BELLMAN CALIBRATION FOR V-LEARNING IN OFFLINE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Calibration is a standard tool for assessing and correcting the reliability of predictions in supervised learning, but is less developed for long-horizon value prediction in reinforcement learning, where the prediction target is not directly observed. We extend calibration to the Bellman fixed-point setting by introducing *Bellman calibration*, which requires the Bellman equation to hold conditionally on the predicted value. Equivalently, among states assigned the same predicted value, the average one-step Bellman target should equal that prediction. Because this target itself depends on the value predictor, calibration becomes a fixed-point problem. We define a calibration error that quantifies systematic Bellman residuals conditional on the predicted value and develop doubly robust estimators from off-policy data. We establish a calibration–refinement bound that relates calibration error to value-prediction error, separating the component that post-hoc calibration can correct from information lost by the original predictor. We then develop *Fitted Bellman Calibration*, a model-agnostic post-hoc procedure that iteratively recalibrates a learned value predictor using one-dimensional histogram or isotonic regression, without refitting the underlying high-dimensional value function. We establish finite-sample guarantees showing one-dimensional nonparametric calibration rates for histogram and isotonic calibration, together with value-error guarantees for fixed-partition calibration, without imposing realizability or Bellman-completeness assumptions on the calibrated predictor class.

1 INTRODUCTION

Many sequential decision problems require predicting the long-term return of a target policy from different states using data generated under another policy. In a Markov decision process (MDP), this corresponds to estimating the state-dependent value function of a target policy π from off-policy data. Such predictions arise across applications: clinicians evaluate proposed treatment rules (Van Calster et al., 2019); online platforms estimate customer lifetime value and retention under alternative recommendation strategies (Theocharous et al., 2015; Maystre & Russo, 2022; Xue et al., 2025); and economists assess the downstream effects of counterfactual programs (Rust, 1987; Cowgill & Tucker, 2020). Systematic miscalibration can distort these long-term predictions and lead to poor decisions.

Flexible predictors such as neural networks and gradient-boosted trees can be systematically miscalibrated under model misspecification, limited data, or distribution shift (Zadrozny & Elkan, 2001; Niculescu-Mizil & Caruana, 2005; Bella et al., 2010; Guo et al., 2017; Błasiok et al., 2023). In long-horizon value prediction, bootstrapping compounds these difficulties: each regression target depends on earlier value estimates, so finite-iteration, approximation, and off-policy errors can propagate into the final predictions (Gordon, 1995; Tsitsiklis & Van Roy, 1996; Munos & Szepesvári, 2008; Farahmand et al., 2010; Agarwal et al., 2021). These errors are difficult to assess because fitted value methods lack a direct supervised validation loss for hyperparameter tuning, model selection, or early stopping. Existing analyses instead control error propagation under assumptions such as Bellman completeness, realizability, linear approximation, or finite state spaces (Chen & Jiang, 2019; Xie et al., 2022; Meyn, 2024; Xie & Jiang, 2021). Minimax methods provide an alternative but require auxiliary function classes and saddle-point optimization, together with appropriate realizability, richness, and coverage conditions (Dai et al., 2018; Uehara et al., 2020; Jin et al., 2021; Xie & Jiang, 2021). This motivates a weaker reliability criterion that targets systematic error in the prediction scale and can be corrected post hoc without refitting the high-dimensional value function.

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.

Calibration provides such a reliability criterion in supervised learning. A predictor is calibrated if, among observations receiving the same prediction, the average outcome equals that prediction; in population terms, a calibrated predictor satisfies $f(X) = \mathbb{E}[Y \mid f(X)]$. For binary classification, this becomes $f(X) = \mathbb{P}\{Y = 1 \mid f(X)\}$, so predicted probabilities agree with event frequencies among observations receiving the same prediction. A central appeal is that systematic miscalibration can often be corrected by a low-dimensional transformation of the model output, without refitting the underlying predictor, using methods such as Platt scaling, histogram binning, or isotonic calibration (Lichtenstein et al., 1977; Platt, 2000; Zadrozny & Elkan, 2001; Vovk et al., 2005; Gupta et al., 2020; Gupta & Ramdas, 2021; Whitehouse et al., 2024). Minimizing a proper prediction loss over a restricted model class does not in general ensure calibration, so post-hoc recalibration can remain useful even after the original model has been trained to minimize its fitting loss (Błasiok et al., 2023).

In this work, we extend calibration to value prediction. We say a value predictor v is *Bellman calibrated* if the Bellman equation holds conditionally on the predicted value:

$$v(S) = \mathbb{E}_\pi[R + \gamma v(S') \,|\, v(S)]\,.$$

That is, the prediction is conditionally unbiased for the one-step Bellman target $R + \gamma v(S')$ induced by the predictor itself. This property is weaker than full Bellman consistency, $v(S) = \mathbb{E}_\pi[R + \gamma v(S') \,|\, S]$, because it constrains only the scalar prediction rather than the entire state-to-value mapping, while still ruling out systematic under- or overestimation within level sets of the predictor.

We first relate Bellman calibration to value-prediction error and show how calibration error can be estimated from off-policy data. We show that Bellman miscalibration isolates a component of value-prediction error that can be corrected post hoc. We then establish a calibration–refinement bound that separates this correctable component from refinement error, which reflects information about the state not retained by the original prediction. When calibration error is zero, the remaining value error is controlled, up to constants, by the error of the best transformation of the original predictor. We also develop doubly robust estimators of calibration error from off-policy data.

Finally, we develop post-hoc correction procedures with finite-sample calibration guarantees. Unlike in supervised calibration, the Bellman target depends on the recalibrated predictor itself, making calibration a fixed-point problem. We develop *Fitted Bellman Calibration*, which iteratively recalibrates one-step Bellman targets using a one-dimensional transformation of the original value predictor. Finite-sample analysis of this fitted iteration would generally require realizability or Bellman completeness of the calibration class. To avoid imposing such assumptions, we use fixed-partition histogram calibration, for which the partition induces a finite aggregated MDP and the corresponding piecewise-constant function class is Bellman complete. The same argument does not apply directly to isotonic calibration because the class of monotone functions is not Bellman complete. We therefore propose a hybrid isotonic–histogram procedure that uses an initial isotonic fit to determine a data-adaptive partition and then applies histogram calibration while holding this partition fixed in subsequent iterations. This retains the finite-sample guarantees of the fixed-partition analysis while avoiding separate selection or tuning of the number of bins.

Contributions. Our main contributions are as follows.

1. We formalize **Bellman calibration**, develop doubly robust estimators of its calibration error, and establish a calibration–refinement bound relating calibration to value-prediction error.
2. We develop **Fitted Bellman Calibration**, which iteratively recalibrates a value predictor using fixed-partition histogram or isotonic regression, together with a hybrid isotonic–histogram procedure for value-error control.
3. We establish **finite-sample guarantees** with one-dimensional nonparametric rates for histogram and isotonic calibration and value-error bounds for fixed-partition calibration, without imposing realizability or Bellman-completeness assumptions on the calibrated predictor class.

1.1 Related work

Post-hoc calibration. Our work builds on the classical literature on calibration and post-hoc recalibration in supervised learning (Zadrozny & Elkan, 2001; 2002; Niculescu-Mizil & Caruana, 2005; Guo et al., 2017; Gupta & Ramdas, 2021). These methods transform the output of a fitted predictor without refitting the underlying model. More recent work extends calibration to causal

quantities, where the target is not directly observed and must be estimated using nuisance functions (van der Laan et al., 2023; Whitehouse et al., 2024). Bellman calibration shares the unobserved-target feature but introduces a fixed-point dependence because the Bellman target changes with the recalibrated value predictor. Existing uses of calibration in reinforcement learning have instead focused on predictive uncertainty in learned dynamics models or conformal uncertainty sets (Malik et al., 2019); our focus is post-hoc reliability of long-horizon value predictions.

Value-function estimation under function approximation. Bellman calibration addresses a different problem from direct value-function estimation. Fitted value methods regress Bellman targets onto a function class, with finite-sample guarantees typically requiring coverage together with Bellman completeness or related approximation conditions (Munos & Szepesvári, 2008; Chen & Jiang, 2019; Chang et al., 2022). Minimax methods use Bellman moment conditions or saddle-point formulations and replace completeness with realizability or richness conditions on the value and critic classes (Uehara et al., 2020; 2023). These methods aim to estimate the value function within a potentially high-dimensional model class. We instead start from an already fitted predictor and modify only its one-dimensional output scale. Our guarantees therefore concern conditional Bellman calibration rather than recovery of the full value function and do not require imposing realizability or Bellman-completeness assumptions on the one-dimensional calibration class.

2 Calibration for Value Functions

2.1 Setup and Notation

We consider an MDP with state space $\mathcal{S}$, discrete action space $\mathcal{A}$, initial-state distribution ρ, transition kernel $P(s' \mid s, a)$, reward function $r_0(s, a)$, and discount factor $\gamma \in [0, 1)$. We observe an i.i.d. calibration sample $\mathcal{C}_n := \{(S_i, A_i, R_i, S_i')\}_{i=1}^n$ collected under a behavior policy b_0, where $S_i \sim \rho$, $A_i \sim b_0(\cdot \mid S_i)$, $S_i' \sim P(\cdot \mid S_i, A_i)$, and $R_i = r_0(S_i, A_i) + \varepsilon_i$, with $\mathbb{E}[\varepsilon_i \mid S_i, A_i] = 0$. For a target policy $\pi(a \mid s)$, let $v_0(s) := \mathbb{E}_\pi[\sum_{t=0}^\infty \gamma^t R_t \mid S_0 = s]$ denote its value function.

For a state–action function f, let $(\pi f)(s) := \sum_{a \in \mathcal{A}} \pi(a \mid s) f(s, a)$, and define $b_0 f$ analogously. For a state function v, let $(Pv)(s, a) := \mathbb{E}[v(S') \mid S = s, A = a]$. Write $w_\pi(a \mid s) := \pi(a \mid s)/b_0(a \mid s)$, $r_{0,\pi} := \pi r_0$, and $P_\pi v := \pi(Pv)$. The policy-evaluation Bellman operator is $\mathcal{T}_\pi v := r_{0,\pi} + \gamma P_\pi v$, and v_0 is its unique bounded fixed point (Bellman, 1952; 1966), $v_0 = \mathcal{T}_\pi v_0$.

For a measure μ on $\mathcal{S}$, write $\|v\|_{2,\mu} := \{\int v(s)^2 \, \mu(ds)\}^{1/2}$, and let $L^2(\rho)$ denote the corresponding space of square-integrable state functions. We use $\|f\|_{n,S,A}$ and $\|g\|_{n,S'}$ for empirical L^2 norms over $\{(S_i, A_i)\}_{i=1}^n$ and $\{S_i'\}_{i=1}^n$, respectively. Finally, $\lesssim$ denotes inequality up to an absolute constant, and $[B] := \{1, \ldots, B\}$.

Q-function calibration. The policy Q-function is obtained as a special case by treating $X = (S, A)$ as the state and $X' = (S', A')$, with $A' \sim \pi(\cdot \mid S')$, as the next state. Then $q_0(S, A) = \mathbb{E}_\pi[R + \gamma q_0(S', A') \mid S, A]$. Thus, all results apply after replacing S by $X = (S, A)$.

2.2 Definition of Bellman Calibration

Let $\hat{v}$ be a value predictor for the target policy π, trained independently of the calibration sample $\mathcal{C}_n$. We treat $\hat{v}$ as fixed and ask whether its predictions satisfy a weaker post-hoc reliability condition. We refer to the scalar $\hat{v}(S)$ as the predicted value, or prediction score.

A natural benchmark is *statewise Bellman consistency*,

$$\hat{v}(S) \approx \mathbb{E}_\pi[R + \gamma \hat{v}(S') \mid S] \,,$$

which requires the Bellman residual to have conditional mean zero given the state. In general, however, this condition is difficult both to assess and to guarantee from offline data. Squared temporal-difference error is not an unbiased estimate of squared Bellman error because it also contains reward and transition noise, while estimating the statewise Bellman target requires a potentially high-dimensional regression. Consequently, validating statewise Bellman consistency requires additional structure, such as double sampling, restrictions on the Bellman target, or Bellman-completeness-type conditions (Chen & Jiang, 2019; Duan et al., 2021; Zitovsky et al., 2023).

We therefore seek a weaker but more tractable reliability condition. Rather than conditioning on the full state, we require

$$\hat{v}(S) \approx \mathbb{E}_\pi[R + \gamma \hat{v}(S') \mid \hat{v}(S)] .$$

This requires predictions to agree, on average, with their one-step Bellman targets among states assigned the same predicted value. It parallels supervised calibration, which requires predictions to be correct on average conditional on their own numerical output (Zadrozny & Elkan, 2002; Niculescu-Mizil & Caruana, 2005; Guo et al., 2017). Conditioning on the scalar prediction rather than the full state reduces the high-dimensional conditional moment to a one-dimensional one, making assessment and post-hoc correction possible without estimating the full state-dependent Bellman target.

For any candidate value function v, define the Bellman-calibration map

$$\Gamma_0(v)(s) := \mathbb{E}_\pi \left[R + \gamma v(S') \mid v(S) = v(s)\right] . \tag{1}$$

We say that v is *Bellman calibrated* if $v(S) = \Gamma_0(v)(S)$ almost surely, or equivalently,

$$\mathbb{E}_\pi[R + \gamma v(S') - v(S) \mid v(S)] = 0 \qquad \text{almost surely.}$$

Bellman calibration also has a reward-model interpretation. Any value function v induces the reward function $r_{\pi,v}(s) := v(s) - \gamma P_\pi v(s)$, under which v satisfies the Bellman equation for policy π. Bellman calibration is equivalent to

$$\mathbb{E}_\pi[R - r_{\pi,v}(S) \mid v(S)] = 0.$$

Hence, within prediction level sets, the reward implied by v agrees on average with the target-policy reward. When $\gamma = 0$, Bellman calibration reduces to ordinary regression calibration (Roth, 2022).

2.3 A Calibration–Refinement Bound for Value Error

The true value function v_0 is Bellman calibrated. Indeed, the Bellman equation and iterated expectations give $v_0(S) = \mathbb{E}_\pi[R + \gamma v_0(S') \mid v_0(S)] = \Gamma_0(v_0)(S)$. Bellman calibration is therefore necessary for exact value prediction, but it is not sufficient. Because calibration conditions only on the scalar prediction $v(S)$, state-level errors can cancel among states assigned the same predicted value (Gupta et al., 2020; van der Laan et al., 2023). Calibration can correct systematic Bellman error within prediction level sets, but it cannot recover state information lost through the reduction $S \mapsto \hat{v}(S)$. The following bound separates these two sources of value error.

For a candidate value function v, define the ℓ^2 **Bellman calibration error**

$$\mathrm{Cal}^2_{\ell^2}(v) := \|v - \Gamma_0(v)\|^2_{2,\rho} = E_\rho\Big[\{E_\pi[R + \gamma v(S') - v(S) \mid v(S)]\}^2\Big] . \tag{2}$$

This criterion is zero exactly when v is Bellman calibrated and quantifies systematic Bellman residuals conditional on the predicted value.

Functions of the original prediction $\hat{v}(S)$ form the class

$$\mathcal{V}_{\hat{v}} := \{g \circ \hat{v} : g : \mathbb{R} \to \mathbb{R} \text{ measurable}\}, \qquad \Pi_{\hat{v}} := L^2(\rho)\text{-projection onto } \mathcal{V}_{\hat{v}}.$$

Define the projected transition operator $P_{\pi,\hat{v}} := \Pi_{\hat{v}} P_\pi$, so that $P_{\pi,\hat{v}} f(s) = E_\pi[f(S') \mid \hat{v}(S) = \hat{v}(s)]$. The corresponding discounted projected concentrability coefficient is

$$\mathcal{C}_{\hat{v},\gamma} := \left\{ (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \left\| \frac{d(\rho P^t_{\pi,\hat{v}})}{d\rho} \right\|^{1/2}_{\infty} \right\}^2 .$$

A1 Projected concentrability. $\mathcal{C}_{\hat{v},\gamma} < \infty$.

When ρ is stationary under P_π, it is also stationary under $P_{\pi,\hat{v}}$ by Lemma S3, and hence $\mathcal{C}_{\hat{v},\gamma} = 1$. More generally, $\mathcal{C}_{\hat{v},\gamma}$ quantifies the mismatch between ρ and the discounted occupancy induced by $P_{\pi,\hat{v}}$. It is small when this occupancy remains well covered by ρ and large when it places substantial mass on regions that are rare under ρ.

