PRESERVING EVERY TARGET ROW DOES NOT PRESERVE PROPAGATED UNCERTAINTY IN RECURSIVE FITTED EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Bootstrap procedures use ensembles of value functions to estimate uncertainty in a policy's value through recursive fitted evaluation. Fixing the target vectors leaves unspecified which model fits each vector, a choice implementations make implicitly. We show that reassigning those vectors among models can collapse ensemble spread by changing their dependence on the models' bootstrap resamples and fitting histories. We prove that reassigning target vectors can change the limiting distribution as ensemble size grows at fixed recursion depth and derive exact finite-ensemble moment recursions for affine fits. In Bootstrap-FQE experiments across six environments, reassignment at every update reduces spread on all 105 datasets and lowers observed coverage of nominal 90% intervals from 88–100% to 4–30% against the evaluation references. Indexing choices that determine which model fits each vector therefore belong to the ensemble's statistical specification.

1 INTRODUCTION

Bootstrap ensembles estimate uncertainty in a policy's value by measuring how the estimate varies across resampled datasets (Hao et al., 2021). In recursive fitted evaluation, each bootstrap replicate, or member, requires a sequence of regressions for which targets are computed from earlier fitted predictions (Ernst et al., 2005; Munos & Szepesvári, 2008). A member's resample consequently influences both its current fit and the targets used in subsequent updates. The spread of the final estimates reflects this recursive propagation of uncertainty. Bootstrapped fitted Q evaluation (Bootstrap-FQE) uses that spread to form confidence intervals, while related ensemble methods use propagated uncertainty to guide exploration and set pessimism in policy optimization (Osband et al., 2016; An et al., 2021; Ghasemipour et al., 2022).

The construction of regression targets is therefore consequential for uncertainty estimation. Comparisons between member-specific and shared pessimistic targets show that target rules can substantially affect ensemble uncertainty and pessimism (Ghasemipour et al., 2022; Li et al., 2026). These comparisons, however, change target values together with their assignment to members. They cannot distinguish the effect of replacing predictions from the effect of changing which member fits each target vector, the targets over the training inputs. Whether preserving the complete collection of target vectors suffices to preserve propagated uncertainty therefore remains open.

We show that ensemble spread and coverage can collapse even when every target vector is preserved. When a member fits targets from its own value function, the same resample and fitting history shape both the supplied vector and the regression that fits it. We change this dependence by randomly permuting complete target vectors among members at each update while each recipient retains its bootstrap resample and fitting history. This intervention, which we call whole-row reassignment, preserves the supplied joint predictions. The distribution of target vectors therefore leaves unspecified an association that can affect the distribution of fitted values.

A second dependence arises within each target vector from one member's resample and fitting history, which shape its targets. When a fitted value draws on targets at several inputs, its variance across members depends on the covariances among those targets as well as on their individual variances. We show that this dependence, which we call coupling, can change the distribution of fitted values even when the target distribution at every input is fixed. Our coupling result extends to re-

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.

cursive fitting a finding that work on joint predictive distributions establishes for decision problems, where accurate marginal predictions can be insufficient (Wen et al., 2022; Osband et al., 2022; 2023).

We measure the effect of reassignment on spread and coverage by running Bootstrap-FQE under two fitting regimes. Exact fitting rules out optimization as the source of this effect, whereas neural fitting shows how large the effect is in practice. Under exact tabular fitting, whole-row reassignment reduces spread and coverage while leaving point estimates identical, with a spread reduction that matches the theoretical prediction. The agreement extends to fixed-feature ridge fits with member-specific fitting maps, which match the exact finite-ensemble moments we derive for affine fits. Under neural fitting across six environments, whole-row reassignment reduces spread on all 105 datasets, by median factors of 2.3 to 6.1, lowering coverage of nominal 90% intervals from 88–100% to 4–30% against the evaluation references. Point estimates can improve even as coverage and interval score deteriorate, so point-estimate accuracy does not reveal this failure. Nor does a single-update comparison, since reassignment can increase spread at one update while reducing it when applied at every update.

Empirically, both coupling and assignment affect spread and coverage across six evaluation environments. Changing assignment alone reduces final ensemble spread on all 105 datasets, lowering coverage of nominal 90% intervals from 88–100% to 4–30%. This reduction under repeated fitting occurs even though the same assignment change can increase spread in single-update comparisons. Coverage and interval score can deteriorate even while the root mean squared error of the point estimate decreases. Changing both coupling and assignment reduces spread by up to $28\times$, cutting coverage to 12% or below. The effects persist for ensembles of 10 to 100 members over a tenfold range of dataset sizes.

Our contributions are as follows:

1. We identify a dependence that the distribution of target vectors leaves unspecified, namely the association between each target vector and its recipient's retained state, which includes the resample, initialization, and fitting history.
2. We prove that assignment and coupling can each change the limiting distribution of value estimates as ensemble size grows at fixed depth, with the assignment result holding even when members retain their fitting histories. For finite ensembles of affine fits, we derive exact recursions for the mean and spread under reassignment at any depth.
3. We show that whole-row reassignment at every update collapses the spread and coverage of Bootstrap-FQE across six environments. Changing coupling as well deepens the collapse. Exact tabular fitting reproduces the whole-row collapse without optimization, reducing spread by the theoretically predicted factor while leaving point estimates identical.

2 Related Work

Joint predictions and target construction. Work on joint predictive distributions and epistemic neural networks distinguishes uncertainty representations that agree pointwise but differ when predictions are queried together (Wen et al., 2022; Osband et al., 2022; 2023). CBOP uses uncertainty in model-based value expansion to construct conservative Bellman targets (Feinberg et al., 2018; Buckman et al., 2018; Jeong et al., 2023). MSG compares shared pessimistic targets with independently computed member targets and analyzes their consequences for ensemble uncertainty (Ghasemipour et al., 2022). Its default implementation retains each critic's initialization and corresponding target row while redrawing random loss weights, represented by the update noise η_t in Equation 1. Its default therefore uses own-row assignment (Appendix D.2). Our whole-row intervention preserves the complete empirical distribution of incoming target vectors and changes only their recipient fitters. It isolates association with persistent state while holding the supplied joint predictions intact.

Off-policy inference. Bootstrap-FQE resamples data and refits complete value functions to quantify policy-value uncertainty (Hao et al., 2021). Differentiable FQE inference obtains uncertainty estimates under assumptions on the fitted procedure (Zhang et al., 2022). Our intervention holds the data and fitting settings fixed and changes how bootstrap members receive their targets. This comparison identifies a dependence required to specify the propagated ensemble law, to which an inferential argument then applies.

Bagging and propagated spread. Bagging averages predictors fit to bootstrap resamples (Breiman, 1996), and deep ensembles use disagreement among independently trained members as predictive uncertainty (Lakshminarayanan et al., 2017). In either case, members are fit independently, so relabeling them leaves both the aggregate and the member distribution unchanged. In recursive fitting, members also supply one another's regression targets, so reassignment changes the fitted members themselves and the law to which they converge. The distinction matters when member spread serves as the uncertainty estimate, because reassignment can improve point-estimate accuracy while worsening coverage and interval score (Section 4.1).

3 Coupling and Assignment in Recursive Fitting

We formalize coupling and assignment, then show that changing assignment alone can alter ensemble spread while preserving every incoming target vector. This effect can persist as ensemble size grows at fixed recursion depth. For affine fits, we derive exact finite-ensemble moment recursions.

3.1 Target rows and retained member state

Fix a dataset, an evaluation policy π, a recursion depth T, and the fitting hyperparameters. At update t, member b supplies a target vector $X_t^{(b)} \in \mathbb{R}^{d_t}$ containing one regression target per training input. In fitted Q evaluation (FQE), this vector $X_t^{(b)} = (y_{i,t}^{(b)})_{i=1}^{d_t}$ contains the Bellman targets computed from the member's fitted value function $Q_t^{(b)}$,

$$y_{i,t}^{(b)} = r_i + \gamma \mathbb{E}_{a \sim \pi(\cdot|s_i')} Q_t^{(b)}(s_i', a),$$

where r_i and s_i' are the observed reward and successor state, γ is the discount, and the continuation term is zero at terminal states. Each member fits its assigned vector using its fixed bootstrap resample D_b. Its *retained state* $R_t^{(b)}$ contains that resample and any fitting history carried into the update, including network weights and optimizer state.

Stacking the M target vectors as rows gives a target matrix in $\mathbb{R}^{M \times d_t}$. We distinguish three rules for constructing the assigned rows (Figure 1). *Own-row* assigns each member the row that the member supplied. *Whole-row* applies a common permutation to all columns, reassigning complete rows among members. *Coordinatewise* applies an independent permutation to each column, preserving the values at each training input while allowing their grouping into rows to change. During reassignment, each recipient keeps its bootstrap resample and current fitting state.

Writing $Y_t^{(b),c}$ for the row assigned to member b under rule c, we have

$$Y_t^{(b),\mathrm{own}} = X_t^{(b)}, \qquad Y_t^{(b),\mathrm{whole}} = X_t^{(\sigma_t(b))}, \qquad \big(Y_t^{(b),\mathrm{coord}}\big)_i = \big(X_t^{(\sigma_{t,i}(b))}\big)_i.$$

Here $i \in \{1, \dots, d_t\}$ indexes training inputs. Whole-row assignment uses a single uniform permutation σ_t of the M member indices; coordinatewise assignment uses independent uniform permutations $\sigma_{t,i}$, one per input. Under either rule, the permutations at each update are sampled independently of earlier permutations, the incoming target matrix, and retained states.

Coordinatewise reassignment can change spread by changing dependence across the targets within a row. For example, the variance of a prediction that averages two targets depends on their covariance (Appendix A.1). Whole-row reassignment preserves every complete row and therefore preserves this dependence. Comparing own-row with whole-row assignment isolates whether spread also depends on which retained member state receives each row.

3.2 Intact rows can yield different spread

Consider two members with fixed fitting maps $\Phi_+(x) = 1 + \gamma x$ and $\Phi_-(x) = -1 + \gamma x$, where $0 < \gamma < 1$. Each row contains a single target, so there is no within-row coupling to change. Starting from target rows equal to zero, the first update produces outputs $+1$ and -1, which become the target rows for the second update.

At the second update, own-row assignment produces outputs $(1+\gamma, -1-\gamma)$. Whole-row assignment samples a uniform permutation of the two incoming targets, producing this same output pair with

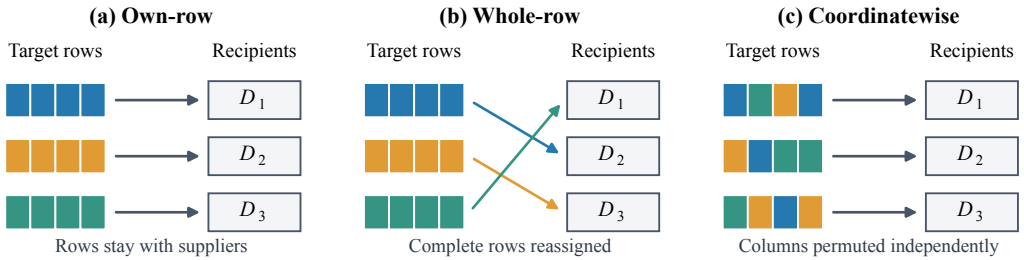

Figure 1: Coupling and assignment on a target array. Colors identify the members supplying target values; boxes identify the recipients, which retain their bootstrap resamples D_b and fitting histories during reassignment. Own-row keeps each row with its supplier. Whole-row reassigns complete rows, preserving the values grouped within each row. Coordinatewise permutes each column independently, preserving the values at each training input while allowing their grouping into rows to change.

probability $1/2$ and $(1-\gamma, -1+\gamma)$ with probability $1/2$. Every output pair has mean zero. Let V_c denote the variance across the two fitted outputs under rule c, using denominator two. Taking expectation over the whole-row permutation gives

$$V_{\text{own}} = (1+\gamma)^2,$$
$$\mathbb{E}[V_{\text{whole}}] = \tfrac{1}{2}(1+\gamma)^2 + \tfrac{1}{2}(1-\gamma)^2 = 1+\gamma^2.$$

Whole-row assignment therefore reduces the expected ensemble variance by 2γ. Both rules use the same incoming targets and fitting maps, so this difference comes entirely from changing which member receives each target row. Preserving complete target rows is therefore insufficient to preserve ensemble spread.

