obtain (63); no equality of the design laws is used. For Gram control, globally whiten as in (58). Retained block matrices are bounded by $Cd_S b_{\mathrm{ret}}/n \leq Cd_S/\mathcal{N}_{\mathrm{ret}}$ and their total expectation is at most Id. Matrix Bernstein controls their deviation by $1/4$ under the sample-size condition in (64). The omitted empirical and population matrices each have norm at most $Cd_S\mathsf{o}_{\mathrm{ret}}$, establishing the sufficiency of the omitted-mass condition $Cd_S\mathsf{o}_{\mathrm{ret}} \leq 1/8$. Transfer this one joint event from the product law with cost $\delta_{\mathrm{coup}}(b_{\mathrm{ret}}, q_{\mathrm{ret}})$.

KRR score and second-moment bounds. Retain the full-domain target-ball and kernel conditions of Proposition H.3. Replace (39)–(40) by

$$s^{\mathrm{RB}}_n = 2\kappa_K\eta + C\kappa_K V_{\max}\left[\mathsf{o}_{\mathrm{ret}} + \sqrt{\{H_Y(\eta) + \log(8/\delta) + 1\}/\mathcal{N}_{\mathrm{ret}}}\right],$$
$$v^{\mathrm{RB}}_n = C\kappa^2_K\left[\mathsf{o}_{\mathrm{ret}} + \sqrt{\{\log(8/\delta) + 1\}/\mathcal{N}_{\mathrm{ret}}}\right].$$

With these substitutions, the radius, prediction and common evaluation bounds hold for the full-sample KRR fit with probability at least $1-\delta-\delta_{\mathrm{coup}}(b_{\mathrm{ret}}, q_{\mathrm{ret}})$. Boundary tuning from Corollary H.1, with N replaced by $\mathcal{N}_{\mathrm{ret}}$ in its verified full-label entropy and radii, gives a bounded fitted norm if additionally $\mathsf{o}_{\mathrm{ret}} = O(\varrho_n)$ and $\delta_{\mathrm{coup}}(b_{\mathrm{ret}}, q_{\mathrm{ret}}) \to 0$. The stricter $\mathsf{o}_{\mathrm{ret}} = o(\varrho_n)$ retains the interior conclusion $B_{K,n} = B_F + o(1)$ when the other interior premises hold.

Apply the Hilbert-valued score and second-moment concentration arguments of Proposition H.3 to the independent copied blocks. Their envelopes are $\kappa_K V_{\max}$ and κ^2_K, and their independent-unit count is $\mathcal{N}_{\mathrm{ret}}$. On the one shared coupling event, decompose the original full-sample statistics into retained and omitted averages:

$$\mathbf{z}_y = (1 - \mathsf{o}_{\mathrm{ret}})\mathbf{z}_{\mathrm{ret},y} + \mathsf{o}_{\mathrm{ret}}\mathbf{z}_{\mathrm{omit},y}, \qquad \widehat{\Sigma}_K = (1 - \mathsf{o}_{\mathrm{ret}})\widehat{\Sigma}_{K,\mathrm{ret}} + \mathsf{o}_{\mathrm{ret}}\widehat{\Sigma}_{K,\mathrm{omit}}.$$

Bounded sections give $\sup_y \|\mathbf{z}_y - \mathbf{z}_{\mathrm{ret},y}\| \leq 2\kappa_K V_{\max}\mathsf{o}_{\mathrm{ret}}$ and $\|\widehat{\Sigma}_K - \widehat{\Sigma}_{K,\mathrm{ret}}\|_{\mathrm{op}} + \|\Sigma_K - \Sigma_{K,\mathrm{ret}}\|_{\mathrm{op}} \leq 4\kappa^2_K\mathsf{o}_{\mathrm{ret}}$. These are precisely the additive terms in the stated radii. The full-sample ridge normal equation and basic inequality are unchanged, so (41)–(42) apply with the substituted radii. Dividing the omission terms by ϱ_n proves the boundary or interior conclusion, respectively. The failure is $\delta + \delta_{\mathrm{coup}}(b_{\mathrm{ret}}, q_{\mathrm{ret}})$, without multiplying the coupling cost by the label cover.

L.5 Dependence scales and value consequences

For bounded joint entropy cost h_n, the proved concentration inputs are $u_n = O(\sqrt{h_n/\mathcal{N}_b})$ and a squared prediction remainder $O(h_n/\mathcal{N}_b)$. Together with pair membership, the numerical-input and policy-optimization bounds in (7), and $\mathcal{R}_n$, these uniform statistical events imply $\mathcal{E}_n$. Theorem G.1 therefore accepts these errors on the same deterministic localized policy class, with dependence failure δ^{mix}_b included in the bound for $\Pr(\mathcal{E}^c_n)$; under stable propagation, Theorem 1 applies. For either the finite-state product class or the complete continuous-state RKHS ball of Appendix K, replace N by $\mathcal{N}_b$ in the entropy balance, retune the deterministic critic classes and tolerances, and choose $M \geq c\log\mathcal{N}_b$ for sufficiently large c. The corresponding exponent is $p = \min\{\vartheta_*, 1/2\}/(2 + D)$, with $D = d_{\mathrm{app}}$ or D_{ct}; retain $\sqrt{\Delta_{\varepsilon_{\mathrm{app}}}}$ in the coupled case. Corollary K.3 gives the analogous exponent when critic factors grow. All information, state-transfer and approximation conditions are imposed at this retuned scale for the same pooled laws. This substitution follows from the full-sample objective and prediction bounds above. The deterministic policy penalty requires no empirical-to-population comparison; the fixed-radius norm-localization condition is retuned at the same effective sample size.

If $\beta_{\mathrm{mix}}(k) \leq C_m \exp(-c_m k^{p_{\mathrm{mix}}})$, $p_{\mathrm{mix}} > 0$, choose

$$b = \min\{T, \lceil [c\log(en)]^{1/p_{\mathrm{mix}}} \rceil\}$$

with sufficiently large c. If uncapped, $\delta_b^{\mathrm{mix}} \le (n/b)\beta_{\mathrm{mix}}(b) \le n^{-2}$; if capped, each subject has a single block and $\delta_b^{\mathrm{mix}} = 0$. Thus $\mathcal{N}_b \asymp N \max\{1, T/\log^{1/p_{\mathrm{mix}}}(en)\}$, up to constants and integer rounding. Assumption 5 is the case $p_{\mathrm{mix}} = 1$, for which $b = O(\log(en))$. Hence the approximation scale $\mathcal{N}_b^{-1/(2+D)}$ equals $b^{1/(2+D)}(NT)^{-1/(2+D)}$. All NT transitions remain in the fitted objectives.

For Example 3 and its continuous-state extension below, take $D = D_{\mathrm{Ada}}$ and $D = D_{\mathrm{ct}}$, respectively. Choose approximation accuracy $\varepsilon_{\mathrm{app},n}$ of order $(NT)^{-1/(2+D)}$ times a sufficiently large logarithmic factor incorporating $b^{1/(2+D)}$, the covering terms, and the confidence level. Require critic squared-loss fitting error $O(\varepsilon_{\mathrm{app},n}^2)$, numerical input error $O(\varepsilon_{\mathrm{app},n})$, and policy optimization error and penalty coefficient $O(\varepsilon_{\mathrm{app},n})$, with the full-domain approximation, covering, localization, and bounded-transfer conditions of the corresponding structural corollary. The uniform logged prediction error $\delta_{\mathrm{Ada},n}$ and critic-update error then satisfy

$$\delta_{\mathrm{Ada},n} = \widetilde{O}((NT)^{-1/(2+D)}), \qquad \varepsilon_{Q,n} = \widetilde{O}((NT)^{-\vartheta_*/(2+D)}).$$

The policy error is $\widetilde{O}(\varepsilon_{\mathrm{app},n}^{1/2} + \sqrt{\Delta_n})$. With $M \ge c\log(NT)$ for sufficiently large fixed c, Theorem 1 gives the value rate in Example 3 and the continuous-state bound below. The one-score choices $d = s_h = s_b = 1$ and $p_{\mathrm{Lip}} = 0$ give $D_{\mathrm{Ada}} = 1$; when $\vartheta_* \ge 1/2$ and $\Delta_n = 0$, the finite-state excess rate is $(NT)^{-1/6}$ up to logarithms.