Theorem 1 (Calibration–refinement bound)**.** *Under A1,*

$$\|\hat{v} - v_0\|_{2,\rho} \le \frac{\sqrt{\mathcal{C}_{\hat{v},\gamma}}}{1-\gamma} \left\{ \mathrm{Cal}_{\ell^2}(\hat{v}) + \min_{\theta:\mathbb{R}\to\mathbb{R}} \|\theta \circ \hat{v} - v_0\|_{2,\rho} \right\} .$$

Theorem 1 separates two sources of value-prediction error. The first term measures systematic Bellman residuals within levels of $\hat{v}(S)$, while the second is the error of the best value predictor based only on $\hat{v}(S)$ and therefore quantifies information about the state not retained by the original prediction. Calibration can remove systematic Bellman error but cannot recover differences in $v_0(S)$ among states assigned the same value by $\hat{v}$. When calibration error is zero, the remaining value error is bounded by the refinement error up to the factor $\sqrt{\mathcal{C}_{\hat{v},\gamma}}/(1-\gamma)$. This parallels classical calibration–refinement decompositions in classification and regression (Murphy, 1973; DeGroot & Fienberg, 1983; van der Laan et al., 2023; Whitehouse et al., 2024), while the concentrability and discount dependence are analogous to those in classical FVI/FQE error bounds (Munos & Szepesvári, 2008).

2.4 Doubly Robust Estimation of Bellman Calibration Error

A practical advantage of Bellman calibration error is that, unlike value-prediction error and refinement error, it can be estimated from one-step transition data. Because calibration is defined under the target policy π while actions are observed under the behavior policy b_0, estimation requires off-policy adjustment. We develop a doubly robust estimator that extends static calibration methods (Xu & Yadlowsky, 2022; van der Laan et al., 2023) to the sequential setting.

Let $\widehat{w}_\pi$ estimate the action importance ratio $w_\pi := \pi/b_0$. For a fixed state-value function v, define $q_v(s,a) := \mathbb{E}[R + \gamma v(S') \mid S = s, A = a]$, and let $\hat{q}_v$ estimate q_v. For $O = (S, A, R, S')$, define

$$\widehat{Y}_v(O) = (\pi \hat{q}_v)(S) + \widehat{w}_\pi(A \mid S)\{R + \gamma v(S') - \hat{q}_v(S,A)\}, \qquad \widehat{\mathcal{T}}_\pi v(s) = \mathbb{E}\Big[\widehat{Y}_v(O) \mid S = s\Big].$$

Theorem 2 (Doubly robust Bellman target)**.** *For any fixed v,*

$$E\Big[\{\widehat{\mathcal{T}}_\pi v - \mathcal{T}_\pi v\}(S) \mid v(S) = t\Big] = E[b_0\{(w_\pi - \widehat{w}_\pi)(\hat{q}_v - q_v)\}(S) \mid v(S) = t].$$

Consequently, $E\{\widehat{Y}_v(O) \mid v(S)\} = E\{(\mathcal{T}_\pi v)(S) \mid v(S)\}$ if either $\widehat{w}_\pi = w_\pi$ or $\hat{q}_v = q_v$.

The construction includes importance weighting ($\hat{q}_v = 0$) and model-based Bellman-target evaluation ($\widehat{w}_\pi = 0$) as special cases. Thus, either nuisance estimator is sufficient for recovering the prediction-conditional Bellman target.

We now use this pseudo-outcome to estimate squared Bellman calibration error. Let $U_i := v(S_i)$ and $\widehat{Y}_i := \widehat{Y}_v(O_i)$. Following Xu & Yadlowsky (2022), define

$$\widehat{\mathrm{Cal}}^2_{\ell^2,\mathrm{db}}(v) := \frac{1}{n}\sum_{i=1}^{n}\{\widehat{Y}_i - U_i\}\{\widehat{m}_{-i}(U_i) - U_i\},$$

where $\widehat{m}_{-i}$ is fit without observation i to estimate $u \mapsto E\{\widehat{Y}_v(O) \mid v(S) = u\}$. The estimator uses the identity $E[\{\widehat{Y}_v - v(S)\}\{m_v(v(S)) - v(S)\}] = E[\{m_v(v(S)) - v(S)\}^2]$, where $m_v(u) := E\{\widehat{Y}_v(O) \mid v(S) = u\}$. By Theorem 2, this target equals $\mathrm{Cal}^2_{\ell^2}(v)$ when either nuisance estimator is correct, with the remaining bias otherwise governed by the product of the two nuisance errors.

3 Post-hoc Bellman Calibration

Given an initial value predictor $\hat{v}$, our goal is to construct a Bellman-calibrated predictor using only a one-dimensional transformation of $\hat{v}$. Unlike supervised calibration, a single recalibration step is not generally sufficient because changing the predictor also changes its Bellman target. This motivates an iterative post-hoc procedure.

3.1 Fitted Bellman Calibration

Algorithm 1 defines *Fitted Bellman Calibration*. Starting from $V_0 = \hat{v}$, iteration k constructs doubly robust Bellman targets $\widehat{Y}_{V_k}(O_i)$ and regresses them on the scalar prediction $\hat{v}(S_i)$ over a calibrator class $\mathcal{G}$, yielding $V_{k+1} = g_{k+1} \circ \hat{v}$. Every iterate therefore remains a one-dimensional transformation of the original predictor. Writing $\mathcal{V}_{\hat{v}}(\mathcal{G}) = \{g \circ \hat{v} : g \in \mathcal{G}\}$, the procedure is a fitted value iteration restricted to functions of $\hat{v}$.

Algorithm 1 Fitted Bellman Calibration

1: **Input:** $\hat{v}$, $\mathcal{C}_n = \{O_i = (S_i, A_i, R_i, S_i')\}_{i=1}^n$, pseudo-outcome $v \mapsto \widehat{Y}_v$, class $\mathcal{G}$, iterations K
2: Initialize $V_0 = \hat{v}$
3: **for** $k = 0, \ldots, K-1$ **do**
4: $\quad V_{k+1} = g_{k+1} \circ \hat{v}$, where $g_{k+1} \in \arg\min_{g \in \mathcal{G}} \sum_{i=1}^n \{\widehat{Y}_{V_k}(O_i) - g(\hat{v}(S_i))\}^2$.
5: **end for**
6: **Output:** V_K

At the population level, the unrestricted iteration targets the fixed point of $\mathcal{B}_{\hat{v}} f = \Pi_{\hat{v}} r_{0,\pi} + \gamma P_{\pi,\hat{v}} f$, which is Bellman calibrated. The population operator is contractive, so $O\{\log(1/\varepsilon)/(1-\gamma)\}$ iterations suffice to reduce the finite-iteration error below ε.

For a generic calibrator class $\mathcal{G}$, however, finite-sample analysis of the fitted iteration would generally require the population fixed point to belong to $\mathcal{V}_{\hat{v}}(\mathcal{G})$ and this class to be closed under the Bellman operator, corresponding to realizability and Bellman completeness (Munos & Szepesvári, 2008). These conditions are stronger than needed for calibration and need not hold for standard calibrator classes. We therefore analyze two constructions that avoid them: fixed-partition histogram calibration and a hybrid isotonic–histogram procedure.

3.2 Histogram Calibration

Histogram calibration instantiates Algorithm 1 with a piecewise-constant calibrator. Partition the initial predictions $\{\hat{v}(S_i)\}_{i=1}^n$ into ordered bins $I_1, \ldots, I_B$, using either empirical quantiles (*equal-mass*) or uniform discretization (*equal-width*), and let $\mathcal{G}_B := \mathrm{span}\{\mathbf{1}_{I_b} : b \in [B]\}$ and $\mathcal{H}_B(\hat{v}) := \{g \circ \hat{v} : g \in \mathcal{G}_B\}$. Thus, every $h \in \mathcal{H}_B(\hat{v})$ is constant within bins of the original prediction. The bins may be data-adaptive, while B is deterministic and may grow with n.

Each iteration averages the fitted Bellman targets within each bin,

$$g_{k+1}(t) = \frac{1}{n_b} \sum_{i:\hat{v}(S_i) \in I_b} \widehat{Y}_i^{(k)}, \qquad t \in I_b,$$

where $n_b := \sum_{i=1}^n \mathbf{1}\{\hat{v}(S_i) \in I_b\}$. Histogram calibration therefore reduces fitted Bellman iteration to a finite aggregation of the original prediction. After sufficiently many iterations, the final predictor approximately satisfies the empirical Bellman calibration condition within each bin,

$$V_K(s) \approx E_n\left[\widehat{Y}_{V_K}(O) \,\middle|\, \hat{v}(S) \in I_b\right], \qquad \hat{v}(s) \in I_b.$$

Bounding the calibration error then reduces to controlling empirical fluctuations in the within-bin calibration errors.

Calibration error. We next give finite-sample, finite-iteration bounds for the Bellman calibration error in (2). The analysis uses the following technical conditions.

C1 Bounded targets. There exists $M < \infty$ such that, almost surely, $\|\hat{v}\|_\infty \le M$ and $\|\widehat{Y}_f\|_\infty \vee \|\mathcal{T}_\pi f\|_\infty \le M$ for every $f \in \{\hat{v}\} \cup \mathcal{V}_{\hat{v}}(\mathcal{G})$ satisfying $\|f\|_\infty \le M$.

C2 Sample splitting. The auxiliary estimators $\widehat{w}_\pi$ and $v \mapsto \hat{q}_v$, together with the initial predictor $\hat{v}$, are trained on data independent of $\mathcal{C}_n$.

C3 Stability of the Bellman-target regression. For all $f, g \in \mathcal{H}_B(\hat{v})$ with $\|f\|_\infty, \|g\|_\infty \le M$, $\|\hat{q}_f - \hat{q}_g\|_{n,S,A} \le L \|f - g\|_{n,S'}$.

Boundedness can be enforced by clipping or truncation, and sample splitting can be replaced by cross-fitting, as in van der Laan et al. (2023). Condition C3 controls the additional complexity from the map $v \mapsto \hat{q}_v$. It holds trivially for the importance-weighted target when $\hat{q}_v = 0$, and with $L = 1$ for cellwise estimates of $E\{R + \gamma f(S') \mid S, A\}$.

Theorem 3 (Calibration error for histogram binning)**.** *Assume C1–C3. Then there is a constant $C < \infty$ such that, for any $K \in \mathbb{N}$, with probability at least $1 - \delta$,*

$$\mathrm{Cal}_{\ell^2}(V_K) \le C \left\{ \sqrt{\frac{B}{n} \log\Big(\frac{n}{B}\Big)} + \sqrt{\frac{\log(1/\delta)}{n}} \right\} + \big\| b_0\{(\widehat{w}_\pi - w_\pi)(\widehat{q}_{V_K} - q_{V_K})\} \big\|_{2,\rho} + \big\| \widehat{\mathcal{T}}_\pi V_K - \widehat{\mathcal{T}}_\pi V_{K-1} \big\|_{2,\rho}.$$

The bound contains three terms. The first is the statistical error of one-dimensional histogram calibration and matches classical rates when the calibration target is observed directly (Gupta et al., 2020; Gupta & Ramdas, 2021; van der Laan & Alaa, 2025). The second is the doubly robust auxiliary-model error from Theorem 2. The third is the finite-iteration error, which decays geometrically as γ^K under the contraction conditions in Appendix J.4, as in fitted value iteration (Munos & Szepesvári, 2008). Notably, the leading statistical term has no explicit dependence on the effective horizon; γ enters through the finite-iteration and auxiliary-model terms. Thus, estimating Bellman calibration remains a one-step conditional prediction problem, although translating calibration error into value-prediction error still depends on discounting and coverage through Theorem 1.

Value error. We next translate the calibration guarantee into a bound on value-prediction error. The result does not require realizability or Bellman completeness of the original value class; instead, it depends on approximation error within the histogram class, together with finite-iteration, statistical, and auxiliary-model errors.

Let $\Pi_B q := \arg\min_{h \in \mathcal{H}_B(\hat{v})} \|q - h\|_{2,\rho}$ denote the $L^2(\rho)$-projection onto $\mathcal{H}_B(\hat{v})$, define $\mathcal{B}_{\pi,B} := \Pi_B \mathcal{T}_\pi$, and let $V_B^\star$ satisfy $V_B^\star = \mathcal{B}_{\pi,B} V_B^\star$. Writing $P_{\pi,B} := \Pi_B P_\pi$, we have $P_{\pi,B} f(s) = E_\pi[f(S') \mid \hat{v}(S) \in I_{\hat{b}(s)}]$, where $\hat{b}(s) = b$ when $\hat{v}(s) \in I_b$. Define the corresponding aggregated discounted concentrability coefficient by $\mathcal{C}_{B,\gamma} := \left\{ (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \left\| \frac{d(\rho P_{\pi,B}^t)}{d\rho} \right\|_\infty^{1/2} \right\}^2$.

C4 Aggregated projected concentrability. $\mathcal{C}_{B,\gamma} < \infty$.

Theorem 4 (Value error for histogram calibration)**.** *Assume C1–C4. Then there exists a constant $C < \infty$ such that, for any $K \in \mathbb{N}$, with probability at least $1 - \delta$,*

$$\|V_K - v_0\|_{2,\rho} \le \gamma^K \|\hat{v} - V_B^\star\|_\infty + \frac{\sqrt{\mathcal{C}_{B,\gamma}}}{1-\gamma} \min_{h \in \mathcal{H}_B(\hat{v})} \|h - v_0\|_{2,\rho} + C \frac{\sqrt{\mathcal{C}_{B,\gamma}}}{1-\gamma} \left[\sqrt{\frac{B}{n} \log\Big(\frac{n}{B}\Big)} + \sqrt{\frac{\log(K/\delta)}{n}} + \max_{0 \le j \le K} \big\| b_0\{(\widehat{w}_\pi - w_\pi)(\widehat{q}_{V_j} - q_{V_j})\} \big\|_{2,\rho} \right].$$

The result avoids realizability and Bellman completeness of the original value class because these properties hold after aggregation. For any $f \in \mathcal{H}_B(\hat{v})$, $P_{\pi,B} f$ is binwise constant in $\hat{v}$, so $P_{\pi,B} f \in \mathcal{H}_B(\hat{v})$ and hence $\mathcal{B}_{\pi,B} f \in \mathcal{H}_B(\hat{v})$. Classical discretization arguments use state aggregation to recover Bellman consistency (Whitt, 1978; Van Roy, 2006; Chen & Jiang, 2019; Xie & Jiang, 2021). Here, we instead aggregate states through the scalar prediction $\hat{v}(S)$ to obtain Bellman calibration.

Theorems 3 and 4 show that the leading statistical difficulty of Bellman calibration is one-dimensional, even when the original value predictor is high-dimensional. With known importance weights and negligible iteration error, histogram calibration achieves Bellman calibration error of order $\sqrt{(B/n)\log(n/B)}$. The value-error bound contains the same stochastic error together with $\sqrt{\mathcal{C}_{B,\gamma}} \min_{h \in \mathcal{H}_B(\hat{v})} \|h - v_0\|_{2,\rho}/(1-\gamma)$, the approximation error from restricting the calibrated predictor to binned functions of $\hat{v}$. Thus, effective-horizon and concentrability penalties enter the value-error bound but not the leading calibration rate.

The number of bins B controls the tradeoff between approximation and estimation error. By the triangle inequality, $\min_{h \in \mathcal{H}_B(\hat{v})} \|h - v_0\|_{2,\rho} \le \|\hat{v} - v_0\|_{2,\rho} + \min_{h \in \mathcal{H}_B(\hat{v})} \|h - \hat{v}\|_{2,\rho}$. The second term is the error from approximating the identity map on the range of $\hat{v}(S)$. For bins of diameter $O(B^{-1})$, this error is $O(B^{-1})$, since the identity map is Lipschitz (Györfi et al., 2002). Balancing this approximation error with the histogram estimation error gives $B \asymp (n/\log n)^{1/3}$ and $\|V_K - v_0\|_{2,\rho} \le \frac{\sqrt{\mathcal{C}_{B,\gamma}}}{1-\gamma} \left[\|\hat{v} - v_0\|_{2,\rho} + O_p\{(\log n/n)^{1/3}\} \right]$, up to the finite-iteration and auxiliary-model terms. The Bellman calibration error converges at rate $O_p\{(\log n/n)^{1/3}\}$.