3.3 Population recursions with retained fitting history

The two-member example shows that assignment can change spread in a finite ensemble. To determine whether this effect persists as M grows, we characterize the joint probability distribution of a member's retained state and supplied target row in an idealized infinite ensemble.

For any joint distribution μ of retained state R and supplied target row $X \in \mathbb{R}^d$, let $\mathcal{A}_c(\mu)$ denote the joint distribution of R and the row Y assigned to the same member under rule c. Write μ^R, μ^X, and μ^{X_j} for the marginals of R, X, and its jth coordinate. With $\otimes$ denoting independent coupling, the population assignment rules are

$$\mathcal{A}_{\text{own}}(\mu) = \mu, \qquad \mathcal{A}_{\text{whole}}(\mu) = \mu^R \otimes \mu^X, \qquad \mathcal{A}_{\text{coord}}(\mu) = \mu^R \otimes \bigotimes_{j=1}^{d} \mu^{X_j}.$$

Whole-row assignment makes the assigned row independent of the recipient state while preserving the joint distribution of target values within a row. Coordinatewise assignment additionally makes the assigned coordinates independent.

After assignment, fitting updates the recipient's retained state and produces the next supplied row. Write $R_t = (S_t, \xi)$, where ξ contains member-specific information held fixed across updates and S_t contains the evolving fitting state. Let h_t map the current fitting state to the supplied target row, and let F_t update the fitting state using the assigned row Y_t, ξ, and member-specific update noise η_t. The member dynamics are

$$X_t = h_t(S_t), \qquad S_{t+1} = F_t(S_t, Y_t, \xi, \eta_t). \tag{1}$$

Because the assigned row enters S_{t+1}, reassignment can affect both the next target row and subsequent fits.

Let $\mathcal{K}_t((r,y),\cdot)$ denote the law of (R_{t+1}, X_{t+1}) induced by these dynamics from retained state r and assigned row y. Starting from a common initial joint law, the population recursion under rule c

is

$$\mu_{t+1}^c = \big(\mathcal{A}_c(\mu_t^c)\big)\mathcal{K}_t, \qquad c \in \{\text{own}, \text{whole}, \text{coord}\}. \tag{2}$$

The fitting kernel is common across rules, so assignment affects the recursion through the joint law passed to fitting.

To express the resulting variation in a fitted quantity, let Ψ be a fixed real-valued Borel-measurable function of (R_T, X_T). We call Ψ a *readout*. The population recursion determines the distribution μ_T^c of (R_T, X_T) after T updates. We define the *propagated law* ν_T^c as the distribution of the readout under μ_T^c,

$$\nu_T^c = \mathrm{Law}_{\mu_T^c}\big(\Psi(R_T, X_T)\big).$$

In FQE, for example, the readout evaluates the fitted network encoded in R_T at the initial-state queries to form a policy-value estimate. The population recursion therefore specifies a distribution ν_T^c of policy-value estimates conditional on the fixed dataset. Assignment can change this distribution while the evaluation policy and policy-value estimand remain fixed. Applying the same readout to a finite ensemble gives M policy-value estimates. The following theorem establishes that their empirical distribution converges to ν_T^c as $M \to \infty$ with T fixed.

Theorem 3.1 (Assignment limits with retained fitting history)**.** *Fix a realization of any shared exogenous randomness, such as a minibatch schedule. Conditional on this realization, assume that all state and noise spaces are Polish and that h_t and F_t are Borel measurable in all arguments. The initial pairs $(S_0^{(b)}, \xi^{(b)})_{b=1}^M$ are i.i.d. The update noises $(\eta_t^{(b)})_{b=1}^M$ are i.i.d. across members, independent across times, and independent of the initial pairs. For whole-row and coordinatewise assignment, the permutations are uniform and independent across updates and, in the coordinatewise case, across coordinates. They are also independent of the initial pairs and update noises. The recursion depth T, row dimensions, maps, and laws of the initial pairs and update noises do not vary with M.*

For each rule c, define the empirical joint and readout laws by

$$\widehat{\mu}_{M,T}^c = \frac{1}{M}\sum_{b=1}^M \delta_{(R_T^{(b),c}, X_T^{(b),c})}, \qquad \widehat{\nu}_{M,T}^c = \frac{1}{M}\sum_{b=1}^M \delta_{\Psi(R_T^{(b),c}, X_T^{(b),c})}.$$

As $M \to \infty$, these laws converge weakly in probability to μ_T^c and ν_T^c, respectively. If $\mu_T^c(\Psi^2) < \infty$, the empirical readout mean and variance also converge in probability to the mean and variance under ν_T^c. See Appendix A.2 for the proof.

Distinct population variances would therefore establish that the assignment effect persists in the large-ensemble limit. A population version of the two-member example gives such a separation. Let each member's intercept be sampled independently from a centered distribution with variance $0 < \tau^2 < \infty$ and held fixed across updates. For a population member with intercept ξ, consider the recursion $X_{t+1} = \xi + \gamma Y_t$ with $X_0 = 0$ and $0 < \gamma < 1$.

Under own-row assignment, $Y_t = X_t$, so $X_T^{\text{own}} = \xi \sum_{s=0}^{T-1} \gamma^s$. Under whole-row assignment, Y_t^{whole} has the same distribution as X_t^{whole} and is independent of ξ at each population update. Hence $\mathrm{Var}(X_{t+1}^{\text{whole}}) = \tau^2 + \gamma^2 \mathrm{Var}(X_t^{\text{whole}})$. The depth-$T$ variances are therefore

$$\mathrm{Var}(X_T^{\text{own}}) = \tau^2 \left(\sum_{s=0}^{T-1} \gamma^s\right)^2, \qquad \mathrm{Var}(X_T^{\text{whole}}) = \tau^2 \sum_{s=0}^{T-1} \gamma^{2s}. \tag{3}$$

For $T \geq 2$, the own-row variance is strictly larger because the squared sum includes positive cross terms (Remark A.3). Both distributions have finite second moments, so Theorem 3.1, applied with $\Psi(R_T, X_T) = X_T$, implies that the empirical variances converge to these distinct limits as $M \to \infty$ at fixed T.

3.4 Exact moments for finite affine ensembles

The distinct limiting variances establish that assignment effects can persist as M grows, but they do not determine the expected spread of a particular finite ensemble. For affine fits with linear target construction, three ensemble moments give an exact recursion for that spread, conditional

on the members' fitting maps. Fix $M \geq 2$ and deterministic initial parameters. With linear target construction, whole-row updates take the form $\theta_b^+ = c_b + A_b\theta_{\sigma(b)}$, where $\theta_b \in \mathbb{R}^q$ is member b's parameter vector and σ is a uniform permutation sampled independently at each update.

Let $\bar{\theta} = M^{-1}\sum_b \theta_b$. Taking expectations over the assignment permutations, define

$$m = \mathbb{E}[\bar{\theta}], \qquad S = \mathrm{Cov}(\bar{\theta}), \qquad C = \mathbb{E}\left[\frac{1}{M}\sum_b (\theta_b - \bar{\theta})(\theta_b - \bar{\theta})^\top\right].$$

Here S describes variation of the ensemble mean across assignment realizations, while C describes expected spread within each ensemble.

To see why these moments suffice, condition on the incoming parameter array and write $C_{\mathrm{emp}} = M^{-1}\sum_b(\theta_b - \bar{\theta})(\theta_b - \bar{\theta})^\top$. Each recipient receives a uniformly selected parameter vector, with conditional mean $\bar{\theta}$ and covariance C_{emp}. Sampling without replacement also determines the dependence between distinct recipients,

$$\mathrm{Cov}\big(\theta_{\sigma(b)}, \theta_{\sigma(c)} \mid \theta_{1:M}\big) = -\frac{C_{\mathrm{emp}}}{M-1}, \qquad b \neq c.$$

These identities determine the conditional first and second moments after applying the fixed affine fitting maps. Averaging over the incoming ensemble then requires only $\mathbb{E}[\bar{\theta}] = m$, $\mathbb{E}[\bar{\theta}\bar{\theta}^\top] = S + mm^\top$, and $\mathbb{E}[C_{\mathrm{emp}}] = C$.

Write $\bar{A} = M^{-1}\sum_b A_b$, $\bar{c} = M^{-1}\sum_b c_b$, $\Delta A_b = A_b - \bar{A}$, and $\Delta c_b = c_b - \bar{c}$. Proposition A.6 in Appendix A.3 gives the exact whole-row recursion

$$\begin{aligned}
m^+ &= \bar{c} + \bar{A}m, \\
S^+ &= \bar{A}S\bar{A}^\top + \frac{1}{M(M-1)}\sum_b \Delta A_b C \Delta A_b^\top, \\
C^+ &= \frac{1}{M}\sum_b (\Delta c_b + \Delta A_b m)(\Delta c_b + \Delta A_b m)^\top \\
&\quad + \frac{1}{M}\sum_b A_b(S+C)A_b^\top - S^+.
\end{aligned}$$

The additional term in S^+ is the variation in the ensemble mean introduced by reassignment; it vanishes when all members share the same operator. Subtracting S^+ in the final equation accounts for centering member spread at the realized ensemble mean. The proposition also gives the coordinate-wise recursion in terms of the same three moments.

Iterating these equations determines the expected sample variance of any scalar readout $a^\top\theta_b$ as $Ma^\top C_T a/(M-1)$. Under the same fixed maps and initialization, own-row fitting produces a deterministic ensemble. Its member variance can therefore be compared exactly with the expected variance under reassignment at any finite depth and ensemble size.

4 Experimental Evaluation

Section 3 proves that coupling and assignment can change ensemble spread but leaves open whether spread changes in practice and whether the ensemble's confidence intervals still cover the true policy value. We first show that reassignment reduces spread and lowers coverage in Bootstrap-FQE with neural networks (Section 4.1). Because the spread reduction could arise from neural optimization rather than from the recursion, we then test whether reassignment still reduces spread under exact tabular fitting, which involves no optimization and thus leaves only the recursion as a possible cause of any reduction (Section 4.2). Finally, reassignment at one update changes both the current fit and the targets of every later update, so we test whether the effect of reassignment can be judged from a single update (Section 4.3).