Continuous-state value rate. For $s \in [0,1]$ and $\mathcal{H} = H^1([0,1]; H^2(0,1))$, let the outer targets depend on the state and d functional scores. With the smoothness orders s_h, s_b and fitted-class Lipschitz exponent p_{Lip} specified in Appendix K.1, the joint complexity exponent is

$$D_{\mathrm{ct}} = \max\{(1+d)/s_h,\ 1/s_b,\ 5 + 5p_{\mathrm{Lip}}/2\}.$$

Under the full-policy-ball approximation, covering, localization, and state-transfer conditions of that appendix, Assumption 5, and the preceding computational calibration,

$$J^\star - J(\widehat{\pi}_M) \le a_{\mathcal{H}} + \widetilde{O}_p\Big((NT)^{-\min\{\vartheta_*,1/2\}/(2+D_{\mathrm{ct}})} + \sqrt{\Delta_n}\Big).$$

For $d = s_h = s_b = 1$, $p_{\mathrm{Lip}} = 0$, $\vartheta_* \ge 1/2$, and $\Delta_n = 0$, one has $D_{\mathrm{ct}} = 5$ and excess value rate $\widetilde{O}_p((NT)^{-1/14})$. Its slower exponent reflects the larger policy-class complexity. These rates are sufficient and are not asserted to be optimal.

If $\beta_{\mathrm{mix}}(k) \le C_m k^{-p_{\mathrm{poly}}}$, $p_{\mathrm{poly}} > 0$, take

$$b = \min\{T, \lceil [C_m n \log^2(en)]^{1/(1+p_{\mathrm{poly}})} \rceil\}.$$

Then $\delta_b^{\mathrm{mix}} \le \log^{-2}(en)$ when uncapped, and

$$\mathcal{N}_b \asymp \max\{N,\ n^{p_{\mathrm{poly}}/(1+p_{\mathrm{poly}})} \log^{-2/(1+p_{\mathrm{poly}})}(en)\}.$$

For prescribed failure ϵ_n, replace $\log^2(en)$ by $1/\epsilon_n$. The confidence level determines the corresponding scale explicitly. With $T \asymp N^{p_{\mathrm{len}}}$, a verified power p of this scale becomes, up to logarithms, $N^{-\max\{1,(1+p_{\mathrm{len}})p_{\mathrm{poly}}/(1+p_{\mathrm{poly}})\}p}$. For unequal lengths the same block proof uses $T_{\max}$ as the cap and $\mathcal{N}_\infty = n/T_{\max}$ in place of N in these maxima. All approximation, information and state-transfer conditions continue to refer to the corresponding pooled laws.

For the retained-block alternative, its proved prediction radius additionally includes $V_{\max}\sqrt{\mathsf{o}_{\mathrm{ret}}}$ and its objective bound includes $O(\mathsf{o}_{\mathrm{ret}})$. Apply the main theorem with these actual inputs and the separately verified norm-localization event; no equality between retained and full design laws is needed. The temporal calculation in Appendix M.6 verifies the stronger main-text rate for Example 1 using the full-sample construction.

M VERIFICATION OF EXAMPLE 1

All calculations in this section concern Example 1, with the same MDP, behavior law, Sobolev-valued policy RKHS, fixed comparison ball, and full functional linear critic kernel.

M.1 The model, policy class, behavior support and moments

Take $\mathcal{S} = \{1, 2\}$ and $\mathcal{A} = \{a \in L^2(0,1) : \|a\|_{L^2} \le B_A\}$. For the discount factor $\gamma \in (0,1)$, define

$$q_\circ(u) = e^u - (e-1), \qquad v_\circ = \frac{q_\circ}{\|q_\circ\|_{L^2}},$$
$$z(a) = \langle v_\circ, a\rangle_{L^2}, \qquad \iota = \frac{\min\{1, \sqrt{\gamma^{-2}-1}\}}{8B_A}.$$

The smooth function $v_\circ \in H^2(0,1)$ has unit L^2 norm and zero integral. Specify the reward and next-state laws by

$$R \mid (S=1, A=a) \sim \text{Bernoulli}(1/4), \qquad R \mid (S=2, A=a) \sim \text{Bernoulli}(3/4),$$
$$P_S(2 \mid 1, a) = \tfrac{1}{2} + \iota z(a), \qquad P_S(2 \mid 2, a) = \tfrac{1}{2} - \iota z(a), \qquad P_S(1 \mid s, a) = 1 - P_S(2 \mid s, a).$$

Use fresh reward and transition randomization at each time, independently conditional on the current state and action.

Let $\mathcal{G} = H^2(0,1)$ with $\|a\|_{\mathcal{G}}^2 = \|a\|_{L^2}^2 + \|a''\|_{L^2}^2$ and $\mathcal{H} = \mathcal{G} \oplus \mathcal{G}$. With $T_{\mathcal{G}}$ defined by $\langle T_{\mathcal{G}} a, b\rangle_{\mathcal{G}} = \langle a, b\rangle_{L^2}$, the L^2-valued operator kernel is $\Gamma(s,t) = \mathbf{1}\{s=t\}T_{\mathcal{G}}$. The base class is the product of the feasible L^2 action balls within $\mathcal{H}$. The analytical comparison class is

$$\Pi_{\mathcal{H}}(B_{\mathcal{H}}) = \left\{\pi \in \Pi : \sum_{s=1}^{2} \|\pi(s)\|_{\mathcal{G}}^2 \le B_{\mathcal{H}}^2\right\}.$$

Consequently every selected action has $\mathcal{G}$ norm at most $B_{\mathcal{H}}$. This localized class is compact for statewise L^2 convergence: bounded H^2 sets are weakly compact and embed compactly into L^2, and the constraints are closed.

Choose $\varsigma_A > 0$ so that $\varsigma_A^2 \sum_{j\ge1} j^{-9/2} \le B_A^2$. With independent symmetric signs ξ_j and the normalized shifted Legendre basis $\{\phi_j\}$, draw $A = \varsigma_A \sum_{j\ge1} j^{-9/4}\xi_j\phi_j$. This behavior action series converges in L^2 and is feasible. Its mean is zero and covariance is $\varsigma_A^2 \mathbf{W}_{9/4}$. Fresh draws are independent of current states and past records. The behavior-averaged transition row is $(1/2, 1/2)$, so uniform initial states give $\nu_X = \frac{1}{2}(\delta_1 + \delta_2) \otimes \mathcal{L}(A)$ and $\nu_+ = \frac{1}{2}(\delta_1 + \delta_2)$ at every time.