3.3 Isotonic Calibration

In supervised learning, isotonic calibration is a tuning-free alternative to histogram binning that learns the partition adaptively under a monotonicity constraint (Zadrozny & Elkan, 2002). A direct extension of Algorithm 1 takes $\mathcal{G} = \mathcal{G}_{\text{iso}}$, the class of nondecreasing functions on $\mathbb{R}$, and regresses the fitted Bellman targets on the initial predictions $\hat{v}(S_i)$ by isotonic regression at each iteration. The fit can be computed by the pool-adjacent-violators algorithm (PAVA) (Best & Chakravarti, 1990). As with histogram calibration, isotonic regression averages Bellman targets over groups of predicted values, but chooses these groups adaptively subject to monotonicity (Barlow & Brunk, 1972). The identity map belongs to $\mathcal{G}_{\text{iso}}$, so the monotonicity constraint does not exclude an already calibrated initial predictor.

C5 Finite variation. There exists $V < \infty$ such that, almost surely, the function $t \mapsto E[(\mathcal{T}_\pi V_K)(S) \mid \hat{v}(S) = t, \mathcal{C}_n]$ has total variation at most V.

Condition C5 controls the complexity of the one-dimensional calibration function and is analogous to controlling the number of bins B in histogram calibration (van der Laan et al., 2023). We do not assume that the calibration function is itself monotone.

Theorem 5 (Calibration error for isotonic calibration)**.** *Assume C1, C2, C3, and C5 with $\mathcal{G} = \mathcal{G}_{\text{iso}}$. Then there exists a constant $C < \infty$ such that, for any $K \in \mathbb{N}$, with probability at least $1 - \delta$,*

$$\mathrm{Cal}_{\ell^2}(V_K) \le C \left\{ n^{-1/3} + \sqrt{\frac{\log(1/\delta)}{n}} \right\} + \left\| b_0 \{ (\widehat{w}_\pi - w_\pi)(\widehat{q}_{V_K} - q_{V_K}) \} \right\|_{2,\rho} + \left\| \widehat{\mathcal{T}}_\pi V_K - \widehat{\mathcal{T}}_\pi V_{K-1} \right\|_{2,\rho}.$$

The isotonic analysis gives the standard one-dimensional $n^{-1/3}$ rate (Chatterjee et al., 2015), comparable to histogram calibration with $B \asymp n^{1/3}$ up to logarithmic factors, without requiring a choice of B. Monotonicity is used to determine how observations are pooled, while the calibration function itself need not be monotone.

The calibration guarantee above does not by itself ensure that isotonic recalibration preserves or improves value-prediction error relative to the initial predictor. The class of monotone transformations is not generally Bellman complete, and the isotonic partition may change across iterations, so the fixed-aggregation argument used for histogram calibration does not apply. To recover a value-error guarantee while retaining adaptive, tuning-free partitioning, we propose a hybrid isotonic–histogram procedure. We fit the isotonic calibrator once, take its plateaus $\{\widehat{I}_b\}_{b=1}^{\widehat{B}}$ as adaptive bins, freeze the partition, and apply the histogram Bellman-calibration updates within these bins. Subsequent iterates then remain in a fixed histogram class, allowing the histogram value-error analysis to apply while preserving the adaptive partition from the initial isotonic fit. If the initial fit has at most B_n plateaus with probability at least $1 - \delta$, the histogram guarantees hold with B replaced by B_n. Under standard isotonic regression conditions, B_n is of order $n^{1/3}$, yielding $n^{-1/3}$ calibration and stochastic value-error rates up to logarithmic, concentrability, and effective-horizon factors. Appendix D.1 gives the procedure and corresponding error bounds.

4 Experimental illustration

We use FQE as a representative fitted Bellman method to study the effect of post-hoc calibration on value prediction. The experiments consider both controlled distortions of fitted values and errors arising from function-class restriction, finite Bellman iteration, and temporal reward shift. The controlled experiments apply known affine or monotone transformations to fitted predictions to isolate correction of the prediction scale. The remaining experiments examine whether calibration improves errors arising in fitted value estimation and whether lower Bellman calibration error translates into lower value-prediction and policy-value error. We also evaluate neural FQE on D4RL continuous-control tasks (Fu et al., 2020).

In the synthetic experiments, ridge-regularized linear FQE is fit on a fixed random-feature representation. The transformation, restricted-feature, and finite-iteration experiments use five-fold cross-calibration, with the uncalibrated comparator given by the same FQE learner refit on the full

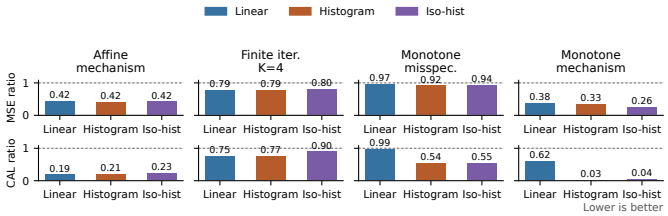


Figure 1: **Post-hoc calibration of fitted value predictions.** Calibrated-to-raw ratios for value MSE (top) and CAL (bottom) in four synthetic regimes; values below one favor calibration.

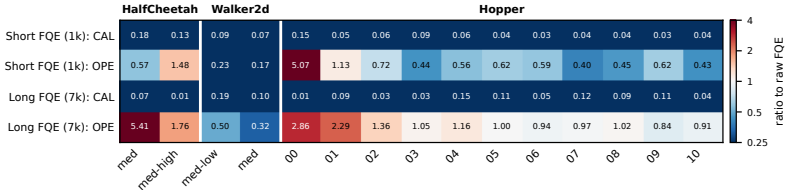


Figure 2: **Bellman calibration and OPE error under neural FQE.** D4RL calibrated-to-raw CAL/OPE ratios at 1,000/7,000 FQE updates; below one favors calibration. Appendix Table S2 reports intervals.

fitting sample (van der Laan et al., 2023). For temporal reward shift, FQE is trained on historical data, the calibrator is fit on a held-out sample from the current environment, and the comparator refits the same learner on pooled historical and current data. Synthetic results are evaluated using true-V MSE and held-out Bellman calibration error (CAL) over 100 replications. For D4RL, we use disjoint trajectory splits, with 60% used to train neural FQE, 20% to fit the calibrator, and 20% for evaluation. The uncalibrated comparator is the same neural FQE fit on the training split without subsequent calibration. Complete experimental details are given in Appendix B.

Calibration substantially reduces value error when the error is systematically related to the fitted prediction (Figure 1). Linear calibration corrects the affine transformation, while histogram and isotonic–histogram calibration correct the monotone nonlinear transformation. Calibration also improves value prediction when FQE is stopped after a small number of Bellman iterations. At the converged ridge-FQE solution, mean MSE remains lower after calibration, but the calibrated estimator no longer improves on raw FQE in a majority of replications, indicating that the gains become less systematic as the base fit improves. Under temporal reward shift, calibration fit on recent data reduces both value error and CAL relative to refitting FQE on pooled historical and current data.

For D4RL, calibration reduces CAL for every target policy at both FQE training lengths (Figure 2). OPE error improves for most policies after 1,000 updates but is mixed after 7,000 updates, when the raw FQE fits are themselves more accurate. Several policies also have substantially lower CAL after calibration without lower OPE error. These results illustrate the distinction in the calibration–refinement analysis. Post-hoc calibration corrects Bellman error determined by the scalar base prediction; its effect on OPE therefore depends on how well that prediction distinguishes state–action pairs with different values under Q^π.

5 Conclusion

Bellman calibration provides an estimable criterion for assessing the reliability of learned value functions by requiring predictions to agree, on average, with target-policy Bellman targets among states assigned the same predicted value. Fitted Bellman Calibration imposes this condition through one-dimensional post-hoc recalibration, with finite-sample guarantees that do not require Bellman completeness or value-function realizability. Its accuracy is governed by the information retained in the original predictor, off-policy coverage, and the accuracy of Bellman-target estimation. Appendix L extends the construction to the Bellman optimality operator, while stationary or occupancy-ratio weighting may improve calibration under substantial distribution shift (van der Laan & Kallus, 2025a;b; 2026).

AI USE STATEMENT

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

REFERENCES

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

Richard E Barlow and Hugh D Brunk. The isotonic regression problem and its dual. *Journal of the American Statistical Association*, 67(337):140–147, 1972.

Antonio Bella, Cèsar Ferri, José Hernández-Orallo, and María José Ramírez-Quintana. Calibration of machine learning models. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, pp. 128–146. IGI Global, 2010.

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

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

Michael J Best and Nilotpal Chakravarti. Active set algorithms for isotonic regression; a unifying framework. *Mathematical Programming*, 47(1):425–439, 1990.

Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. When does optimizing a proper loss yield calibration? *arXiv preprint arXiv:2305.18764*, 2023.

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

Sabyasachi Chatterjee, Adityanand Guntuboyina, and Bodhisattva Sen. On risk bounds in isotonic and other shape restricted regression problems. *The Annals of Statistics*, 43(4):1774–1800, 2015. doi: 10.1214/15-AOS1324.

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

Bo Cowgill and Catherine E Tucker. Algorithmic fairness and economics. *SSRN Electronic Journal*, 2020. doi: 10.2139/ssrn.3361280. URL https://ssrn.com/abstract=3361280.

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

Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

Hang Deng, Qiyang Han, and Cun-Hui Zhang. Confidence intervals for multiple isotonic regression and other monotone models. *The Annals of Statistics*, 49(4):2021–2052, 2021.

Yaqi Duan, Chi Jin, and Zhiyuan Li. Risk bounds and rademacher complexity in batch reinforcement learning. In *International Conference on Machine Learning*, pp. 2892–2902. PMLR, 2021.

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

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

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: Datasets for deep data-driven reinforcement learning, 2020.

Justin Fu, Mohammad Norouzi, Ofir Nachum, George Tucker, Ziyu Wang, Alexander Novikov, Mengjiao Yang, Michael R. Zhang, Yutian Chen, Aviral Kumar, Cosmin Paduraru, Sergey Levine, and Tom Le Paine. Benchmarks for deep off-policy evaluation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=kWSeGEeHvF8.

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

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Chirag Gupta and Aaditya Ramdas. Distribution-free calibration guarantees for histogram binning without sample splitting. In *International Conference on Machine Learning*, pp. 3942–3952. PMLR, 2021.

Chirag Gupta, Aleksandr Podkopaev, and Aaditya Ramdas. Distribution-free binary classification: prediction sets, confidence intervals and calibration. *Advances in Neural Information Processing Systems*, 33:3711–3723, 2020.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer, 2002.

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

Sarah Lichtenstein, Baruch Fischhoff, and Lawrence D Phillips. Calibration of probabilities: The state of the art. In *Decision Making and Change in Human Affairs: Proceedings of the Fifth Research Conference on Subjective Probability, Utility, and Decision Making, Darmstadt, 1–4 September, 1975*, pp. 275–324. Springer, 1977.

Ali Malik, Volodymyr Kuleshov, Jiaming Song, Danny Nemer, Harlan Seymour, and Stefano Ermon. Calibrated model-based deep reinforcement learning. In *International Conference on Machine Learning*, pp. 4314–4323. PMLR, 2019.

Lucas Maystre and Daniel Russo. Temporally-consistent survival analysis. *Advances in Neural Information Processing Systems*, 35:10671–10683, 2022.

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

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

Allan H Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12 (4):595–600, 1973. URL https://www.jstor.org/stable/26176769.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pp. 625–632, 2005.

John C. Platt. Probabilities for SV machines. In Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans (eds.), *Advances in Large-Margin Classifiers*, chapter 5, pp. 61–74. MIT Press, 2000. doi: 10.7551/mitpress/1113.003.0008.

Aaron Roth. Uncertain: Modern topics in uncertainty estimation, 2022. URL https://www.cis.upenn.edu/~aaroth/uncertainty-notes.pdf. Unpublished lecture notes.

John Rust. Optimal replacement of gmc bus engines: An empirical model of harold zurcher. *Econometrica: Journal of the Econometric Society*, pp. 999–1033, 1987.

Georgios Theocharous, Philip S Thomas, and Mohammad Ghavamzadeh. Ad recommendation systems for life-time value optimization. In *Proceedings of the 24th International Conference on World Wide Web Companion*, pp. 1305–1310, 2015. doi: 10.1145/2740908.2741998.

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

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

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

Ben Van Calster, David J. McLernon, Maarten van Smeden, Laure Wynants, and Ewout W. Steyerberg. Calibration: the achilles heel of predictive analytics. *BMC Medicine*, 17(1):230, 2019. ISSN 1741-7015. doi: 10.1186/s12916-019-1466-7. On behalf of Topic Group "Evaluating diagnostic tests and prediction models" of the STRATOS initiative.

Lars van der Laan and Ahmed Alaa. Generalized venn and venn-abers calibration with applications in conformal prediction. *arXiv preprint arXiv:2502.05676*, 2025.

Lars van der Laan and Ahmed M Alaa. Self-calibrating conformal prediction. *Advances in Neural Information Processing Systems*, 37:107138–107170, 2024.

Lars van der Laan and Nathan Kallus. Fitted q-evaluation without bellman completeness via occupancy weighting. *arXiv preprint arXiv:2512.23805*, 2025a. URL https://arxiv.org/abs/2512.23805.

Lars van der Laan and Nathan Kallus. Soft fitted q-iteration without bellman completeness: Occupancy reweighting and temperature annealing. *arXiv preprint arXiv:2512.23927*, 2025b. URL https://arxiv.org/abs/2512.23927.

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

Lars van der Laan, Ernesto Ulloa-Pérez, Marco Carone, and Alex Luedtke. Causal isotonic calibration for heterogeneous treatment effects. In *International Conference on Machine Learning*, pp. 34831–34854. PMLR, 2023.

Lars van der Laan, Ziming Lin, Marco Carone, and Alex Luedtke. Stabilized inverse probability weighting via isotonic calibration. *arXiv preprint arXiv:2411.06342*, 2024.

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

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

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

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Justin Whitehouse, Christopher Jung, Vasilis Syrgkanis, Bryan Wilder, and Zhiwei Steven Wu. Orthogonal causal calibration. *arXiv preprint arXiv:2406.01933*, 2024.

Ward Whitt. Approximations of dynamic programs, i. *Mathematics of Operations Research*, 3(3): 231–243, 1978.

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

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

Yizhe Xu and Steve Yadlowsky. Calibration error for heterogeneous treatment effects. In *International Conference on Artificial Intelligence and Statistics*, pp. 9280–9303. PMLR, 2022.

Zhenghai Xue, Qingpeng Cai, Bin Yang, Lantao Hu, Peng Jiang, Kun Gai, and Bo An. Auro: Reinforcement learning for adaptive user retention optimization in recommender systems. In *Proceedings of the ACM on Web Conference 2025*, pp. 391–401, 2025.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *ICML*, pp. 609–616, 2001.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 694–699, 2002.

Joshua P. Zitovsky, Daniel de Marchi, Rishabh Agarwal, and Michael Rene Kosorok. Revisiting Bellman errors for offline model selection. In *International Conference on Machine Learning*, pp. 43369–43406. PMLR, 2023.

A Supplementary Experimental Results

Table S1: **Direct V-space calibration under temporal reward shift, relative to pooled all-data FQE.** The pooled baseline fits the same random-feature FQE learner to 2,000 historical and 100 recent current-regime transitions. Ratios below one are better. Brackets are paired percentile-bootstrap 95% intervals over 100 matched replications; win rates are reported as true-V MSE/Bellman CAL.

Method	Rel. true-V MSE	Rel. Bellman CAL error	Win rate
Historical-only predictor	1.065 [1.058, 1.072]	1.072 [1.068, 1.076]	0.02/0.01
Pooled all-data predictor	1.000	1.000	–
Recent linear V-calibration	0.671 [0.644, 0.696]	0.769 [0.682, 0.869]	0.98/0.79
Recent isotonic V-calibration	0.667 [0.642, 0.692]	0.982 [0.813, 1.225]	0.97/0.62

Table S2: **Target-policy-stratified training sensitivity.** CAL and OPE entries are ratios to the matched uncalibrated FQE predictor; values below one favor calibration. Brackets are paired-bootstrap 95% intervals over 20 matched trajectory splits. Reference returns are used only to compute OPE error.