Throughout, we define spread as the standard deviation (SD) of the members' policy-value estimates, obtained by averaging each member's predictions over the same logged initial states with equal weights. To test whether the intervals bounded by the 5th and 95th percentiles of the members'

	SD reduction		Coverage		
Environment	Whole	Coord.	Own-row	Whole	Coord.
CartPole	6.08	28.06	23/25	1/25	0/25
Cliff Walking	5.82	10.98	24/25	3/25	0/25
Sepsis	3.77	5.14	22/25	3/25	3/25
Hopper	2.27	7.20	10/10	3/10	0/10
Walker2d	5.14	15.78	10/10	1/10	0/10
HalfCheetah	3.59	6.59	10/10	1/10	0/10

Table 1: Neural-network ensembles. SD reduction is the median paired ratio of own-row to reassigned SD; coverage counts nominal 90% intervals containing the horizon-capped evaluation reference. Whole and Coord. denote whole-row and coordinatewise reassignment, which both reduce spread on all 105 datasets.

estimates contain the true policy value in 90% of datasets, we measure coverage, the fraction of datasets for which the interval contains the experiment's evaluation reference. Coverage is measured against the evaluation policy's return from rollouts truncated at a fixed horizon (Appendix B.3 varies the horizon). Because wide intervals can achieve high coverage, we also report interval scores, which jointly penalize interval width and missed references. Appendix B gives the full protocols.

4.1 Assignment alone reduces spread and coverage

We compare the three rules on 25 datasets from each of CartPole, Cliff Walking, and Sepsis and 10 from each D4RL (Fu et al., 2020) medium datasets for Hopper, Walker2d, and HalfCheetah. Each dataset contains $N = 10$ episodes, and each run fits $M = 50$ neural-network members. Paired runs on a dataset differ only in the rule, sharing initializations, bootstrap resamples, minibatch orders, and numbers of updates. Reassignment draws new permutations at every update, and recipients keep network weights and Adam state across updates.

Whole-row reassignment keeps every target vector intact yet reduces spread on all 105 datasets, with median paired reductions of $2.27\times$–$6.08\times$ across environments (Table 1; Appendix C.3 varies ensemble and dataset size). Across environments, coverage is 88%–100% under own-row assignment but only 4%–30% under whole-row reassignment, and the mean interval score worsens in every environment despite the narrower intervals. On Hopper, coverage falls even as the root mean squared error of the ensemble mean drops from 8.769 to 6.809 (Appendix D).

Changing coupling further reduces spread. Coordinatewise reassignment produces larger median SD reductions than whole-row reassignment in every environment (Table 1). Because coordinatewise reassignment also changes assignment, we change only coupling in a separate CartPole experiment, keeping own-row assignment and varying whether dropout masks are shared across inputs when members compute targets. Independent masks reduce spread relative to shared masks on all 25 datasets, although both procedures give the same prediction distribution at each input for fixed network weights (Appendix C.4). Dependence within target vectors therefore affects spread even under own-row assignment.

4.2 Assignment effects persist under exact fitting

To isolate the effects of reassignment in Section 4.1 from neural optimization, we use exact tabular fitting in a one-state, two-action MDP. Under own-row assignment, member b follows $x_{t+1}^{(b)} = \bar{r}_b + \gamma x_t^{(b)}$, with $\bar{r}_b$ the mean reward in member b's bootstrap resample, and whole-row reassignment permutes the continuation values $x_t^{(b)}$ while retaining each recipient's $\bar{r}_b$. Both updates follow the scalar recursion of Section 3.3 with intercept $\bar{r}_b$, so Equation 3 implies large-depth variances of $\tau^2/(1-\gamma)^2$ under own-row assignment and $\tau^2/(1-\gamma^2)$ under whole-row reassignment, with τ^2 the variance of $\bar{r}_b$ across members. The predicted own-row-to-whole-row spread ratio therefore approaches $\sqrt{(1+\gamma)/(1-\gamma)}$ as depth grows. We test the prediction at $\gamma \in \{0, 0.5, 0.8, 0.95\}$, fitting $M = 100$ members for 200 updates on each of 25 datasets (Appendix C.1).

	Own-row / whole-row spread		Coverage	
γ	Predicted RMS ratio	Measured SD ratio	Own-row	Whole-row
0	1.000	1.000 [1.000, 1.000]	21/25	21/25
0.5	1.732	1.753 [1.692, 1.776]	21/25	12/25
0.8	3.000	3.016 [2.901, 3.134]	21/25	6/25
0.95	6.245	6.212 [5.901, 6.445]	21/25	1/25

Table 2: Exact tabular fitting at depth 200. Predicted ratios compare RMS spreads from expected empirical variances; measured ratios are median paired SD ratios with 95% dataset-bootstrap intervals. Coverage counts nominal 90% intervals containing the true depth-200 value.

Reassignment still reduces spread, and the measured SD ratio closely follows the prediction, growing with the discount from one at $\gamma = 0$ (Table 2). Own-row spread is larger because own-row assignment repeatedly pairs $\bar{r}_b$ with a continuation value built from member b's resample, so resampling error in $\bar{r}_b$ accumulates. The point estimates, however, are identical, since all members share γ and permuting the continuation values leaves the ensemble mean unchanged.[1] Coverage of the true depth-200 value nevertheless falls at $\gamma = 0.95$ from 21/25 under own-row assignment to 1/25 under whole-row reassignment, so assignment changes spread and coverage even with exact fits and identical point estimates.

4.3 Single-update responses do not predict repeated effects

To test whether the effect of reassignment can be judged from a single update, we copy the own-row ensembles of Section 4.1 at seven checkpoints for CartPole, Cliff Walking, and Sepsis. Starting from identical incoming targets and recipient states, each copy takes one update under a different rule with matched fitting randomness. We compare the resulting changes in spread with runs that apply each rule at every update for 500 updates.

After one update, whole-row reassignment *increases* spread in 19 of the 21 environment–checkpoint comparisons (Figure 2a), whereas applying whole-row reassignment at every update *decreases* final spread relative to own-row assignment on every dataset in the three environments (Figure 2b). Coordinatewise reassignment reduces spread in both comparisons. A single update therefore does not reveal the effect of reassigning at every update.

5 Discussion and Limitations

Implications for practice. Our experiments show that the uncertainty Bootstrap-FQE reports depends on which member fits each target vector and on which predictions share a vector, even when the distribution of targets at each input stays unchanged. Published comparisons of target rules report changed outcomes but do not isolate or explain the dependency behind the change. Li et al. (2026) report excessive pessimism in several settings when ensemble FQE uses shared-minimum rather than per-critic targets, in a comparison that includes retrospective clinical data. Rerunning the official implementation on Walker2d-medium-v2, we find that the single configuration flag controlling the switch reduces the reported standard deviation from 4.90 to 1.01 after 25,000 updates, a 4.9-fold reduction that reaches roughly ninefold at earlier checkpoints, and moves the value estimate from 115.57 to -158.07 (Appendix D.3). Because a shared minimum also changes the target values, the comparison cannot separate assignment from the change in targets. REDQ fits every critic to a target built from a random subset of target critics, and its published ablation with a single sampled critic supplies one member's complete target vector to every critic (Chen et al., 2021). Such broadcasts change which member's vector each critic fits, but they also duplicate that vector, so they change the collection of target vectors as well as their assignment (Appendix D.2). Whole-row reassignment changes only assignment, keeping every target vector intact, so the reductions in spread and coverage in Section 4.1 show that assignment alone can change the reported uncertainty. The confidence intervals, exploration, and pessimism that published methods derive from ensemble

[1] Appendix C.2 confirms the finite-M moment predictions of Section 3.4 for fixed-feature ridge fits with member-specific continuation operators.

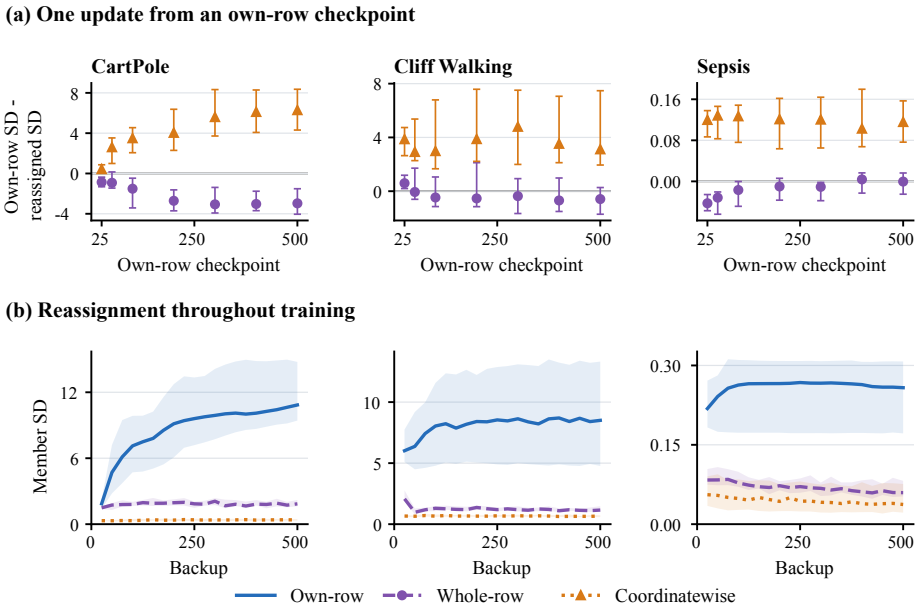

Figure 2: Spread under reassignment on three environments. (a) Paired change in spread, own-row SD minus reassigned SD, after one reassigned update from each saved own-row checkpoint; negative values mean reassignment increased spread. (b) Member SD when each rule is applied at every update. Points and lines show medians over datasets, and bars and bands show dataset IQRs.

uncertainty (Hao et al., 2021; Osband et al., 2016; An et al., 2021; Ghasemipour et al., 2022) can therefore change with target rules as small as one configuration flag.

Limitations. Our results show that reassignment lowers coverage, not that own-row intervals are calibrated. Own-row coverage is measured against references truncated at a fixed horizon, and the intervals ignore sampling variation in the logged initial states, so near-nominal own-row coverage need not hold for the full policy value (Appendix B.3). The theory does not establish calibration either, because the large-ensemble limits hold the dataset and the number of updates fixed. The theory also does not predict the size of neural effects and assumes that members update separately, which excludes shared-parameter designs such as HyperAgent (Li et al., 2024).

6 Conclusion

Recursive fitted evaluation propagates bootstrap uncertainty through regression targets, but the targets alone leave unspecified which member fits each target vector and which predictions share a vector. Both choices change the propagated uncertainty even when the distribution of targets at every input is preserved, and, as we prove, the change can persist as the ensemble grows. In Bootstrap-FQE across six environments, reassigning intact target vectors lowered coverage of nominal 90% intervals to 4%–30%, against 88%–100% under own-row assignment, even when point estimates improved.

Published methods that use propagated uncertainty for confidence intervals, exploration, and pessimism construct targets in different ways. Replacing per-critic targets with a shared minimum, for example, changed the reported standard deviation of a published ensemble method for confidence estimation five- to ninefold on a standard benchmark. Decisions based on propagated uncertainty therefore depend on how targets are constructed. Any procedure that propagates uncertainty through regression targets must state which member fits each target vector and which predictions share a vector, and specifying the propagated uncertainty must precede, rather than replace, an argument for statistical validity.

AI USE STATEMENT

We used generative AI tools to assist in the writing of proofs. In each case, we specified the proof strategy and key steps in an outline, and generative AI tools completed the derivations between these steps. Additionally, we used generative AI tools for sentence-level editing to improve readability without altering technical content. We also used generative AI tools to locate the code repositories of the published implementations analyzed in Appendix D.2.

We have reviewed all AI-assisted work.

REPRODUCIBILITY STATEMENT

Appendix B gives the experimental details. Appendices C–D provide the numerical comparisons and the repositories of the published implementations we inspected.

