WHEN DOES CURVATURE HELP OFFLINE VALUE SELECTION?

Anonymous authors
Paper under double-blind review

ABSTRACT

Repeated Bellman updates fit the same finite dataset more closely, but need not improve the true value estimate. We study risk calibration and estimator selection with fixed linear features. Joint reward, transition, and Gram-matrix uncertainty changes both the variance and the mean of a finite iterate; omitting the mean correction can miss an order-$1/n$ risk term. We derive a same-data risk score for smooth value estimators and extend its guarantees to ordinary independent and identically distributed (iid) transitions through local conditioning and an explicit fallback on ill-conditioned datasets. For one-state problems, a characterization for any sample size and finite candidate bank identifies when the correction's downward shift improves selection and when it hurts. We also show why an endpoint-scale oracle bound alone cannot certify improvement over convergence. Matched experiments evaluate 19,800 independent datasets, including 7,200 iid datasets with overlapping features, using matched ridge, normalized-difference tournaments, paired intervals, and independent pilot comparators for each candidate family. Curvature gives modest, statistically resolved gains within both checkpoints and ridge, but independent pilots reveal much larger remaining selection gaps. Severe small-sample miscalibration and the cost of covariance estimation remain important limits. We report guards that reject the population and conservative confidence checks that certify no fits; neither calibration nor useful selection follows from an observed inverse margin alone.

1 INTRODUCTION

An offline learner can apply more Bellman updates without collecting new observations. These updates bring the iterate closer to the dataset's fixed point, but not necessarily to the true value function. This raises a concrete question: *which checkpoint should we keep when all updates use the same finite data?*

Even a stable, exactly solved update can reuse the same sampling error many times. Its targets contain predictions from the preceding iterate, so model errors affect later predictions. Convergence answers whether the numerical procedure has finished, not whether finishing is statistically preferable to retaining an earlier estimate. The true values needed to resolve that question are unavailable to an offline selector.

Consider a state that either returns to itself or terminates. An overestimate of its return probability increases the value assigned to future visits. Further updates reuse that probability; they provide no new evidence about its accuracy. More data may justify more updates, whereas a small dataset can favor a finite checkpoint. With several states or shared features, the answer also depends on where uncertainty originates and how it propagates—information not captured by sample count or convergence rate alone.

We study fixed-policy evaluation with fixed linear features. This removes policy changes, maximization bias, and neural optimization so that dataset reuse can be examined directly. The setting builds on temporal-difference learning, least-squares evaluation, and fitted batch methods (Sutton, 1988; Bradtke & Barto, 1996; Boyan, 2002; Ernst et al., 2005; Le et al., 2019). It is narrower than the full offline-control problem (Levine et al., 2020). A checkpoint here is one exact fitted Bellman update, not a neural training epoch. Our aim is to explain statistical risk and checkpoint choice in

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.

this controlled setting, rather than to claim that every offline reinforcement learning (RL) algorithm should stop early.

Neither instability nor early stopping is a new phenomenon. Linear temporal-difference (TD) convergence is classical (Tsitsiklis & Van Roy, 1997), and Perdomo et al. (2023) already characterize offline fitted Q-iteration (FQI) and least-squares temporal-difference (LSTD) learning using matrix powers, stability, and statistical complexity. Early stopping and explicit regularization also have extensive precedents (Raskutti et al., 2014; Kolter & Ng, 2009). Our starting observation is more specific: **at a biased finite checkpoint, estimated transitions can shift the mean of the iterate by an amount that affects risk at the same order as its variance.** Reward, transition, and feature-Gram errors also come from the same observations and can cancel. Propagating variance alone, or breaking this dependence, can assign the wrong risk even in a tiny realizable process.

These observations lead to three questions: can we approximate a *fixed* checkpoint's expected error, use those approximations to select from the same data, and certify the selected estimate's error with high probability? An answer to one need not answer the others. For example, minimizing unbiased scores can favor a checkpoint whose score happens to be unusually low. We keep these distinctions explicit.

Contributions. We first derive a joint empirical-moment expansion with an explicit remainder, then specialize smooth risk correction to dependent value estimates (Sections 3–4). A realizable cancellation example explains why Gram uncertainty cannot generally be omitted. A scalar finite-bank characterization covers arbitrary sample sizes and discounts, with an explicit noisy witness where curvature improves *selected risk* and a boundary beyond which it harms. It is not a general coupled-feature selection law.

Second, the sampling theorem uses conditioning only near the population moments, with a defined zero fallback elsewhere. It permits ordinary iid transitions whose individual Gram matrices are singular. The associated endpoint-scale oracle rate controls risk but does not prove the benefit of correction: its constants cannot certify improvement over the endpoint. We state this limitation rather than treating an order-$1/n$ remainder as evidence of useful adaptation.

Third, we compare candidate families with the same all-data score and check normalization, oracle choice, and covariance estimation. The experiments include overlapping features with random Gram matrices, normalized-difference LSTD-Tournament (Liu et al., 2025), and matched ridge. Same-score ridge wins the aggregate comparisons; full covariance helps with population information in the overlapping-feature suite but can cost accuracy when estimated. The main result is therefore a theory of *which terms determine risk*, with explicit evidence about when estimating them improves selection. Neither the correction nor the endpoint oracle rate replaces established risk-estimation principles. The method computes the endpoint and bank, so no computational saving over a small direct solve is claimed.

2 What changes when one dataset is reused?

2.1 The fitted update and the prediction target

A dataset observation contains a current state (or state–action pair), a reward, and a next state. The target policy, discount $0 < \gamma < 1$, and feature map are fixed before fitting. Let $\phi, \phi' \in \mathbb{R}^d$ be current and target-policy next features, with terminal feature zero. For ordinary least squares, one update fits the targets $r_i + \gamma \phi_i'^{\top} \widehat{\theta}_k$ to the current features:

$$\widehat{\theta}_{k+1} \in \arg\min_{\theta} \sum_i \left(\phi_i^{\top}\theta - r_i - \gamma \phi_i'^{\top}\widehat{\theta}_k\right)^2. \tag{1}$$

Hats mark quantities fitted from data. The feature-Gram matrix records how current features overlap. It appears in the same normal equations as the reward and next-feature moments, so their uncertainty must be considered jointly.

The population fitted update and its endpoint are

$$A\theta_{k+1} = F\theta_k + g, \qquad B = A^{-1}F, \quad b = A^{-1}g, \quad v = (A-F)^{-1}g, \tag{2}$$

where $A = \mathbb{E}\phi\phi^\top$, $F = \gamma\mathbb{E}\phi\phi'^\top$, and $g = \mathbb{E}\phi r$. We assume the displayed inverses exist and $\rho(B) < 1$ when considering convergence. Under linear realizability, v represents the true value. Otherwise it is a projected fixed point; statements about it do not automatically become true-value guarantees. Such representation and distributional distinctions matter in offline RL (Chen & Jiang, 2019; Xie & Jiang, 2021).

The empirical update replaces all three moments by estimates from the same data, starts at $\widehat{\theta}_0 = 0$, and reuses them at every step. When defined, its endpoint is $\widehat{v} = (\widehat{A} - \widehat{F})^{-1}\widehat{g}$. A small empirical Bellman residual shows that the fitted equations have been solved, not that $\widehat{v}$ is close to v. Population stability does not guarantee empirical stability; the local theorem uses a conditioning gate and defines the estimator even when a dataset fails it.

We measure $R_K = \mathbb{E}\left\|\widehat{\theta}_K - v\right\|_M^2$, where $M \succeq 0$ is a fixed evaluation matrix. For example, $M = \mathbb{E}_\nu\phi\phi^\top$ measures squared prediction error under an evaluation distribution ν, which need not equal the sampling distribution. Transforming both features and M consistently leaves the error unchanged. This prevents an apparent amplification effect from being created by a change of parameter units.

A finite candidate set $\mathcal{K}$ is specified in advance and may contain $K = \infty$, the exact empirical endpoint solve. Expectations are over independent datasets, not optimizer seeds. Here n counts independent observation *blocks*, each possibly containing several transitions whose within-block dependence is retained. Methods receive the same observation budget within each design.

2.2 A USEFUL BASELINE: KNOWN DYNAMICS

If B is known and only b is estimated, write $y = b + \varepsilon$, $\mathbb{E}\varepsilon = 0$, $\mathrm{Cov}(\varepsilon) = \Omega$, and $S_K = \sum_{j<K} B^j$. Direct expansion gives

$$R_K = \left\|B^K v\right\|_M^2 + \mathrm{tr}(MS_K\Omega S_K^\top). \tag{3}$$

The first term is unfinished population iteration. The second is dataset noise propagated through the same dynamics at every update. In particular, it contains cross terms between different propagation steps: the noise at one step is not a fresh independent draw. For scalar $B = \rho \in (0, 1)$ and $\Omega = \sigma^2/n$, this variance is $\sigma^2(1 - \rho^K)^2/[n(1 - \rho)^2]$, approximately σ^2K^2/n when $K(1 - \rho) \ll 1$. The bias decreases in this scalar example while the variance increases. Their balance depends on signal size and data quality; the matrix setting need not have a monotone or U-shaped risk curve. This familiar identity is a foundation, not our novelty claim. With a *fixed total observation budget*, reuse can have lower variance than disjoint per-update estimates (Proposition 13); fresh data are not free.

A realizable funnel makes the geometry explicit. Set $B_a = \rho I + aue^\top$, where u indicates m upstream states and e selects a downstream state, with $0 < \rho$, $0 \le a$, and $\rho + a \le \gamma$. The corresponding transition matrix is B_a/γ, with missing mass terminating. The family can preserve the true values, reward means, noise covariance, and eigenvalues while changing the propagated risk (Proposition 14). To see why, put the true downstream value at zero and the reward noise only at that state. Changing upstream-to-downstream coupling then leaves the true upstream values unchanged, but spreads downstream estimation error into more predictions. With uniform evaluation weights, the converged variance ratio to the uncoupled process is $1 + ma^2/(1 - \rho)^2$. Thus the same scalar relaxation rate can coexist with different statistical costs of convergence. This distinguishes processes with the same spectrum, while the richer complexity measures of Perdomo et al. (2023); Duan et al. (2020; 2024) can distinguish them.

Stopping is not merely numerical damping. If a damped update $\theta_{k+1} = (1 - \alpha)\theta_k + \alpha(\widehat{B}\theta_k + \widehat{b})$ converges with $\alpha \neq 0$, its limit still solves the same empirical equations. Damping changes the path; by itself it does not change the converged estimator. Keeping a finite checkpoint, in contrast, deliberately accepts some unfinished iteration in exchange for a different statistical estimator.

3 WHY VARIANCE ALONE IS NOT ENOUGH

3.1 FOLLOW THE FITTED ESTIMATOR, NOT SEPARATE NOISE SOURCES

With known dynamics, the iterate is linear in reward noise. Estimating A or F makes it nonlinear through matrix inversion and repeated multiplication. Even centered moment errors can then move its average. Expanding the *whole fitted update* preserves that effect and cancellation between inputs.

Write the moment errors as $\Delta A, \Delta F, \Delta g$, and put $U = A^{-1}\Delta A$, $V = A^{-1}\Delta F$, $w = A^{-1}\Delta g$. Let $x_{k+1} = Bx_k + b$ be the population path. The first two perturbations obey

$$u_{k+1} = Bu_k + Vx_k + w - Ux_{k+1}, \qquad u_0 = 0, \tag{4}$$

$$q_{k+1} = Bq_k + Vu_k - Uu_{k+1}, \qquad q_0 = 0. \tag{5}$$

Here u_k is linear in the joint data error, and q_k is quadratic. The exact remainder $r_k = \widehat{\theta}_k - x_k - u_k - q_k$ satisfies

$$(I + U)r_{k+1} = (B + V)r_k + Vq_k - Uq_{k+1}. \tag{6}$$

These recursions follow by substituting $\widehat{\theta}_k = x_k + u_k + q_k + r_k$ into the empirical normal equations and collecting terms by their degree in the moment error. The term $-Ux_{k+1}$ accounts for the random Gram matrix at first order; its interaction with u_{k+1} matters at second order. Dropping either term changes the expansion, even when the empirical Gram remains well conditioned. Theorem 15 supplies an explicit finite-horizon bound on the remainder. No independence between the three empirical moments is used.

For centered moments, $\mathbb{E}u_K = 0$, but generally $\mathbb{E}q_K \neq 0$. With $d_K = x_K - v$, the risk is

$$R_K = \underbrace{\|d_K\|_M^2}_{\text{iteration bias}} + \underbrace{\mathbb{E}\,\|u_K\|_M^2}_{\text{variance}} + \underbrace{2d_K^\top M\mathbb{E}q_K}_{\text{mean correction}} + \mathcal{R}_K. \tag{7}$$

The mean correction is order $1/n$ when d_K remains nonzero. On a uniformly regular bounded domain, the remainder is order $n^{-3/2}$; its constants are given in Appendix O. The word *curvature* refers to this nonlinear change in the estimator's mean, not to a Hessian of a neural loss.

The correction is signed: it can raise or lower the risk approximation depending on whether the mean shift points away from or toward the target. It is not an additional nonnegative variance penalty. If the population checkpoint is already exact, $d_K = 0$ and this particular order-$1/n$ contribution vanishes, even though the empirical estimator can still have a mean shift. This explains why an endpoint-only analysis need not expose the leading correction at an unfinished checkpoint.

3.2 ONE STATE ALREADY EXPOSES THE MISSING TERM

Suppose a state returns to itself with probability p, otherwise terminates, with known reward r. Let $\rho = \gamma p$, $\widehat{\rho} = \gamma\,\mathrm{Bin}(n, p)/n$. At three updates, $f(z) = r(1 + z + z^2)$. Writing $d = f(\rho) - r/(1 - \rho)$ and $\mu_j = \mathbb{E}(\widehat{\rho} - \rho)^j$, the exact risk is

$$d^2 + \left[r^2(1 + 2\rho)^2 + 2dr\right]\mu_2 + 2r^2(1 + 2\rho)\mu_3 + r^2\mu_4. \tag{8}$$

The term $2dr\mu_2$ is absent from a variance-only approximation. For $r > 0$, the population checkpoint underestimates the infinite-horizon value: $d = -r\rho^3/(1 - \rho) < 0$. Since $f'' = 2r > 0$, transition uncertainty raises the average finite iterate and partly offsets that negative bias. Its effect on risk is therefore negative here, despite contributing positive variance. Exact binomial calculations isolate the missing term's $1/n$ order; including it leaves an n^{-2} error *in this polynomial example*, not in every model (Figure 4). The example requires neither action maximization nor representation error.

3.3 JOINT UNCERTAINTY CAN ALSO CANCEL EXACTLY

Consider two states with scalar features $(1, 2)$. State one pays v and terminates; state two pays $2(1 - \gamma)v$ and returns to itself. If the sample fraction from state two is q, then

$$\widehat{A} = 1 + 3q, \quad \widehat{F} = 4\gamma q, \quad \widehat{g} = (\widehat{A} - \widehat{F})v. \tag{9}$$

Thus *every empirical endpoint equals* v. Full joint derivatives correctly give zero endpoint noise. Keeping the fitted $\widehat{A}$ but incorrectly freezing its uncertainty gives a spurious derivative $3v/[1+(3-4\gamma)q]$. For the candidate bank $\{0,\infty\}$, the resulting score can discard the exact endpoint. At $p=.5$, $\gamma=.9$, and $n=8$, it does so with probability $246/256$. Proposition 11 proves the selection condition. The ablation freezes the Gram only in the *uncertainty calculation*, not in the fitted estimator. Its perturbations break a relation between current-feature, next-feature, and reward moments that holds for every observation. This intentionally noise-free diagnostic does not claim that a Gram ablation always fails.

4 Calibrating and selecting value estimates

Let $Z_i = (\operatorname{vec} A_i, \operatorname{vec} F_i, g_i)$ be iid raw moment observations, $\mu = \mathbb{E}Z_i$, and $\Gamma = \operatorname{Cov}(Z_i)$. The sample mean is $\bar{Z}$ and the unbiased sample covariance is $\widehat{\Gamma}$. A Z_i may be one ordinary transition; it need not contain all states. Write f_K for the checkpoint map and $h(Z) = (A-F)^{-1}g$ for the endpoint. At a regular empirical mean, compute

$$J_K = Df_K, \quad J_\infty = Dh, \quad \widehat{\beta} = \frac{D^2h : \widehat{\Gamma}}{2n}, \quad V_{K\infty} = \frac{J_K\widehat{\Gamma}J_\infty^\top}{n}, \quad V_\infty = \frac{J_\infty\widehat{\Gamma}J_\infty^\top}{n}. \tag{10}$$

The colon contracts the endpoint Hessian with the covariance. All derivatives here are evaluated at $\bar{Z}$. The full score is

$$\widehat{R}_K^{(2)} = \|f_K - h\|_M^2 + 2\operatorname{tr}(MV_{K\infty}) - \operatorname{tr}(MV_\infty) + 2(f_K-h)^\top M\widehat{\beta}. \tag{11}$$

The first term compares two estimates from the same data. The trace terms correct their shared noise; the last term corrects the endpoint's mean shift. The score may be negative. Dropping the last term defines the no-curvature ablation. Neither score uses true values.

4.1 Why the candidate curvature cancels, but the endpoint curvature does not

Taylor-expand $\|f_K - h\|_M^2$ and $\|f_K - v\|_M^2$ at μ. Both expansions contain $(f_K(\mu)-v)^\top M(D^2f_K : \Gamma)/n$, so this term cancels in their difference. The cross-covariance and endpoint-curvature terms do not cancel. This calculation explains both the correction in (11) and its applicability to *other smooth candidate maps*, including ridge. It is a concrete specialization of risk-estimation principles, not an assertion that risk correction itself is new (Stein, 1981; Eldar, 2009).

