# FROM PARTIAL IDENTIFICATION TO DIRECTED EXPLORATION IN WORLD MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

World models support planning by predicting action consequences, but data collected by a controller may lack sufficient action variation, leaving different predictions for untried actions. We study world models with linear Gaussian latent dynamics and recoverable states. We fully characterize which action effects existing data determine and which remain ambiguous, and quantify how prediction accuracy on observed trajectories constrains predictions under changed actions. This characterization guides the identification of action directions requiring additional exploration. We determine the minimum number of directions needed to resolve the ambiguity and optimize probing strength across these directions to improve the accuracy of action effect estimates. The resulting method combines existing trajectories and new interventions to estimate action effects. We further analyze how errors in learning state representations and selecting probe directions affect the final estimation accuracy. With matched interaction counts and action perturbation budgets, directed acquisition reduces control cost by about $53\%$ against balanced full-space exploration under missing and weak coverage. Robot manipulation and pixel experiments further show improvements in control success and action response prediction, respectively.

## 1 INTRODUCTION

World models support planning by predicting the consequences of different actions from the same state (Sutton, 1991; Ha & Schmidhuber, 2018; Hafner et al., 2019; 2025; Hansen et al., 2024; Zhou et al., 2025). However, their training trajectories often come from controllers that select actions according to the state, coupling action variation with state variation (Gustavsson et al., 1977; Forssell & Ljung, 1999; Levine et al., 2020). When action variation conditional on state is insufficient, multiple world models can explain the recorded trajectories equally well while predicting different consequences of changing the action (Zhang et al., 2026). Figure 1 illustrates such explanations through braking and passive drag. Accurate prediction of recorded trajectories therefore does not determine all action consequences needed for planning. We need to establish which action comparisons the data support and which action effects require further exploration.

Under full conditional action coverage and conditions for representation recovery, Zhang et al. (2026) establish joint recovery of states and controlled dynamics in world models. Their analysis also exhibits prediction error amplification under weak coverage and indistinguishable models when action variation is missing. Under partial coverage, it remains to characterize completely the action effects determined by existing data and the remaining ambiguity, and to determine the interventions that resolve it. These questions are linked because the action effects constrained by existing data determine which information exploration needs to add. We address them for world models with linear Gaussian latent dynamics under conditions ensuring state recovery, and study how a limited action perturbation budget can improve estimation accuracy.

The central insight is that two models fitting the same behavior data can disagree on action consequences only along directions that the controller never varied independently, so perturbing those directions can distinguish them. Feedback can produce recorded action variation in every direction while hiding some action effects in the state response, preventing the action distribution alone from revealing the missing directions. We subtract the part of each recorded action predictable from the state and use the resulting conditional action residuals to distinguish directions with independent



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.

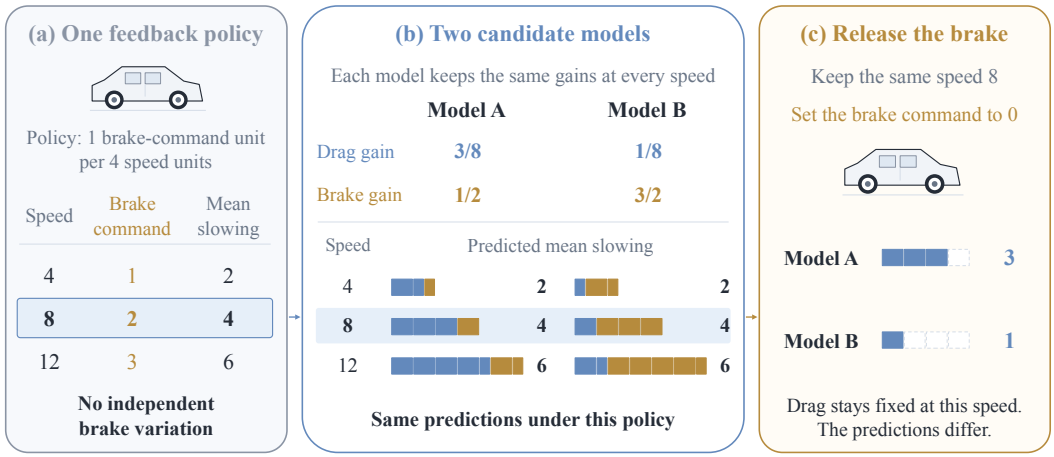



Figure 1: Two candidate models with fixed parameters predict the same mean slowing under the recorded feedback policy, but attribute it differently to drag and braking. Holding speed fixed and releasing the brake leaves each model's drag contribution unchanged and removes its braking contribution, yielding different predictions. Independently varying the action can distinguish these two models, both compatible with the existing data.

variation from those without excitation. Behavior data determine the action effects in the former directions, while the effects in the latter can still vary across compatible models. This distinction lets us characterize all models compatible with the behavior data and determine the action variation that new interventions must supply. The resulting method retains the transition components determined by existing trajectories, focuses probes on action directions whose effects remain insufficiently determined, and allocates probe strength using existing information and action costs. New data estimate the effects in these directions, which combine with the retained components to give a complete transition estimate. Figure 2 shows the procedure.

Our contributions are as follows.

1. **Characterizing identified and ambiguous action effects.** For linear Gaussian latent systems, we characterize all affine dynamics compatible with behavior data, identifying the transition components fixed by existing trajectories and the action effects that remain ambiguous. We give sharp bounds on how behavior prediction error constrains predictions under changed actions, distinguishing the consequences of missing and weak action excitation for identification.
2. **Locating action directions that need exploration.** We characterize when feedback prevents the marginal action distribution from revealing gaps in coverage, and establish the relation between conditional action residuals and undetermined action effects. This relation lets us estimate directions requiring further exploration from data and quantify how finite samples and errors in learned coordinates affect these estimates.
3. **Allocating probe budgets and estimating action effects.** We determine the minimum number of intervention directions needed to resolve ambiguity. In the analyzed regression and information design problems, we derive optimal perturbation allocations accounting for existing information and action costs. We construct a two-stage estimator combining existing trajectories with new interventions, establish error guarantees, and experimentally evaluate prediction and control performance.

## 2 Related Work and Motivation

### 2.1 Related Work

**State and action-effect identification.** Recovering a state representation and identifying action effects are distinct problems. Classical closed-loop identification already establishes the role of feedback and independent excitation (Gustavsson et al., 1977; Forssell & Ljung, 1999; Bazanella et al., 2010). Latent models additionally require informative coordinates. Matching the observation

