MODEL SELECTION FOR OFF-POLICY EVALUATION VIA PAIRWISE BELLMAN VALIDATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Model selection is a central challenge in off-policy evaluation (OPE) because fitted Q-functions and occupancy ratios lack supervised validation losses. Standard approaches based on Bellman-residual estimation or minimax criteria often require auxiliary regressions or critic classes that must themselves be learned and tuned. We introduce *pairwise Bellman validation* (PBV), a tuning-free selection procedure among a finite collection of fitted Q-functions and occupancy-ratio estimators that uses only held-out transitions and requires no auxiliary validation models. The construction is motivated by an oracle critic given by a candidate's estimation error, which, when used to test the Bellman residual, controls the error of interest. Because this oracle critic is unknown, we approximate it using pairwise differences between candidates. Under misspecification, these pairwise approximations are imperfect, so we add penalties that account for the resulting critic-approximation error. For Q-function selection, we introduce target-aligned PBV, which uses marginalized importance weighting to evaluate the pairwise tests under the discounted target-occupancy distribution. When occupancy-ratio estimation is undesirable, we propose an unweighted variant closely related to LSTD-Tournament, but it requires an additional conditioning assumption linking offline Bellman tests to Q-function error. For occupancy-ratio selection, we introduce adjoint PBV, which applies the same pairwise construction to the adjoint Bellman equation. For all three procedures, we establish population and finite-sample oracle inequalities relative to the best candidates in possibly misspecified classes.

1 INTRODUCTION

Off-policy evaluation (OPE) estimates the value of a target policy using data collected under a different policy. In discounted infinite-horizon Markov decision processes (MDPs), two central nuisance functions are the target-policy action-value function, or Q-function, and the discounted occupancy ratio (Uehara et al., 2026). These functions underlie direct, importance-weighted, and doubly robust OPE, as well as semiparametrically efficient inference (Jiang & Li, 2016; Xie et al., 2019; Kallus & Uehara, 2020; 2022; van der Laan et al., 2025). They are characterized, respectively, by the Bellman and adjoint Bellman equations (Uehara et al., 2021). Existing methods for learning Q-functions and occupancy ratios exploit this fixed-point structure through fitted iteration and minimax methods that test Bellman residuals against a class of critic functions (Munos & Szepesvári, 2008; Nachum et al., 2019; Uehara et al., 2020; 2021; van der Laan & Kallus, 2026).

Model selection is a fundamental practical challenge in OPE. Choices of function class, regularization, optimization, and hyperparameters often produce finite classes of candidate estimates, yet logged data provide no supervised validation loss for either nuisance. For the Q-function, squared temporal-difference error is a seemingly natural criterion but is generally invalid for comparing candidates under stochastic transitions because its noise variance depends on the candidate (Baird III, 1995). Existing approaches therefore typically rely either on direct estimation of the Bellman residual or on testing whether its moments vanish against a class of critic functions. Both approaches introduce additional estimation and tuning problems. Bellman-residual validation requires fitting and tuning a separate regression for each candidate (Zitovsky et al., 2023). Minimax moment methods instead test a candidate's Bellman residual against a class of critic functions and use the worst-case moment to measure violation of the Bellman equation (Uehara et al., 2021; Miyaguchi, 2026). The critic class must be specified, while fitting the critic typically introduces an additional saddle-point

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.

optimization problem. A recent alternative is LSTD-Tournament (Liu et al., 2025), which compares candidates pairwise using LSTDQ and aggregates the resulting comparisons into a tournament score. This avoids a separately specified critic class, but its theoretical guarantees require the class to contain the true Q-function as well as matrix-conditioning assumptions linking the offline distribution to the target-policy dynamics.

In this work, we develop a tuning-free minimax model-selection procedure for finite candidate classes using only held-out transitions. Rather than learning or tuning a separate critic class, we construct the critics directly from the candidate class and evaluate each candidate by its worst-case empirical Bellman moment over these critics. Validation therefore reduces to a finite collection of Bellman moment tests, with no auxiliary validation models. Whereas LSTD-Tournament derives its pairwise comparisons from two-candidate LSTDQ normal equations, our procedure derives them from minimax Bellman-moment testing with candidate-induced critics.

The construction is motivated by an oracle critic that would be available if the target function were known. For Q-function selection, the oracle critic for candidate Q_i is its estimation error $Q_i - Q^\pi$, and testing the Bellman residual against this critic directly controls Q-function error. For occupancy-ratio selection, the corresponding oracle critic for candidate w_i is $\log(w_i/w^\pi)$. Contraction properties of the Bellman and adjoint Bellman operators relate these oracle tests to squared L^2 and generalized-KL error, respectively. Because the oracle critics depend on the unknown target, we replace them with pairwise differences between candidates. For Q-function selection, candidate Q_j gives the critic $Q_i - Q_j$; for occupancy-ratio selection, candidate w_j gives $\log(w_i/w_j)$. A comparator close to the truth then gives a pairwise critic close to the oracle critic, with exact recovery under realizability. We test each candidate against all such pairwise critics and include a penalty for comparator approximation error. We call the resulting procedure *pairwise Bellman validation* (PBV), which selects the candidate with the smallest worst-case penalized Bellman test.

For Q-function selection, testing a candidate's Bellman residual against critics under the offline distribution need not control Q-function error, because the target-policy Bellman operator need not be contractive in the corresponding L^2 norm. We therefore propose target-aligned PBV (TA-PBV), which uses marginalized importance weighting so that the pairwise tests are evaluated under the discounted target-occupancy distribution, where the contraction property of the Bellman operator implies Q-function error control. We also develop an unweighted variant that avoids occupancy-ratio estimation but requires a pairwise coercivity condition linking the offline Bellman tests to Q-function error, analogous to matrix-conditioning assumptions in LSTD-Tournament and linear TD learning (Bradtke & Barto, 1996; Tsitsiklis & Van Roy, 1997; Lagoudakis & Parr, 2003; Amortila et al., 2026). For occupancy-ratio selection, we propose adjoint PBV (A-PBV), which applies the pairwise construction to the adjoint Bellman equation using log-ratio critics. No analogous reweighting is needed because the adjoint contraction used to control generalized-KL error holds under the offline distribution. Under misspecification, the penalty yields oracle inequalities relative to the best candidate, using a quadratic penalty for Q-function selection and a Hellinger penalty for occupancy-ratio selection.

Contributions. Our contributions are threefold.

1. We develop a common theory for pairwise selection under Bellman and adjoint Bellman equations that allows heterogeneous and misspecified classes.
2. For **A-PBV and TA-PBV**, we establish joint finite-sample oracle inequalities that separate occupancy-ratio and Q-function approximation error, quantify propagation of ratio-selection error, and attain $\sqrt{\log n/n}$ validation error up to logarithmic dependence on class cardinality.
3. For **unweighted PBV**, we establish a finite-sample oracle inequality under pairwise coercivity, yielding a weighting-free guarantee for Q-function selection.

1.1 Related work

Model selection for Q-functions and occupancy ratios. Existing OPE model-selection methods have focused primarily on Q-functions, using Bellman residuals, candidate-dependent projections, or auxiliary regressions and critics in place of the unavailable supervised loss (Farahmand & Szepesvári, 2011; Xie & Jiang, 2021; Zhang & Jiang, 2021; Lee et al., 2022; Zitovsky et al., 2023; Miyaguchi, 2026). These approaches generally introduce additional estimation or structural requirements beyond the candidate class itself; LSTD-Tournament is an important exception and

is discussed separately below. By contrast, model selection among already-fitted occupancy-ratio estimators has received substantially less attention, despite extensive work on occupancy-ratio estimation (Nachum et al., 2019; Uehara et al., 2020; 2021; van der Laan & Kallus, 2026). A-PBV addresses this post-hoc selection problem directly. Extended discussion appears in Appendix E.

Closest comparison: LSTD-Tournament. LSTD-Tournament (Liu et al., 2025) constructs validation criteria directly from a finite class of candidate Q-functions using pairwise LSTDQ. For each pair (Q_i, Q_j), it treats (Q_i, Q_j) as a two-dimensional feature map and evaluates the corresponding LSTDQ normal-equation residuals (Lagoudakis & Parr, 2003). If $Q^\pi \in \{Q_i, Q_j\}$, then Q^π is realizable in this span, and invertibility of the pair-specific LSTD matrix links the residual to Q-function error. Taking the worst residual across comparators gives a tournament score for each candidate. Its theoretical guarantee therefore requires the class to contain Q^π and depends on the conditioning of the pair-specific LSTD matrices. In their experiments, Liu et al. (2025) use the normalized-difference feature map $[Q_i/c_i, (Q_j - Q_i)/c_{ij}]^\top$ for numerical stability, which is close in form to an unpenalized version of unweighted PBV but arises from pairwise LSTDQ.

PBV instead starts from the oracle critic $Q_i - Q^\pi$. Replacing the unknown Q^π with a comparator Q_j gives the pairwise critic $Q_i - Q_j$, and a quadratic penalty accounts for comparator approximation error. This yields oracle inequalities relative to the best candidate in the class without requiring Q^π to belong to the class. Unweighted PBV requires a pairwise coercivity condition analogous to the matrix-conditioning requirement in LSTD-Tournament, whereas discounted-occupancy weighting removes this condition. Applying the same construction to the adjoint Bellman equation gives pairwise selection of occupancy-ratio estimators.

2 OPE SETUP

2.1 MDP AND VALIDATION DATA

Consider a time-homogeneous Markov decision process with state space $\mathcal{S}$, action space $\mathcal{A}$, transition kernel $P(\mathrm{d}s' \mid s, a)$, reward R, and discount factor $\gamma \in [0, 1)$. Write $X = (S, A)$ and let $r(x) := \mathbb{E}[R \mid X = x]$ denote the mean reward function. Fix a target policy $\pi(\mathrm{d}a \mid s)$ and an initial state–action distribution d_0, and let $P_\pi(\mathrm{d}s', \mathrm{d}a' \mid s, a) := P(\mathrm{d}s' \mid s, a)\pi(\mathrm{d}a' \mid s')$ denote the induced state–action transition kernel. For any state–action function f, define $(\pi f)(s') := \int f(s', a')\,\pi(\mathrm{d}a' \mid s')$ and $(P_\pi f)(s, a) := \int (\pi f)(s')\,P(\mathrm{d}s' \mid s, a)$. The goal of OPE is to estimate the discounted value $V^\pi := \mathbb{E}_\pi[\sum_{t=0}^\infty \gamma^t R_t \mid X_0 \sim d_0]$, where $\mathbb{E}_\pi$ denotes expectation over trajectories generated by the MDP under π and initialized from d_0.

For model selection, we use a sample-splitting setup. Candidate classes and nuisance estimates are fitted on a training sample and evaluated on an independent validation sample $\mathcal{D}_n := \{(X_t, R_t, S'_t)\}_{t=1}^n$. We condition on the training sample and assume that $(X_t)_{t\geq 1}$ is a stationary Markov chain with invariant offline distribution ν and transition operator $(P_b f)(x) := \mathbb{E}[f(X_{t+1}) \mid X_t = x]$. This includes trajectories from a stationary behavior policy $b(\mathrm{d}a \mid s)$; for independent transitions, $(P_b f)(x) = \mathbb{E}_\nu[f(X)]$. In both cases, the conditional law of (R_t, S'_t) given X_t is the one-step MDP law. For any state–action distribution ρ, square-integrable f, g, and positive integer m, write $\langle f, g\rangle_\rho := \mathbb{E}_\rho[f(X)g(X)]$ and $\|f\|_{2,\rho} := \langle f, f\rangle_\rho^{1/2}$, and set $[m] := \{1, \ldots, m\}$.

2.2 BELLMAN AND ADJOINT BELLMAN EQUATIONS

The two nuisance functions we study are the target Q-function and the discounted occupancy ratio. Both can be used to recover the policy value and satisfy Bellman and adjoint Bellman equations, respectively. The target Q-function is $Q^\pi(x) := \mathbb{E}_\pi[\sum_{t=0}^\infty \gamma^t R_t \,|\, X_0 = x]$, and recovers the policy value through $V^\pi = \mathbb{E}_{d_0}[Q^\pi(X)]$. It is characterized by the Bellman fixed-point equation

$$Q^\pi = \mathcal{B}_\pi Q^\pi, \qquad \mathcal{B}_\pi Q := r + \gamma P_\pi Q, \tag{1}$$

where $\mathcal{B}_\pi$ is the Bellman operator (Sutton & Barto, 1998). The normalized discounted occupancy ratio is

$$w^\pi := \frac{\mathrm{d}d_{\pi,\gamma}}{\mathrm{d}\nu}, \qquad d_{\pi,\gamma} := (1-\gamma)\sum_{t=0}^\infty \gamma^t \mathcal{L}_\pi(X_t), \qquad X_0 \sim d_0,$$

where $\mathcal{L}_\pi(X_t)$ denotes the marginal law of X_t under the target policy. We assume $d_{\pi,\gamma} \ll \nu$. The ratio yields $V^\pi = (1-\gamma)^{-1}\mathbb{E}_{d_{\pi,\gamma}}[r(X)] = (1-\gamma)^{-1}\mathbb{E}_\nu[w^\pi(X)r(X)]$. Moreover, w^π satisfies the adjoint Bellman fixed-point equation (van der Laan & Kallus, 2026)

$$\mathcal{B}_\pi^* w^\pi = w^\pi, \qquad \mathcal{B}_\pi^* w := (1-\gamma)\frac{\mathrm{d}d_0}{\mathrm{d}\nu} + \gamma P_\pi^* w, \qquad P_\pi^* w := \frac{\mathrm{d}\{(w\nu)P_\pi\}}{\mathrm{d}\nu}, \tag{2}$$

which follows from the occupancy identity $d_{\pi,\gamma} = (1-\gamma)d_0 + \gamma d_{\pi,\gamma}P_\pi$ (Nachum et al., 2019).

The terminology "adjoint" reflects that P_π^* is the adjoint of P_π under the ν-inner product, so $\langle P_\pi^* w, Q\rangle_\nu = \langle w, P_\pi Q\rangle_\nu$ (Uehara et al., 2021).

3 Validation for Q-function selection

We first examine why standard validation criteria fail for Q-function selection and then develop the pairwise Bellman validation construction. Section 4 presents the resulting PBV algorithms.

3.1 Limitations of standard validation criteria

Let $Q_1, \ldots, Q_K$ be a finite class of candidate Q-functions fitted on the training sample. A natural approach, which we refer to as *Bellman-residual validation*, selects candidates using the mean-squared Bellman residual under the offline distribution ν (Antos et al., 2008; Farahmand & Szepesvári, 2011):

$$\mathcal{L}_{\mathrm{BR}}(Q) := \|Q - \mathcal{B}_\pi Q\|_{2,\nu}^2.$$

Because Q^π is a Bellman fixed point, $\mathcal{L}_{\mathrm{BR}}(Q^\pi) = 0$, so the minimum over a class containing Q^π is attained at Q^π. The criterion is not directly observable because $(\mathcal{B}_\pi Q)(X)$ depends on unknown conditional laws. Without an MDP model, evaluating it generally requires estimating the Bellman residual $Q_i - \mathcal{B}_\pi Q_i$, for example by separately regressing $Q_i(X) - R - \gamma(\pi Q_i)(S')$ on X for each candidate (Zitovsky et al., 2023), introducing additional computation, tuning, and estimation error.

A natural observable alternative is squared one-step temporal-difference (TD) error (Sutton, 1988),

$$\mathcal{L}_{\mathrm{TD}}(Q) := \mathbb{E}_\nu\left[\{Q(X) - R - \gamma(\pi Q)(S')\}^2\right],$$

which is directly estimable from held-out transitions. Although the TD residual is conditionally unbiased for the Bellman residual, $\mathbb{E}[Q(X) - R - \gamma(\pi Q)(S') \mid X] = Q(X) - (\mathcal{B}_\pi Q)(X)$, squaring introduces a conditional-variance term:

$$\mathcal{L}_{\mathrm{TD}}(Q) = \mathcal{L}_{\mathrm{BR}}(Q) + \mathbb{E}_\nu[\mathrm{Var}(R + \gamma(\pi Q)(S') \mid X)].$$

For $\gamma > 0$ and stochastic transitions, this term generally depends on Q, so minimizing squared TD error need not minimize the Bellman-residual error (Zitovsky et al., 2023). Avoiding this bias leads to the classical double-sampling obstruction: unbiased estimation of the squared Bellman residual requires two conditionally independent targets at the same state–action pair (Baird III, 1995; Feng et al., 2019).

Minimax Bellman-moment methods provide a natural way around this obstruction and a basis for model selection. At the true Q-function, the TD residual has conditional mean zero given X. Multiplying this residual by a critic $f(X)$ and averaging therefore gives a moment condition that vanishes at Q^π, as in minimax Q-learning and related methods (Antos et al., 2008; Feng et al., 2019; Uehara et al., 2020; Zanette & Wainwright, 2022). In particular, because Q^π is a Bellman fixed point, $\langle f, Q^\pi - \mathcal{B}_\pi Q^\pi\rangle_\nu = 0$ for every suitable critic f. These moments are observable from one-step transitions because

$$\mathbb{E}_\nu[f(X)\{Q(X) - R - \gamma(\pi Q)(S')\}] = \langle f, Q - \mathcal{B}_\pi Q\rangle_\nu.$$

Given a critic class $\mathcal{F}$, this motivates the minimax population criterion

$$\mathcal{L}_{\mathcal{F}}(Q) := \sup_{f \in \mathcal{F}:\ \|f\|_{2,\nu} \le 1} \left|\mathbb{E}_\nu[f(X)\{Q(X) - R - \gamma(\pi Q)(S')\}]\right|,$$

which measures the largest violation over the critic class. Its empirical analogue replaces the expectation by a validation-sample average.

The main practical challenge is choosing a critic class $\mathcal{F}$ for which these tests distinguish among candidate Q-functions. If $\mathcal{F}$ contains the normalized Bellman residual $(Q_i - \mathcal{B}_\pi Q_i)/\|Q_i - \mathcal{B}_\pi Q_i\|_{2,\nu}$ for each $i \in [K]$, then Cauchy–Schwarz and inclusion of this critic give $\mathcal{L}_{\mathcal{F}}(Q_i) = \|Q_i - \mathcal{B}_\pi Q_i\|_{2,\nu}$ (Uehara et al., 2021). Under this choice of critic class, minimax moment validation recovers Bellman-residual validation without directly estimating $\mathcal{B}_\pi Q_i$. The remaining difficulty is constructing, regularizing, and tuning an informative critic class: restrictive classes may fail to distinguish among candidates, whereas richer classes increase statistical and computational complexity (Uehara et al., 2020; Zanette & Wainwright, 2022).

3.2 From oracle critics to pairwise differences

The preceding discussion shows that minimax model selection hinges on choosing informative critics. Standard minimax analyses choose critic classes so that the resulting worst-case moments recover or approximate Bellman-residual validation. We instead identify ideal critics for which the corresponding moments directly control candidate estimation error rather than Bellman-residual error, and show that these critics can be approximated using pairwise differences within the candidate class. The construction rests on three observations.

Our first observation is that the candidate error $\Delta_i := Q_i - Q^\pi$ can itself serve as the critic. Testing the Bellman residual against Δ_i gives the oracle criterion

$$\begin{aligned}\mathcal{L}^\star(Q_i) &:= \mathbb{E}_\nu[\Delta_i(X)\,\{Q_i(X) - R - \gamma(\pi Q_i)(S')\}] \\ &= \langle \Delta_i, Q_i - \mathcal{B}_\pi Q_i\rangle_\nu = \langle \Delta_i, (I - \gamma P_\pi)\Delta_i\rangle_\nu,\end{aligned}$$

where the final equality follows from the fixed-point identity $\mathcal{B}_\pi Q^\pi = Q^\pi$. In contrast to squared Bellman-residual criteria, $\mathcal{L}^\star(Q_i)$ is linear in the Bellman residual and, if Δ_i were known, could be estimated directly from one-step transitions without estimating $\mathcal{B}_\pi Q_i$.

Whereas standard minimax criteria are constructed to recover the Bellman-residual error $\mathcal{L}_{\mathrm{BR}}(Q_i)$, $\mathcal{L}^\star(Q_i)$ is constructed to be directly comparable to the candidate mean-squared error $\|\Delta_i\|_{2,\nu}^2$. Since $\mathcal{L}^\star(Q_i) = \langle \Delta_i, (I - \gamma P_\pi)\Delta_i\rangle_\nu$, a lower bound in terms of $\|\Delta_i\|_{2,\nu}^2$ follows if the quadratic form induced by $I - \gamma P_\pi$ is coercive along the candidate-error directions, meaning that, for some $\kappa_\nu > 0$ and every $i \in [K]$,

$$\langle \Delta_i, (I - \gamma P_\pi)\Delta_i\rangle_\nu \geq \kappa_\nu \|\Delta_i\|_{2,\nu}^2. \tag{3}$$

A reverse bound follows from the standard one-step concentrability condition $C_{\nu,\pi} := \|\mathrm{d}(\nu P_\pi)/\mathrm{d}\nu\|_\infty < \infty$. In particular, Cauchy–Schwarz gives, for every $i \in [K]$,

$$\langle \Delta_i, (I - \gamma P_\pi)\Delta_i\rangle_\nu \leq (1 + \gamma\sqrt{C_{\nu,\pi}})\|\Delta_i\|_{2,\nu}^2.$$

Consequently, $\mathcal{L}^\star(Q_i)$ is comparable, up to constants, to $\|\Delta_i\|_{2,\nu}^2$. It follows that any oracle selector $i^\star_{\mathrm{oracle}} \in \arg\min_{1\leq i\leq K} \mathcal{L}^\star(Q_i)$ satisfies

$$\|Q_{i^\star_{\mathrm{oracle}}} - Q^\pi\|_{2,\nu} \leq \left(\frac{1 + \gamma\sqrt{C_{\nu,\pi}}}{\kappa_\nu}\right)^{1/2} \min_{1\leq i\leq K} \|Q_i - Q^\pi\|_{2,\nu}.$$

Thus, if the oracle criterion were available, minimizing it would select a candidate whose error is within a fixed factor of the minimum error in the class.

Our second observation is that pairwise differences within the candidate class approximate these oracle critics. Indeed, for any Q_j, $Q_i - Q_j = \Delta_i - \Delta_j$. Under realizability, if $Q_{i^\star} = Q^\pi$, then $Q_i - Q_{i^\star} = \Delta_i$ exactly. More generally, $Q_i - Q_j$ differs from the oracle critic Δ_i by $-\Delta_j$, so a comparison candidate close to Q^π provides an accurate approximation. Because the best comparison candidate is unknown, we evaluate the Bellman residual of each Q_i against all K pairwise critics $\{Q_i - Q_j : j \in [K]\}$ and take the largest resulting moment as its worst-case score. We then select the candidate with the smallest worst-case score. Thus, the K^2 candidate–critic pairs are determined entirely by the candidate class and require no auxiliary critic fitting.

Pairwise critics remove the need to learn a separate critic class, but they do not by themselves ensure that Bellman moments under the offline distribution control Q-function error. The pairwise criterion still requires the coercivity condition in (3), which depends on the relationship between the offline distribution and the target-policy dynamics. A simple sufficient condition is

$\|P_\pi \Delta_i\|_{2,\nu} \le c\|\Delta_i\|_{2,\nu}$ for every $i \in [K]$ and some $c < \gamma^{-1}$, in which case Cauchy–Schwarz gives (3) with $\kappa_\nu = 1-\gamma c$. Under one-step concentrability, one may take $c = \sqrt{C_{\nu,\pi}}$, so coercivity holds whenever $C_{\nu,\pi} < \gamma^{-2}$. As a special case, under stationary on-policy sampling, $C_{\nu,\pi} = 1$ and P_π is an $L^2(\nu)$ contraction, so coercivity holds automatically with $\kappa_\nu = 1-\gamma$. When Q^π and the candidate Q-functions lie in a common linear span, this condition is analogous to matrix-conditioning assumptions in linear TD learning (Bradtke & Barto, 1996; Tsitsiklis & Van Roy, 1997; Lagoudakis & Parr, 2003; Amortila et al., 2026). Under a general offline distribution, however, coercivity need not hold.

Our third observation is that marginalized importance weighting by $w^\pi = \mathrm{d}d_{\pi,\gamma}/\mathrm{d}\nu$ avoids the offline-law coercivity requirement by evaluating the Bellman criterion under the discounted target occupancy distribution. Under this measure, γP_π is a $\sqrt{\gamma}$-contraction in $L^2(d_{\pi,\gamma})$ (van der Laan & Kallus, 2025), and hence

$$(1-\sqrt{\gamma})\|\Delta_i\|^2_{2,d_{\pi,\gamma}} \le \langle \Delta_i, (I-\gamma P_\pi)\Delta_i\rangle_{d_{\pi,\gamma}} \le (1+\sqrt{\gamma})\|\Delta_i\|^2_{2,d_{\pi,\gamma}}.$$

The oracle criterion is therefore comparable, up to constants, to squared prediction error under the target occupancy without the coercivity assumption in (3). The tradeoff is that the occupancy ratio w^π must be estimated sufficiently accurately.

Combining target-occupancy weighting with pairwise difference critics yields our target-aligned selector. For candidate i and comparator j, let $h_{ij} := Q_i - Q_j$ and define the target-aligned pairwise moment

$$\mathcal{M}^{\mathrm{TA}}_{ij} := \mathbb{E}_\nu[w^\pi(X) h_{ij}(X)\{Q_i(X) - R - \gamma(\pi Q_i)(S')\}] = \langle h_{ij}, Q_i - \mathcal{B}_\pi Q_i\rangle_{d_{\pi,\gamma}}.$$

Candidate i is scored by its worst-case moment $\max_{1\le j\le K} \mathcal{M}^{\mathrm{TA}}_{ij}$. Under realizability, minimizing this score recovers the target Q-function. Indeed, if $Q_{i^\star} = Q^\pi$ for some $i^\star \in [K]$, then $h_{ii^\star} = \Delta_i$ is the oracle critic, so

$$\max_{1\le j\le K} \mathcal{M}^{\mathrm{TA}}_{ij} \ge \mathcal{M}^{\mathrm{TA}}_{ii^\star} \ge (1-\sqrt{\gamma})\|\Delta_i\|^2_{2,d_{\pi,\gamma}}.$$

For $i = i^\star$, the Bellman residual vanishes and every pairwise moment is zero. The minimum worst-case score is therefore zero, and any population minimizer satisfies $\|\Delta_i\|_{2,d_{\pi,\gamma}} = 0$.

4 Pairwise Bellman validation algorithms

We now turn the population construction in Section 3.2 into implementable model-selection procedures. The additional steps are to account for critic-approximation error under misspecification and, for TA-PBV, to estimate the discounted occupancy ratio.