Environment	Policy	Short FQE (1k)		Long FQE (7k)	
		CAL	OPE	CAL	OPE
HalfCheetah	med	0.18 [0.10, 0.27]	0.57 [0.42, 0.74]	0.07 [0.03, 0.13]	5.41 [3.96, 7.94]
	med-high	0.13 [0.04, 0.28]	1.48 [1.24, 1.72]	0.01 [0.00, 0.03]	1.76 [1.70, 1.82]
Walker2d	med-low	0.09 [0.05, 0.14]	0.23 [0.15, 0.31]	0.19 [0.13, 0.26]	0.50 [0.38, 0.62]
	med	0.07 [0.04, 0.10]	0.17 [0.13, 0.22]	0.10 [0.07, 0.15]	0.32 [0.24, 0.41]
Hopper	00	0.15 [0.07, 0.27]	5.07 [4.58, 5.54]	0.01 [0.00, 0.01]	2.86 [2.12, 3.48]
	01	0.05 [0.02, 0.08]	1.13 [0.67, 1.63]	0.09 [0.03, 0.15]	2.29 [1.65, 3.12]
	02	0.06 [0.03, 0.09]	0.72 [0.57, 0.87]	0.03 [0.02, 0.05]	1.36 [1.21, 1.47]
	03	0.09 [0.06, 0.13]	0.44 [0.33, 0.55]	0.03 [0.02, 0.06]	1.05 [0.87, 1.20]
	04	0.06 [0.03, 0.10]	0.56 [0.45, 0.67]	0.15 [0.07, 0.26]	1.16 [0.97, 1.31]
	05	0.04 [0.02, 0.06]	0.62 [0.50, 0.74]	0.11 [0.05, 0.17]	1.00 [0.80, 1.17]
	06	0.03 [0.01, 0.05]	0.59 [0.45, 0.73]	0.05 [0.03, 0.09]	0.94 [0.75, 1.12]
	07	0.04 [0.02, 0.07]	0.40 [0.30, 0.50]	0.12 [0.06, 0.20]	0.97 [0.76, 1.14]
	08	0.04 [0.02, 0.06]	0.45 [0.32, 0.57]	0.09 [0.04, 0.15]	1.02 [0.85, 1.17]
	09	0.03 [0.01, 0.05]	0.62 [0.52, 0.71]	0.11 [0.05, 0.17]	0.84 [0.63, 1.04]
	10	0.04 [0.02, 0.07]	0.43 [0.35, 0.51]	0.04 [0.02, 0.06]	0.91 [0.74, 1.08]

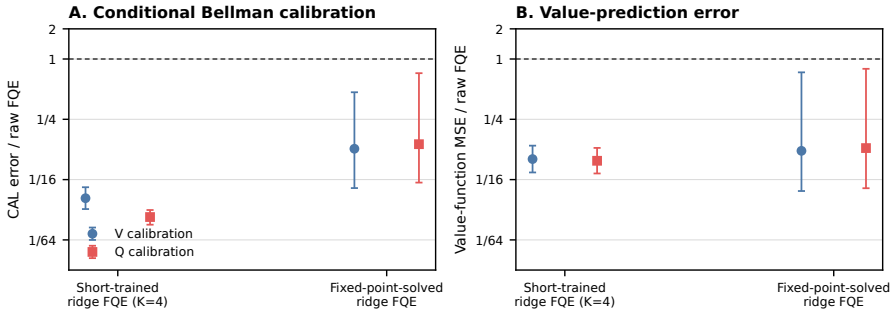


Figure S1: **Calibration across base-FQE training regimes.** Ratios to the matched raw ridge-FQE predictor over 100 replications; points and bars are ratios of means and paired-bootstrap 95% intervals. At $K = 4$, calibration improves CAL and true-V MSE in nearly every replication. At the FQE fixed point, mean true-V MSE remains lower after calibration, although fewer than half of replications improve.

B Additional Experimental Details

B.1 Experiment design

Common simulation design. The simulation studies use discounted MDPs with continuous states $S_t \in \mathbb{R}^d$, actions $A_t \in \{1, 2, 3\}$, and softmax target and behavior policies. The target policy has logits $s^\top w_a^\pi/\tau_\pi$, while the behavior policy has logits $s^\top w_a^b/\tau_b$, with $w_a^b = w_a^\pi - \Delta u_a$. Initial states satisfy $S_0 \sim N(\mu_0, I_d)$. Except where specified below, transitions and mean rewards are

$$S_{t+1} = \tanh\{MS_t + \alpha_{A_t} + 0.22\sin(S_t) + 0.08\cos(S_t^{\text{rev}})\} + \sigma_\varepsilon \varepsilon_t,$$

and

$$r_0(s, a) = s^\top \beta + 0.65\sin(s_1) + 0.35\cos(s_2) + s^\top \eta_a + \rho_a - 0.06\|s\|_2^2.$$

Observed rewards add mean-zero Gaussian noise to r_0. Reference values $V^\pi(s)$ are computed from independent target-policy Monte Carlo rollouts using the conditional mean reward. Training data, calibration data, Bellman-evaluation data, evaluation states, and Monte Carlo rollouts are generated independently. Results use 100 independent replications unless stated otherwise.

Controlled prediction distortions. To isolate correction of the prediction scale, we first fit a common random-feature FQE predictor and then transform its predictions using

$$D_{\text{aff}}(q) = 2 + 0.45q \quad \text{or} \quad D_{\text{sat}}(q) = 0.2\tanh(q/0.2).$$

The first produces a known affine distortion and the second a monotone nonlinear distortion.

Affine misspecification. We use the common nonlinear MDP with reward

$$r_{\text{aff}}(s, a) = 1.25 + 0.7\, r_0(s, a)$$

and fit linear FQE with a restricted feature class. This produces a setting in which an affine correction of the fitted value scale is appropriate.

Finite Bellman iteration. To study error from incomplete fitted iteration, we use

$$S_{t+1} = 0.92S_t + \alpha_{A_t} + 0.03\sin(S_t) + \sigma_\varepsilon \varepsilon_t,$$

with $\gamma = 0.985$, horizon 120, and reward

$$r_{\text{fin}}(s, a) = 0.8 + 0.9\tanh(s_1) + 0.25s_2 + \rho_a + 0.15\tanh(s_1 + 0.5s_2) + s^\top \eta_a.$$

Random-feature FQE is stopped after $K = 4$ Bellman iterations.

For the training-sensitivity experiment, we hold the sample, random-feature map, and ridge penalty fixed and compare $K = 4$ with the ridge-FQE fixed point. If Z and Z' denote the current and next state–action feature matrices, including an unpenalized intercept, and D is the ridge penalty matrix, the fixed-point coefficients solve

$$\{Z^\top Z + \lambda D - \gamma Z^\top Z'\}\theta = Z^\top R.$$

Temporal reward shift. The historical and current environments have the same transition law, policies, and state distribution but differ in reward:

$$r_{\text{current}}(s,a) = 2.0 + 1.5\, r_{\text{old}}(s,a).$$

Random-feature FQE is trained on $n_{\text{old}} = 2{,}000$ historical transitions, and the calibration map is fit on $n_{\text{recent}} = 100$ current-regime transitions. Evaluation uses independent current-regime data. The comparator refits the same FQE learner on the pooled 2,100 historical and current-regime transitions.

Monotone misspecification. We transform the reward according to

$$r_{\text{mono}}(s,a) = \text{sign}\{r_0(s,a)\}\sqrt{|r_0(s,a)| + 1}$$

and fit linear FQE with the same restricted feature class. This produces a nonlinear monotone distortion of the fitted value scale that need not be removed by an affine calibration map.

Cross-calibration. Each method first fits a Q-function to the offline behavior-policy data and forms

$$\widehat{V}(s) = \sum_a \pi(a \mid s)\widehat{Q}(s,a).$$

We use linear or random-feature FQE depending on the experiment. Outside the temporal-shift setting, calibration uses five-fold cross-fitting (van der Laan et al., 2023). For fold k, the Q-function is trained on the other four folds, and the held-out fold contributes

$$\left\{\widehat{V}^{(-k)}(S_i), \widehat{V}^{(-k)}(S_i'), R_i\right\}.$$

We pool these out-of-fold observations across folds and fit a single calibration map g. At evaluation states, the calibrated fold-specific predictions are aggregated by

$$\widehat{V}_{\text{cal}}(s) = \text{median}_{k=1,\ldots,K_{\text{fold}}}\, g\left\{\widehat{V}^{(-k)}(s)\right\}.$$

The uncalibrated comparator refits the same Q-learner on the full training sample.

Calibration updates. For each calibration observation, let

$$\omega_i = \frac{\pi(A_i \mid S_i)}{b(A_i \mid S_i)}.$$

The action-ratio weights are clipped at 20 and normalized to have empirical mean one. Starting from the identity map, we iterate

$$g_{t+1} \in \arg\min_g \sum_i \omega_i \left[g\{\widehat{V}(S_i)\} - \left\{R_i + \gamma g_t\{\widehat{V}(S_i')\}\right\}\right]^2.$$

We use affine, histogram, isotonic, and hybrid isotonic–histogram calibration.

Evaluation metrics. The scalar OPE estimate averages $g\{\widehat{V}(S_0)\}$ over independent initial states. Value-prediction error is measured by

$$\mathbb{E}\left[\left\{g(\widehat{V}(S)) - V^\pi(S)\right\}^2\right],$$

using independent evaluation states and Monte Carlo estimates of $V^\pi(S)$.

Bellman calibration error is evaluated on an independent transition sample. We partition $g\{\widehat{V}(S)\}$ into prediction-quantile cells and, within each cell, compute the squared difference between the weighted mean prediction and weighted mean Bellman target

$$r_0(S,A) + \gamma g\{\widehat{V}(S')\}.$$

The cellwise errors are averaged using the action-ratio weights.

Simulation sizes. The primary simulation comparisons use 2,500 Bellman-evaluation transitions and 30 quantile cells; the training-sensitivity experiment in Figure S1 uses 5,000 transitions and 50 cells. True-V MSE uses the same evaluation states, with one Monte Carlo rollout per state in the primary comparisons and two in the training-sensitivity experiment. The controlled-distortion settings use $n = 400$, $d = 6$, $\gamma = 0.90$, and horizon 30. The affine and monotone misspecification settings use $n = 500$, $d = 6$, and linear FQE. The finite-iteration setting uses $n = 1{,}200$, $d = 4$, $\gamma = 0.985$, and horizon 120.

C D4RL Q-Function Calibration Benchmark

Data and sample splitting. We apply post-hoc calibration to fitted Q-functions on the HalfCheetah-medium, Walker2d-medium, and Hopper-medium D4RL datasets (Fu et al., 2020). We evaluate two fixed target policies for HalfCheetah and Walker2d and all eleven Hopper target policies from the Deep OPE benchmark (Fu et al., 2021), which also provides Monte Carlo reference returns. For each policy, we use 20 random trajectory splits and two neural-FQE checkpoints, giving 600 fitted models. Within each split, 60% of trajectories fit the base Q-function, 20% fit the calibrator, and 20% are held out for evaluation. We cap each dataset at 192 trajectories and use $\gamma = 0.99$ for HalfCheetah and Walker2d and $\gamma = 0.995$ for Hopper.

Within each split and policy, the raw and calibrated predictors share the same base Q-function and evaluation sample. Reference returns are used only to compute scalar OPE error.

Q-function calibration. Let $\widehat{Q}(s, a)$ denote the fitted Q-function. We learn a one-dimensional map g from the scalar predictions $\widehat{Q}(S_i, A_i)$, using Bellman targets $Y_i(g) = R_i + \gamma \mathbb{E}_{A' \sim \pi(\cdot|S_i')} g\{\widehat{Q}(S_i', A')\}$. The corresponding scalar value estimate averages $g\{\widehat{Q}(S_0, A_0)\}$ over initial states and target-policy actions.

We calibrate Q directly so that the calibration sample remains the logged state–action distribution. Equivalently, this is value calibration on the Markov chain with state $X = (S, A)$ and next state $X' = (S', A')$, $A' \sim \pi(\cdot \mid S')$. This avoids the additional state-distribution reweighting that would arise from first forming $\widehat{V}(S) = \mathbb{E}_{A \sim \pi(\cdot|S)} \widehat{Q}(S, A)$.

Base learner. The base learner is neural FQE with two hidden layers of width 256. We evaluate checkpoints after 1,000 and 7,000 gradient updates, using learning rate 3×10^{-4}, target-network rate 0.005, and batch size 256. Within each split–policy pair, the checkpoints use the same initialization, architecture, and data split.

Metrics and uncertainty. We report held-out CAL error and scalar OPE absolute error. CAL is evaluated using ten prediction-quantile cells, with tied step-function predictions kept in the same cell, and the exact within-cell leave-one-out debiased moment. OPE error is the absolute difference between the estimated and reference policy values. Since state–action reference values for Q^π are unavailable, we do not report Q-function MSE. Table entries are ratios to the matched raw Q-function, with paired-bootstrap intervals over the 20 splits computed separately for each target policy.

D Additional Details and Extensions

D.1 Hybrid Isotonic–Histogram Calibration

Fitted isotonic Bellman calibration adaptively pools nearby predicted values through the monotonicity constraint, but the induced partition may change across Bellman iterations. The iterates therefore need not correspond to a single aggregated Bellman operator.

The hybrid isotonic–histogram procedure instead selects the partition once. It first regresses the initial Bellman target isotonicly on $\hat{v}(S)$, uses the flat regions of the fitted isotonic function as bins, and then runs histogram Bellman calibration on this fixed partition. This retains data-adaptive binning while recovering the fixed-partition structure used in the histogram analysis.

Algorithm S1 Iso–Hist Fitted Bellman Calibration

Input: Value predictor $\hat{v}$, calibration data $\mathcal{C}_n$, Bellman target-estimate map $v \mapsto \widehat{Y}_v$, iterations K
Stage 1: select partition
  $\widehat{Y}_i^{(0)} \leftarrow \widehat{Y}_{\hat{v}}(O_i)$
  $g_{\text{iso}} \leftarrow \arg\min_{g \in \mathcal{G}_{\text{iso}}} \sum_{i=1}^{n} (\widehat{Y}_i^{(0)} - g(\hat{v}(S_i)))^2$
  Extract bins $\hat{I}_1, \ldots, \hat{I}_{\hat{B}}$ from the flat regions of g_{iso}
  Define $\widehat{\mathcal{G}}_{\hat{B}} := \{g : g \text{ is constant on each } \hat{I}_b\}$
  $V_0 \leftarrow g_{\text{iso}} \circ \hat{v}$
Stage 2: fixed-partition calibration
for $k = 0, \ldots, K-1$ **do**
  $V_{k+1} = g_{k+1} \circ \hat{v}$, where $g_{k+1} \in \arg\min_{g \in \widehat{\mathcal{G}}_{\hat{B}}} \sum_{i=1}^{n} \{\widehat{Y}_{V_k}(O_i) - g(\hat{v}(S_i))\}^2$
end for
Output: V_K

Once the partition is fixed, the analysis of Section 3.2 applies conditionally on the selected partition. The stochastic terms depend on a deterministic high-probability bound on the number of selected bins, while the finite-iteration term starts from the isotonic warm start. The same argument applies to other data-adaptive binning methods that select and then freeze a partition before Bellman iteration.

Theorem S1 (Calibration and Estimation Error for Alg. S1)**.** *Let V_K denote the output of Algorithm S1. Suppose the isotonic regression step produces at most $\hat{B}$ bins, where $\hat{B} \le B_n$ for some deterministic sequence B_n with probability at least $1-\delta$. Assume C1–C4 and C5, with Condition C4 imposed on the selected partition. Then, with probability at least $1-2\delta$, the guarantees in Theorems 3 and 4 hold for the selected fixed partition, with B in the stochastic terms replaced by B_n. In Theorem 4, the finite-iteration term is replaced by $\gamma^K \|g_{\text{iso}} \circ \hat{v} - V^\star_{\hat{B}}\|_\infty$, where $V^\star_{\hat{B}}$ is the projected Bellman fixed point for the selected partition.*

Existing isotonic regression theory gives $\widehat{B} = O(n^{1/3})$ under standard smoothness conditions (Deng et al., 2021); the number of constant segments can also be constrained directly (van der Laan & Alaa, 2024). For the value-error bound, the selected histogram class must also approximate the identity map on the range of $\hat{v}$. When $\hat{v}$ is already close to v_0, this approximation error is small; substantial isotonic pooling can otherwise increase the refinement error.