On a bounded C^3 domain, the fixed-score calibration error is at most $C_{\mathrm{cal}}n^{-3/2}$ (Theorem 18). C_{cal} is explicit and contains the inverse-map derivatives; it can be large. Exact unbiasedness is generally impossible with a fixed finite number of Bernoulli transitions: expectations of sample statistics are polynomials in the transition probability, whereas value risk contains a resolvent denominator (Theorem 17). These statements concern expected scores for *fixed* candidates, not the loss after minimizing noisy scores (Tibshirani & Rosset, 2019; Cauchois et al., 2021; Bellec & Zhang, 2021).

4.2 A local result for ordinary iid data

Requiring inverses throughout the convex hull of raw observations would exclude ordinary unregularized least squares in $d>1$, because that hull contains the rank-one matrices $\phi\phi^\top$. We instead assume inverse margins only near the population moments and define the estimator on the remaining datasets.

Fix thresholds $a_0, c_0 > 0$ and radius T before fitting. If either $\lambda_{\min}(\widehat{A}) < a_0$ or $\sigma_{\min}(\widehat{A}-\widehat{F}) < c_0$, return zero. Otherwise minimize (11) and clip the selected parameter to the ball of radius T. Each fixed candidate is guarded in the same way. The gate needs only empirical matrices, not μ. Strict population margins above both thresholds are required. A threshold above a population margin instead creates persistent fallback bias (Proposition 10). For use beyond the synthetic designs, fix feature scaling first. Known lower bounds $a_L, c_L > 0$ can set $a_0 = a_L/2, c_0 = c_L/2$; independent-pilot bounds are an alternative, not a guarantee of useful margins. The radius T needs a candidate envelope, not just a target bound (Appendix G).

Suppose $\|Z_i - \mu\| \le R$, the maps and their first three derivatives are bounded on a ball of radius $2r$ around μ, and the gate passes and clipping is inactive on the smaller ball of radius r. Appendix B

gives sufficient population singular-value margins. Bounded differences gives

$$\mathbb{P}(\left\|\bar{Z}-\mu\right\| > r) \le t_n(r) := \exp\left[-\frac{n}{2}\left(\frac{r}{R}-\frac{1}{\sqrt{n}}\right)_+^2\right]. \tag{12}$$

Theorem 4 and Proposition 5 then yield

$$|\mathbb{E}\widetilde{S}_K - \widetilde{R}_K| \le C_{\mathrm{cal}} n^{-3/2} + C_{\mathrm{bad}}(n) t_n(r), \tag{13}$$

$$\mathbb{E}\left\|\widetilde{f}_{\widehat{K}} - v\right\|_M^2 \le (1+\eta)\min_K \widetilde{R}_K + c_\eta Q_n + d_\eta P_n + 4m_0 T^2 t_n(r). \tag{14}$$

Here $\widetilde{R}_K$ is the risk of a guarded fixed candidate, and $\widetilde{S}_K$ is its bounded reported score. All constants and the reporting convention are stated explicitly in the appendix. $Q_n, P_n = O(1/n)$ retain joint endpoint covariance. The proof uses unconditional moment identities before adding the bad-event cost; it does not incorrectly assume that covariance remains unbiased after conditioning on a well-conditioned sample. Singular individual observations and singular datasets are allowed.

What the oracle rate does not establish. The comparison underlying the second inequality is an endpoint-scale control, not evidence that curvature achieves useful adaptation. In the global regular case, $R_\infty = \mathcal{V}/n + O(n^{-3/2})$, where $\mathcal{V} = \mathrm{tr}(MDh(\mu)\Gamma Dh(\mu)^\top)$. For $\mathcal{V} > 0$, the right side of the global endpoint bound is *at least eight times this leading endpoint risk*, even after optimizing its multiplicative parameter (Proposition 7). It cannot certify beating convergence asymptotically. The scalar lower bound in Proposition 3 concerns known operators and does not establish optimality of the correction. We therefore use a direct noisy comparison, rather than that rate, to establish a selection benefit.

4.3 A FINITE-BANK SELECTION CHARACTERIZATION, WITH BOTH SIGNS

For one state with reward one, $n \ge 2$ Bernoulli transitions, any $0 < \gamma < 1$, and any finite ordered checkpoint bank, let a_j, b_j be the no-curvature and full selected predictions at count j. Theorem 8 proves $b_j \le a_j$ and

$$R_0 - R_2 = \sum_{j=0}^{n} \mathbb{P}_p(J=j)(a_j - b_j)\left(a_j + b_j - \frac{2}{1-\gamma p}\right). \tag{15}$$

The correction tilts selection toward smaller values; it helps exactly when the true value is below the switch-weighted midpoint. After multiplication by $1-\gamma p$, the sign is that of a polynomial of degree at most $n+1$. Sample size and bank size change the switching counts, not just a variance coefficient. This is a scalar characterization, not a general rule for multiple features.

The $n = 8, \gamma = 4/5, \{2, \infty\}$ witness is its simplest case: the rules differ only at count seven, and curvature helps for $0 < p < 455/604$ and hurts above that boundary (Proposition 6). At $p = .5$, risks are approximately .13 (full), .22 (no curvature), and .27 (endpoint); fixed two-step evaluation is still better at .09. For any fixed finite bank containing the endpoint, the advantage eventually becomes exponentially small at fixed $p > 0$; the explicit sufficient sample size can be enormous (Corollary 9). Neither result establishes adaptation to the best fixed candidate.

Matched ridge and implementation. For $M \succ 0$ and $T_B = I - \widehat{B}$, consider $f_\lambda = (T_B^\top M T_B + \lambda M)^{-1} T_B^\top M \widehat{b}$. Substituting Df_λ into the *same* score uses all observations for both candidate families. Appendix E gives analytic derivatives and tests. We also select over the union of the two banks. The method computes the endpoint and candidate bank, so its statistical comparisons are not claims of computational savings over a small direct solve.

5 CONTROLLED COMPARISONS AND PRACTICAL LIMITS

We separate four questions: whether correction can change selection beneficially; whether this happens beyond tabular aggregation; whether it survives matched baselines; and whether covariance estimation explains the ablation results. Simulator truth is used for evaluation and separately labeled population diagnostics, never by an operational selector.

Table 1: Mean per-setting risk relative to the same endpoint (guarded for iid, equal to one). Lower is better. The last three rows choose one fixed candidate using the same independent pilot datasets, not test truth; they are approximate fixed-oracle diagnostics, not operational baselines. Paired curvature effects are reported in Table 2.

Method	Tabular	Weighted	Sampling	iid
Full checkpoint score	.52	.57	.59	.57
No curvature	.54	.58	.61	.58
Freeze Gram uncertainty	.52	.51	.53	.55
Ridge, same full-data score	.51	.48	.51	.53
Union, same full-data score	.48	.51	.53	.54
LSTD normalized difference	.59	.56	.64	.68
Ridge, holdout and refit	.69	.68	.68	.78
Fixed eight updates	.60	.68	.70	.86
Pilot-fixed checkpoint	.29	.33	.35	.42
Pilot-fixed ridge	.31	.27	.28	.40
Pilot-fixed union	.26	.26	.28	.40

5.1 Data designs and the appropriate fixed comparator

The controlled grids contain 5,400 stratified tabular, 3,600 positive-weight random-Gram, and 3,600 random-count guaranteed-coverage datasets. The latter two use rotated group indicators and lumpable dynamics: they are tabular over groups, *not* overlapping-feature evidence. Comparisons reuse the same datasets, not additional replications.

A fourth suite has 7,200 datasets across 48 settings: 24 states, four or eight overlapping trigonometric features, cyclic or asymmetric dynamics, and $n \in \{64, 256, 1024\}$ iid transitions. No state is mandatory. Rewards are bounded and value-realizable. More distinct feature rows than dimensions rule out an invertible transformation to group indicators. Each raw Gram has rank one and the empirical Gram varies.

The candidate grid contains 16 checkpoints, from zero to 128 updates and the endpoint, and ten ridge parameters. All methods receive the same total data budget. The tournament baseline selects on an independent half-data holdout and refits the selected checkpoint on all data. Independent-candidate guarantees are not claimed for the refitted output.

The theorem's comparator is $\min_K \mathbb{E}\ell_K$, *not* $\mathbb{E}\min_K \ell_K$. An independent 1,000-dataset pilot per setting selects one fixed checkpoint, ridge parameter, and union member; each index is tested on disjoint datasets. These 132,000 auxiliary datasets are reused across families, not new replications. The pilot is an approximation, not an exact oracle; reported intervals condition on its choices.

5.2 Matched selectors change the interpretation

We use the normalized-difference LSTD-Tournament variant used in the main experiments of Liu et al. (2025). Its witnesses are q_i/c_i and $(q_j - q_i)/c_{ji}$, with empirical standard deviations on the holdout; TD residuals use unnormalized candidates. Zero witnesses contribute zero; other denominators are floored at 10^{-10} times the largest candidate holdout root-mean-square value, itself bounded below by 10^{-30}. Ridge also receives the same full-data covariance and curvature correction as checkpoints. This isolates early stopping from the benefit of avoiding a split. We also include a ridge holdout comparator.

Results under the matched comparisons The full score improves on its no-curvature ablation within both candidate families, with paired intervals below. However, same-data ridge is better in all four grids, and the union does not uniformly outperform ridge alone. The normalized-difference tournament does not produce a uniform winner. The appropriate conclusion is that the score is a way to assess *smooth value estimators*, not that checkpointing is the superior regularizer. In iid data, the checkpoint–ridge gap, multiplied by 100, is 3.49: 1.89 comes from pilot-fixed families and 1.60 from pilot-relative selection gaps. The union has the best pilot comparator ($.40$) yet its adaptive score ($.54$) loses to ridge ($.53$). A stronger candidate bank does not ensure better selection.

Table 2: Full minus no-curvature normalized risk, multiplied by 100, with paired 95% Monte Carlo intervals. Negative favors curvature within the same family. All four ridge contrasts remain below zero after a four-comparison Bonferroni adjustment; effects are small relative to the pilot selection gaps.

Suite	Checkpoint curvature effect	Ridge curvature effect
Tabular	-2.20 [-2.67, -1.73]	-2.01 [-2.46, -1.56]
Weighted Gram	-.89 [-1.20, -.58]	-1.07 [-1.48, -.66]
Sampling Gram	-1.74 [-2.17, -1.31]	-1.27 [-1.70, -.83]
iid overlap	-1.16 [-1.41, -.91]	-1.13 [-1.38, -.87]

5.3 Paired differences and population diagnostics

For each fixed configuration, we divide its mean method risk by its mean endpoint risk, then average configurations equally. Paired delta-method 95% intervals preserve covariance among both methods and the denominator. They quantify Monte Carlo uncertainty for these fixed grids, not variation over an unseen task population. Table 2 reports within-family curvature effects. Ridge gains survive a four-suite Bonferroni adjustment, but remain small relative to its .13–.22 pilot gaps. These are asymptotic Monte Carlo statements, not new-task guarantees.

We replace the empirical covariance by the *exact population covariance*, while retaining the fitted candidates and sample-point derivatives. We also examine separate and joint endpoint-curvature replacements; these truth-based diagnostics are not offline algorithms. In the two grouped random-Gram suites, true covariance reduces but *does not reverse* the advantage of freezing Gram uncertainty: the remaining paired gaps are approximately .03 and .03. In contrast, in the genuinely overlapping-feature iid suite, population covariance *reverses* the empirical frozen-Gram advantage: full minus frozen becomes $-.03$ with paired interval $[-.04, -.02]$. Population curvature does not uniformly improve results. Covariance-estimation cost matters in iid data but does not alone explain the grouped ablation advantage. Random derivatives, higher-order terms, and selection remain entangled.

Prespecified half-covariance and half-Gram checks retain unfavorable cases: halving all terms is worse, while half-Gram lies between full and frozen in the grouped grids. These are descriptive ablations, not truth-tuned recommendations.

5.4 Small samples and fallback behavior are part of the result

The scalar warning is important enough to state in the main text. At $p = .7$, $\gamma = .95$, reward .1, $K = 3$, and $n = 8$, expected corrected score minus true risk is .10, while risk is only 8.54×10^{-3}. This exact binomial calculation exposes a severe failure near a small resolvent denominator; an $n^{-3/2}$ statement with large derivative constants is not quantitatively reassuring there. The diagnostic ratio $\gamma\sqrt{p(1-p)/n}/(1-\gamma p)$ measures transition uncertainty relative to the denominator, but is not by itself a calibration certificate. Boundary events can matter even when it is modest.

The fixed gate ($a_0 = .03, c_0 = 3/200, T = 10$) returns zero on 891 of 7,200 iid datasets. At $n = 64, 256, 1024$, the fallback fractions are 23.4%, 10.7%, 3.0%, respectively. No output clipping was activated. All datasets that fail the gate remain in the risk averages. In this suite "endpoint" means the guarded LSTD solution, not an unconditional convergence claim. The primary gate has positive population margins throughout the grid. In contrast, sensitivity threshold $c_0 = .02$ exceeds $29/1600$ in the cyclic $d = 8, \gamma = .95$ populations, causing persistent fallback at both signal levels. Empirical margins are only diagnostics. A conservative support-bound calculation certifies both inverses on none of the 7,200 datasets, despite 6,309 empirical gate passes. Appendix G distinguishes population assumptions from observable diagnostics and certificates.

6 Relation to prior work and scope

Evaluation and operator structure. LSTD and least-squares policy iteration (LSPI) establish the linear fixed-point perspective (Bradtke & Barto, 1996; Boyan, 2002; Lagoudakis & Parr, 2003); off-policy trace methods extend it (Geist & Scherrer, 2014). Fitted-value analyses track finite-sample

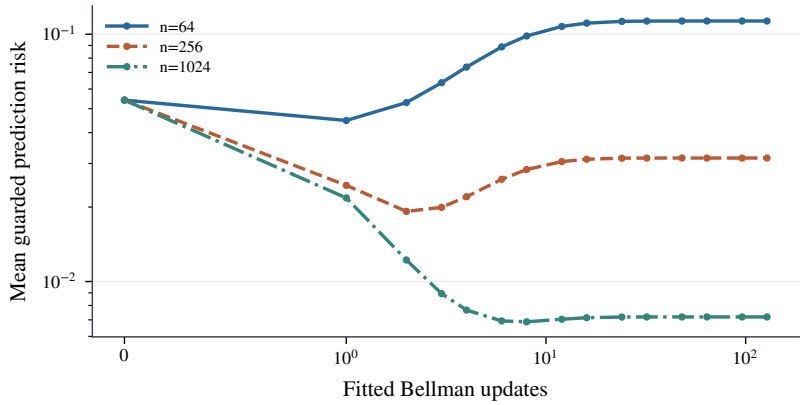


Figure 1: Ordinary iid transitions with overlapping features: asymmetric dynamics, $d = 8$, $\gamma = .95$, and signal .35. More data shifts the useful range of updates. Each curve averages 150 datasets, including fallbacks. The full grid, not this illustration, determines Table 1.

error propagation and distribution shift (Munos & Szepesvári, 2008), while representation-based analyses study stability (Ghosh & Bellemare, 2020). These results already go beyond spectral radius. We focus on the mean and covariance of a *finite empirical iterate*, retaining joint raw-moment errors.

Risk estimation and stopping. Stein's unbiased risk estimate (SURE) and its extensions, early stopping, and affine aggregation underpin this work (Eldar, 2009; Raskutti et al., 2014; Dalalyan & Salmon, 2012). Higher-order Stein formulas quantify Gaussian risk-estimate uncertainty (Bellec & Zhang, 2021); our correction instead follows the nonlinear endpoint under uncertain transitions. Instance-dependent RL confidence and stopping are studied by Xia et al. (2023). Statistical bootstrapping for fitted Q-evaluation (FQE) measures shared-data uncertainty (Hao et al., 2021), using resampling rather than Bellman updates. We add a joint expansion, local iid guarantees, and explicit selection examples.

Other offline evaluation and selection routes. Doubly robust and minimax methods address important complementary value-estimation problems (Jiang & Li, 2016; Uehara et al., 2020). Modern offline selectors include Bellman-error validation, LSTD-Tournament, and the duality-based residual estimator (DRE) (Zitovsky et al., 2023; Liu et al., 2025; Miyaguchi, 2026). Recent work also studies stopping with prior data (Romeo et al., 2026) and feature-gradient covariance control (Qiao et al., 2026). Our empirical comparisons use LSTD-Tournament and ridge baselines.

Limitations and conclusion. The analysis assumes fixed features, bounded independent observations or blocks, and local inverse margins. It excludes learned representations, long dependent trajectories, policy maximization, and downstream control. The guard exposes difficult datasets but creates persistent bias when it rejects the population; passing it does not certify calibration. Curvature can improve selected risk, but gains depend on the setting, calibration constants can be large, and simpler covariance estimates or regularizers can win. Convergence, risk calibration, and successful selection remain distinct properties.

AI USE STATEMENT

In this work, we used generative AI tools to draft sections of the paper and polish writing. Additionally, we used generative AI tools for editing code. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Pierre C. Bellec and Cun-Hui Zhang. Second-order Stein: SURE for SURE and other applications in high-dimensional inference. *The Annals of Statistics*, 49(4):1864–1903, 2021. doi: 10.1214/20-AOS2005. URL https://arxiv.org/abs/1811.04121.

Justin A. Boyan. Technical update: Least-squares temporal difference learning. *Machine Learning*, 49(2–3):233–246, 2002. doi: 10.1023/A:1017936530646. URL https://link.springer.com/article/10.1023/A:1017936530646.