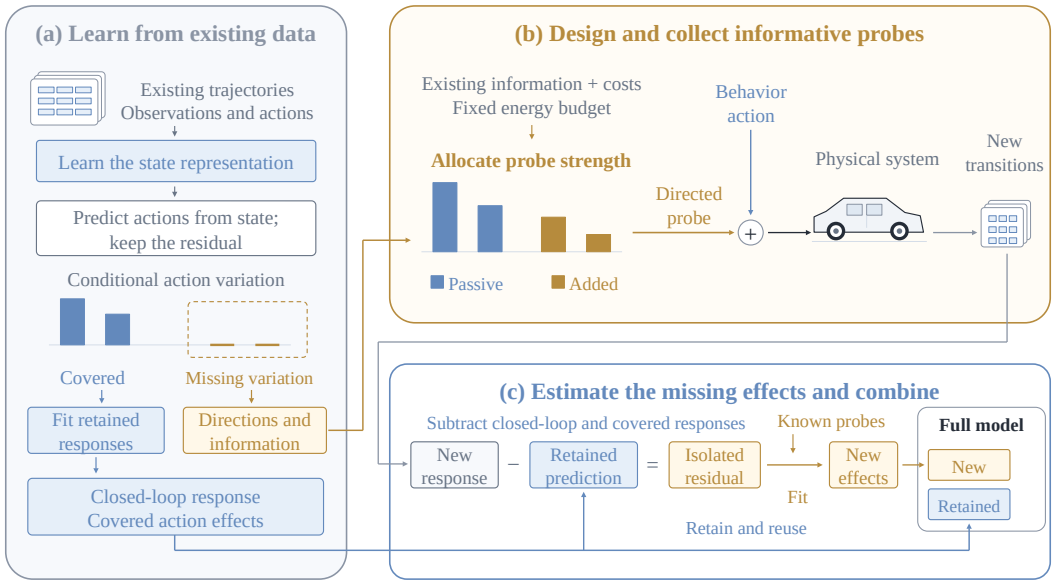



Figure 2: Two-stage estimation under missing conditional action coverage. Action residuals reveal the unexcited directions. Probes add independent variation in these directions. Subtracting the estimated closed-loop response and covered action effects from new transitions gives a regression for the remaining effect. Combining the retained and newly estimated effects gives dynamics estimates. Information allocation accounts for passive information and action costs, as specified in (18).

distribution alone need not identify them (Hyvärinen & Pajunen, 1999; Locatello et al., 2019), while temporal or auxiliary structure and interventions supply additional identifying assumptions (Hyvärinen & Morioka, 2016; Khemakhem et al., 2020; Ahuja et al., 2023). Mechanism sparsity and observed intervention targets provide other routes to latent identification (Lachapelle et al., 2022; Lippe et al., 2022). Closest to our setting, Zhang et al. (2026) separate predictable-signal conditions for state recovery from conditional action coverage. Under their stationary linear–Gaussian, observation and predictor assumptions, positive signal separation and full conditional coverage yield joint recovery up to a rotation; their constructions also expose weak-coverage amplification and missing-coverage counterexamples. We retain the signal-separation premise and characterize the entire compatible affine set under partial coverage, including its invariants and a dimension threshold for masking by the marginal action law.

**Control without full identification.** Partial identification describes what assumptions and observations determine through a set of compatible parameters (Manski, 2003; Tamer, 2010). In control, persistently exciting trajectories can span a system's behavior (Willems et al., 2005), whereas data informativity asks whether a specified control task is possible even without unique dynamics (van Waarde et al., 2020). In offline reinforcement learning, methods address distribution shift through uncertainty penalties and pessimistic models (Levine et al., 2020; Yu et al., 2020; Kidambi et al., 2020); causal-confusion work uses targeted interventions to distinguish explanations learned from behavior (de Haan et al., 2019). These perspectives separate useful control from complete identification. Our affine fiber specifies which action responses all compatible transitions share and where they can disagree. When the action response of interest remains ambiguous, additional acquisition aims to distinguish the compatible responses.

**Exploration and model uncertainty.** Bayesian experimental design maximizes expected utility (Chaloner & Verdinelli, 1995); identification-oriented design trades input cost against estimation accuracy (Pukelsheim, 2006; Bombois et al., 2006). Active methods select informative inputs for linear systems (Wagenmaker & Jamieson, 2020) or plan trajectories to excite known nonlinear features (Mania et al., 2022), with task-specific objectives when estimation serves control (Wagenmaker et al., 2021). Our design first removes the effects already identified by passive data. In the resulting Gaussian regression, we allocate a fixed expected energy budget to minimize the minimax risk of

estimating the remaining action effects. Curiosity and disagreement methods use prediction error or differences among learned models to guide interaction (Pathak et al., 2017; Burda et al., 2019; Shyam et al., 2019; Pathak et al., 2019; Sekar et al., 2020). Ensemble agreement need not certify an identified response. Members can share an extrapolation outside conditional coverage, and fitting variability can create disagreement inside it. The proposed screen complements these methods by diagnosing independent action variation.

### 2.2 From Identification Results to Intervention Design

The complete affine identified set characterizes the action effects determined by behavior data and those that additional interaction must distinguish. An executable intervention design must also learn the state representation and missing action directions from finite data. Representation and subspace estimation errors affect the choice of intervention directions, while sample and energy allocation affect the accuracy of estimating the remaining action effects. We therefore study how to turn the identified set into an acquisition rule, propagate representation and finite-sample errors to the resulting estimates, and quantify accuracy gains under given interaction and energy budgets. Weak coverage also matters. Even when no direction is exactly missing, passive information can vary substantially across directions. Acquisition should account for this variation and allocate energy to action directions with insufficient information.

## 3 Setting and identification target

Fix a latent dimension $d \geq 1$ and action dimension $p \geq 2$. Let

$$Z \sim \mathcal{N}(0, I_d), \qquad E \sim \mathcal{N}(0, \Sigma), \qquad U = KZ + E, \tag{1}$$

$$Z^+ = AZ + BU + \Xi, \qquad Z^+ \sim \mathcal{N}(0, I_d), \tag{2}$$

where $E \perp Z$, the centered Gaussian noise $\Xi$ is independent of $(Z, U)$, and $A \in \mathbb{R}^{d\times d}$, $B \in \mathbb{R}^{d\times p}$, $K \in \mathbb{R}^{p\times d}$. The action covariance $\Sigma \succeq 0$ has rank $r$ with $1 \leq r < p$; write $S := \mathrm{range}(\Sigma)$ for the excited action subspace and $P_S$ for its orthogonal projector. A centered Gaussian is supported on the range of its covariance, so $E \in S$ almost surely and every point of $S$ lies in its support.

The observation is $X = g(Z)$, where $g$ is unknown and *state sufficient*, meaning that some measurable decoder $s$ satisfies $s(g(Z)) = Z$ almost surely. Thus $g$ has a left inverse on the data support. The next observation uses the same map, $X^+ = g(Z^+)$. An encoder $h$ is admissible when it is measurable and population whitened,

$$\mathbb{E}h(X) = 0, \qquad \mathrm{Cov}(h(X)) = I_d. \tag{3}$$

Stationarity gives the same two moments for $h(X^+)$, and (3) imposes no Gaussianity on the encoder output.

Let $C := AZ + BU$ be the predictable part of the next state, $R := \mathrm{Cov}(C)$ and $\rho := \lambda_{\max}(R)$. Stationarity and noise independence give $I_d = R + \mathrm{Cov}(\Xi)$, so $0 \preceq R \preceq I_d$. The representation premise is

$$\gamma_{\mathrm{rep}} := \lambda_{\min}(R) - \lambda_{\max}(R)^2 > 0, \tag{4}$$

which forces $R \succ 0$, $\rho < 1$ and $R - \rho^2 I_d \succeq \gamma_{\mathrm{rep}} I_d \succ 0$. For features normalized to unit variance, measure predictability by the minimum mean squared error when predicting the next feature from the current state and action. Under the Gaussian model, (4) ensures that even the least predictable linear state feature has lower error than every purely nonlinear feature, meaning one orthogonal to all linear state features (Proposition 10). This gap gives the prediction objective a preference for preserving every linear state direction.

For an encoder $h$ and a predictor $F$ in a class $\mathcal{F}$, consider

$$\mathcal{L}(h, F) := \mathbb{E}\|h(X^+) - F(h(X), U)\|_2^2. \tag{5}$$

With $h_Q(x) = Qs(x)$ on the data support, we require only target realizability,

$$F_Q(y, u) := QAQ^\top y + QBu \in \mathcal{F} \quad \text{for every orthogonal } Q. \tag{6}$$

This asks for the true affine predictor only at target encoders, not the conditional mean for every admissible encoder.