E Stationary Measures and Contraction Results

Lemma S2 (Bellman contraction under a stationary measure)**.** *Let P be a Markov operator with stationary distribution μ (i.e., $\mu = \mu P$), and define the Bellman operator $\mathcal{T} f := r + \gamma P f$. Then P is nonexpansive in $L^2(\mu)$:*

$$\|P(f-g)\|_{2,\mu} \;\le\; \|f-g\|_{2,\mu}.$$

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

$$\|\mathcal{T}(f-g)\|_{2,\mu} \;\le\; \gamma\,\|f-g\|_{2,\mu}.$$

Proof. Because μ is stationary for P, the operator P is nonexpansive in $L^2(\mu)$. Jensen's inequality implies

$$(Ph)^2 \le P(h^2) \qquad \text{pointwise.}$$

Integrating both sides with respect to μ and using $\mu = \mu P$ yields

$$\|Ph\|_{2,\mu}^2 = \int (Ph)^2\, d\mu \le \int P(h^2)\, d\mu = \int h^2\, d\mu = \|h\|_{2,\mu}^2.$$

Now apply this with $h = f - g$. Since the reward function r cancels,

$$\mathcal{T} f - \mathcal{T} g = \gamma\, P(f-g).$$

Therefore,

$$\|\mathcal{T} f - \mathcal{T} g\|_{2,\mu} = \gamma \|P(f-g)\|_{2,\mu} \le \gamma \|f-g\|_{2,\mu},$$

which proves the claim. □

Lemma S3 (Stationarity under coarsening)**.** *Let P be a Markov kernel on $\mathcal{S}$ and let λ be stationary for P (i.e., $\lambda P = \lambda$). For any measurable coarsening map $g : \mathcal{S} \to \mathbb{R}$, define*

$$P_g f(s) := \mathbb{E}[\, f(S') \mid g(S) = g(s)\,],$$

where (S, S') has joint law $\lambda(ds)\, P(s, ds')$. Then λ is stationary for P_g.

Proof. Let $(S, S') \sim \lambda(ds)\, P(s, ds')$, so $S \sim \lambda$ and $S' \mid S \sim P(S, \cdot)$. Then

$$\begin{aligned}\int P_g f(s)\, \lambda(ds) &= \mathbb{E}_{\lambda,P}[P_g f(S)] \\ &= \mathbb{E}_{\lambda,P}[\mathbb{E}_{\lambda,P}[f(S') \mid g(S)]] \\ &= \mathbb{E}_{\lambda,P}[f(S')].\end{aligned}$$

Since λ is stationary for P, the marginal of S' is again λ, hence $\mathbb{E}_{\lambda,P}[f(S')] = \int f(s)\, \lambda(ds)$. □

F Calibration–Refinement Proofs

Set

$$\mathcal{B}_{\pi,\hat{v}} q := \Pi_{\hat{v}} \mathcal{T}_\pi q, \qquad V_{\hat{v}}^\star \in \mathcal{V}_{\hat{v}}, \qquad V_{\hat{v}}^\star = \mathcal{B}_{\pi,\hat{v}} V_{\hat{v}}^\star.$$

Theorem S4 (Sharper calibration–refinement bound)**.** *Under A1,*

$$\|\hat{v} - v_0\|_{2,\rho} \le \|\hat{v} - V_{\hat{v}}^\star\|_{2,\rho} + \frac{\sqrt{\mathcal{C}_{\hat{v},\gamma}}}{1-\gamma} \min_{\theta:\mathbb{R}\to\mathbb{R}} \|\theta \circ \hat{v} - v_0\|_{2,\rho}.$$

Proof of Theorem S4. Define $P_{\pi,\hat{v}} := \Pi_{\hat{v}} P_\pi$ and $r_{\pi,\hat{v}} := \Pi_{\hat{v}} r_\pi$. Because v_0 and $V_{\hat{v}}^\star$ satisfy the fixed-point equations

$$v_0 = \mathcal{T}_\pi v_0, \qquad V_{\hat{v}}^\star = \mathcal{B}_{\pi,\hat{v}} V_{\hat{v}}^\star = r_{\pi,\hat{v}} + \gamma\, P_{\pi,\hat{v}} V_{\hat{v}}^\star,$$

we begin by writing

$$V_{\hat{v}}^\star - v_0 = \big(\mathcal{B}_{\pi,\hat{v}} V_{\hat{v}}^\star - \mathcal{B}_{\pi,\hat{v}} v_0\big) + \big(\mathcal{B}_{\pi,\hat{v}} v_0 - \mathcal{T}_\pi v_0\big).$$

Using linearity on differences,

$$\mathcal{B}_{\pi,\hat{v}} V_{\hat{v}}^\star - \mathcal{B}_{\pi,\hat{v}} v_0 = \gamma\, P_{\pi,\hat{v}}(V_{\hat{v}}^\star - v_0),$$

so

$$V_{\hat{v}}^\star - v_0 = \gamma\, P_{\pi,\hat{v}}(V_{\hat{v}}^\star - v_0) + \big(\mathcal{B}_{\pi,\hat{v}} v_0 - v_0\big),$$

where we used $\mathcal{T}_\pi v_0 = v_0$ for the second term. Since $\mathcal{B}_{\pi,\hat{v}} v_0 = \Pi_{\hat{v}} v_0$, this gives

$$V_{\hat{v}}^\star - v_0 = \gamma P_{\pi,\hat{v}}(V_{\hat{v}}^\star - v_0) + (\Pi_{\hat{v}} - I) v_0.$$

Let $a := (\Pi_{\hat{v}} - I) v_0$. Iterating the display yields

$$V_{\hat{v}}^\star - v_0 = \sum_{t=0}^{\infty} \gamma^t P_{\pi,\hat{v}}^t a.$$

For each $t \ge 0$, set

$$c_t := \left\| \frac{d(\rho P_{\pi,\hat{v}}^t)}{d\rho} \right\|_\infty.$$

Jensen's inequality implies

$$\|P_{\pi,\hat{v}}^t a\|_{2,\rho} \le \sqrt{c_t}\, \|a\|_{2,\rho}.$$

Therefore,

$$\|V_{\hat{v}}^\star - v_0\|_{2,\rho} \le \frac{\sqrt{\mathcal{C}_{\hat{v},\gamma}}}{1-\gamma} \|\Pi_{\hat{v}} v_0 - v_0\|_{2,\rho}. \tag{$\star$}$$

Since $\Pi_{\hat{v}}$ is the $L^2(\rho)$-projection onto $\{\theta \circ \hat{v} : \theta : \mathbb{R} \to \mathbb{R}\}$,

$$\|\Pi_{\hat{v}} v_0 - v_0\|_{2,\rho} = \min_{\theta:\mathbb{R}\to\mathbb{R}} \|\theta \circ \hat{v} - v_0\|_{2,\rho}.$$

Finally, by the triangle inequality,

$$\|\hat{v} - v_0\|_{2,\rho} \le \|\hat{v} - V_{\hat{v}}^\star\|_{2,\rho} + \|V_{\hat{v}}^\star - v_0\|_{2,\rho},$$

Combining this display with ($\star$) yields the stated result. □

Proof of Theorem 1. Let $P_{\pi,\hat{v}} := \Pi_{\hat{v}} P_\pi$. As in the proof of Theorem S4,

$$\|V^\star_{\hat{v}} - v_0\|_{2,\rho} \le \frac{\sqrt{\mathcal{C}_{\hat{v},\gamma}}}{1-\gamma} \min_{\theta:\mathbb{R}\to\mathbb{R}} \|\theta \circ \hat{v} - v_0\|_{2,\rho}. \tag{$\star$}$$

Moreover,

$$\hat{v} - V^\star_{\hat{v}} = \hat{v} - \mathcal{B}_{\pi,\hat{v}}\hat{v} + \gamma P_{\pi,\hat{v}}(\hat{v} - V^\star_{\hat{v}}),$$

and therefore

$$\hat{v} - V^\star_{\hat{v}} = \sum_{t=0}^{\infty} \gamma^t P^t_{\pi,\hat{v}}\{\hat{v} - \mathcal{B}_{\pi,\hat{v}}\hat{v}\}.$$

Since $\mathcal{B}_{\pi,\hat{v}}\hat{v} = \Gamma_0(\hat{v})$, the concentrability argument above yields

$$\|\hat{v} - V^\star_{\hat{v}}\|_{2,\rho} \le \frac{\sqrt{\mathcal{C}_{\hat{v},\gamma}}}{1-\gamma} \mathrm{Cal}_{\ell^2}(\hat{v}). \tag{$\dagger$}$$

The triangle inequality and displays ($\star$) and ($\dagger$) give the claim. □

G Proof of Theorem 2

Proof of Theorem 2. Fix v and treat the nuisance estimates as fixed functions. By definition,

$$\widehat{\mathcal{T}}_\pi v(s) = \mathbb{E}\Big[(\pi\hat{q}_v)(S) + \widehat{w}_\pi(A \mid S)\big\{R + \gamma\, v(S') - \hat{q}_v(S,A)\big\} \;\Big|\; S = s\Big].$$

Since $(\pi\hat{q}_v)(S)$ depends only on S,

$$\mathbb{E}\big[(\pi\hat{q}_v)(S) \mid S = s\big] = (\pi\hat{q}_v)(s).$$

For the second term, using the definition of b_0 and $q_v := r_0 + \gamma P v$,

$$\begin{aligned}
&\mathbb{E}\Big[\widehat{w}_\pi(A \mid S)\big\{R + \gamma\, v(S') - \hat{q}_v(S,A)\big\} \;\Big|\; S = s\Big] \\
&\quad = b_0\Big\{\widehat{w}_\pi(\cdot \mid s)\,\mathbb{E}\big[R + \gamma\, v(S') - \hat{q}_v(s,\cdot) \mid S = s, A = \cdot\big]\Big\}(s) \\
&\quad = b_0\big\{\widehat{w}_\pi(q_v - \hat{q}_v)\big\}(s).
\end{aligned}$$

Hence

$$\widehat{\mathcal{T}}_\pi v(s) = (\pi\hat{q}_v)(s) + b_0\big\{\widehat{w}_\pi(q_v - \hat{q}_v)\big\}(s).$$

Subtracting $(\mathcal{T}_\pi v)(s) = (\pi q_v)(s)$ yields

$$\begin{aligned}
\widehat{\mathcal{T}}_\pi v(s) - (\mathcal{T}_\pi v)(s) &= (\pi\hat{q}_v - \pi q_v)(s) + b_0\big\{\widehat{w}_\pi(q_v - \hat{q}_v)\big\}(s) \\
&= (\pi(\hat{q}_v - q_v))(s) - b_0\big\{\widehat{w}_\pi(\hat{q}_v - q_v)\big\}(s).
\end{aligned}$$

Using $(\pi f)(s) = b_0\{w_\pi f\}(s)$ with $f = \hat{q}_v - q_v$ gives

$$(\pi(\hat{q}_v - q_v))(s) = b_0\{w_\pi(\hat{q}_v - q_v)\}(s),$$

so

$$\widehat{\mathcal{T}}_\pi v(s) - (\mathcal{T}_\pi v)(s) = b_0\big\{(w_\pi - \widehat{w}_\pi)(\hat{q}_v - q_v)\big\}(s).$$

Taking conditional expectations given $v(S) = t$ proves the claim. □

H Notation and Maximal Inequalities

Let P_0 denote the joint distribution of (S, A, R, S') induced by the behavior policy, and let P_n denote the empirical measure of the calibration sample $\mathcal{C}_n$.

For a state function f, define the empirical L^2 norms with respect to the state and next-state samples by

$$\|f\|_{n,S} := \Big(\tfrac{1}{n}\sum_{i=1}^n f(S_i)^2\Big)^{1/2}, \quad \|f\|_{n,S'} := \Big(\tfrac{1}{n}\sum_{i=1}^n f(S'_i)^2\Big)^{1/2},$$

where S_i and S_i' denote the observed states and next states, respectively, in the calibration sample $\mathcal{C}_n$.

Further define

$$\widehat{\mathcal{G}} := \Big\{ (f_1 - f_2)\big(\widehat{Y}_{f_2} - f_2\big) : \ f_1, f_2 \in \mathcal{H}_B(\hat{v}) \Big\}.$$

By assumption, both $\hat{v}$ and the Bellman target-estimate map $v \mapsto \widehat{Y}_v$ are fixed conditional on the training data, which is independent of the calibration sample $\mathcal{C}_n$. Consequently, the classes $\mathcal{H}_B(\hat{v})$ and $\widehat{\mathcal{G}}$ are non-random conditional on the training dataset.

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

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

where the supremum is taken over all discrete probability distributions Q.

Write $x \lessapprox y$ when x is bounded above by y up to a constant that depends only on the global constants in our conditions.

H.1 Local maximal inequality

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

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

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

$$\mathcal{R}_n(\mathcal{F}, \delta) := \mathbb{E}\left[\sup_{\substack{f \in \mathcal{F} \\ \|f\|_{2,\bar{P}} \le \delta}} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(O_i) \right],$$

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

The next result states Lemma 11 of Foster & Syrgkanis (2023) in the present notation.

Lemma S5 (Local maximal inequality)**.** *Let $\mathcal{F}$ be a star-shaped class of functions satisfying $\sup_{f \in \mathcal{F}} \|f\|_\infty \le M$. Let $\delta > 0$ satisfy the critical radius condition $\mathcal{R}_n(\mathcal{F}, \delta) \le \delta^2$. Suppose further that $n^{-1/2}\sqrt{\log\log(1/\delta)} = o(\delta)$. Then there exists a universal constant $C > 0$ such that, for all $u \ge 1$, with probability at least $1 - e^{-u^2}$, every $f \in \mathcal{F}$ satisfies*

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

Proof. Apply Lemma 11 of Foster & Syrgkanis (2023) with the notation above. □

The following lemma bounds the localized Rademacher complexity by the uniform entropy integral.

Lemma S6. *Let $\mathcal{F}$ be a star-shaped class of functions such that $\sup_{f \in \mathcal{F}} \|f\|_\infty \le M$. Then, for every $\delta > 0$,*

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

where the implicit constant depends only on M.

Proof of Lemma S6. Apply Theorem 2.1 of van der Vaart & Wellner (2011) to the $L^2(Q)$-localized class, after rescaling its envelope to one, and use Rademacher symmetrization. In terms of the absolute entropy integral $\mathcal{J}(\delta, \mathcal{F})$, its expectation bound is

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

which proves the lemma. □

I Proof of Theorem 3

I.1 Technical lemmas

Lemma S7. *Under our conditions,*

$$\mathcal{J}\left(\delta, \widehat{\mathcal{G}}\right) \lesssim \delta \sqrt{B \log(1/\delta)},$$

where the implicit constant is independent of B.