4.1 Adjoint PBV and target-aligned PBV

To implement TA-PBV, we first select an occupancy ratio using adjoint PBV (A-PBV), which applies the pairwise construction to the adjoint Bellman equation. We briefly review the construction here and defer further details and intuition to Appendix A. As in the Q-function case, the adjoint Bellman moment for any fixed critic is directly estimable from one-step transitions and independent draws from d_0 (Nachum et al., 2019).

For nonnegative integrable functions p and q, define the generalized KL divergence

$$\mathsf{D}_\nu(p\|q) := \mathbb{E}_\nu\left[p(X)\log\frac{p(X)}{q(X)} - p(X) + q(X)\right].$$

Unlike the usual KL divergence, D_ν does not require p and q to integrate to one and therefore permits unnormalized ratio candidates. The Bregman three-point identity (Bauschke & Combettes, 2017), together with contraction of the adjoint Bellman operator in generalized KL divergence (van der Laan & Kallus, 2026), gives

$$\begin{aligned}\langle w - \mathcal{B}^*_\pi w, \log(w/w^\pi)\rangle_\nu &= \mathsf{D}_\nu(w\|w^\pi) - \mathsf{D}_\nu(\mathcal{B}^*_\pi w\|w^\pi) + \mathsf{D}_\nu(\mathcal{B}^*_\pi w\|w)\\ &\ge (1-\gamma)\mathsf{D}_\nu(w\|w^\pi).\end{aligned}$$

Algorithm 1 Adjoint and target-aligned pairwise Bellman validation

Require: Ratio candidates $w_1, \dots, w_M$; Q-function candidates $Q_1, \dots, Q_K$; validation trajectory $\mathcal{D}_n$; independent draws $X_{0,1}, \dots, X_{0,n} \sim d_0$ (or exact d_0 moments); discount factor γ

1: Select $\widehat{i}_w$ as the smallest index in $\arg\min_{1\le a\le M} \max_{1\le b\le M} \widehat{\mathcal{S}}^{\mathrm{pen}}_{ab}$ and set $\widehat{w} := w_{\widehat{i}_w}$.

2: Select $\widehat{i}_Q$ as the smallest index in $\arg\min_{1\le i\le K} \max_{1\le j\le K} \widehat{\mathcal{M}}^{\mathrm{pen}}_{ij}(\widehat{w})$ and set $\widehat{Q} := Q_{\widehat{i}_Q}$.

3: **return** $(\widehat{w}, \widehat{Q})$.

Thus, the oracle critic is $\log(w/w^\pi)$. We approximate it using pairwise log-ratio differences $\ell_{ab} := \log(w_a/w_b)$ and select an occupancy-ratio candidate using the resulting adjoint Bellman moments.

Under misspecification, the pairwise critics only approximate their oracle counterparts. Hellinger and quadratic penalties account for this approximation error in the occupancy-ratio and Q-function stages, respectively. To implement A-PBV and TA-PBV, let $X_{0,1}, \dots, X_{0,n}$ be independent draws from d_0. If expectations under d_0 are known, the corresponding empirical averages may instead be replaced by their population values. Define the penalized empirical pairwise scores

$$\begin{aligned}
\widehat{\mathcal{S}}^{\mathrm{pen}}_{ab} &:= \frac{1}{n}\sum_{t=1}^{n} w_a(X_t)\{\ell_{ab}(X_t) - \gamma(\pi\ell_{ab})(S'_t)\} - \frac{1-\gamma}{n}\sum_{t=1}^{n}\ell_{ab}(X_{0,t}) \\
&\quad - \frac{1-\gamma}{4n}\sum_{t=1}^{n}\{\sqrt{w_a(X_t)} - \sqrt{w_b(X_t)}\}^2, \\
\widehat{\mathcal{M}}^{\mathrm{pen}}_{ij}(w) &:= \frac{1}{n}\sum_{t=1}^{n} w(X_t)h_{ij}(X_t)\{Q_i(X_t) - R_t - \gamma(\pi Q_i)(S'_t)\} \\
&\quad - \frac{1-\sqrt{\gamma}}{4n}\sum_{t=1}^{n} w(X_t)h_{ij}(X_t)^2,
\end{aligned}$$

where $h_{ij} := Q_i - Q_j$. The coefficient $1-\sqrt{\gamma}$ in the Q-function penalty matches the lower contraction bound under the discounted target occupancy and yields an oracle comparison without requiring realizability. Algorithm 1 combines the two stages and evaluates M^2+K^2 pairwise scores, with no auxiliary validation model or critic fitting.

4.2 Unweighted PBV

When target-occupancy weighting is unavailable or unstable, unweighted PBV evaluates both the Bellman moment and its quadratic penalty directly under the offline distribution. Define

$$\begin{aligned}
\widehat{\mathcal{M}}^{\mathrm{unw}}_{ij} &:= \frac{1}{n}\sum_{t=1}^{n}\left[h_{ij}(X_t)\{Q_i(X_t) - R_t - \gamma(\pi Q_i)(S'_t)\} - \frac{1}{4}h_{ij}(X_t)\{h_{ij}(X_t) - \gamma(\pi h_{ij})(S'_t)\}\right], \\
\widehat{i}^{\mathrm{unw}}_Q &:= \min\arg\min_{1\le i\le K}\max_{1\le j\le K}\widehat{\mathcal{M}}^{\mathrm{unw}}_{ij}. \qquad (4)
\end{aligned}$$

By iterated expectation, the population penalty is $\frac{1}{4}\langle h_{ij}, (I-\gamma P_\pi)h_{ij}\rangle_\nu$, the same quadratic form appearing in the coercivity condition. Using this form directly avoids requiring the coercivity constant κ_ν as an input to the procedure, although coercivity remains necessary for the corresponding error guarantee.

5 Finite-sample guarantees

5.1 Guarantee for A-PBV and TA-PBV

We now establish finite-sample oracle inequalities for A-PBV, TA-PBV, and unweighted PBV. To accommodate temporal dependence in the validation trajectory, we impose the following spectral-gap condition, which holds with $g_b = 1$ for independent transitions (Jiang et al., 2025).

(A1) *Behavior-process spectral gap.* There exists $g_b \in (0, 1]$ such that

$$\|P_b f\|_{2,\nu} \le (1 - g_b)\|f\|_{2,\nu} \qquad \text{for every real-valued } f \in L^2(\nu) \text{ with } \mathbb{E}_\nu f = 0. \tag{5}$$

For adjoint PBV, the next condition ensures uniform boundedness of the pairwise log-ratio critics.

(A2) *Target-relative support and comparability.* Let $\mathcal{X}^\pi := \{x : w^\pi(x) > 0\}$. Each candidate satisfies $w_i = 0$ ν-almost surely on $(\mathcal{X}^\pi)^c$, and there exists $\Lambda_w \ge 1$ such that

$$\Lambda_w^{-1} \le \frac{w_i(X)}{w^\pi(X)} \le \Lambda_w \qquad \nu\text{-almost surely on } \mathcal{X}^\pi, \qquad i = 1, \dots, M.$$

Define the approximation errors $\epsilon_w := \min_{a\in[M]} \sqrt{\mathsf{D}_\nu(w_a\|w^\pi)}$ and $\epsilon_Q := \min_{i\in[K]} \|Q_i - Q^\pi\|_{2,d_{\pi,\gamma}}$, and the validation error rates $r_w(\delta) := \{\log(4n/\delta)\log(24M/\delta)/(ng_b)\}^{1/2}$ and $r_{Q,\mathrm{TA}}(\delta) := \{\log(12MK/\delta)/(ng_b)\}^{1/2}$.

Theorem 5.1 (A-PBV and TA-PBV oracle guarantee)**.** *Suppose Conditions A1 and A2 hold, together with the technical conditions in Appendix B.1. Then there exist finite constants $C_{\mathrm{ratio}}, C_{\mathrm{TA}} > 0$, independent of n, M, K, δ, and γ, such that, for every $\delta \in (0, 1)$, with probability at least $1 - \delta$, the output of Algorithm 1 satisfies*

$$\sqrt{\mathsf{D}_\nu(\widehat{w}\|w^\pi)} \le C_{\mathrm{ratio}} \left\{ \frac{\epsilon_w}{1 - \sqrt{\gamma}} + \frac{r_w(\delta/2)}{1 - \gamma} \right\}. \tag{6}$$

$$\|\widehat{Q} - Q^\pi\|_{2,d_{\pi,\gamma}} \le C_{\mathrm{TA}} \left[\frac{\epsilon_Q + r_{Q,\mathrm{TA}}(\delta/2)}{1 - \sqrt{\gamma}} + \frac{1}{1 - \sqrt{\gamma}} \left\{ \frac{\epsilon_w}{1 - \sqrt{\gamma}} + \frac{r_w(\delta/2)}{1 - \gamma} \right\} \right]. \tag{7}$$

Temporal dependence reduces the effective validation sample size from n to ng_b, while dependence on the class cardinality is only logarithmic. As $\gamma \to 1$, the direct validation terms in both the ratio and Q-function stages scale at the standard order $(1 - \gamma)^{-1}$. Error in the selected ratio incurs an additional factor of order $(1 - \gamma)^{-1}$ when propagated into Q-function selection, yielding an overall $(1 - \gamma)^{-2}$ dependence for the propagated ratio-error term.

For Q-functions, Cauchy–Schwarz gives the plug-in value bound $|\mathbb{E}_{d_0}[\widehat{Q}(X)] - V^\pi| \le c_{0,\pi}\|\widehat{Q} - Q^\pi\|_{2,d_{\pi,\gamma}}$, where $c_{0,\pi} := \|\mathrm{d}d_0/\mathrm{d}d_{\pi,\gamma}\|_{2,d_{\pi,\gamma}}$, and the Q-function error is bounded in (7). For occupancy ratios, if $r \in L^2(d_{\pi,\gamma})$, the analogous plug-in estimator satisfies $\left|(1-\gamma)^{-1}\mathbb{E}_\nu[\widehat{w}(X)r(X)] - V^\pi\right| \le (1 - \gamma)^{-1}\sqrt{2\Lambda_w}\,\|r\|_{2,d_{\pi,\gamma}}\sqrt{\mathsf{D}_\nu(\widehat{w}\|w^\pi)}$, where the generalized-KL error is bounded in (6). A doubly robust bound is given in Appendix C.

5.2 Unweighted PBV

Unweighted PBV avoids occupancy-ratio weighting but instead requires pairwise coercivity and one-step concentrability under the offline law.

(A3) *One-step concentrability and pairwise coercivity.* The one-step law satisfies $\nu P_\pi \ll \nu$ with $C_{\nu,\pi} := \|\mathrm{d}(\nu P_\pi)/\mathrm{d}\nu\|_\infty < \infty$, and there exists $\kappa_\nu \in (0, 1]$ such that

$$\langle h_{ij}, (I - \gamma P_\pi)h_{ij}\rangle_\nu \ge \kappa_\nu \|h_{ij}\|_{2,\nu}^2 \qquad \text{for every } i, j \in [K].$$

Theorem 5.2 (Unweighted PBV oracle guarantee)**.** *Suppose Conditions A1, A3, B2, B3, and B5 hold. Then there exists a finite constant $C_{Q,\mathrm{unw}} > 0$, depending only on H_Q, G_Q, σ_Q, such that, for every $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\|Q_{\widehat{i}_Q^{\mathrm{unw}}} - Q^\pi\|_{2,\nu} \le \left\{ 1 + \frac{2(1 + \gamma\sqrt{C_{\nu,\pi}})}{\kappa_\nu} \right\} \min_{1\le i\le K} \|Q_i - Q^\pi\|_{2,\nu} + \frac{C_{Q,\mathrm{unw}}}{\kappa_\nu} \sqrt{\frac{\log(12K/\delta)}{ng_b}}.$$

6 Experimental investigation

We evaluate PBV on Hopper (Liu et al., 2025) and Deep OPE/D4RL (Fu et al., 2021; 2020). We compare unweighted PBV, unweighted PBV without the quadratic penalty, squared temporal-difference error, and LSTD-Tournament using the normalized-difference implementation of Liu

Table 1: Selected-Q performance at $n_Q = 3{,}200$ (mean; lower is better). Target Q-RMSE is under the discounted target occupancy. Bold marks the smallest mean within each study and benchmark.

Study	Selector	Absolute value error		Target Q-RMSE	
		Hopper	Deep OPE	Hopper	Deep OPE
Broad offline	Unweighted PBV	**4.2**	264.8	11.2	412.7
	Unweighted PBV (no penalty)	4.8	265.1	**10.7**	**402.1**
	LSTD-Tournament	5.6	**258.8**	12.7	411.6
	Squared TD error	9.4	276.2	18.7	461.9
Controlled overlap	Unweighted PBV	25.5	83.2	68.7	135.9
	Unweighted PBV (no penalty)	15.1	75.9	42.7	131.3
	TA-PBV ($T = 5$)	**14.5**	**69.3**	34.2	136.5
	LSTD-Tournament	15.9	74.7	47.1	**128.4**
	Squared TD error	15.5	177.8	**30.6**	343.0

et al. (2025); full details are given in Appendix D. We use both the original broad offline datasets and controlled-overlap datasets constructed to study target-occupancy weighting when coverage is not the dominant source of error. In the controlled-overlap datasets, the logging policy samples one benchmark policy uniformly for each trajectory, and each component policy is treated in turn as the target. Hence, $1/10$ of Hopper trajectories and $1/11$ of Deep OPE trajectories are generated on-policy for each target, and the corresponding trajectory and discounted occupancy ratios are bounded by 10 and 11, respectively.

Each Q-class contains 15 candidates. Broad Hopper uses the candidates provided with the benchmark; Deep OPE and controlled Hopper use two-hidden-layer neural FQE networks with width 64, 128, or 256, saved after 50, 100, 200, 400, or 800 FQE iterations. For TA-PBV, A-PBV first selects among 28 fitted occupancy-ratio evaluation (FORE) candidates, and the selected ratio is then used to validate the Q-candidates on an independent sample. We use $n_Q = 3{,}200$ validation samples. Following Sinha et al. (2022), we temper the selected ratio as $w^{1/5}/\mathbb{P}_{n,w}[w^{1/5}]$. Sensitivity analyses and additional experimental details are given in Appendix D.

On the broad offline benchmarks, unweighted PBV has the lowest mean value error on Hopper, while LSTD-Tournament has the lowest mean value error on Deep OPE (Table 1). We also evaluated TA-PBV at $T = 5$ in the broad-offline setting; its mean value errors were 5.7 on Hopper and 264.0 on Deep OPE, with no consistent improvement from occupancy weighting under weak coverage. At the policy level, unweighted PBV has lower value error than LSTD-Tournament in 65 of 120 Hopper comparisons and 114 of 220 Deep OPE comparisons (Figure S1).

Under controlled overlap, TA-PBV has the lowest mean value error on both benchmarks: 14.5 on Hopper and 69.3 on Deep OPE, compared with 25.5 and 83.2 for unweighted PBV and 15.9 and 74.7 for LSTD-Tournament. Squared TD error has the lowest target-Q RMSE on Hopper, while LSTD-Tournament has the lowest target-Q RMSE on Deep OPE. Target-occupancy weighting therefore improves value-based selection in the controlled-overlap setting. In the broad offline setting, where the estimated ratios are much more variable, it provides no corresponding improvement.

7 Discussion

PBV uses pairwise candidate differences as critics for both Q-function and occupancy-ratio selection, avoiding a separately learned critic class. It does not remove the basic difficulties of infinite-horizon OPE. TA-PBV addresses distribution shift through discounted-occupancy weighting but requires overlap and accurate occupancy-ratio estimation, whereas unweighted PBV avoids importance weighting at the cost of pairwise coercivity under the offline distribution. The guarantees also deteriorate with the effective horizon as $\gamma \to 1$. Important extensions include greater robustness to weak overlap and long horizons, as well as selection beyond finite classes. The strong empirical performance of LSTD-Tournament further supports the use of candidate-induced critics, despite its different LSTDQ derivation and realizability-based theory.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript writing and editing, as well as proof revision and checking. The authors reviewed and verified all AI-assisted content and take full responsibility for the paper.

REFERENCES

Felix Abramovich, Daniela De Canditiis, and Marianna Pensky. Solution of linear ill-posed problems by model selection and aggregation. *Electronic Journal of Statistics*, 12(1):1822–1841, 2018. doi: 10.1214/18-EJS1447. URL https://doi.org/10.1214/18-EJS1447.

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society and Oxford University Press, 2000. doi: 10.1090/mmono/191. URL https://doi.org/10.1090/mmono/191.

Philip Amortila, Audrey Huang, Akshay Krishnamurthy, and Nan Jiang. A unifying view of coverage in linear off-policy evaluation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/858e773f5c99a4186b548f3f531fb07e-Abstract-Conference.html.

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

Leemon C. Baird III. Residual algorithms: Reinforcement learning with function approximation. In Armand Prieditis and Stuart J. Russell (eds.), *Machine Learning Proceedings 1995*, pp. 30–37. Morgan Kaufmann, 1995. doi: 10.1016/B978-1-55860-377-6.50013-X. URL https://doi.org/10.1016/B978-1-55860-377-6.50013-X.

Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer, Cham, 2 edition, 2017. doi: 10.1007/978-3-319-48311-5. URL https://doi.org/10.1007/978-3-319-48311-5.

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

L. Cavalier, G. K. Golubev, D. Picard, and A. B. Tsybakov. Oracle inequalities for inverse problems. *The Annals of Statistics*, 30(3):843–874, 2002. doi: 10.1214/aos/1028674843. URL https://doi.org/10.1214/aos/1028674843.

Matej Cief, Branislav Kveton, and Michal Kompan. Cross-validated off-policy evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(15):16073–16081, 2025. doi: 10.1609/aaai.v39i15.33765. URL https://ojs.aaai.org/index.php/AAAI/article/view/33765.

Dong Dai, Philippe Rigollet, and Tong Zhang. Deviation optimal learning using greedy Q-aggregation. *The Annals of Statistics*, 40(3):1878–1905, 2012. doi: 10.1214/12-AOS1025. URL https://doi.org/10.1214/12-AOS1025.

Victor H. de la Peña. A general class of exponential inequalities for martingales and ratios. *The Annals of Probability*, 27(1):537–564, 1999. doi: 10.1214/aop/1022677271. URL https://doi.org/10.1214/aop/1022677271.

Amir-massoud Farahmand and Csaba Szepesvári. Model selection in reinforcement learning. *Machine Learning*, 85(3):299–332, 2011. doi: 10.1007/s10994-011-5254-7. URL https://doi.org/10.1007/s10994-011-5254-7.

Yihao Feng, Lihong Li, and Qiang Liu. A kernel loss for solving the bellman equation. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/48e59000d7dfcf6c1d96ce4a603ed738-Abstract.html.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975. doi: 10.1214/aop/1176996452. URL https://doi.org/10.1214/aop/1176996452.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: datasets for deep data-driven reinforcement learning. *CoRR*, abs/2004.07219, 2020. URL https://arxiv.org/abs/2004.07219.

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.

Audrey Huang and Nan Jiang. Occupancy-based policy gradient: Estimation, convergence, and optimality. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/010c855df402b443e0c16e5b7434e74c-Paper-Conference.pdf.

Audrey Huang, Jinglin Chen, and Nan Jiang. Reinforcement learning in low-rank MDPs with density features. In *International Conference on Machine Learning*, pp. 13710–13752, 2023. URL https://proceedings.mlr.press/v202/huang23a.html.

Alexander Irpan, Kanishka Rao, Konstantinos Bousmalis, Chris Harris, Julian Ibarz, and Sergey Levine. Off-policy evaluation via off-policy classification. In *Advances in Neural Information Processing Systems*, volume 32, pp. 5438–5449, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/b5b03f06271f8917685d14cea7c6c50a-Abstract.html.

Bai Jiang, Qiang Sun, and Jianqing Fan. Bernstein's inequalities for general markov chains. *arXiv preprint arXiv:1805.10721v4*, 2025. URL https://arxiv.org/abs/1805.10721v4.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 652–661, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/jiang16.html.

Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63, 2020. URL http://jmlr.org/papers/v21/19-827.html.

Nathan Kallus and Masatoshi Uehara. Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *Oper. Res.*, 70(6):3282–3302, 2022. doi: 10.1287/OPRE.2021.2249. URL https://doi.org/10.1287/opre.2021.2249.

Michail G. Lagoudakis and Ronald Parr. Least-squares policy iteration. *J. Mach. Learn. Res.*, 4: 1107–1149, 2003. URL https://jmlr.org/papers/v4/lagoudakis03a.html.

Hui Lan and Vasilis Syrgkanis. Causal Q-aggregation for CATE model selection. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4366–4374. PMLR, 2024. URL https://proceedings.mlr.press/v238/lan24a.html.

Jonathan N. Lee, George Tucker, Ofir Nachum, Bo Dai, and Emma Brunskill. Oracle inequalities for model selection in offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 28194–28207, 2022. doi: 10.52202/068431-2044. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b51693c2ba5b5ddf67429966576fb962-Abstract-Conference.html.

Zihao Li, Hui Lan, Vasilis Syrgkanis, Mengdi Wang, and Masatoshi Uehara. Regularized DeepIV with model selection. arXiv preprint arXiv:2403.04236, 2024. URL https://arxiv.org/abs/2403.04236.

Pai Liu, Lingfeng Zhao, Shivangi Agarwal, Jinghan Liu, Audrey Huang, Philip Amortila, and Nan Jiang. Model selection for off-policy evaluation: New algorithms and experimental protocol. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.neurips.cc/paper_files/paper/2025/hash/3d4dc72d715bd6415d356293079adf3d-Abstract-Conference.html.

Kohei Miyaguchi. Duality-based residual estimation for fully offline value-based reinforcement learning. In Emtiyaz Khan, Yingzhen Li, Arno Solin, and Aaditya Ramdas (eds.), *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 1576–1584. PMLR, 02–05 May 2026. URL https://proceedings.mlr.press/v300/miyaguchi26a.html.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(27):815–857, 2008. URL http://jmlr.org/papers/v9/munos08a.html.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/cf9a242b70f45317ffd281241fa66502-Paper.pdf.

Allen Nie, Yannis Flet-Berliac, Deon R. Jordan, William Steenbergen, and Emma Brunskill. Data-efficient pipeline for offline reinforcement learning with limited data. In *Advances in Neural Information Processing Systems*, volume 35, pp. 14810–14823, 2022. doi: 10.52202/068431-1077. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/5ee7ed60a7e8169012224dec5fe0d27f-Abstract-Conference.html.

Allen Nie, Yash Chandak, Christina J. Yuan, Anirudhan Badrinath, Yannis Flet-Berliac, and Emma Brunskill. OPERA: Automatic offline policy evaluation with re-weighted aggregates of multiple estimators. In *Advances in Neural Information Processing Systems*, volume 37, pp. 103652–103680, 2024. doi: 10.52202/079017-3292. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/bba9be4bc526c5d515a9d3c16ccfe138-Abstract-Conference.html.

Tom Le Paine, Cosmin Paduraru, Andrea Michi, Caglar Gulcehre, Konrad Zolna, Alexander Novikov, Ziyu Wang, and Nando de Freitas. Hyperparameter selection for offline reinforcement learning. *arXiv preprint arXiv:2007.09055*, 2020. URL https://arxiv.org/abs/2007.09055.

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

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, 1974. doi: 10.1111/j.2517-6161.1974.tb00994.x. URL https://doi.org/10.1111/j.2517-6161.1974.tb00994.x.

Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. Doubly robust off-policy evaluation with shrinkage. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9167–9176. PMLR, 2020a. URL https://proceedings.mlr.press/v119/su20a.html.

Yi Su, Pavithra Srinath, and Akshay Krishnamurthy. Adaptive estimator selection for off-policy evaluation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9196–9205. PMLR, 2020b. URL https://proceedings.mlr.press/v119/su20d.html.

Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44, 1988. doi: 10.1007/BF00115009. URL https://doi.org/10.1007/BF00115009.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998. ISBN 9780262193986. URL https://mitpress.mit.edu/9780262193986/reinforcement-learning/.

Shengpu Tang and Jenna Wiens. Model selection for offline reinforcement learning: Practical considerations for healthcare settings. In *Proceedings of the 6th Machine Learning for Healthcare Conference*, volume 149 of *Proceedings of Machine Learning Research*, pp. 2–35. PMLR, 2021. URL https://proceedings.mlr.press/v149/tang21a.html.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2139–2148, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/thomasa16.html.

John N. Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997. doi: 10.1109/9.580874. URL https://doi.org/10.1109/9.580874.

Takuma Udagawa, Haruka Kiyohara, Yusuke Narita, Yuta Saito, and Kei Tateno. Policy-adaptive estimator selection for off-policy evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8):10025–10033, 2023. doi: 10.1609/aaai.v37i8.26195. URL https://ojs.aaai.org/index.php/AAAI/article/view/26195.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9659–9668. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/uehara20a.html.

Masatoshi Uehara, Masaaki Imaizumi, Nan Jiang, Nathan Kallus, Wen Sun, and Tengyang Xie. Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency. *CoRR*, abs/2102.02981, 2021. URL https://arxiv.org/abs/2102.02981.

Masatoshi Uehara, Chengchun Shi, and Nathan Kallus. A review of off-policy evaluation in reinforcement learning. *Statistical Science*, 41(3):561–581, 2026. doi: 10.1214/25-STS985. URL https://doi.org/10.1214/25-STS985.

Lars van der Laan and Nathan Kallus. Fitted Q-evaluation without Bellman completeness via occupancy weighting. *CoRR*, abs/2512.23805, 2025. doi: 10.48550/ARXIV.2512.23805. URL https://doi.org/10.48550/arXiv.2512.23805.

Lars van der Laan and Nathan Kallus. Fitted occupancy-ratio evaluation without bellman completeness. *arXiv preprint arXiv:2607.05375*, 2026. doi: 10.48550/arXiv.2607.05375. URL https://arxiv.org/abs/2607.05375.

Lars van der Laan, David Hubbard, Allen Tran, Nathan Kallus, and Aurélien Bibaut. Semiparametric double reinforcement learning with applications to long-term causal inference. *arXiv preprint arXiv:2501.06926*, 2025. doi: 10.48550/arXiv.2501.06926. URL https://arxiv.org/abs/2501.06926.

Mark J. van der Laan and Sandrine Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. Working Paper 130, Division of Biostatistics, University of California, Berkeley, November 2003. URL https://biostats.bepress.com/ucbbiostat/paper130/.

Mark J. van der Laan, Eric C. Polley, and Alan E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1):Article 25, 2007. doi: 10.2202/1544-6115.1309. URL https://doi.org/10.2202/1544-6115.1309.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudík. Optimal and adaptive off-policy evaluation in contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3589–3597. PMLR, 2017. URL https://proceedings.mlr.press/v70/wang17a.html.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. doi: 10.1016/S0893-6080(05)80023-1. URL https://doi.org/10.1016/S0893-6080(05)80023-1.