After recovering and undoing the encoder rotation, we compare affine maps in the original latent coordinates. Define

$$\mathcal{I}(A, B) := \{(A', B') : A'Z + B'U = AZ + BU \text{ a.s.}\}. \tag{7}$$

Centering identifies any compatible affine intercept as zero. Two pairs are equivalent when their on-policy conditional means agree; using the same independent noise $\Xi$ then gives the same $Z^+$ under the behavior law.

Point identification means that (7) is a singleton, and partial identification asks for the whole set when it is not (Manski, 2003; Tamer, 2010). The behavior law in identified state coordinates determines $K$ through $KZ = \mathbb{E}[U \mid Z]$ and $S$ as the support subspace of $U - KZ$. Finite-sample benchmarks use $n$ independent stationary triples $(Z_i, U_i, Z_i^+)$. The learned-coordinate estimators observe $(X_i, U_i, X_i^+)$ and use an independently trained encoder. We write $\|\cdot\|_{\mathrm{op}}$ and $\|\cdot\|_F$ for operator and Frobenius norm and $\Sigma^\dagger$ for the Moore-Penrose inverse.

## 4 State Representation Recovery

Action effects can be analyzed only in a state coordinate that expresses the dynamics. Under the signal-separation, realizability and normalization conditions above, the population predictive objective recovers that coordinate even when conditional action coverage is singular.

**Theorem 1** (Population state recovery)**.** *Assume* (1)–(5)*, target realizability* (6)*, state sufficiency of $g$, and the separation condition* (4)*. Then every population global minimizer has an orthogonal matrix $Q \in \mathbb{R}^{d\times d}$ such that*

$$h(g(z)) = Qz \quad \textit{for standard-Gaussian-almost every } z. \tag{8}$$

The conclusion covers every centered, whitened measurable encoder regardless of $\operatorname{rank}\Sigma$. The proof separates the Gaussian channel's degree-one and higher-order modes by (4) (Proposition 10). If an admissible $h$ and predictor $F$ satisfy $\mathcal{L}(h, F) \le d - \operatorname{tr} R + \varepsilon$, some orthogonal $Q$ gives $\mathbb{E}\|h(g(Z)) - QZ\|_2^2 \le 2\varepsilon/\gamma_{\mathrm{rep}}$ (Theorem 11, Appendix A), with separate terms for imperfect whitening and an empirical objective (Propositions 12 and 13).

Even in this recovered coordinate, missing action variation can leave several transition parameters consistent with the behavior law. We next characterize that remaining ambiguity.

## 5 Partial Identification and Counterfactual Prediction

Write the behavior action as its state-predicted part plus a residual. Substituting $U = KZ + E$ into the transition gives

$$AZ + BU = (A + BK)Z + BE,$$

and the data vary $Z$ in every latent direction while varying $E$ only inside $S$. The joint support is therefore

$$\operatorname{supp}(Z, U) = \{(z, Kz + e) : z \in \mathbb{R}^d,\ e \in S\}. \tag{9}$$

An action effect outside $S$ can be changed if the state matrix compensates for it. Indeed, a difference of affine means on (9) is $(\Delta_A + \Delta_B K)z + \Delta_B e$. Independent variation of $z$ and $e \in S$ makes this zero exactly when $\Delta_A = -\Delta_B K$ and $\Delta_B P_S = 0$ (Lemma 14, Appendix B).

**Theorem 2** (The affine identified fiber)**.** *Under the assumptions of Theorem 1, transform the learned coordinates by $Q^\top$. Then the affine identified set is*

$$\boxed{\mathcal{I}(A, B) = \{(A - DK, B + D) : D \in \mathbb{R}^{d\times p},\ DP_S = 0\}.} \tag{10}$$

*Its dimension is $\dim \mathcal{I}(A, B) = d(p - r)$, and the restriction of $B$ to $S^\perp$ can be changed to any linear map $S^\perp \to \mathbb{R}^d$, accompanied by the state-matrix correction in* (10)*.*

The fiber has an equally explicit complement.

**Proposition 3** (Complete invariants)**.** *For two affine pairs $(A, B)$ and $(A', B')$, equality of their on-policy means is equivalent to*

$$A' + B'K = A + BK, \qquad B'P_S = BP_S. \tag{11}$$

Behavior data determine the closed-loop response and the action response on $S$, with one free $d$-vector per unexcited action coordinate (cf. Gustavsson et al., 1977; Bazanella et al., 2010) (Appendix B.2). Every fiber member is a controlled transition giving the same next state on each record, since $DP_S = 0$ implies $DE = 0$ almost surely.

Define $\mathcal{R}(A', B') := \mathbb{E}\|Z^+ - A'Z - B'U\|_2^2$. Anchoring this prediction risk at the true conditional mean converts the fiber into a continuous modulus.

**Theorem 4** (Exact excess risk and sharp counterfactual modulus)**.** *Let $\Delta_A = A' - A$, $\Delta_B = B' - B$, and $H = \Delta_A + \Delta_B K$. Then*

$$\mathcal{R}(A', B') - \mathcal{R}(A, B) = \|H\|_F^2 + \operatorname{tr}(\Delta_B \Sigma \Delta_B^\top). \tag{12}$$

*Consequently, excess risk at most $\varepsilon$ implies*

$$\|\Delta_A + \Delta_B K\|_F \le \sqrt{\varepsilon}, \qquad \|\Delta_B P_S\|_F \le \sqrt{\varepsilon/\lambda_+}, \tag{13}$$

*where $\lambda_+ := \lambda_{\min}(\Sigma|_S) > 0$. For every $z \in \mathbb{R}^d$ and residual action $e \in \mathbb{R}^p$, write $\ell_S(z, e) := \|z\|_2^2 + e^\top \Sigma^\dagger e$. The exact worst-case disagreement is*

$$\sup_{\mathcal{R}(A',B') - \mathcal{R}(A,B) \le \varepsilon} \|Hz + \Delta_B e\|_2 = \begin{cases} \sqrt{\varepsilon \ell_S(z, e)}, & e \in S, \\ +\infty, & e \notin S. \end{cases} \tag{14}$$

*The formula includes $\varepsilon = 0$, where it is zero on the behavior support and unbounded at every residual point off that support.*

Both branches are attainable. For a unit residual action along an excited eigenvector at $z = 0$, the modulus $\sqrt{\varepsilon/\lambda}$ distinguishes identification at every positive eigenvalue from instability as $\lambda \to 0$. Along an unexcited direction it is infinite even at zero excess risk. A near-optimal fit remains near an exact fiber member (Corollary 17).

The fiber, invariants and span criterion hold for any law with $\operatorname{supp}(Z, E) = \mathbb{R}^d \times S$ (Proposition 16). Additional behavior data leave action effects along unexcited directions undetermined. We next consider interventions that identify these effects.

## 6 Directed Exploration and Action Effect Estimation

The procedure in Figure 2 first estimates the invariants and excited subspace, then adds interventions that remove the remaining fiber. Missing directions concern the action variation left after conditioning on state, which the marginal covariance $\Sigma_U = \operatorname{Cov}(U) = KK^\top + \Sigma$ need not reveal.