Steven J. Bradtke and Andrew G. Barto. Linear least-squares algorithms for temporal difference learning. *Machine Learning*, 22:33–57, 1996. doi: 10.1007/BF00114723. URL https://link.springer.com/article/10.1007/BF00114723.

Lawrence D. Brown and Mark G. Low. A constrained risk inequality with applications to nonparametric functional estimation. *The Annals of Statistics*, 24(6):2524–2535, 1996. doi: 10.1214/aos/1032181166.

Maxime Cauchois, Alnur Ali, and John Duchi. A comment and erratum on "excess optimism: How biased is the apparent error of an estimator tuned by SURE?". *arXiv preprint arXiv:2112.14353*, 2021. URL https://arxiv.org/abs/2112.14353.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1042–1051, 2019. URL https://proceedings.mlr.press/v97/chen19e.html.

Arnak Dalalyan and Joseph Salmon. Sharp oracle inequalities for aggregation of affine estimators. *The Annals of Statistics*, 40(4):2327–2355, 2012. doi: 10.1214/12-AOS1038. URL https://arxiv.org/abs/1104.3969.

Yaqi Duan, Zeyu Jia, and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2701–2709, 2020. URL https://proceedings.mlr.press/v119/duan20b.html.

Yaqi Duan, Mengdi Wang, and Martin J. Wainwright. Optimal policy evaluation using kernel-based temporal difference methods. *The Annals of Statistics*, 52(5):1927–1952, 2024. doi: 10.1214/24-AOS2399. URL https://arxiv.org/abs/2109.12002.

John Duchi and Feng Ruan. A constrained risk inequality for general losses. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 802–810, 2021. URL https://proceedings.mlr.press/v130/duchi21a.html.

Yonina C. Eldar. Generalized SURE for exponential families: Applications to regularization. *IEEE Transactions on Signal Processing*, 57(2):471–481, 2009. doi: 10.1109/TSP.2008.2008212. URL https://doi.org/10.1109/TSP.2008.2008212.

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://jmlr.org/papers/v6/ernst05a.html.

Matthieu Geist and Bruno Scherrer. Off-policy learning with eligibility traces: A survey. *Journal of Machine Learning Research*, 15(10):289–333, 2014. URL https://www.jmlr.org/papers/v15/geist14a.html.

Dibya Ghosh and Marc G. Bellemare. Representations for stable off-policy reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3556–3565, 2020. URL https://proceedings.mlr.press/v119/ghosh20b.html.

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, 2021. URL https://proceedings.mlr.press/v139/hao21b.html.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 652–661, 2016. URL https://proceedings.mlr.press/v48/jiang16.html.

J. Zico Kolter and Andrew Y. Ng. Regularization and feature selection in least-squares temporal difference learning. In *Proceedings of the 26th International Conference on Machine Learning*, pp. 521–528, 2009. doi: 10.1145/1553374.1553442. URL https://zicokolter.com/publications/kolter2009regularization.pdf.

Michail G. Lagoudakis and Ronald Parr. Least-squares policy iteration. *Journal of Machine Learning Research*, 4:1107–1149, 2003. URL https://www.jmlr.org/papers/v4/lagoudakis03a.html.

Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3703–3712, 2019. URL https://proceedings.mlr.press/v97/le19a.html.

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

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, pp. 42942–42976, 2025. doi: 10.52202/085713-1430. URL https://proceedings.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 *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 1576–1584, 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 https://jmlr.org/papers/v9/munos08a.html.

Juan C. Perdomo, Akshay Krishnamurthy, Peter Bartlett, and Sham Kakade. A complete characterization of linear estimators for offline policy evaluation. *Journal of Machine Learning Research*, 24 (284):1–50, 2023. URL https://jmlr.org/papers/v24/22-0341.html.

Nan Qiao, Sheng Yue, Shuning Wang, Yongheng Deng, and Ju Ren. Less is more: Clustered cross-covariance control for offline RL. *arXiv preprint arXiv:2601.20765*, 2026. URL https://arxiv.org/abs/2601.20765.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Early stopping and non-parametric regression: An optimal data-dependent stopping rule. *Journal of Machine Learning Research*, 15(11):335–366, 2014. URL https://jmlr.org/papers/v15/raskutti14a.html.

Carlo Romeo, Girolamo Macaluso, Alessandro Sestini, and Andrew D. Bagdanov. SOPE: Stabilizing off-policy evaluation for online RL with prior data. *arXiv preprint arXiv:2605.05863*, 2026. URL https://arxiv.org/abs/2605.05863.

Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981. doi: 10.1214/aos/1176345632.

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://link.springer.com/article/10.1007/BF00115009.

Ryan J. Tibshirani and Saharon Rosset. Excess optimism: How biased is the apparent error of an estimator tuned by SURE? *Journal of the American Statistical Association*, 114(526):697–712, 2019. doi: 10.1080/01621459.2018.1429276. URL https://arxiv.org/abs/1612.09415.

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://www.mit.edu/~jnt/Papers/J063-97-bvr-td.pdf.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and Q-function learning for off-policy evaluation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9659–9668, 2020. URL https://proceedings.mlr.press/v119/uehara20a.html.

Eric Xia, Koulik Khamaru, Martin J. Wainwright, and Michael I. Jordan. Instance-dependent confidence and early stopping for reinforcement learning. *Journal of Machine Learning Research*, 24(392):1–43, 2023. URL https://jmlr.org/papers/v24/23-0646.html.

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, 2021. URL https://proceedings.mlr.press/v139/xie21d.html.

Joshua P. Zitovsky, Daniel de Marchi, Rishabh Agarwal, and Michael R. 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, 2023. URL https://proceedings.mlr.press/v202/zitovsky23a.html.

A ENDPOINT-SCALE COMPARISON UNDER GLOBAL REGULARITY

This section states a global sufficient condition and proves the endpoint comparison used in the local iid analysis. The support-hull condition applies to the guaranteed-coverage grouped designs, but rank-one individual Gram matrices in dimension greater than one invalidate it for ordinary iid transitions. Appendix B gives a local condition that permits these singular individual observations.

A.1 THE KEY REDUCTION: COMPARE WITH A CORRECTED ENDPOINT

Set $t = h(\bar{Z}) - \widehat{\beta}$ and $p_K = 2\,\mathrm{tr}(MV_{K\infty})$. Completing the square in (11) shows that its minimizer also minimizes $\|f_K - t\|_M^2 + p_K$. The following elementary comparison works even though everything was estimated from the same data.

Lemma 1 (Penalized-endpoint comparison)**.** *For arbitrary candidates a_K, target v, endpoint t, and real penalties p_K, let $\widehat{K}$ minimize $\|a_K - t\|_M^2 + p_K$. For every $\eta > 0$,*

$$\left\|a_{\widehat{K}} - v\right\|_M^2 \le (1+\eta)\min_K \|a_K - v\|_M^2 + c_\eta \|t - v\|_M^2 + d_\eta\,\mathrm{osc}(p), \tag{16}$$

where $c_\eta = (2+\eta)^2/\eta$, $d_\eta = (2+\eta)/2$, and $\mathrm{osc}(p) = \max_K p_K - \min_K p_K$.

This is a penalized least-squares comparison, not a new general aggregation principle; sharper results are available for other statistical models and aggregation procedures (Dalalyan & Salmon, 2012). Its usefulness here is that it avoids concentration of each dependent checkpoint score. The right side uses the unknown target and is an analysis tool, not a computable confidence certificate.

Assumption 1 (Uniform regularity)**.** $n \ge 2$ and $\|Z_i - \mu\|_2 \le R$ almost surely. On a neighborhood of the convex hull of the observation support, all f_K, h are C^2, their first derivatives have norm at most L_1, and $D^2 h$ has norm at most L_2. Every inverse defining these maps exists on this domain. Constants are uniform over the fixed candidate family. Derivative norms are Euclidean multilinear operator norms.

The condition is stronger than invertibility on the observed sample. Our random-Gram constructions satisfy it through positive coverage and aggregation; we do not infer it merely from successful numerical inversion. Higher derivatives and bounded map values are additionally needed for Theorem 18.

Theorem 2 (Expected oracle bound with joint endpoint noise)**.** *Under Assumption 1, define*

$$G = \mathrm{tr}\,\Gamma, \qquad \mathcal{V} = \mathrm{tr}\left(MDh(\mu)\Gamma Dh(\mu)^\top\right), \tag{17}$$

$$a = \sqrt{m_0}L_2R^2(1+\sqrt{3}/2), \qquad Q_n = \left(\sqrt{\mathcal{V}/n} + a/n\right)^2, \tag{18}$$

$$P_n = \frac{4\sqrt{m_0}L_1\sqrt{G\mathcal{V}}}{n} + \frac{8m_0L_1L_2R^3}{n^{3/2}}. \tag{19}$$

The selector in (11) *satisfies, for every $\eta > 0$,*

$$\boxed{\mathbb{E}\left\|\widehat{\theta}_{\widehat{K}} - v\right\|_M^2 \le (1+\eta)\min_{K\in\mathcal{K}} R_K + c_\eta Q_n + d_\eta P_n.} \tag{20}$$

The leading remainder is order $1/n$, not $1/\sqrt{n}$. Its endpoint quantity $\mathcal{V}$ preserves covariance cancellations among reward, transition, and Gram errors. The result has no explicit $\log|\mathcal{K}|$ term, but the common derivative bounds can depend on the family, dimension, coverage, and horizon. It is a fixed-dimensional, multiplicative oracle inequality, not an exact-leading-constant or dimension-minimax result. Constants can be large. Curvature improves fixed-score calibration; it is *not* necessary for every endpoint-based rule to attain an order-$1/n$ bound.

Proof idea. Lemma 1 reduces the task to endpoint error and penalty variation. Taylor expansion and $\mathbb{E}\left\|\bar{Z} - \mu\right\|^4 \le 3R^4/n^2$ give $\mathbb{E}\,\|t - v\|_M^2 \le Q_n$. Cauchy–Schwarz, $\mathbb{E}\widehat{\Gamma} = \Gamma$, and the Lipschitz derivative of h give $\mathbb{E}\,\mathrm{osc}(p) \le P_n$. No independence between $\bar{Z}$ and $\widehat{\Gamma}$ is used. Appendix H gives every constant.

A.2 A LOWER BOUND IN THE KNOWN-OPERATOR SUBCLASS

Proposition 3 (A bounded scalar lower bound). *Let $n \geq 16$, known $B = \rho \in [0, 1)$, and iid rewards $X_i \in \{-1, 1\}$ with mean μ. Put $\tau^2 = 1/[n(1-\rho)^2]$, $v_\mu = \mu/(1-\rho)$, and $c = 1/(16e^4)$. For every estimator T and $0 < \eta \leq 1$,*

$$\max_{\mu \in \{0, 2/\sqrt{n}\}} \left\{ \mathbb{E}_\mu (T - v_\mu)^2 - (1+\eta) \min \left[v_\mu^2, (1-\mu^2)\tau^2 \right] \right\} \geq c\tau^2. \tag{21}$$

The comparator consists only of zero and the empirical endpoint. A constrained-risk argument (Brown & Low, 1996; Duchi & Ruan, 2021) proves the result. It matches the $n^{-1}(1-\rho)^{-2}$ order in the known-operator scalar subclass. It does not establish a matching general matrix lower bound or a sharp result for unknown transitions.

B ORDINARY IID TRANSITIONS: LOCAL CONDITIONING AND A DEFINED FALLBACK

The whole-support condition in Assumption 1 does not hold for raw, unregularized least squares with $d > 1$: every observation has Gram contribution $\phi\phi^\top$ of rank at most one. Population invertibility does not change this fact. We instead require local population inverse margins and define the estimator on all datasets through a guard and fallback.

Write $Z_i = (\operatorname{vec} A_i, \operatorname{vec} F_i, g_i)$, with $n \geq 2$ independent, identically distributed observations, $\mu = \mathbb{E} Z_i$, $\Gamma = \operatorname{Cov}(Z_i)$, and $\|Z_i - \mu\| \leq R$. One Z_i may now be *one transition*. Let $\mathcal{B}_{2r} = \{z : \|z - \mu\| \leq 2r\}$ for a deterministic $r > 0$. The ball is taken within the moment space with symmetric A. Only inverse maps on this ball, not at individual observations, must be regular.

B.1 A FULLY SPECIFIED GUARDED ESTIMATOR

Choose positive constants a_0, c_0, T before observing the data. Compute the raw empirical moments and accept if

$$\lambda_{\min}(\widehat{A}) \geq a_0, \qquad \sigma_{\min}(\widehat{A} - \widehat{F}) \geq c_0. \tag{22}$$

On acceptance, compute the ordinary fitted checkpoint maps and their same-data scores. Choose the smallest-index minimizing score, and project its parameter vector onto the Euclidean ball of radius T. If the gate fails, return zero. Define each guarded fixed candidate $\widetilde{f}_K$ by the same gate and projection, without selection. These are estimators on *every* dataset, including singular ones. The thresholds are supplied analysis/design parameters, not estimated population margins disguised as a data-driven guarantee. Projection is merely a safeguard; the following event makes it inactive. The thresholds and Euclidean projection refer to a fixed feature coordinate system; they are not invariant to arbitrary feature rescaling. Coordinate-invariance tests concern the unguarded score and prediction norm, with M transformed consistently. The iid experiment fixes its feature normalization in advance.

Assume all f_K, h are C^3 on a neighborhood of $\mathcal{B}_{2r}$, with norms at most T and derivative bounds L_1, L_2, L_3 there. Assume that $\mathcal{B}_r$ satisfies the gate. A sufficient condition is

$$\lambda_{\min}(A(\mu)) > a_0 + 2r, \qquad \sigma_{\min}(A(\mu) - F(\mu)) > c_0 + 2\sqrt{2}r.$$

The constants 2 supply slack for the larger ball; they are not essential. Weyl's singular-value inequality and $\|\Delta A - \Delta F\|_{\mathrm{op}} \leq \sqrt{2}\,\|\Delta Z\|$ prove this sufficient condition. The maps are finite rational functions for a fixed checkpoint bank, so inverse margins on the ball imply finite derivative bounds. The endpoint is the LSTD solution; calling it an iteration limit additionally requires empirical stability. The guarded theorem itself only uses the map, not that additional assertion.

Put $E_n = \{\|\bar{Z} - \mu\| \leq r\}$ and

$$t_n(r) = \exp\left[-\frac{n}{2} \left(\frac{r}{R} - \frac{1}{\sqrt{n}} \right)_+^2 \right]. \tag{23}$$

Then $\mathbb{P}(E_n^c) \leq t_n(r)$. Indeed $\mathbb{E}\,\|\bar{Z} - \mu\| \leq R/\sqrt{n}$; replacing one observation changes this norm by at most $2R/n$. Bounded differences gives (23), using the trivial bound one if $r \leq R/\sqrt{n}$. No independence is asserted after conditioning on E_n.

Theorem 4 (Local iid oracle extension). *Under the local conditions above, let* $\widetilde{R}_K = \mathbb{E}\left\|\widetilde{f}_K - v\right\|_M^2$ *and* $v = h(\mu)$. *Use the constants* Q_n, P_n, c_η, d_η *in Theorem 2, but with the* local *derivative bounds. The guarded selector satisfies*

$$\mathbb{E}\left\|\widetilde{f}_{\widehat{K}} - v\right\|_M^2 \le (1+\eta)\min_K \widetilde{R}_K + c_\eta Q_n + d_\eta P_n + 4m_0T^2 t_n(r). \tag{24}$$

It allows singular individual Gram contributions and positive probability of a singular empirical Gram. It does not require one observation of every state or feature direction in every block.

Proof. On E_n, the gate passes and clipping is inactive for every candidate. Apply the pathwise penalized-endpoint lemma there. Its endpoint Taylor remainder and penalty estimates only use the segment from μ to $\bar{Z}$, hence the local bounds suffice. For example,

$$\mathbb{E}[\mathbf{1}_{E_n}\|t - v\|_M^2] \le Q_n, \qquad \mathbb{E}[\mathbf{1}_{E_n}\operatorname{osc}(p)] \le P_n.$$

To verify these inequalities, retain the indicator until applying Cauchy–Schwarz or Minkowski, then dominate nonnegative moment terms by their *unconditional* expectations. In particular $\mathbb{E}\widehat{\Gamma} = \Gamma$ is used only unconditionally; $\mathbb{E}[\widehat{\Gamma} \mid E_n] = \Gamma$ would generally be false. For any fixed comparator j, its good-event expected loss is at most $\widetilde{R}_j$. Outside E_n, both target and output have norm at most T, so loss is at most $4m_0T^2$. Add this contribution and minimize the bound over deterministic j. □

B.2 Local calibration, including the cost of the bad event

For risk reporting, set the score to zero on gate failure and clip each computed score to $[-S, S]$ on acceptance. This clipping is optional for selection; either minimizing the clipped scores or the unclipped scores has the same good-event rule and satisfies Theorem 4. Choose

$$D_0 = m_0(3L_1^2 + 2TL_2), \qquad S \ge 4m_0T^2 + D_0R^2.$$

On E_n, clipping is inactive because the gap norm is at most $2T$ and $\operatorname{tr}\widehat{\Gamma}/n \le 2R^2/n \le R^2$. Denote the resulting reported score by $\widetilde{S}_K$.