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11404–11413. PMLR, 2021. URL https://proceedings.mlr.press/v139/xie21d.html.

Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/4ffb0d2ba92f664c2281970110a2e071-Paper.pdf.

Mengjiao Yang, Bo Dai, Ofir Nachum, George Tucker, and Dale Schuurmans. Offline policy selection under uncertainty. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 4376–4396. PMLR, 2022. URL https://proceedings.mlr.press/v151/yang22a.html.

Andrea Zanette and Martin J. Wainwright. Bellman residual orthogonalization for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3137–3151, 2022. doi: 10.52202/068431-0227. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/14ecbfb2216bab76195b60bfac7efb1f-Abstract-Conference.html.

Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. GenDICE: Generalized offline estimation of stationary values. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HkxlcnVFwB.

Siyuan Zhang and Nan Jiang. Towards hyperparameter-free policy selection for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12864–12875, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/6add07cf50424b14fdf649da87843d01-Abstract.html.

Joshua P. Zitovsky, Daniel De Marchi, Rishabh Agarwal, and Michael Rene Kosorok. Revisiting Bellman errors for offline model selection. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 43369–43406. PMLR, 2023. URL https://proceedings.mlr.press/v202/zitovsky23a.html.

Appendix roadmap. The full adjoint PBV construction and its standalone guarantees appear in Section A. The remaining conditions for the joint result and the standalone guarantees for TA-PBV and unweighted PBV are stated in Section B. The doubly robust extension appears in Appendix C, followed by experimental details and additional related work in Appendices D and E. The concentration result used throughout is isolated in Appendix F. Proofs of the occupancy-ratio, TA-PBV, unweighted Q-function, joint TA-PBV, and doubly robust guarantees follow in that order.

A ADJOINT PBV FOR OCCUPANCY-RATIO SELECTION

The first stage of Algorithm 1 selects the discounted occupancy ratio used by its TA-PBV stage. We consider a finite class of candidates $w_1, \ldots, w_M$ fitted on the training sample and, conditional on that sample, treat them as fixed and independent of the validation data. As in the Q-function setting, we use a minimax criterion based on adjoint Bellman moments and construct the critics directly from the candidate class.

A.1 FROM ORACLE LOG-RATIO CRITICS TO PAIRWISE SELECTION

For every bounded critic f, the adjoint Bellman residual satisfies the moment identity (Nachum et al., 2019)

$$\mathbb{E}_\nu[\{w(X) - (\mathcal{B}_\pi^* w)(X)\} f(X)] = \mathbb{E}_\nu[w(X)\{f(X) - \gamma(\pi f)(S')\}] - (1-\gamma)\mathbb{E}_{d_0}[f(X)]. \quad \text{(S1)}$$

In particular, this moment vanishes at $w = w^\pi$ for every bounded critic f. We measure occupancy-ratio error using the generalized Kullback–Leibler divergence, defined for nonnegative integrable functions p and q by

$$\mathsf{D}_\nu(p\|q) := \mathbb{E}_\nu\left[p(X)\log\frac{p(X)}{q(X)} - p(X) + q(X)\right]. \quad \text{(S2)}$$

Unlike the standard KL divergence, D_ν does not require p and q to integrate to one and therefore accommodates unnormalized candidates. We use $0\log(0/q) = 0$ and $p\log(p/0) = +\infty$ for $p > 0$. A natural measure of fixed-point error is $\mathsf{D}_\nu(\mathcal{B}_\pi^* w_i \| w_i)$, which vanishes when w_i satisfies the adjoint Bellman equation. Moreover, the adjoint Bellman operator is a contraction in generalized KL divergence (van der Laan & Kallus, 2026):

$$\mathsf{D}_\nu(\mathcal{B}_\pi^* w\|w^\pi) \le \gamma \mathsf{D}_\nu(w\|w^\pi). \quad \text{(S3)}$$

We therefore seek a critic whose adjoint Bellman moment directly controls generalized-KL error. In the Q-function case, pairing the Bellman residual with $Q - Q^\pi$ yields a criterion comparable to squared L^2 error. For occupancy ratios, the relevant error is generalized KL, which is the Bregman divergence generated pointwise by $\phi(u) = u\log u - u$:

$$\mathsf{D}_\nu(p\|q) = \mathbb{E}_\nu[\phi\{p(X)\} - \phi\{q(X)\} - \phi'\{q(X)\}\{p(X) - q(X)\}],$$

where $\phi'(u) = \log u$ is the corresponding dual coordinate (Amari & Nagaoka, 2000). The natural analogue of the Q-function error is therefore the dual-coordinate error $\phi'\{w(X)\} - \phi'\{w^\pi(X)\} = \log\{w(X)/w^\pi(X)\}$. By the Bregman three-point identity (Bauschke & Combettes, 2017),

$$\begin{aligned}
&\mathbb{E}_\nu[\{w(X) - (\mathcal{B}_\pi^* w)(X)\}\{\phi'(w(X)) - \phi'(w^\pi(X))\}] \\
&\quad = \mathsf{D}_\nu(w\|w^\pi) - \mathsf{D}_\nu(\mathcal{B}_\pi^* w\|w^\pi) + \mathsf{D}_\nu(\mathcal{B}_\pi^* w\|w) \\
&\quad \ge \mathsf{D}_\nu(w\|w^\pi) - \mathsf{D}_\nu(\mathcal{B}_\pi^* w\|w^\pi).
\end{aligned}$$

By this inequality and the contraction property of $\mathcal{B}_\pi^*$ in (S3),

$$\mathbb{E}_\nu[\{w(X) - (\mathcal{B}_\pi^* w)(X)\}\log\{w(X)/w^\pi(X)\}] \ge (1-\gamma)\mathsf{D}_\nu(w\|w^\pi).$$

This motivates the oracle criterion

$$\mathcal{S}^\star(w) := \mathbb{E}_\nu\left[w(X)\left\{\log\frac{w(X)}{w^\pi(X)} - \gamma\left(\pi\log\frac{w}{w^\pi}\right)(S')\right\}\right] - (1-\gamma)\mathbb{E}_{d_0}\left[\log\frac{w(X)}{w^\pi(X)}\right], \quad \text{(S4)}$$

which, by (S1), directly controls generalized-KL error. Unlike in the Q-function case, the adjoint contraction holds under the offline law ν, so no distributional correction or additional coercivity condition is required.

Under Condition A2, the pairwise log-ratio critics satisfy $|\log\{w_i(X)/w_j(X)\}| \le 2\log\Lambda_w$ ν-almost surely for every $i, j \in [M]$. Under this boundedness, the following two-sided comparison holds.

Lemma S1 (Oracle log-ratio equivalence)**.** *Suppose Condition A2 holds. Then, for every* $i \in \{1, \dots, M\}$,

$$(1-\gamma)\mathsf{D}_\nu(w_i\|w^\pi) \le \mathcal{S}^\star(w_i) \le 2\sqrt{2}\,\Lambda_w(1+\sqrt{\gamma})\,\mathsf{D}_\nu(w_i\|w^\pi). \tag{S5}$$

By Lemma S1, a minimizer of the oracle score has generalized-KL error within a fixed factor of the best candidate in the class. The remaining obstacle is that the oracle critic $\log(w_i/w^\pi)$ is unknown. As in the Q-function construction, we replace it by pairwise critics from the candidate class. Define $\ell_{ij} := \log\frac{w_i}{w_j}$ and the corresponding pairwise score

$$\mathcal{S}_{ij} := \mathbb{E}_\nu[w_i(X)\{\ell_{ij}(X) - \gamma(\pi\ell_{ij})(S')\}] - (1-\gamma)\mathbb{E}_{d_0}[\ell_{ij}(X)].$$

We then select the candidate with the smallest worst-case score $\max_{1\le j\le M}\mathcal{S}_{ij}$. Under realizability, if $w_{i^\star} = w^\pi$, then $\ell_{ii^\star} = \log(w_i/w^\pi)$ is the oracle critic and hence

$$\max_{1\le j\le M}\mathcal{S}_{ij} \ge \mathcal{S}_{ii^\star} = \mathcal{S}^\star(w_i) \ge (1-\gamma)\mathsf{D}_\nu(w_i\|w^\pi).$$

By contrast, $\mathcal{S}_{i^\star j} = 0$ for every j, since $w_{i^\star} = w^\pi$ satisfies the adjoint Bellman equation. Thus, minimizing the worst-case pairwise score recovers w^π under realizability.

A.2 Penalized score and selector

The remaining challenge is misspecification. When w^π is not contained in the class, pairwise log ratios only approximate the oracle critics. We therefore penalize each pairwise score to account for the error from using a comparison candidate w_j in place of w^π. We use Hellinger distance because it provides a symmetric measure of separation between the two candidate ratios that can be controlled by their generalized-KL errors relative to w^π. For nonnegative p and q, define the squared Hellinger distance

$$H^2_\nu(p,q) := \mathbb{E}_\nu\left[\{\sqrt{p(X)} - \sqrt{q(X)}\}^2\right].$$

For each ordered pair (i, j), define the penalized population score

$$\mathcal{S}^{\mathrm{pen}}_{ij} := \mathcal{S}_{ij} - \frac{1-\gamma}{4}H^2_\nu(w_i, w_j).$$

The coefficient is matched to the generalized-KL contraction constant $1-\gamma$ so that the error from replacing the oracle critic by a pairwise critic can be absorbed in the oracle comparison. The A-PBV population selector is

$$\widetilde{i}_w \in \arg\min_{1\le i\le M}\max_{1\le j\le M}\mathcal{S}^{\mathrm{pen}}_{ij}.$$

To implement A-PBV, we evaluate empirical versions of the penalized pairwise scores using held-out one-step transitions $\{(X_t, S'_t)\}_{t=1}^n$ and independent initial state–action draws $\{X_{0,t}\}_{t=1}^n$ from d_0. Algorithm S1 summarizes the resulting procedure, which requires only M^2 directly computable pairwise scores and no additional nuisance or critic fitting.

A.3 Theoretical guarantees for occupancy-ratio selection

Population oracle inequality. We first establish a population oracle inequality for A-PBV. The candidate class may be entirely misspecified, and the selected candidate is compared with the best member of the class in generalized KL.

Theorem S2 (Population A-PBV oracle inequality)**.** *Suppose Condition A2 holds. Then every population minimizer* $\widetilde{i}_w$ *satisfies*

$$\sqrt{\mathsf{D}_\nu(w_{\widetilde{i}_w}\|w^\pi)} \le \left\{\frac{4\sqrt{2}\,\Lambda_w}{1-\sqrt{\gamma}} + \frac{2}{3}\right\}\min_{1\le i\le M}\sqrt{\mathsf{D}_\nu(w_i\|w^\pi)}. \tag{S6}$$

Algorithm S1 Adjoint PBV selector for occupancy ratios

Require: Candidate ratios $w_1, \dots, w_M$; validation transitions $\{(X_t, S'_t)\}_{t=1}^n$; initial draws $\{X_{0,t}\}_{t=1}^n$; discount factor γ

1: **for** each pair $(i, j) \in \{1, \dots, M\}^2$ **do**

2: Set $\ell_{ij} := \log(w_i/w_j)$ and compute

$$\widehat{\mathcal{S}}_{ij}^{\mathrm{pen}} := \frac{1}{n}\sum_{t=1}^{n} w_i(X_t)\{\ell_{ij}(X_t) - \gamma(\pi\ell_{ij})(S'_t)\} - \frac{1-\gamma}{n}\sum_{t=1}^{n}\ell_{ij}(X_{0,t}) - \frac{1-\gamma}{4n}\sum_{t=1}^{n}\{\sqrt{w_i(X_t)} - \sqrt{w_j(X_t)}\}^2.$$

3: **end for**

4: Let $\widehat{i}_w$ be the smallest index in $\arg\min_{1\le i\le M}\max_{1\le j\le M}\widehat{\mathcal{S}}_{ij}^{\mathrm{pen}}$.

5: **return** $w_{\widehat{i}_w}$.

Finite-sample oracle inequality. For empirical validation, we additionally impose a tail condition on the target ratio and require independent draws from the initial distribution.

(C1) *Subexponential target ratio.* There exists $K_w \in (0, \infty)$ such that $\mathbb{E}_\nu\left[\exp\left\{\frac{w^\pi(X)}{K_w}\right\}\right] \le 2$.

(C2) *Independent initial-distribution draws.* Conditional on the training sample, the variables $X_{0,1}, \dots, X_{0,n}$ are independent and identically distributed according to d_0 and are independent of the validation trajectory.

Under Condition C1, the target occupancy ratio is subexponential; together with Condition A2, each candidate ratio is also subexponential. The tail condition is needed only for finite-sample concentration. If all candidates vanish off $\mathcal{X}^\pi$ and $0 < w_{\min} \le w_i, w^\pi \le w_{\max}$ on $\mathcal{X}^\pi$, then Conditions A2 and C1 hold with $\Lambda_w = w_{\max}/w_{\min}$ and $K_w = w_{\max}/\log 2$. In Condition C2, we assume access to samples from the target state–action distribution d_0 used to initialize the discounted occupancy distribution and to evaluate the adjoint Bellman moments. If the target initial-state law equals the state marginal ν_S of ν, one may draw $S_{0,t} \sim \nu_S$ from offline states independently of the validation sample and then sample $A_{0,t} \sim \pi(\cdot \mid S_{0,t})$, yielding $d_0(\mathrm{d}s, \mathrm{d}a) = \nu_S(\mathrm{d}s)\pi(\mathrm{d}a \mid s)$. If expectations under d_0 are known exactly, they may replace the empirical initial-sample averages in Algorithm S1, and Condition C2 is unnecessary.

Theorem S3 (Finite-sample A-PBV oracle inequality)**.** *Suppose Conditions A1, A2, and C1 hold, and suppose either Condition C2 holds or the required d_0 moments are known exactly. Then there is a universal constant $C_1 > 0$ such that, for every $\delta \in (0, 1)$, conditional on the training sample, with probability at least $1 - \delta$, the output of Algorithm S1 satisfies*

$$\sqrt{\mathsf{D}_\nu(w_{\widehat{i}_w}\|w^\pi)} \le \left\{\frac{4\sqrt{2}\,\Lambda_w}{1-\sqrt{\gamma}} + \frac{2}{3}\right\}\min_{1\le i\le M}\sqrt{\mathsf{D}_\nu(w_i\|w^\pi)} + \frac{C_1\Lambda_w^{3/2}\sqrt{K_w}}{1-\gamma}\sqrt{\frac{\log(4n/\delta)\log(24M/\delta)}{ng_b}}. \tag{S7}$$

The bound consists of the error of the best candidate and a validation term of order $\sqrt{\log(n/\delta)\log(M/\delta)/(ng_b)}$. Temporal dependence enters through the effective sample size ng_b, while the additional logarithmic factor reflects the subexponential, rather than bounded, target ratio. Unlike the Q-function bound, there is no weight-estimation term because the adjoint Bellman moments are directly observable for each occupancy-ratio candidate. The selector is consistent whenever the best candidate converges to w^π in generalized KL and $\log n \log M = o(ng_b)$.

A ratio-based plug-in value bound follows from Theorem S3. Specifically, if $r \in L^2(d_{\pi,\gamma})$, then, by Condition A2 and Lemma S12,

$$\left|\frac{1}{1-\gamma}\mathbb{E}_\nu\left[w_{\widehat{i}_w}(X)r(X)\right] - V^\pi\right| \le \frac{\sqrt{2\Lambda_w}}{1-\gamma}\|r\|_{2,d_{\pi,\gamma}}\sqrt{\mathsf{D}_\nu(w_{\widehat{i}_w}\|w^\pi)}.$$

B THEORETICAL GUARANTEES FOR Q-FUNCTION SELECTION

We now establish population and finite-sample guarantees for the target-aligned and unweighted PBV selectors.

B.1 REGULARITY CONDITIONS FOR THE JOINT TA-PBV GUARANTEE

For Theorem 5.1, we also impose the following conditions in addition to Conditions A1 and A2.

(J1) *Target-occupancy second moments.* The candidate functions and rewards satisfy

$$\max_{1\le i\le K} \|Q_i\|^2_{2,d_{\pi,\gamma}} + \mathbb{E}_\nu[w^\pi(X)R^2] < \infty.$$

(J2) *Bounded candidate differences and conditional TD means.* For $\varepsilon_i := Q_i(X) - R - \gamma(\pi Q_i)(S')$, there are finite constants H_Q and G_Q such that, almost surely,

$$\max_{1\le i,j\le K} |h_{ij}(X)| \le H_Q, \qquad \max_{1\le i\le K} |\mathbb{E}[\varepsilon_i \mid X]| \le G_Q.$$

(J3) *Conditional sub-Gaussian TD noise.* There is a finite constant $\sigma_Q > 0$ such that, almost surely,

$$\max_{1\le i\le K} \mathbb{E}\left[\exp\left\{\frac{(\varepsilon_i - \mathbb{E}[\varepsilon_i \mid X])^2}{\sigma_Q^2}\right\}\middle| X\right] \le 2.$$

(J4) *Uniform occupancy-ratio candidate bound.* There is a finite constant W_Q such that

$$0 \le w_a(X) \le W_Q \qquad \nu\text{-almost surely}, \qquad a = 1, \dots, M.$$

(J5) *Subexponential target ratio.* There is a finite constant K_w such that

$$\mathbb{E}_\nu\left[\exp\left\{\frac{w^\pi(X)}{K_w}\right\}\right] \le 2.$$

(J6) *Initial-distribution moments.* Either the required expectations under d_0 are known exactly, or, conditional on training, $X_{0,1}, \dots, X_{0,n}$ are independent draws from d_0 and are independent of the shared validation trajectory.

B.2 TARGET-ALIGNED PBV

Population oracle inequality. We first establish a population oracle inequality that justifies the quadratic penalty used by TA-PBV. Unlike the realizable argument in Section 3.2, we allow the entire candidate class to be misspecified and compare the selected candidate with the best member of the class. We require only the following second-moment condition.

(B1) *Second moments under the target occupancy.* The candidate functions and rewards satisfy

$$\max_{1\le i\le K} \|Q_i\|^2_{2,d_{\pi,\gamma}} + \mathbb{E}_\nu[w^\pi(X)R^2] < \infty.$$

Condition B1 holds, for example, when the rewards and candidate functions are uniformly bounded.

Theorem S4 (Population TA-PBV oracle inequality)**.** *Suppose Condition B1 holds. Then every population minimizer $\widetilde{i}_Q$ satisfies*

$$\|Q_{\widetilde{i}_Q} - Q^\pi\|_{2,d_{\pi,\gamma}} \le \left\{\frac{7}{6}\frac{1+\sqrt{\gamma}}{1-\sqrt{\gamma}} + \frac{1}{2}\right\} \min_{1\le i\le K} \|Q_i - Q^\pi\|_{2,d_{\pi,\gamma}}. \tag{S8}$$

Thus, every minimizer of the penalized population criterion has error within a factor depending only on γ of that of the best candidate, without requiring realizability. In particular, if Q^π belongs to the candidate class, every population minimizer equals Q^π in $L^2(d_{\pi,\gamma})$.

Finite-sample oracle inequality. We next account for the two additional sources of error in the practical selector: estimation of w^π and sampling variability in the pairwise scores computed on the validation sample. For each candidate, define the TD residual

$$\varepsilon_i := Q_i(X) - R - \gamma(\pi Q_i)(S').$$

The following conditions are also required for the finite-sample analysis.

(B2) *Bounded candidate differences and conditional TD means.* There are finite constants H_Q and G_Q such that, almost surely,

$$\max_{1\le i,j\le K} |h_{ij}(X)| \le H_Q, \qquad \max_{1\le i\le K} |\mathbb{E}[\varepsilon_i \mid X]| \le G_Q.$$

(B3) *Conditional sub-Gaussian TD noise.* There is a finite constant $\sigma_Q > 0$ such that, almost surely,

$$\max_{1\le i\le K} \mathbb{E}\left[\exp\left\{\frac{(\varepsilon_i - \mathbb{E}[\varepsilon_i \mid X])^2}{\sigma_Q^2}\right\}\middle| X\right] \le 2.$$

(B4) *Bounded fitted weights.* There is a finite constant W_Q such that $0 \le \widehat{w}(X) \le W_Q$, ν-almost surely.

Under Condition A1, Bernstein concentration holds for the stationary validation process with effective sample size of order ng_b. Independent validation transitions correspond to $g_b = 1$ (Jiang et al., 2025). Under Condition B2, the pairwise directions and the predictable component of the TD residual are bounded. Under Condition B3, its stochastic component is conditionally subgaussian, while the realized residual may be unbounded. If $|\varepsilon_i| \le \bar{G}_Q$ almost surely for every i, then the conditional-mean bound in Condition B2 holds with $G_Q = \bar{G}_Q$, and Condition B3 holds with $\sigma_Q = 2\bar{G}_Q/\sqrt{\log 2}$. Finally, the weighted validation scores are bounded under Condition B4, preventing large fitted weights from dominating their concentration.

Theorem S5 (Finite-sample TA-PBV oracle inequality)**.** *Suppose Condition A1 and Conditions B1–B4 hold, and suppose the Q-validation fold is independent of all data used to construct $\widehat{w}$. There are finite constants $C_{Q,\mathrm{wt}}, C_{Q,\mathrm{val}} > 0$, depending only on H_Q, G_Q, σ_Q, and W_Q, such that, for every $\delta \in (0,1)$, conditional on the training sample and the data used to construct $\widehat{w}$, with probability at least $1-\delta$, the TA-PBV selector in Algorithm 1 satisfies*

$$\begin{aligned}\|Q_{\widehat{i}_Q} - Q^\pi\|_{2,d_{\pi,\gamma}} \le{}& \left\{\frac{7}{6}\frac{1+\sqrt{\gamma}}{1-\sqrt{\gamma}} + \frac{1}{2}\right\} \min_{1\le i\le K} \|Q_i - Q^\pi\|_{2,d_{\pi,\gamma}} \\ &+ \frac{C_{Q,\mathrm{wt}}}{1-\sqrt{\gamma}}\left[\mathbb{E}_\nu\left\{\frac{\{\widehat{w}(X) - w^\pi(X)\}^2}{w^\pi(X)}\right\}\right]^{1/2} \\ &+ \frac{C_{Q,\mathrm{val}}}{1-\sqrt{\gamma}}\sqrt{\frac{\log(12K/\delta)}{ng_b}}. \end{aligned} \tag{S9}$$

Explicit choices for $C_{Q,\mathrm{wt}}$ and $C_{Q,\mathrm{val}}$ are given in Appendix H.3. The bound in Theorem S5 consists of the best-candidate error, a target-relative L^2 error for the fitted occupancy ratio, and a validation term of order $\sqrt{\log(12K/\delta)/(ng_b)}/(1-\sqrt{\gamma})$. Thus, temporal dependence enters through the effective sample size ng_b, while the class size enters only logarithmically. For fixed γ and envelope constants, the validation term vanishes whenever $\log K = o(ng_b)$. Consistency follows when the best-candidate error and relative weight error also vanish. If $\widehat{w} = w^\pi$, the weight-error term is identically zero.

A plug-in value-error bound also holds. Since $d_{\pi,\gamma} \ge (1-\gamma)d_0$, we have $c_{\mathrm{conc}} := \|\mathrm{d}d_0/\mathrm{d}d_{\pi,\gamma}\|_\infty \le (1-\gamma)^{-1}$, and Cauchy–Schwarz gives $|\mathbb{E}_{d_0}[Q_{\widehat{i}_Q}(X)] - V^\pi| \le \sqrt{c_{\mathrm{conc}}}\,\|Q_{\widehat{i}_Q} - Q^\pi\|_{2,d_{\pi,\gamma}}$.

B.3 Unweighted PBV

The unweighted selector avoids occupancy-ratio estimation but instead requires one-step concentrability and coercivity along the pairwise candidate directions under the offline law ν.

(B5) *Second moments under the offline distribution.* The true action-value function, candidate functions, and rewards satisfy

$$\|Q^\pi\|_{2,\nu}^2 + \max_{1\le i\le K} \|Q_i\|_{2,\nu}^2 + \mathbb{E}_\nu[R^2] < \infty.$$

By the one-step concentrability clause of Condition A3, the distribution shift induced by the target policy relative to the offline law ν is bounded. Pairwise coercivity is the analogue of the lower bound in (3): $I - \gamma P_\pi$ must be coercive along the candidate differences h_{ij}. By the coercivity clause, the penalty in (S86) is at least $\kappa_\nu \|h_{ij}\|_{2,\nu}^2/4$. Because coercivity is required only over the finite collection $\{h_{ij} : i, j \in [K]\}$, it can be assessed empirically from validation data.

When the candidates lie in a common linear span, pairwise coercivity is closely related to the standard positive-definiteness conditions arising in linear TD learning (Tsitsiklis & Van Roy, 1997; Amortila et al., 2026). Indeed, if $Q_i(x) = \phi(x)^\top \theta_i$, then $h_{ij}(x) = \phi(x)^\top(\theta_i - \theta_j)$, and generalized positive definiteness is required for $\mathbb{E}_\nu[\phi(X)\{\phi(X) - \gamma(P_\pi\phi)(X)\}^\top]$ relative to $\mathbb{E}_\nu[\phi(X)\phi(X)^\top]$ along the pairwise directions $\{\theta_i - \theta_j\}$. If Q^π lies in the same span, (3) is the analogous requirement along $\{\theta_i - \theta^\pi\}$. Under an arbitrary offline distribution, however, neither requirement need hold.

The corresponding population bound is given in Theorem S23 in Appendix I.1. We state here the finite-sample guarantee for the empirical selector.

Theorem S6 (Finite-sample oracle inequality for unweighted PBV)**.** *Suppose Conditions A1, A3, B2, B3, and B5 hold. There is a finite constant $C_{Q,\mathrm{unw}} > 0$, depending only on H_Q, G_Q, and σ_Q, such that, for every $\delta \in (0,1)$, conditional on the training sample, with probability at least $1-\delta$, the selector in (4) satisfies*

$$\|Q_{\widehat{i}_Q^{\mathrm{unw}}} - Q^\pi\|_{2,\nu} \le \left\{1 + \frac{2(1+\gamma\sqrt{C_{\nu,\pi}})}{\kappa_\nu}\right\} \min_{1\le i\le K} \|Q_i - Q^\pi\|_{2,\nu} + \frac{C_{Q,\mathrm{unw}}}{\kappa_\nu}\sqrt{\frac{\log(12K/\delta)}{ng_b}}. \tag{S10}$$

An explicit choice of $C_{Q,\mathrm{unw}}$ is given in Appendix I.3. The validation term has the same $\sqrt{\log K/(ng_b)}$ order as for target-aligned PBV, and there is no weight-estimation term. The trade-off is dependence on the offline coercivity constant κ_ν and an oracle comparison in $L^2(\nu)$. For fixed $\kappa_\nu > 0$ and envelope constants, consistency holds when the best-candidate error vanishes and $\log K = o(ng_b)$.