For every support point of $\mathcal{L}(A)$, its jth coefficient has magnitude $\varsigma_A j^{-9/4}$. Indeed the series is a continuous map from the compact sign product space into L^2, since its squared tails converge uniformly; its image is closed and is the support of the pushforward product-sign law. Thus $\sum_j j^4 |\langle A, \phi_j\rangle|^2 = \varsigma_A^2 \sum_j j^{-1/2} = \infty$ on the entire support. Every action supplied by $\Pi_{\mathcal{H}}(B_{\mathcal{H}})$ instead satisfies the finite bound $\|a\|_\zeta \le \sqrt{c_{\mathcal{G}}} B_{\mathcal{H}}$ at $\zeta = 2$.

To show a positive distance, choose a finite m with $\varsigma_A(\sum_{j\le m} j^{-1/2})^{1/2} > 2\sqrt{c_{\mathcal{G}}} B_{\mathcal{H}}$. For a support point x and a policy-selected action a, the triangle inequality in the weighted first-m coordinates gives

$$m^2 \|x - a\|_{L^2} \ge \left(\sum_{j\le m} j^4 |\langle x-a, \phi_j\rangle|^2\right)^{1/2} \ge \varsigma_A \left(\sum_{j\le m} j^{-1/2}\right)^{1/2} - \sqrt{c_{\mathcal{G}}} B_{\mathcal{H}} > 0.$$

This one bound applies to all actions selected by the localized policy class and all support points.

M.2 Full linear critics and spectral transfer

The critic RKHS of $K\{(s,a),(s',a')\} = \mathbf{1}\{s = s'\}(1 + \langle a, a'\rangle_{L^2})$ is isometric to $(\mathbb{R} \oplus L^2)^2$. Each critic has an intercept and arbitrary L^2 slope at each state. The bounded action ball makes $K(x,x) \le 1 + B_A^2$. The feature span is the whole coefficient space, and

$$\Sigma = \tfrac{1}{2}\,\text{diag}(1, \varsigma_A^2 \mathbf{W}_{9/4}, 1, \varsigma_A^2 \mathbf{W}_{9/4}), \qquad \mathbf{W} = \text{diag}(1, \mathbf{W}_{9/4}, 1, \mathbf{W}_{9/4}).$$

For every coefficient direction, $\langle d, \Sigma d\rangle \ge \frac{1}{2}\min\{1, \varsigma_A^2\}\langle d, \mathbf{W} d\rangle$. This is a quadratic-form comparison in the same coordinates. For $\vartheta = 8/9$, $2(9/4)\vartheta = 4$ and

$$\int \sup_{\pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})} \|\mathbf{W}^{-\vartheta/2} K_{(s,\pi(s))}\|^2 \nu_+(ds) \le 1 + c_{\mathcal{G}} B_{\mathcal{H}}^2.$$

Proposition F.1 therefore gives uniform Hölder transfer for raw coefficient secants in any fixed radius ball. When the target is bounded in $[0, V_{\max}]$, clipping a fitted raw critic does not increase its pointwise error to the target.

The same family has no finite uniform linear error ratio. For $j \geq 3$ put $p_j(u) = \int_0^u (u-t)\phi_j(t)\,dt$. Writing $c_j = \{2\sqrt{(2j-1)(2j+1)}\}^{-1}$, Legendre integration gives

$$p_j = c_j c_{j+1}\phi_{j+2} - (c_j^2 + c_{j-1}^2)\phi_j + c_{j-1}c_{j-2}\phi_{j-2}. \tag{65}$$

For $j \geq 2$, the Legendre recurrence and derivative identities (NIST, 2026, Section 14.10) give $\int_0^u \phi_j(t)\,dt = c_j\phi_{j+1}(u) - c_{j-1}\phi_{j-1}(u)$; the constant is zero at $u = 0$ by endpoint values. Integrating again proves (65). The coefficients have order j^{-2}, $p_j'' = \phi_j$, and hence $\|p_j\|_{L^2} = O(j^{-2})$, $\|p_j\|_{\mathcal{G}} = O(1)$ and $|\langle p_j, \phi_j\rangle| \asymp j^{-2}$.

A single sufficiently small positive multiple of every p_j belongs to $\Pi_{\mathcal{H}}(B_{\mathcal{H}})$, taking the other state's action to be zero. Indeed, $\sup_j \|p_j\|_{\mathcal{G}} < \infty$, so one fixed multiplier can satisfy both action feasibility and the policy-norm bound. For a fixed amplitude $\kappa_0 > 0$ chosen small enough that clipping remains inactive, the critic differences $\Delta_j(s,a) = \kappa_0 \mathbf{1}\{s=1\}\langle a, \phi_j\rangle$ then satisfy

$$\epsilon_j = \kappa_0 \varsigma_A j^{-9/4}/\sqrt{2}, \qquad cj^{-2} \leq \upsilon_j \leq Cj^{-2}.$$

The upper bound uses the Sobolev spectral embedding and the lower bound the preceding feasible actions. Thus $\upsilon_j/\epsilon_j \asymp j^{1/4} \to \infty$. Also $\upsilon_j/\epsilon_j^{\vartheta'} \to \infty$ for every $\vartheta' > 8/9$, so this exponent cannot be improved uniformly over these bounded coefficient secants. They can be realized between bounded critics by adding a fixed interior intercept and using the stated small amplitude κ_0.

M.3 State stability and the globally optimal policy

The smooth normalized function $\upsilon_\circ$ has unit L^2 norm, zero first Legendre coefficient and infinitely many nonzero coefficients; otherwise the exponential minus its mean would be a polynomial. For all feasible actions, $|z(a)| \leq B_A$ and $\iota B_A \leq 1/8$, so all transition probabilities lie in $[3/8, 5/8]$. Uniform ν_+ gives, for every probability occupancy,

$$C_{\mathrm{occ}} \leq \sqrt{2}, \qquad \sum_{t=1}^{2} \frac{P_S(t \mid s,a)^2}{1/2} = 1 + 4\iota^2 z(a)^2 \leq 1 + 4\iota^2 B_A^2.$$

The state-density calculation in Appendix B gives $C_B \leq \sqrt{1 + 4\iota^2 B_A^2}$ for every residual, including those involving $Q^\dagger$. The chosen ι implies $\chi^2 \leq \gamma^2[1 + \min\{1, \gamma^{-2} - 1\}/16] < 1$.

For a continuation vector $c \in [0, V_{\max}]^2$, the full-domain targets are

$$\begin{aligned} Q_c(1,a) &= \tfrac{1}{4} + \gamma(c_1 + c_2)/2 + \gamma\iota(c_2 - c_1)\langle \upsilon_\circ, a\rangle, \\ Q_c(2,a) &= \tfrac{3}{4} + \gamma(c_1 + c_2)/2 - \gamma\iota(c_2 - c_1)\langle \upsilon_\circ, a\rangle. \end{aligned} \tag{66}$$

At the global optimum the two maximized continuation terms coincide, so $V^\star(2) - V^\star(1) = 1/2$. The actions $\pi^\star(1) = B_A\upsilon_\circ$ and $\pi^\star(2) = -B_A\upsilon_\circ$ maximize those terms by Cauchy–Schwarz. This gives a solution of the global Bellman equations, which is unique by discounted contraction. The actions lie in the base class Π. Moreover, $\|\pi^\star\|_{\mathcal{H}}^2 = 2B_A^2\|\upsilon_\circ\|_{\mathcal{G}}^2$, so the condition $B_{\mathcal{H}}^2 \geq 2B_A^2\|\upsilon_\circ\|_{\mathcal{G}}^2$ places them in $\Pi_{\mathcal{H}}(B_{\mathcal{H}})$. Thus $J_{\mathcal{H}}^\star = J^\star$ equals the global value and $a_{\mathcal{H}} = 0$ in this example.