Proposition 5 (Local fixed-checkpoint calibration). *With* $C_{\mathrm{cal}} = m_0R^3[(2\sqrt{3}+4)TL_3 + (5\sqrt{3}+16)L_1L_2]$,

$$|\mathbb{E}\widetilde{S}_K - \widetilde{R}_K| \le C_{\mathrm{cal}}n^{-3/2} + C_{\mathrm{bad}}(n)t_n(r), \tag{25}$$

$$C_{\mathrm{bad}}(n) = S + 4m_0T^2 + 4m_0TL_1R + D_0R^2(1+2/n). \tag{26}$$

The bound is uniform over the fixed family. Its conditioning and discount dependence remains in L_j, r, T*; it is not a blanket small-sample guarantee.*

Proof. Use the functions g, ℓ, C in Appendix R. Let $e = \bar{Z} - \mu$, and form the polynomial

$$P(e, \widehat{\Gamma}) = D(g-\ell)(\mu)e + \tfrac{1}{2}D^2(g-\ell)(\mu)[e,e] + C(\mu, \widehat{\Gamma})/n.$$

The cancellation identity (71), centering, and unbiased sample covariance give $\mathbb{E}P = 0$ without conditioning. On E_n the same Taylor and Lipschitz estimates as in Theorem 18 bound the expected remainder by $C_{\mathrm{cal}}n^{-3/2}$. On all observations, $\|e\| \le R$ and $\operatorname{tr}\widehat{\Gamma} \le 2R^2$. The first-derivative difference is $-2Dh(\mu)^\top M(f_K(\mu) - h(\mu))$, with norm at most $4m_0TL_1$. The half-Hessian difference and covariance-linear correction have norms at most D_0 by (71). Consequently

$$|P| \le 4m_0TL_1R + D_0R^2(1+2/n).$$

Outside E_n, the reported score minus guarded loss has absolute value at most $S + 4m_0T^2$. Substitute $\mathbb{E}P = 0$, bound the difference from P on E_n^c, and use (23). □

C A NOISY REGIME WHERE CURVATURE CHANGES SELECTION FOR THE BETTER

This result compares the *actual selected estimator*, not the calibration of fixed scores and not the endpoint-scale bound. Consider one state with deterministic reward one and unknown return probability $p \in (0,1)$. Take $\gamma = 4/5$, eight independent transitions, $M = 1$, and the bank $\{2, \infty\}$. Let $J \sim \mathrm{Bin}(8,p)$, $q = J/8$, $z = 4q/5$, and $w = (4/5)^2 q(1-q)/7$. Define $f = 1+z$, $h = (1-z)^{-1}$, $h' = (1-z)^{-2}$, and $\beta = w(1-z)^{-3}$. The no-curvature and full score differences against the endpoint are

$$D_0(q) = (f-h)^2 + 2wh'(1-h'), \tag{27}$$

$$D_2(q) = D_0(q) + 2(f-h)\beta. \tag{28}$$

Choose the finite checkpoint for a nonpositive difference, resolving ties in its favor.

Proposition 6 (Eight-count noisy witness)**.** *The two rules differ only when $J = 7$. At that count the no-curvature rule chooses $10/3$ and the full rule chooses $17/10$. If R_0, R_2 are their expected prediction risks, then*

$$R_0 - R_2 = 8p^7(1-p)\frac{49}{30}\left(\frac{151}{30} - \frac{2}{1-4p/5}\right). \tag{29}$$

Thus curvature strictly improves selection for $0 < p < 455/604$, ties at $455/604$, and harms selection for $455/604 < p < 1$. At $p = 1/2$,

$$R_2 = \frac{121}{900}, \qquad R_0 = \frac{6371}{28800}, \qquad R_0 - R_2 = \frac{833}{9600}.$$

Proof. Substitute each rational $q \in \{0, 1/8, \ldots, 1\}$ in (28). The signs agree at all counts except seven, where $D_2 < 0 < D_0$; at zero the identical candidates tie. All signs can be checked directly from the following rational values; no simulation is used:

J	0	1	2	3	4	5	6	7	8
$\operatorname{sgn} D_0$	0	$-$	$-$	$-$	$-$	$-$	$-$	$+$	$+$
$\operatorname{sgn} D_2$	0	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$+$

In particular $D_0(7/8) = 3409/8100$ and $D_2(7/8) = -6391/8100$. At seven, $z = 7/10$, $f = 17/10$, and $h = 10/3$. The risk difference at this count is

$$(h-v)^2 - (f-v)^2 = (h-f)(h+f-2v), \quad v = (1-4p/5)^{-1}.$$

Multiplication by $\mathbb{P}(J = 7) = 8p^7(1-p)$ gives (29). The bracket is positive exactly when $p < 455/604$. Summing all nine rational losses at $p = 1/2$ gives the stated risks. □

The endpoint risk at $p = 1/2$ is $158645/580608 \simeq .27$, whereas the fixed two-step risk is $41/450 \simeq .09$. The full rule beats both the no-curvature rule and convergence, but not the best fixed checkpoint in this particular environment. At $J = 8$, empirical transition variance is zero and both selectors retain a poor endpoint. The theorem is a constructive regime result, not a universal adaptation claim or a new minimax rate. It was found by an exploratory scalar search and subsequently proved exactly; it is not a pre-registered benchmark win.

D WHAT THE ENDPOINT-SCALE ORACLE INEQUALITY CANNOT PROVE

Proposition 7 (Nondegenerate endpoint-scale obstruction)**.** *Suppose the global assumptions hold and $\mathcal{V} > 0$. Then $R_\infty = \mathcal{V}/n + O(n^{-3/2})$. For any $\eta > 0$, the right-hand side of* (20) *is at least $8\mathcal{V}/n$. Thus its ratio to R_∞ has lower limit at least eight, even if η is optimized with n.*

Proof. At the population endpoint the zeroth-order error is zero, so (76) gives the endpoint expansion. Also $Q_n \geq \mathcal{V}/n$, $P_n \geq 0$, and $c_\eta = \eta + 4 + 4/\eta \geq 8$. The comparator risk is nonnegative. Division by the endpoint expansion proves the assertion. □

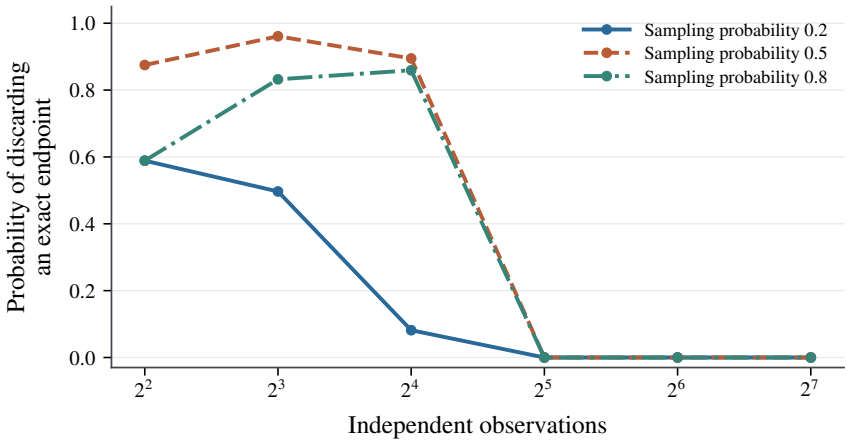


Figure 2: Exact two-state calculation. Freezing Gram uncertainty can favor zero over an endpoint that is correct on every dataset. The full joint score never makes this error in this construction. Probabilities are computed by binomial enumeration, not estimated by simulation.

The bound controls the scale of the selected risk but does not certify an improvement over convergence in this nondegenerate asymptotic regime. Its scalar lower bound does not change that conclusion and does not establish optimality of curvature correction. The direct noisy comparison in Proposition 6, not the order-$1/n$ rate, supplies a selection advantage. When $\mathcal{V} = 0$, the stated obstruction is inapplicable; special cancellations may permit smaller remainders. Those are separately analyzed cases, not evidence that the generic constants are tight.

E Matched all-data scores for ridge candidates

Assume $M \succ 0$. Set $T_B = I - \widehat{B}$. For both same-data and holdout selection, use $\lambda \in \{0, 10^{-5}, 10^{-4}, 10^{-3}, 3 \times 10^{-3}, .01, .03, .1, .3, 1\}$ and define

$$f_\lambda = (T_B^\top M T_B + \lambda M)^{-1} T_B^\top M \widehat{b}.$$

We use the same h, joint moment covariance, and endpoint curvature as for checkpoints. No observation split is used by either corrected-score method. The score formula (11) applies to any smooth candidate map f, since candidate curvature cancels in (71). This isolates the choice of estimator family from the choice of selection data.

For a centered raw direction (a, f, g), write

$$\dot{B} = \widehat{A}^{-1}(f - a\widehat{B}), \quad \dot{b} = \widehat{A}^{-1}(g - a\widehat{b}), \quad \dot{T} = -\dot{B}.$$

If $G_\lambda = T_B^\top M T_B + \lambda M$, differentiate its normal equations:

$$\dot{f}_\lambda = G_\lambda^{-1}\{\dot{T}^\top M(\widehat{b} - T_B f_\lambda) + T_B^\top M(\dot{b} - \dot{T} f_\lambda)\}. \tag{30}$$

Summing direction products gives its covariance with the endpoint. The implementation tests these derivatives against central differences for random nondiagonal Gram matrices. Ridge, checkpoints, and their union each minimize the identical full score on their respective prespecified banks. A result for one bank is not credited as a distinct data-budget advantage.

F Selection boundaries beyond one count and one sample size

The eight-observation example is constructive, not a multidimensional selection law. The next result extends the *scalar* comparison to any sample size and finite checkpoint bank. It makes the dependence on the bank explicit without claiming that curvature must help.

Let the reward be one, $J \sim \mathrm{Bin}(n, p)$ with $n \geq 2$, and $0 < \gamma < 1$. At count j, put $z_j = \gamma j/n$, $w_j = \gamma^2(j/n)(1 - j/n)/(n-1)$, $h_j = (1 - z_j)^{-1}$, and $\beta_j = w_j h_j^3$. A checkpoint k predicts

$$f_k(z) = \sum_{t=0}^{k-1} z^t, \qquad f_0 = 0, \qquad f_\infty = h.$$

The finite ordered bank $\mathcal{B} \subset \{0, 1, \ldots, \infty\}$ is fixed before sampling. Write a_j for the no-curvature selected prediction and b_j for the full-score selected prediction, using the same earliest-index tie convention. Both scores include the shared covariance correction. Thus the only difference is the term $2(f_k - h_j)\beta_j$.

Theorem 8 (Arbitrary-sample scalar bank characterization)**.** *For every count, $b_j \leq a_j$. With $d_j = a_j - b_j \geq 0$ and $v(p) = (1 - \gamma p)^{-1}$,*

$$R_0 - R_2 = \sum_{j=0}^{n} \binom{n}{j} p^j (1-p)^{n-j} d_j \{a_j + b_j - 2v(p)\}, \tag{31}$$

$$= \frac{H_{n,\mathcal{B}}(p)}{1 - \gamma p},$$

$$H_{n,\mathcal{B}}(p) = \sum_{j=0}^{n} \binom{n}{j} p^j (1-p)^{n-j} d_j \{(1 - \gamma p)(a_j + b_j) - 2\}. \tag{32}$$

Here $H_{n,\mathcal{B}}$ is a polynomial of degree at most $n+1$. It may be identically zero. If $W(p) = \sum_j \mathbb{P}_p(J = j) d_j > 0$, curvature helps precisely when

$$v(p) < \frac{\sum_j \mathbb{P}_p(J = j) d_j (a_j + b_j)}{2W(p)}.$$

Consequently it helps throughout any range where the true value is below every switched pair's midpoint, and harms throughout any range where it is above every such midpoint.

Proof. For a fixed count, let s_k be the no-curvature score. The full score is $s_k + 2\beta_j f_k$ plus a candidate-independent constant. Let i_0, i_2 be the two minimizers. Their optimality gives

$$s_{i_0} \leq s_{i_2}, \qquad s_{i_2} + 2\beta_j f_{i_2} \leq s_{i_0} + 2\beta_j f_{i_0}.$$

If $\beta_j > 0$, subtracting proves $f_{i_2} \leq f_{i_0}$. If $\beta_j = 0$, the rules have identical scores and tie convention. Taking the expectation of $(a_j - v)^2 - (b_j - v)^2 = d_j(a_j + b_j - 2v)$ proves (31). Multiplication by $1 - \gamma p > 0$ proves (32); all a_j, b_j depend on $n, \gamma, \mathcal{B}, j$, not on p. The remaining conclusions follow by dividing by $2W(p) > 0$. A positive constant reward multiplies every risk difference by its square and leaves score rankings unchanged. □

This result characterizes a regime through countwise decisions rather than fitting a power law. Increasing n changes both those decisions and their probabilities. Enlarging the bank changes its switched pairs. Neither operation has a universal sign. For rational γ, the countwise decisions are computable in exact rational arithmetic. General root isolation is not needed to evaluate the risk on a specified parameter grid; our reported expectations use stable binomial summation after exact decisions.

Corollary 9 (A fixed bank eventually favors its endpoint)**.** *Suppose $p > 0$, $\infty \in \mathcal{B}$, and $K_{\max}$ is the largest finite checkpoint. Set*

$$D_* = \frac{(\gamma p/2)^{2K_{\max}}}{(1 - \gamma p/2)^2}.$$

If $n - 1 > \gamma^2/\{D_*(1-\gamma)^4\}$, both rules select the endpoint whenever $J/n \geq p/2$, and

$$|R_0 - R_2| \leq \frac{\exp(-np^2/2)}{(1-\gamma)^2}.$$

A bank containing only the endpoint has identically zero difference.

Proof. For $q \geq p/2$ and any finite bank member, $(h - f_k)^2 \geq D_*$. Also $0 \leq f_k' \leq h'$, $h - f_k \leq h$, $w \leq \gamma^2/[4(n-1)]$, and $h \leq (1-\gamma)^{-1}$. The full-score difference against the endpoint is bounded below by

$$D_* - 2wh^4 - 2wh^4 \geq D_* - \frac{\gamma^2}{(n-1)(1-\gamma)^4} > 0.$$

The no-curvature difference has the same lower bound. Hoeffding's scalar lower-tail inequality bounds $\mathbb{P}(J/n < p/2)$ by $\exp(-np^2/2)$. Every selected prediction and v lies in $[0, (1-\gamma)^{-1}]$, so their squared-loss difference is bounded by $(1-\gamma)^{-2}$ in absolute value. Multiply by the disagreement-event bound. □

The explicit threshold can be enormous for a long bank or large discount. This is an asymptotic separation statement, not a claim that a practically relevant sample size has been reached. It explains why a persistent finite-bank curvature advantage cannot be inferred from its order-$1/n$ expansion alone. Growing banks and local parameter sequences are outside this corollary.

What extends to several dimensions? In any dimension, with the fitted bank held fixed, the score difference is a linear tilt $2f_i^\top M\widehat{\beta}_h$ plus a constant. If g is the gap from the no-curvature minimizer to its runner-up and

$$g > \max_i 2f_i^\top M\widehat{\beta}_h - \min_i 2f_i^\top M\widehat{\beta}_h,$$

that minimizer cannot change. This follows directly by bounding each change in a pairwise score difference. It is an observable *no-switch* diagnostic, not a certificate of lower true risk. For independent scalar coordinates, a diagonal evaluation metric and a Cartesian-product bank make losses and scores additive, so (31) adds coordinatewise. Coupled feature geometries do not reduce to this argument. We therefore do not promote the scalar theorem to a general multidimensional prediction of benefit.

G Population gates, pilot information, and observable certificates

A numerical threshold and a population assumption answer different questions. Let $C = A - F$, $a_* = \lambda_{\min}(A)$, and $c_* = \sigma_{\min}(C)$. Empirical gates accept when their estimated counterparts exceed a_0, c_0. The local theorem requires *strictly positive population margins* $a_* - a_0$ and $c_* - c_0$, in addition to local smoothness and inactive clipping.

Proposition 10 (Persistent fallback is a different estimator)**.** *For iid integrable moments, fixed thresholds, and bounded outputs, a strict population gate failure ($a_* < a_0$ or $c_* < c_0$) implies that the fallback probability tends to one. With zero fallback, risk tends to $\|v\|_M^2$. Strict population acceptance implies the gate-failure probability tends to zero. Equality at a threshold has no such general conclusion.*

Proof. The sample means converge almost surely. Continuity of the smallest eigenvalue of a symmetric matrix and the smallest singular value gives eventual failure or acceptance under a strict margin. On eventual failure the estimate is zero. Bounded output and target norms allow dominated convergence for the risk. This statement does not remove possible persistent clipping on accepted datasets. □

For the cyclic family at $d = 8, \gamma = .95$, $a_* = 1/8$ and $c_* = (1 - .9\gamma)/8 = 29/1600$. Therefore .02 is population-rejected, while the primary $3/200$ has positive margin $1/320$. Both signal settings are affected by .02: two of sixteen configurations at each sample size. Its limiting aggregate fallback fraction is $2/16 = 12.5\%$, not zero.

G.1 What information can set thresholds?

Known structural lower bounds. Fix and document the feature scale first: absolute thresholds change under reparameterization. Given valid lower bounds $a_L \leq a_*$ and $c_L \leq c_*$ with $a_L, c_L > 0$, one admissible choice is $a_0 = a_L/2, c_0 = c_L/2$. These margins may be known from a sampling

design. In on-policy evaluation with a stationary current/next feature distribution, Cauchy–Schwarz gives

$$x^\top F x \le \gamma x^\top A x, \qquad \sigma_{\min}(A - F) \ge (1-\gamma)\lambda_{\min}(A).$$

The second inequality follows from $\|Cx\|\,\|x\| \ge x^\top C x$. The same argument applies to a dominated next-feature second moment. It does *not* follow from discount alone in arbitrary off-policy data. The iid synthetic gates were set using known design margins, not claimed to be a default for unknown environments.