**Proposition 5** (Feedback masking of the excited subspace)**.** *Let $1 \le r < p$. If $d + r \ge p$, then every $\Sigma_U \succ 0$ and every $r$-dimensional $S \subseteq \mathbb{R}^p$ are realized by some $K$ and some $\Sigma \succeq 0$ with $\operatorname{range}(\Sigma) = S$, so no function of the marginal action law determines $S$. If $d + r < p$, then $\Sigma_U$ is singular with $\ker(\Sigma_U) = \ker(K^\top) \cap S^\perp$ of dimension at least $p - d - r$, so the marginal law does determine a nonzero part of $S^\perp$. A nonsingular marginal action covariance is therefore compatible with every $r$-dimensional excited subspace exactly when $d + r \ge p$.*

When the state dimension is at least the number of missing action directions, a marginal screen has no guarantee, although it can succeed when the realized feedback gain does not mask them. Residual moments estimate the invariants without choosing a fiber member. With $Y = Z^+$, they satisfy

$$\begin{aligned} M &= A + BK = \mathbb{E}[YZ^\top], \\ H_B &= B\Sigma = \mathbb{E}[YU^\top] - MK^\top. \end{aligned} \tag{15}$$

Their sample analogues use only covariance matrices, with errors of order $L\sqrt{(2d + p + \log(1/\delta))/n}$ for $L$ the operator norm of $\operatorname{Cov}(Z, U, Y)$. Once $\lambda_+$ exceeds four times this tolerance, thresholding $\widehat{\Sigma}$ returns rank exactly $r$, and the errors of the subspace and of $\widehat{H}_B \widehat{\Sigma}_r^\dagger \to BP_S$ scale with the moment error over $\lambda_+$, so the smallest nonzero action eigenvalue controls the sensitivity to weak coverage (cf. Simchowitz et al., 2018; Ziemann et al., 2023) (Theorem 20, Appendix C.4).

Exact state coordinates recover $S$ almost surely when $n \ge d + r$ (Proposition 18). A spectral gap controls sensitivity to coordinate and sampling errors and defines the subspace of weak excitation

under full-rank coverage. An $\varepsilon$-optimal representation adds encoder error of order $\sqrt{\varepsilon/\gamma_{\mathrm{rep}}}$, which Theorem 22 propagates to $K, M, S$ and $BP_S$.

Let $q = p - r$ and let $V \in \mathbb{R}^{p\times q}$ be an orthonormal basis of $S^\perp$. Add independent scalar perturbations along $v_1, \ldots, v_k$. Their residual support is $S_k = \operatorname{span}(S, v_1, \ldots, v_k)$. The support calculation yields the following theorem.

**Theorem 6** (Span criterion and minimum achievable count)**.** *In identified latent coordinates the affine transition fiber under $(Z, U_k)$ is $\mathcal{I}_k(A, B) = \{(A - DK, B + D) : DP_{S_k} = 0\}$, of dimension $d(p - \dim S_k)$, so it is a singleton precisely when*

$$S_k = \mathbb{R}^p, \qquad \textit{and} \qquad \min\{k : \mathcal{I}_k(A, B) \textit{ a singleton}\} = p - r. \tag{16}$$

*Every choice with $k < p - r$ leaves a nonzero affine ambiguity.*

The count is achievable only when the directions satisfy the span condition (cf. Gevers et al., 2009); stationary retraining and admissible energy are treated in Appendix D.

In exact latent coordinates with known $K$, $M = A + BK$ and $G_S = BP_S$, subtracting the closed-loop response and the action response on $S$ isolates the missing effect $T = BV$ in the oracle regression

$$R_i = Y_i - MZ_i - G_S(U_i - KZ_i) = Tx_i + \xi_i, \tag{17}$$

With noise covariance trace $\nu$ and independent Gaussian probes $x_i \sim \mathcal{N}(0, \Omega)$, we minimize estimation risk under a fixed expected energy budget.

For $\Omega \succ 0$ and $n > q + 1$ the minimax risk is $\mathfrak{R}_n(\Omega) = \nu \operatorname{tr}(\Omega^{-1})/(n - q - 1)$, attained by ordinary least squares (Mourtada, 2022) and infinite when $\operatorname{rank}(\Omega) < q$ or $n \le q + 1$. Under an action cost $W \succ 0$ and total expected energy $n \operatorname{tr}(W\Omega) \le \mathcal{E}$ (cf. Pukelsheim, 2006; Bombois et al., 2006), the unique optimal covariance is $\Omega_\star = \mathcal{E}\, W^{-1/2}/\{n \operatorname{tr}(W^{1/2})\}$, with risk $\nu n\{\operatorname{tr}(W^{1/2})\}^2/\{\mathcal{E}(n - q - 1)\}$ (Theorem 30, Appendix E.1).

With partial passive information, we allocate new measurements on a specified $k$-dimensional action space by minimizing the A-optimal objective $\operatorname{tr}[(\Lambda + J)^{-1}]$. Here $\Lambda \succeq 0$ is passive information and $J \succeq 0$ added information, with $\Lambda + J \succ 0$ and $\operatorname{tr}(WJ) \le \mathcal{E}$. For Gaussian sampling, $J = n\Omega$ is the expected information from $n$ probes. When $W \succ 0$ and $\Lambda$ share an eigenbasis, with respective eigenvalues $w_j, \lambda_j$, Proposition 39 gives a water-filling rule (cf. Cover & Thomas, 2006),

$$j_j^\star = \left(c\, w_j^{-1/2} - \lambda_j\right)_+, \qquad \textstyle\sum_j w_j j_j^\star = \mathcal{E}, \tag{18}$$

so direction $j$ receives energy exactly when $\lambda_j < c\, w_j^{-1/2}$. On the $q$-dimensional missing quotient, $\Lambda = 0$ gives $J_\star = \mathcal{E}W^{-1/2}/\operatorname{tr}(W^{1/2})$ and $J_\star/n = \Omega_\star$ (Appendix E.5).

For positive-definite task weights on the missing quotient with $\Lambda = 0$, the same design is optimal exactly when those weights are isotropic (Proposition 40, Appendix E.6) (Gevers & Ljung, 1986; Hjalmarsson, 2005; Wagenmaker et al., 2021).

To implement the regression, one data split estimates the representation and $K, M, G_S, S$. An independent batch probes $\widehat{S}^\perp$ and fits the observed residual in (17) by least squares. Theorem 33 combines process noise, parameter, encoder and subspace errors in the bounds for $A, B$. In exact coordinates with independent stationary resets, the conditional risk splits into projection error and probe variance governed by the probe Gram matrix. The optimal Gram matrix has the same allocation shape as $n\Omega_\star$, including with estimated nuisance parameters (Proposition 34).

The experiments evaluate residual spectral screening with probes of fixed norm and energy allocation with deterministic Gram designs (Section 7, Appendix G.3).

## 7 Experiments

We evaluate coverage diagnosis, energy allocation and the prediction and control value of directed acquisition. *Directed* probes directions with low conditional residual variance, using the full action space when selection is unstable. FullBalanced balances energy across coordinates, RandomFull samples random directions, and Global variants use marginal action geometry. Controlled and robot comparisons match transitions, perturbation energy and fitters through antipodal probes at shared resets. Details appear in Appendices G, H, I and J.

### 7.1 CONTROLLED SYSTEMS

We vary action coverage on the same linear Gaussian systems and learn a predictive encoder from nonlinear observations, inspired by I-JEPA's representation prediction (Assran et al., 2023). Conditional residuals reveal coverage gaps masked by feedback in raw actions. Directed lowers tracking cost under missing, weak and anisotropic coverage (Figure 3a). Selecting equally many directions from marginal geometry increases cost against FullBalanced (Table 2), supporting the importance of their orientation. Isotropic coverage selects no preferred subspace. These comparisons connect conditional coverage to where additional excitation benefits control.