M.4 Full-linear ridge fitting for coverage and value

Every possible continuation from any clipped critic and any policy is a vector $c \in [0, V_{\max}]^2$. The labels $Y_c = R + \gamma c_{S'}$ have log covering number at most $2\log(1 + CV_{\max}/\epsilon)$, independently of the fitted critic radius. Formula (66) places every target in a fixed ball of the full critic RKHS: intercepts and slopes are bounded by constants depending only on γ, ι. All numerical directions of the fitted critic remain available; the known $\upsilon_\circ$ only defines the target dynamics.

Use $\delta_N = N^{-2}$, net accuracy N^{-1} and ridge $\varrho_n = c_\varrho\sqrt{\log(eN)/N}$ for a sufficiently large fixed c_ϱ. The score and operator bounds of Proposition H.3 are $O(\varrho_n)$. A KRR objective gap $O(\varrho_n)$ gives $B_{K,n} = O(1)$ and $\delta_{K,n} = \widetilde{O}(N^{-1/4})$ on one full-label event. No positive minimum eigenvalue of

the infinite covariance is used. Its bounded raw coefficients also give a common action-Lipschitz constant for all fitted clipped critics on this event. Put

$$\eta_N := C\delta_{K,n}^{8/9} = \widetilde{O}(N^{-2/9}), \tag{67}$$

where C is a sufficiently large fixed constant covering the uniform coefficient radii. The spectral calculation above gives $\varepsilon_{Q,n} \le \eta_N$.

The same calculation also supplies a policy-norm-dependent bound before localization is known. Write a raw fitted critic and its target at state s as $\widehat{b}_{s,c} + \langle \widehat{d}_{s,c}, a\rangle$ and $b_{s,c} + \langle d_{s,c}, a\rangle$, respectively. Because the logged signs are centered and have diagonal covariance, the prediction bound implies, uniformly in s and c,

$$|\widehat{b}_{s,c} - b_{s,c}| \le C\delta_{K,n}, \qquad \|\mathbf{W}_{9/4}^{1/2}(\widehat{d}_{s,c} - d_{s,c})\|_{L^2} \le C\delta_{K,n}.$$

The fitted and target RKHS radii also give $\|\widehat{d}_{s,c} - d_{s,c}\|_{L^2} = O(1)$. Interpolating these two slope norms at exponent $8/9$, pairing against the order-two spectral norm of a, and using the Sobolev–spectral embedding yields

$$|\widehat{Q}_c(s,a) - Q_c(s,a)| \le C\eta_N\{1 + \|a\|_{\mathcal{G}}\}, \qquad \|a\|_{L^2} \le B_A, \quad a \in H^2(0,1). \tag{68}$$

Clipping does not enlarge the left side because $Q_c \in [0, V_{\max}]$. Unlike a supremum over an arbitrary critic ball, (68) uses the prediction-error bound for the actual KRR fits and therefore becomes sharper as N grows.

M.5 Policy-norm self-localization and the value rate

For each continuation vector c, the target in (66) has a pointwise maximizing action $a_c^\star(s) \in \{0, \pm B_A v_\circ\}$ and

$$\|a_c^\star(s)\|_{\mathcal{G}}^2 \le B_0 := B_A^2 \|v_\circ\|_{\mathcal{G}}^2.$$

A conforming spline action $\bar{a}_{c,d_u}(s)$ can be chosen with $\mathcal{G}$ norm at most $B_0^{1/2}$ and $\|\bar{a}_{c,d_u}(s) - a_c^\star(s)\|_{L^2} \le C d_u^{-2}$.

Let $\widehat{\mu}_s = n^{-1}\sum_{i,t} \mathbf{1}\{S_{i,t+1} = s\}$. Independent bounded subject averages imply $\Pr\{\min_s \widehat{\mu}_s < 1/4\} \le 4e^{-N/8}$. On the complementary event the empirical policy objective is statewise separable. The critic term at state s has factor $\widehat{\mu}_s$, whereas the product-space norm contributes the unweighted term $\lambda_{\pi,n}\|a\|_{\mathcal{G}}^2$. Let $R(a) = \|a\|_{\mathcal{G}}^2$. Comparing a computed statewise update $\widehat{a}_s$ with $\bar{a}_{c,d_u}(s)$, using target optimality and (68), gives

$$\lambda_{\pi,n} R(\widehat{a}_s) \le C\eta_N\{1 + \sqrt{R(\widehat{a}_s)}\} + C\{d_u^{-2} + \lambda_{\pi,n} B_0 + \tau_{\text{solver}} + \epsilon_{\text{num}}\}. \tag{69}$$

Here statewise solver errors are obtained directly, or from a total nonnegative gap by dividing by $\min_s \widehat{\mu}_s$; a uniform numerical objective discrepancy is charged twice and absorbed into ϵ_{num}.

Choose $\lambda_{\pi,n} \asymp \eta_N$, $d_u^{-2} = O(\eta_N)$ and $\tau_{\text{solver}} + \epsilon_{\text{num}} = O(\eta_N)$. The conforming product-spline resolution $d_u \gtrsim N^{1/9}$ suffices for this calibration with the logarithmic majorant in (67). Writing $x = \sqrt{R(\widehat{a}_s)}$ and dividing (69) by $\lambda_{\pi,n}$ gives $x^2 \le C_1 x + C_2$ with fixed constants. Hence all returned actions have one deterministic $\mathcal{G}$-norm bound, uniformly over labels and iterations. Taking $B_{\mathcal{H}}$ large enough for the two-state product norm and for the global optimum places both in the same fixed $\Pi_{\mathcal{H}}(B_{\mathcal{H}})$, so $a_{\mathcal{H}} = 0$.

For policy improvement, every action occurring in the pointwise envelope of $\Pi_{\mathcal{H}}(B_{\mathcal{H}})$ has squared $\mathcal{G}$ norm at most $B_{\mathcal{H}}^2$. Regularized near-optimality, conforming spline approximation and the bounded fitted-critic Lipschitz constant therefore give directly

$$\sup_{\pi \in \Pi_{\mathcal{H}}(B_{\mathcal{H}})} \widehat{Q}_c\{s, \pi(s)\} - \widehat{Q}_c(s, \widehat{a}_s) \le C\{d_u^{-2} + \lambda_{\pi,n} B_{\mathcal{H}}^2 + \tau_{\text{solver}} + \epsilon_{\text{num}}\} = O(\eta_N). \tag{70}$$

This is a pointwise bound, so it supplies $\varepsilon_{\pi,n} = O(\eta_N)$ without the generic square-root conversion in Lemma D.1. The initialization $Q = 0$ obeys the same bounds by comparison with the zero action.

Equations (67), (69) and (70), together with the verified state constants and $M \ge c \log N$, establish the independent-subject rate $\widetilde{O}_p(N^{-2/9})$. The following subsection sharpens this to the rate stated in the continuation of Example 1. Thus the full linear critic in Example 1 supplies the coverage, localization and value calculations.

M.6 Temporal refinement for Example 1

The independent action draws and reward–transition randomization verify the dependence condition used for the stronger main-text rate. Given complete records through time t, S_{t+1} is known, but averaging the independent action draw and transition at $t+1$ gives $\mathcal{L}(S_{t+2} \mid Z_1, \ldots, Z_t) = (1/2, 1/2)$. All future draws depend on the past only through this reset state. Consequently the complete-record absolute regularity coefficient is zero at lags $k \geq 2$. The comparison in Lemma L.1 uses the coefficient at lag $b+1$, so block length $b = 1$ gives effective sample size $n = NT$ and zero comparison failure.