Proof. By Condition C3,

$$\|\hat{q}_f - \hat{q}_g\|_{n,S,A} \le L \|f - g\|_{n,S'}$$

almost surely. For each s, by Jensen's inequality,

$$|\pi h(s)|^2 = \left| \sum_a \pi(a \mid s) h(a, s) \right|^2 \le \sum_a \pi(a \mid s) |h(a, s)|^2,$$

so $\|\pi h\|_{n,S} \le \|h\|_{n,S,A}$. Hence, we also have

$$\|\pi(\hat{q}_f - \hat{q}_g)\|_{n,S} \le \|\hat{q}_f - \hat{q}_g\|_{n,S,A} \le L \|f - g\|_{n,S'}.$$

Hence, by the definition of $\widehat{Y}_f$ and boundedness of $\widehat{w}_\pi$,

$$\|\widehat{Y}_f - \widehat{Y}_g\|_{L^2(P_n)} \lesssim \|f - g\|_{n,S'}.$$

Take $f, g \in \widehat{\mathcal{G}}$ with $f = (f_1 - f_2)\left(\widehat{Y}_{f_2} - f_2\right)$ and $g = (g_1 - g_2)\left(\widehat{Y}_{g_2} - g_2\right)$ for $f_1, f_2, g_1, g_2 \in \mathcal{H}_B(\hat{v})$. By Lipschitz continuity of multiplication and boundedness of nuisances, we have

$$\|f - g\|_{L^2(P_n)} \lesssim \|f_1 - g_1\|_{n,S} + \|f_2 - g_2\|_{n,S} + \|f_2 - g_2\|_{n,S'}.$$

Taking the supremum over all discrete distributions Q, we obtain

$$\|f - g\|_{L^2(P_n)} \lesssim \sup_Q \|f_1 - g_1\|_{L^2(Q)} + \sup_Q \|f_2 - g_2\|_{L^2(Q)}.$$

Hence, by preservation of entropy integrals in van der Vaart & Wellner (1996),

$$\log N(\varepsilon, \widehat{\mathcal{G}}, L^2(P_n)) \lesssim \sup_Q \log N(\varepsilon, \mathcal{H}_B(\hat{v}), L^2(Q)).$$

Taking the supremum over discrete distributions Q on both sides yields the uniform covering number bound

$$\sup_Q \log N(\varepsilon, \widehat{\mathcal{G}}, L^2(Q)) \lesssim \sup_Q \log N(\varepsilon, \mathcal{H}_B(\hat{v}), L^2(Q)).$$

The class $\mathcal{H}_B(\hat{v})$ satisfies

$$\sup_Q \log N\big(\varepsilon, \mathcal{H}_B(\hat{v}), L^2(Q)\big) \le \sup_Q \log N\big(\varepsilon, \widetilde{\mathcal{F}}_B, L^2(Q \circ \hat{v}^{-1})\big),$$

where $Q \circ \hat{v}^{-1}$ denotes the pushforward of Q under $\hat{v}$. Hence,

$$\sup_Q \log N\big(\varepsilon, \mathcal{H}_B(\hat{v}), L^2(Q)\big) \le \sup_Q \log N\big(\varepsilon, \widetilde{\mathcal{F}}_B, L^2(Q)\big),$$

where, by a slight abuse of notation, the supremum on the right-hand side is taken over all discrete probability distributions Q on $\mathbb{R}$. The class $\widetilde{\mathcal{F}}_B$ consists of all piecewise-constant functions on $\mathbb{R}$ taking at most B values. Therefore, $\widetilde{\mathcal{F}}_B$ has VC–subgraph dimension $O(B)$, and van der Vaart & Wellner (1996) implies

$$\begin{aligned} \sup_Q \log N(\varepsilon, \mathcal{H}_B(\hat{v}), L^2(Q)) &\lesssim \sup_Q \log N\big(\varepsilon, \widetilde{\mathcal{F}}_B, L^2(Q)\big) \\ &\lesssim B \log(1/\varepsilon). \end{aligned}$$

Consequently, $\mathcal{J}(\delta, \mathcal{H}_B(\hat{v})) \lesssim \delta \sqrt{B \log(1/\delta)}$. □

I.2 Proof of Theorem 3

Proof of Theorem 3. Denote $O_i := (S_i, A_i, R_i, S_i')$. The first-order optimality conditions for $\theta_n^{(K)}$ imply that, for each bin $b \in [B]$,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{I_b}(\hat{v}(S_i))\{\widehat{Y}_{V_{K-1}}(O_i) - V_K(S_i)\} = 0.$$

Hence, for any $f \in \mathcal{G}_B$, which is a linear combination of these indicators,

$$\frac{1}{n}\sum_{i=1}^{n} f(\hat{v}(S_i))\{\widehat{Y}_{V_{K-1}}(O_i) - V_K(S_i)\} = 0. \tag{S2}$$

Moreover, for any function $g : \mathbb{R} \to \mathbb{R}$, we may write $g \circ V_K = f \circ \hat{v}$ for some f, and therefore, for all such g,

$$\frac{1}{n}\sum_{i=1}^{n} g(V_K(S_i))\{\widehat{Y}_{V_{K-1}}(O_i) - V_K(S_i)\} = 0.$$

Noting that $\Gamma_0(V_K) - V_K$ is of the form $g \circ V_K$ for some function g, we have

$$\begin{aligned} 0 = \frac{1}{n}\sum_{i=1}^{n} &\Big\{\Gamma_0(V_K)(S_i) - V_K(S_i)\Big\} \\ &\times \Big\{\widehat{Y}_{V_{K-1}}(O_i) - V_K(S_i)\Big\}. \end{aligned}$$

Adding and subtracting P_0 yields

$$\begin{aligned} &P_0\Big\{\Gamma_0(V_K) - V_K\Big\}\Big\{\widehat{Y}_{V_K} - V_K\Big\} \\ &= (P_0 - P_n)\Big\{\Gamma_0(V_K) - V_K\Big\}\Big\{\widehat{Y}_{V_{K-1}} - V_K\Big\} \\ &\quad + P_0\Big\{\Gamma_0(V_K) - V_K\Big\}\Big\{\widehat{Y}_{V_K} - \widehat{Y}_{V_{K-1}}\Big\}. \end{aligned} \tag{S3}$$

We rewrite the left-hand side of (S3). Adding and subtracting and applying the law of total expectation, the calibration error decomposes as

$$\begin{aligned} &\|\Gamma_0(V_K) - V_K\|^2 \\ &= P_0\Big\{\Gamma_0(V_K) - V_K\Big\}\Big\{\widehat{Y}_{V_K} - V_K\Big\} \\ &\quad + P_0\Big\{\Gamma_0(V_K) - V_K\Big\}\Big\{\Gamma_0(V_K) - \widehat{Y}_{V_K}\Big\}. \end{aligned} \tag{S4}$$

The first term on the right-hand side can be decomposed as in (S3). By the law of total expectation, the proof of Theorem 2, and the Cauchy–Schwarz inequality, the second term satisfies

$$\begin{aligned} &\Big|P_0\Big\{\Gamma_0(V_K) - V_K\Big\}\Big\{\widehat{Y}_{V_K} - \Gamma_0(V_K)\Big\}\Big| \\ &= \Big|P_0\Big\{\Gamma_0(V_K) - V_K\Big\}\Big\{\widehat{Y}_{V_K} - \mathcal{T}_\pi(V_K)\Big\}\Big| \\ &\leq \|\Gamma_0(V_K) - V_K\| \left\|b_0\{(\widehat{w}_\pi - w_\pi)(\widehat{q}_{V_K} - q_{V_K})\}\right\|_{2,\rho}. \end{aligned}$$

Hence, (S3) and (S4) together imply that

$$\begin{aligned} &\|\Gamma_0(V_K) - V_K\|^2 \\ &\leq (P_0 - P_n)\Big\{\Gamma_0(V_K) - V_K\Big\}\Big\{\widehat{Y}_{V_{K-1}} - V_K\Big\} \\ &\quad + \|\Gamma_0(V_K) - V_K\| \left\|\widehat{\mathcal{T}}_\pi V_K - \widehat{\mathcal{T}}_\pi V_{K-1}\right\| \\ &\quad + \|\Gamma_0(V_K) - V_K\| \left\|b_0\{(\widehat{w}_\pi - w_\pi)(\widehat{q}_{V_K} - q_{V_K})\}\right\|_{2,\rho}. \end{aligned} \tag{S5}$$

where $\widehat{\mathcal{T}}_\pi v(s) := \mathbb{E}[\widehat{Y}_v(O) \mid S = s]$. Here, the second term on the right-hand side follows from (S3), noting that

$$\begin{aligned}
&\left|P_0\Big\{\Gamma_0(V_K) - V_K\Big\}\Big\{\widehat{Y}_{V_K} - \widehat{Y}_{V_{K-1}}\Big\}\right| \\
&= \left|P_0\Big\{\Gamma_0(V_K) - V_K\Big\}\,\mathbb{E}\left[\widehat{Y}_{V_K} - \widehat{Y}_{V_{K-1}} \mid S\right]\right| \\
&\le \|\Gamma_0(V_K) - V_K\|\,\left\|\mathbb{E}\left[\widehat{Y}_{V_K} - \widehat{Y}_{V_{K-1}} \mid S\right]\right\| \\
&= \|\Gamma_0(V_K) - V_K\|\,\left\|\widehat{\mathcal{T}}_\pi V_K - \widehat{\mathcal{T}}_\pi V_{K-1}\right\|_{2,\rho}.
\end{aligned}$$

The empirical-process term on the right-hand side of (S5) is controlled as follows. The product

$$\big(\Gamma_0(V_K) - V_K\big)\big(\widehat{Y}_{V_{K-1}} - V_K\big)$$

lies in a uniformly bounded subset of the class

$$\widehat{\mathcal{G}} := \big\{(f_1 - f_2)\big(\widehat{Y}_{f_2} - f_2\big) : f_1, f_2 \in \mathcal{H}_B(\hat{v})\big\},$$

Lemma S7 gives $\mathcal{J}\left(\delta, \widehat{\mathcal{G}}\right) \lesssim \delta\sqrt{B\log(1/\delta)}$. By assumption, $\widehat{\mathcal{G}}$ is a fixed, nonrandom function class conditional on the training data, which is independent of the calibration sample $\mathcal{C}_n$. Hence, applying Lemma S6 conditional on the training data, we obtain

$$\begin{aligned}
\mathcal{R}_n(\widehat{\mathcal{G}}, \delta) &\lesssim \frac{1}{\sqrt{n}}\,\mathcal{J}(\delta, \widehat{\mathcal{G}})\left(1 + \frac{\mathcal{J}(\delta, \widehat{\mathcal{G}})}{\delta^2\sqrt{n}}\right) \\
&\lesssim \frac{1}{\sqrt{n}}\,\delta\sqrt{B\log(1/\delta)}\left(1 + \frac{\delta\sqrt{B\log(1/\delta)}}{\delta^2\sqrt{n}}\right) \\
&\lesssim \frac{\delta\sqrt{B\log(1/\delta)}}{\sqrt{n}} + \frac{B\log(1/\delta)}{n}.
\end{aligned}$$

The choice

$$\delta_n := \sqrt{\frac{B}{n}\,\log\left(\frac{n}{B}\right)},$$

satisfies

$$\mathcal{R}_n(\widehat{\mathcal{G}}, \delta_n) \lesssim \delta_n^2.$$

Thus a sufficiently large universal multiple of δ_n satisfies the critical-radius condition in Lemma S5. Because the subsequent bounds suppress universal constants, we continue to denote this multiple by δ_n.

Applying Lemma S5 conditional on the training data with $\mathcal{F} := \widehat{\mathcal{G}}$ shows that, with probability at least $1 - e^{-u^2}$, every $f \in \widehat{\mathcal{G}}$ satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\big(f(O_i) - \mathbb{E}[f(O_i)]\big) \lesssim \delta_n^2 + \delta_n\|f\|_{2,P_0} + \frac{u\|f\|_{2,P_0}}{\sqrt{n}} + \frac{u^2}{n}.$$

Choosing $u = \sqrt{\log(1/\eta)}$ gives $1 - e^{-u^2} = 1 - \eta$. Hence, with probability at least $1 - \eta$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\big(f(O_i) - \mathbb{E}[f(O_i)]\big) \\
&\lesssim \delta_n^2 + \delta_n\|f\|_{2,P_0} + \frac{\sqrt{\log(1/\eta)}\,\|f\|_{2,P_0}}{\sqrt{n}} + \frac{\log(1/\eta)}{n}.
\end{aligned}$$

Boundedness of the nuisance estimates gives

$$\left\|\{\Gamma_0(V_K) - V_K\}\{\widehat{Y}_{V_{K-1}} - V_K\}\right\| \lesssim \|\Gamma_0(V_K) - V_K\|_{2,\rho}.$$

Hence, with probability at least $1-\eta$,

$$\begin{aligned}(P_n - P_0)\Big\{\Gamma_0(V_K) - V_K\Big\}\Big\{\widehat{Y}_{V_{K-1}} - V_K\Big\} \\ \lesssim\ \delta_n^2\ +\ \delta_n\|\Gamma_0(V_K) - V_K\| \\ +\ \frac{\sqrt{\log(1/\eta)}\,\|\Gamma_0(V_K) - V_K\|}{\sqrt{n}}\ +\ \frac{\log(1/\eta)}{n},\end{aligned}$$

where the implicit constants do not depend on B.

Combining the above with (S5), we find with probability at least $1-\eta$,

$$\begin{aligned}&\|\Gamma_0(V_K) - V_K\|^2 \\ &\lesssim\ \delta_n^2\ +\ \delta_n\|\Gamma_0(V_K) - V_K\| \\ &\quad +\ \frac{\sqrt{\log(1/\eta)}}{\sqrt{n}}\,\|\Gamma_0(V_K) - V_K\|\ +\ \frac{\log(1/\eta)}{n} \\ &\quad +\ \|\Gamma_0(V_K) - V_K\|\,\big\|\widehat{\mathcal{T}}_\pi V_K - \widehat{\mathcal{T}}_\pi V_{K-1}\big\| \\ &\quad +\ \|\Gamma_0(V_K) - V_K\|\,\big\|b_0\{(\widehat{w}_\pi - w_\pi)\,(\widehat{q}_{V_K} - q_{V_K})\}\big\|_{2,\rho}.\end{aligned} \tag{S6}$$

The inequality in (S6) implies that, with probability at least $1-\eta$, the calibration error satisfies

$$\begin{aligned}\|\Gamma_0(V_K) - V_K\|\ \lesssim\ \delta_n\ +\ \sqrt{\frac{\log(1/\eta)}{n}} \\ +\ \big\|\widehat{\mathcal{T}}_\pi V_K - \widehat{\mathcal{T}}_\pi V_{K-1}\big\| \\ +\ \big\|b_0\{(\widehat{w}_\pi - w_\pi)\,(\widehat{q}_{V_K} - q_{V_K})\}\big\|_{2,\rho}.\end{aligned}$$

Recall that $\delta_n := \sqrt{\frac{B}{n}\,\log\left(\frac{n}{B}\right)}$. Then, with probability at least $1-\eta$,

$$\begin{aligned}\|\Gamma_0(V_K) - V_K\|\ \lesssim\ \sqrt{\frac{B}{n}\,\log\left(\frac{n}{B}\right)}\ +\ \sqrt{\frac{\log(1/\eta)}{n}} \\ +\ \big\|\widehat{\mathcal{T}}_\pi V_K - \widehat{\mathcal{T}}_\pi V_{K-1}\big\| \\ +\ \big\|b_0\{(\widehat{w}_\pi - w_\pi)\,(\widehat{q}_{V_K} - q_{V_K})\}\big\|_{2,\rho}.\end{aligned}$$

□

J Proof of Theorem 4

J.1 Additional notation

Define the bin-index map $\hat{b}(s)$ by setting $\hat{b}(s) = b$ whenever $\hat{v}(s) \in I_b$. Let Π_B denote the binning projection

$$(\Pi_B g)(s) := E\Big[g(S)\,\Big|\,\hat{v}(S) \in I_{\hat{b}(s)}\Big].$$

Denote

$$\mathcal{B}_{\pi,B}(g)(s) := (\Pi_B \mathcal{T}_\pi g)(s).$$

Define $V_B^\star$ as the fixed point satisfying

$$\mathcal{B}_{\pi,B}(V_B^\star) = V_B^\star.$$

Such a fixed point exists by Banach's fixed point theorem because $\mathcal{B}_{\pi,B}(g) \in \mathcal{H}_B(\hat{v})$ whenever $g \in \mathcal{H}_B(\hat{v})$, and the operator is a γ-contraction in the sup-norm.

Recall the projected transition operator

$$(P_{\pi,B}h)(s) = \mathbb{E}_\pi\left[h(S')\,\middle|\,\hat{v}(S) \in I_{\hat{b}(s)}\right].$$

It holds that

$$P_{\pi,B} \;:=\; \Pi_B P_\pi,$$

and

$$\Pi_B \mathcal{T}_\pi(f) - \Pi_B \mathcal{T}_\pi(g) = \gamma\, P_{\pi,B}(f-g).$$

J.2 Supporting lemmas

We begin with the deterministic ingredients for the histogram value bound.

Throughout, histogram bins mean the retained cells of the fitted partition: empty empirical cells are merged with a neighboring cell or discarded, and B denotes the number of retained cells. Population projections are defined on the corresponding non-null cells; ρ-null cells are identified in $L^2(\rho)$ and omitted from the index set.