A similar plug-in value-error bound holds in the offline norm. If $d_0 \ll \nu$ and $c_{0,\nu} := \|\mathrm{d}d_0/\mathrm{d}\nu\|_{2,\nu} < \infty$, then $|\mathbb{E}_{d_0}[Q_{\widehat{i}_Q^{\mathrm{unw}}}(X)] - V^\pi| \le c_{0,\nu}\|Q_{\widehat{i}_Q^{\mathrm{unw}}} - Q^\pi\|_{2,\nu}$.

C End-to-end doubly robust OPE

This appendix extends the joint TA-PBV procedure to end-to-end OPE. Its first stage selects an occupancy-ratio candidate, and its second stage uses that ratio to select a Q-function candidate on the same validation fold. The selected nuisances are then evaluated using a doubly robust estimator on an evaluation fold.

Let the candidate classes $Q_1, \dots, Q_K$ and $w_1, \dots, w_M$ be fitted on a training sample. Let $\mathcal{D}_{\mathrm{sel}}$ and $\mathcal{D}_V$ denote the shared selection and evaluation transition folds. Let $\mathcal{D}_{0,\mathrm{sel}}$ and $\mathcal{D}_{0,V}$ denote the corresponding initial-distribution samples when those moments are evaluated empirically.

Let $\mathbb{P}_{n,V}$ denote the empirical average over the final transition fold, and let $\mathbb{P}_{n,0,V}$ denote the empirical average over $\mathcal{D}_{0,V}$. The estimator returned by Algorithm S2 is the doubly robust estimator (Jiang & Li, 2016; Kallus & Uehara, 2020):

$$\widehat{V}_{\mathrm{DR}} := \mathbb{P}_{n,0,V}\widehat{Q} + \frac{1}{1-\gamma}\mathbb{P}_{n,V}\left[\widehat{w}(X)\{R + \gamma(\pi\widehat{Q})(S') - \widehat{Q}(X)\}\right]. \tag{S11}$$

The following conditions are imposed for the end-to-end analysis.

(D1) *Validation laws.* Each transition fold follows the stationary validation setup of Section 2.1, has invariant distribution ν, and has spectral gap at least g_b. Each empirical initial-distribution sample consists of independent draws from d_0.

Algorithm S2 Sequential PBV doubly robust OPE

Require: Candidate classes $Q_1, \ldots, Q_K$ and $w_1, \ldots, w_M$ fitted only on a training sample; mutually independent transition folds $\mathcal{D}_{\mathrm{sel}}, \mathcal{D}_V$, each of size n and independent of the training sample; when d_0 moments are empirical, initial samples $\mathcal{D}_{0,\mathrm{sel}}, \mathcal{D}_{0,V}$, each of size n and independent of all other samples; discount factor γ

1: Run both stages of Algorithm 1 on $\mathcal{D}_{\mathrm{sel}}$, using $\mathcal{D}_{0,\mathrm{sel}}$ or exact d_0 moments, and set $(\widehat{w}, \widehat{Q})$ to the returned pair.

2: Evaluate $(\widehat{Q}, \widehat{w})$ on $\mathcal{D}_V$ and $\mathcal{D}_{0,V}$, or using exact d_0 moments.

3: **return** $\widehat{V}_{\mathrm{DR}}$ defined in (S11).

(D2) *Selector regularity.* Conditions B1, B2, B3, A2, and C1 hold on the selection fold, with constants $H_Q, G_Q, \sigma_Q, \Lambda_w$, and K_w as defined there.

(D3) *Uniform occupancy-ratio candidate bound.* There is a finite constant W_Q such that

$$0 \le w_j(X) \le W_Q \qquad \nu\text{-almost surely}, \qquad j = 1, \ldots, M.$$

(D4) *Initial-value evaluation.* Either $\mathbb{E}_{d_0}[Q_i(X)]$ is known exactly for every candidate, or there is a finite constant $B_{Q,0}$ such that

$$\max_{1 \le i \le K} |Q_i(X_0)| \le B_{Q,0} \qquad d_0\text{-almost surely}.$$

In the exact-moment case, set $B_{Q,0} := 0$ and interpret $\mathbb{P}_{n,0,V}\widehat{Q}$ in (S11) as $\mathbb{E}_{d_0}[\widehat{Q}(X)]$.

The finite ratio class permits a uniform analysis of both PBV stages on the shared selection fold, while the independent evaluation fold makes both selected nuisances fixed when analyzing the final estimator. For the selected ratio, Condition B4 holds by Condition D3.

For the statement below, let

$$\epsilon_w := \min_{1 \le j \le M} \sqrt{\mathsf{D}_\nu(w_j \| w^\pi)}, \quad \epsilon_Q := \min_{1 \le i \le K} \|Q_i - Q^\pi\|_{2, d_{\pi,\gamma}},$$
$$\rho_\delta := \frac{\log\{12n(K \vee M)/\delta\}}{\sqrt{n g_b}}.$$

Corollary S7 (Doubly robust oracle bound for joint TA-PBV)**.** *Run Algorithm S2 and suppose Conditions D1–D4 hold. Then, conditional on the training sample, for every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$|\widehat{V}_{\mathrm{DR}} - V^\pi| \lesssim \frac{\epsilon_w \epsilon_Q}{(1-\gamma)^3} + \frac{\epsilon_w^2}{(1-\gamma)^4} + \left\{ \frac{\epsilon_Q}{(1-\gamma)^3} + \frac{\epsilon_w}{(1-\gamma)^4} + \frac{1}{1-\gamma} \right\} \rho_\delta + \frac{\rho_\delta^2}{(1-\gamma)^4}. \tag{S12}$$

Here $\lesssim$ suppresses only universal constants and the fixed structural constants in Conditions D2–D4.

The first two terms are the product and squared occupancy component of the second-order nuisance remainder. The third term collects the $1/\sqrt{n g_b}$ validation errors and their interactions with approximation error, while the fourth is the resulting $1/(n g_b)$ term. Reuse of the final evaluation data for selection is not covered; establishing a cross-fitted analogue would require an additional dependence argument.

D Additional experimental details

D.1 Benchmark designs

Broad offline benchmarks. For the Hopper benchmark of Liu et al. (2025), we use the four model-free perturbation experiments, each with ten target-policy checkpoints evaluated under three environment settings, giving 120 environment–policy combinations. For Deep OPE, we use all eleven target policies of Fu et al. (2021) with the random, medium, medium-replay, medium-expert, and expert D4RL-v2 datasets (Fu et al., 2020) for HalfCheetah, Hopper, Walker2d, and Ant, giving $4 \times 5 \times 11 = 220$ environment–dataset–policy combinations.

Uniform-mixture controlled overlap. The controlled benchmark uses the same logging-policy construction for Hopper and Deep OPE. If $\pi_1, \ldots, \pi_L$ are the benchmark policies, the logging policy draws J uniformly from $\{1, \ldots, L\}$ at the start of each trajectory and then follows π_J for the entire trajectory. Each π_j is evaluated in turn as the target against this common logging distribution. Since $\mathbb{P}_{\log} = L^{-1} \sum_{j=1}^{L} \mathbb{P}_{\pi_j}$, the target-to-logging trajectory density ratio is bounded by L. The same bound holds after marginalization for the discounted occupancy ratio. Thus, every target satisfies a known overlap bound under the same logging distribution.

For Hopper, $L = 10$, and all ten targets are evaluated in the gravity and action-noise perturbation experiments, each under three environment settings, giving 60 environment–target combinations. For Deep OPE, $L = 11$, and all eleven targets are evaluated in each of four environments, giving 44 combinations. We use five independent occupancy-ratio data replicates for each combination. The occupancy-ratio training and validation samples are independent of the samples used to fit and select the Q-function. Within each environment–target combination, the fitted Q-function class and its validation sample are held fixed across the five occupancy-ratio replicates. Each training or validation sample contains at least $51{,}200$ transitions. Deep OPE splits preserve complete trajectories and contain at least three trajectories from every mixture component.

D.2 Candidate classes and selectors

Q-function candidates. Broad Hopper uses the 15 candidate Q-functions from Liu et al. (2025). These candidates are induced by a grid of perturbed Hopper simulators, with either gravity or transition-noise level varied across the grid. For each target policy, candidate Q-values are computed by Monte Carlo rollouts in the corresponding simulator rather than by fitting FQE. For Deep OPE and controlled Hopper, we fit a 15-candidate neural FQE grid on an independent sample containing at least $51{,}200$ transitions.[1] Each Q-network is a multilayer perceptron with two SiLU hidden layers of width 64, 128, or 256. For each width, checkpoints are saved after 50, 100, 200, 400, and 800 fitted-Q iterations, giving 15 candidates in total. Each iteration comprises 20 AdamW updates with minibatches of size 512, learning rate 3×10^{-4}, weight decay 10^{-4}, and Huber loss. Inputs are standardized, the target network is updated after each fitted-Q iteration, and no early stopping is used. All fits use $\gamma = 0.99$. Deep OPE splits preserve complete trajectories. For stochastic target policies, conditional action expectations are approximated using eight draws per transition and 32 draws per reset state.

Occupancy-ratio candidates. TA-PBV uses neural FORE estimators of the discounted target-to-logging occupancy ratio. The log ratio is represented by a ReLU network with either one or two 64-unit hidden layers. Each architecture is trained full-batch with AdamW at learning rate 3×10^{-4} or 10^{-3} and weight decay 10^{-4}. Snapshots after 25, 50, 100, 200, 400, 800, and 1,600 updates give 28 FORE candidates. A-PBV selects among these candidates on an independent occupancy-ratio validation sample. Before validation, fitted ratios are floored at 10^{-6} and rescaled on the training sample to satisfy the empirical constant adjoint moment. The selected ratio estimator is then held fixed when TA-PBV selects the Q-function. For the primary analysis, we temper the selected ratio as $w_T = w^{1/T} / \mathbb{P}_{n,w}[w^{1/T}]$ with $T = 5$, where the normalizing constant is computed on the ratio-training sample and then held fixed. Target-policy returns and selected Q-function performance are not used in occupancy-ratio selection. For broad Hopper, the ratio-training and ratio-validation samples each contain $3{,}200$ transitions; for broad Deep OPE, each contains $25{,}600$ transitions. The controlled-overlap comparisons use $51{,}200$ transitions in each sample.

LSTD-Tournament comparator. For candidate i, let $\mathbb{P}_{n,Q}$ denote the empirical average over the Q-function validation sample and write $\varepsilon_i = Q_i(X) - R - \gamma(\pi Q_i)(S')$. We implement the normalized-difference criterion of Liu et al. (2025):

$$\widehat{i}_{\mathrm{LSTD}} \in \arg\min_{1 \le i \le K} \max \left\{ \frac{|\mathbb{P}_{n,Q}[Q_i(X)\varepsilon_i]|}{\widehat{\sigma}_n(Q_i)}, \max_{j \neq i} \frac{|\mathbb{P}_{n,Q}[(Q_i - Q_j)(X)\varepsilon_i]|}{\widehat{\sigma}_n(Q_i - Q_j)} \right\},$$

[1] The HalfCheetah and Walker2d medium-replay datasets contain fewer than $102{,}400$ usable transitions. For these broad Deep OPE settings, we form disjoint trajectory pools and complete each nominal block by sampling complete trajectories within its pool; at least 90% of each block consists of distinct source rows.

where $\widehat{\sigma}_n(f)^2 = \mathbb{P}_{n,Q}[(f - \mathbb{P}_{n,Q} f)^2]$. Unlike PBV, the criterion uses absolute moments and no quadratic penalty. The empirical variance uses divisor n, following Liu et al. (2025). The criterion is undefined if any required empirical standard deviation is zero, and ties are resolved by candidate order. We apply this closed-form criterion directly to the fitted candidate predictions in both benchmarks; no pair-specific LSTDQ model is fitted.

D.3 Validation and evaluation

The main comparisons use $n_Q = 3{,}200$. We additionally evaluate the controlled-overlap selectors at $n_Q \in \{3{,}200, 6{,}400, 12{,}800, 25{,}600, 51{,}200\}$. Hopper uses nested prefixes of a fixed random permutation, so the largest sample uses the full validation pool in each of five independent occupancy-ratio replicates. Deep OPE samples complete trajectories with replacement until reaching the nominal size and uses 96 validation resamples per setting.

The 90% intervals in Table S1 use 4,000 percentile-bootstrap draws over the benchmark hierarchy after averaging over the inner validation resamples. For Hopper, the hierarchy consists of perturbation condition, target policy, and occupancy-ratio replicate when present. For Deep OPE, it consists of environment–dataset combination (or environment under controlled overlap), target policy, and occupancy-ratio replicate when present. Each level receives equal weight.

For fitted FQE candidates, the reset-value estimate averages $Q(S_0, A_0)$ over $51{,}200$ independent reset-state draws, integrating over target actions as described above. Reference values are supplied by the original broad Hopper benchmark and otherwise are Monte Carlo means from 1,024 target-policy trajectories. Hopper and Deep OPE trajectories have horizons 1,024 and 1,000, respectively, and end at termination. Value errors are reported on the original return scale.

For episodic tasks, discounted occupancies are defined over preterminal state–action pairs, with no additional pairs contributed after termination.

In all four benchmark settings, we evaluate the L^2 error of the selected Q-function under the discounted target occupancy in addition to absolute policy-value error. We draw $N_{\mathrm{eval}} = 4{,}096$ state–action pairs $X_j = (S_j, A_j)$ independently from the discounted target occupancy and generate two conditionally independent continuation returns $G_j^{(1)}$ and $G_j^{(2)}$ from each X_j. We estimate the target-occupancy Q-MSE by

$$\widehat{\mathcal{E}}_Q(\widehat{Q}) = \frac{1}{N_{\mathrm{eval}}} \sum_{j=1}^{N_{\mathrm{eval}}} \{\widehat{Q}(X_j) - G_j^{(1)}\}\{\widehat{Q}(X_j) - G_j^{(2)}\}.$$

Because $\mathbb{E}[G_j^{(k)} \mid X_j] = Q^\pi(X_j)$ and the two continuation returns are conditionally independent,

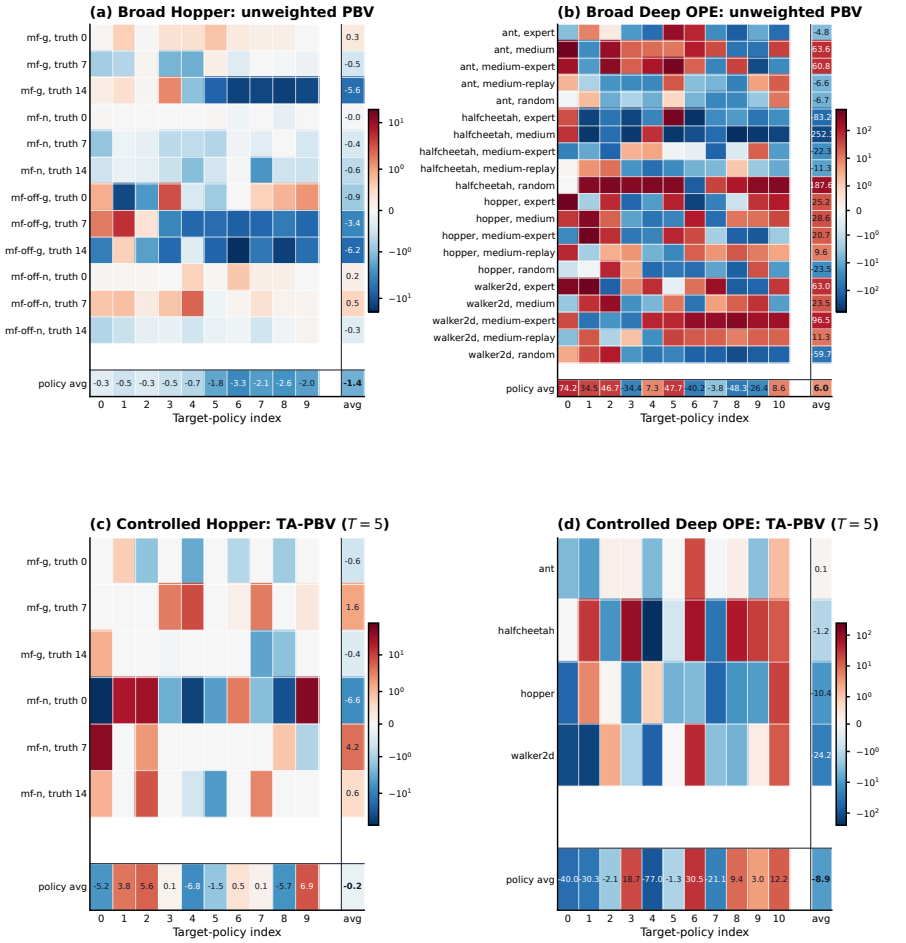
$$\mathbb{E}\Big[\{\widehat{Q}(X) - G^{(1)}\}\{\widehat{Q}(X) - G^{(2)}\} \mid X\Big] = \{\widehat{Q}(X) - Q^\pi(X)\}^2.$$

Thus, conditional on the selected Q-function, $\widehat{\mathcal{E}}_Q(\widehat{Q})$ is unbiased for its squared L^2 error under the discounted target occupancy. Table 1 reports the square root of the mean of these estimates over benchmark settings and validation resamples when present, denoted target Q-RMSE. We take the square root after averaging because individual unbiased MSE estimates can be negative due to Monte Carlo error; the corresponding mean MSE is the square of the reported entry. Target-policy simulation is used only for evaluation after model selection. For broad Hopper, the same target-occupancy evaluation sample is used for the corresponding on- and off-policy logging conditions. For broad Deep OPE, the same environment–target evaluation sample is used across the five D4RL dataset qualities.

D.4 Supplementary analyses

TA-PBV has the lowest mean value error at every validation sample size on both controlled-overlap benchmarks. Its mean error decreases from 14.5 to 12.8 on Hopper and from 69.3 to 52.4 on Deep OPE as n_Q increases from $3{,}200$ to $51{,}200$.

Target-component coverage. To vary coverage directly while holding the Q-class fixed, we replace each validation transition with an independent draw from the discounted target occupancy



Each tile is a prespecified comparison; blue (negative) favors PBV and red (positive) favors LSTD. Right and bottom margins are row and policy means. Color scales are panel-specific symmetric logs.

Figure S1: Policy-level difference in absolute value error relative to LSTD-Tournament; negative values favor PBV. Broad panels use $n_Q = 3{,}200$, and controlled-overlap panels use $n_Q = 51{,}200$. Panels (a)–(b) show unweighted PBV on the broad benchmarks, and panels (c)–(d) show TA-PBV with $T = 5$ under controlled overlap. Margins show means across rows and target policies.

Table S1: Controlled-overlap absolute value error across the Q-validation sample-size grid (mean [90% hierarchical bootstrap interval]). Lower is better; boldface marks the smallest mean in each column.

	n_Q				
Selector	3,200	6,400	12,800	25,600	51,200
Hopper					
Unweighted PBV	25.5 [11.9,43.9]	24.8 [11.3,42.7]	23.2 [10.6,41.6]	22.8 [10.6,39.9]	22.2 [9.7,40.0]
Unweighted PBV (no penalty)	15.1 [8.1,24.3]	14.5 [7.3,24.5]	14.3 [7.3,23.3]	13.9 [6.9,23.2]	13.9 [6.5,23.0]
TA-PBV ($T = 5$)	**14.5 [8.0,22.4]**	**13.1 [7.0,20.8]**	**13.1 [6.8,20.5]**	**12.8 [6.8,20.0]**	**12.8 [6.8,20.2]**
LSTD-Tournament	15.9 [7.2,27.5]	14.1 [6.3,24.6]	13.4 [6.1,23.0]	13.3 [5.9,23.3]	13.0 [5.7,22.8]
Squared TD error	15.5 [8.4,22.9]	15.7 [8.7,23.6]	15.4 [8.8,23.1]	15.4 [8.2,23.1]	15.5 [8.4,23.1]
Deep OPE					
Unweighted PBV	83.2 [36.7,139.3]	79.3 [35.7,130.8]	75.4 [34.2,126.0]	72.8 [33.1,120.1]	71.2 [32.9,117.8]
Unweighted PBV (no penalty)	75.9 [30.1,136.9]	71.4 [29.4,126.6]	66.5 [29.1,117.1]	62.8 [27.7,109.4]	61.2 [26.8,106.7]
TA-PBV ($T = 5$)	**69.3 [27.5,129.5]**	**62.8 [25.2,115.4]**	**58.1 [24.1,105.2]**	**55.0 [23.0,100.4]**	**52.4 [22.8,94.1]**
LSTD-Tournament	74.7 [30.4,135.5]	70.1 [29.5,126.9]	66.0 [29.3,116.0]	63.6 [28.8,109.3]	61.3 [28.5,105.2]
Squared TD error	177.8 [46.7,354.1]	178.5 [46.4,347.0]	179.0 [46.9,368.2]	179.6 [47.4,366.5]	179.6 [47.2,353.0]

Table S2: TA-PBV temperature sensitivity under controlled overlap at $n_Q = 3{,}200$. Value-error entries are means over benchmark settings and validation resamples. ESS entries are median (minimum) relative effective sample sizes, ESS/n, across benchmark settings. The $T = 5$ specification follows Sinha et al. (2022).

	Value error		ESS fraction	
T	Hopper	Deep OPE	Hopper	Deep OPE
1	16.0	102.2	.099 (<.001)	.022 (<.001)
2.5	15.7	76.6	.106 (.066)	.447 (<.001)
5	14.0	69.8	.198 (.167)	.757 (<.001)
7.5	13.7	68.9	.409 (.387)	.868 (.002)
10	14.0	69.4	.610 (.571)	.918 (.053)
∞	16.0	74.4	1.000 (1.000)	1.000 (1.000)

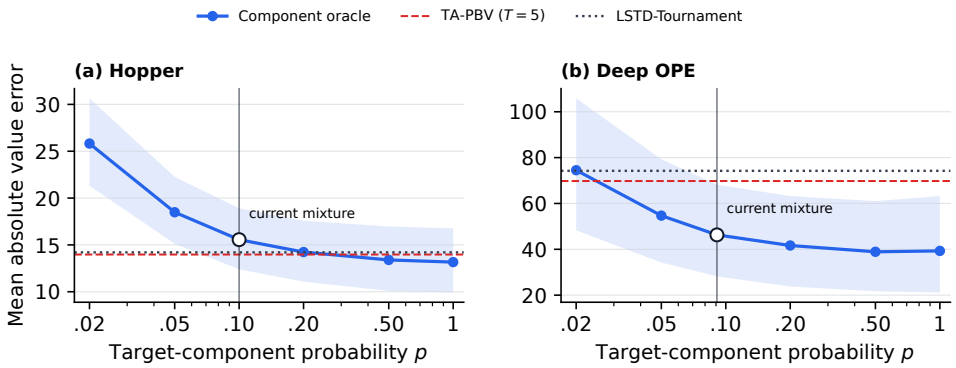


Figure S2: Value error as the target-policy component of the validation distribution increases. Bands are 95% intervals clustered over the 60 Hopper or 44 Deep OPE settings. Vertical lines mark the target-component probability in the controlled-overlap benchmark; horizontal lines show TA-PBV with $T = 5$ under the original logging distribution and LSTD-Tournament.

with probability $p \in \{.02, .05, .10, .20, .50, 1\}$ for Hopper; for Deep OPE, $1/11$ replaces .10. Time within each target-policy trajectory is sampled proportional to γ^t. Evaluation quantities are not used for selection. Value error decreases as the target component of the validation distribution increases and changes little beyond $p = .5$, with a larger reduction on Deep OPE (Figure S2).

E Additional related work

Section 1.1 compares PBV with existing Bellman validation methods and LSTD-Tournament. Here we place the problem in the broader literatures on OPE model selection, occupancy-ratio estimation, supervised aggregation, and statistical inverse problems.

Model selection for Q-functions and occupancy ratios. Existing OPE model-selection methods have focused primarily on Q-functions, replacing the unavailable supervised loss with structured Bellman criteria. BErMin estimates Bellman error with complexity penalties; BVFT and BVFT-PE use projected Bellman errors over candidate-dependent state aggregations; and ModBE uses relative square-loss tests for nested model classes (Farahmand & Szepesvári, 2011; Xie & Jiang, 2021; Zhang & Jiang, 2021; Lee et al., 2022). Supervised Bellman validation instead fits a one-step-target regression for each candidate, while Duality-based Residual Estimation learns residual critics and selects among critic configurations on an additional split (Zitovsky et al., 2023; Miyaguchi, 2026). PBV instead uses candidate-induced critics directly, avoiding candidate-dependent projection, nestedness, and auxiliary critic fitting. Model selection for occupancy ratios has received substantially less attention, despite extensive work on estimating them through DICE, minimax weight learning, and fitted occupancy-ratio methods (Nachum et al., 2019; Uehara et al., 2020; 2021; van der Laan

& Kallus, 2026). A-PBV addresses the distinct post-hoc problem of selecting among already-fitted ratio candidates.

Model-selection targets in OPE. Model selection arises at several levels in off-policy evaluation. One line of work selects or combines scalar value estimates, including MAGIC, SWITCH, shrinkage estimators, SLOPE, proxy-task and cross-validation methods, and OPERA (Thomas & Brunskill, 2016; Wang et al., 2017; Su et al., 2020a;b; Udagawa et al., 2023; Cief et al., 2025; Nie et al., 2024). A second line uses OPE to rank policies or tune complete offline-RL pipelines (Irpan et al., 2019; Paine et al., 2020; Tang & Wiens, 2021; Yang et al., 2022; Nie et al., 2022). Zhang & Jiang (2021) primarily study offline policy selection but also use BVFT-PE to select among candidate Q^π estimates for a fixed policy. We instead study fixed-policy nuisance selection: from independently fitted finite classes, we select the Q-function or occupancy-ratio estimate used for OPE. This differs from selecting scalar value estimates or policies because neither nuisance is observed, so their target losses are not directly available.

Occupancy-ratio estimation and selection. Marginalized importance sampling, DICE methods, and minimax weight learning estimate discounted occupancy ratios from adjoint Bellman equations, typically using prescribed ratio and critic classes (Xie et al., 2019; Nachum et al., 2019; Zhang et al., 2020; Uehara et al., 2020; 2021). FORE is an occupancy-ratio analogue of FQE that estimates the ratio through fitted adjoint Bellman recursion without requiring Bellman completeness (van der Laan & Kallus, 2026), while FORC is a related finite-horizon method that does require Bellman completeness (Huang et al., 2023; Huang & Jiang, 2024). These methods construct an occupancy-ratio estimate within a specified function class. A-PBV instead selects post hoc among already-fitted candidates using pairwise log-ratio critics and Hellinger-penalized adjoint Bellman scores.