An independent pilot can provide conservative lower bounds. Suppose $\|\phi\|, \|\phi'\| \le L$ and $|r| \le R_{\max}$ are declared support bounds. For m independent pilot observations define

$$t_m = \frac{1 + \sqrt{2\log(3/\delta)}}{\sqrt{m}}, \qquad e_A = L^2 t_m, \quad e_C = (1+\gamma)L^2 t_m, \quad e_g = L R_{\max} t_m.$$

Then simultaneously, with probability at least $1-\delta$, $\left\|\widehat{A} - A\right\|_F \le e_A$, $\left\|\widehat{C} - C\right\|_F \le e_C$, and $\|\widehat{g} - g\| \le e_g$. Hence

$$a_L = \lambda_{\min}(\widehat{A}) - e_A, \qquad c_L = \sigma_{\min}(\widehat{C}) - e_C$$

are lower confidence bounds. Positive values can set half-margin thresholds for an *independent fitting dataset*; conditional on a successful pilot the fixed-threshold local argument applies. Failure to obtain positive bounds means *uncertified*, not singular. Pilot cost and its failure probability must be added, and any output bound must be capped by a prespecified finite ceiling to control pilot failure. Reusing the same sample to select thresholds does not automatically inherit the fixed-threshold theorem.

For completeness, the displayed concentration bound follows without estimating a variance. If $\|X\|_F \le K$, then $\mathbb{E}\left\|\bar{X} - \mathbb{E}X\right\|_F \le K/\sqrt{m}$ by the second moment. Replacing one sample changes this norm by at most $2K/m$. The bounded-differences inequality therefore yields radius $K[1 + \sqrt{2\log(1/\epsilon)}]/\sqrt{m}$ at failure probability ϵ. Apply it to the three moments and use the union bound. Sharper matrix concentration can be substituted when its variance inputs are available; the simple bounds are deliberately conservative, not tuned to these results.

What should the output radius mean? If the target is realizable, rewards are bounded, and $A \succeq a_L I$, then

$$\|v\| \le \frac{R_{\max}}{(1-\gamma)\sqrt{a_L}}.$$

Indeed $v^\top A v = \mathbb{E}[V^\pi(S)^2] \le [R_{\max}/(1-\gamma)]^2$. This bounds the *target*, not every estimated checkpoint. The calibration proof additionally needs clipping to be inactive on the good neighborhood for every candidate. One must check a local candidate envelope (using inverse margins and the fixed bank), or choose T larger than that envelope. A smaller safety cap can be sensible but changes the estimator; unmodified smooth-score calibration then does not apply at clipping points. In the experiments $T = 10$ was a declared numerical cap, never a reward-derived deployment guarantee; clipping events are reported. A tighter radius cannot be justified merely because it lowers simulator risk.

G.2 A CONSERVATIVE OBSERVABLE COMPARISON CERTIFICATE

Even without a full score-calibration certificate, support bounds can sometimes bound actual value error. On the simultaneous moment event above, let $\widehat{c} = \sigma_{\min}(\widehat{C}) > e_C$ and $\widehat{h} = \widehat{C}^{-1}\widehat{g}$. Then

$$\left\|\widehat{h} - v\right\| \le r_v := \frac{e_g + e_C\left\|\widehat{h}\right\|}{\widehat{c} - e_C}. \tag{33}$$

This follows from $C(\widehat{h} - v) = (C - \widehat{C})\widehat{h} + (\widehat{g} - g)$ and $\sigma_{\min}(C) \ge \widehat{c} - e_C$. For any two data-dependent candidates a, b, including clipped outputs,

$$\|a - v\|_M^2 - \|b - v\|_M^2 \in D_h \pm 2\,\|M(a-b)\|\, r_v, \quad D_h = \left\|a - \widehat{h}\right\|_M^2 - \left\|b - \widehat{h}\right\|_M^2. \tag{34}$$

The difference of squared losses proves this interval. Because one common confidence ball contains v, it holds simultaneously for the whole bank without assuming independent candidates or adding a union bound over them. A positive lower endpoint favors b; an interval containing zero does not resolve the comparison. This is a *realized-loss comparison* certificate, not a certificate that the same-data risk score estimates its expectation accurately.

Using the declared reward bound $R_{\max} = 2$ and known feature maximum in the iid suite, these conservative 95% moment bounds certify both inverses on *zero of 7,200 datasets*. Empirical conditioning passes on 6,309. We report both outcomes: an observed acceptable inverse margin is much weaker than a population confidence statement. No improved empirical method is claimed from a certificate that is vacuous here.

H ENDPOINT-SCALE ORACLE BOUND: FULL PROOF

This appendix proves the endpoint-scale selection bound directly. The known-operator identities, full empirical-moment remainder, fixed-score approximation, and the separate Gaussian confidence bound are proved in subsequent appendices. The distinctions between these statements are important: an expected oracle inequality, a calibrated expected score, and a high-probability certificate are not interchangeable.

H.1 COMPLETING THE SQUARE AND COMPARING CANDIDATES

Put $a_K = f_K(\bar{Z})$, $h = h(\bar{Z})$, $\beta = \widehat{\beta}$, and $t = h - \beta$. Expansion gives

$$\|a_K - h\|_M^2 + 2(a_K - h)^\top M\beta = \|a_K - t\|_M^2 - \|\beta\|_M^2 ,$$
$$\widehat{R}_K^{(2)} = \|a_K - t\|_M^2 + p_K - \operatorname{tr}(MV_\infty) - \|\beta\|_M^2 .$$

The last two terms do not depend on K, so the minimizers agree even if penalties or scores are negative.

Proof of Lemma 1. Fix any comparator j. The minimizing property and expansion around v give

$$\ell_{\widehat{K}} \le \ell_j + 2(a_{\widehat{K}} - a_j)^\top M(t - v) + p_j - p_{\widehat{K}}, \qquad \ell_k = \|a_k - v\|_M^2 .$$

Cauchy–Schwarz in the M seminorm yields

$$\ell_{\widehat{K}} \le \ell_j + 2(\sqrt{\ell_{\widehat{K}}} + \sqrt{\ell_j})e + \operatorname{osc}(p), \quad e = \|t - v\|_M .$$

Apply $2xy \le \alpha x^2 + y^2/\alpha$ twice, with $\alpha = \eta/(2+\eta) \in (0,1)$. Therefore

$$(1-\alpha)\ell_{\widehat{K}} \le (1+\alpha)\ell_j + 2e^2/\alpha + \operatorname{osc}(p).$$

Division by $1-\alpha$ gives the constants in (16); then minimize over j. The argument is deterministic. It uses neither independent candidates nor positive penalties, and remains valid for singular $M \succeq 0$. □

H.2 MOMENT BOUNDS USED IN THE PROOF

Let $X_i = Z_i - \mu$ and $e_n = n^{-1}\sum_i X_i$. The observations are centered, independent, and bounded by R. Thus

$$\mathbb{E}\,\|e_n\|^2 = \operatorname{tr}\Gamma/n, \qquad \mathbb{E}\,\|e_n\|^4 \le 3R^4/n^2, \qquad \mathbb{E}\widehat{\Gamma} = \Gamma, \qquad \operatorname{tr}\widehat{\Gamma} \le 2R^2. \tag{35}$$

For completeness, independence and centering give

$$\mathbb{E}\,\|\textstyle\sum_i X_i\|^4 = n\mathbb{E}\,\|X_1\|^4 + n(n-1)\big[(\operatorname{tr}\Gamma)^2 + 2\operatorname{tr}(\Gamma^2)\big] \le 3n^2R^4.$$

Also $(n-1)\widehat{\Gamma} = \sum_i X_iX_i^\top - ne_ne_n^\top$, so its trace is at most nR^2; $n/(n-1) \le 2$. These statements do not require Gaussian data or independence between mean and covariance.

H.3 Endpoint error retains the joint covariance

Write $J = Dh(\mu)$. Taylor's theorem along the segment from μ to $\bar{Z}$ gives

$$h(\bar{Z}) - h(\mu) = Je_n + r_h, \qquad \|r_h\| \le L_2 \|e_n\|^2 /2.$$

Since $\widehat{\Gamma}$ is positive semidefinite, its spectral decomposition and the multilinear derivative norm imply

$$\left\|\widehat{\beta}\right\| = \left\|D^2 h(\bar{Z}) : \widehat{\Gamma}/(2n)\right\| \le L_2 \operatorname{tr}\widehat{\Gamma}/(2n) \le L_2 R^2/n.$$

Minkowski's inequality in the space of square-integrable random vectors, after applying $M^{1/2}$, now yields

$$\begin{aligned}\left(\mathbb{E}\,\|t - h(\mu)\|_M^2\right)^{1/2} &\le \left(\mathbb{E}\,\|Je_n\|_M^2\right)^{1/2} + \left(\mathbb{E}\,\|r_h\|_M^2\right)^{1/2} + \left(\mathbb{E}\left\|\widehat{\beta}\right\|_M^2\right)^{1/2} \\ &\le \sqrt{\mathcal{V}/n} + \frac{\sqrt{3m_0}L_2R^2}{2n} + \frac{\sqrt{m_0}L_2R^2}{n}.\end{aligned}$$

This proves $\mathbb{E}\,\|t - v\|_M^2 \le Q_n$. In particular, the leading quantity is $\mathcal{V} = \operatorname{tr}(MJ\Gamma J^\top)$, not a sum of separately propagated marginal variances. Cancellations within $J\Gamma J^\top$ survive.

H.4 Penalty variation without a union bound

For every candidate,

$$\begin{aligned}|p_K| &= \frac{2}{n}\left|\operatorname{tr}\left(MDf_K(\bar{Z})\widehat{\Gamma}Dh(\bar{Z})^\top\right)\right| \\ &\le \frac{2\sqrt{m_0}L_1}{n}\sqrt{\operatorname{tr}\widehat{\Gamma}}\,\left\|M^{1/2}Dh(\bar{Z})\widehat{\Gamma}^{1/2}\right\|_F.\end{aligned}$$

Use $\left\|Dh(\bar{Z}) - J\right\|_{\mathrm{op}} \le L_2\|e_n\|$, then $\operatorname{osc}(p) \le 2\max_K |p_K|$. Taking expectations gives

$$\begin{aligned}\mathbb{E}\operatorname{osc}(p) &\le \frac{4\sqrt{m_0}L_1}{n}\left[\mathbb{E}\left\{\sqrt{\operatorname{tr}\widehat{\Gamma}}\,\left\|M^{1/2}J\widehat{\Gamma}^{1/2}\right\|_F\right\} + \sqrt{m_0}L_2\mathbb{E}\{\|e_n\|\operatorname{tr}\widehat{\Gamma}\}\right] \\ &\le \frac{4\sqrt{m_0}L_1\sqrt{G\mathcal{V}}}{n} + \frac{8m_0L_1L_2R^3}{n^{3/2}} = P_n.\end{aligned}$$

The first expectation is at most $\sqrt{G\mathcal{V}}$ by Cauchy–Schwarz and $\mathbb{E}\widehat{\Gamma} = \Gamma$. The second is at most $2R^2\sqrt{\mathbb{E}\,\|e_n\|^2} \le 2R^3/\sqrt{n}$. The estimates hold simultaneously for all candidates through the common derivative bound; no union bound over K is used.

Proof of Theorem 2. Apply Lemma 1 to the completed-square score. Take expectations, use the endpoint and penalty estimates just proved, and use $\mathbb{E}\min_K \ell_K \le \min_K \mathbb{E}\ell_K = \min_K R_K$. This gives (20). □

A derivative-only corollary and what the rate does not say. Replacing $\mathcal{V} \le m_0L_1^2R^2$ gives the simpler endpoint bound

$$\mathbb{E}\,\|t - v\|_M^2 \le m_0(L_1R/\sqrt{n} + L_2R^2/n)^2, \qquad \mathbb{E}\operatorname{osc}(p) \le 4m_0L_1^2R^2/n,$$

where the first display follows directly from the Lipschitz bound on h, without the Taylor refinement. Either route gives an order-$1/n$ remainder. The geometry-aware version is useful when joint covariance makes $\mathcal{V}$ small. If $\mathcal{V} = 0$, its displayed remainder is $O(n^{-3/2}) + O(n^{-2})$; this is not a claim that every such instance has that exact risk. In the cancellation construction below the endpoint error and penalties are exactly zero.

The theorem does not estimate R, L_1, L_2, Γ, or $\mathcal{V}$ from the data and then certify their global validity. They are analysis quantities. It is therefore inappropriate to describe a numerical evaluation of the theorem using the true target as an offline confidence certificate. A high-probability bound for the distinct known-operator Gaussian rule is stated separately in Theorem 16.

H.5 A MATCHING SCALAR ORDER BY A CONSTRAINED-RISK ARGUMENT

Proof of Proposition 3. Let P_0 be the law of n independent fair signs and P_1 the law with mean $\mu_1 = 2/\sqrt{n}$. For $n \geq 16$, both are valid reward laws. Include any estimator randomization as an independent variable with the same law under both models. The likelihood ratio is

$$L = \prod_{i=1}^{n}(1 + \mu_1 X_i), \qquad \mathbb{E}_0 L^2 = (1 + \mu_1^2)^n = (1 + 4/n)^n \leq e^4.$$

At $\mu = 0$, the zero candidate has risk zero. If $\mathbb{E}_0 T^2 \geq c\tau^2$, the result follows immediately. Otherwise,

$$|\mathbb{E}_1 T| = |\mathbb{E}_0 LT| \leq (\mathbb{E}_0 L^2 \mathbb{E}_0 T^2)^{1/2} < e^2\sqrt{c}\,\tau = \tau/4.$$

The alternative target is $v_1 = 2\tau$. Jensen's inequality implies $\mathbb{E}_1(T - v_1)^2 \geq (7\tau/4)^2 = 49\tau^2/16$. The empirical endpoint has risk $(1 - \mu_1^2)\tau^2 \leq \tau^2$, so subtracting $(1 + \eta)$ times the best comparator risk leaves at least $17\tau^2/16 > c\tau^2$. Infinite-risk cases make the inequality immediate. This is a specialization of constrained-risk reasoning, not a new general statistical inequality (Brown & Low, 1996; Duchi & Ruan, 2021). □

For the known scalar operator, $h(\mu) = \mu/(1 - \rho)$, $L_2 = 0$, $\mathcal{V} = \mathrm{Var}(X)/(1 - \rho)^2$, and the uniform first derivative of all checkpoints is at most $(1 - \rho)^{-1}$. The upper bound therefore has the same $n^{-1}(1 - \rho)^{-2}$ order as the lower bound. Constants and the full multidimensional geometry are not matched by this construction.

I AN EXACT RANDOM-GRAM CANCELLATION AND ITS FAILURE UNDER ABLATION

Proposition 11 (Spurious uncertainty from freezing the Gram)**.** *In the two-state construction* (9)*, sample current states independently with second-state probability $p \in (0, 1)$ and use the bank $\{0, \infty\}$, scalar prediction norm $M = 1$, and $v \neq 0$. Let*

$$q = \mathrm{Bin}(n, p)/n, \quad D = 1 + (3 - 4\gamma)q, \quad W = q(1 - q)/(n - 1).$$

The full joint score always selects the exact endpoint. The derivative ablation that holds $\widehat{A}$ fixed, while retaining the actual fitted parameters, selects zero whenever

$$W > \frac{D^2}{18 + 24\gamma}. \tag{36}$$

For $n = 8$, $p = 1/2$, and $\gamma = .9$, this event has probability $246/256$.

Proof. The two per-observation moment triples are $(1, 0, v)$ and $(4, 4\gamma, 4(1 - \gamma)v)$. Their average has the form (9). Since $D > 0$ for $\gamma < 1$ and $q \in [0, 1]$, the endpoint is exactly v throughout the support hull. Its derivative and every higher derivative along this one-dimensional observed direction are zero. Thus the sample covariance contractions in the full score vanish. The two scores are v^2 and zero.

For the ablation, perturb F and g along their observed direction while holding the fitted A fixed. Then

$$h'_{\mathrm{frozen}} = \frac{(3 - 4\gamma)v + 4\gamma v}{D} = \frac{3v}{D}, \qquad \frac{1}{2}h''_{\mathrm{frozen}} = \frac{12\gamma v}{D^2}.$$

The empirical mean covariance of the Bernoulli direction is W. Consequently,

$$\widehat{V}_{\infty,\mathrm{frozen}} = 9v^2 W/D^2, \qquad \widehat{\beta}_{\mathrm{frozen}} = 12\gamma v W/D^2.$$

The scores at zero and the endpoint are respectively

$$s_0 = v^2 - (9 + 24\gamma)v^2 W/D^2, \qquad s_\infty = 9v^2 W/D^2.$$

Comparing these gives (36). With the stated parameters it holds exactly for binomial counts $2, 3, 4, 5, 6, 7$, whose coefficients sum to $28 + 56 + 70 + 56 + 28 + 8 = 246$. □

This example is linearly realizable: the true state values are $(v, 2v)$, rewards are deterministic, and the empirical scalar Bellman operator is $4\gamma q/(1 + 3q) \leq \gamma$. It is not caused by divergence, omitted states causing a singular Gram, or inaccurate transition observations. The only error is treating correlated components of the empirical fit as if one component had no uncertainty.

Exact finite-checkpoint comparison. For this same process,

$$B(q) = \frac{4\gamma q}{1+3q}, \qquad f_K(q) = v(1-B(q)^K), \qquad B'(p) = \frac{4\gamma}{(1+3p)^2}, \quad B''(p) = -\frac{24\gamma}{(1+3p)^3}.$$

Exact risk is computed by enumerating $q = j/n$ with binomial probabilities. The variance-only expansion of risk is $[f_K(p) - v]^2 + f_K'(p)^2 p(1-p)/n$; the second-order expansion adds $[f_K(p) - v] f_K''(p) p(1-p)/n$. These expressions describe a random-Gram fitted map. The expansion need not improve every small-sample case.