With unequal action costs, weighted allocation within the selected subspace improves action effect estimation over uniform allocation at the same transition count and weighted energy, in both exact and learned coordinates (Tables 3 and 5). With deterministic designs, probe variance decreases inversely with energy; at fixed energy and Gram matrix, increasing transition count leaves risk nearly unchanged. This agrees with the decomposition of estimation error into projection error and probe variance (Proposition 34).

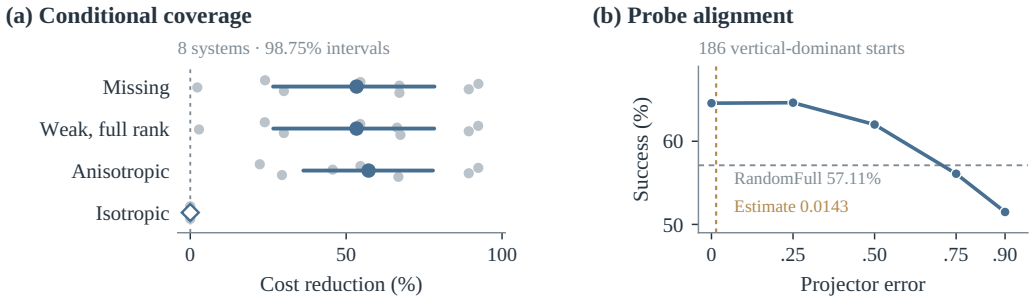



Figure 3: Coverage and probe alignment. Left shows cost reductions against FullBalanced with eight system means and $98.75\%$ bootstrap intervals. Isotropic coverage uses identical data. Right shows direction sensitivity on 186 PickCube starts dominated by vertical motion.

### 7.2 ROBOT MANIPULATION

In `PickCube-v1` (Tao et al., 2024), horizontal behavior probes leave vertical motion unexcited. Directed estimates this direction and improves success and final distance over RandomFull and EnsembleDisagreement (Lakshminarayanan et al., 2017; Sekar et al., 2020) under model predictive control (Rawlings et al., 2017) (Table 1). Its performance approaches that of the oracle direction. Larger direction errors reduce the advantage (Figure 3b), linking acquisition value to alignment with the needed response. GlobalMean also finds a nearly vertical direction here, giving similar control performance.

Table 1: PickCube control over 330 starts, averaged over acquisition budgets of 16, 32 and 64 transitions. Bold identifies Directed. Oracle uses the known vertical direction; uncertainty intervals appear in Table 7 and Figure 5.

| Acquisition | Success (%) ↑ | Final distance (mm) ↓ |
|---|---|---|
| RandomFull | 65.03 | 30.163 |
| GlobalMean | 69.52 | 28.033 |
| EnsembleDisagreement | 60.91 | 31.269 |
| **Directed** | **69.66** | **27.995** |
| Oracle | 69.67 | 27.992 |

### 7.3 RESPONSE PREDICTION AND CONTROL FROM PIXEL OBSERVATIONS

In `cheetah-run` (Tassa et al., 2018), policies share a frozen pixel encoder, matched episode and commanded energy budgets, and a common behavior prior for planning. Directed lowers response error against FullBalanced at both budgets (Figure 4). At ten episodes it also improves return in each

seed; at 30, return gains vary across seeds. EnsembleDisagreement has the lowest mean response error and GlobalGap the highest mean return at ten episodes. These differing rankings separate the benefit of learning action responses from how the planner uses them.

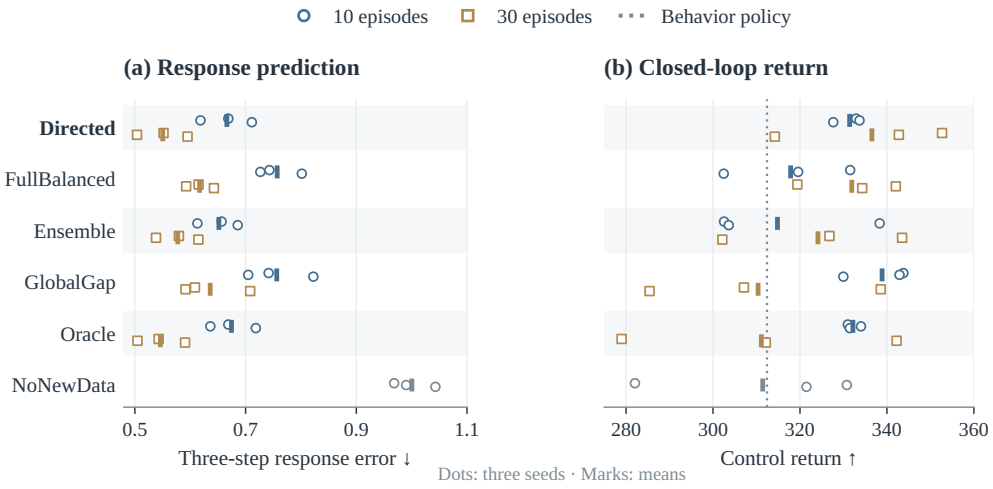



Figure 4: Pixel response prediction and control. Left shows normalized squared three-step response error on the imposed $S^{\perp}$; right shows return. Dots are three seeds; marks are means. The dotted line shows the common behavior policy. NoNewData uses passive data alone.

### 7.4 Action Coverage and Counterfactual Error in Offline Data

On Minari locomotion and Adroit human datasets (Younis et al., 2024; Fu et al., 2020), simulator branches evaluate responses along directions with different excitation. Most locomotion datasets have larger relative errors in less excited directions; response amplitude explains the reverse ordering on Halfcheetah-expert (Table 9). In Adroit, history improves prediction without uniformly improving poorly excited responses (Figure 6). Observation information and action coverage therefore capture different aspects of counterfactual prediction quality.

### 7.5 Performance beyond the Modeling Assumptions

We test nonlinear dynamics and behavior, heavy-tailed noise, state-dependent excitation and acquisition without paired resets. Effects vary at 24 added transitions; Directed lowers mean cost against FullBalanced in all tested conditions at 96 transitions (Table 13). Response interpolation and controller comparisons connect this budget dependence to response estimation error and controller sensitivity (Tables 18 and 16).

## 8 Conclusion

For linear Gaussian world models under partial action coverage, we characterize the affine dynamics compatible with behavior data and the action effects they share. Conditional action residuals expose the remaining ambiguity, including when feedback masks it in raw actions. This characterization determines the interventions needed for identification and leads to optimal energy allocation for action-effect estimation in the Gaussian regression, with error bounds for estimation in learned coordinates.

Controlled experiments connect residual coverage diagnosis to lower control cost through directed acquisition. Robot manipulation and pixel experiments further show improvements in control success and action response prediction. Together, these results support directing additional interaction toward action effects that behavior data leave poorly determined.

### AI USE STATEMENT

We used generative AI tools to assist with drafting, editing, and presentation during manuscript preparation. The authors reviewed and revised all AI-assisted material and take responsibility for the accuracy, claims, and final content of this work.

### ETHICS STATEMENT

This work combines theoretical analysis with experiments in simulated environments and public benchmark datasets. Robot manipulation and pixel control are evaluated in simulation. Experiments with human demonstrations use existing public Adroit datasets (Fu et al., 2020).

### REPRODUCIBILITY STATEMENT

The theoretical setting and assumptions are stated in the paper, with proofs in Appendices A to F. Appendices G, H, I, and J describe the systems and datasets, data splits, models and controllers, acquisition budgets, numbers of seeds, and evaluation procedures. They also specify how results and uncertainty intervals are computed and aggregated.