Apply Proposition L.2 with label resolution n^{-1}, confidence level n^{-2}, ridge coefficient ϱ_n of order $\sqrt{\log(en)/n}$, and KRR regularized empirical squared-loss objective gap $O(\varrho_n)$. The fitted norm stays bounded and $\delta_{K,n} = \widetilde{O}(n^{-1/4})$. Replacing the scale N by n in (67), write

$$\eta_n = C\delta_{K,n}^{8/9} = \widetilde{O}(n^{-2/9}).$$

Use $\lambda_{\pi,n} \asymp \eta_n$, $d_u^{-2} = O(\eta_n)$, and statewise solver and numerical objective errors of order $O(\eta_n)$. The full-sample objective bound in Proposition L.1, applied to state indicators, also places $\min_s \widehat{\mu}_s \geq 1/4$ on an event with probability tending to one at this scale. Intersecting it with the full-label KRR event allows (69) and (70) to give the same fixed-radius policy localization and

$$\varepsilon_{Q,n} + \varepsilon_{\pi,n} = \widetilde{O}(n^{-2/9}).$$

The intersected statistical events establish $\mathcal{R}_n$ and the required uniform bounds, giving $\Pr(\mathcal{E}_n) \to 1$. Theorem 1, with $a_{\mathcal{H}} = 0$ and $M \geq c \log n$ for sufficiently large fixed c, therefore gives

$$J^\star - J(\widehat{\pi}_M) = \widetilde{O}_p((NT)^{-2/9}),$$

as stated in the continuation of Example 1. The preceding independent-subject calculation remains available without using the temporal refinement.

M.7 Numerical observations without imposing smooth observed paths

Finite point evaluations do not define a bounded observation map on arbitrary L^2 equivalence classes. Accordingly, the exact-score result places no smoothness requirement on observed actions, whereas an implemented numerical score retains its discrepancy from observing and integrating the action function. A vanishing uniform point-quadrature error over the whole L^2 ball is not available.

Here is a useful sufficient condition that does not require a smooth observed action. Suppose the supplied observation record provides cell averages $\bar{A}_k = |I_k|^{-1} \int_{I_k} A(u)du$ and the implemented score is $\sum_k |I_k| \bar{A}_k \beta(t_k)$, $t_k \in I_k$. If the learned basis function is $\mathrm{Lip}_{\beta,n}$-Lipschitz and the largest cell width is h, then

$$\left| \sum_k |I_k| \bar{A}_k \beta(t_k) - \langle A, \beta \rangle \right| \leq \|A\|_{L^2} \|\beta_h - \beta\|_{L^2} \leq B_A \mathrm{Lip}_{\beta,n} h.$$

Here $\beta_h(u) = \beta(t_k)$ on I_k. The cell-average definition and linearity of integration give

$$\sum_k |I_k| \bar{A}_k \beta(t_k) - \langle A, \beta \rangle = \sum_k \int_{I_k} A(u)\{\beta(t_k) - \beta(u)\}du = \langle A, \beta_h - \beta \rangle.$$

Cauchy–Schwarz bounds its absolute value by $\|A\|_{L^2} \|\beta_h - \beta\|_{L^2}$. The Lipschitz property gives

$$\|\beta_h - \beta\|_{L^2}^2 \leq \mathrm{Lip}_{\beta,n}^2 \sum_k \int_{I_k} |u - t_k|^2 du \leq \mathrm{Lip}_{\beta,n}^2 h^2 \sum_k |I_k| = \mathrm{Lip}_{\beta,n}^2 h^2.$$

Taking square roots proves the bound without any derivative of A. This calculation applies only to records supplying cell averages or a justified equivalent reconstruction; raw point samples are not converted into averages. Other observation schemes retain their separately verified numerical discrepancy.

N FURTHER RELATED WORK

Classical FQI alternates regression and greedy improvement (Ernst et al., 2005), with finite-sample value analysis based on Bellman error propagation (Munos & Szepesvári, 2008; Farahmand et al., 2010). Lin et al. (2026, Sections 3.2–3.3) use KRR and average B-spline policy optimization for functional-action learning. Their physical-activity application treats a distribution of daily steps as the action and establishes optimal-policy existence alongside empirical results. We analyze the finite-sample value of the learned policy. Interactive functional control provides another setting: Pan et al. (2018) use descriptor regularity to reduce policy covering numbers while collecting new transitions during training, and Pirmorad et al. (2021) control spatial forcing functions. Our setting uses fixed offline data.

In the conference version of Antos et al. (2007), Assumptions A1–A2 require a compact Euclidean action set and a positive behavior-density lower bound. Their equation (3) fits labels with inverse-density-weighted squared loss. Assumption A6 lower-bounds the fraction of every ℓ^1 pyramid around an action that remains inside the action set, while Assumptions A7–A8 impose action-Lipschitz transition, reward and critic functions. Combined with density positivity, this geometry gives polynomial-order behavior probability to small ℓ^1 neighborhoods and lets nearby logged actions control a selected action. We instead regress under the observed law and use directional information plus RKHS smooth evaluation, so no logged action need be metrically close to a policy action. Their Theorem 3.2 has a sample term proportional, up to logarithms and other constants, to $N^{-1/\{4(d_A+1)\}}$. Substituting $d_A = d_{u,N} \sim c \log N$ gives $N^{-1/\{4(d_{u,N}+1)\}} \to e^{-1/(4c)}$. This only shows that the stated fixed-dimensional bound does not directly give functional consistency; it does not rule out growing-dimensional FQI. Their final value theorem uses an exactly greedy output policy, whereas our output is the learned average-objective policy. Functional neighborhood probabilities also differ from finite-dimensional volume calculations (Delaigle & Hall, 2010; Mas, 2012), although density ratios relative to specified probability laws can exist.

Pessimistic nonlinear fitted-Q analysis uses structural conditions beyond differentiability (Yin et al., 2023; Di et al., 2024). Simultaneous control of critics and data-dependent Bellman targets is central to deep nonparametric offline-RL bounds (Nguyen-Tang et al., 2022); here the labels also range over the policy class and the sampling units are subjects or separated blocks. Intrinsic-dimensional fixed-policy evaluation (Ji et al., 2023) provides related function-class-relative mismatch ideas, but does not include policy approximation and optimization. AdaFNN learns the basis functions forming its functional scores (Yao et al., 2021), complementing classical functional representations (Ramsay & Silverman, 2005; Hsing & Eubank, 2015). We impose an explicit label-uniform L^2 factorization for approximation and separately prove the learned-basis secant identity; neither step alone supplies the information condition.

The model-based coefficient of Uehara & Sun (2022, Definition 1) is the worst, over candidate transition models, ratio between squared total-variation transition error under a comparator-policy occupancy and under the offline law. Definition 1 of Xie et al. (2021) is the model-free analogue formed from squared Bellman-residual norms under a target law and the data law. Both coefficients can be smaller than a state–action density-ratio bound, and our construction follows the same function-class-relative principle. The difference is the transferred error family and the target functional: we use actual critic secants, take a policy supremum inside a next-state integral over one fixed RKHS policy class, and allow the information constant to depend on the prescribed evaluation accuracy. Consequently, a uniform nonlinear bound can yield vanishing evaluation error even when the corresponding uniform linear error ratio is infinite. The low-rank specialization of Uehara & Sun (2022) also includes an action-probability ratio. The importance of dynamics beyond static feature coverage appears in linear off-policy evaluation: Wang et al. (2021) show exponential horizon dependence despite realizability and feature coverage, Perdomo et al. (2023) characterize distinct conditions for FQI and LSTD to succeed, and Amortila et al. (2026) develop feature-dynamics coverage for LSTDQ. These results motivate distinguishing target approximation, directional information, and state propagation in our functional-action analysis.