Lemma S8 (Sup-norm contraction of the aggregated Bellman operator)**.** *For any bounded $f, g : \mathcal{S} \to \mathbb{R}$,*

$$\|\mathcal{B}_{\pi,B} f - \mathcal{B}_{\pi,B} g\|_\infty \le \gamma \|f-g\|_\infty.$$

Proof. Since the rewards cancel on differences,

$$\mathcal{B}_{\pi,B} f - \mathcal{B}_{\pi,B} g = \gamma P_{\pi,B}(f-g).$$

The operator $P_{\pi,B}$ is a Markov averaging operator. Hence, for any bounded h,

$$|P_{\pi,B} h(s)| \le E_\pi[|h(S')| \mid \hat{v}(S) \in I_{\hat{b}(s)}] \le \|h\|_\infty.$$

Taking $h = f - g$ proves the claim. □

Lemma S9 (Approximation error under aggregated concentrability)**.** *Assume C4. Then*

$$\|V_B^\star - v_0\|_{2,\rho} \le \frac{\sqrt{\mathcal{C}_{B,\gamma}}}{1-\gamma}\, \|(I - \Pi_B) v_0\|_{2,\rho}.$$

Proof. Because $v_0 = \mathcal{T}_\pi v_0$ and $V_B^\star = \Pi_B \mathcal{T}_\pi V_B^\star$,

$$V_B^\star - v_0 = \gamma P_{\pi,B}(V_B^\star - v_0) + (\Pi_B - I) v_0.$$

Let $a := (\Pi_B - I) v_0$. Iterating the display and using boundedness gives the resolvent expansion

$$V_B^\star - v_0 = \sum_{t=0}^{\infty} \gamma^t P_{\pi,B}^t a.$$

For any $t \ge 0$, set

$$c_t := \left\| \frac{d(\rho P_{\pi,B}^t)}{d\rho} \right\|_\infty.$$

Jensen's inequality for the Markov operator $P_{\pi,B}^t$ gives

$$\|P_{\pi,B}^t a\|_{2,\rho}^2 \le \int P_{\pi,B}^t(a^2)\, d\rho = \int a^2\, d(\rho P_{\pi,B}^t) \le c_t \|a\|_{2,\rho}^2.$$

Therefore,

$$\|V_B^\star - v_0\|_{2,\rho} \le \sum_{t=0}^{\infty} \gamma^t \sqrt{c_t}\, \|a\|_{2,\rho} = \frac{\sqrt{\mathcal{C}_{B,\gamma}}}{1-\gamma} \|(I - \Pi_B) v_0\|_{2,\rho}.$$

□

Lemma S10 (Inexact iterations under aggregated concentrability)**.** *Assume C4. Let $\{\eta_k\}_{k \ge 1}$ be any sequence such that*

$$\|V_k - \mathcal{B}_{\pi,B}(V_{k-1})\|_{2,\rho} \le \eta_k \qquad \text{for all } k.$$

For $m \ge 0$, set

$$c_m := \left\| \frac{d(\rho P_{\pi,B}^m)}{d\rho} \right\|_\infty.$$

Then, for any $K \geq 1$,

$$\|V_K - V_B^\star\|_{2,\rho} \leq \gamma^K \|\hat{v} - V_B^\star\|_\infty + \sum_{j=1}^{K} \gamma^{K-j} \sqrt{c_{K-j}}\, \eta_j.$$

Consequently,

$$\|V_K - V_B^\star\|_{2,\rho} \leq \gamma^K \|\hat{v} - V_B^\star\|_\infty + \frac{\sqrt{\mathcal{C}_{B,\gamma}}}{1-\gamma} \max_{1 \leq j \leq K} \eta_j.$$

Proof. Let

$$e_k := V_k - \mathcal{B}_{\pi,B}(V_{k-1}).$$

Since $V_B^\star = \mathcal{B}_{\pi,B} V_B^\star$,

$$V_k - V_B^\star = \gamma P_{\pi,B}(V_{k-1} - V_B^\star) + e_k.$$

Iterating this recursion yields

$$V_K - V_B^\star = \gamma^K P_{\pi,B}^K (\hat{v} - V_B^\star) + \sum_{j=1}^{K} \gamma^{K-j} P_{\pi,B}^{K-j} e_j.$$

The first term is bounded by $\gamma^K \|\hat{v} - V_B^\star\|_\infty$ because $P_{\pi,B}$ is a Markov operator. For the remaining terms, the same Jensen and change-of-measure argument used in Lemma S9 gives

$$\|P_{\pi,B}^{K-j} e_j\|_{2,\rho} \leq \sqrt{c_{K-j}}\, \|e_j\|_{2,\rho} \leq \sqrt{c_{K-j}}\, \eta_j.$$

The displayed bounds follow by the triangle inequality and the definition of $\mathcal{C}_{B,\gamma}$. □

Lemma S11 (Error bound for inexact updates)**.** *There exists $C < \infty$ such that, with probability at least $1 - \delta$, uniformly over $k = 1, \ldots, K$,*

$$\begin{aligned} &\|V_k - \mathcal{B}_{\pi,B}(V_{k-1})\|_{2,\rho} \\ &\quad \leq C \left\{ \sqrt{\frac{B}{n} \log\left(\frac{n}{B}\right)} + \sqrt{\frac{\log(K/\delta)}{n}} + \left\| b_0\{(\widehat{w}_\pi - w_\pi)\,(\widehat{q}_{V_{k-1}} - q_{V_{k-1}})\} \right\|_{2,\rho} \right\} \end{aligned}$$

Proof. Equation S2 in the proof of Theorem 3 shows that, for any transformation $f \in \mathcal{G}_B$ and any k,

$$\frac{1}{n} \sum_{i=1}^{n} f(\hat{v}(S_i)) \{ \widehat{Y}_{V_{k-1}}(O_i) - V_k(S_i) \} = 0.$$

Applying the above with $f \circ \hat{v} = V_B^{\star(k)} - V_k$, where

$$V_B^{\star(k)} := \mathcal{B}_{\pi,B}(V_{k-1}),$$

we obtain

$$\frac{1}{n} \sum_{i=1}^{n} (V_B^{\star(k)} - V_k)(S_i) \left\{ \widehat{Y}_{V_{k-1}}(O_i) - V_k(S_i) \right\} = 0.$$

Adding and subtracting P_0, we find

$$\begin{aligned} &P_0 \left[(V_B^{\star(k)} - V_k) \left\{ \widehat{Y}_{V_{k-1}} - V_k \right\} \right] \\ &= (P_0 - P_n) \left[(V_B^{\star(k)} - V_k) \left\{ \widehat{Y}_{V_{k-1}} - V_k \right\} \right]. \end{aligned} \tag{S7}$$

First, we study the left-hand side of (S7). We have

$$\begin{aligned} &P_0 \left[(V_B^{\star(k)} - V_k) \left\{ \widehat{Y}_{V_{k-1}} - V_k \right\} \right] \\ &= \|V_B^{\star(k)} - V_k\|_{2,\rho}^2 \\ &\quad + P_0 \left[(V_B^{\star(k)} - V_k) \left\{ \widehat{Y}_{V_{k-1}} - V_B^{\star(k)} \right\} \right]. \end{aligned}$$

Since $V_B^{\star(k)} = \Pi_B \mathcal{T}_\pi V_{k-1}$ and $V_B^{\star(k)} - V_k \in \mathcal{H}_B(\hat{v})$, the projection identity gives

$$P_0\left[(V_B^{\star(k)} - V_k)\{V_B^{\star(k)} - \mathcal{T}_\pi V_{k-1}\}\right] = 0.$$

Thus

$$\begin{aligned} &P_0\left[(V_B^{\star(k)} - V_k)\left\{\widehat{Y}_{V_{k-1}} - V_B^{\star(k)}\right\}\right] \\ &= P_0\left[(V_B^{\star(k)} - V_k)\left\{\widehat{Y}_{V_{k-1}} - \mathcal{T}_\pi V_{k-1}\right\}\right]. \end{aligned}$$

By Theorem 2 and arguing as in the proof of Theorem 3, it follows that

$$\begin{aligned} &\left|P_0\left[(V_B^{\star(k)} - V_k)\left\{\widehat{Y}_{V_{k-1}} - V_B^{\star(k)}\right\}\right]\right| \\ &\leq \|V_B^{\star(k)} - V_k\|_{2,\rho} \left\|b_0\{(\widehat{w}_\pi - w_\pi)\,(\widehat{q}_{V_{k-1}} - q_{V_{k-1}})\}\right\|_{2,\rho}. \end{aligned}$$

Putting it all together,

$$\begin{aligned} \|V_B^{\star(k)} - V_k\|_{2,\rho}^2 &\leq P_0\left[(V_B^{\star(k)} - V_k)\left\{\widehat{Y}_{V_{k-1}} - V_k\right\}\right] \\ &+ \|V_B^{\star(k)} - V_k\|_{2,\rho} \left\|b_0\{(\widehat{w}_\pi - w_\pi)\,(\widehat{q}_{V_{k-1}} - q_{V_{k-1}})\}\right\|_{2,\rho}. \end{aligned}$$

Hence, by (S7),

$$\begin{aligned} &\|V_B^{\star(k)} - V_k\|_{2,\rho}^2 \\ &\quad\leq (P_0 - P_n)\left[(V_B^{\star(k)} - V_k)\left\{\widehat{Y}_{V_{k-1}} - V_k\right\}\right] \\ &\quad+ \|V_B^{\star(k)} - V_k\|_{2,\rho} \left\|b_0\{(\widehat{w}_\pi - w_\pi)\,(\widehat{q}_{V_{k-1}} - q_{V_{k-1}})\}\right\|_{2,\rho}. \end{aligned} \tag{S8}$$

Next, we obtain a high-probability bound for the first term on the right-hand side of (S8). Applying Lemmas S5, S6, and S7, and arguing as in the proof of Theorem 3, we obtain that, with probability at least $1-\delta$,

$$\begin{aligned} &\left|(P_0 - P_n)\left[(V_B^{\star(k)} - V_k)\{\widehat{Y}_{V_{k-1}} - V_k\}\right]\right| \\ &\qquad\lesssim \delta_n^2 + \delta_n \|V_B^{\star(k)} - V_k\|_{2,\rho} \\ &\qquad\quad+ \frac{\sqrt{\log(1/\delta)}\,\|V_B^{\star(k)} - V_k\|_{2,\rho}}{\sqrt{n}} + \frac{\log(1/\delta)}{n}, \end{aligned}$$

where $\delta_n := \sqrt{(B/n)\,\log(n/B)}$, and the implicit constants do not depend on B.

Plugging this high probability bound into (S8), we obtain that, with probability at least $1-\delta$,

$$\begin{aligned} &\|V_B^{\star(k)} - V_k\|_{2,\rho}^2 \\ &\lesssim \delta_n^2 + \delta_n \|V_B^{\star(k)} - V_k\|_{2,\rho} \\ &\quad+ \frac{\sqrt{\log(1/\delta)}\,\|V_B^{\star(k)} - V_k\|_{2,\rho}}{\sqrt{n}} + \frac{\log(1/\delta)}{n} \\ &\quad+ \|V_B^{\star(k)} - V_k\|_{2,\rho} \left\|b_0\{(\widehat{w}_\pi - w_\pi)\,(\widehat{q}_{V_{k-1}} - q_{V_{k-1}})\}\right\|_{2,\rho}. \end{aligned}$$

Recalling that $\delta_n := \sqrt{\frac{B}{n}\,\log\left(\frac{n}{B}\right)}$, the inequality above implies that, with probability at least $1-\delta$,

$$\begin{aligned} \|V_B^{\star(k)} - V_k\|_{2,\rho} &\lesssim \sqrt{\frac{B}{n}\,\log\left(\frac{n}{B}\right)} + \sqrt{\frac{\log(1/\delta)}{n}} \\ &\quad+ \left\|b_0\{(\widehat{w}_\pi - w_\pi)\,(\widehat{q}_{V_{k-1}} - q_{V_{k-1}})\}\right\|_{2,\rho}. \end{aligned}$$

Taking the failure probability to be δ/K and applying a union bound over $k = 1, \ldots, K$ proves the result. □

J.3 Proof of Theorem 4

Proof of Theorem 4. For $C < \infty$ large enough, define

$$\eta_k = C\Bigg\{\sqrt{\frac{B}{n}\log\Big(\frac{n}{B}\Big)} + \sqrt{\frac{\log(K/\delta)}{n}} + \big\|b_0\{(\widehat{w}_\pi - w_\pi)\,(\widehat{q}_{V_{k-1}} - q_{V_{k-1}})\}\big\|_{2,\rho}\Bigg\}$$

By Lemma S11, with probability at least $1-\delta$, uniformly over $k = 1, \dots, K$,

$$\|V_k - \mathcal{B}_{\pi,B}(V_{k-1})\|_{2,\rho} \le \eta_k.$$

Lemma S10 therefore gives

$$\|V_K - V_B^\star\|_{2,\rho} \le \gamma^K \|\hat{v} - V_B^\star\|_\infty + \frac{\sqrt{\mathcal{C}_{B,\gamma}}}{1-\gamma} \max_{1\le j\le K} \eta_j.$$

Combining this with Lemma S9 and the triangle inequality yields

$$\|V_K - v_0\|_{2,\rho} \le \gamma^K \|\hat{v} - V_B^\star\|_\infty + \frac{\sqrt{\mathcal{C}_{B,\gamma}}}{1-\gamma}\left[\|(I-\Pi_B)v_0\|_{2,\rho} + \max_{1\le j\le K}\eta_j\right].$$

Since Π_B is the $L^2(\rho)$ projection onto $\mathcal{H}_B(\hat{v})$,

$$\|(I-\Pi_B)v_0\|_{2,\rho} = \min_{h\in\mathcal{H}_B(\hat{v})} \|h - v_0\|_{2,\rho}.$$

Substituting the definition of η_k and absorbing constants completes the proof. □

J.4 Bound on Successive Iteration Errors

Lemma S12 (Bound on Successive Iteration Errors)**.** *Under the conditions of Theorem 4,*

$$\begin{aligned}&\|V_K - V_{K-1}\|_{2,\rho}\\ &\lesssim \gamma^{K-1}\|\hat{v} - V_B^\star\|_\infty\\ &\quad + \frac{\sqrt{\mathcal{C}_{B,\gamma}}}{1-\gamma} C\Bigg(\sqrt{\frac{B}{n}\log\Big(\frac{n}{B}\Big)} + \sqrt{\frac{\log(K/\delta)}{n}}\\ &\qquad + \max_{1\le j\le K}\big\|b_0\{(\widehat{w}_\pi - w_\pi)\,(\widehat{q}_{V_{j-1}} - q_{V_{j-1}})\}\big\|_{2,\rho}\Bigg).\end{aligned}$$

Proof. Lemma S10 and Lemma S11 give the same bound for $\|V_m - V_B^\star\|_{2,\rho}$, for $m = K-1$ and $m = K$, with $\gamma^m\|\hat{v} - V_B^\star\|_\infty$ as the finite-iteration term. By the triangle inequality,

$$\|V_K - V_{K-1}\|_{2,\rho} \le \|V_{K-1} - V_B^\star\|_{2,\rho} + \|V_K - V_B^\star\|_{2,\rho}.$$

Substituting the two bounds and absorbing constants proves the claim. □

K Proofs of Theorems 5 and S1

Proof of Theorem 5. The first-order optimality conditions for isotonic regression (equivalently, its interpretation as a histogram estimator) imply that for any function $f : \mathbb{R} \to \mathbb{R}$ that is a linear combination of the indicator functions defining the isotonic partition,

$$\frac{1}{n}\sum_{i=1}^n f\big(V_K(S_i)\big)\Big\{\widehat{Y}_{V_{K-1}}(O_i) - V_K(S_i)\Big\} = 0.$$

See, for example, the proof of Lemma C.1 in van der Laan et al. (2023) and van der Laan et al. (2024). Choosing f appropriately yields

$$0 = \frac{1}{n}\sum_{i=1}^{n}\Big\{\Gamma_0(V_K)(S_i) - V_K(S_i)\Big\} \times \Big\{\widehat{Y}_{V_{K-1}}(O_i) - V_K(S_i)\Big\},$$

which is the same basic equality as (S3) in the proof of Theorem 3. The remainder of the argument proceeds along the same lines with minor modifications.

Specifically, let $\mathcal{F}_{TV}$ denote the union of $\mathcal{G}_{\text{iso}}$, which is uniformly bounded by $2M$ under Condition C1, with all functions of bounded total variation bounded by the constant C in Condition C5. By van der Vaart & Wellner (1996), this class satisfies the uniform entropy integral bound $\mathcal{J}(\delta, \mathcal{F}_{TV}) \lesssim \sqrt{\delta}$.