J Full raw-moment implementation and domain validity

The calculation uses iid blocks (A_i, F_i, g_i) and their means $(\widehat{A}, \widehat{F}, \widehat{g})$. Define scaled centered directions

$$(a_i, f_i, g_i^\circ) = (A_i - \widehat{A}, F_i - \widehat{F}, g_i - \widehat{g})/\sqrt{n(n-1)}.$$

Their outer products sum to the unbiased covariance of the sample mean. This factorization retains arbitrary within-block covariance without constructing a matrix of size $(2d^2+d)^2$.

Let $H = (\widehat{A} - \widehat{F})^{-1}$, $h = H\widehat{g}$, and $E_i = f_i - a_i$. The endpoint first-order direction and mean-curvature estimate are

$$U_{\infty,i} = H(g_i^\circ + E_i h), \qquad \widehat{\beta} = H \sum_i E_i U_{\infty,i}, \qquad V_\infty = \sum_i U_{\infty,i} U_{\infty,i}^\top. \tag{37}$$

To verify the curvature formula, differentiate $(A-F)h = g$ twice in one direction. The second derivative is $2HE_iU_{\infty,i}$; summing half of these derivatives gives (37).

Starting from $x_0 = 0$ and $U_{0,i} = 0$, propagate

$$x_{k+1} = \widehat{B} x_k + \widehat{b}, \tag{38}$$

$$U_{k+1,i} = \widehat{B} U_{k,i} + \widehat{A}^{-1}\big(g_i^\circ + f_i x_k - a_i x_{k+1}\big), \tag{39}$$

$$V_{K\infty} = \sum_i U_{K,i} U_{\infty,i}^\top. \tag{40}$$

Differentiating $Ax_{k+1} = Fx_k + g$ proves the recursion. It is essential that the Gram term uses x_{k+1}, not x_k. The factorized and explicit-covariance implementations agree algebraically. Tests compare them and check both first and second derivatives by finite differences, including a random non-diagonal Gram.

The stored input and centered matrix directions require $O(nd^2)$ memory; current first-order directions require $O(nd)$. Dense propagation costs $O(K_{\max} nd^2 + d^3)$ arithmetic after the moments are formed. The reference implementation additionally retains $O(|\mathcal{K}|nd)$ directions for diagnostics; streaming just the needed scores would remove this extra storage. This is not a neural-scale or sublinear-time algorithm. Ridge and the converged direct solve are essential inexpensive comparators at these dimensions.

J.1 Why the random-Gram experiments satisfy the support assumption

Each compressed construction has d groups, two physical states per group, and a fixed orthogonal matrix Q. Features are group indicators transformed by Q. The population process is lumpable: both physical states in a group have the same distribution over next groups and the same group value. The true value therefore lies in the feature span.

For every empirical block and every convex combination of blocks, the Gram has the form $Q^\top C Q$, where C is diagonal with strictly positive group masses. The cross moment has the form $Q^\top C\gamma T Q$, where T is nonnegative and substochastic. Therefore

$$\widehat{B} = Q^\top \gamma T Q, \qquad \rho(\widehat{B}) \le \gamma < 1.$$

In the weighted construction, all regression weights lie in $\{.5, 1.5\}$ and each group is sampled. In the unweighted sampling construction, each block includes one mandatory sample per group and d

additional samples drawn uniformly over groups, giving group masses at least $1/(2d)$. Both yield a deterministic positive lower bound on C over the support hull. Rewards and all raw entries are bounded. The hull is compact; inverse determinants do not vanish there; continuity then gives an open neighborhood with bounded inverse derivatives of any fixed finite order. The endpoint and all retained checkpoint maps consequently satisfy the required regularity.

This proves validity for these particular experiments. It is not a general statement that random offline features produce stable fitted iteration. The arbitrary-feature off-policy stability problem is deliberately left outside the scope.

K Setting and exact persistent-error geometry

We fix a target policy, a feature map $\phi \in \mathbb{R}^d$, and an evaluation matrix $M \succeq 0$. Define $\|x\|_M^2 = x^\top M x$. For linear fitted policy evaluation the population normal equations are

$$A\theta_{k+1} = F\theta_k + g, \qquad B = A^{-1}F, \quad b = A^{-1}g, \tag{41}$$

where $A = \mathbb{E}[\phi\phi^\top]$, $F = \gamma\mathbb{E}[\phi\phi'^\top]$, and $g = \mathbb{E}[\phi r]$. The next feature may include the target-policy action expectation. We assume A invertible and $\rho(B) < 1$ when discussing convergence. The target is

$$v = (I - B)^{-1}b, \qquad L = (I - B)^{-1}. \tag{42}$$

Under realizability, v represents the target policy value. Without realizability our formulas concern the projected fixed point, not automatically the true value function. The empirical iteration reuses $(\widehat{A}, \widehat{F}, \widehat{g})$ at every update. Unless stated otherwise initialization is zero. A checkpoint $K = \infty$ means the exact empirical fixed-point solve.

K.1 Known operator

Suppose B is known and $y = b + \varepsilon$ with $\mathbb{E}\varepsilon = 0$, $\mathrm{Cov}(\varepsilon) = \Omega$. Let $S_K = \sum_{j=0}^{K-1} B^j$, $S_0 = 0$, and $S_\infty = L$. Write $R_K = \mathbb{E}\left\|\widehat{\theta}_K - v\right\|_M^2$.

Proposition 12 (Exact risk and coordinate invariance)**.** *For a deterministic initialization independent of the data,*

$$R_K = \left\|B^K(\theta_0 - v)\right\|_M^2 + \mathrm{tr}(MS_K\Omega S_K^\top). \tag{43}$$

At $K = \infty$ the bias is zero. Under any invertible coordinate change $\widetilde{\theta} = T\theta$, the same risk is obtained using $\widetilde{B} = TBT^{-1}$, $\widetilde{\Omega} = T\Omega T^\top$, and $\widetilde{M} = T^{-\top}MT^{-1}$.

The amplification profile is $\mathcal{A}_K = \mathrm{tr}(MS_K\Omega S_K^\top)$. It depends on the alignment of propagation, observation uncertainty, and the evaluation norm. It cannot in general be recovered from the eigenvalues of B alone.

For scalar $B = \rho \in (0, 1)$ and $\Omega = \sigma^2/n$, the variance is $\sigma^2(1 - \rho^K)^2/[n(1 - \rho)^2]$. Before the relaxation time it grows approximately as K^2/n, not as K/n. This reflects repeated accumulation of the same error. It does *not* establish that data reuse is wasteful.

Proposition 13 (Equal-observation-budget comparison)**.** *In the known-operator common-reward-mean model, let N independent reward vectors have individual covariance Σ. Set $A_j = M^{1/2}B^j\Sigma^{1/2}$. Using their common mean at all K updates has variance $\left\|\sum_{j<K} A_j\right\|_F^2/N$. Using disjoint fresh estimates with continuous sample allocations $n_j > 0$, $\sum_j n_j = N$, has minimum variance*

$$\frac{(\sum_{j<K}\|A_j\|_F)^2}{N} \;\geq\; \frac{\left\|\sum_{j<K} A_j\right\|_F^2}{N}. \tag{44}$$

Zero-weight updates can be omitted. Integer allocations cannot improve on the continuous minimum.

Thus the relevant motivation for stopping is regularization of a finite-data estimator, not a blanket argument for collecting equally accurate fresh estimates every round at no cost.

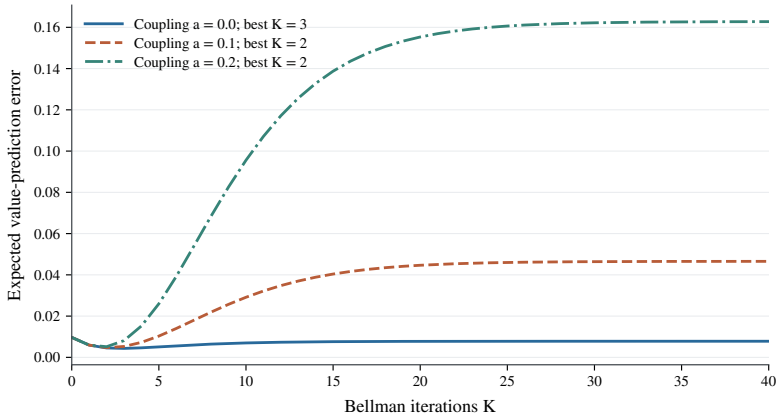


Figure 3: Exact risk in a matched-spectrum funnel. Coupling changes the statistically preferred checkpoint and endpoint risk although the spectrum, target, reward means, and reward-noise covariance are held fixed. Here $m = 31$, $\rho = .75$, $\gamma = .95$, $c = .1$, and $\tau^2 = 1/64$. This is an analytic construction, not a neural-training experiment.

K.2 A REALIZABLE MATCHED-SPECTRUM FUNNEL

There are m upstream states, a downstream state, and a zero-valued terminal state. Let $u = (1, \dots, 1, 0)^\top$, let e select the downstream state, and set

$$B_a = \rho I + aue^\top, \quad 0 < \rho, \quad 0 \le a, \quad \rho + a \le \gamma < 1. \tag{45}$$

The substochastic transition matrix is B_a/γ; missing mass terminates. Choose $v = cu$, $b = (1-\rho)cu$, $M = I/(m+1)$, and $\Omega = \tau^2 ee^\top$. The same reward means, true values, evaluation matrix, and reward covariance apply to all a.

Proposition 14 (Exact funnel profile)**.** *All eigenvalues of B_a equal ρ. With $s_K = \sum_{j<K} \rho^j$ and $t_K = \sum_{j=1}^{K-1} j\rho^{j-1}$,*

$$R_K(a) = \frac{mc^2}{m+1}\rho^{2K} + \frac{\tau^2}{m+1}\left(s_K^2 + ma^2 t_K^2\right). \tag{46}$$

For $K \ge 1$, the variance ratio to $a = 0$ is $1 + ma^2(t_K/s_K)^2$; at convergence it is $1 + ma^2/(1-\rho)^2$.

Appendix N.4 also gives a two-checkpoint separation for pre-data schedules that see only the common summary. This is a restricted scheduling lower bound, not a lower bound for arbitrary data-adaptive selectors. The construction does not contradict sup-norm Bellman contraction, and it does not separate our profile from the richer Lyapunov analysis of Perdomo et al. (2023).

L ESTIMATED OPERATORS: VARIANCE IS NOT THE WHOLE CORRECTION

Write $(\widehat{A}, \widehat{F}, \widehat{g}) = (A + \Delta A, F + \Delta F, g + \Delta g)$, all from the same data. Define $U = A^{-1}\Delta A$, $V = A^{-1}\Delta F$, and $w = A^{-1}\Delta g$. The population path is $x_{k+1} = Bx_k + b$, $x_0 = 0$.

Theorem 15 (Joint empirical-moment expansion)**.** *Suppose A and $A + \Delta A$ are invertible. Define*

$$u_{k+1} = Bu_k + Vx_k + w - Ux_{k+1}, \tag{47}$$

$$q_{k+1} = Bq_k + Vu_k - Uu_{k+1}, \tag{48}$$

with $u_0 = q_0 = 0$. Then u and q are respectively homogeneous linear and quadratic functions of the joint moment perturbation. The exact remainder $r_k = \widehat{\theta}_k - x_k - u_k - q_k$ satisfies

$$(I+U)r_{k+1} = (B+V)r_k + Vq_k - Uq_{k+1}. \tag{49}$$

For an induced norm, let $G_H = \sum_{j<H} \|B^j\|$, $\widehat{G}_H = \sum_{j<H} \left\|\widehat{B}^j\right\|$, $X_H = \max_{k\le H} \|x_k\|$, $\eta = \|U\| + \|V\|$, $D = \|w\| + \eta X_H$, and $\alpha = \left\|(I+U)^{-1}\right\|$. Then

$$\max_{k\le H} \|r_k\| \le \alpha \widehat{G}_H G_H^2 \eta^2 D. \tag{50}$$

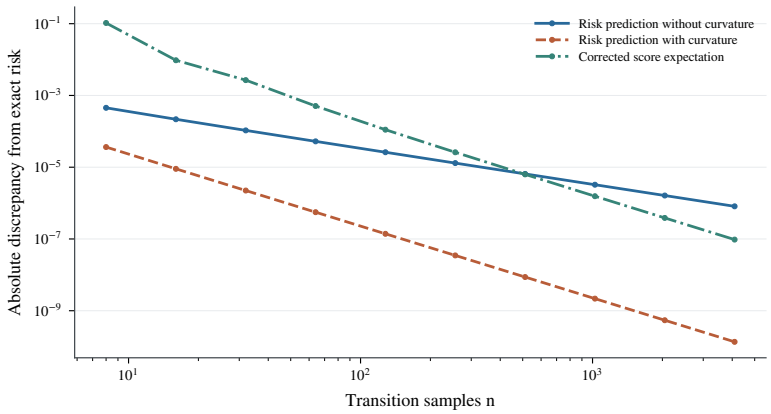


Figure 4: Exact binomial enumeration exposes the missing mean correction. The variance-only approximation has a leading order-$1/n$ error. The corrected approximation has order-$1/n^2$ error in this scalar polynomial case. The displayed case is $p = .7$, $\gamma = .95$, $r = .1$, and $K = 3$. No Monte Carlo estimation is used in this figure.

No independence between U, V, w is assumed.

For centered empirical moments, $\mathbb{E}u_K = 0$. Let $d_K = x_K - v$. Expanding squared prediction error gives

$$R_K = \|d_K\|_M^2 + \mathbb{E}\,\|u_K\|_M^2 + 2d_K^\top M \mathbb{E}q_K + \mathcal{R}_K. \tag{51}$$

The third term is a mean-curvature correction, not a variance term. If $u = O_{L^2}(n^{-1/2})$, $q = O_{L^2}(n^{-1})$, and $r = O_{L^2}(n^{-3/2})$ on a uniformly regular domain, then $\mathcal{R}_K = O(n^{-3/2})$ for bounded d_K. Appendix O gives an explicit bound. Stability and moment regularity must be justified, not inferred by conditioning on a favorable sample event while continuing to assume centered perturbations.

L.1 An exactly enumerable one-state example

Let a single nonterminal state self-transition with unknown probability p, otherwise terminate. The mean reward $r \neq 0$ is known. Put $\rho = \gamma p$, $\widehat{\rho} = \gamma\,\mathrm{Bin}(n, p)/n$, and $v = r/(1-\rho)$. At checkpoint three, $f(z) = r(1 + z + z^2)$. If $d = f(\rho) - v$ and $\mu_j = \mathbb{E}(\widehat{\rho} - \rho)^j$, then

$$R_3 = d^2 + \left[r^2(1+2\rho)^2 + 2dr\right]\mu_2 + 2r^2(1+2\rho)\mu_3 + r^2\mu_4. \tag{52}$$

Here $\mu_2 = \gamma^2 p(1-p)/n$, while $\mu_3, \mu_4 = O(n^{-2})$. Since $d \neq 0$ for $0 < p < 1$, omitting $2dr\mu_2$ leaves an order-$1/n$ error. Including it leaves an exact order-$1/n^2$ remainder in this polynomial example. The general theorem does not assert a universal n^{-2} remainder.

M Risk scores and what selection can guarantee

A good approximation to each fixed checkpoint's expected risk is not automatically a guarantee for the checkpoint selected by minimizing that approximation. SURE-tuning optimism and corrections to its analysis are documented by Tibshirani & Rosset (2019); Cauchois et al. (2021). We state separate fixed-score and adaptive-selection results.

M.1 Known operator: an exact score and simultaneous Gaussian bounds

For zero initialization define

$$\widehat{R}_K = \|S_K y - Ly\|_M^2 + 2\,\mathrm{tr}(MS_K\Omega L^\top) - \mathrm{tr}(ML\Omega L^\top). \tag{53}$$

Direct expansion gives $\mathbb{E}\widehat{R}_K = R_K$ for any centered error with covariance Ω. This is a linear-estimation/SURE specialization (Eldar, 2009), not a new unbiased-estimation principle. The same

formula works with any fixed affine candidate in place of $S_K y$, permitting SURE tuning of ridge candidates when the operator is known.

Theorem 16 (Known-operator Gaussian selection). *Let $\varepsilon \sim N(0, \Omega)$ and let the fixed finite checkpoint family have size m. For $\eta \in (0,1)$, $\delta \in (0,1)$, set $t = \log(8m/\delta)$ and*

$$H_C = \Omega^{1/2} C^\top M C \Omega^{1/2}, \tag{54}$$

$$p(C) = 2 \|H_C\|_F \sqrt{t} + 2t(1+\eta^{-1}) \|H_C\|_{\mathrm{op}}, \tag{55}$$

$$a_K = p(S_K - L), \qquad c_K = p(S_K), \tag{56}$$

$$U_K = \frac{1+\eta}{1-\eta}(\widehat{R}_K + a_K) + c_K. \tag{57}$$

With probability at least $1-\delta$, all realized losses $\ell_K = \left\|\widehat{\theta}_K - v\right\|_M^2$ satisfy $\ell_K \le U_K$. Moreover, $\widehat{K} \in \arg\min_K U_K$ satisfies

$$\ell_{\widehat{K}} \le \min_{K \in \mathcal{K}} \left\{ \frac{(1+\eta)^2}{1-\eta} R_K + \frac{2(1+\eta)}{1-\eta} a_K + c_K \right\}. \tag{58}$$

The rule uses known B, Ω, M and observed y, but not the unknown b.

These bounds are comparator-local but conservative. They do not say that the adaptively minimized score is unbiased. Nor do they justify replacing B and Ω by arbitrary estimates.