REFERENCES

Gaon An, Seungyong Moon, Jang-Hyun Kim, and Hyun Oh Song. Uncertainty-based offline reinforcement learning with diversified Q-ensemble. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://papers.neurips.cc/paper_files/paper/2021/hash/3d3d286a8d153a4a58156d0e02d8570c-Abstract.html.

Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996. doi: 10.1007/BF00058655.

Jacob Buckman, Danijar Hafner, George Tucker, Eugene Brevdo, and Honglak Lee. Sample-efficient reinforcement learning with stochastic ensemble value expansion. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/f02208a057804ee16ac72ff4d3cec53b-Abstract.html.

Xinyue Chen, Che Wang, Zijian Zhou, and Keith Ross. Randomized ensembled double Q-learning: Learning fast without a model. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2101.05982.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6(18):503–556, 2005. URL https://www.jmlr.org/papers/v6/ernst05a.html.

Vladimir Feinberg, Alvin Wan, Ion Stoica, Michael I. Jordan, Joseph E. Gonzalez, and Sergey Levine. Model-based value estimation for efficient model-free reinforcement learning, 2018. URL https://arxiv.org/abs/1803.00101.

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

Seyed Kamyar Seyed Ghasemipour, Shixiang Shane Gu, and Ofir Nachum. Why so pessimistic? estimating uncertainties for offline rl through ensembles, and why their independence matters. In *Advances in Neural Information Processing Systems*, volume 35, pp. 18267–18281, 2022.

Botao Hao, Xiang Ji, Yaqi Duan, Hao Lu, Csaba Szepesvári, and Mengdi Wang. Bootstrapping fitted q-evaluation for off-policy inference. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4074–4084. PMLR, 2021.

Jihwan Jeong, Xiaoyu Wang, Michael Gimelfarb, Hyunwoo Kim, Baher Abdulhai, and Scott Sanner. Conservative bayesian model-based value expansion for offline policy optimization. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=dNqxZgyjcYA.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6402–6413, 2017.

Jason Li, Muhammad Hamza Yousuf, Juhyun Seo, Jens Lehmann, and Sahar Vahdati. Test-time confidence quantification using Q-ensembles for offline RL. *IEEE Access*, 14:61867–61880, 2026. doi: 10.1109/ACCESS.2026.3685907.

Yingru Li, Jiawei Xu, Lei Han, and Zhi-Quan Luo. Q-star meets scalable posterior sampling: Bridging theory and practice via HyperAgent. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024.

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

Michael Oberst and David Sontag. Counterfactual off-policy evaluation with Gumbel-Max structural causal models. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4881–4890, 2019. URL https://proceedings.mlr.press/v97/oberst19a.html.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped DQN. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://papers.neurips.cc/paper/6501-deep-exploration-via-bootstrapped-dqn.

Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Xiuyuan Lu, and Benjamin Van Roy. Evaluating high-order predictive distributions in deep learning. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 1552–1560. PMLR, 2022.

Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Morteza Ibrahimi, Xiuyuan Lu, and Benjamin Van Roy. Approximate thompson sampling via epistemic neural networks. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 1586–1595. PMLR, 2023.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

Zheng Wen, Ian Osband, Chao Qin, Xiuyuan Lu, Morteza Ibrahimi, Vikranth Dwaracherla, Mohammad Asghari, and Benjamin Van Roy. From predictions to decisions: The importance of joint predictive distributions. *arXiv preprint arXiv:2107.09224*, 2022.

Yue Wu, Shuangfei Zhai, Nitish Srivastava, Joshua M. Susskind, Jian Zhang, Ruslan Salakhutdinov, and Hanlin Goh. Uncertainty weighted actor-critic for offline reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11319–11328. PMLR, 2021.

Ruiqi Zhang, Xuezhou Zhang, Chengzhuo Ni, and Mengdi Wang. Off-policy fitted q-evaluation with differentiable function approximators: Z-estimation and inference theory. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 26713–26749. PMLR, 2022.

A Coupling, Assignment Laws, and Proofs

Coordinate coupling determines the law supplied to a fit; assignment determines its dependence on the recipient's retained state. The results below establish the resulting large-ensemble limits, derive exact finite-ensemble affine moments, and justify the supporting experimental diagnostics.

A.1 Coordinate coupling changes propagated spread

A two-target average suffices to show why coordinate marginals do not determine a propagated law. Let Z be uniform on $\{-1, 1\}$. The target rows (Z, Z) and $(Z, -Z)$ have identical coordinate marginals, but averaging the coordinates gives Z and zero, respectively. Thus even a fixed affine fit can produce different output laws from the same target marginals. The covariance calculation below quantifies this effect and gives the prediction used by the fixed-smoother experiment.

Proposition A.1 (Covariance and effective support)**.** *Let $X \in \mathbb{R}^d$ have finite covariance Σ, and let $X^{\otimes}$ have independent coordinates with the same marginals. For every $a \in \mathbb{R}^d$,*

$$\mathrm{Var}(a^\top X) - \mathrm{Var}(a^\top X^{\otimes}) = 2 \sum_{i<j} a_i a_j \Sigma_{ij}.$$

The difference is nonnegative when every summand is nonnegative, but has no universal sign. If each coordinate has variance $\varsigma^2 > 0$ and each distinct pair has covariance $\rho\varsigma^2$, then weights satisfying $\mathbf{1}^\top w = 1$ give

$$\mathrm{Var}(w^\top X) = \varsigma^2[\rho + (1-\rho)\|w\|_2^2].$$

For uniform averaging over K coordinates, the output SD is $\varsigma\sqrt{\rho + (1-\rho)/K}$. Perfect alignment ($\rho = 1$) therefore has $\sqrt{K}$ times the SD of independent coordinates ($\rho = 0$). For nonnegative normalized weights, the aligned-to-independent variance ratio is the effective support $d_{\mathrm{eff}}(w) = 1/\|w\|_2^2$, equal to K under uniform averaging.

Proof. The variance is $a^\top \Sigma a$, while independent reconstruction retains only the diagonal of Σ. In the equicorrelated case, substitute $\Sigma = \varsigma^2[(1-\rho)I + \rho\mathbf{1}\mathbf{1}^\top]$ and use $\mathbf{1}^\top w = 1$. □

A.2 Assignment limits with retained fitting history

Use the dynamics and assumptions of Theorem 3.1, conditional on any shared exogenous schedule. Write $Z_t^{(b)} = (R_t^{(b)}, X_t^{(b)})$, $\hat{Q}_{M,t} = M^{-1}\sum_b \delta_{Z_t^{(b)}}$, and let μ_t^c obey Equation 2. The initial fitted state and persistent resample may be dependent within a member. Across members their pairs are i.i.d., and the private update noises and assignment permutations have the independence stated in the main theorem. All dimensions, maps, and primitive laws remain fixed as M grows. Let $d_{\mathrm{TV}}(P, Q) = \sup_B |P(B) - Q(B)|$ and let d_{BL} be the bounded-Lipschitz metric for a compatible metric on the state space.

Theorem A.2 (Finite-ensemble convergence)**.** *For whole-row assignment set $r_t = 1$, and for coordinatewise assignment set $r_t = d_t$. Define $\ell_0 = 1$ and $\ell_{t+1} = (1 + r_t)\ell_t$. For $k \le M$, put $n = k\ell_t$ and*

$$\varepsilon_{M,t,k} = \begin{cases} \min\left\{1, \dfrac{n(n-1) - k(k-1)}{2(M-n+1)}\right\}, & M \ge n, \\ 1, & M < n. \end{cases} \tag{4}$$

Then

$$d_{\mathrm{TV}}\Big(\mathrm{Law}(Z_t^{(1)}, \dots, Z_t^{(k)}), (\mu_t^c)^{\otimes k}\Big) \le \varepsilon_{M,t,k}. \tag{5}$$

Own-row members are exactly independent, with $\varepsilon_{M,t,k} = 0$. For $M \ge 2$ and bounded Borel f,

$$\mathbb{E}|\hat{Q}_{M,t} f - \mu_t^c f|^2 \le 4\|f\|_\infty^2 \big(M^{-1} + \varepsilon_{M,t,2}\big). \tag{6}$$

At fixed depth, the empirical joint law and the empirical law of any finite real-valued Borel readout converge weakly in probability to their population laws. If g is finite real-valued and Borel with $\mu_t^c|g| < \infty$, then $\hat{Q}_{M,t} g \to_p \mu_t^c g$. In particular, a finite limiting second moment of the readout gives convergence of its empirical mean and variance, as asserted in Theorem 3.1.

Proof. Ancestry and collisions. A member's state at time $t+1$ depends on its own state at time t and on r_t donor states. Unfolding the dependencies of a root at time t to time zero gives at most ℓ_t initial leaves. Keep repeated appearances along the recipient's own history in one label block, carrying that member's initial pair and time-indexed private noises. Each donor query introduces a new block unless its label has already appeared. A forest of k tagged roots therefore requires at most $n = k\ell_t$ blocks or attempted introductions.

A collision occurs when a donor query reaches an already discovered block. Before a collision, every queried permutation argument is unexposed. With j labels discovered and a images of the current permutation exposed, $a \le j$, and the next image is uniform among $M - a$ labels. Its collision probability is at most $j/(M-j)$. Summing over $j = k, \dots, n-1$ and using $M - j \ge M - n + 1$ gives Equation 4; bounds of one are immediate. Repeated uses of the recipient's own history are retained, not counted as collisions.

The explored graph and collision event depend only on permutations, independently of primitive values. On every collision-free graph, distinct blocks carry independent initial pairs and noise sequences. Each recipient retains the joint law of its state and resample, while whole-row donors are independent of that history and coordinatewise donors are also independent across coordinates. Applying the update maps therefore gives exactly Equation 2. Distinct roots have disjoint histories, so their conditional joint law is $(\mu_t^c)^{\otimes k}$ on every such graph. Consequently the tagged law is $(1 - p_{\mathrm{bad}})(\mu_t^c)^{\otimes k} + p_{\mathrm{bad}}\rho$, where ρ is a probability law and $p_{\mathrm{bad}} \le \varepsilon_{M,t,k}$. This proves Equation 5.

Empirical laws. The member array is exchangeable. Expanding the empirical squared error into diagonal and two-member terms, then applying the preceding mixture, gives Equation 6. A countable convergence-determining family on the Polish state space and subsequence extraction give weak convergence in probability. Apply the same argument to bounded tests composed with a Borel readout to obtain its empirical-law limit. No continuity of the update or readout maps is required.

Unbounded moments. Let I_b indicate a collision in member b's history and put $B_M = \sum_b I_b$. For $M \ge 2\ell_t$, the one-member collision bound gives $\mathbb{E}B_M \le C_t = \ell_t(\ell_t - 1)$. At fixed depth there are finitely many possible full ancestry equality patterns. Each pattern has an output law ρ_j independent of M, since the pattern fixes all primitive reuses and its probability does not depend on primitive values. For $V_M = \max_{b:I_b=1} |g(Z_t^{(b)})|$, with empty maximum zero, a union bound gives

$$\Pr(V_M > L) \le C_t \max_j \rho_j(|g| > L), \qquad \frac{1}{M}\sum_b I_b |g(Z_t^{(b)})| \le \frac{B_M V_M}{M} = O_p(M^{-1}). \tag{7}$$

Indeed, $B_M = O_p(1)$ and $V_M = O_p(1)$ because g is finite under each of finitely many pattern laws. No integrability assumption on collision histories is needed.