Let

$$m_K(t) := E[(\mathcal{T}_\pi V_K)(S) \mid \hat{v}(S) = t, \mathcal{C}_n].$$

Condition C5 states that m_K has bounded total variation. Since

$$\Gamma_0(V_K)(S) = E\{m_K(\hat{v}(S)) \mid V_K(S)\},$$

and V_K is a nondecreasing function of $\hat{v}$, Lemma 6 of van der Laan et al. (2024) implies that the coarsened curve $\Gamma_0(V_K)$ also has bounded total variation and lies in $\mathcal{F}_{TV,\hat{v}} := \{f \circ \hat{v} : f \in \mathcal{F}_{TV}\}$. Thus, it holds that

$$\big(\Gamma_0(V_K) - V_K\big)\big(\widehat{Y}_{V_{K-1}} - V_K\big)$$

lies in a uniformly bounded subset of the class

$$\widehat{\mathcal{G}} := \Big\{(f_1 - f_2)\big(\widehat{Y}_{f_2} - f_2\big) : \ f_1, f_2 \in \mathcal{F}_{TV,\hat{v}}\Big\}.$$

Arguing as in the proof of Theorem 3, we have

$$\mathcal{J}(\delta, \widehat{\mathcal{G}}) \lesssim \mathcal{J}(\delta, \mathcal{F}_{TV}) \lesssim \sqrt{\delta}.$$

The proof now follows directly from Theorem 3 with this new choice of $\widehat{\mathcal{G}}$ and its associated critical radius $\delta_n = n^{-1/3}$ for monotone functions. □

Proof of Theorem S1. On the event $\{\widehat{B} \le B_n\}$, replace $\mathcal{G}_B$ in the proofs of Theorems 3 and 4 by the class of all piecewise-constant functions with at most B_n constant segments. The uniform entropy bound in Lemma S7 therefore applies to the selected partition, so the stochastic terms in both results hold with $B = B_n$. The deterministic approximation and concentrability terms are those of the selected fixed partition.

Algorithm S1 initializes the fixed-partition iteration at $V_0 = g_{\text{iso}} \circ \hat{v}$. Hence the contraction argument in the proof of Theorem 4 gives $\gamma^K \|g_{\text{iso}} \circ \hat{v} - V^\star_{\widehat{B}}\|_\infty$ in place of $\gamma^K \|\hat{v} - V^\star_B\|_\infty$. A union bound with the event $\{\widehat{B} \le B_n\}$ yields probability at least $1 - 2\delta$, proving the stated result. □

L Bellman Calibration for Optimal Control

Optimal control replaces the target-policy Bellman operator by the nonlinear Bellman optimality operator. We first define Bellman calibration for this operator, then derive a calibration–refinement bound, and finally analyze the binned projected fixed point. The proofs are collected in the final subsection.

Throughout this section, condition on the training sample used to construct the initial predictor $\widehat{Q}$. Assume that $\mathcal{A}$ is finite and that r_0 and $\widehat{Q}$ are bounded. Let $X = (S, A) \sim \nu$, where $\nu(ds, da) = \rho(ds)b_0(da \mid s)$, and define

$$\|Q\|_{2,\nu} := \big\{\mathbb{E}_\nu[Q(X)^2]\big\}^{1/2}.$$

For a stochastic policy η, write

$$(P_\eta Q)(s,a) := \mathbb{E}\left[\sum_{a' \in \mathcal{A}} \eta(a' \mid S')Q(S',a') \,\middle|\, S=s, A=a\right].$$

The Bellman optimality operator is

$$(\mathcal{T}^\star Q)(s,a) := r_0(s,a) + \gamma \mathbb{E}\left[\max_{a' \in \mathcal{A}} Q(S',a') \,\middle|\, S=s, A=a\right],$$

and the optimal Q-function $Q^\star$ is its unique bounded fixed point:

$$Q^\star = \mathcal{T}^\star Q^\star.$$

L.1 Bellman Calibration for Control

Bellman calibration for control requires the optimality Bellman residual to vanish after conditioning on the scalar prediction $Q(X)$. For a candidate Q, define the calibration map

$$\Gamma_0^\star(Q)(x) := \mathbb{E}\left[R + \gamma \max_{a' \in \mathcal{A}} Q(S',a') \,\middle|\, Q(X) = Q(x)\right]. \tag{S9}$$

We say that Q is *Bellman calibrated for control* if

$$Q(X) = \Gamma_0^\star(Q)(X) \qquad \text{almost surely,}$$

or equivalently,

$$\mathbb{E}\left[R + \gamma \max_{a' \in \mathcal{A}} Q(S',a') - Q(X) \,\middle|\, Q(X)\right] = 0 \qquad \text{almost surely.}$$

The optimal Q-function is Bellman calibrated. Indeed, by the Bellman optimality equation and iterated expectations,

$$Q^\star(X) = \mathbb{E}\left[R + \gamma \max_{a'} Q^\star(S',a') \,\middle|\, Q^\star(X)\right] = \Gamma_0^\star(Q^\star)(X).$$

Thus, as in policy evaluation, Bellman calibration is a necessary but weak optimality condition: it requires the Bellman optimality residual to vanish only after conditioning on the scalar prediction.

For a candidate Q, define the ℓ^2 optimality Bellman calibration error

$$\mathrm{Cal}^2_{\ell^2,\star}(Q) := \|Q - \Gamma_0^\star(Q)\|^2_{2,\nu} = \mathbb{E}_\nu\left[\left\{\mathbb{E}\left[R + \gamma \max_{a'} Q(S',a') - Q(X) \,\middle|\, Q(X)\right]\right\}^2\right]. \tag{S10}$$

This criterion is zero exactly when Q is Bellman calibrated for control.

L.2 A Calibration–Refinement Bound for Control

The Q-functions generated by the original prediction $\widehat{Q}(X)$ form the class

$$\mathcal{Q}_{\widehat{Q}} := \left\{g \circ \widehat{Q} : g : \mathbb{R} \to \mathbb{R} \text{ measurable}\right\}, \qquad \Pi_{\widehat{Q}} := L^2(\nu)\text{-projection onto } \mathcal{Q}_{\widehat{Q}}.$$

The corresponding refinement error is

$$\mathrm{Ref}_{\ell^2,\star}(\widehat{Q}) := \min_{\theta:\mathbb{R}\to\mathbb{R}} \|\theta \circ \widehat{Q} - Q^\star\|_{2,\nu} = \|\Pi_{\widehat{Q}} Q^\star - Q^\star\|_{2,\nu}. \tag{S11}$$

It measures the optimal-value information lost by reducing $X = (S,A)$ to the scalar prediction $\widehat{Q}(X)$.

The next lemma gives the key pairwise reduction. The interpolating policy depends on (Q_1, Q_2), so $\mathcal{T}^\star$ remains nonlinear. For the fixed pair $(\widehat{Q}, Q^\star)$, however, the lemma yields the linear error recursion needed below.

Lemma S13 (Interpolating-policy representation)**.** *For any bounded Q_1, Q_2, there exists a stochastic policy η_{Q_1,Q_2} such that*

$$\mathcal{T}^\star Q_1 - \mathcal{T}^\star Q_2 = \gamma P_{\eta_{Q_1,Q_2}}(Q_1 - Q_2). \tag{S12}$$

For a stochastic policy η, define the projected transition

$$P_{\eta,\widehat{Q}} := \Pi_{\widehat{Q}} P_\eta.$$

Thus,

$$P_{\eta,\widehat{Q}} f(x) = \mathbb{E}_\eta\Big[f(S', A') \,\Big|\, \widehat{Q}(X) = \widehat{Q}(x)\Big], \qquad A' \sim \eta(\cdot \mid S').$$

Define the control projected concentrability coefficient

$$\mathcal{C}^\star_{\widehat{Q},\gamma} := \sup_\eta \left\{ (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \left\| \frac{d(\nu P^t_{\eta,\widehat{Q}})}{d\nu} \right\|_\infty^{1/2} \right\}^2, \tag{S13}$$

where the supremum is over stochastic policies on $\mathcal{A}$.

AC1 Projected concentrability for control. $\mathcal{C}^\star_{\widehat{Q},\gamma} < \infty$.

Because the interpolating policy depends on the pair of Q-functions, Condition AC1 is uniform over policies. The proof uses only the policy associated with $(\widehat{Q}, Q^\star)$.

Theorem S14 (Calibration–refinement bound for control)**.** *Under Condition AC1,*

$$\begin{aligned} \|\widehat{Q} - Q^\star\|_{2,\nu} &\le \frac{\sqrt{\mathcal{C}^\star_{\widehat{Q},\gamma}}}{1-\gamma} \left\{ \mathrm{Cal}^2_{\ell^2,\star}(\widehat{Q}) + \mathrm{Ref}^2_{\ell^2,\star}(\widehat{Q}) \right\}^{1/2} \\ &\le \frac{\sqrt{\mathcal{C}^\star_{\widehat{Q},\gamma}}}{1-\gamma} \left\{ \mathrm{Cal}_{\ell^2,\star}(\widehat{Q}) + \mathrm{Ref}_{\ell^2,\star}(\widehat{Q}) \right\}. \end{aligned} \tag{S14}$$

The bound separates two sources of error. The calibration term measures the optimality Bellman residual within levels of $\widehat{Q}$; the refinement term measures variation in $Q^\star(X)$ not captured by $\widehat{Q}(X)$. Their squared sum reflects their $L^2(\nu)$-orthogonality before dynamic error propagation. Calibration can therefore correct systematic errors along the scale of $\widehat{Q}$, but it cannot recover discarded state–action information.

L.3 Binned Calibration and the Projected Optimality Fixed Point

To connect the preceding bound to histogram calibration, let $I_1, \dots, I_B$ be fixed bins of the original prediction $\widehat{Q}(X)$, and define

$$\mathcal{G}_B := \mathrm{span}\{\mathbf{1}_{I_b} : b \in [B]\}, \qquad \mathcal{H}_B(\widehat{Q}) := \{g \circ \widehat{Q} : g \in \mathcal{G}_B\}.$$

Let $\widehat{b}(x) = b$ whenever $\widehat{Q}(x) \in I_b$, and let Π_B denote the $L^2(\nu)$-projection onto $\mathcal{H}_B(\widehat{Q})$:

$$(\Pi_B f)(x) = \mathbb{E}_\nu\Big[f(X) \,\Big|\, \widehat{Q}(X) \in I_{\widehat{b}(x)}\Big].$$

Define the binned projected optimality operator

$$\mathcal{B}^\star_B := \Pi_B \mathcal{T}^\star.$$

By construction,

$$\mathcal{B}^\star_B f \in \mathcal{H}_B(\widehat{Q}) \qquad \text{for every bounded } f.$$

Thus the projected Bellman update is closed on the binned class. Since Π_B is a conditional-expectation operator,

$$\|\mathcal{B}^\star_B f - \mathcal{B}^\star_B g\|_\infty \le \gamma \|f - g\|_\infty.$$

Hence $\mathcal{B}^\star_B$ has a unique bounded fixed point, denoted $Q^\star_B$:

$$Q^\star_B = \mathcal{B}^\star_B Q^\star_B = \Pi_B \mathcal{T}^\star Q^\star_B. \tag{S15}$$

The fixed point satisfies the binwise optimality Bellman equation

$$Q_B^\star(x) = \mathbb{E}\Big[R + \gamma \max_{a'} Q_B^\star(S', a') \,\Big|\, \widehat{Q}(X) \in I_{\widehat{b}(x)}\Big].$$

Because $Q_B^\star(X)$ is itself measurable with respect to the bin index, iterated expectations imply

$$Q_B^\star(X) = \mathbb{E}\Big[R + \gamma \max_{a'} Q_B^\star(S', a') \,\Big|\, Q_B^\star(X)\Big].$$

Thus $Q_B^\star$ is Bellman calibrated for control.

For a stochastic policy η, define

$$P_{\eta,B} := \Pi_B P_\eta,$$

and the aggregated control concentrability coefficient

$$\mathcal{C}_{B,\gamma}^\star := \sup_\eta \left\{ (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \left\| \frac{d(\nu P_{\eta,B}^t)}{d\nu} \right\|_\infty^{1/2} \right\}^2. \tag{S16}$$

Theorem S15 (Approximation error of the binned control fixed point)**.** *Suppose* $\mathcal{C}_{B,\gamma}^\star < \infty$*. Then*

$$\|Q_B^\star - Q^\star\|_{2,\nu} \le \frac{\sqrt{\mathcal{C}_{B,\gamma}^\star}}{1-\gamma} \min_{h \in \mathcal{H}_B(\widehat{Q})} \|h - Q^\star\|_{2,\nu}. \tag{S17}$$

The approximation bound contains only the best binned approximation to $Q^\star$. It has no Bellman-completeness term because $\Pi_B \mathcal{T}^\star$ maps every bounded function into $\mathcal{H}_B(\widehat{Q})$.

The theorem supplies the deterministic approximation term. A finite-sample result would additionally bound the one-step regression error for

$$R + \gamma \max_{a'} Q_k(S', a'),$$

and its propagation around the binned fixed point $Q_B^\star$.

L.4 Proofs for the Optimal-Control Extension

Proof of Lemma S13. Fix s and write

$$d(s) := \max_a Q_1(s,a) - \max_a Q_2(s,a), \qquad \Delta(s,a) := Q_1(s,a) - Q_2(s,a).$$

Because

$$\min_a \Delta(s,a) \le d(s) \le \max_a \Delta(s,a),$$

there exists a probability vector $\eta_{Q_1,Q_2}(\cdot \mid s)$ such that

$$d(s) = \sum_a \eta_{Q_1,Q_2}(a \mid s) \Delta(s,a).$$

Since the action space is finite, the weights may be chosen measurably in s. Taking the conditional expectation over S' gives (S12). □

Proof of Theorem S14. Let

$$\Delta := \widehat{Q} - Q^\star, \qquad c := \widehat{Q} - \Gamma_0^\star(\widehat{Q}), \qquad r := \Pi_{\widehat{Q}} Q^\star - Q^\star.$$

By iterated expectations,

$$\Gamma_0^\star(\widehat{Q}) = \Pi_{\widehat{Q}} \mathcal{T}^\star \widehat{Q}.$$

Using $Q^\star = \mathcal{T}^\star Q^\star$,

$$\Delta = \widehat{Q} - \Pi_{\widehat{Q}} \mathcal{T}^\star \widehat{Q} + \Pi_{\widehat{Q}} \left(\mathcal{T}^\star \widehat{Q} - \mathcal{T}^\star Q^\star \right) + \Pi_{\widehat{Q}} Q^\star - Q^\star.$$

By Lemma S13, there exists a stochastic policy η such that

$$\Delta = c + \gamma P_{\eta,\widehat{Q}} \Delta + r. \tag{S18}$$

Iterating (S18) gives

$$\Delta = \sum_{t=0}^{\infty} \gamma^t P_{\eta,\widehat{Q}}^t (c + r).$$

For

$$c_t(\eta) := \left\| \frac{d(\nu P_{\eta,\widehat{Q}}^t)}{d\nu} \right\|_\infty,$$

Jensen's inequality gives

$$\|P_{\eta,\widehat{Q}}^t (c + r)\|_{2,\nu} \le \sqrt{c_t(\eta)} \|c + r\|_{2,\nu}.$$

Hence Condition AC1 implies

$$\|\Delta\|_{2,\nu} \le \frac{\sqrt{\mathcal{C}^\star_{\widehat{Q},\gamma}}}{1-\gamma} \|c + r\|_{2,\nu}.$$

Now $c \in \mathcal{Q}_{\widehat{Q}}$, while $r = (\Pi_{\widehat{Q}} - I)Q^\star$ is orthogonal to $\mathcal{Q}_{\widehat{Q}}$. Therefore,

$$\|c + r\|_{2,\nu}^2 = \|c\|_{2,\nu}^2 + \|r\|_{2,\nu}^2 = \mathrm{Cal}^2_{\ell^2,\star}(\widehat{Q}) + \mathrm{Ref}^2_{\ell^2,\star}(\widehat{Q}).$$

This proves the first inequality; the second follows from $\sqrt{a^2 + b^2} \le a + b$. □

Proof of Theorem S15. Let

$$\Delta_B := Q_B^\star - Q^\star, \qquad r_B := \Pi_B Q^\star - Q^\star.$$

Using

$$Q_B^\star = \Pi_B \mathcal{T}^\star Q_B^\star, \qquad Q^\star = \mathcal{T}^\star Q^\star,$$

we obtain

$$\Delta_B = \Pi_B \left(\mathcal{T}^\star Q_B^\star - \mathcal{T}^\star Q^\star \right) + r_B.$$

By Lemma S13, there exists a stochastic policy η such that

$$\Delta_B = \gamma P_{\eta,B} \Delta_B + r_B.$$

Iterating gives

$$\Delta_B = \sum_{t=0}^{\infty} \gamma^t P_{\eta,B}^t r_B.$$

The same Jensen and concentrability argument used in Theorem S14 therefore yields

$$\|\Delta_B\|_{2,\nu} \le \frac{\sqrt{\mathcal{C}^\star_{B,\gamma}}}{1-\gamma} \|r_B\|_{2,\nu}.$$

Since Π_B is the $L^2(\nu)$-projection onto $\mathcal{H}_B(\widehat{Q})$,

$$\|r_B\|_{2,\nu} = \min_{h \in \mathcal{H}_B(\widehat{Q})} \|h - Q^\star\|_{2,\nu},$$

which proves (S17). □