M.2 Why exact unbiasedness can fail with unknown transitions

Theorem 17 (Fixed-finite-sample unbiasedness obstruction). *Observe exactly $n < \infty$ independent Bernoulli(p) transitions from the one-state process above, with known $r \neq 0$ and $0 < \gamma < 1$. For any estimator $\widehat{v}$ depending only on these observations and independent auxiliary randomness, with finite conditional second moments, no integrable statistic has expectation*

$$\mathbb{E}_p\big(\widehat{v} - r/(1-\gamma p)\big)^2 \tag{59}$$

for every p in an open subinterval of $(0,1)$. Furthermore, no exactly unbiased statistic estimates the risk difference between checkpoints zero and one for all such p.

The obstruction is elementary but relevant: expectations of finite-Bernoulli-sample statistics are polynomials in p, whereas the risk contains a nonremovable resolvent pole. It is not a minimax lower bound on checkpoint selection; indeed, the particular two-checkpoint difference has an easy sign. It explains why an estimated-transition score must be qualified as an approximation unless stronger information or another observation model is available.

M.3 Curvature-corrected same-data score

Let $Z_1, \dots, Z_n$ be independent, identically distributed primitive-moment blocks, $\mu = \mathbb{E} Z_i$, $\Gamma = \mathrm{Cov}(Z_i)$, $\bar{Z} = n^{-1} \sum_i Z_i$, and

$$\widehat{\Gamma} = \frac{1}{n-1} \sum_i (Z_i - \bar{Z})(Z_i - \bar{Z})^\top. \tag{60}$$

A block may contain one sample from every independent state stratum, or jointly dependent observations such as a sampled transition permutation. Define the smooth population checkpoint map $f_K(\mu)$ and endpoint map $h(\mu)$; the target is $h(\mu)$. The primitive moments are (A, F, g) for random-Gram FQE, or (B, b) when the Gram is fixed. All derivatives below are evaluated at $\bar{Z}$. Set

$$J_K = Df_K(\bar{Z}), \quad J_\infty = Dh(\bar{Z}), \quad \widehat{\beta}_\infty = \frac{1}{2n} D^2 h(\bar{Z}) : \widehat{\Gamma}, \tag{61}$$

$$V_{K\infty} = J_K \widehat{\Gamma} J_\infty^\top / n, \qquad V_\infty = J_\infty \widehat{\Gamma} J_\infty^\top / n. \tag{62}$$

The Hessian contraction is componentwise. Our score is

$$\boxed{\begin{aligned}\widehat{R}_K^{(2)} = &\left\|f_K(\bar{Z}) - h(\bar{Z})\right\|_M^2 + 2\,\mathrm{tr}(MV_{K\infty}) - \mathrm{tr}(MV_\infty)\\ &+ 2\big(f_K(\bar{Z}) - h(\bar{Z})\big)^\top M\widehat{\beta}_\infty.\end{aligned}} \tag{63}$$

The covariance correction accounts for reuse of the same observations by the checkpoint and endpoint. The final term corrects the endpoint's second-order mean. It involves the endpoint curvature, not just the checkpoint curvature: the checkpoint mean corrections cancel when comparing gap risk to target risk. Appendix R proves this cancellation explicitly.

Assumption 2 (Bounded smooth moment domain)**.** $n \geq 2$, $\|Z_i - \mu\|_2 \leq R$ almost surely, and the convex hull of the observation support has a neighborhood on which all f_K, h are three times continuously differentiable. Uniformly over that hull and the fixed checkpoint family,

$$\begin{aligned}\|f_K\|, \|h\| \leq L_0, \qquad \|Df_K\|, \|Dh\| \leq L_1,\\ \left\|D^2 f_K\right\|, \left\|D^2 h\right\| \leq L_2, \quad \left\|D^3 f_K\right\|, \left\|D^3 h\right\| \leq L_3.\end{aligned}$$

Derivative norms are Euclidean multilinear operator norms. Inverses defining the maps must exist throughout this domain. Write $m_0 = \|M\|_{\mathrm{op}}$.

Theorem 18 (Explicit score approximation)**.** *Under Assumption 2, for each fixed K,*

$$\left|\mathbb{E}\widehat{R}_K^{(2)} - R_K\right| \leq \beta_n, \tag{64}$$

$$\beta_n = \frac{m_0 R^3}{n^{3/2}}\left[(2\sqrt{3}+4)L_0L_3 + (5\sqrt{3}+16)L_1L_2\right].$$

The bound is uniform over the fixed family and retains reward–transition and cross-coordinate covariance. It does not require the sample mean and sample covariance to be independent.

Corollary 19 (Conservative estimated-operator selection)**.** *Under Assumption 2, define*

$$\begin{aligned}A_0 &= 16m_0L_0L_1R, \qquad D_0 = 8m_0L_0L_1R,\\ B_0 &= 4m_0(8L_1L_2 + 2L_0L_3)R^3 + 12m_0(3L_1^2 + 2L_0L_2)R^2.\end{aligned}$$

For $m = |\mathcal{K}|$, $\widehat{K} \in \arg\min_K \widehat{R}_K^{(2)}$, and $t = \log(4m/\delta)$, with probability at least $1-\delta$,

$$\ell_{\widehat{K}} \leq \min_K R_K + \left[2(A_0 + B_0/n) + D_0\right]\sqrt{t/(2n)} + 2\beta_n. \tag{65}$$

This alternative high-probability guarantee is intentionally loose: its leading adaptive remainder is $O(\sqrt{\log(m)/n})$, not the order-$1/n$ expected-risk guarantee in Theorem 2. The two statements have different probability modes and leading constants. The score-bias rate in Theorem 18 must not be advertised as an equally fast selected-risk guarantee. Near an ill-conditioned resolvent or Gram matrix, the smoothness constants can be large or fail to exist.

M.4 Implementation and computational scope

For fixed-Gram moments, stack $z = (b_s, B_{s,:})_{s=1}^d$ and define the injection $T(x)\Delta z = \Delta b + \Delta B x$. Then

$$J_{k+1} = BJ_k + T(x_k), \quad J_0 = 0, \qquad J_\infty = LT(v), \qquad \beta_\infty = L\mathbb{E}[\Delta B(J_\infty \Delta z)]. \tag{66}$$

The expectation contracts the covariance of the empirical mean. We implement both full covariance and a faster independent-row version. For independent state strata, sample covariance is projected onto the known block-diagonal structure; the theorem and proof continue to hold under this positive trace-preserving projection. Dependence inside a block must not be discarded. Overlapping-feature comparisons use the full primitive covariance.

The algorithm fits the empirical moments, solves the empirical endpoint, propagates checkpoint values and sensitivities, and minimizes (63). No true value or online return enters this selection. The procedure computes the endpoint and candidate bank before selection; it does not establish computational savings from sequential stopping. Endpoint solves and covariance calculations add work. The empirical comparison includes the direct fixed-point solve precisely because, in small problems, it is cheap. Damping alone cannot change that solve's statistical error: whenever a damped update converges, its unique fixed point remains $(I - \widehat{B})^{-1}\widehat{b}$.

N PROOFS FOR THE KNOWN-OPERATOR MODEL

N.1 EXACT RISK AND INVARIANCE

Proof of Proposition 12. Subtract $v = Bv + b$ from the update. Iterating gives

$$\widehat{\theta}_K - v = B^K(\theta_0 - v) + S_K\varepsilon.$$

The first term is deterministic and the second has mean zero. Expanding the M-quadratic form gives (43). Stability makes $B^K \to 0$ and $S_K \to L$, proving the endpoint statement. Under the coordinate change, $\widetilde{S}_K = TS_KT^{-1}$. The bias quadratic form is unchanged, and cyclicity of trace gives

$$\operatorname{tr}(\widetilde{M}\widetilde{S}_K\widetilde{\Omega}\widetilde{S}_K^\top) = \operatorname{tr}(MS_K\Omega S_K^\top).$$

No orthogonality of T is required. □

N.2 EQUAL-BUDGET REUSE

Proof of Proposition 13. The common sample mean has covariance Σ/N, giving the stated reuse variance. If the jth propagated update uses a disjoint mean with n_j observations, these estimates are independent and their propagated variances add:

$$V_{\text{fresh}} = \sum_{j<K} \frac{\|A_j\|_F^2}{n_j}.$$

Cauchy–Schwarz implies

$$\left(\sum_j \|A_j\|_F\right)^2 = \left(\sum_j \frac{\|A_j\|_F}{\sqrt{n_j}}\sqrt{n_j}\right)^2 \le NV_{\text{fresh}}.$$

Equality is achieved by the continuous allocation $n_j \propto \|A_j\|_F$ on the nonzero terms. The triangle inequality gives $\left\|\sum_j A_j\right\|_F \le \sum_j \|A_j\|_F$. Combining proves the result. This argument compares common-mean observations under the same total observation count. It says nothing about trajectories with different temporal information or an estimated operator. □

N.3 SCALAR PREFERRED CHECKPOINT

For $\theta_0 = 0$, $v = b/(1-\rho)$, and $a = \sigma^2/[n(1-\rho)^2]$,

$$R_K = v^2\rho^{2K} + a(1-\rho^K)^2.$$

Let $z = \rho^K \in [0,1]$. Minimizing the quadratic $v^2z^2 + a(1-z)^2$ yields $z_* = a/(v^2+a)$ when $v \neq 0$ and $a > 0$. The continuous preferred checkpoint is

$$K_*^{\text{cont}} = \frac{\log(a/(v^2+a))}{\log\rho}.$$

For the integer problem evaluate its neighboring feasible integers and the permitted boundary checkpoints. If $v = 0$, checkpoint zero has zero risk. If $a = 0$ and $v \neq 0$, convergence is preferred. These elementary limiting cases prevent a universal claim that one finite stopping time is always best.

N.4 FUNNEL RISK AND RESTRICTED SCHEDULING SEPARATION

Proof of Proposition 14. Since $e^\top u = 0$, $(ue^\top)^2 = 0$. The binomial expansion therefore terminates:

$$B_a^j = \rho^j I + ja\rho^{j-1}ue^\top \quad (j \ge 1).$$

All diagonal entries of B_a are ρ and its strictly upper-triangular perturbation is nilpotent, so all eigenvalues equal ρ. Also $B_av = \rho v$, hence $B_a^Kv = c\rho^Ku$. Finally,

$$S_Ke = s_Ke + at_Ku.$$

The vectors e, u are orthogonal with squared norms $1, m$. Substituting into Proposition 12 gives (46). Dividing the variance by its $a = 0$ value gives the ratio, and the endpoint uses $s_\infty = (1-\rho)^{-1}$ and $t_\infty = (1-\rho)^{-2}$. □

Corollary 20 (A pre-data spectrum-only scheduling limitation)**.** *Consider just the choices* $\{1, \infty\}$*, fixed* $a > 0$*, and the two instances* B_0, B_a*. Let* $s = (1-\rho)^{-1}$*,* $t = (1-\rho)^{-2}$ *and choose*

$$c^2 = \frac{\tau^2}{m\rho^2}\left(s^2 - 1 + \frac{1}{2}ma^2t^2\right), \qquad \Delta = \frac{\tau^2 ma^2 t^2}{2(m+1)}.$$

Any possibly randomized pre-data *schedule depending only on the common spectrum, reward mean, true value, reward covariance, and evaluation matrix has expected excess risk at least* $\Delta/2$ *in one of these instances, relative to the better of the two fixed checkpoints.*

Proof. Equation (46) gives

$$R_1(0) = R_1(a) = \frac{R_\infty(0) + R_\infty(a)}{2},$$

with $R_1(0) - R_\infty(0) = R_\infty(a) - R_1(a) = \Delta$. A pre-data rule with access only to the common summary chooses checkpoint one with the same probability q in both instances, independently of the subsequently drawn estimation error. Its excess risks are $q\Delta$ and $(1-q)\Delta$. At least one is at least $\Delta/2$. □

The independence restriction in the proof is essential. One cannot extend this argument to a rule that observes the realized reward noise merely by replacing q by its marginal probability: selected losses and checkpoint choice would be correlated. This corollary concerns restricted resource schedules, not general adaptive estimation.

N.5 Damping does not regularize the converged estimator

For $\alpha \neq 0$, a fixed point of

$$\theta^+ = (1-\alpha)\theta + \alpha(\widehat{B}\theta + \widehat{b})$$

satisfies $(I - \widehat{B})\theta = \widehat{b}$. If this system has a unique solution and the damped iteration converges, that limit is the undamped empirical fixed point. Damping can change convergence or the finite-checkpoint path; it cannot change the unique converged estimator without an additional modification.

O Full empirical-moment expansion and risk remainder

Proof of Theorem 15. Multiplication of the empirical normal equations by A^{-1} gives

$$(I+U)\widehat{\theta}_{k+1} = (B+V)\widehat{\theta}_k + b + w.$$

Insert $\widehat{\theta} = x + u + q + r$. Terms of degree zero yield the population recursion. Terms of degree one yield (47), and terms of degree two yield (48). The remaining identity is exactly (49). Because the recursions begin at zero, induction proves the homogeneity of u and q.

For $k \le H$, the forcing in the u recursion has norm at most $\|w\| + \eta X_H = D$, so convolution with powers of B gives $\max\|u_k\| \le G_H D$. The forcing in the q recursion has norm at most $\eta \max\|u_k\|$, giving $\max\|q_k\| \le G_H^2 \eta D$. Multiplying (49) by $(I+U)^{-1}$ gives a recursion with propagation matrix $\widehat{B} = (I+U)^{-1}(B+V)$ and forcing bounded by $\alpha\eta\max\|q_k\|$. Its convolution bound is (50). For $H = 0$ every term is zero. No probabilistic assumption was used. □

Lemma 21 (Explicit quadratic-risk remainder)**.** *Assume the joint perturbations are centered, and the random inverses and moments below are well-defined with finite second moments. Let*

$$U_2 = (\mathbb{E}\,\|u_K\|_M^2)^{1/2}, \quad Q_2 = (\mathbb{E}\,\|q_K\|_M^2)^{1/2}, \quad T_2 = (\mathbb{E}\,\|r_K\|_M^2)^{1/2}.$$

The remainder in (51) *satisfies*

$$|\mathcal{R}_K| \le 2U_2Q_2 + Q_2^2 + 2(\|d_K\|_M + U_2 + Q_2)T_2 + T_2^2. \tag{67}$$

Proof. Linearity in the centered perturbation gives $\mathbb{E}u_K = 0$. Expand $\|d_K + u_K + q_K + r_K\|_M^2$ and remove the three displayed terms in (51). Bound $2\mathbb{E}u_K^\top M q_K$ by $2U_2Q_2$ and $\mathbb{E}\,\|q_K\|_M^2$ by Q_2^2. Bound the cross term with r_K by Cauchy–Schwarz and the L^2 triangle inequality, then add T_2^2. The seminorm induced by $M \succeq 0$ suffices. □

This result alone is not a probabilistic bound for arbitrary empirical inverse matrices. A stable compact moment domain or another explicit integrability argument is needed to turn the deterministic remainder into rates. Restricting to a favorable sample event can destroy the centering used above; we do not silently make that substitution.

O.1 Exact scalar moments

Put $h = \widehat{\rho} - \rho$. Since f is quadratic,

$$f(\rho + h) - v = d + r(1 + 2\rho)h + rh^2.$$

Squaring and using $\mathbb{E}h = 0$ gives (52). For the average of n Bernoulli variables,

$$\begin{aligned}
\mu_2 &= \frac{\gamma^2 p(1-p)}{n}, \\
\mu_3 &= \frac{\gamma^3 p(1-p)(1-2p)}{n^2}, \\
\mu_4 &= \gamma^4 \left[\frac{3p^2(1-p)^2}{n^2} + \frac{p(1-p)(1-6p(1-p))}{n^3} \right].
\end{aligned}$$

These follow by expanding centered sums: only terms with no singly occurring independent centered variable survive. The exact distribution calculation independently computes $\sum_{j=0}^{n} \mathbb{P}(\mathrm{Bin}(n,p) = j)(f(\gamma j/n) - v)^2$. Its equality to the moment expression is unit tested.

P Unbiasedness obstruction with unknown transitions

Proof of Theorem 17. Condition on each binary observation sequence $x \in \{0,1\}^n$ and integrate out any independent auxiliary randomness. The conditional mean of an integrable statistic is a finite number t_x independent of p. Its expectation is

$$\sum_{x \in \{0,1\}^n} t_x p^{|x|} (1-p)^{n-|x|},$$

a polynomial of degree at most n. The same argument gives polynomials $a(p) = \mathbb{E}_p \widehat{v}^2$ and $b(p) = \mathbb{E}_p \widehat{v}$. Therefore

$$R(p) = a(p) - \frac{2rb(p)}{1 - \gamma p} + \frac{r^2}{(1 - \gamma p)^2}.$$

If a polynomial $T(p)$ equaled this risk on an open interval, multiplying by $(1 - \gamma p)^2$ would give a polynomial identity. At $p = 1/\gamma$, the left side and the first two terms on the right vanish, but the final term is $r^2 \neq 0$, a contradiction. Evaluation at $1/\gamma > 1$ is algebraic: equality on the statistical interval already implies identity of polynomials.