O EXPERIMENTAL DESIGN AND IMPLEMENTATION

Cohorts and estimands. The experiment uses 20 independent datasets and their associated training runs, with random seeds 0 through 19 for reproducibility, at each of five sample sizes, $n = NT \in \{2000, 4000, 8000, 16000, 32000\}$ with $T = 20$ and $N \in \{100, 200, 400, 800, 1600\}$. FQI with functional actions uses either AdaFNN or Nyström KRR. For each approximator, sample size, and seed, five candidate policies use $\lambda_{\Omega,n}/s_\Omega \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ at $M = 20$ and $\gamma = .95$, where the deterministic conversion factor s_Ω is defined in Appendix O.1. The sample-size panels use one selected candidate per approximator, sample size, and seed. The AdaFNN state-dependent constant-torque comparator is a separate, full $M = 20$ FQI fit for every seed and sample size. It uses the corresponding offline dataset and common random numbers for evaluation.

Environment and behavior. Our Pendulum extension has $S = (\cos\theta, \sin\theta, \omega)$ and bounded torque-function action $a : [0, 1] \to [-2, 2]$, executed over 0.5 physical time units. The entire function is chosen at the start of each decision interval; the constant-torque comparator restricts it to $a(u) \equiv c(S)$. In normalized within-action time u,

$$\frac{d\theta}{du} = 0.5\omega, \qquad \frac{d\omega}{du} = 0.5\left\{\frac{3g}{2}\sin\theta + 3a(u)\right\}, \qquad g = 10,$$

where the pendulum mass and length are both one. Velocity is restricted to $[-8, 8]$. Fourth-order Runge–Kutta uses 128 substeps. Stochastic transitions add wrapped angle noise with standard deviation 0.03 and truncated angular-velocity noise with standard deviation 0.05. Reward is one minus the integrated angle, velocity, and torque cost,

$$\frac{\int_0^1 \{\mathrm{wrap}(\theta(u))^2 + 0.1\omega(u)^2 + 0.001a(u)^2\}\, du}{\pi^2 + 0.1(8)^2 + 0.001(2)^2},$$

clipped to $[0, 1]$ for numerical tolerance. The behavior action combines a baseline controller, one persistent subject-level Matérn-$5/2$ Gaussian-process realization, and an independent step-level Matérn-$5/2$ realization. The step amplitude is 0.35. Subjects are independent while decisions within a subject share its persistent component. Sample-size variants are nested subject prefixes of the same seed-specific data pool.

Critic and policy optimization. AdaFNN uses four learned global bases, 128-point action quadrature, and a three-hidden-layer width-256 ReLU outer network. Its critic is trained with early stopping. The functional policy is a bounded, state-dependent cubic B-spline with $p_u = 12$ coefficients; full-covariance CMA-ES (Hansen & Ostermeier, 2001) optimizes the policy objective for AdaFNN. The state-dependent constant-torque comparator uses the common critic family and offline dataset but restricts each decision to $a_s(u) \equiv c(s)$. Its within-action curvature vanishes, so the curvature penalty does not affect its policy objective. KRR uses the Nyström approximation (Williams & Seeger, 2000), fixed Gaussian state lengthscales $(1, 1, 2)$, an action lengthscale equal to the training-action median quadrature-weighted L^2 distance, and landmark ranks $512, 1024, 1536, 2304, 3584$ across the five sample sizes. The KRR critic ridge coefficient is selected within each FQI iteration by fivefold subject-grouped cross-validation of the fixed regression target, without policy-value information. For KRR policy optimization, we use Adam (Kingma & Ba, 2015) following the implementation of Lin et al. (2026), with bounded outputs enforced by our policy parameterization. KRR and AdaFNN use a common policy-coefficient grid and environment; their critic approximators remain distinct.

O.1 DIMENSIONLESS EMPIRICAL TOTAL-CURVATURE POLICY OBJECTIVE

For the cubic B-spline action $\pi_C(s)(u) = B_{p_u}(u)^\top c_C(s)$, write

$$W_{p_u} = \int_0^1 B''_{p_u}(u) B''_{p_u}(u)^\top\, du, \qquad \widehat{\Omega}_{\mathcal{D}}(\pi_C) = \frac{1}{n}\sum_{j=1}^{n} c_C(S_j^+)^\top W_{p_u} c_C(S_j^+).$$

This is the uncentered integral of squared second derivatives, averaged over all observed next states; it is not a centered across-state functional variance. Let $\mathcal{I}_\Phi$ index a fixed subset of the training next states and define the fitted value term at FQI iteration m by

$$\widehat{\Phi}_m(\pi_C) := \frac{1}{|\mathcal{I}_\Phi|}\sum_{j\in\mathcal{I}_\Phi} \widehat{Q}_m\{S_j^+, \pi_C(S_j^+)\}.$$

The implemented dimensionless objective is

$$\mathcal{J}_{\lambda_{\Omega,n},m}(C) = \frac{\widehat{\Phi}_m(\pi_C)}{Q_{\text{scale}}} - \frac{\lambda_{\Omega,n}}{s_\Omega}\frac{\widehat{\Omega}_{\mathcal{D}}(\pi_C)}{\Omega_{\text{scale}}(p_u)},$$
$$Q_{\text{scale}} = (1-\gamma)^{-1} = 20, \qquad \Omega_{\text{scale}}(p_u) = 4p_u\lambda_{\max}(W_{p_u}), \qquad s_\Omega = \frac{Q_{\text{scale}}}{\Omega_{\text{scale}}(p_u)}.$$

The value term averages the fitted critic over this fixed subset, whereas empirical total curvature averages over all n training next states. The scales are deterministic for fixed p_u and γ and do not depend on realized curvature or a particular seed. Because the policy coefficients lie in $[-2,2]^{p_u}$, $\widehat{\Omega}_{\mathcal{D}}(\pi_C) \le \Omega_{\text{scale}}(p_u)$. Multiplying the objective by Q_{scale} gives $\widehat{\Phi}_m(\pi_C) - \lambda_{\Omega,n}\widehat{\Omega}_{\mathcal{D}}(\pi_C)$. The experiment therefore replaces direct computation of the complete policy-RKHS norm in Algorithm 1 by empirical total curvature as a surrogate for its within-action regularity; it does not add a second penalty. The tuning grid is expressed in the dimensionless ratio $\lambda_{\Omega,n}/s_\Omega$. Appendix J proves that the resulting spline class satisfies the same spectral smoothness envelope used in coverage transfer and gives conditions for sharper penalty-localized envelopes.

At $n = 8000$, increasing the curvature-penalty coefficient from the smallest to the largest candidate value reduces the mean AdaFNN policy curvature from about 13,625 to numerical zero, while the number of selected policies outside the constant-within-interval subclass falls from 7 to 0 across the 20 datasets.