## REFERENCES

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 372–407, 2023.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Alexandre Sanfelice Bazanella, Michel Gevers, and Ljubiša Mišković. Closed-loop identification of MIMO systems: A new look at identifiability and experiment design. *European Journal of Control*, 16(3):228–239, 2010. doi: 10.3166/ejc.16.228-239.

Xavier Bombois, Gérard Scorletti, Michel Gevers, Paul M. J. Van den Hof, and Roland Hildebrand. Least costly identification experiment for control. *Automatica*, 42(10):1651–1662, 2006. doi: 10.1016/j.automatica.2006.05.016.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In *International Conference on Learning Representations*, 2019.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, 2nd edition, 2006. doi: 10.1002/047174882X.

Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Urban Forssell and Lennart Ljung. Closed-loop identification revisited. *Automatica*, 35(7):1215–1241, 1999. doi: 10.1016/S0005-1098(99)00022-9.

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

Michel Gevers and Lennart Ljung. Optimal experiment designs with respect to the intended model application. *Automatica*, 22(5):543–554, 1986. doi: 10.1016/0005-1098(86)90064-6.

Michel Gevers, Alexandre Sanfelice Bazanella, Xavier Bombois, and Ljubisa Miskovic. Identification and the information matrix: How to get just sufficiently rich? *IEEE Transactions on Automatic Control*, 54(12):2828–2840, 2009. doi: 10.1109/TAC.2009.2034199.

Ivar Gustavsson, Lennart Ljung, and Torsten Söderström. Identification of processes in closed loop—identifiability and accuracy aspects. *Automatica*, 13(1):59–75, 1977. doi: 10.1016/0005-1098(77)90009-7.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870, 2018. URL https://proceedings.mlr.press/v80/haarnoja18b.html.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565, 2019.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025. doi: 10.1038/s41586-025-08744-2.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024.

Håkan Hjalmarsson. From experiment design to closed-loop control. *Automatica*, 41(3):393–438, 2005. doi: 10.1016/j.automatica.2004.11.021.

Aapo Hyvärinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217, 2020.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. MOReL: Model-based offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21810–21823, 2020.

Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E. Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pp. 428–484, 2022.

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

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

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Stratis Gavves. CITRIS: Causal identifiability from temporal intervened sequences. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 13557–13603, 2022.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 4114–4124, 2019.

Horia Mania, Michael I. Jordan, and Benjamin Recht. Active learning for nonlinear system identification with guarantees. *Journal of Machine Learning Research*, 23(32):1–30, 2022.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer Series in Statistics. Springer, New York, 2003. doi: 10.1007/b97478.

Jaouad Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *The Annals of Statistics*, 50(4):2157–2178, 2022. doi: 10.1214/22-AOS2181.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2778–2787, 2017.

Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5062–5071, 2019.

Friedrich Pukelsheim. *Optimal Design of Experiments*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2006. doi: 10.1137/1.9780898719109.

James B. Rawlings, David Q. Mayne, and Moritz M. Diehl. *Model Predictive Control: Theory, Computation, and Design*. Nob Hill Publishing, 2nd edition, 2017.

Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8583–8592, 2020.

Pranav Shyam, Wojciech Jaśkowski, and Faustino Gomez. Model-based active exploration. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5779–5788, 2019.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I. Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 439–473, 2018.

Richard S. Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, 2(4):160–163, 1991. doi: 10.1145/122344.122377.

Elie Tamer. Partial identification in econometrics. *Annual Review of Economics*, 2(1):167–195, 2010. doi: 10.1146/annurev.economics.050708.143401.

Stone Tao, Fanbo Xiang, Arth Shukla, Yuzhe Qin, Xander Hinrichsen, Xiaodi Yuan, Chen Bao, Xinsong Lin, Yulin Liu, Tse kai Chan, Yuan Gao, Xuanlin Li, Tongzhou Mu, Nan Xiao, Arnav Gurha, Viswesh Nagaswamy Rajesh, Yong Woo Choi, Yen-Ru Chen, Zhiao Huang, Roberto Calandra, Rui Chen, Shan Luo, and Hao Su. ManiSkill3: GPU parallelized robotics simulation and rendering for generalizable embodied AI, 2024. URL https://arxiv.org/abs/2410.00425.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. DeepMind control suite, 2018. URL https://arxiv.org/abs/1801.00690.

Henk J. van Waarde, Jaap Eising, Harry L. Trentelman, and M. Kanat Camlibel. Data informativity: A new perspective on data-driven analysis and control. *IEEE Transactions on Automatic Control*, 65(11):4753–4768, 2020. doi: 10.1109/TAC.2020.2966717.

Andrew Wagenmaker and Kevin Jamieson. Active learning for identification of linear dynamical systems. In *Proceedings of the 33rd Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3487–3582, 2020.

Andrew Wagenmaker, Max Simchowitz, and Kevin Jamieson. Task-optimal exploration in linear dynamical systems. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10641–10652, 2021.

Jan C. Willems, Paolo Rapisarda, Ivan Markovsky, and Bart L. M. De Moor. A note on persistency of excitation. *Systems & Control Letters*, 54(4):325–329, 2005. doi: 10.1016/j.sysconle.2004.09.003.

Omar G. Younis, Rodrigo Perez-Vicente, John U. Balis, Will Dudley, Alex Davey, and Jordan K. Terry. Minari. Zenodo, 2024. URL https://doi.org/10.5281/zenodo.13767625.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. MOPO: Model-based offline policy optimization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 14129–14142, 2020.

Xiangteng Zhang, Yang Guan, Bo Zhang, Hongyang Li, Ya-Qin Zhang, and Shengbo Eben Li. On the identifiability of controlled world models, 2026. URL https://arxiv.org/abs/2607.22430.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79115–79135, 2025.

Ingvar Ziemann, Anastasios Tsiamis, Bruce Lee, Yassir Jedra, Nikolai Matni, and George J. Pappas. A tutorial on the non-asymptotic theory of system identification. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pp. 8921–8939, 2023. doi: 10.1109/CDC49753.2023.10383489.

## A Representation recovery

We prove state recovery and its quantitative form through the Hermite spectrum of the Gaussian prediction channel.

*Remark* 7 (Spectral condition in whitened coordinates). In the stationary coordinates of Section 3, write $Q_\xi = \mathrm{Cov}(\Xi)$ and let $a = \lambda_{\min}(Q_\xi)$ and $b = \lambda_{\max}(Q_\xi)$. Since $R = I_d - Q_\xi$,

$$\gamma_{\mathrm{rep}} = (1-b) - (1-a)^2 = 2a - a^2 - b, \qquad \gamma_{\mathrm{rep}} > 0 \iff b < 2a - a^2.$$

This compares prediction errors directly. For centered features of unit variance, the largest optimal prediction error among linear state features is $b$. The smallest such error among features orthogonal to every linear state feature is $2a - a^2$ (Proposition 10). Both errors refer to predicting the next feature from the current state and action. Their strict separation favors preserving every linear state direction in the prediction objective.

The condition implies $0 < a \le b < 1$. For a fixed smallest noise variance $a$, it bounds the largest by $2a - a^2$. The covariance here is measured after whitening the stationary state. Moreover, with $M = A + BK$,

$$R = MM^\top + B\Sigma B^\top, \qquad R \succ 0 \iff \mathrm{range}(M) + \mathrm{range}(B\Sigma^{1/2}) = \mathbb{R}^d.$$

Thus every state direction receives predictable signal through the closed-loop dynamics or the excited actions. The spectral inequality additionally orders the linear and nonlinear prediction errors.