A collision-free member has conditional law μ_t^c, so its expected empirical tail contribution is at most $\mu_t^c(|g|\mathbf{1}_{\{|g|>L\}})$. Apply Equation 6 to a bounded truncation of g, Markov's inequality to this tail, and Equation 7 to collision histories. Taking $M \to \infty$ and then $L \to \infty$ proves the moment claim. Apply the claim to the readout and its square for empirical variance. Own-row uses independent primitive families and the integrable i.i.d. law of large numbers. □

Remark A.3 (Scalar assignment separation). For the additive-intercept recursion in Equation 3, the variance gap is $2\tau^2 \sum_{0 \le r < s < T} \gamma^{r+s} > 0$ for $T \ge 2$ and $0 < \gamma < 1$. The own-row and whole-row variances tend to $\tau^2/(1-\gamma)^2$ and $\tau^2/(1-\gamma^2)$ as depth grows. Thus their limiting SD ratio is $\sqrt{(1+\gamma)/(1-\gamma)}$, the prediction used in the exact-tabular experiment.

Remark A.4 (Persistent target separation). The reverse triangle inequality gives $d_{\mathrm{BL}}(\hat{\nu}_{M,T}^c, \nu_T^{c'}) \to_p d_{\mathrm{BL}}(\nu_T^c, \nu_T^{c'})$, so distinct population laws remain separated as ensemble size grows. These limits fix the dataset and depth; they establish neither bootstrap coverage as data grow nor a useful finite-ensemble error bound at the reported neural depths.

Remark A.5 (Bootstrap function replicates). Retaining each bootstrap resample with its fitted function through target construction and refitting implements own-row assignment. Warm-started fitting also retains learned parameters and optimizer state. This specifies the propagated law without establishing bootstrap validity.

A.3 EXACT MOMENTS CONDITIONAL ON A FINITE OPERATOR BANK

Fix $M \geq 2$, matrices $F_b \in \mathbb{R}^{q\times d}$ and $G \in \mathbb{R}^{d\times q}$, intercepts $c_b \in \mathbb{R}^q$, and $A_b = F_bG$. Whole-row updates are $\theta_b^+ = c_b + A_b\theta_{\sigma(b)}$. Coordinatewise updates are $\theta_b^+ = c_b + F_by_b$, where $(y_b)_j = (G\theta_{\sigma_j(b)})_j$. The permutations are uniform, independent across coordinates and of the incoming array, and fresh at each update. Expectations condition on this fixed bank and average over assignments and any incoming randomness. The incoming parameters may be dependent but have finite second moments.

Proposition A.6 (Conditional finite-ensemble moment closure)**.** *For $\bar\theta = M^{-1}\sum_b \theta_b$, define*

$$m = \mathbb{E}\bar\theta, \qquad S = \mathrm{Cov}(\bar\theta), \qquad C = \mathbb{E}\left[\frac{1}{M}\sum_b(\theta_b-\bar\theta)(\theta_b-\bar\theta)^\top\right].$$

Bars denote arithmetic averages over members. Put $\Delta A_b = A_b - \bar A$, $\Delta c_b = c_b - \bar c$, and

$$H(m) = \frac{1}{M}\sum_b(\Delta c_b + \Delta A_b m)(\Delta c_b + \Delta A_b m)^\top.$$

For whole-row assignment, set

$$J = \frac{1}{M}\sum_b A_bCA_b^\top, \qquad J_0 = \bar A C\bar A^\top,$$
$$V = \frac{J-J_0}{M-1} = \frac{1}{M(M-1)}\sum_b \Delta A_bC\Delta A_b^\top.$$

The exact moment recursion is

$$m^+ = \bar c + \bar A m,$$
$$S^+ = \bar A S\bar A^\top + V,$$
$$C^+ = H(m) + \frac{1}{M}\sum_b \Delta A_bS\Delta A_b^\top + \frac{M-2}{M-1}J + \frac{1}{M-1}J_0.$$

For coordinatewise assignment, use $D_G = \mathrm{diag}(GCG^\top)$, $J = M^{-1}\sum_b F_bD_GF_b^\top$, $J_0 = \bar F D_G\bar F^\top$, and $V = (J-J_0)/(M-1)$. The sum-of-squares expression for V then uses $F_b - \bar F$ and D_G in place of ΔA_b and C.

For readout $a^\top\theta_b$, the ensemble mean has expectation $a^\top m$ and variance $a^\top Sa$; the expected sample member variance is $Ma^\top Ca/(M-1)$. With deterministic initialization, these expectations average over assignment schedules and are not medians of realized standard deviations.

Proof. Condition on the incoming array and write $C_{\mathrm{emp}} = M^{-1}\sum_b(\theta_b-\bar\theta)(\theta_b-\bar\theta)^\top$. Each whole-row donor has mean $\bar\theta$ and covariance C_{emp}. Distinct donors have cross covariance $-C_{\mathrm{emp}}/(M-1)$, since the sum of centered outer products over ordered distinct pairs is $-MC_{\mathrm{emp}}$. Independent column permutations replace $GC_{\mathrm{emp}}G^\top$ by its diagonal. The fitted ensemble mean therefore has conditional expectation $\bar c + \bar A\bar\theta$ and conditional covariance $V(C_{\mathrm{emp}})$, viewing J, J_0, V as linear maps of the covariance argument. Total expectation and covariance give m^+ and S^+; centering the operators gives the sum-of-squares expression for V.

The expected average outer product around the conditional ensemble mean is $H(\bar\theta) + J(C_{\mathrm{emp}})$. Centering at the realized ensemble mean subtracts $V(C_{\mathrm{emp}})$. Thus C^+ follows from

$$\mathbb{E}H(\bar\theta) = H(m) + \frac{1}{M}\sum_b \Delta A_bS\Delta A_b^\top, \qquad J - V = \frac{M-2}{M-1}J + \frac{1}{M-1}J_0.$$

Every term in C^+ is positive semidefinite. The readout formulas follow by linearity and the sample-variance normalization. Fresh permutations permit exact iteration, including for fixed time-dependent banks. □

For ridge FQE, let Φ be the training-feature matrix, r the reward vector, $\lambda > 0$ the penalty, and $W_b = \mathrm{diag}(w_{b1},\dots,w_{bn})/n$ the bootstrap weight matrix. Then $F_b = (\Phi^\top W_b\Phi + \lambda I)^{-1}\Phi^\top W_b$ and $c_b = F_br$. Row i of G is the discounted, nonterminal evaluation-policy expectation of successor features. Reward-fit initialization gives $\theta_{b,0} = c_b$, $m_0 = \bar c$, $S_0 = 0$, and $C_0 = M^{-1}\sum_b \Delta c_b\Delta c_b^\top$.

A.4 ACCUMULATION AND MEAN-SHIFT DIAGNOSTICS

The exact moments determine total spread. To isolate accumulation across backups, fix the bank and deterministic initialization, with $m_0 = M^{-1}\sum_b \theta_{b,0}$ and $m_{t+1} = \bar{c} + \bar{A}m_t$. Propagate deviations through the common operator $\bar{A}$ while retaining each recipient's centered forcing,

$$f_{b,0} = \theta_{b,0} - m_0, \qquad f_{b,t+1} = \Delta c_b + \Delta A_b m_t,$$
$$z_{b,0} = f_{b,0}, \qquad z_{b,t+1} = f_{b,t+1} + \bar{A} z_{\sigma_t(b),t}.$$

Use identity assignment for own-row and fresh uniform permutations for whole-row. Both surrogate ensembles have zero mean. For readout a, let $h_{b,s} = a^\top \bar{A}^{T-s} f_{b,s}$ and define

$$V_T^{\mathrm{coh}} = \frac{1}{M-1}\sum_b \left(\sum_{s=0}^{T} h_{b,s}\right)^2, \qquad V_T^{\mathrm{inc}} = \frac{1}{M-1}\sum_b \sum_{s=0}^{T} h_{b,s}^2.$$

Their difference, $2(M-1)^{-1}\sum_b \sum_{s<r} h_{b,s}h_{b,r}$, records covariance between contributions from different backups. Positive cross-backup covariance increases V_T^{coh}, regardless of the individual contributions' signs.

Write Var_M and Cov_M for sample moments with denominator $M-1$. Couple each full recursion and surrogate through the same assignments, set $e_{b,t} = \theta_{b,t} - m_t - z_{b,t}$, and define $V_T^{\mathrm{rem}} = \mathbb{E}\,\mathrm{Var}_M(a^\top e_{b,T}^{\mathrm{whole}})$.

Proposition A.7 (Exact whole-row variance decomposition)**.** *The surrogate has own-row variance V_T^{coh} and expected whole-row variance V_T^{inc}. Its whole-row remainder has zero expected sample cross covariance with the surrogate, so*

$$W_T := \mathbb{E}\,\mathrm{Var}_M(a^\top \theta_{b,T}^{\mathrm{whole}}) = V_T^{\mathrm{inc}} + V_T^{\mathrm{rem}}.$$

Proof. Identity assignment gives $a^\top z_{b,T}^{\mathrm{own}} = \sum_s h_{b,s}$. Under whole-row assignment, permutation averaging eliminates each cross term between the deterministic centered forcing and assigned surrogate rows, giving expected variance V_T^{inc}. Subtracting the surrogate from the full recursion gives

$$e_{b,t+1} = A_b e_{\sigma_t(b),t} + \Delta A_b z_{\sigma_t(b),t}, \qquad e_{b,0} = 0.$$

Let $\mathcal{F}_t$ contain the bank and preceding assignments, let $\bar{e}_t = M^{-1}\sum_b e_{b,t}$, and set $K_t = \mathrm{Cov}_M(e_t, z_t)$. Since $\bar{z}_t = 0$ and $\sum_b \Delta A_b = 0$, permutation averaging gives $\mathbb{E}[\bar{e}_{t+1} \mid \mathcal{F}_t] = \bar{A}\bar{e}_t$ and

$$\mathbb{E}[K_{t+1} \mid \mathcal{F}_t] = \bar{A}K_t\bar{A}^\top + \frac{1}{M-1}\sum_b A_b \bar{e}_t f_{b,t+1}^\top.$$

The forcing is deterministic and $e_{b,0} = 0$, so induction gives $\mathbb{E}\bar{e}_t = 0$ and $\mathbb{E}K_t = 0$. Expanding the full variance proves the decomposition. □

Proposition A.6 determines W_T, hence $V_T^{\mathrm{rem}} = W_T - V_T^{\mathrm{inc}}$. The coupled recursions for e and z also determine joint diagnostics.

Projection onto the own-row surrogate. For $V_T^{\mathrm{coh}} > 0$, set $u_b = \sum_s h_{b,s}$, $v_b = a^\top \theta_{b,T}^{\mathrm{own}}$, and $\beta_T = \mathrm{Cov}_M(u,v)/V_T^{\mathrm{coh}}$. The signed coefficient measures linear alignment with the surrogate. When $\mathrm{Var}_M(v) > 0$, the aligned share is the squared sample correlation $R_T^2 = \mathrm{Corr}_M(u,v)^2 = \beta_T^2 V_T^{\mathrm{coh}}/\mathrm{Var}_M(v)$. The least-squares residual $v_b - \bar{v} - \beta_T u_b$ is orthogonal to u; the raw own-row remainder need not be. These diagnostics measure linear agreement across computed responses, not predictive fidelity or a unique causal mechanism.