Tuning, independent evaluation, and uncertainty. For each candidate fit, 1,000 length-100 Monte Carlo trajectories estimate unpenalized normalized policy value for validation. The largest validation mean selects one positive coefficient, with exact ties assigned to the smaller coefficient. Because the experiment concerns policy improvement, simulator rollouts provide these value estimates without fitting a separate policy evaluator such as FQE. The selected policy is evaluated on a disjoint set of 1,000 trajectories. Validation and evaluation use independent random numbers; within each set, candidate coefficients and approximators share common random numbers for a given data/training seed. The state-dependent constant-torque comparator is evaluated using the same random numbers as the selected functional policy. All reported values estimate $(1-\gamma)\mathbb{E}\sum_{t=0}^{99}\gamma^t R_t$ without subtracting the policy penalty. Monte Carlo sampling error remains, and at $\gamma = .95$ the normalized reward tail beyond 100 decisions is at most $\gamma^{100} \simeq 0.00592$. Pointwise 95% percentile intervals use 10,000 resamples of the 20 independent datasets and their associated training runs; trajectories and held-out diagnostic pairs are not resampled as independent training datasets. The figures show means for value and medians for critic-relative diagnostics.

P Additional AdaFNN and KRR results

Selected-policy sample-size curves. The mean AdaFNN values at $n = 2000, 4000, 8000, 16000, 32000$ are $0.587, 0.593, 0.605, 0.657, 0.702$, compared with $0.531, 0.532, 0.535, 0.559, 0.560$ for the state-dependent constant-torque comparator. Seed-level mean differences are descriptive rather than a claim about every dataset. For KRR, the selected-policy means on the common grid are $0.710, 0.695, 0.682, 0.658, 0.645$.

At the smallest sample size, KRR's mean normalized policy value is higher than AdaFNN's (0.710 versus 0.587). As sample size grows, their value trends differ: AdaFNN improves while KRR declines. The KRR diagnostics in Appendix P.1 show increasing relative feature leverage and an overall increase in critic-energy ratios over the same sample sizes. The latter trend is not monotone, with a small dip from $n = 4000$ to $n = 8000$ in Figure S4(c). This coexistence motivates examining critic representation and its relevant difference directions when interpreting evaluation stability. The curves describe performance under the respective tuning procedures, with the two critic families reported separately.

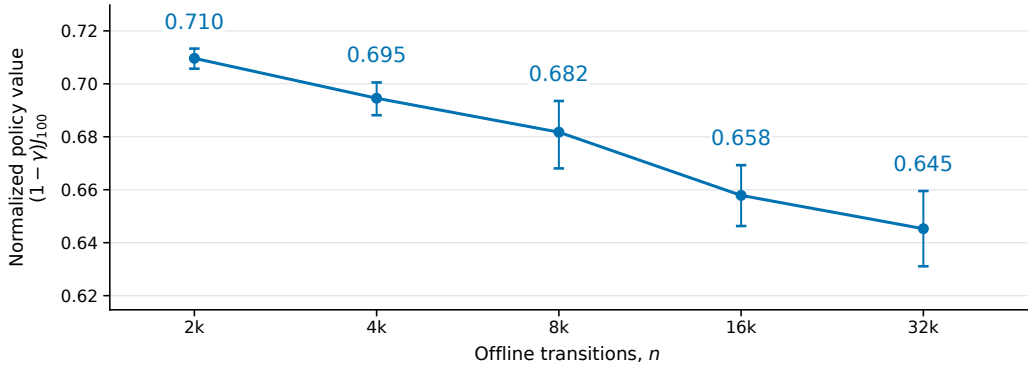

Figure S1: Nyström KRR normalized policy value across the sample sizes in Figure 1(a), with independent evaluation after Monte Carlo validation of the curvature coefficient. Points are means across the 20 datasets with pointwise 95% percentile-bootstrap intervals. Landmark rank also increases with n as specified in Appendix O; the curve describes the implemented learning procedure, not an isolated effect of sample size.

Paired policy-value differences. Figure S2 compares the selected functional policy and the separately trained constant-torque policy within each data/training seed. Mean differences increase from 0.056 at $n = 2000$ to 0.141 at $n = 32000$. Their pointwise intervals lie above zero at all five sample sizes. These paired comparisons quantify the policy-value gain of the functional-policy learning procedure under the common offline-data and evaluation design.

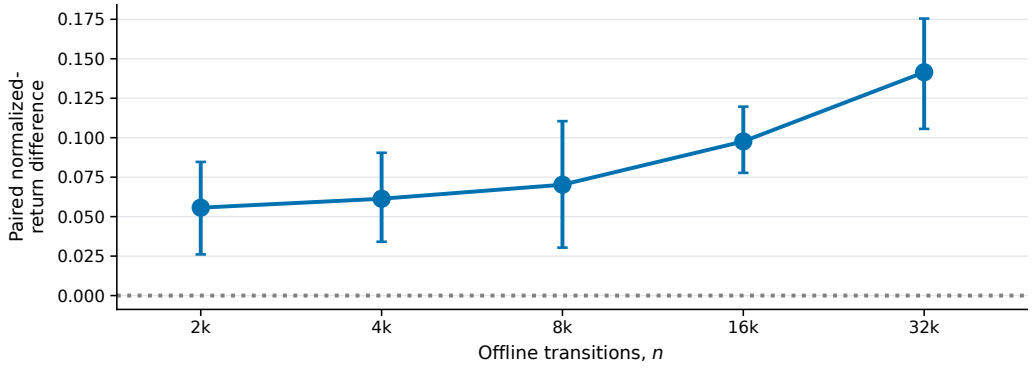

Figure S2: AdaFNN functional-minus-constant difference in normalized policy value, paired within seed using common random numbers for evaluation. Points are mean differences; bars are pointwise 95% percentile intervals from 10,000 whole-seed bootstrap resamples. All sample sizes from a seed are resampled together, but the displayed intervals are not a simultaneous confidence band.

P.1 Empirical locality and critic differences

Independent diagnostic sample and measurement. The diagnostics evaluate the fitted critics $\widehat{Q}_{19}$ and $\widehat{Q}_{20}$ on an independent evaluation cohort generated for each fit from the same environment and behavior policy as the offline data. At sample size $n = NT$, the cohort contains N trajectories with $T = 20$ decisions, yielding $q = NT = n$ evaluation states: $q \in \{2000, 4000, 8000, 16000, 32000\}$. Both the learned action and the observed behavior action are evaluated at each of these q states. The training and evaluation trajectories are independent. Write $\mathcal{N}_k(s)$ for the indices of the k nearest training states to s, ordered by

$$d_S(s, s')^2 = (s_1 - s'_1)^2 + (s_2 - s'_2)^2 + (s_3 - s'_3)^2/4,$$

with ties broken by training-record order. For a query action a at state s, the nearest-action L^2 distance is

$$d_k(s,a) = \min_{(i,t)\in\mathcal{N}_k(s)} \left[\int_0^1 \{a(u) - A_{i,t}(u)\}^2 \, du\right]^{1/2},$$

computed by quadrature on the common 128-point action grid. At each evaluation state, the learned action $\widehat{\pi}_{20}(s)$ and the observed behavior action are compared with the candidate set $\mathcal{N}_k(s)$. Each fit contributes the median distance for each action class; Figure 1(b) reports the median across the 20 datasets, with $k = 32$. The behavior-action baseline is not the constant-torque comparator. This finite-sample distance describes closeness to recorded actions at nearby states, not probability mass or a population support gap.