**Example 8** (Isotropic process noise)**.** If $Q_\xi = \sigma^2 I_d$ with $0 < \sigma^2 < 1$, stationarity gives $R = (1-\sigma^2) I_d$ and $\gamma_{\mathrm{rep}} = \sigma^2(1-\sigma^2) > 0$. For $\sigma^2 = 1/2$, every linear feature of unit variance has optimal prediction error $1/2$, while the best centered feature of unit variance orthogonal to all linear features has error $3/4$. The separation follows from the isotropic noise covariance and is compatible with singular action coverage.

**Lemma 9** (Controlled Hermite separation)**.** *Let $Y = C + N \sim \mathcal{N}(0, I_d)$, where $C \sim \mathcal{N}(0, R)$, $N \sim \mathcal{N}(0, I_d - R)$, and $C \perp N$. Suppose*

$$\lambda_{\min}(R) > \lambda_{\max}(R)^2.$$

*For every measurable $\phi$ satisfying $\mathbb{E}\phi(Y) = 0$ and $\mathrm{Cov}(\phi(Y)) = I_d$,*

$$\mathbb{E}\big\|\mathbb{E}[\phi(Y) \mid C]\big\|_2^2 \le \mathrm{tr}\, R. \tag{19}$$

*Equality holds precisely for $\phi(y) = Qy$ almost surely with $Q$ orthogonal. Quantitatively, if the left side of* (19) *is at least $\mathrm{tr}\, R - \varepsilon$, then some orthogonal $Q$ satisfies*

$$\mathbb{E}\|\phi(Y) - QY\|_2^2 \le \frac{2\varepsilon}{\gamma_{\mathrm{rep}}}. \tag{20}$$

*Proof.* Because $\phi(Y)$ is centered, its multivariate Hermite expansion under standard Gaussian measure has no degree-zero component. Write the orthogonal decomposition as

$$\begin{aligned}\phi &= \phi_1 + \phi_2 + \cdots, \\ \phi_1(y) &= Wy, \\ W &:= \mathbb{E}[\phi(Y)Y^\top].\end{aligned}$$

The whitening constraint yields

$$\sum_{n\geq 1} \|\phi_n\|_{L^2}^2 = d, \qquad \|\phi_1\|_{L^2}^2 = \|W\|_F^2. \tag{21}$$

The strict spectral premise makes $R$ positive definite. Set $X = R^{-1/2}C$, so $X \sim \mathcal{N}(0, I_d)$ and $\sigma(X) = \sigma(C)$. In an eigenbasis of $R$, the cross-covariance between $Y$ and $X$ is

$$\operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_d}),$$

where $\lambda_j$ are the eigenvalues of $R$. The Gaussian conditional-expectation operator maps a normalized Hermite monomial of multi-index $\alpha$ to the corresponding monomial multiplied by

$$\prod_{j=1}^{d} \lambda_j^{\alpha_j/2}.$$

Different total degrees remain orthogonal after conditioning. The squared norm of the degree-one contribution is

$$\mathbb{E}\|\mathbb{E}[WY \mid C]\|_2^2 = \mathbb{E}\|WC\|_2^2 = \operatorname{tr}(WRW^\top). \tag{22}$$

Let $\rho = \lambda_{\max}(R)$. For every total degree $n \geq 2$, the squared operator norm is at most $\rho^n \leq \rho^2$, since stationarity gives $\rho \leq 1$. Using (21),

$$\mathbb{E}\big\|\mathbb{E}[\phi(Y) \mid C]\big\|_2^2 \leq \operatorname{tr}(WRW^\top) + \rho^2(d - \|W\|_F^2). \tag{23}$$

The joint covariance of $(\phi(Y), Y)$ is positive semidefinite and has identity diagonal blocks. Hence every singular value of $W$ is at most one, so $0 \preceq W^\top W \preceq I_d$. Define

$$H := R - \rho^2 I_d \succeq \gamma_{\mathrm{rep}} I_d \succ 0.$$

The right side of (23) equals

$$\begin{aligned}\rho^2 d + \operatorname{tr}(HW^\top W) &\leq \rho^2 d + \operatorname{tr} H &(24)\\ &= \operatorname{tr} R, &(25)\end{aligned}$$

which proves (19).

It remains to characterize equality. The gap in (25) is

$$\operatorname{tr}\big(H(I_d - W^\top W)\big) \geq \gamma_{\mathrm{rep}} \operatorname{tr}(I_d - W^\top W). \tag{26}$$

If (19) is an equality, every intervening upper bound is an equality, so (26) vanishes. Positivity of $H$ gives $W^\top W = I_d$. Thus $W$ is orthogonal and $\|W\|_F^2 = d$. Equation (21) then forces every Hermite component of degree at least two to vanish. Therefore $\phi(Y) = WY$ almost surely. Conversely, an orthogonal $Q$ gives

$$\mathbb{E}\|\mathbb{E}[QY \mid C]\|_2^2 = \mathbb{E}\|QC\|_2^2 = \operatorname{tr} R.$$

For the quantitative claim, energy at least $\operatorname{tr} R - \varepsilon$ and (26) give

$$d - \|W\|_F^2 \leq \varepsilon/\gamma_{\mathrm{rep}}.$$

All singular values $s_j(W)$ lie in $[0, 1]$. If $Q$ is an orthogonal polar factor of $W$ (completed arbitrarily on a zero singular subspace), then

$$\|W - Q\|_F^2 = \sum_j (1 - s_j(W))^2 \leq \sum_j (1 - s_j(W)^2) = d - \|W\|_F^2.$$

The higher-order Hermite part $\phi - WY$ is orthogonal to every linear function, so

$$\mathbb{E}\|\phi(Y) - QY\|_2^2 = (d - \|W\|_F^2) + \|W - Q\|_F^2 \leq 2\varepsilon/\gamma_{\mathrm{rep}}.$$

□

**Proposition 10** (Exact Hermite band separation)**.** *In the channel of Lemma 9, let $Tf = \mathbb{E}[f(Y) \mid C]$. On the degree-one Hermite subspace, the eigenvalues of $T^*T$ are $\lambda_1(R), \dots, \lambda_d(R)$. On the orthogonal sum of all degrees at least two, its operator norm is $\rho^2$. Hence*

$$\inf_{f \in \mathcal{H}_1,\ \|f\|_2=1} \|Tf\|_2^2 - \sup_{f \in \mathcal{H}_{\geq 2},\ \|f\|_2=1} \|Tf\|_2^2 = \gamma_{\mathrm{rep}}.$$

*Thus $\gamma_{\mathrm{rep}} > 0$ is necessary and sufficient for strict uniform separation of every linear Hermite mode from every nonlinear Hermite mode. At equality the two spectral bands touch, so no positive stability constant based on degree separation exists.*

*Proof of Proposition 10.* Diagonalize $R$ and standardize $C$ as in Lemma 9. For a normalized Hermite monomial with multi-index $\alpha \neq 0$, the squared singular value of $T$ is

$$\lambda^\alpha := \prod_{j=1}^{d} \lambda_j(R)^{\alpha_j}.$$

For $|\alpha| = 1$ these values are exactly the eigenvalues of $R$. For $|\alpha| \geq 2$, stationarity gives $0 \leq \lambda_j(R) \leq \rho \leq 1$, whence $\lambda^\alpha \leq \rho^{|\alpha|} \leq \rho^2$. Equality is attained by the degree-two monomial in an eigen-coordinate having eigenvalue $\rho$ (using the normalized second Hermite polynomial). Thus the two extremal values are $\lambda_{\min}(R)$ and $\rho^2$, and their difference is precisely $\gamma_{\mathrm{rep}}$. If the difference is nonpositive, the two invariant Hermite subspaces have no strict ordering; if it is positive, the displayed gap is uniform and exact. □