Accumulated mean shifts. Let $m_t^{\mathrm{own}} = M^{-1}\sum_b \theta_{b,t}^{\mathrm{own}}$ and $g_t = M^{-1}\sum_b \Delta A_b(\theta_{b,t}^{\mathrm{own}} - m_t^{\mathrm{own}})$. Averaging the own-row update gives $m_{t+1}^{\mathrm{own}} = \bar{c} + \bar{A}m_t^{\mathrm{own}} + g_t$. Subtracting the expected whole-row recursion and unrolling from the same initialization gives

$$a^\top(m_T^{\mathrm{own}} - m_T) = \sum_{t=0}^{T-1} a^\top \bar{A}^{T-1-t} g_t.$$

Each operator–row covariance g_t contributes to the endpoint shift through the subsequent mean operators.

To separate scalar gain from changes in value-function shape, take $a \neq 0$, write $v_{b,t} = a^\top \theta^{\mathrm{own}}_{b,t}$, and choose a common profile p_t with $a^\top p_t = 1$. Define

$$r_{b,t} = (\theta^{\mathrm{own}}_{b,t} - m^{\mathrm{own}}_t) - (v_{b,t} - \bar{v}_t)p_t,$$
$$q_{t,T} = (\bar{A}^{T-1-t})^\top a, \qquad k_{b,t,T} = q_{t,T}^\top A_b p_t.$$

The residual has $a^\top r_{b,t} = 0$ and sums to zero across members; $k_{b,t,T}$ maps a unit of current value along p_t to the final readout. The choice $p_t = a/(a^\top a)$ always applies. The reported diagnostics instead use $p_t = m^{\mathrm{own}}_t/\bar{v}_t$, with nonzero means, giving $r_{b,t} = \theta^{\mathrm{own}}_{b,t} - v_{b,t}p_t$. Substitution into g_t yields

$$q_{t,T}^\top g_t = \frac{M-1}{M}\,\mathrm{Cov}_M(k_{b,t,T}, v_{b,t}) + \frac{1}{M}\sum_b q_{t,T}^\top \Delta A_b r_{b,t}. \tag{8}$$

The gain–value covariance and shape contribution depend on the profile, but their sum does not. Positive accumulated gain–value covariance lowers the expected whole-row center unless an opposing shape contribution dominates.

A.5 Fresh resampling and retained fitting history

The fresh-resampling and shared-target comparisons use a special case of the retained-state model. Let $\xi_b \overset{\mathrm{iid}}{\sim} \Xi$, independently of $X_0^{(b)} \overset{\mathrm{iid}}{\sim} P_0$, and set $X_{t+1}^{(b)} = \Phi_t(Y_t^{(b)}, \xi_b, \eta_t^{(b)})$, with jointly Borel maps and the private-noise assumptions of Theorem 3.1. Write $K_t^\xi(y,\cdot) = \mathrm{Law}(\Phi_t(y,\xi,\eta_t))$. The update retains ξ_b but uses no other recipient history. Setting $S_t = X_t$ and letting F_t ignore its first argument embeds this model in Theorem A.2, so no separate convergence argument is needed. In this specialization, μ_t^c and $\hat{Q}_{M,t}$ denote the population and empirical laws of the projected pair (ξ, X_t). In particular, the whole-row joint law obeys

$$\mu^{\mathrm{whole}}_{t+1}(d\xi, dz) = \Xi(d\xi)\int K_t^\xi(y, dz)\, P_t^{\mathrm{whole}}(dy), \qquad P_t^c = (\mu_t^c)^X. \tag{9}$$

Coordinatewise assignment replaces P_t^c by $\mathcal{R}_{d_t}(P_t^c) = \bigotimes_j (P_t^c)_j$.

Corollary A.8 (Fresh state and retained history)**.** *In this model, independently redrawing $\xi_b^{(t)} \sim \Xi$ before each own-row update leaves members independent, with row recursion*

$$P^{\mathrm{fresh}}_{t+1} = P^{\mathrm{fresh}}_t \bar{K}_t, \qquad \bar{K}_t(y,\cdot) = \int K_t^\xi(y,\cdot)\,\Xi(d\xi). \tag{10}$$

From the same initial row law, this equals the persistent whole-row population row law.

With retained fitting history, let $L_t(s,\xi,y;\cdot) = \mathrm{Law}(F_t(s,y,\xi,\eta_t))$ and let Ξ be the initial persistent-state marginal. Redrawing only ξ independently at each update, while retaining S_t and its own row $h_t(S_t)$, instead gives

$$\rho^{\mathrm{fresh}}_{t+1}(ds') = \int\!\!\int L_t(s,\xi,h_t(s);ds')\,\Xi(d\xi)\,\rho_t^{\mathrm{fresh}}(ds), \tag{11}$$

where ρ_t^{fresh} is the state law of the independent members. If L_t is independent of s, the row laws again agree with persistent whole-row from equal initial row laws. With recipient memory, neither equality nor a variance ordering holds in general.

Proof. Integrating the independent fresh state gives Equation 10, which is the row marginal of Equation 9, and similarly gives Equation 11. If L_t ignores s, both row recursions use the same averaged kernel. Otherwise, take $h(s) = s$, a centered initial state with variance $\varsigma^2 > 0$, and $F(s,y,\xi) = s + ay$. Fresh own-row has variance $(1+a)^2\varsigma^2$ after one update, whereas persistent whole-row has variance $(1+a^2)\varsigma^2$. Their difference $2a\varsigma^2$ can have either sign. □

For affine refits, the averaged kernel also links the output mean and covariance.

Corollary A.9 (Affine population moments)**.** *In the model without additional recipient history, let $\Phi_t(y, \xi, \eta_t) = b_t(\xi) + A_t(\xi)y + \eta_t$. Suppose A_t, b_t, η_t, and the incoming row have finite second moments, with η_t independent of the state and assigned row. Under whole-row assignment, writing $m_t = \mathbb{E}X_t$, $\Sigma_t = \mathrm{Cov}(X_t)$, and suppressing ξ in the operators gives*

$$\begin{aligned} m_{t+1} &= \mathbb{E}b_t + \mathbb{E}[A_t]m_t + \mathbb{E}\eta_t, \\ \Sigma_{t+1} &= \mathbb{E}[A_t\Sigma_t A_t^\top] + \mathrm{Cov}(b_t + A_t m_t) + \mathrm{Cov}(\eta_t). \end{aligned} \tag{12}$$

Coordinatewise assignment replaces Σ_t inside the expectation by $\mathrm{diag}(\Sigma_t)$. Thus the two mean trajectories agree from equal initial means, while spread also depends on the mean through $\mathrm{Cov}(b_t + A_t m_t)$.

Proof. The assigned whole row is independent of (A_t, b_t). Conditional on ξ, its fitted mean and covariance are $b_t + A_t m_t + \mathbb{E}\eta_t$ and $A_t\Sigma_t A_t^\top + \mathrm{Cov}(\eta_t)$. Total expectation and total covariance give Equation 12. Independent coordinates preserve the input mean and diagonalize its covariance. □

A.6 Shared targets in the implementation comparisons

Broadcasting one target $S_{M,t}$ to every fitter generally changes the empirical row law. In the fixed-state model of Section A.5, define $H_t(y)(d\xi, dz) = \Xi(d\xi)K_t^\xi(y, dz)$ and $\bar{K}_t(y, \cdot) = H_t(y)^X$. A concentrating target gives a deterministic limiting ensemble law; a target selected from finitely many members can leave a random limit.

Proposition A.10 (Shared-target limits)**.** *Assume the fixed-state model. Choose the common target before the current update noises, with its joint law with the ensemble invariant under member relabeling.*

(a) *If $S_{M,t} \to_p s_t$ for deterministic s_t and K_t^ξ is jointly measurable and Feller in its row argument for every ξ, then $\hat{Q}_{M,t+1} \to H_t(s_t)$ weakly in probability. Any fixed number of tagged pairs converge weakly to independent copies of $H_t(s_t)$.*

For a shared mean, concentration follows inductively if $\mathbb{E}\|X_0\|^p < \infty$ for some $p > 1$ and

$$\int \|z\|^p K_t^\xi(y, dz) \le c_t(1 + \|y\|^p + w_t(\xi)), \qquad c_t < \infty, \quad w_t \ge 0, \quad \Xi w_t < \infty.$$

The recursion is $s_t = \int x\, P_t^{\mathrm{shared}}(dx)$ and $P_{t+1}^{\mathrm{shared}} = H_t(s_t)^X$, with $P_0^{\mathrm{shared}} = P_0$.

(b) *Fix $\kappa \ge 1$. Let $S_{M,t,j} = s_{t,j}((X_{t,j}^{(b)})_{b \in J_{t,j}})$ for symmetric Borel functions $s_{t,j} \colon \mathbb{R}^\kappa \to \mathbb{R}$ and uniform κ-subsets $J_{t,j}$, independent across coordinates and times and of all primitives. Borel kernels suffice in this case. Define*

$$\begin{aligned} \tau_t(P) &= \bigotimes_{j=1}^{d_t} (s_{t,j})_\# P_j^{\otimes\kappa}, \\ s_0 &\sim \tau_0(P_0), \qquad \mathrm{Law}(s_{t+1} \mid s_{0:t}) = \tau_{t+1}(\bar{K}_t(s_t, \cdot)). \end{aligned} \tag{13}$$

Then $d_{\mathrm{BL}}(\hat{Q}_{M,t+1}, H_t(S_{M,t})) \to_p 0$ and

$$(S_{M,0:t}, \hat{Q}_{M,t+1}) \Rightarrow (s_{0:t}, H_t(s_t)).$$

Conditional on the limiting target history, finitely many tagged pairs are independent with law $H_t(s_t)$. Their unconditional limit is the mixture $\mathbb{E}[H_t(s_t)^{\otimes k}]$, which need not be a product law.

If one subset supplies all coordinates, replace $\tau_t(P)$ by the law of $(s_{t,j}(V_{1j}, \ldots, V_{\kappa j}))_j$ for i.i.d. rows $V_1, \ldots, V_\kappa \sim P$. With $\kappa = 1$ and identity statistics, a common subset broadcasts one intact row; independent coordinate subsets broadcast a row with independent coordinates.

Proof. For a bounded pair test f, set $g_f(y, \xi) = \int f(\xi, z)K_t^\xi(y, dz)$. Conditional on the incoming ensemble, past randomness and common target, the current noises are independent. Thus $\hat{Q}_{M,t+1}f$ has conditional mean $M^{-1}\sum_b g_f(S_{M,t}, \xi_b)$ and variance at most $\|f\|_\infty^2/M$.

For part (a), take bounded continuous f. The Feller assumption makes $g_f(\cdot, \xi)$ continuous. Set $\omega_\rho(\xi) = \sup_{\|y-s_t\| \le \rho} |g_f(y, \xi) - g_f(s_t, \xi)|$. A countable dense subset makes this supremum measurable, and dominated convergence gives $\Xi\omega_\rho \to 0$ as $\rho \downarrow 0$. On $\{\|S_{M,t} - s_t\| \le \rho\}$, replacing $S_{M,t}$ by s_t changes the conditional mean by at most $M^{-1} \sum_b \omega_\rho(\xi_b)$. The i.i.d. state law of large numbers applies to this bound and to $M^{-1} \sum_b g_f(s_t, \xi_b)$. Taking $M \to \infty$ and then $\rho \downarrow 0$ proves convergence of continuous test integrals, and a countable determining family gives weak convergence in probability. Exchangeability compares products of tagged tests with products of empirical averages up to $O(k^2/M)$ from repeated indices, giving the independent tagged limit. For shared means, Jensen's inequality and the drift bound give uniform row pth moments by induction. Truncation controls first-moment tails uniformly, so weak convergence gives concentration of the next shared mean.