For checkpoints zero and one the estimates are respectively 0 and r, independent of the transition data. Their risk difference is

$$R_0(p) - R_1(p) = \frac{2r^2}{1 - \gamma p} - r^2.$$

Multiplication by $1 - \gamma p$ and evaluation at $p = 1/\gamma$ yields the analogous contradiction. The conclusion is about exact unbiased numerical estimation with a fixed finite observation count. It does not assert difficulty of deciding the sign, nor rule out consistent biased estimates, approximate risk scores, or observation models with unbounded randomized sample sizes. □

Q Exact score and Gaussian simultaneous selection

Q.1 Fixed-checkpoint unbiasedness

Let $C_K = S_K - L$ and $a = C_K b$. Expanding the squared gap gives

$$\mathbb{E} \left\| C_K y \right\|_M^2 = \|a\|_M^2 + \mathrm{tr}(M C_K \Omega C_K^\top).$$

Since M, Ω are symmetric,

$$\begin{aligned}\operatorname{tr}(MC_K\Omega C_K^\top) + 2\operatorname{tr}(MS_K\Omega L^\top) - \operatorname{tr}(ML\Omega L^\top)\\ = \operatorname{tr}(MS_K\Omega S_K^\top).\end{aligned}$$

Thus $\mathbb{E}\widehat{R}_K = R_K$. An offset in a fixed affine candidate changes its deterministic bias, not this covariance cancellation. The candidate maps must be fixed independently of y; learned maps need extra derivative terms.

Q.2 Gaussian quadratic concentration

Lemma 22. *For $z \sim N(0, I)$, a positive semidefinite matrix H, a vector h, and $t > 0$, outside a set of probability at most $4e^{-t}$,*

$$|2h^\top z| \le 2\sqrt{2t}\,\|h\|, \qquad |z^\top Hz - \operatorname{tr} H| \le 2\,\|H\|_F\sqrt{t} + 2\,\|H\|_{\mathrm{op}}\,t.$$

Proof. The first inequality is the scalar Gaussian tail bound applied to $h^\top z$. Diagonalizing H and using rotational invariance yields

$$\log \mathbb{E}e^{s(z^\top Hz - \operatorname{tr} H)} = \frac{1}{2}\sum_i [-\log(1-2s\lambda_i) - 2s\lambda_i] \le \frac{s^2\,\|H\|_F^2}{1 - 2s\,\|H\|_{\mathrm{op}}}$$

for $0 < s < (2\,\|H\|_{\mathrm{op}})^{-1}$. The elementary logarithm bound follows by its nonnegative power series. Chernoff optimization gives the upper-tail threshold $2\,\|H\|_F\sqrt{t} + 2\,\|H\|_{\mathrm{op}}\,t$. Applying the negative-parameter bound, whose log-moment generating function is at most $s^2\,\|H\|_F^2$ for $s < 0$, gives the lower threshold $2\,\|H\|_F\sqrt{t}$. A union bound for the two Gaussian tails and the two quadratic tails gives the stated probability. Zero matrices or vectors follow by continuity. □

Proof of Theorem 16. Write $\varepsilon = \Omega^{1/2}z$; this remains valid for singular Ω. Both the centered score and the centered loss are the sum of a Gaussian linear form and a centered positive semidefinite quadratic form. The deterministic bias is the same vector $a = (S_K - L)b$. For the score use $C = S_K - L$, and for the loss use $C = S_K$. In either case $H_C = \Omega^{1/2}C^\top MC\Omega^{1/2}$ and the linear coefficient is $h = \Omega^{1/2}C^\top Ma$. Therefore

$$\|h\| \le \sqrt{\|H_C\|_{\mathrm{op}}}\,\|a\|_M.$$

Young's inequality gives

$$2\sqrt{2t}\,\|h\| \le \eta\,\|a\|_M^2 + \frac{2t}{\eta}\,\|H_C\|_{\mathrm{op}}.$$

Since $R_K \ge \|a\|_M^2$, Lemma 22 and a union bound over two forms per checkpoint give, with probability at least $1 - 8me^{-t}$,

$$|\widehat{R}_K - R_K| \le \eta R_K + a_K, \qquad |\ell_K - R_K| \le \eta R_K + c_K \quad \text{for all } K. \tag{68}$$

In particular $(1-\eta)R_K \le \widehat{R}_K + a_K$, hence

$$\ell_K \le (1+\eta)R_K + c_K \le U_K.$$

The other side of the score inequality implies

$$U_K \le \frac{(1+\eta)^2}{1-\eta}R_K + \frac{2(1+\eta)}{1-\eta}a_K + c_K.$$

Apply this simultaneously to the minimizer of U and every comparator. With $t = \log(8m/\delta)$ this proves both claims. The proof never assumes checkpoint independence. □

R Smooth unknown-operator score: complete finite-sample proof

We prove the general smooth-map result because the same calculation applies to fixed-Gram moments and to raw (A, F, g) moments when their inverses are uniformly regular. Throughout this appendix all vector norms are Euclidean, matrix covariance contractions use the nuclear norm when needed, and $m_0 = \|M\|_{\mathrm{op}}$.

R.1 Moment and sample-covariance bounds

Let $X_i = Z_i - \mu$, $e = \bar{Z} - \mu$, and $\|X_i\| \le R$. Independence and centering give

$$\mathbb{E}\,\|e\|^2 \le R^2/n, \qquad \mathbb{E}\,\|e\|^4 \le 3R^4/n^2, \qquad \mathbb{E}\,\|e\|^3 \le \sqrt{3}R^3/n^{3/2}. \tag{69}$$

To verify the fourth moment, write

$$\Big\|\sum_i X_i\Big\|^2 = \sum_i \|X_i\|^2 + 2\sum_{i<j}\langle X_i, X_j\rangle.$$

After squaring and taking expectations, mixed terms with a singly occurring centered variable vanish. The remaining diagonal fourth moments, squared-norm cross terms, and squared-inner-product terms are bounded by $nR^4 + 3n(n-1)R^4 \le 3n^2R^4$. Divide by n^4. The third-moment bound follows by Cauchy–Schwarz.

The sample covariance is positive semidefinite and unbiased, with

$$\mathbb{E}\widehat{\Gamma} = \Gamma, \qquad \operatorname{tr}\widehat{\Gamma} = \frac{\sum_i \|X_i\|^2 - n\,\|e\|^2}{n-1} \le 2R^2. \tag{70}$$

If one observation is replaced, the mean changes by at most $2R/n$. The difference of covariance matrices has nuclear norm at most $12R^2/n$: the difference of the uncentered rank-one sums is at most $2R^2$, and the difference of $nee^\top$ is at most $4R^2$, all divided by $n-1$. These deterministic bounds do not require independence between sample mean and covariance.

R.2 Cancellation of the checkpoint curvature

Fix a checkpoint, suppress its subscript, and set $d = f(\mu) - h(\mu)$. Define

$$g(x) = \|f(x) - h(x)\|_M^2, \qquad \ell(x) = \|f(x) - h(\mu)\|_M^2.$$

The half-Hessian contraction difference is

$$\begin{aligned}\frac{1}{2}(D^2 g - D^2\ell)(\mu) : \Gamma = &-2\operatorname{tr}(MDf(\mu)\Gamma Dh(\mu)^\top)\\ &+\operatorname{tr}(MDh(\mu)\Gamma Dh(\mu)^\top) - d^\top M(D^2h(\mu):\Gamma).\end{aligned} \tag{71}$$

Both quantities contain the checkpoint mean term $d^\top M(D^2 f : \Gamma)$, so that term cancels. The endpoint mean term does not. Define

$$\begin{aligned}C(x,G) = &\,2\operatorname{tr}(MDf(x)GDh(x)^\top) - \operatorname{tr}(MDh(x)GDh(x)^\top)\\ &+ (f(x)-h(x))^\top M(D^2h(x):G).\end{aligned} \tag{72}$$

Then the score is $g(\bar{Z}) + C(\bar{Z}, \widehat{\Gamma})/n$. Equation (71) identifies the correction at population moments.

R.3 Derivative bounds

For a vector function a, the third derivative of $\|a\|_M^2$ in a unit direction is

$$2\langle a, MD^3a\rangle + 6\langle Da, MD^2a\rangle.$$

Under Assumption 2, this yields

$$\|D^3 g\| \le m_0(8L_0L_3 + 24L_1L_2), \qquad \|D^3\ell\| \le m_0(4L_0L_3 + 6L_1L_2). \tag{73}$$

The bounds hold along every unit direction, which suffices for the one-dimensional Taylor formula on segments; the mixed multilinear form is bounded by the same product estimates.

For $G \succeq 0$, differentiating (72) and bounding each covariance contraction by $\operatorname{tr} G$ gives

$$|C(x,G) - C(y,G)| \le m_0(8L_1L_2 + 2L_0L_3)\operatorname{tr}(G)\,\|x-y\|. \tag{74}$$

Indeed, the first trace contributes $4L_1L_2$, the second $2L_1L_2$, and the endpoint term $2L_1L_2 + 2L_0L_3$. For arbitrary symmetric $G - G'$, linearity and the nuclear/operator-norm dual bound give

$$|C(x,G) - C(x,G')| \le m_0(3L_1^2 + 2L_0L_2)\,\|G - G'\|_*. \tag{75}$$

These are dimension-free in the supplied regularity constants, which themselves can depend strongly on dimension and horizon.

R.4 EXPECTED-SCORE APPROXIMATION

Proof of Theorem 18. Taylor-expand $g(\mu + e)$ and $\ell(\mu + e)$ to second order. The constant terms agree, first-order expectations vanish, and $\mathbb{E}ee^\top = \Gamma/n$. Equation (71) gives

$$\mathbb{E}g(\bar{Z}) - \mathbb{E}\ell(\bar{Z}) = -C(\mu, \Gamma)/n + \zeta,$$

where (73) and Taylor's factor $1/6$ imply

$$|\zeta| \le m_0(2L_0L_3 + 5L_1L_2)\mathbb{E}\left\|e\right\|^3.$$

Because C is linear in its second argument and $\widehat{\Gamma}$ is unbiased, $\mathbb{E}C(\mu, \widehat{\Gamma}) = C(\mu, \Gamma)$. The remaining plug-in term is bounded using (74), (70), and $\mathbb{E}\left\|e\right\| \le R/\sqrt{n}$:

$$\frac{1}{n}\mathbb{E}|C(\bar{Z}, \widehat{\Gamma}) - C(\mu, \widehat{\Gamma})| \le \frac{2m_0(8L_1L_2 + 2L_0L_3)R^3}{n^{3/2}}.$$

Combining with (69) yields (64). Dependence between e and $\widehat{\Gamma}$ was bounded deterministically rather than discarded. The constants are common to all checkpoints, establishing uniformity of the fixed-checkpoint statement. □

For completeness, the same calculation without the endpoint comparison gives the explicit population risk expansion

$$\begin{aligned} R_K = \ & \|d_K\|_M^2 + \frac{1}{n}\operatorname{tr}(MDf_K(\mu)\Gamma Df_K(\mu)^\top) \\ & + \frac{1}{n}d_K^\top M(D^2 f_K(\mu) : \Gamma) + \zeta_K, \\ |\zeta_K| \le \ & \frac{\sqrt{3}m_0R^3}{n^{3/2}}\left(\frac{2}{3}L_0L_3 + L_1L_2\right). \end{aligned} \tag{76}$$

This is another direct way to see why variance alone misses an order-$1/n$ term at a biased checkpoint.

R.5 ADAPTIVE SELECTION BY BOUNDED DIFFERENCES

Proof of Corollary 19. The gradient of g has norm at most $8m_0L_0L_1$. A single observation replacement therefore changes its value by at most $16m_0L_0L_1R/n = A_0/n$. For the correction C/n, split the change into a mean change holding the old covariance fixed and a covariance change holding the new mean fixed. Equations (74)–(75) and the replacement bounds give an additional change of at most B_0/n^2. The loss map ℓ has gradient norm at most $4m_0L_0L_1$, giving replacement bound D_0/n.

For independent observations and a function with coordinate replacement bounds c_i, the bounded-differences inequality follows by exposing observations sequentially, applying the conditional Hoeffding bound to each martingale difference of range length at most c_i, and multiplying the moment generating functions. It gives

$$\mathbb{P}(|T - \mathbb{E}T| > u) \le 2\exp\left(-\frac{2u^2}{\sum_i c_i^2}\right).$$

Applying this to every score and loss, a union bound with $t = \log(4m/\delta)$ yields

$$\begin{aligned} |\widehat{R}_K^{(2)} - \mathbb{E}\widehat{R}_K^{(2)}| &\le (A_0 + B_0/n)\sqrt{t/(2n)}, \\ |\ell_K - R_K| &\le D_0\sqrt{t/(2n)} \quad \text{simultaneously for all } K. \end{aligned}$$

Theorem 18 turns the first line into a uniform deviation from R_K with an added β_n. The argmin inequality gives

$$R_{\widehat{K}} \le \min_K R_K + 2(A_0 + B_0/n)\sqrt{t/(2n)} + 2\beta_n.$$

Here $R_{\widehat{K}}$ is the deterministic expected-risk table evaluated at a random index; it is not equated to expected selected loss. The second simultaneous inequality transfers this table comparison to the actual realized loss, giving (65). □

R.6 INDEPENDENT-ROW COVARIANCE PROJECTION

Suppose a block consists of independent state strata, so Γ is block diagonal. Let $\mathcal{P}$ remove all off-block entries. Then $\mathcal{P}(\widehat{\Gamma}) \succeq 0$, has the same trace, and is unbiased for Γ. Moreover $\mathcal{P}$ is a contraction in nuclear norm: it is an average of orthogonal sign-conjugations that flip blocks, so the triangle inequality and unitary invariance give $\|\mathcal{P}(H)\|_* \le \|H\|_*$. Thus the trace, mean, and replacement bounds used above remain valid after projection. This justifies the faster independent-row implementation under its stated sampling design. It does not justify deleting genuine within-block or cross-stratum dependence.

R.7 CONCRETE SMOOTHNESS CONSTANTS FOR TABULAR EVALUATION

The assumptions are nonvacuous. Suppose $B \ge 0$, $\|B\|_\infty \le \gamma < 1$, and $\|b\|_\infty \le r_{\max}$ throughout the moment domain. Use the input norm $\max(\|\Delta b\|_\infty, \|\Delta B\|_\infty)$. The Neumann series and its differentiated terms give, uniformly over all finite checkpoints and the endpoint,

$$
\begin{aligned}
L_{0,\infty} &= \frac{r_{\max}}{1-\gamma},\\
L_{1,\infty} &= \frac{1}{1-\gamma} + \frac{r_{\max}}{(1-\gamma)^2},\\
L_{2,\infty} &= \frac{2}{(1-\gamma)^2} + \frac{2r_{\max}}{(1-\gamma)^3},\\
L_{3,\infty} &= \frac{6}{(1-\gamma)^3} + \frac{6r_{\max}}{(1-\gamma)^4}.
\end{aligned}
$$

For example, one derivative either differentiates b or one of the j matrix factors; summing $\sum_j \gamma^j$ and $\sum_j j\gamma^{j-1}$ yields L_1. Higher derivatives count ordered distinct matrix-factor choices, and possibly one reward derivative, giving the displayed factorial sums. Partial sums for finite checkpoints are bounded by the corresponding infinite sums.

For row-major Euclidean primitive coordinates in dimension d, conservative choices in Assumption 2 are $L_0 = \sqrt{d}L_{0,\infty}$ and $L_j = d^{(j+1)/2}L_{j,\infty}$. This follows because the induced matrix infinity norm is at most $\sqrt{d}$ times its Frobenius norm, while vector Euclidean output norm is at most $\sqrt{d}$ times its infinity norm. These deliberately loose constants illustrate the horizon/dimension cost. We do not claim horizon-free or practically tight selection bounds. For random empirical Gram matrices one additionally needs uniform nonsingularity on the full moment domain; the tabular argument does not provide it automatically.

S INFLUENCE IMPLEMENTATION AND COVARIANCE BOOKKEEPING

For fixed-Gram coordinates $z = (b_s, B_{s,:})_s$, Δz includes both reward and transition perturbations. Differentiating the population recursion yields $J_{k+1} = BJ_k + T(x_k)$. Differentiating $v = Bv + b$ yields $J_\infty = LT(v)$. The quadratic endpoint term solves

$$q_\infty = Bq_\infty + \Delta B(J_\infty \Delta z),$$

so its expectation is $\beta_\infty = L\mathbb{E}[\Delta B(J_\infty \Delta z)]$. This proves (66). When the operator is known and only rewards vary, $\Delta B = 0$, $J_K = S_K$, and $\beta_\infty = 0$; the score reduces exactly to (53).

For independent row strata, let W_s be the covariance of the empirical mean of row s, with primitive ordering $[r_s, \gamma e_{s'}]$, including the zero vector at termination. Put $z(x) = (1, x^\top)^\top$ and $v = Lb$. Then

$$
\begin{aligned}
C_\infty &= L\,\mathrm{diag}\left(z(v)^\top W_s z(v)\right)L^\top,\\
C_{k+1,\infty} &= BC_{k,\infty} + \mathrm{diag}\left(z(x_k)^\top W_s z(v)\right)L^\top, \quad C_{0,\infty} = 0,\\
(\beta_\infty) &= Lt, \qquad t_s = \sum_j L_{js}\ \mathrm{Cov}\left(\Delta B_{sj}, \Delta b_s + \Delta B_{s,:}v\right).
\end{aligned}
$$

The cross covariance recursion uses the same dataset noise at every checkpoint. It is not the recursion for independent per-update innovations. Sample-mean covariance is estimated as

$$\widehat{W}_s = \frac{1}{n(n-1)} \sum_{i=1}^{n} (z_{si} - \bar{z}_s)(z_{si} - \bar{z}_s)^\top.$$

Using individual-observation covariance without dividing by n would inflate the score by a factor n. Tests verify the fast rowwise result against full primitive-covariance contraction, and verify reduction to known-operator SURE when transition uncertainty is zero.

The general primitive dimension for fixed-Gram features is $p = d(d+1)$. Explicit Jacobians require $O(dp)$ memory per checkpoint, and full covariance requires $O(p^2)$ storage. Full matrix products can be costly as d grows. The implementation streams the recursion between selected checkpoints; scalability to neural value estimators is not established. The independent-row specialization avoids storing cross-row covariance that is known to be zero, not covariance that is merely inconvenient to estimate.