Sensitivity to neighborhood size. Figure S3 reports the comparison for $k \in \{16, 32, 64, 128\}$, keeping the fitted policies, selected penalty coefficients, and evaluation queries fixed. Since $\mathcal{N}_k(s) \subseteq \mathcal{N}_{k+1}(s)$, we necessarily have $d_{k+1}(s,a) \le d_k(s,a)$. A decrease with k is therefore a consequence of minimizing over more candidates, not evidence of improved coverage. Larger neighborhoods also weaken state locality: at $k = n$, the distance is to the entire action sample regardless of state, and need not be zero. The relevant sensitivity contrast is between learned and behavior actions at a fixed k. The median learned-action distance remains larger at every sample size and every displayed k. This supports the qualitative contrast over the tested neighborhood sizes, without asserting that it persists for arbitrary k or establishes a population support gap.

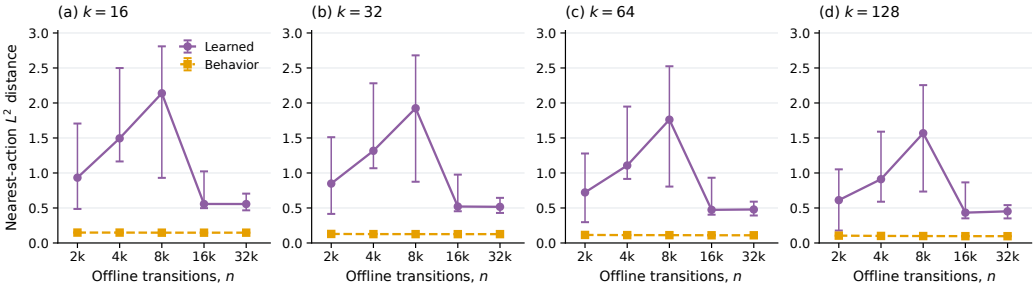

Figure S3: AdaFNN nearest-action L^2 distances across sample sizes for (a) $k = 16$, (b) $k = 32$, (c) $k = 64$, and (d) $k = 128$ nearest training states. Each fit contributes the median nearest-action L^2 distance over its independent $q = n$ evaluation states; points are medians across 20 independent datasets, with pointwise 95% percentile intervals from 10,000 whole-seed bootstrap resamples. Panel (b) reproduces Figure 1(b). All panels use the same vertical scale.

Adjacent critic differences. For each fit, the adjacent critic difference is $h = \widehat{Q}_{20} - \widehat{Q}_{19}$. The critic-energy ratio ER_h defined in Section 5 can be written as

$$\mathsf{ER}_h = \frac{\mathrm{mean}_{(s,a)\in\mathcal{X}_{\mathrm{pol}}} h(s,a)^2}{\mathrm{mean}_{(s,a)\in\mathcal{D}} h(s,a)^2},$$

where $\mathcal{X}_{\mathrm{pol}}$ denotes the held-out learned-policy graph and $\mathcal{D}$ the logged design, with every logged record counted. The displayed center is the median of the ratios from the 20 datasets, not a ratio of aggregated energies. All selected fits have positive design-energy denominators; no denominator regularization is applied. Values near one mean that this realized adjacent critic difference has comparable empirical squared energy on the two finite designs. This is a diagnostic for one adjacent difference per fit, rather than uniform control of the fitted-versus-target family in Section 3.

KRR representation diagnostics. Figure S4 gives the KRR sample-size counterparts of the locality and adjacent-critic summaries, together with ridge leverage in the fitted Nyström representation. Learned-policy leverage is normalized by the median behavior-action leverage from the corresponding fit and independent evaluation cohort; each fit contributes its median normalized learned-policy leverage. For each fitted KRR critic, the deterministic Nyström feature map and the coefficients of $\widehat{Q}_{19}$ and $\widehat{Q}_{20}$ determine the adjacent-difference and leverage geometry.

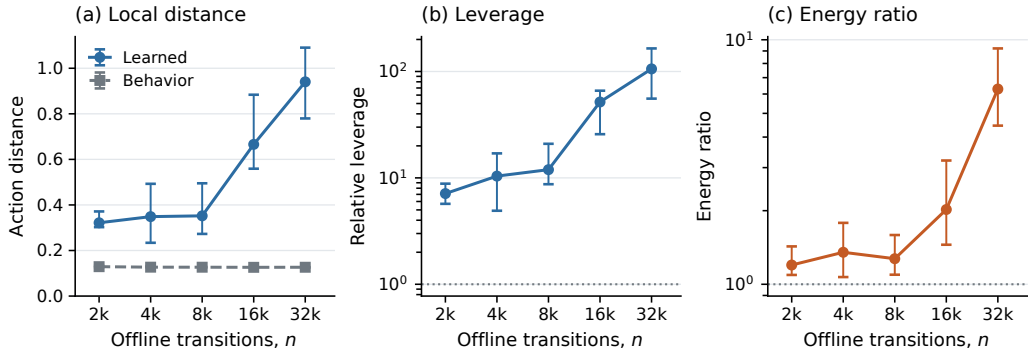

Figure S4: Nyström KRR policy-action diagnostics across the five sample sizes: (a) nearest-action L^2 distance, (b) relative feature leverage, and (c) critic-energy ratio. Every fit uses an independent cohort of $q = n$ evaluation states. Points are medians over the 20 datasets with pointwise 95% percentile-bootstrap intervals across whole seeds. The offline state–action design and critic iterations 19 and 20 are common to AdaFNN and KRR; the critic representation and ridge geometry are approximator-specific. Panels (b)–(c) use logarithmic vertical scales.

Absolute adjacent-critic energies. The ratio alone does not specify the scale of the critic update. Figure S5 therefore displays its numerator and denominator separately for the selected fits. Each quantity is a mean squared difference for $\widehat{Q}_{20} - \widehat{Q}_{19}$, not a squared error relative to a true or Bellman-target critic.

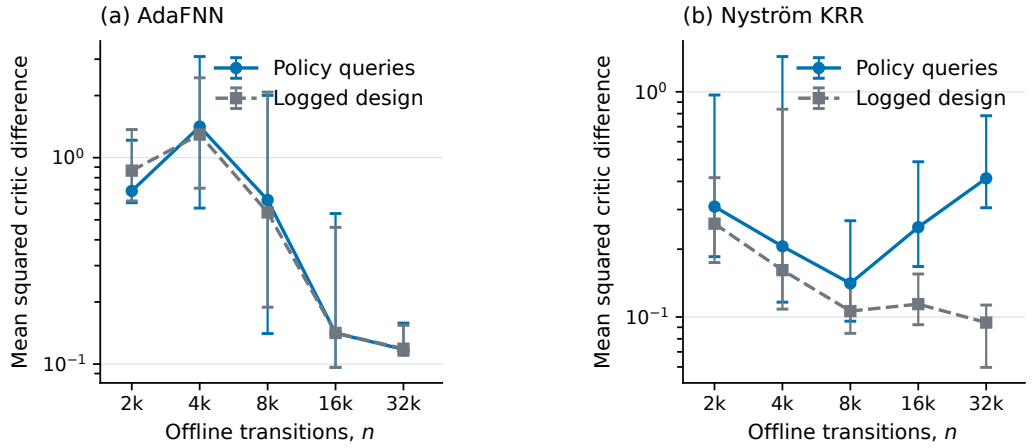

Figure S5: Absolute adjacent-critic energies on independent held-out learned-policy state–action pairs and the logged design. The held-out mean evaluates the learned action at $q = n$ states for every fit. Points are medians of the energies from the 20 datasets with pointwise 95% percentile-bootstrap intervals across whole seeds. Each panel uses a logarithmic vertical scale. The median of the per-fit ratios in Figure 1(c) is generally not the ratio of these plotted medians.