For part (b), condition on the subsets through time t. Their union $\mathcal{I}_t$ has size at most $L_t = \kappa \sum_{s=0}^{t} d_s$, and the target history depends only on primitives with labels in $\mathcal{I}_t$. States outside that union remain i.i.d. Ξ and independent of the history. Their centered empirical g_f-sum has conditional variance at most $\|f\|_\infty^2/M$; labels inside contribute at most $2L_t\|f\|_\infty/M$. Including current-noise variation gives

$$\mathbb{E}|\hat{Q}_{M,t+1}f - H_t(S_{M,t})f| \le \frac{2\|f\|_\infty}{\sqrt{M}} + \frac{2L_t\|f\|_\infty}{M}.$$

The subset groups are disjoint and avoid any fixed k tags except with probability $O((L_t + k)^2/M)$. On a disjoint configuration, each selected group uses primitives independent of prior target history. Induction gives Equation 13, and the final tags use independent primitives outside the union. This proves total-variation convergence of target history and tagged pairs to the stated mixture. A common subset uses one group per update, replacing L_t by $\kappa(t + 1)$.

Adjoining the Borel map H_t preserves total-variation convergence of the target history. To establish empirical-measure tracking, choose a compact set carrying all but ϵ of $\mathbb{E}H_t(s_t)$. Total-variation convergence controls the expected mass of $H_t(S_{M,t})$ outside that set; the displayed test bound controls the empirical mass. On the compact set, the bounded-Lipschitz unit ball has a finite uniform ϵ-net. Applying the test bound to this net and using Markov's inequality gives d_{BL} tracking as $M \to \infty$ and then $\epsilon \downarrow 0$. Joint convergence with the target history follows. □

B EXPERIMENTAL PROTOCOL

B.1 PAIRED NEURAL COMPARISONS AND FITTING

Our comparison uses 25 independently generated datasets, which we call parents, in each of CartPole, Cliff Walking, and Sepsis, and ten episode subsamples of each D4RL medium-v2 dataset for Hopper, Walker2d, and HalfCheetah. Each parent contains $N = 10$ episodes and is fit by $M = 50$ separately parameterized networks under each assignment rule, with paired runs sharing initializations, fixed episode-bootstrap resamples, and minibatch orders. Every member runs 500 Bellman backups without a stopping rule.

All six neural environments use discount $\gamma = 0.99$ and Q networks with two ReLU hidden layers of width 64 and a scalar linear output, trained in single precision by Adam with learning rate 0.001, momentum coefficients $(0.9, 0.999)$, numerical constant 10^{-8}, and no weight decay, for 30 reward-regression epochs and eight epochs per Bellman backup. Every member keeps separate parameters, losses, and Adam accumulators through all fits, even when members are stacked in one tensor.

Member b draws N episode indices with replacement and minimizes the minibatch mean of $w_{bi}(Q_b(s_i, a_i) - y_{bi})^2$, where w_{bi} is the multiplicity of the episode containing transition i, without renormalizing by the sum of bootstrap weights. Each epoch divides one permutation of the parent's n transitions into $\max\{1, \lfloor n/256 \rfloor\}$ nearly equal minibatches, shared by all members and assignment rules. At each backup, all members compute targets over the full parent dataset before assignment, so conditional on the shared minibatch schedule the update has the retained-state form of Theorem A.2.

All assignment rules clip targets to the same range, $[0, 100]$ for CartPole, $[-5000, 0]$ for Cliff Walking, $[-1, 1]$ for Sepsis, and $[\min\{0, r_{\min}/(1-\gamma)\}, \max\{0, r_{\max}/(1-\gamma)\}]$ for MuJoCo, with reward extrema from the full published dataset. Physical termination sets the continuation value to zero, whereas a collection time limit does not.

Single-update comparison. For each small environment, we save the 25 own-row ensembles after backups 25, 50, 100, 200, 300, 400, and 500. Each saved ensemble then takes one own-row, whole-row, or coordinatewise update from copied member weights, Adam accumulators, and fitting random-number state, with the same recipient resamples and minibatch order. The two exceptions to the whole-row spread increase in Figure 2a are Cliff Walking at backup 25 and Sepsis at backup 400, and the IQRs there describe variation across datasets rather than across repeated assignments at a fixed checkpoint.

B.2 ENVIRONMENTS, POLICIES, AND REFERENCE RETURNS

CartPole uses CartPole-v1, with collection capped at 500 transitions. Its evaluation policy takes the greedy action of a fixed DQN with probability 0.55 and a uniform action otherwise. Cliff Walking uses a 4×12 grid in which, with probability 0.1, the chosen action is replaced by a uniform action. Ordinary transitions receive reward -1; entering the cliff receives -50 and terminates, as does reaching the goal. Collection is capped at 500 transitions, and the evaluation policy is greedy with respect to tabular Q-learning run for 50,000 episodes. Sepsis is the simulator of Oberst & Sontag (2019), with seven observed categorical variables, latent diabetes, and eight treatment actions. Death and discharge produce terminal rewards -1 and $+1$, and logged episodes are capped at 20 transitions. Its evaluation policy mixes a uniform policy equally with a greedy policy that does not observe glucose or diabetes. For each small environment, the behavior policy at shift δ is $(1 - \delta)\pi + \delta\,\mathrm{Unif}(\mathcal{A})$, with $\delta = 0.5$ in the main interventions.

The CartPole reference 67.4196 (standard error 0.0247) averages 1,048,576 default-reset rollouts capped at 500 transitions, and the Sepsis reference -0.643299 (standard error 0.000868) averages 524,288 default-reset simulations. Cliff Walking's reference -17.1612 is computed by an exact policy-evaluation solve. Thus the collection cap does not define the value horizon for Cliff Walking or Sepsis.

The MuJoCo evaluation policies are the SAC checkpoints that collected the D4RL medium-v2 datasets, run deterministically by applying $\tanh$ to the policy mean. Each reference averages 1,024 default-reset episodes capped at 1,000 transitions. This gives 226.842 for Hopper, 183.416 for

Walker2d, and 389.201 for HalfCheetah, with Monte Carlo standard errors 0.0482, 1.14, and 0.987. Each parent draws ten complete episodes without replacement, and different parents may share episodes. All random streams use fixed NumPy seeds, and the policy weights, configurations, and intervention implementation are included with the code.

B.3 Readouts, intervals, and reference horizons

Because each member's readout averages over the parent's logged initial states while coverage uses the common default-reset reference, variation in those empirical initial states also contributes to parent-level variation, separately from the reported reference Monte Carlo errors.

For a nominal $1 - \alpha = 90\%$ member interval $[L, U]$ and reference y, the central interval score is

$$S_\alpha(L, U; y) = (U - L) + \frac{2}{\alpha}(L - y)_+ + \frac{2}{\alpha}(y - U)_+. \tag{14}$$

Lower score is better. Spread ratios are computed within paired parents before aggregation, so a median paired ratio need not equal the ratio of displayed medians.

Reference horizon and coverage sensitivity. Because the Bellman backup continues past collection time limits and ends only at physical termination, the CartPole and MuJoCo references, which are capped at those limits, are truncated approximations to the value it targets. Let V_H retain rewards at times $0, \ldots, H - 1$ and let V_∞ continue until physical termination. CartPole's unit rewards give

$$0 \le V_\infty - V_{500} \le \frac{0.99^{500}}{1 - 0.99} = 0.657048\ldots. \tag{15}$$

With the saved neural intervals held fixed, the own-row, whole-row, and coordinatewise coverage counts remain $23/25$, $1/25$, and $0/25$ for every candidate reference in $[67.4196, 68.07665]$ and in its widening by 1.96 reported Monte Carlo standard errors on each side, $[67.371188, 68.125061]$. This sensitivity check uses a normal approximation for reference simulation error and is not an exact confidence bound for V_∞.

For MuJoCo, rewards certified to satisfy $|R_t| \le R_{\max}$ at all future times would give $|V_\infty - V_{1000}| \le R_{\max}\gamma^{1000}/(1 - \gamma) = 0.00431713\ldots\, R_{\max}$, but the dataset reward extrema used for clipping do not certify such a bound. We therefore report MuJoCo coverage and interval scores against the capped reference, without establishing their equality with continuation-value metrics.

C Supporting Validation and Sensitivity

C.1 Exact tabular fitting

The exact-fitting experiment in Table 2 has one nonterminal state to which both actions return, a uniform behavior policy, an evaluation policy that always selects action 1, and independent Gaussian rewards with variance one and means -0.1 and 0.2 for actions 0 and 1. Each of 25 datasets contains 400 independent one-transition observation windows, and each of $M = 100$ members draws 400 windows with replacement once and solves the weighted tabular regression exactly from zero action values, without clipping or regularization, for $\gamma \in \{0, 0.5, 0.8, 0.95\}$ and depths $1, 5, 20, 100, 200$.

The member readout is $Q_T(s, 1)$, and coverage compares linearly interpolated member percentiles with the depth-T value $0.2 \sum_{t=0}^{T-1} \gamma^t$. Because all members share the continuation coefficient γ, the own-row and whole-row ensemble means are identical. The 95% intervals for the median paired SD ratios exponentiate percentile bounds on the median log ratio from 10,000 dataset-bootstrap draws.

C.2 Finite-ensemble moment validation

We validate Proposition A.6 using fixed-feature ridge FQE on the same 25 CartPole parents with $M = 50$, $N = 10$, shift 0.5, and 500 backups, reusing the neural experiment's policy, discount, bootstrap bank, and logged-initial-state readout and fitting in double precision without clipping. The feature vector $\phi(z) = (1, \sqrt{2/64}\cos(\Omega^\top z + \beta))$ of the parent-standardized observation–action input $z \in \mathbb{R}^p$ uses $\Omega_{ij} \sim N(0, 1/p)$ and $\beta_j \sim \mathrm{Unif}(0, 2\pi)$ drawn once and shared across parents, and ridge minimizes $n^{-1} \sum_i w_{bi}(\phi_i^\top \theta - y_i)^2 + 0.001\|\theta\|^2$, penalizing the intercept as well.

Cholesky factorization gives the affine operators, and separate QR regressions reproduce their fitted outputs with a maximum relative SD error of 4.42×10^{-13} across all 525 combinations of parent, assignment rule, and checkpoint, verifying the fitting representation for fixed assignment schedules.

The assignment-averaged comparison fixes each parent's bank of 50 affine operators, initializes from the reward-only fits, and applies Proposition A.6 through 500 backups, while sampled affine recursions average the ensemble mean and sample member variance over 256 whole-row schedules per parent with a fresh permutation at every backup. Table 3 averages each quantity over the 25 parents, and its standard errors, built from across-schedule standard deviations divided by $\sqrt{256}$, quantify assignment sampling rather than uncertainty over parent datasets.

	Ensemble mean			Sample member variance		
Backup	Exact	Sampled	SE	Exact	Sampled	SE
25	21.905264	21.905225	0.000128	0.074247	0.073840	0.000212
50	32.992893	32.992580	0.000291	0.417035	0.417534	0.001287
100	41.857125	41.857229	0.000466	1.106599	1.110669	0.003302
200	45.128302	45.127364	0.000552	1.505191	1.511185	0.004387
300	45.431258	45.430047	0.000556	1.546419	1.544526	0.004499
400	45.462808	45.462571	0.000556	1.550600	1.540479	0.004409
500	45.466385	45.466225	0.000556	1.551056	1.543071	0.004463

Table 3: Exact finite-bank predictions and sampled whole-row moments in fixed-feature Bootstrap-FQE (Proposition A.6). Each entry averages over the same 25 datasets. Member variances use denominator $M - 1$. Sampled values average 256 assignment schedules per dataset; SE denotes Monte Carlo standard error of this average.