Supervised model selection and aggregation. In supervised prediction, held-out outcomes make the risk of a fixed candidate directly estimable. Cross-validation uses this risk to select algorithms or tuning parameters, while stacking and Super Learner combine candidates using cross-validated predictions (Stone, 1974; Wolpert, 1992; van der Laan & Dudoit, 2003; van der Laan et al., 2007). Model-selection aggregation instead seeks to match the best element of a finite dictionary; under squared loss, Q-aggregation attains sharp oracle inequalities with logarithmic remainders (Dai et al., 2012; Lan & Syrgkanis, 2024). These supervised criteria are unavailable here because neither Q^π nor w^π is observed, squared temporal-difference error is not a valid target loss, and offline-distribution Bellman moments need not control target-policy error. PBV instead constructs observable pairwise criteria and selects a supplied candidate rather than forming an aggregate. Its oracle inequalities account for misspecification and temporal dependence, as well as occupancy-ratio error for TA-PBV or pairwise coercivity for unweighted PBV.

Model selection in linear inverse problems. Related model-selection challenges arise in statistical linear inverse problems, where instability of the inverse operator motivates operator-dependent risk estimates and penalties (Cavalier et al., 2002; Abramovich et al., 2018). For example, Regularized DeepIV uses an out-of-sample Tikhonov criterion to obtain oracle guarantees under misspecification in nonparametric instrumental-variable regression (Li et al., 2024). In general, small forward error need not control structural error without regularization and source or link conditions. The Bellman structures considered here are better posed in the relevant target geometries: the discounted target-policy transition operator is contractive in target-occupancy L^2, while the adjoint Bellman update is contractive in generalized KL. These contractions convert pairwise Bellman moments into target-error bounds without a Tikhonov bias–variance choice or generic inverse-problem source conditions. This does not eliminate structural assumptions: TA-PBV requires coverage and occupancy-ratio accuracy, A-PBV requires ratio comparability and tail control, and unweighted PBV requires pairwise coercivity under the offline distribution.

F Concentration along the validation trajectory

The dependence argument used for all three selectors combines Bernstein concentration for bounded additive functionals of a Markov chain with martingale concentration for the one-step transition innovation. Let $\mathcal{F}_t := \sigma(X_1, R_1, S'_1, \ldots, X_{t-1}, R_{t-1}, S'_{t-1}, X_t)$. Under the validation setup, conditional on $\mathcal{F}_t$, the pair (R_t, S'_t) follows the MDP reward and transition law given X_t.

Lemma S8 (Trajectory Bernstein bounds)**.** *Let $U_1, \dots, U_n$ be a stationary Markov chain with invariant law μ and transition operator P_U. Suppose, for some $g \in (0,1]$, $\|P_U f\|_{2,\mu} \le (1-g)\|f\|_{2,\mu}$ for every centered real-valued $f \in L^2(\mu)$. Let $(\mathcal{F}_t)_{t=1}^{n+1}$ be a filtration such that U_t is $\mathcal{F}_t$-measurable and Y_t is $\mathcal{F}_{t+1}$-measurable. Suppose there is a Markov kernel K_Y, not depending on t, such that $\mathcal{L}(Y_t \mid \mathcal{F}_t) = K_Y(U_t, \cdot)$ for every t. For $u \ge 1$, set $r_g(u) := \sqrt{u/(ng)}$.*

(i) If $|Y_t| \le B$ almost surely and $\mathbb{E}(Y_t^2) \le v^2$, then, with probability at least $1 - 6e^{-u}$,

$$\left| \frac{1}{n} \sum_{t=1}^{n} Y_t - \mathbb{E} Y_t \right| \le C_{\mathrm{M}} \{ v r_g(u) + B r_g(u)^2 \}. \tag{S13}$$

(ii) Suppose $Y_t = a(U_t)\{d(U_t) + \zeta_t\}$, where $|a| \le B$, $|d| \le D$, $\mathbb{E}_\mu(a^2) \le v^2$, $\sigma > 0$, and

$$\mathbb{E}(\zeta_t \mid \mathcal{F}_t) = 0, \qquad \mathbb{E}\left[\exp\left\{ \frac{\zeta_t^2}{\sigma^2} \right\} \middle| \mathcal{F}_t \right] \le 2. \tag{S14}$$

Then, with probability at least $1 - 6e^{-u}$,

$$\left| \frac{1}{n} \sum_{t=1}^{n} Y_t - \mathbb{E} Y_t \right| \le C_{\mathrm{M}} (D + \sigma) \{ v r_g(u) + B r_g(u)^2 \}. \tag{S15}$$

Here $C_{\mathrm{M}} > 0$ is a universal constant.

Proof of Lemma S8. We use the following consequence of Theorem 1 of Jiang et al. (2025). If f is a bounded, time-independent function and $\mathbb{E}_\mu f = 0$, then, for a universal constant $C > 0$,

$$\left| \frac{1}{n} \sum_{t=1}^{n} f(U_t) \right| \le C \left\{ \|f\|_{2,\mu} \sqrt{\frac{u}{ng}} + \|f\|_\infty \frac{u}{ng} \right\} \tag{S16}$$

outside an event of probability $2e^{-u}$. For the proof, also write $r_1(u) := \sqrt{u/n}$.

For part (i), let $m(x) := \int y \, K_Y(x, \mathrm{d}y)$ and $\eta_t = Y_t - m(U_t)$. The centered function $m - \mathbb{E} Y_t$ has variance at most v^2 and absolute value at most $2B$. Thus, by (S16), its empirical average is bounded by $C\{v r_g(u) + B r_g(u)^2\}$. Next, define $q(x) := \int \{y - m(x)\}^2 K_Y(x, \mathrm{d}y)$. By conditional variance contraction,

$$0 \le q \le B^2, \qquad \mathbb{E}_\mu q \le v^2, \qquad \mathrm{Var}_\mu(q) \le B^2 v^2.$$

By (S16), applied to $q - \mathbb{E}_\mu q$, outside an event of probability $2e^{-u}$,

$$\frac{1}{n} \sum_{t=1}^{n} q(U_t) \le C\{v + B r_g(u)\}^2. \tag{S17}$$

The sequence $\{\eta_t\}$ is a martingale difference with conditional variance process $\{q(U_t)\}$ and $|\eta_t| \le 2B$. By the two-sided form of Freedman's inequality (Freedman, 1975) and (S17),

$$\left| \frac{1}{n} \sum_{t=1}^{n} \eta_t \right| \le C\{v r_1(u) + B r_1(u) r_g(u) + B r_1(u)^2\}$$

outside an event of probability $4e^{-u}$. After combining this bound with the bound for the predictable component, the sharper intermediate bound is

$$C\{v r_g(u) + B r_g(u)^2 + v r_1(u) + B r_1(u) r_g(u) + B r_1(u)^2\}.$$

Since $r_1(u) \le r_g(u)$ for $g \le 1$, (S13) follows.

For part (ii), apply (S16) first to the centered function ad. Its variance is at most $D^2 v^2$ and its absolute value is at most $2BD$, so its empirical average is bounded by $CD\{v r_g(u) + B r_g(u)^2\}$. Apply (S16) a second time to $a^2 - \mathbb{E}_\mu(a^2)$. Since $\mathbb{E}_\mu(a^4) \le B^2 v^2$,

$$\frac{1}{n} \sum_{t=1}^{n} a(U_t)^2 \le C\{v + B r_g(u)\}^2 \tag{S18}$$

outside an event of probability $2e^{-u}$. Under (S14), for every integer $k \geq 2$,

$$\mathbb{E}\big(|a(U_t)\zeta_t|^k \,\big|\, \mathcal{F}_t\big) \leq \frac{k!}{2} C^2\sigma^2 a(U_t)^2 (C\sigma B)^{k-2}.$$

By the martingale Bernstein inequality under this conditional moment condition (de la Peña, 1999) and (S18),

$$\left|\frac{1}{n}\sum_{t=1}^{n} a(U_t)\zeta_t\right| \leq C\sigma\{vr_1(u) + Br_1(u)r_g(u) + Br_1(u)^2\}$$

outside an event of probability $4e^{-u}$. After adding the predictable bound, the sharper intermediate bound is

$$C\left[D\{vr_g(u) + Br_g(u)^2\} + \sigma\{vr_1(u) + Br_1(u)r_g(u) + Br_1(u)^2\}\right].$$

Because $r_1(u) \leq r_g(u)$, (S15) follows. □

G Proofs for occupancy-ratio model selection

The proof follows the main-text argument. By generalized-KL contraction and the three-point identity, the oracle log-ratio moment is at least $(1-\gamma)\mathsf{D}_\nu(w\|w^\pi)$. Because $H_\nu(p,q) = \|\sqrt{p} - \sqrt{q}\|_{2,\nu}$ and $H_\nu^2(p,q) \leq \mathsf{D}_\nu(p\|q)$, the pairwise Hellinger distance is bounded by the sum of the two square-root generalized-KL errors. Under Condition A2, log-ratio differences are bounded in terms of this distance by (S24); the pairwise score and its leading validation deviation then satisfy the linear bounds (S33) and (S44). The quadratic Hellinger penalty absorbs these terms; population optimality and localized concentration then reduce the oracle inequalities to quadratic comparisons.

G.1 Generalized-KL contraction and adjoint residual identities

The following contraction was established in the FORE analysis of van der Laan & Kallus (2026). We include the short generalized-KL argument for completeness.

Lemma S9 (Generalized-KL contraction of the adjoint update)**.** *Let w and $\widetilde{w}$ be nonnegative integrable functions whose adjoint updates exist as ν-densities. Then*

$$\mathsf{D}_\nu(\mathcal{B}_\pi^* w\|\mathcal{B}_\pi^*\widetilde{w}) \leq \gamma\mathsf{D}_\nu(w\|\widetilde{w}), \tag{S19}$$

whenever the divergence on the right is finite.

Proof of Lemma S9. Write D for generalized KL between finite nonnegative measures. This divergence is jointly convex and obeys the data-processing inequality for Markov kernels. Let $\mu = w\nu$ and $\widetilde{\mu} = \widetilde{w}\nu$. By (2), joint convexity, and data processing,

$$\begin{aligned}\mathsf{D}_\nu(\mathcal{B}_\pi^* w\|\mathcal{B}_\pi^*\widetilde{w}) &= \mathsf{D}((1-\gamma)d_0 + \gamma\mu P_\pi \,\|\, (1-\gamma)d_0 + \gamma\widetilde{\mu}P_\pi) \\ &\leq (1-\gamma)\mathsf{D}(d_0\|d_0) + \gamma\mathsf{D}(\mu P_\pi\|\widetilde{\mu}P_\pi) \\ &\leq \gamma\mathsf{D}(\mu\|\widetilde{\mu}) = \gamma\mathsf{D}_\nu(w\|\widetilde{w}),\end{aligned}$$

which is (S19). □

Lemma S10 (Generalized-KL three-point identity for the adjoint residual)**.** *Let w and v be nonnegative integrable functions that are positive on a common measurable support and vanish ν-almost surely off that support. Set $\log(w/v) = 0$ on their common-zero set, and suppose $w^+ := \mathcal{B}_\pi^* w$ exists as a nonnegative ν-density. If every divergence in (S20) is finite, then*

$$\mathbb{E}_\nu\left[(w - w^+)\log\frac{w}{v}\right] = \mathsf{D}_\nu(w\|v) - \mathsf{D}_\nu(w^+\|v) + \mathsf{D}_\nu(w^+\|w). \tag{S20}$$

In particular, if $v = w^\pi$, then

$$\mathcal{S}^\star(w) \geq (1-\gamma)\mathsf{D}_\nu(w\|w^\pi). \tag{S21}$$

Proof of Lemma S10. Finiteness of $\mathsf{D}_\nu(w^+\|v)$ implies that w^+ vanishes off the common support of w and v. Every integrand below therefore vanishes there. Expanding the three generalized-KL terms and canceling their linear parts gives

$$\begin{aligned}\mathsf{D}_\nu(w\|v) - \mathsf{D}_\nu(w^+\|v) + \mathsf{D}_\nu(w^+\|w) &= \mathbb{E}_\nu\left[w\log\frac{w}{v} - w^+\log\frac{w^+}{v} + w^+\log\frac{w^+}{w}\right] \\ &= \mathbb{E}_\nu\left[(w-w^+)\log\frac{w}{v}\right].\end{aligned} \tag{S22}$$

The final expression in (S22) is the left-hand side of (S20), so that identity follows. To derive (S21), set $v = w^\pi$. By (S1) and (S4), this expression is $\mathcal{S}^\star(w)$. Now use $\mathcal{B}_\pi^* w^\pi = w^\pi$ in Lemma S9. By that lemma, the middle divergence in (S20) is at most $\gamma\mathsf{D}_\nu(w\|w^\pi)$, while the final divergence is nonnegative. Thus, (S21) follows. □

G.2 Target-relative ratio bounds

Write

$$e_i := \sqrt{\mathsf{D}_\nu(w_i\|w^\pi)}, \qquad H_{ij} := H_\nu(w_i, w_j),$$

throughout the remainder of this appendix.

Lemma S11 (Relative generalized-KL inequalities)**.** *Let $\Lambda \geq 1$. For every $z \in [\Lambda^{-1}, \Lambda]$,*

$$(z-1)^2 \leq 2\Lambda\{z\log z - z + 1\}. \tag{S23}$$

For $z, \widetilde{z} \in [\Lambda^{-1}, \Lambda]$,

$$|\log z - \log\widetilde{z}| \leq 2\sqrt{\Lambda}|\sqrt{z} - \sqrt{\widetilde{z}}|. \tag{S24}$$

Finally, for $p, q \geq 0$,

$$(\sqrt{p} - \sqrt{q})^2 \leq p\log\frac{p}{q} - p + q \tag{S25}$$

whenever the right-hand side is finite.

Proof of Lemma S11. The function $\phi(z) = z\log z - z + 1$ satisfies $\phi''(z) = 1/z \geq 1/\Lambda$ on $[\Lambda^{-1}, \Lambda]$. By the strong-convexity inequality at $z = 1$, (S23) holds. By the mean-value theorem applied to $\log u$ on $u \in [\Lambda^{-1/2}, \Lambda^{1/2}]$, together with $\log z - \log\widetilde{z} = 2(\log\sqrt{z} - \log\sqrt{\widetilde{z}})$, (S24) holds.

For (S25), set $t = p/q$ when $q > 0$. After division by q, the difference between the right- and left-hand sides is

$$t\log t - 2t + 2\sqrt{t} = 2\sqrt{t}\{\sqrt{t}\log\sqrt{t} - \sqrt{t} + 1\} \geq 0.$$

When $q > 0$, the case $p = 0$ follows by continuity. When $q = 0$, finiteness of the right-hand side requires $p = 0$ under the convention in (S2), and both sides vanish. □

Lemma S12 (Occupancy-ratio plug-in value bound)**.** *Suppose Condition A2 holds and $r \in L^2(d_{\pi,\gamma})$. Then, for every $i \in \{1, \ldots, M\}$,*

$$\left|\frac{1}{1-\gamma}\mathbb{E}_\nu[w_i(X)r(X)] - V^\pi\right| \leq \frac{\sqrt{2\Lambda_w}}{1-\gamma}\|r\|_{2,d_{\pi,\gamma}}\sqrt{\mathsf{D}_\nu(w_i\|w^\pi)}. \tag{S26}$$

Proof of Lemma S12. Fix $i \in \{1, \ldots, M\}$ and set $z_i = w_i/w^\pi$ on $\mathcal{X}^\pi$. By Condition A2, both w_i and w^π vanish off $\mathcal{X}^\pi$. By the occupancy representation of V^π and the identity $d_{\pi,\gamma} = w^\pi\nu$,

$$\left|\frac{1}{1-\gamma}\mathbb{E}_\nu[w_i(X)r(X)] - V^\pi\right| = \frac{1}{1-\gamma}\left|\mathbb{E}_{d_{\pi,\gamma}}[(z_i(X)-1)r(X)]\right|.$$

By Cauchy–Schwarz and (S23),

$$\begin{aligned}\left|\mathbb{E}_{d_{\pi,\gamma}}[(z_i(X)-1)r(X)]\right| &\leq \|r\|_{2,d_{\pi,\gamma}}\|z_i - 1\|_{2,d_{\pi,\gamma}}, \\ \|z_i - 1\|_{2,d_{\pi,\gamma}}^2 &\leq 2\Lambda_w\mathbb{E}_{d_{\pi,\gamma}}[z_i\log z_i - z_i + 1] \\ &= 2\Lambda_w\mathsf{D}_\nu(w_i\|w^\pi).\end{aligned}$$

Combining these inequalities proves (S26). □

Proof of Lemma S1. Fix $i \in \{1, \dots, M\}$ and abbreviate $w = w_i$. The comparison with w^π is preserved by the adjoint update. Indeed,

$$\Lambda_w^{-1} w^\pi \nu \le (1-\gamma) d_0 + \gamma (w\nu) P_\pi \le \Lambda_w w^\pi \nu, \tag{S27}$$

because $w^\pi \nu = (1-\gamma) d_0 + \gamma (w^\pi \nu) P_\pi$ and the corresponding bounds on $w\nu$ relative to $w^\pi \nu$ hold by Condition A2. Thus, the middle measure in (S27) admits a ν-density $w^+ := \mathcal{B}_\pi^* w$. The ratios w/w^π and w^+/w^π are bounded above and away from zero on $\mathcal{X}^\pi$, so the hypotheses of Lemma S10 hold. By the lower bound in that lemma, (S21) holds.

For the upper bound, define

$$z := \frac{w}{w^\pi}, \qquad \bar{z} := \frac{w^+}{w^\pi},$$

setting both ratios equal to one on $(\mathcal{X}^\pi)^c$.

Apply (S23) to z and $\bar{z}$ under the probability measure $w^\pi \nu = d_{\pi,\gamma}$. By the triangle inequality and (S19),

$$\begin{aligned} \|z - \bar{z}\|_{2,d_{\pi,\gamma}} &\le \|z-1\|_{2,d_{\pi,\gamma}} + \|\bar{z}-1\|_{2,d_{\pi,\gamma}} \\ &\le \sqrt{2\Lambda_w} \left\{ \sqrt{\mathsf{D}_\nu(w\|w^\pi)} + \sqrt{\mathsf{D}_\nu(w^+\|w^\pi)} \right\} \\ &\le \sqrt{2\Lambda_w}(1+\sqrt{\gamma}) \sqrt{\mathsf{D}_\nu(w\|w^\pi)}. \end{aligned} \tag{S28}$$

Moreover, by the log-Hellinger comparison (S24) and the generalized-KL control of Hellinger distance in (S25),

$$\|\log z\|_{2,d_{\pi,\gamma}} \le 2\sqrt{\Lambda_w} \sqrt{\mathsf{D}_\nu(w\|w^\pi)}. \tag{S29}$$

By (S28), (S29), (S1), (S4), and Cauchy–Schwarz,

$$|\mathcal{S}^\star(w)| \le 2\sqrt{2}\,\Lambda_w (1+\sqrt{\gamma}) \mathsf{D}_\nu(w\|w^\pi). \tag{S30}$$

The upper bound in (S5) follows from (S30). Since i was arbitrary, the result holds for every candidate. □

Lemma S13 (Pairwise occupancy-score bounds)**.** *Suppose Condition A2 holds. Then, for every ordered pair (i,j),*

$$\mathcal{S}_{ij} \ge (1-\gamma) e_i^2 - 2\sqrt{2}\,\Lambda_w (1+\sqrt{\gamma}) e_i e_j, \tag{S31}$$

$$H_{ij} \le e_i + e_j, \tag{S32}$$

$$|\mathcal{S}_{ij}| \le 2\sqrt{2}\,\Lambda_w (1+\sqrt{\gamma}) e_i H_{ij}. \tag{S33}$$

Proof of Lemma S13. Let $w_i^+ := \mathcal{B}_\pi^* w_i$, $z_i = w_i/w^\pi$, and $\bar{z}_i = w_i^+/w^\pi$, setting both relative ratios equal to one on $(\mathcal{X}^\pi)^c$. By (S28),

$$\|z_i - \bar{z}_i\|_{2,d_{\pi,\gamma}} \le \sqrt{2\Lambda_w}(1+\sqrt{\gamma}) e_i. \tag{S34}$$

By the log-Hellinger comparison (S24) and the generalized-KL control of Hellinger distance in (S25),

$$\|\log z_j\|_{2,d_{\pi,\gamma}} \le 2\sqrt{\Lambda_w}\, e_j. \tag{S35}$$

By Cauchy–Schwarz, (S34), and (S35),

$$\left| \mathbb{E}_\nu \left[(w_i - w_i^+) \log \frac{w_j}{w^\pi} \right] \right| \le 2\sqrt{2}\,\Lambda_w (1+\sqrt{\gamma}) e_i e_j. \tag{S36}$$

By (S1), the definition of $\mathcal{S}_{ij}$, and $\ell_{ij} = \log(w_i/w^\pi) - \log(w_j/w^\pi)$,

$$\mathcal{S}_{ij} = \mathcal{S}^\star(w_i) - \mathbb{E}_\nu \left[(w_i - w_i^+) \log \frac{w_j}{w^\pi} \right].$$

Combining this identity with (S21) and (S36) proves (S31).

After integration, $H_\nu(p,q) \le \sqrt{\mathsf{D}_\nu(p\|q)}$ by (S25). Therefore, by the triangle inequality in $L^2(\nu)$,

$$H_{ij} \le H_\nu(w_i, w^\pi) + H_\nu(w_j, w^\pi) \le e_i + e_j,$$

which proves (S32). Finally, by (S24),

$$\|\ell_{ij}\|_{2,d_{\pi,\gamma}} \le 2\sqrt{\Lambda_w}\, H_{ij}. \tag{S37}$$

By (S34) and (S37), Cauchy–Schwarz applied to $\mathcal{S}_{ij} = \mathbb{E}_\nu[(w_i - w_i^+)\ell_{ij}]$ gives (S33). □

For the penalized scores, write

$$T_i^w := \max_{1\le j\le M} \mathcal{S}_{ij}^{\mathrm{pen}},$$

and define the population comparison factor

$$R_w := \frac{2}{3}\left[\frac{2\sqrt{2}\,\Lambda_w}{1-\sqrt{\gamma}} + \frac{1}{2} + \sqrt{4\left\{\frac{2\sqrt{2}\,\Lambda_w}{1-\sqrt{\gamma}}\right\}^2 + \frac{2\sqrt{2}\,\Lambda_w}{1-\sqrt{\gamma}} + 1}\right]. \tag{S38}$$

Lemma S14 (Population A-PBV score bounds)**.** *Suppose Condition A2 holds. Then, for every* $i, k \in \{1, \dots, M\}$,

$$T_i^w \ge \mathcal{S}_{ik}^{\mathrm{pen}} \ge \frac{3(1-\gamma)}{4} e_i^2 - \left\{2\sqrt{2}\,\Lambda_w(1+\sqrt{\gamma}) + \frac{1-\gamma}{2}\right\} e_i e_k - \frac{1-\gamma}{4} e_k^2, \tag{S39}$$

$$T_k^w \le \frac{8\Lambda_w^2(1+\sqrt{\gamma})^2}{1-\gamma} e_k^2. \tag{S40}$$

Moreover,

$$R_w \le \frac{4\sqrt{2}\,\Lambda_w}{1-\sqrt{\gamma}} + \frac{2}{3}. \tag{S41}$$

Proof of Lemma S14. By the pairwise lower bound (S31) and the triangle bound (S32),

$$\mathcal{S}_{ik}^{\mathrm{pen}} \ge (1-\gamma)e_i^2 - 2\sqrt{2}\,\Lambda_w(1+\sqrt{\gamma})e_i e_k - \frac{1-\gamma}{4}(e_i+e_k)^2,$$

which is (S39). For every j, by the pairwise upper bound (S33),

$$\mathcal{S}_{kj}^{\mathrm{pen}} \le 2\sqrt{2}\,\Lambda_w(1+\sqrt{\gamma})e_k H_{kj} - \frac{1-\gamma}{4} H_{kj}^2.$$

Completing the square in H_{kj},

$$\begin{aligned} &2\sqrt{2}\,\Lambda_w(1+\sqrt{\gamma})e_k H_{kj} - \frac{1-\gamma}{4} H_{kj}^2 \\ &\quad = \frac{8\Lambda_w^2(1+\sqrt{\gamma})^2}{1-\gamma} e_k^2 - \frac{1-\gamma}{4}\left\{H_{kj} - \frac{4\sqrt{2}\,\Lambda_w(1+\sqrt{\gamma})}{1-\gamma} e_k\right\}^2. \end{aligned}$$

Hence, $\mathcal{S}_{kj}^{\mathrm{pen}} \le 8\Lambda_w^2(1+\sqrt{\gamma})^2 e_k^2/(1-\gamma)$. Taking the maximum over j proves (S40).