Supporting affine diagnostics. On the same operator banks, the common-operator surrogate in Proposition A.7 has positive net covariance across backup contributions at all seven checkpoints. At backup 500, the median own-row-to-expected-whole-row variance ratio is 47.59 with IQR $[28.66, 63.53]$, the surrogate's accumulation ratio has median 73.71, and its squared correlation with the full own-row response has median 66.4% and range 14.9%–92.0%, which measures alignment rather than a unique causal attribution. In the mean-shift decomposition of equation 8, with $p_t = m_t^{\mathrm{own}}/\bar{v}_t$ and every $\bar{v}_t$ nonzero, accumulated gain–value covariance is positive on all 25 banks

at backup 500, yet opposing shape contributions reverse the mean-shift direction on two banks, so both signs of mean movement coexist with the same spread ordering.

Fixed-smoother check. On a 128-point grid, each of 93 interior rows uniformly averages K Gaussian coordinates of unit variance and equicorrelation $\rho \in \{0, 0.25, 0.5, 0.75, 1\}$, and row SDs from 200,000 draws are averaged over rows. For $K \in \{4, 9, 16, 25, 36\}$, the measured SD ratios between aligned ($\rho = 1$) and independent ($\rho = 0$) coordinates differ from the ratios $2, 3, 4, 5, 6$ predicted by Proposition A.1 by 0.206%–0.237%. The aligned SD estimate is shared across K, so these comparisons are not independent confirmations.

C.3 Data-budget and ensemble-size sensitivity

The CartPole sensitivity experiment varies $N \in \{10, 20, 50, 100\}$ at $M = 50$ and $M \in \{10, 25, 50, 100\}$ at $N = 10$, with the same 25 parent seeds and other settings as Appendix B.1, reusing the original runs for $N = 10$, $M = 50$. Datasets are nested prefixes of a 100-episode parent, bootstrap draws and member initializations are nested across ensemble sizes, and whole-row permutations are shared across episode budgets at fixed M. Because each member averages over the current dataset's N logged initial states, the data-budget sweep changes the empirical readout as well as the training dataset.

Whole-row reassignment reduces the paired-median spread, lowers coverage, and raises the mean interval score in every setting (Table 4), although one parent at $N = 20$, $M = 50$ and one at $N = 10$, $M = 10$ reverse the spread ordering.

N	M	SD ratio [IQR]	Own-row larger	Coverage	Mean score
10	50	6.08 [4.88, 9.07]	25/25	$23/25 \to 1/25$	$39.07 \to 269.79$
20	50	4.35 [2.91, 6.84]	24/25	$25/25 \to 6/25$	$22.96 \to 88.93$
50	50	5.14 [4.37, 6.85]	25/25	$23/25 \to 1/25$	$14.09 \to 57.23$
100	50	5.09 [4.30, 5.73]	25/25	$24/25 \to 5/25$	$11.51 \to 30.83$
10	10	7.17 [4.56, 10.06]	24/25	$21/25 \to 4/25$	$41.69 \to 232.57$
10	25	7.25 [4.18, 8.98]	25/25	$24/25 \to 3/25$	$37.33 \to 265.54$
10	100	6.41 [4.94, 7.62]	25/25	$23/25 \to 1/25$	$39.58 \to 273.99$

Table 4: Neural CartPole sensitivity after 500 backups, with 25 paired parents per setting. SD ratios are paired medians of own-row divided by whole-row member SD, with parent IQRs in brackets. Own-row larger counts parents whose own-row SD exceeds whole-row SD. Arrows run from own-row to whole-row for 0.90 coverage counts and mean interval score, with lower score better. The first block varies N at $M = 50$; the second varies M at $N = 10$, sharing the first row as its $M = 50$ baseline.

C.4 Coupling through dropout masks

The coupling-only comparison fits $M = 25$ networks to each of 25 CartPole parents for 500 backups under own-row assignment, with the architecture, optimizer, and epoch schedule of Appendix B.1, dropout probability 0.1, and standard elementwise dropout during training. At target construction, locked masking shares one mask per hidden layer and member across queried inputs and reuses it for that backup's initial-state readout, whereas batched masking draws masks independently across queried input rows. Both rules refresh query masks at every backup.

For fixed network weights, both masking rules give the same prediction distribution at each input, yet locked masking gives greater member SD on all 25 parents, with a median paired locked-to-batched ratio of 5.89. Locked and batched masking give median value estimates of 43.34 and 51.85 and nominal 90% coverage of $8/25$ and $0/25$, so preserving coupling increases spread in this comparison without establishing calibration.

D Additional Interval Results and Implementations

D.1 Neural Bootstrap-FQE endpoints

Whole-row and coordinatewise reassignment worsen the mean interval score in every environment of Table 5, even on Hopper, where point-estimate RMSE improves from 8.769 under own-row assignment to 6.809 under whole-row and 5.921 under coordinatewise assignment. Better point estimation therefore does not imply better uncertainty estimates.

Environment	Assignment	Estimate	Member SD	Width	Mean score
CartPole	Own-row	66.88	10.85	33.68	39.07
	Whole-row	48.15	1.85	5.49	269.79
	Coordinatewise	58.76	0.39	1.16	234.71
Cliff Walking	Own-row	-23.45	8.51	20.40	28.37
	Whole-row	-21.57	1.16	3.58	60.27
	Coordinatewise	-21.89	0.67	1.98	73.65
Sepsis	Own-row	-0.58	0.26	0.76	1.01
	Whole-row	-0.59	0.06	0.18	2.90
	Coordinatewise	-0.59	0.04	0.11	2.98
Hopper	Own-row	218.67	7.68	23.39	22.77
	Whole-row	220.34	2.99	9.52	46.62
	Coordinatewise	221.46	0.95	2.95	81.72
Walker2d	Own-row	185.74	44.85	117.86	131.65
	Whole-row	222.39	8.79	29.09	560.58
	Coordinatewise	225.39	2.79	8.59	761.45
HalfCheetah	Own-row	396.12	21.23	66.86	67.99
	Whole-row	405.55	6.13	19.75	160.44
	Coordinatewise	404.80	3.48	10.43	234.15

Table 5: Neural Bootstrap-FQE endpoints on 25 paired datasets per small environment and ten per MuJoCo task. Estimates, member SDs, and widths are medians across datasets; interval scores are means. Width and score use nominal 90% intervals and the evaluation references in Appendix B.2. Lower score is better.

D.2 Published constructor implementations

Published implementations choose coupling through sampling and batching, and assignment through target indexing or sharing. Bootstrap-FQE carries one resample through a complete fitted-Q sequence and scalar readout (Hao et al., 2021), which implements own-row assignment.[2] Vectorized implementations preserve this procedure by keeping each member's resample, fitting history, target row, and readout aligned along the member axis.

MSG retains member-specific initialization and own-row targets by default and draws fresh random loss weights at every update. Its `shared_lcb` option instead broadcasts a lower-confidence target to all critics.[3]

UWAC (Wu et al., 2021; Srivastava et al., 2014) tiles each input batch across stochastic samples and applies independent elementwise dropout, producing product-style rather than function-level prediction rows.[4]

[2] aditya1909-bit/ess-ope-diagnostics, commit `5ae681c`, `episode_bootstrap` and `fitted_q_evaluation`.

[3] MSG learning code, commit `952b8d85`, preserves stacked member order and draws loss weights with `jax.random.uniform`; configuration defaults select independent targets and random weighting.

[4] `https://github.com/apple/ml-uwac`, commit `003911f`, `FlattenDropout_Mlp.multiple` and the inherited `F.dropout` call.

HyperAgent's DeepSea default draws one target index per transition and shares it across the independently drawn loss indices.[5] Disabling `state_dependent_noise_target` uses one target index for the whole batch, and enabling `dependent_target_noise` matches target and loss indices.[6]

NIMI ensemble FQE uses own-row targets by default, and setting `clip_q_values=True` broadcasts the minimum Bellman target to all critics. Its evaluator reports the across-critic variance averaged over initial states and its square root.[7] EDAC/SAC-N likewise broadcasts the minimum target to critics that retain separate states.[8]

CBOP's `indep_sampling` option changes the dependence among samples in its uncertain Bellman targets (Jeong et al., 2023).[9]

REDQ fits separately initialized critics to a common target built from a random subset of target critics (Chen et al., 2021).[10] In its published ablation with $M = 1.5$, half of the updates broadcast one critic's complete target vector to every critic. Library implementations of REDQ in TorchRL and Tianshou, and REDQ-style targets in RLPD, SERL, and TD-MPC2, make the same choice by indexing a leading member axis of target parameters or outputs, reducing across it, and broadcasting the result to all online critics.[11] SERL draws its two target indices with replacement, so with ten critics about one update in ten is a single-member broadcast.

D.3 Target-flag comparison in ensemble FQE

We run NIMI ensemble FQE under each `clip_q_values` setting (Appendix D.2) on the full Walker2d-medium-v2 dataset with training seed 2020 and identical initial parameters for 20 critics, and we report the first 25,000 updates.

At 25,000 updates, shared-minimum targets reduce the reported SD from 4.904 to 1.007, a factor of 4.869, while changing the value estimate from 115.57 to -158.07 (Table 6). Within the reported window, the own-row-to-shared-minimum SD ratio peaks at 9.245 at 11,000 updates. This comparison uses one paired training seed, and because sharing the minimum also changes the target values, it does not isolate assignment.

	Value estimate		Reported SD		
Update	Own-row	Shared-min	Own-row	Shared-min	SD ratio
0	-0.03	-0.03	0.58	0.58	1.00
11000	51.65	-4.68	2.42	0.26	9.25
25000	115.57	-158.07	4.90	1.01	4.87

Table 6: NIMI ensemble FQE on Walker2d-medium-v2, seed 2020, through 25,000 updates. Selected checkpoints show initialization, the largest SD ratio within this window, and the 25,000-update endpoint of the comparison. The ratio divides own-row by shared-minimum SD.

[5] HyperAgent, Appendix C.2 and Eq. (11), `https://arxiv.org/html/2402.10228v5`.

[6] HyperAgent policy code and DeepSea defaults, commit `014902cd`.

[7] NIMI ensemble FQE, commit `6d38532f`, `EnsembleFQE._critic_loss`, `eval`, and `learn`; YAML defaults.

[8] EDAC trainer, commit `198d5708`, `SACTrainer.train_from_torch`.

[9] `https://github.com/jihwan-jeong/CBOP`, commit `3fb693e`. The Python signature defaults to `True`; the official reproduction command selects `false`.

[10] `https://github.com/watchernyu/REDQ`, commit `7b5d1bf`, `redq/algos/redq_sac.py`, lines 9–19, 40–51, 148–164, and 200–211.

[11] TorchRL `b8b6f0c`, `torchrl/objectives/redq.py`; Tianshou `f240205`, `tianshou/algorithm/modelfree/redq.py`; RLPD `c90fd4b`, `rlpd/networks/ensemble.py`; SERL `1fa2af7`, `serl_launcher/agents/continuous/sac.py`; TD-MPC2 `e9f5932`, `tdmpc2/common/world_model.py`.