To bound R_w, note that $2\sqrt{2}\,\Lambda_w/(1-\sqrt{\gamma}) \ge 1$ and that $\sqrt{4t^2+t+1} \le 2t + 1/2$ for every $t \ge 1$. By this elementary inequality and (S38), (S41) holds. □

G.3 Population oracle inequality

Proof of Theorem S2. Choose $k \in \arg\min_{1\le i\le M} e_i$, and write $x = e_{\widetilde{i}_w}$ and $y = e_k$. By population optimality, $T_{\widetilde{i}_w}^w \le T_k^w$. Combining this comparison with Lemma S14,

$$\frac{3(1-\gamma)}{4} x^2 \le \left\{2\sqrt{2}\,\Lambda_w(1+\sqrt{\gamma}) + \frac{1-\gamma}{2}\right\} xy + \left\{\frac{1-\gamma}{4} + \frac{8\Lambda_w^2(1+\sqrt{\gamma})^2}{1-\gamma}\right\} y^2. \tag{S42}$$

If $y = 0$, then $x = 0$ by (S42). If $y > 0$, let $r = x/y$ and $t = 2\sqrt{2}\,\Lambda_w/(1-\sqrt{\gamma})$. Then (S42) is equivalent to

$$\frac{3}{4} r^2 - (t + \tfrac{1}{2})r - (t^2 + \tfrac{1}{4}) \le 0,$$

whose nonnegative root is R_w as defined in (S38). Thus, $x \le R_w y$. By (S41), (S6) follows. □

G.4 Localized validation deviation

For any $\delta \in (0,1)$, define

$$\tau_{w,\delta} := K_w \log(4n/\delta),$$
$$U_{w,\delta} := 2(1-\gamma)\log \Lambda_w + \left\{ 2\Lambda_w(1+\gamma)\log \Lambda_w + \frac{1-\gamma}{4}\left(\sqrt{\Lambda_w} - \frac{1}{\sqrt{\Lambda_w}}\right)^2 \right\} \tau_{w,\delta},$$
$$V_{w,\delta} := 16\Lambda_w^3(1+\gamma)\tau_{w,\delta} + 16(1-\gamma)\Lambda_w + \frac{(1-\gamma)^2\tau_{w,\delta}}{4}\left(\sqrt{\Lambda_w} - \frac{1}{\sqrt{\Lambda_w}}\right)^2. \tag{S43}$$

Lemma S15 (Localized occupancy-score deviation)**.** *Suppose Conditions A1, A2, and C1 hold, and suppose either Condition C2 holds or the d_0 moments are evaluated exactly. Let $\mathcal{P} \subseteq \{1,\dots,M\}^2$ be any nonempty collection fixed conditional on the training sample. For every $\delta \in (0,1)$, conditional on the training sample, with probability at least $1-\delta$, simultaneously for all $(i,j) \in \mathcal{P}$,*

$$\left|\widehat{\mathcal{S}}_{ij}^{\mathrm{pen}} - \mathcal{S}_{ij}^{\mathrm{pen}}\right| \le C_{\mathrm{M}}\sqrt{V_{w,\delta}}\, H_{ij}\sqrt{\frac{\log(12|\mathcal{P}|/\delta)}{ng_b}} + C_{\mathrm{M}}U_{w,\delta}\frac{\log(12|\mathcal{P}|/\delta)}{ng_b}. \tag{S44}$$

Proof of Lemma S15. Condition on the training sample. For a fixed ordered pair, write

$$g_{ij} := \sqrt{w_i} - \sqrt{w_j}$$

and define

$$\Psi_{ij} := w_i(X)\{\ell_{ij}(X) - \gamma(\pi\ell_{ij})(S')\} - (1-\gamma)\ell_{ij}(X_0) - \frac{1-\gamma}{4}g_{ij}(X)^2. \tag{S45}$$

Then $\widehat{\mathcal{S}}_{ij}^{\mathrm{pen}} = \mathbb{P}_{n,w}\Psi_{ij}$ and $\mathcal{S}_{ij}^{\mathrm{pen}} = \mathbb{E}\Psi_{ij}$, where $\mathbb{P}_{n,w}$ averages the paired observations $(X_t, S'_t, X_{0,t})$. When the d_0 moment is evaluated exactly, replace $\ell_{ij}(X_0)$ in (S45) by $\mathbb{E}_{d_0}[\ell_{ij}(X)]$; the resulting empirical and population scores have the same difference. The bounds below continue to hold by Jensen's inequality.

Envelope. Let

$$W := w^\pi(X), \qquad z_i := \frac{w_i}{w^\pi}, \qquad q_{ij} := (\sqrt{z_i} - \sqrt{z_j})^2.$$

For the validation trajectory, write $W_t := w^\pi(X_t)$. Here and below, each z_i is set to one on $(\mathcal{X}^\pi)^c$. By Condition A2 and Lemma S11,

$$|\ell_{ij}| \le 2\log \Lambda_w, \qquad |\ell_{ij}| \le 2\sqrt{\Lambda_w q_{ij}}, \qquad q_{ij} \le \left(\sqrt{\Lambda_w} - \frac{1}{\sqrt{\Lambda_w}}\right)^2, \tag{S46}$$

ν-almost surely. The initial-state and next-state laws are also supported where these inequalities hold. For the initial state, by (2) and the fixed-point identity for w^π,

$$(1-\gamma)d_0 \le w^\pi \nu. \tag{S47}$$

If $\gamma > 0$, then, by the comparison in (S27),

$$\gamma(w_i\nu)P_\pi \le (\mathcal{B}_\pi^* w_i)\nu \le \Lambda_w w^\pi \nu. \tag{S48}$$

The next-state term is identically zero when $\gamma = 0$.

We now bound the four terms in (S45) separately. The current-state and next-state terms satisfy

$$|w_i(X)\ell_{ij}(X)| \le 2\Lambda_w W \log \Lambda_w, \qquad \gamma|w_i(X)(\pi\ell_{ij})(S')| \le 2\gamma\Lambda_w W \log \Lambda_w.$$

The initial-state term is at most $2(1-\gamma)\log\Lambda_w$ in absolute value. This remains true in the exact-d_0 branch because the absolute value of its expectation is bounded by the same envelope. Finally, because $g_{ij}^2 = Wq_{ij}$, the penalty term is at most

$$\frac{1-\gamma}{4}W\left(\sqrt{\Lambda_w} - \frac{1}{\sqrt{\Lambda_w}}\right)^2.$$

After adding these four bounds, we obtain

$$|\Psi_{ij}| \le 2(1-\gamma)\log\Lambda_w + \left\{2\Lambda_w(1+\gamma)\log\Lambda_w + \frac{1-\gamma}{4}\left(\sqrt{\Lambda_w} - \frac{1}{\sqrt{\Lambda_w}}\right)^2\right\} W. \tag{S49}$$

Second moment. We truncate the target occupancy ratio at $\tau_{w,\delta}$ and set

$$\Psi_{ij}^{\mathrm{tr}} := \Psi_{ij}\mathbf{1}\{W \le \tau_{w,\delta}\}, \qquad |\Psi_{ij}^{\mathrm{tr}}| \le U_{w,\delta}.$$

The second inequality follows from (S49). We next bound the second moment of each of its four components in terms of $H_{ij}^2 = \mathbb{E}_\nu[Wq_{ij}]$.

For the current-state term, by (S46) and $W\mathbf{1}\{W \le \tau_{w,\delta}\} \le \tau_{w,\delta}$,

$$\mathbb{E}\big[w_i(X)^2\ell_{ij}(X)^2\mathbf{1}\{W \le \tau_{w,\delta}\}\big] \le 4\Lambda_w^3\tau_{w,\delta}H_{ij}^2.$$

The next-state term vanishes when $\gamma = 0$. When $\gamma > 0$, by conditional Jensen's inequality and (S48),

$$\begin{aligned}\mathbb{E}\big[\gamma^2 w_i(X)^2\{(\pi\ell_{ij})(S')\}^2\mathbf{1}\{W \le \tau_{w,\delta}\}\big] &\le \gamma^2\Lambda_w\tau_{w,\delta}\mathbb{E}_\nu\big[w_i P_\pi(\ell_{ij}^2)\big] \\ &\le \gamma\Lambda_w^2\tau_{w,\delta}\mathbb{E}_\nu[W\ell_{ij}^2] \\ &\le 4\gamma\Lambda_w^3\tau_{w,\delta}H_{ij}^2.\end{aligned}$$

For an empirical initial-state sample, by (S47),

$$(1-\gamma)^2\mathbb{E}\big[\ell_{ij}(X_0)^2\mathbf{1}\{W \le \tau_{w,\delta}\}\big] \le (1-\gamma)^2\mathbb{E}_{d_0}[\ell_{ij}^2] \le 4(1-\gamma)\Lambda_w H_{ij}^2.$$

In the exact-d_0 branch, the square of $(1-\gamma)\mathbb{E}_{d_0}[\ell_{ij}]$ is bounded by the same right-hand side by Jensen's inequality. Thus, the stated bound holds under either evaluation of the initial-state term.

For the penalty term, $g_{ij}^4 = W^2q_{ij}^2 \le \tau_{w,\delta}\left(\sqrt{\Lambda_w} - 1/\sqrt{\Lambda_w}\right)^2 Wq_{ij}$ on $\{W \le \tau_{w,\delta}\}$. Combining these four component bounds with $(z_1+z_2+z_3+z_4)^2 \le 4\sum_{\ell=1}^4 z_\ell^2$ yields

$$\mathrm{Var}(\Psi_{ij}^{\mathrm{tr}}) \le \mathbb{E}[(\Psi_{ij}^{\mathrm{tr}})^2] \le V_{w,\delta}H_{ij}^2. \tag{S50}$$

Concentration. When Condition C2 holds, the pairs $U_t = (X_t, X_{0,t})$ form a stationary Markov chain with invariant law $\nu \otimes d_0$ and spectral gap at least g_b. Indeed, for centered $f \in L^2(\nu \otimes d_0)$, define $\bar{f}(x) := \mathbb{E}_{d_0}[f(x, X_0)]$, which is centered under ν. If P_U is the transition operator of U_t, then, by (5) and conditional Jensen's inequality,

$$\|P_U f\|_{2,\nu\otimes d_0} = \|P_b\bar{f}\|_{2,\nu} \le (1-g_b)\|\bar{f}\|_{2,\nu} \le (1-g_b)\|f\|_{2,\nu\otimes d_0}.$$

When the d_0 moments are evaluated exactly, take $U_t = X_t$; its absolute spectral gap is at least g_b by Condition A1. In either case, conditional on U_t, the remaining randomness in Ψ_{ij}^{tr} comes only from the time-homogeneous one-step MDP transition. Its conditional law is therefore governed by a kernel that does not depend on t.

By Lemma S8(i), with $B = U_{w,\delta}$ and $v = \sqrt{V_{w,\delta}}H_{ij}$ as established in (S50), and a union bound with $u = \log(12|\mathcal{P}|/\delta)$, there is a simultaneous event of probability at least $1-\delta/2$. With $u_{\mathcal{P},\delta} := \log(12|\mathcal{P}|/\delta)$, on this event,

$$\left|(\mathbb{P}_{n,w} - \mathbb{E})\Psi_{ij}^{\mathrm{tr}}\right| \le C_{\mathrm{M}}\sqrt{V_{w,\delta}}H_{ij}\sqrt{\frac{u_{\mathcal{P},\delta}}{ng_b}} + C_{\mathrm{M}}U_{w,\delta}\frac{u_{\mathcal{P},\delta}}{ng_b}, \tag{S51}$$

simultaneously for all $(i,j) \in \mathcal{P}$.

Removal of truncation. It remains to remove the truncation. Under Condition C1,

$$\mathbb{P}_\nu(W > u) \le 2e^{-u/K_w}, \qquad \mathbb{E}_\nu[W\mathbf{1}\{W > u\}] \le 2(u+K_w)e^{-u/K_w}. \tag{S52}$$

Moreover, since $\mathbb{E}_\nu W = 1$, by Jensen's inequality, $K_w \ge 1/\log 2$. Hence $\tau_{w,\delta} \ge 2$ and $\tau_{w,\delta} + K_w \le 2\tau_{w,\delta}$. By (S49) and (S52),

$$\left|\mathbb{E}(\Psi_{ij} - \Psi_{ij}^{\mathrm{tr}})\right| \le \frac{U_{w,\delta}}{n}. \tag{S53}$$

Also, by a union bound and the first inequality in (S52), $\mathbb{P}\{\max_{1\le t\le n} W_t > \tau_{w,\delta}\} \le \delta/2$. On the complementary event, $\mathbb{P}_{n,w}\Psi_{ij} = \mathbb{P}_{n,w}\Psi_{ij}^{\mathrm{tr}}$ for every pair (i,j). Combining (S51) and (S53) on the intersection of the concentration event and the no-truncation event has failure probability at most $\delta/2 + \delta/2 = \delta$. Because $g_b \le 1$ and $u_{\mathcal{P},\delta} \ge 1$, (S44) follows after increasing the universal constant C_{M} if necessary. □

G.5 Finite-sample oracle inequality

Proof of Theorem S3. Empirical quadratic. Let k be the smallest index in $\arg\min_i e_i$, and define

$$\mathcal{P}_k^w := \{(i,k) : 1 \le i \le M\} \cup \{(k,j) : 1 \le j \le M\}.$$

The collection $\mathcal{P}_k^w$ is fixed conditional on the training sample and has at most $2M-1$ elements. By Lemma S15, the localized bound holds on $\mathcal{P}_k^w$ with its logarithmic factor enlarged to $\log(24M/\delta)$. On the resulting event, write

$$x := e_{\widehat{i}_w}, \qquad y := e_k, \qquad \mathfrak{r}_w := \sqrt{\frac{\log(24M/\delta)}{n g_b}}.$$

For the pair $(\widehat{i}_w, k)$, combine the population lower bound (S39) with the localized deviation bound (S44). In its leading term, $H_{\widehat{i}_w k} \le x + y$ by (S32). Hence,

$$\begin{aligned}\max_j \widehat{\mathcal{S}}^{\mathrm{pen}}_{\widehat{i}_w j} \ge{}& \frac{3(1-\gamma)}{4}x^2 - \left\{2\sqrt{2}\,\Lambda_w(1+\sqrt{\gamma}) + \frac{1-\gamma}{2}\right\}xy - \frac{1-\gamma}{4}y^2 \\ &- C_{\mathrm{M}}\sqrt{V_{w,\delta}}(x+y)\mathfrak{r}_w - C_{\mathrm{M}}U_{w,\delta}\mathfrak{r}_w^2.\end{aligned} \tag{S54}$$

For the oracle candidate k, each population score is bounded by the concave quadratic in H_{kj} from (S33), and its empirical deviation is bounded by (S44). Maximizing that quadratic yields

$$\max_j \widehat{\mathcal{S}}^{\mathrm{pen}}_{kj} \le \frac{\left\{2\sqrt{2}\,\Lambda_w(1+\sqrt{\gamma})y + C_{\mathrm{M}}\sqrt{V_{w,\delta}}\mathfrak{r}_w\right\}^2}{1-\gamma} + C_{\mathrm{M}}U_{w,\delta}\mathfrak{r}_w^2. \tag{S55}$$

By the empirical optimality of $\widehat{i}_w$ and (S54)–(S55),

$$\begin{aligned}\frac{3(1-\gamma)}{4}x^2 \le{}& \left[\left\{2\sqrt{2}\,\Lambda_w(1+\sqrt{\gamma}) + \frac{1-\gamma}{2}\right\}y + C_{\mathrm{M}}\sqrt{V_{w,\delta}}\mathfrak{r}_w\right]x \\ &+ \left\{\frac{1-\gamma}{4} + \frac{8\Lambda_w^2(1+\sqrt{\gamma})^2}{1-\gamma}\right\}y^2 \\ &+ C_{\mathrm{M}}\sqrt{V_{w,\delta}}\left\{1 + \frac{4\sqrt{2}\,\Lambda_w(1+\sqrt{\gamma})}{1-\gamma}\right\}y\mathfrak{r}_w \\ &+ \left\{2C_{\mathrm{M}}U_{w,\delta} + \frac{C_{\mathrm{M}}^2 V_{w,\delta}}{1-\gamma}\right\}\mathfrak{r}_w^2.\end{aligned} \tag{S56}$$

Root comparison. Setting the validation terms in (S56) to zero recovers the population quadratic in (S42). The factor R_w in (S38) is the positive root of the corresponding normalized quadratic. Its derivative at R_w is

$$g_w := \frac{3(1-\gamma)}{2}R_w - 2\sqrt{2}\,\Lambda_w(1+\sqrt{\gamma}) - \frac{1-\gamma}{2}.$$

By (S38), this derivative equals

$$\left[\left\{2\sqrt{2}\,\Lambda_w(1+\sqrt{\gamma}) + \frac{1-\gamma}{2}\right\}^2 + 3(1-\gamma)\left\{\frac{1-\gamma}{4} + \frac{8\Lambda_w^2(1+\sqrt{\gamma})^2}{1-\gamma}\right\}\right]^{1/2},$$

so $g_w > 0$. We choose the following coefficient to dominate, respectively, the terms proportional to $y\mathfrak{r}_w$ and $\mathfrak{r}_w^2$ in the root comparison:

$$\begin{aligned}C_w := \max\Bigg\{&\frac{C_{\mathrm{M}}\sqrt{V_{w,\delta}}}{g_w}\left\{R_w + 1 + \frac{4\sqrt{2}\,\Lambda_w(1+\sqrt{\gamma})}{1-\gamma}\right\}, \\ &\frac{2}{3(1-\gamma)}\left[C_{\mathrm{M}}\sqrt{V_{w,\delta}} + \left\{C_{\mathrm{M}}^2 V_{w,\delta} + 3(1-\gamma)\left\{2C_{\mathrm{M}}U_{w,\delta} + \frac{C_{\mathrm{M}}^2 V_{w,\delta}}{1-\gamma}\right\}\right\}^{1/2}\right]\Bigg\}.\end{aligned} \tag{S57}$$

Move all terms in (S56) to the left and denote the resulting quadratic by

$$\begin{aligned} p(u) := \frac{3(1-\gamma)}{4}u^2 &- \left[\left\{2\sqrt{2}\,\Lambda_w(1+\sqrt{\gamma}) + \frac{1-\gamma}{2}\right\} y + C_{\mathrm{M}}\sqrt{V_{w,\delta}}\mathfrak{r}_w\right] u \\ &- \left\{\frac{1-\gamma}{4} + \frac{8\Lambda_w^2(1+\sqrt{\gamma})^2}{1-\gamma}\right\} y^2 \\ &- C_{\mathrm{M}}\sqrt{V_{w,\delta}}\left\{1 + \frac{4\sqrt{2}\,\Lambda_w(1+\sqrt{\gamma})}{1-\gamma}\right\} y\mathfrak{r}_w \\ &- \left\{2C_{\mathrm{M}}U_{w,\delta} + \frac{C_{\mathrm{M}}^2 V_{w,\delta}}{1-\gamma}\right\}\mathfrak{r}_w^2. \end{aligned}$$

Thus, by (S56), $p(x) \le 0$. We evaluate this polynomial at $x_0 = R_w y + C_w \mathfrak{r}_w$ to show that x_0 lies at or beyond its positive root. By the definition of R_w,

$$\begin{aligned} \frac{3(1-\gamma)}{4}R_w^2 &- \left\{2\sqrt{2}\,\Lambda_w(1+\sqrt{\gamma}) + \frac{1-\gamma}{2}\right\} R_w \\ &- \left\{\frac{1-\gamma}{4} + \frac{8\Lambda_w^2(1+\sqrt{\gamma})^2}{1-\gamma}\right\} = 0. \end{aligned}$$

Direct substitution then yields

$$\begin{aligned} p(x_0) = C_w g_w y\mathfrak{r}_w &\\ &- C_{\mathrm{M}}\sqrt{V_{w,\delta}}\left\{R_w + 1 + \frac{4\sqrt{2}\,\Lambda_w(1+\sqrt{\gamma})}{1-\gamma}\right\} y\mathfrak{r}_w \\ &+ \left\{\frac{3(1-\gamma)}{4}C_w^2 - C_{\mathrm{M}}\sqrt{V_{w,\delta}}C_w\right\}\mathfrak{r}_w^2 \\ &- \left\{2C_{\mathrm{M}}U_{w,\delta} + \frac{C_{\mathrm{M}}^2 V_{w,\delta}}{1-\gamma}\right\}\mathfrak{r}_w^2 \ge 0. \end{aligned} \tag{S58}$$

The first entry in the maximum defining C_w makes the coefficient of $y\mathfrak{r}_w$ nonnegative, and the second entry makes the coefficient of $\mathfrak{r}_w^2$ nonnegative. Hence, the final inequality in (S58) holds by (S57). The quadratic p has positive leading coefficient and $p(0) < 0$. Consequently, the set $\{u \ge 0 : p(u) \le 0\}$ ends at its unique positive root. By the inequality $p(x) \le 0$ and (S58),

$$x \le R_w y + C_w \mathfrak{r}_w. \tag{S59}$$

Simplification. It remains to simplify C_w. Since $\tau_{w,\delta} \ge 2$,

$$\left(\sqrt{\Lambda_w} - \frac{1}{\sqrt{\Lambda_w}}\right)^2 \le \Lambda_w, \qquad 2\log\Lambda_w \le 2\Lambda_w,$$

by the definitions in (S43),

$$\begin{aligned} U_{w,\delta} &\le \frac{21}{4}\Lambda_w^2\tau_{w,\delta}, \\ V_{w,\delta} &\le \frac{161}{4}\Lambda_w^3\tau_{w,\delta}. \end{aligned}$$

Consequently,

$$C_{\mathrm{M}}\sqrt{V_{w,\delta}} \le C_a\Lambda_w^{3/2}\sqrt{\tau_{w,\delta}}, \qquad C_{\mathrm{M}}U_{w,\delta} \le C_b\Lambda_w^2\tau_{w,\delta},$$

for universal constants $C_a, C_b > 0$. By (S41) and $(1-\sqrt{\gamma})^{-1} \le 2/(1-\gamma)$, $R_w \le 12\Lambda_w/(1-\gamma)$. By the displayed formula for g_w, $g_w \ge 2\sqrt{2}\,\Lambda_w(1+\sqrt{\gamma}) + (1-\gamma)/2 > 2\Lambda_w$. After substituting these bounds into (S57), we obtain

$$C_w \le \frac{C_1\Lambda_w^{3/2}\sqrt{\tau_{w,\delta}}}{1-\gamma} \tag{S60}$$

for a universal constant $C_1 > 0$. By (S60), (S41), and (S59), (S7) follows. □

H PROOFS FOR TA-PBV

H.1 POPULATION SCORE BOUNDS

The proof follows the main-text construction. Under $d_{\pi,\gamma}$, γP_π is a contraction, so the Bellman quadratic is coercive and the Bellman residual is bounded in $L^2(d_{\pi,\gamma})$. For $h_{ik} = \Delta_i - \Delta_k$, the penalized score separates into the oracle quadratic and a perturbation involving the comparison error Δ_k. If the selected error is not already within the claimed oracle factor, the perturbation is bounded by the coercivity slack. With the negative quadratic penalty, the score of the best candidate has a concave upper bound. The finite-sample proof inserts a localized validation bound into the same selected-versus-comparator comparison.

Recall that

$$\Delta_i := Q_i - Q^\pi.$$

Lemma S16 (Discounted-occupancy operator bounds)**.** *For every $f \in L^2(d_{\pi,\gamma})$,*

$$\|\gamma P_\pi f\|_{2,d_{\pi,\gamma}} \le \sqrt{\gamma}\|f\|_{2,d_{\pi,\gamma}}, \tag{S61}$$

$$\langle f, (I - \gamma P_\pi) f\rangle_{d_{\pi,\gamma}} \ge (1 - \sqrt{\gamma})\|f\|^2_{2,d_{\pi,\gamma}}, \tag{S62}$$

$$\|(I - \gamma P_\pi) f\|_{2,d_{\pi,\gamma}} \le (1 + \sqrt{\gamma})\|f\|_{2,d_{\pi,\gamma}}. \tag{S63}$$

Proof. If $\gamma = 0$, then $\gamma P_\pi = 0$ and $I - \gamma P_\pi = I$, so all three bounds hold with equality. Suppose $\gamma > 0$. By conditional Jensen's inequality,

$$(P_\pi f)^2 \le P_\pi(f^2).$$

By the discounted-occupancy identity $d_{\pi,\gamma} = (1-\gamma)d_0 + \gamma d_{\pi,\gamma} P_\pi$,

$$\gamma \mathbb{E}_{d_{\pi,\gamma}}\big[P_\pi(f^2)\big] = \mathbb{E}_{d_{\pi,\gamma}}(f^2) - (1-\gamma)\mathbb{E}_{d_0}(f^2) \le \mathbb{E}_{d_{\pi,\gamma}}(f^2). \tag{S64}$$

Therefore,

$$\begin{aligned}
\|\gamma P_\pi f\|^2_{2,d_{\pi,\gamma}} &= \gamma^2 \mathbb{E}_{d_{\pi,\gamma}}\big[(P_\pi f)^2\big] \\
&\le \gamma^2 \mathbb{E}_{d_{\pi,\gamma}}\big[P_\pi(f^2)\big] \\
&\le \gamma \mathbb{E}_{d_{\pi,\gamma}}(f^2) \\
&= \gamma \|f\|^2_{2,d_{\pi,\gamma}},
\end{aligned}$$

which proves (S61).

For the coercivity bound, by Cauchy–Schwarz and (S61),

$$\begin{aligned}
\langle f, (I - \gamma P_\pi) f\rangle_{d_{\pi,\gamma}} &= \|f\|^2_{2,d_{\pi,\gamma}} - \langle f, \gamma P_\pi f\rangle_{d_{\pi,\gamma}} \\
&\ge \|f\|^2_{2,d_{\pi,\gamma}} - \|f\|_{2,d_{\pi,\gamma}} \|\gamma P_\pi f\|_{2,d_{\pi,\gamma}} \\
&\ge (1 - \sqrt{\gamma})\|f\|^2_{2,d_{\pi,\gamma}},
\end{aligned}$$

which proves (S62). Finally, by the triangle inequality and (S61),

$$\begin{aligned}
\|(I - \gamma P_\pi) f\|_{2,d_{\pi,\gamma}} &\le \|f\|_{2,d_{\pi,\gamma}} + \|\gamma P_\pi f\|_{2,d_{\pi,\gamma}} \\
&\le (1 + \sqrt{\gamma})\|f\|_{2,d_{\pi,\gamma}},
\end{aligned}$$

which proves (S63). □

The following refinement will be used to control the error from replacing the oracle critic Δ_i by the pairwise critic $h_{ik} = \Delta_i - \Delta_k$.

Lemma S17 (Coercivity-slack bound)**.** *For nonzero $f \in L^2(d_{\pi,\gamma})$, define the coercivity slack*

$$a(f) := \langle f, (I - \gamma P_\pi) f\rangle_{d_{\pi,\gamma}} - (1 - \sqrt{\gamma})\|f\|^2_{2,d_{\pi,\gamma}}.$$

Then

$$\left\| \left(I - \gamma P_\pi - \frac{1 - \sqrt{\gamma}}{2} I \right) f \right\|_{2,d_{\pi,\gamma}} \le \frac{1 - \sqrt{\gamma}}{2} \|f\|_{2,d_{\pi,\gamma}} + \frac{1 + \sqrt{\gamma}}{1 - \sqrt{\gamma}} \frac{a(f)}{\|f\|_{2,d_{\pi,\gamma}}}. \tag{S65}$$

Proof. By (S62), $a(f) \geq 0$. Since

$$I - \gamma P_\pi - \frac{1-\sqrt{\gamma}}{2} I = \frac{1+\sqrt{\gamma}}{2} I - \gamma P_\pi,$$

after expanding the squared norm and applying (S61), we obtain

$$\left\| \left(I - \gamma P_\pi - \frac{1-\sqrt{\gamma}}{2} I \right) f \right\|_{2,d_{\pi,\gamma}}^2 \leq \frac{(1-\sqrt{\gamma})^2}{4} \|f\|_{2,d_{\pi,\gamma}}^2 + (1+\sqrt{\gamma}) a(f).$$

On the other hand,

$$\left\{ \frac{1-\sqrt{\gamma}}{2} \|f\|_{2,d_{\pi,\gamma}} + \frac{1+\sqrt{\gamma}}{1-\sqrt{\gamma}} \frac{a(f)}{\|f\|_{2,d_{\pi,\gamma}}} \right\}^2 = \frac{(1-\sqrt{\gamma})^2}{4} \|f\|_{2,d_{\pi,\gamma}}^2 + (1+\sqrt{\gamma}) a(f) + \left\{ \frac{1+\sqrt{\gamma}}{1-\sqrt{\gamma}} \frac{a(f)}{\|f\|_{2,d_{\pi,\gamma}}} \right\}^2.$$

The final term is nonnegative, so comparison of the two squared quantities proves (S65). □

Lemma S18 (Second moments and score integrability)**.** *Suppose Condition B1 holds. Then $Q^\pi \in L^2(d_{\pi,\gamma})$, every $Q_i - Q^\pi$ belongs to $L^2(d_{\pi,\gamma})$, and the expectations defining $\mathcal{M}_{ij}^{\mathrm{pen}}$ are absolutely integrable for every ordered pair (i,j).*

Proof. Let $r(x) := \mathbb{E}(R \mid X = x)$. By conditional Jensen's inequality, $\|r\|_{2,d_{\pi,\gamma}}^2 \leq \mathbb{E}_\nu[w^\pi(X) R^2] < \infty$. By Lemma S16, $\|\gamma P_\pi\|_{L^2(d_{\pi,\gamma}) \to L^2(d_{\pi,\gamma})} \leq \sqrt{\gamma} < 1$, so $I - \gamma P_\pi$ is invertible on $L^2(d_{\pi,\gamma})$. Hence, by the Bellman equation, $Q^\pi = (I - \gamma P_\pi)^{-1} r \in L^2(d_{\pi,\gamma})$. By Condition B1, every $\Delta_i = Q_i - Q^\pi$ and $h_{ij} = Q_i - Q_j = \Delta_i - \Delta_j$ belongs to $L^2(d_{\pi,\gamma})$.

It remains to verify integrability of the one-step score. For $\gamma > 0$, by conditional Jensen's inequality and (S64), $\mathbb{E}_\nu[w^\pi(X)\{(\pi Q_i)(S')\}^2] \leq \gamma^{-1} \|Q_i\|_{2,d_{\pi,\gamma}}^2 < \infty$; when $\gamma = 0$, the continuation term is identically zero. Hence $Q_i(X) - R - \gamma(\pi Q_i)(S')$ is square integrable under the w^π-weighted transition law. By Cauchy–Schwarz and $h_{ij} \in L^2(d_{\pi,\gamma})$, $\mathbb{E}_\nu[w^\pi(X)|h_{ij}(X)||Q_i(X) - R - \gamma(\pi Q_i)(S')|] < \infty$. Finally, $\mathbb{E}_\nu[w^\pi(X) h_{ij}(X)^2] = \|h_{ij}\|_{2,d_{\pi,\gamma}}^2 < \infty$, so both terms defining $\mathcal{M}_{ij}^{\mathrm{pen}}$ are absolutely integrable. □

For each $i \in [K]$, define the worst-case population score

$$T_i := \max_{1 \leq j \leq K} \mathcal{M}_{ij}^{\mathrm{pen}}.$$

Lemma S19 (Population score comparison)**.** *Suppose Condition B1 holds. Fix a candidate $i \in [K]$ and a comparator $k \in [K]$ satisfying*

$$\|\Delta_i\|_{2,d_{\pi,\gamma}} \geq \frac{1+\sqrt{\gamma}}{1-\sqrt{\gamma}} \|\Delta_k\|_{2,d_{\pi,\gamma}}. \tag{S66}$$

Then

$$T_i \geq \mathcal{M}_{ik}^{\mathrm{pen}} \geq \frac{3(1-\sqrt{\gamma})}{4} \|\Delta_i\|_{2,d_{\pi,\gamma}}^2 - \frac{1-\sqrt{\gamma}}{2} \|\Delta_i\|_{2,d_{\pi,\gamma}} \|\Delta_k\|_{2,d_{\pi,\gamma}} - \frac{1-\sqrt{\gamma}}{4} \|\Delta_k\|_{2,d_{\pi,\gamma}}^2. \tag{S67}$$

Moreover, for every $j \in [K]$,

$$\mathcal{M}_{kj}^{\mathrm{pen}} \leq (1+\sqrt{\gamma}) \|\Delta_k\|_{2,d_{\pi,\gamma}} \|\Delta_k - \Delta_j\|_{2,d_{\pi,\gamma}} - \frac{1-\sqrt{\gamma}}{4} \|\Delta_k - \Delta_j\|_{2,d_{\pi,\gamma}}^2, \tag{S68}$$

and consequently

$$T_k \leq \frac{(1+\sqrt{\gamma})^2}{1-\sqrt{\gamma}} \|\Delta_k\|_{2,d_{\pi,\gamma}}^2. \tag{S69}$$

Proof. Because $Q_i - \mathcal{B}_\pi Q_i = (I - \gamma P_\pi)\Delta_i$ and $d_{\pi,\gamma} = w^\pi \nu$, we obtain after conditioning on X

$$\mathcal{M}_{ij}^{\mathrm{pen}} = \langle h_{ij}, (I - \gamma P_\pi)\Delta_i \rangle_{d_{\pi,\gamma}} - \frac{1-\sqrt{\gamma}}{4} \|h_{ij}\|_{2,d_{\pi,\gamma}}^2. \tag{S70}$$

We first lower-bound the score of Q_i by comparing it with Q_k. Write

$$x := \|\Delta_i\|_{2,d_{\pi,\gamma}}, \qquad y := \|\Delta_k\|_{2,d_{\pi,\gamma}}.$$

If $x = 0$, then $y = 0$ by (S66), and $\mathcal{M}_{ik}^{\mathrm{pen}} = 0$ by (S70); this proves (S67). Suppose $x > 0$.

Since

$$h_{ik} = \Delta_i - \Delta_k,$$

the pairwise critic is the oracle critic Δ_i perturbed by the comparator error $-\Delta_k$. Let $a_i := a(\Delta_i)$ denote the coercivity slack from Lemma S17. After substituting $h_{ik} = \Delta_i - \Delta_k$ into (S70) and expanding the quadratic penalty, we obtain

$$\mathcal{M}_{ik}^{\mathrm{pen}} = \frac{3(1-\sqrt{\gamma})}{4} x^2 + a_i - \left\langle \Delta_k, \left(I - \gamma P_\pi - \frac{1-\sqrt{\gamma}}{2} I \right) \Delta_i \right\rangle_{d_{\pi,\gamma}} - \frac{1-\sqrt{\gamma}}{4} y^2. \tag{S71}$$

By Lemma S17,

$$\left\| \left(I - \gamma P_\pi - \frac{1-\sqrt{\gamma}}{2} I \right) \Delta_i \right\|_{2,d_{\pi,\gamma}} \le \frac{1-\sqrt{\gamma}}{2} x + \frac{1+\sqrt{\gamma}}{1-\sqrt{\gamma}} \frac{a_i}{x}.$$

By Cauchy–Schwarz,

$$\left| \left\langle \Delta_k, \left(I - \gamma P_\pi - \frac{1-\sqrt{\gamma}}{2} I \right) \Delta_i \right\rangle_{d_{\pi,\gamma}} \right| \le \frac{1-\sqrt{\gamma}}{2} xy + \frac{1+\sqrt{\gamma}}{1-\sqrt{\gamma}} \frac{y}{x} a_i.$$

By (S66),

$$1 - \frac{1+\sqrt{\gamma}}{1-\sqrt{\gamma}} \frac{y}{x} \ge 0.$$

Substitute the preceding cross-term bound into (S71) and discard the resulting nonnegative multiple of a_i. Then

$$\mathcal{M}_{ik}^{\mathrm{pen}} \ge \frac{3(1-\sqrt{\gamma})}{4} x^2 - \frac{1-\sqrt{\gamma}}{2} xy - \frac{1-\sqrt{\gamma}}{4} y^2.$$

Since $T_i \ge \mathcal{M}_{ik}^{\mathrm{pen}}$, this proves (S67).

For the upper bound, fix $k, j \in [K]$ and set $u := \|\Delta_k - \Delta_j\|_{2,d_{\pi,\gamma}}$. By (S70), Cauchy–Schwarz, and (S63),

$$\mathcal{M}_{kj}^{\mathrm{pen}} \le (1+\sqrt{\gamma}) \|\Delta_k\|_{2,d_{\pi,\gamma}} u - \frac{1-\sqrt{\gamma}}{4} u^2,$$

which proves (S68). Completing the square,

$$\begin{aligned} &(1+\sqrt{\gamma}) \|\Delta_k\|_{2,d_{\pi,\gamma}} u - \frac{1-\sqrt{\gamma}}{4} u^2 \\ &\quad = \frac{(1+\sqrt{\gamma})^2}{1-\sqrt{\gamma}} \|\Delta_k\|_{2,d_{\pi,\gamma}}^2 - \frac{1-\sqrt{\gamma}}{4} \left\{ u - \frac{2(1+\sqrt{\gamma})}{1-\sqrt{\gamma}} \|\Delta_k\|_{2,d_{\pi,\gamma}} \right\}^2. \end{aligned}$$

Maximizing over $u \ge 0$ therefore gives

$$T_k \le \frac{(1+\sqrt{\gamma})^2}{1-\sqrt{\gamma}} \|\Delta_k\|_{2,d_{\pi,\gamma}}^2,$$

proving (S69). □

Proof of Theorem S4. Choose

$$k \in \arg\min_{1 \le i \le K} \|\Delta_i\|_{2,d_{\pi,\gamma}},$$

and write

$$x := \|\Delta_{\widetilde{i}_Q}\|_{2,d_{\pi,\gamma}}, \qquad y := \|\Delta_k\|_{2,d_{\pi,\gamma}}, \qquad t := \frac{1+\sqrt{\gamma}}{1-\sqrt{\gamma}}.$$

If $x \le ty$, the claimed bound follows since $t \le 7t/6 + 1/2$. Suppose instead that $x > ty$. Then (S66) holds with $i = \widetilde{i}_Q$. Since $\widetilde{i}_Q$ minimizes T_i, by Lemma S19,

$$\frac{3(1-\sqrt{\gamma})}{4}x^2 - \frac{1-\sqrt{\gamma}}{2}xy - \frac{1-\sqrt{\gamma}}{4}y^2 \le T_{\widetilde{i}_Q} \le T_k \le \frac{(1+\sqrt{\gamma})^2}{1-\sqrt{\gamma}}y^2.$$

After rearranging and dividing by $(1-\sqrt{\gamma})/4$,

$$3x^2 \le 2xy + (1+4t^2)y^2.$$

Solving this quadratic inequality for x yields

$$x \le \frac{1}{3}\left\{1 + 2\sqrt{1+3t^2}\right\} y. \tag{S72}$$

Since $t \ge 1$,

$$(7t+1)^2 - 16(1+3t^2) = (t-1)(t+15) \ge 0,$$

and hence

$$\frac{1}{3}\left\{1 + 2\sqrt{1+3t^2}\right\} \le \frac{7}{6}t + \frac{1}{2}. \tag{S73}$$

By (S72) and (S73), after substituting the definition of t,

$$x \le \left\{\frac{7}{6}\frac{1+\sqrt{\gamma}}{1-\sqrt{\gamma}} + \frac{1}{2}\right\} y,$$

which proves (S8). □

H.2 Concentration bound for the Q-function criterion

Define the relative fitted-weight error by

$$\rho_w^2 := \mathbb{E}_\nu\left[\frac{\{\widehat{w}(X) - w^\pi(X)\}^2}{w^\pi(X)}\right]. \tag{S74}$$

The conventions $0/0 := 0$ and $a/0 := +\infty$ for $a > 0$ apply in (S74).

Lemma S20 (Localized validation-score deviation)**.** *Suppose Condition A1 and Conditions B2–B4 hold. There is a universal constant $C_0 > 0$ such that, for every $\delta \in (0,1)$ and every nonempty collection $\mathcal{P} \subseteq \{1,\dots,K\}^2$ fixed conditional on the training sample, with probability at least $1-\delta$, simultaneously for all $(i,j) \in \mathcal{P}$,*

$$\begin{aligned}
\left|\widehat{\mathcal{M}}_{ij}^{\mathrm{pen}} - \mathcal{M}_{ij}^{\mathrm{pen}}\right| \le{}& \left(G_Q + \frac{1-\sqrt{\gamma}}{4}H_Q\right)\rho_w\|h_{ij}\|_{2,d_{\pi,\gamma}} \\
&+ C_0\sqrt{W_Q}\left(G_Q + \frac{1-\sqrt{\gamma}}{4}H_Q + \sigma_Q\right)\|h_{ij}\|_{2,d_{\pi,\gamma}} \\
&\quad\times\sqrt{\frac{\log(6|\mathcal{P}|/\delta)}{ng_b}} \\
&+ C_0\frac{\sqrt{W_Q}}{2}H_Q\left(G_Q + \frac{1-\sqrt{\gamma}}{4}H_Q + \sigma_Q\right)\rho_w \\
&\quad\times\sqrt{\frac{\log(6|\mathcal{P}|/\delta)}{ng_b}} \\
&+ C_0 W_Q H_Q\left(G_Q + \frac{1-\sqrt{\gamma}}{4}H_Q + \sigma_Q\right)\frac{\log(6|\mathcal{P}|/\delta)}{ng_b}.
\end{aligned} \tag{S75}$$

Proof of Lemma S20. If $\rho_w = +\infty$, the conclusion holds by the convention following (S74). Hence, suppose $\rho_w < +\infty$. Condition on the training sample, and let $\mathbb{P}_{n,Q}$ denote the empirical average over the Q-validation transitions. For each $i \in [K]$, set

$$\mu_i(X) := \mathbb{E}[\varepsilon_i \mid X], \qquad \xi_i := \varepsilon_i - \mu_i(X).$$

For an ordered pair (i, j), define

$$Z_{ij} := \widehat{w}(X)h_{ij}(X)\left\{\varepsilon_i(X, R, S') - \frac{1-\sqrt{\gamma}}{4}h_{ij}(X)\right\},$$

so that $\widehat{\mathcal{M}}_{ij}^{\mathrm{pen}} = \mathbb{P}_{n,Q}Z_{ij}$. Write

$$A_{ij}(X) := \widehat{w}(X)h_{ij}(X), \qquad m_{ij}(X) := \mu_i(X) - \frac{1-\sqrt{\gamma}}{4}h_{ij}(X).$$

Then $Z_{ij} = A_{ij}(m_{ij} + \xi_i)$ and, under Condition B2,

$$|m_{ij}| \le G_Q + \frac{1-\sqrt{\gamma}}{4}H_Q.$$

Under Condition B4,

$$|A_{ij}| \le W_Q H_Q, \tag{S76}$$

$$\begin{aligned}\mathbb{E}_\nu(A_{ij}^2) &\le W_Q\left\{\|h_{ij}\|_{2,d_{\pi,\gamma}}^2 + H_Q\rho_w\|h_{ij}\|_{2,d_{\pi,\gamma}}\right\} \\ &\le W_Q\left\{\|h_{ij}\|_{2,d_{\pi,\gamma}} + \frac{H_Q\rho_w}{2}\right\}^2.\end{aligned} \tag{S77}$$

To obtain the second bound, by $\widehat{w}^2 \le W_Q\widehat{w}$ and Cauchy–Schwarz,

$$\left|\mathbb{E}_\nu[(\widehat{w} - w^\pi)h_{ij}^2]\right| \le \rho_w\{\mathbb{E}_{d_{\pi,\gamma}}(h_{ij}^4)\}^{1/2} \le H_Q\rho_w\|h_{ij}\|_{2,d_{\pi,\gamma}}.$$

Fitted-weight bias. Because $\mathbb{E}(\xi_i \mid X) = 0$, the difference between the conditional population score based on $\widehat{w}$ and the target-weighted score is

$$\mathbb{E}Z_{ij} - \mathcal{M}_{ij}^{\mathrm{pen}} = \mathbb{E}_\nu\left[\{\widehat{w}(X) - w^\pi(X)\}h_{ij}(X)m_{ij}(X)\right]. \tag{S78}$$

By Cauchy–Schwarz in (S78), (S74), and $d_{\pi,\gamma} = w^\pi\nu$,

$$\left|\mathbb{E}Z_{ij} - \mathcal{M}_{ij}^{\mathrm{pen}}\right| \le \left(G_Q + \frac{1-\sqrt{\gamma}}{4}H_Q\right)\rho_w\|h_{ij}\|_{2,d_{\pi,\gamma}}. \tag{S79}$$

Empirical fluctuation. Under the validation-data assumptions in Section 2.1, the transition noise ξ_i is a martingale difference, with conditional law determined by the time-homogeneous MDP reward and transition kernels. Under Condition A1, the validation chain has the spectral gap required by Lemma S8. Apply Lemma S8(ii) with

$$a = A_{ij}, \qquad d = m_{ij}, \qquad \zeta = \xi_i, \qquad B = W_Q H_Q, \qquad D = G_Q + \frac{1-\sqrt{\gamma}}{4}H_Q.$$

The multiplier envelope and second-moment conditions hold by (S76) and (S77), respectively. The sub-Gaussian innovation condition in (S14) holds by Condition B3. After applying the lemma to each ordered pair in $\mathcal{P}$ with $u = \log(6|\mathcal{P}|/\delta)$, taking a union bound, and increasing the universal constant C_0, we obtain

$$\begin{aligned}|\mathbb{P}_{n,Q}Z_{ij} - \mathbb{E}Z_{ij}| \le{}& C_0\sqrt{W_Q}\left(G_Q + \frac{1-\sqrt{\gamma}}{4}H_Q + \sigma_Q\right) \\ &\times\sqrt{\frac{\log(6|\mathcal{P}|/\delta)}{ng_b}}\left\{\|h_{ij}\|_{2,d_{\pi,\gamma}} + \frac{H_Q\rho_w}{2}\right\} \\ &+ C_0 W_Q H_Q\left(G_Q + \frac{1-\sqrt{\gamma}}{4}H_Q + \sigma_Q\right)\frac{\log(6|\mathcal{P}|/\delta)}{ng_b}\end{aligned} \tag{S80}$$

simultaneously for all $(i, j) \in \mathcal{P}$. By (S80) and (S79), (S75) follows. □

H.3 FINITE-SAMPLE ORACLE INEQUALITY

Let C_0 be the universal constant in Lemma S20, and define

$$C_{Q,\mathrm{wt}} := 4\sqrt{2}\left(G_Q + \frac{H_Q}{4}\right) + 2\sqrt{2C_0 H_Q\left(G_Q + \frac{H_Q}{4} + \sigma_Q\right)\left(W_Q + \frac{\sqrt{W_Q}}{8}\right)},$$

$$C_{Q,\mathrm{val}} := 4\sqrt{2}\,C_0\sqrt{W_Q}\left(G_Q + \frac{H_Q}{4} + \sigma_Q\right) + 2\sqrt{2C_0 H_Q\left(G_Q + \frac{H_Q}{4} + \sigma_Q\right)\left(W_Q + \frac{\sqrt{W_Q}}{8}\right)}.$$

These constants depend only on the envelopes in Conditions B2–B4.

Proof of Theorem S5. If $\rho_w = +\infty$, (S9) holds by the convention following (S74). Hence, suppose $\rho_w < +\infty$.

Localized event. Let k be the smallest index minimizing $\|\Delta_i\|_{2,d_{\pi,\gamma}}$, and define

$$\mathcal{P}_k := \{(i,k) : 1 \le i \le K\} \cup \{(k,j) : 1 \le j \le K\}.$$

The collection $\mathcal{P}_k$ is fixed conditional on the training sample and has at most $2K-1$ elements. Apply Lemma S20 to $\mathcal{P}_k$, and enlarge its logarithmic factor to $\log(12K/\delta)$. On the resulting event, abbreviate

$$x := \|\Delta_{\widehat{i}_Q}\|_{2,d_{\pi,\gamma}}, \qquad y := \|\Delta_k\|_{2,d_{\pi,\gamma}}, \qquad t := \frac{1+\sqrt{\gamma}}{1-\sqrt{\gamma}}.$$

For the deviation bound, set

$$\mathfrak{r}_Q := \sqrt{\frac{\log(12K/\delta)}{n g_b}},$$

$$\eta := \left(G_Q + \frac{1-\sqrt{\gamma}}{4}H_Q\right)\rho_w + C_0\sqrt{W_Q}\left(G_Q + \frac{1-\sqrt{\gamma}}{4}H_Q + \sigma_Q\right)\mathfrak{r}_Q,$$

$$v := C_0 H_Q\left(G_Q + \frac{1-\sqrt{\gamma}}{4}H_Q + \sigma_Q\right)\left\{\frac{\sqrt{W_Q}}{2}\rho_w\mathfrak{r}_Q + W_Q\mathfrak{r}_Q^2\right\}.$$

If $x \le ty$, then $t \le 7t/6 + 1/2$ implies

$$x \le \left\{\frac{7}{6}\frac{1+\sqrt{\gamma}}{1-\sqrt{\gamma}} + \frac{1}{2}\right\} y.$$

This is the claimed bound because the two remainder terms are nonnegative. Suppose henceforth that $x > ty$. Then (S66) holds with $i = \widehat{i}_Q$.

Empirical comparison. Since $\|h_{\widehat{i}_Q k}\|_{2,d_{\pi,\gamma}} \le x + y$, by (S75), applied to $(\widehat{i}_Q, k)$, and (S67),

$$\max_j \widehat{\mathcal{M}}^{\mathrm{pen}}_{\widehat{i}_Q j} \ge \frac{3(1-\sqrt{\gamma})}{4}x^2 - \frac{1-\sqrt{\gamma}}{2}xy - \frac{1-\sqrt{\gamma}}{4}y^2 - \eta(x+y) - v. \tag{S81}$$

For the oracle comparator k, each population pairwise score is bounded by (S68), and its empirical deviation is bounded by (S75). Maximizing the resulting concave quadratic in $\|\Delta_k - \Delta_j\|_{2,d_{\pi,\gamma}}$ yields

$$\max_j \widehat{\mathcal{M}}^{\mathrm{pen}}_{kj} \le \frac{\{(1+\sqrt{\gamma})y + \eta\}^2}{1-\sqrt{\gamma}} + v. \tag{S82}$$

Quadratic/root bound. By the empirical optimality of $\widehat{i}_Q$ and (S81)–(S82),

$$\frac{3(1-\sqrt{\gamma})}{4}x^2 \le \left\{\frac{1-\sqrt{\gamma}}{2}y+\eta\right\}x+\left\{\frac{1-\sqrt{\gamma}}{4}+\frac{(1+\sqrt{\gamma})^2}{1-\sqrt{\gamma}}\right\}y^2+\left\{1+\frac{2(1+\sqrt{\gamma})}{1-\sqrt{\gamma}}\right\}\eta y + \frac{\eta^2}{1-\sqrt{\gamma}}+2v. \tag{S83}$$

Solving (S83) for its nonnegative root, we obtain

$$x \le \frac{2}{3(1-\sqrt{\gamma})}\left[\frac{1-\sqrt{\gamma}}{2}y+\eta+\left\{\begin{array}{l}\{3(1+\sqrt{\gamma})^2+(1-\sqrt{\gamma})^2\}y^2\\ \quad+2\{3(1+\sqrt{\gamma})+2(1-\sqrt{\gamma})\}\eta y\\ \quad+4\eta^2+6(1-\sqrt{\gamma})v\end{array}\right\}^{1/2}\right]. \tag{S84}$$

Because $(1+\sqrt{\gamma})/(1-\sqrt{\gamma}) \ge 1$,

$$2\{3(1+\sqrt{\gamma})+2(1-\sqrt{\gamma})\} \le 5\sqrt{3(1+\sqrt{\gamma})^2+(1-\sqrt{\gamma})^2}.$$

Indeed, the difference between the squares of the right- and left-hand sides is

$$3\{(1+\sqrt{\gamma})-(1-\sqrt{\gamma})\}\{13(1+\sqrt{\gamma})-3(1-\sqrt{\gamma})\} \ge 0.$$

Thus, the radicand in (S84) is at most

$$\left\{\sqrt{3(1+\sqrt{\gamma})^2+(1-\sqrt{\gamma})^2}\,y+\frac{5}{2}\eta\right\}^2+6(1-\sqrt{\gamma})v.$$

By $\sqrt{u+v} \le \sqrt{u}+\sqrt{v}$,

$$\begin{aligned} x &\le \frac{1}{3}\left[1+2\sqrt{1+3\left\{\frac{1+\sqrt{\gamma}}{1-\sqrt{\gamma}}\right\}^2}\right]y+\frac{7}{3(1-\sqrt{\gamma})}\eta+\frac{2\sqrt{6}}{3}\sqrt{\frac{v}{1-\sqrt{\gamma}}} \\ &\le \left\{\frac{7}{6}\frac{1+\sqrt{\gamma}}{1-\sqrt{\gamma}}+\frac{1}{2}\right\}y+\frac{4\sqrt{2}}{1-\sqrt{\gamma}}\eta+2\sqrt{\frac{2v}{1-\sqrt{\gamma}}}, \end{aligned}$$

where the second inequality follows from (S73). Substituting the definitions of η and v, it remains to note that

$$v \le C_0 H_Q\left(G_Q+\frac{1-\sqrt{\gamma}}{4}H_Q+\sigma_Q\right)\left(W_Q+\frac{\sqrt{W_Q}}{8}\right)(\rho_w+\mathfrak{r}_Q)^2. \tag{S85}$$

Indeed, (S85) follows from $\rho_w\mathfrak{r}_Q \le (\rho_w+\mathfrak{r}_Q)^2/4$ and $\mathfrak{r}_Q^2 \le (\rho_w+\mathfrak{r}_Q)^2$. Because $1-\sqrt{\gamma} \le 1$, $1/\sqrt{1-\sqrt{\gamma}} \le 1/(1-\sqrt{\gamma})$ and

$$G_Q+\frac{1-\sqrt{\gamma}}{4}H_Q \le G_Q+\frac{H_Q}{4},$$

and the same inequality holds after adding σ_Q. After substitution in the preceding root bound and application of (S73), we obtain

$$x \le \left\{\frac{7}{6}\frac{1+\sqrt{\gamma}}{1-\sqrt{\gamma}}+\frac{1}{2}\right\}y+\frac{C_{Q,\mathrm{wt}}}{1-\sqrt{\gamma}}\rho_w+\frac{C_{Q,\mathrm{val}}}{1-\sqrt{\gamma}}\mathfrak{r}_Q.$$

Substituting the definitions of x, y, $\mathfrak{r}_Q$, and ρ_w establishes (S9). □

I Proofs for unweighted pairwise Bellman validation

The proof uses the same selected-versus-comparator structure as the target-aligned argument. Under one-step coverage, the Bellman residual is bounded in $L^2(\nu)$; under pairwise coercivity, each candidate difference is controlled by its Bellman quadratic. By these two bounds, the selected candidate

has a lower score bound and the best candidate has a concave upper score bound. The finite-sample proof inserts a localized validation bound into this same comparison.

For each (i, j), define

$$\mathcal{M}_{ij}^{\mathrm{unw}} := \mathbb{E}_\nu[h_{ij}(X)\{Q_i(X) - R - \gamma(\pi Q_i)(S')\}] - \frac{1}{4}\mathbb{E}_\nu[h_{ij}(X)\{h_{ij}(X) - \gamma(\pi h_{ij})(S')\}]. \tag{S86}$$

The corresponding population selector is

$$\widetilde{i}_Q^{\mathrm{unw}} \in \arg\min_{1\le i\le K}\max_{1\le j\le K}\mathcal{M}_{ij}^{\mathrm{unw}}.$$

I.1 Population score bounds

Lemma S21 (Operator bounds under the offline distribution)**.** *Suppose Condition B5 and the one-step concentrability clause of Condition A3 hold. Then, for every $f \in L^2(\nu)$,*

$$\begin{aligned}\|P_\pi f\|_{2,\nu} &\le \sqrt{C_{\nu,\pi}}\|f\|_{2,\nu},\\ \|(I-\gamma P_\pi)f\|_{2,\nu} &\le (1+\gamma\sqrt{C_{\nu,\pi}})\|f\|_{2,\nu}.\end{aligned} \tag{S87}$$

In particular, all population scores in (S86) are finite.

Proof. By Jensen's inequality and the one-step concentrability clause of Condition A3,

$$\|P_\pi f\|_{2,\nu}^2 \le \int P_\pi(f^2)\,\mathrm{d}\nu = \int f^2\,\mathrm{d}(\nu P_\pi) \le C_{\nu,\pi}\|f\|_{2,\nu}^2,$$

which proves (S87). The second bound follows from the triangle inequality and (S87). It remains to verify that both expectations in (S86) are finite. By Condition B5, $Q_i, Q^\pi, h_{ij} \in L^2(\nu)$ and $\mathbb{E}_\nu[R^2] < \infty$. Moreover, by conditional Jensen's inequality and (S87),

$$\mathbb{E}_\nu\big[\{(\pi Q_i)(S')\}^2\big] \le \int P_\pi(Q_i^2)\,\mathrm{d}\nu \le C_{\nu,\pi}\|Q_i\|_{2,\nu}^2,$$

and the identical bound holds with Q_i replaced by h_{ij}. Thus, Cauchy–Schwarz applies separately to the products of h_{ij} with Q_i, R, $(\pi Q_i)(S')$, and $(\pi h_{ij})(S')$. Each term in (S86) is therefore absolutely integrable. □

For the remaining population argument, write

$$\Delta_i := Q_i - Q^\pi, \qquad T_i^{\mathrm{unw}} := \max_{1\le j\le K}\mathcal{M}_{ij}^{\mathrm{unw}},$$

and define the bilinear form

$$\mathcal{B}_\nu(f,g) := \langle f, (I-\gamma P_\pi)g\rangle_\nu.$$

By Lemma S21 and Cauchy–Schwarz,

$$|\mathcal{B}_\nu(f,g)| \le (1+\gamma\sqrt{C_{\nu,\pi}})\|f\|_{2,\nu}\|g\|_{2,\nu}. \tag{S88}$$

Lemma S22 (Deterministic unweighted score bounds)**.** *Suppose Conditions A3 and B5 hold. For every pair (i,k),*

$$T_i^{\mathrm{unw}} \ge \frac{3\kappa_\nu}{4}\|h_{ik}\|_{2,\nu}^2 - (1+\gamma\sqrt{C_{\nu,\pi}})\|h_{ik}\|_{2,\nu}\|\Delta_k\|_{2,\nu}. \tag{S89}$$

Moreover, for every k, j,

$$\mathcal{M}_{kj}^{\mathrm{unw}} \le (1+\gamma\sqrt{C_{\nu,\pi}})\|\Delta_k\|_{2,\nu}\|h_{kj}\|_{2,\nu} - \frac{\kappa_\nu}{4}\|h_{kj}\|_{2,\nu}^2, \tag{S90}$$

and hence

$$T_k^{\mathrm{unw}} \le \frac{(1+\gamma\sqrt{C_{\nu,\pi}})^2}{\kappa_\nu}\|\Delta_k\|_{2,\nu}^2. \tag{S91}$$

Proof. By the Bellman fixed-point equation and iterated expectation,

$$\mathcal{M}_{ij}^{\mathrm{unw}} = \mathcal{B}_\nu(h_{ij}, \Delta_i) - \frac{1}{4}\mathcal{B}_\nu(h_{ij}, h_{ij}). \tag{S92}$$

Because $\Delta_i = h_{ik} + \Delta_k$, by (S92),

$$\mathcal{M}_{ik}^{\mathrm{unw}} = \frac{3}{4}\mathcal{B}_\nu(h_{ik}, h_{ik}) + \mathcal{B}_\nu(h_{ik}, \Delta_k).$$

By Condition A3, the quadratic term is bounded below, and the absolute cross term is bounded by (S88). Since $T_i^{\mathrm{unw}} \geq \mathcal{M}_{ik}^{\mathrm{unw}}$, this establishes (S89).

With first index k, (S92) reduces to

$$\mathcal{M}_{kj}^{\mathrm{unw}} = \mathcal{B}_\nu(h_{kj}, \Delta_k) - \frac{1}{4}\mathcal{B}_\nu(h_{kj}, h_{kj}).$$

The first term is bounded above by (S88); after multiplication by $-1/4$, the second term is bounded above under Condition A3. Thus, (S90) holds. Completing the square,

$$\begin{aligned}
&(1+\gamma\sqrt{C_{\nu,\pi}})\|\Delta_k\|_{2,\nu}u - \frac{\kappa_\nu}{4}u^2 \\
&\quad = \frac{(1+\gamma\sqrt{C_{\nu,\pi}})^2}{\kappa_\nu}\|\Delta_k\|_{2,\nu}^2 - \frac{\kappa_\nu}{4}\left\{u - \frac{2(1+\gamma\sqrt{C_{\nu,\pi}})}{\kappa_\nu}\|\Delta_k\|_{2,\nu}\right\}^2.
\end{aligned}$$

Maximizing over $u \geq 0$ therefore gives (S91). □

Theorem S23 (Population oracle inequality for unweighted PBV)**.** *Suppose Conditions A3 and B5 hold. Then every population minimizer $\widetilde{i}_Q^{\mathrm{unw}}$ satisfies*

$$\|Q_{\widetilde{i}_Q^{\mathrm{unw}}} - Q^\pi\|_{2,\nu} \leq \left\{1 + \frac{2(1+\gamma\sqrt{C_{\nu,\pi}})}{\kappa_\nu}\right\} \min_{1\leq i\leq K}\|Q_i - Q^\pi\|_{2,\nu}. \tag{S93}$$

Proof. Choose $k \in \arg\min_i \|\Delta_i\|_{2,\nu}$, and set

$$u := \|h_{\widetilde{i}_Q^{\mathrm{unw}},k}\|_{2,\nu}, \qquad y := \|\Delta_k\|_{2,\nu}.$$

By population optimality, $T_{\widetilde{i}_Q^{\mathrm{unw}}}^{\mathrm{unw}} \leq T_k^{\mathrm{unw}}$. Combining this comparison with Lemma S22,

$$\frac{3\kappa_\nu}{4}u^2 - (1+\gamma\sqrt{C_{\nu,\pi}})uy \leq \frac{(1+\gamma\sqrt{C_{\nu,\pi}})^2}{\kappa_\nu}y^2. \tag{S94}$$

Solving (S94) for its nonnegative root gives $u \leq 2(1+\gamma\sqrt{C_{\nu,\pi}})y/\kappa_\nu$. Hence, by the triangle inequality,

$$\|\Delta_{\widetilde{i}_Q^{\mathrm{unw}}}\|_{2,\nu} \leq u + y \leq \left\{1 + \frac{2(1+\gamma\sqrt{C_{\nu,\pi}})}{\kappa_\nu}\right\} y.$$

Since k minimizes $\|\Delta_i\|_{2,\nu}$, (S93) follows. □

I.2 Concentration bound for the unweighted Q-function criterion

Lemma S24 (Localized unweighted-score deviation)**.** *Suppose Conditions A1, B2, and B3 hold. There is a universal constant $C_0 > 0$ such that, for every $\delta \in (0,1)$ and every nonempty collection $\mathcal{P} \subseteq \{1,\dots,K\}^2$ fixed conditional on the training sample, with probability at least $1-\delta$, simultaneously for all $(i,j) \in \mathcal{P}$,*

$$\left|\widehat{\mathcal{M}}_{ij}^{\mathrm{unw}} - \mathcal{M}_{ij}^{\mathrm{unw}}\right| \leq C_0\left(G_Q + \sigma_Q + \frac{H_Q}{2}\right)\left\{\|h_{ij}\|_{2,\nu}\sqrt{\frac{\log(6|\mathcal{P}|/\delta)}{ng_b}} + H_Q\frac{\log(6|\mathcal{P}|/\delta)}{ng_b}\right\}. \tag{S95}$$

Proof. Condition on the training sample, and let $\mathbb{P}_{n,Q}$ denote the empirical average over the Q-validation transitions. For a fixed ordered pair (i, j), set

$$\mu_i(X) := \mathbb{E}[\varepsilon_i \mid X], \qquad \xi_i := \varepsilon_i - \mu_i(X),$$

and define

$$d_{ij}(X) := \mu_i(X) - \frac{1}{4}\{h_{ij}(X) - \gamma(P_\pi h_{ij})(X)\},$$
$$\eta_{ij} := \frac{\gamma}{4}\{(\pi h_{ij})(S') - (P_\pi h_{ij})(X)\}, \qquad \zeta_{ij} := \xi_i + \eta_{ij}.$$

For this pair, the score summand decomposes as

$$Z_{ij} := h_{ij}(X)\{d_{ij}(X) + \zeta_{ij}\}, \qquad \widehat{\mathcal{M}}_{ij}^{\mathrm{unw}} = \mathbb{P}_{n,Q} Z_{ij}, \qquad \mathcal{M}_{ij}^{\mathrm{unw}} = \mathbb{E} Z_{ij}.$$

By the validation setup in Section 2.1, $\mathbb{E}(\zeta_{ij} \mid X) = 0$, and its conditional law is governed by the time-homogeneous MDP kernel. Under Condition B2,

$$|d_{ij}(X)| \le G_Q + \frac{H_Q}{2}, \qquad |\eta_{ij}| \le \frac{H_Q}{2}.$$

By $(x+y)^2 \le 2x^2 + 2y^2$, Condition B3, and conditional Jensen's inequality,

$$\mathbb{E}\left[\exp\left\{\frac{\zeta_{ij}^2}{8(\sigma_Q + H_Q/2)^2}\right\}\middle| X\right] \le e^{1/4}\left\{\mathbb{E}\left[\exp\left\{\frac{\xi_i^2}{\sigma_Q^2}\right\}\middle| X\right]\right\}^{1/4} \le (2e)^{1/4} < 2.$$

We may therefore apply Lemma S8(ii) directly to Z_{ij} with

$$a = h_{ij}, \quad d = d_{ij}, \qquad \zeta = \zeta_{ij},$$
$$B = H_Q, \quad D = G_Q + \frac{H_Q}{2}, \quad \sigma = \sqrt{8}\left(\sigma_Q + \frac{H_Q}{2}\right),$$
$$v = \|h_{ij}\|_{2,\nu}.$$

Apply the lemma with $u = \log(6|\mathcal{P}|/\delta)$ to each pair and take a union bound over $\mathcal{P}$. After substituting the displayed choices of B, D, σ, and v and enlarging the universal constant C_0, we obtain (S95). □

I.3 Finite-sample oracle inequality

Let C_0 be the universal constant in Lemma S24, and define

$$C_{Q,\mathrm{unw}} := 2C_0\left(G_Q + \sigma_Q + \frac{H_Q}{2}\right) + \frac{2}{3}\sqrt{6C_0 H_Q\left(G_Q + \sigma_Q + \frac{H_Q}{2}\right)}.$$

Proof of Theorems 5.2 and S6. Localized event. Let k be the smallest index in $\arg\min_i \|\Delta_i\|_{2,\nu}$, and define

$$\mathcal{P}_k^{\mathrm{unw}} := \{(i,k) : 1 \le i \le K\} \cup \{(k,j) : 1 \le j \le K\}.$$

The collection $\mathcal{P}_k^{\mathrm{unw}}$ is fixed conditional on the training sample and has at most $2K-1$ elements. By Lemma S24, with conditional probability at least $1-\delta$, the localized bound holds on $\mathcal{P}_k^{\mathrm{unw}}$ with its logarithmic factor enlarged to $\log(12K/\delta)$. On this event, abbreviate

$$u := \|h_{\widehat{i}_Q^{\mathrm{unw}},k}\|_{2,\nu}, \qquad y := \|\Delta_k\|_{2,\nu}, \qquad \mathfrak{r}_Q := \sqrt{\frac{\log(12K/\delta)}{n g_b}}.$$

Also set

$$\eta := C_0\left(G_Q + \sigma_Q + \frac{H_Q}{2}\right)\mathfrak{r}_Q, \qquad v := C_0 H_Q\left(G_Q + \sigma_Q + \frac{H_Q}{2}\right)\mathfrak{r}_Q^2.$$

Empirical comparison. For the selected candidate, by (S89) and the localized deviation bound,

$$\max_j \widehat{\mathcal{M}}_{\widehat{i}_Q^{\mathrm{unw}},j}^{\mathrm{unw}} \ge \frac{3\kappa_\nu}{4}u^2 - \{(1 + \gamma\sqrt{C_{\nu,\pi}})y + \eta\}u - v. \tag{S96}$$

For the comparator k, each population score is bounded by (S90), and the corresponding empirical deviation is bounded on the localized event. Thus, for every j,

$$\widehat{\mathcal{M}}_{kj}^{\mathrm{unw}} \le \{(1+\gamma\sqrt{C_{\nu,\pi}})y+\eta\}\|h_{kj}\|_{2,\nu} - \frac{\kappa_\nu}{4}\|h_{kj}\|_{2,\nu}^2 + v.$$

Maximizing the right-hand side over the nonnegative value of $\|h_{kj}\|_{2,\nu}$ yields

$$\max_j \widehat{\mathcal{M}}_{kj}^{\mathrm{unw}} \le \frac{\{(1+\gamma\sqrt{C_{\nu,\pi}})y+\eta\}^2}{\kappa_\nu} + v. \tag{S97}$$

Quadratic/root bound. By the defining optimality of $\widehat{i}_Q^{\mathrm{unw}}$ and (S96)–(S97),

$$\frac{3\kappa_\nu}{4}u^2 - \{(1+\gamma\sqrt{C_{\nu,\pi}})y+\eta\}u - \frac{\{(1+\gamma\sqrt{C_{\nu,\pi}})y+\eta\}^2}{\kappa_\nu} - 2v \le 0. \tag{S98}$$

Solving (S98) and using $\sqrt{a+b} \le \sqrt{a}+\sqrt{b}$, we obtain

$$u \le \frac{2\{(1+\gamma\sqrt{C_{\nu,\pi}})y+\eta\}}{\kappa_\nu} + \frac{2\sqrt{6}}{3}\sqrt{\frac{v}{\kappa_\nu}}.$$

Finally, by the triangle inequality, the bound $\kappa_\nu \le 1$ in Condition A3, and the definitions of η, v, and $C_{Q,\mathrm{unw}}$,

$$\|\Delta_{\widehat{i}_Q^{\mathrm{unw}}}\|_{2,\nu} \le \left\{1 + \frac{2(1+\gamma\sqrt{C_{\nu,\pi}})}{\kappa_\nu}\right\} y + \frac{C_{Q,\mathrm{unw}}}{\kappa_\nu}\mathfrak{r}_Q.$$

Since k minimizes $\|\Delta_i\|_{2,\nu}$, substituting the definitions of y and $\mathfrak{r}_Q$ establishes (S10). □

J Proofs for joint TA-PBV and doubly robust OPE

J.1 Proof of the joint TA-PBV guarantee

Proof of Theorem 5.1. Apply Theorem S3 to the shared validation trajectory with sample size n and failure probability $\delta/2$. Using its exact constants and $1 \le (1-\sqrt{\gamma})^{-1}$, we obtain (6) for a constant C_{ratio} with the stated dependence. On that event, by Condition A2 and Lemma S11,

$$\begin{aligned}\left[\mathbb{E}_\nu\left\{\frac{\{\widehat{w}(X)-w^\pi(X)\}^2}{w^\pi(X)}\right\}\right]^{1/2} &\le \sqrt{2\Lambda_w}\,\mathsf{D}_\nu(\widehat{w}\|w^\pi)^{1/2} \\ &\le \sqrt{2\Lambda_w}\,C_{\mathrm{ratio}}\left\{\frac{\epsilon_w}{1-\sqrt{\gamma}} + \frac{r_w(\delta/2)}{1-\gamma}\right\}.\end{aligned}$$

The integrand on the left is defined to be zero on $\{x : w^\pi(x) = 0\}$, where both functions vanish.

For each $a \in [M]$, let $\widehat{i}_Q(a)$ be the Q-index selected from the shared trajectory when the second-stage score uses the fixed weight w_a. Conditional on training, every w_a is fixed and independent of the validation trajectory. Apply Theorem S5 with $\widehat{w} = w_a$, sample size n, and failure probability $\delta/(2M)$. By a union bound over a, there is an event of probability at least $1-\delta/2$ on which the conclusion holds simultaneously for all $a \in [M]$. The logarithmic factor in that conclusion is $\log\{12K/(\delta/(2M))\} = \log(24MK/\delta)$, which equals the factor in $r_{Q,\mathrm{TA}}(\delta/2)$.

On this simultaneous event, take $a = \widehat{i}_w$; this is valid even though $\widehat{i}_w$ was selected from the same trajectory. The resulting selector is exactly $\widehat{i}_Q$, and its relative weight error is bounded by the first display in this proof on the ratio event. Intersecting the two events and using $1+\sqrt{\gamma} \le 2$, then enlarging C_{TA} to absorb the fixed structural constants, we obtain (7). The two events need not be independent; their intersection has probability at least $1-\delta$ by a union bound.

□

J.2 Proof of the doubly robust oracle bound

Proof of Corollary S7. For this proof, set $\ell = \log\{12n(K \vee M)/\delta\}$, so that $\rho_\delta = \ell/\sqrt{ng_b}$.

Selection events. We first control both selected nuisance functions on the shared selection fold. On $\mathcal{D}_{\mathrm{sel}}$, the hypotheses of Theorem S3 hold by Algorithm S2 and Conditions D1 and D2. Apply that theorem with failure probability $\delta/3$. Since $1/(1-\sqrt{\gamma}) \le 2/(1-\gamma)$ and

$$\sqrt{\log(12n/\delta)\log(72M/\delta)} \le \sqrt{2}\,\ell,$$

there is an event $\mathcal{E}_w$ with conditional probability at least $1-\delta/3$ on which

$$\sqrt{\mathsf{D}_\nu(\widehat{w}\|w^\pi)} \lesssim \frac{\epsilon_w + \rho_\delta}{1-\gamma}. \tag{S99}$$

The logarithmic factors in the Q-selection and evaluation bounds are also bounded by universal multiples of ℓ. To convert the generalized-KL bound into the relative L^2 error required for Q-selection, set $z = \widehat{w}/w^\pi$ on $\mathcal{X}^\pi = \{x : w^\pi(x) > 0\}$ and set $z = 1$ off $\mathcal{X}^\pi$. By Condition D2 and Lemma S11,

$$\begin{aligned}\mathbb{E}_\nu\left[\frac{\{\widehat{w}(X) - w^\pi(X)\}^2}{w^\pi(X)}\right] &= \mathbb{E}_\nu\left[w^\pi(X)\{z(X)-1\}^2\right] \\ &\le 2\Lambda_w \mathsf{D}_\nu(\widehat{w}\|w^\pi).\end{aligned} \tag{S100}$$

Both $\widehat{w}$ and w^π vanish off $\mathcal{X}^\pi$ by Condition D2, so the integrand on the left is defined to be zero there.

For each $a \in [M]$, let $\widehat{i}_Q(a)$ be the Q-index selected on $\mathcal{D}_{\mathrm{sel}}$ when the second-stage score uses w_a. Conditional on training, w_a is fixed and independent of the selection trajectory. By Conditions D1–D3, we may apply Theorem S5 to each w_a with failure probability $\delta/(3M)$. By a union bound, there is an event $\mathcal{E}_Q$ of conditional probability at least $1-\delta/3$ on which all M conclusions hold. In particular, the conclusion holds for $a = \widehat{i}_w$, even though that index uses the same selection trajectory. The best-candidate and validation terms are of order $\epsilon_Q/(1-\gamma)$ and $\rho_\delta/(1-\gamma)$, respectively. By (S99) and (S100), the fitted-weight term is of order $(\epsilon_w+\rho_\delta)/(1-\gamma)^2$. Consequently, on $\mathcal{E}_w \cap \mathcal{E}_Q$,

$$\|\widehat{Q} - Q^\pi\|_{2,d_{\pi,\gamma}} \lesssim \frac{\epsilon_Q + \rho_\delta}{1-\gamma} + \frac{\epsilon_w + \rho_\delta}{(1-\gamma)^2}. \tag{S101}$$

Doubly robust remainder. We next express the population value error as a product of the two nuisance errors. For fixed candidates $Q \in \{Q_1, \dots, Q_K\}$ and $w \in \{w_1, \dots, w_M\}$, define the population functional

$$\Psi(Q,w) := \mathbb{E}_{d_0}[Q(X)] + \frac{1}{1-\gamma}\mathbb{E}_\nu[w(X)\{R + \gamma(\pi Q)(S') - Q(X)\}].$$

Let $\Delta_Q = Q - Q^\pi$. By the Bellman equation (1) and the fixed-point moment identity obtained from (S1) with $w = w^\pi$ and $f = \Delta_Q$, the following expansion holds. Its expectations are finite by Conditions D2 and D3, together with Lemma S18.

$$\begin{aligned}\Psi(Q,w) - V^\pi &= \mathbb{E}_{d_0}[\Delta_Q(X)] - \frac{1}{1-\gamma}\mathbb{E}_\nu[w(X)(I-\gamma P_\pi)\Delta_Q(X)] \\ &= \frac{1}{1-\gamma}\mathbb{E}_\nu[\{w^\pi(X) - w(X)\}(I-\gamma P_\pi)\Delta_Q(X)].\end{aligned}$$

In particular, the bias of Ψ vanishes if either $Q = Q^\pi$ or $w = w^\pi$; hence, Ψ is doubly robust.

By weighted Cauchy–Schwarz and Lemma S16,

$$\begin{aligned}|\Psi(Q,w) - V^\pi| &\le \frac{1}{1-\gamma}\left\{\mathbb{E}_\nu\left[\frac{\{w(X) - w^\pi(X)\}^2}{w^\pi(X)}\right]\right\}^{1/2} \|(I-\gamma P_\pi)\Delta_Q\|_{2,d_{\pi,\gamma}} \\ &\le \frac{1+\sqrt{\gamma}}{1-\gamma}\left\{\mathbb{E}_\nu\left[\frac{\{w(X) - w^\pi(X)\}^2}{w^\pi(X)}\right]\right\}^{1/2} \|Q - Q^\pi\|_{2,d_{\pi,\gamma}}.\end{aligned} \tag{S102}$$

For $(Q,w) = (\widehat{Q}, \widehat{w})$, substitute the relative weight bound (S99) and the Q-error bound (S101) into (S102). We obtain

$$\begin{aligned}|\Psi(\widehat{Q},\widehat{w}) - V^\pi| &\lesssim \frac{\epsilon_w + \rho_\delta}{(1-\gamma)^2}\left\{\frac{\epsilon_Q + \rho_\delta}{1-\gamma} + \frac{\epsilon_w + \rho_\delta}{(1-\gamma)^2}\right\} \\ &\lesssim \frac{\epsilon_w\epsilon_Q}{(1-\gamma)^3} + \frac{\epsilon_w^2}{(1-\gamma)^4} + \left\{\frac{\epsilon_Q}{(1-\gamma)^3} + \frac{\epsilon_w}{(1-\gamma)^4}\right\}\rho_\delta + \frac{\rho_\delta^2}{(1-\gamma)^4},\end{aligned} \tag{S103}$$

where the second inequality follows from $1-\gamma \le 1$.

Evaluation error. Finally, conditional on all preceding folds, $\widehat{Q}$ and $\widehat{w}$ are fixed and independent of $\mathcal{D}_V$ by Algorithm S2. Under the empirical branch of Condition D4, by Hoeffding's inequality, with conditional probability at least $1-\delta/6$,

$$|\mathbb{P}_{n,0,V}\widehat{Q} - \mathbb{E}_{d_0}\widehat{Q}| \lesssim \rho_\delta. \tag{S104}$$

Under the exact-moment branch, both sides of (S104) are zero by convention. For each $i \in [K]$, set

$$\mu_i(X) := \mathbb{E}[\varepsilon_i \mid X], \qquad \xi_i := \varepsilon_i - \mu_i(X).$$

For the transition term, set

$$Y_t := \widehat{w}(X_t^V)\{R_t^V + \gamma(\pi\widehat{Q})(S_t'^{,V}) - \widehat{Q}(X_t^V)\}.$$

The representation in Lemma S8(ii) holds with

$$a = \widehat{w}, \qquad d = -\mu_{\widehat{i}_Q}, \qquad \zeta = -\xi_{\widehat{i}_Q}.$$

The spectral-gap, pointwise, envelope, and conditional sub-Gaussian requirements hold by Conditions D1–D3. Moreover, by Condition D2 and $\mathbb{E}_\nu[w^\pi(X)] = 1$,

$$\|\widehat{w}\|_{2,\nu}^2 \le W_Q \mathbb{E}_\nu[\widehat{w}(X)] \le W_Q \Lambda_w.$$

By Lemma S8(ii), with $u = \log(36/\delta)$, the following bound holds with conditional probability at least $1-\delta/6$:

$$|\mathbb{P}_{n,V}Y - \mathbb{E}Y| \lesssim \rho_\delta + \rho_\delta^2. \tag{S105}$$

On the two corresponding events, by (S104) and (S105),

$$|\widehat{V}_{\mathrm{DR}} - \Psi(\widehat{Q}, \widehat{w})| \lesssim \frac{\rho_\delta + \rho_\delta^2}{1-\gamma}.$$

Adding this evaluation error to the population remainder in (S103), we obtain (S12). Here $1-\gamma \le 1$ is used only to absorb $\rho_\delta^2/(1-\gamma)$ into $\rho_\delta^2/(1-\gamma)^4$. By a sequential union bound over $\mathcal{E}_w$, $\mathcal{E}_Q$, and the two final evaluation events, the total failure probability is at most $\delta/3 + \delta/3 + \delta/6 + \delta/6 = \delta$, which proves the corollary. □