*Proof of Theorem 1.* For an admissible encoder, let $\phi = h \circ g$. The inputs $Z$ and $Z^+$ have the same standard Gaussian law, while stationarity and admissibility give $\mathbb{E}\phi(Z) = \mathbb{E}\phi(Z^+) = 0$ and $\mathrm{Cov}(\phi(Z)) = \mathrm{Cov}(\phi(Z^+)) = I_d$. No Gaussian law is imposed on either encoder output. Projection onto all measurable functions of $(\phi(Z), U)$ gives, for every predictor $F$,

$$\mathcal{L}(h, F) \geq d - \mathbb{E}\big\|\mathbb{E}[\phi(Z^+) \mid \phi(Z), U]\big\|_2^2, \tag{27}$$

with equality when that conditional mean lies in $\mathcal{F}$; the argument uses only this inequality and the affine witnesses in (6).

Conditional expectation is an $L^2$ projection. Because

$$\sigma(\phi(Z), U) \subseteq \sigma(Z, U),$$

monotonicity of projection norms gives

$$\mathbb{E}\|\mathbb{E}[\phi(Z^+) \mid \phi(Z), U]\|_2^2 \leq \mathbb{E}\|\mathbb{E}[\phi(Z^+) \mid Z, U]\|_2^2. \tag{28}$$

Given $(Z, U)$, the next state is $C + \Xi$, and its conditional distribution depends on $(Z, U)$ through $C = AZ + BU$. Therefore the right side is

$$\mathbb{E}\|\mathbb{E}[\phi(Z^+) \mid C]\|_2^2 \leq \mathrm{tr}\, R$$

by Lemma 9. Every admissible encoder–predictor pair therefore has loss at least $d - \mathrm{tr}\, R$.

For any orthogonal $Q$, choose $\phi(z) = Qz$ and the realizable predictor $F_Q$ from (6). Then

$$F_Q(QZ, U) = Q(AZ + BU),$$

and its loss is $\mathbb{E}\|Q\Xi\|^2 = d - \mathrm{tr}\, R$. Hence a global minimizer attains equality in Lemma 9. The equality characterization gives $h(g(z)) = Qz$ almost surely. No invertibility property of $\Sigma$ was used. □

**Theorem 11** (Quantitative representation stability)**.** *Under the Gaussian channel and margin assumptions of Theorem 1, let $h$ be admissible and let $F$ be any measurable predictor. If, for some $\varepsilon \geq 0$,*

$$\mathcal{L}(h, F) \leq d - \mathrm{tr}\, R + \varepsilon, \tag{29}$$

*then an orthogonal $Q$ exists such that*

$$\mathbb{E}\|h(g(Z)) - QZ\|_2^2 \leq \frac{2\varepsilon}{\gamma_{\mathrm{rep}}}. \tag{30}$$

*Proof of Theorem 11.* Let

$$V_\phi := \mathbb{E}\|\mathbb{E}[\phi(Z^+) \mid \phi(Z), U]\|_2^2.$$

Equations (27) and (29) imply $V_\phi \geq \operatorname{tr} R - \varepsilon$. Projection monotonicity then implies that the conditional energy given $C$ is also at least $\operatorname{tr} R - \varepsilon$. Apply the quantitative part of Lemma 9 to $Y = Z^+$. Since $Z$ and $Z^+$ have the same standard-Gaussian law, its conclusion is exactly (30) for $h(g(Z)) = \phi(Z)$. □

### A.1 Whitening and empirical optimization

**Proposition 12** (Approximate post-whitening and empirical optimization)**.** *Let $\varphi(X)$ be centered with covariance $G$ satisfying $\|G - I_d\|_{\mathrm{op}} \leq \delta < 1$, and put $\bar{\varphi} = G^{-1/2}\varphi$. If $\bar{\varphi}$ and a predictor satisfy* (29)*, then for some orthogonal $Q$,*

$$\{\mathbb{E}\|\varphi(X) - QZ\|_2^2\}^{1/2} \leq \sqrt{\frac{2\varepsilon}{\gamma_{\mathrm{rep}}}} + \sqrt{d}\{1 - \sqrt{1-\delta}\}. \tag{31}$$

**Proposition 13** (Empirical optimization)**.** *Let $\widehat{\mathcal{L}}_N$ be an empirical objective on any admissible encoder–predictor class containing an affine target witness. If on an event $\sup_{(h,F)} |\widehat{\mathcal{L}}_N(h,F) - \mathcal{L}(h,F)| \leq a_N$ and an optimization procedure returns $(\widehat{h}, \widehat{F})$ within $\varepsilon_{\mathrm{opt}}$ of the empirical infimum, then on that event*

$$\inf_{Q^\top Q = I_d} \mathbb{E}\|\widehat{h}(X) - QZ\|_2^2 \leq \frac{2(2a_N + \varepsilon_{\mathrm{opt}})}{\gamma_{\mathrm{rep}}}. \tag{32}$$

*Proof of Propositions 12 and 13.* The post-whitened map $\bar{\varphi}$ is centered with identity covariance, so Theorem 11 gives an orthogonal $Q$ with

$$\|\bar{\varphi}(X) - QZ\|_{L^2} \leq \sqrt{2\varepsilon/\gamma_{\mathrm{rep}}}.$$

Since $\varphi = G^{1/2}\bar{\varphi}$ and $\mathrm{Cov}(\bar{\varphi}) = I_d$,

$$\|\varphi - \bar{\varphi}\|_{L^2} = \|G^{1/2} - I_d\|_F \leq \sqrt{d}\{1 - \sqrt{1-\delta}\}.$$

The $L^2$ triangle inequality proves (31).

For the empirical assertion, write $\mathcal{L}_\star = d - \operatorname{tr} R$. The affine witness attains $\mathcal{L}_\star$, while the projection argument shows that no admissible pair has smaller population loss. Uniform deviation and empirical $\varepsilon_{\mathrm{opt}}$-suboptimality give

$$\mathcal{L}(\widehat{h}, \widehat{F}) \leq \widehat{\mathcal{L}}_N(\widehat{h}, \widehat{F}) + a_N \leq \mathcal{L}_\star + 2a_N + \varepsilon_{\mathrm{opt}}.$$

Apply Theorem 11 with $\varepsilon = 2a_N + \varepsilon_{\mathrm{opt}}$. □

## B Geometry of the Affine Identified Set

**Lemma 14** (Gaussian-support annihilator)**.** *For $\Delta_A \in \mathbb{R}^{d\times d}$ and $\Delta_B \in \mathbb{R}^{d\times p}$, the following three statements are equivalent.*

*(i)* $\Delta_A Z + \Delta_B U = 0$ *almost surely;*
*(ii)* $\Delta_A z + \Delta_B u = 0$ *on every point of* $\mathrm{supp}(Z, U)$*;*
*(iii)* $\Delta_A = -\Delta_B K$ *and* $\Delta_B P_S = 0$*.*

*Proof of Lemma 14.* A centered Gaussian with covariance $\Sigma$ has topological support $S = \mathrm{range}(\Sigma)$. Independence of $Z$ and $E$, followed by the continuous map

$$(z, e) \longmapsto (z, Kz + e),$$

gives

$$\mathrm{supp}(Z, U) = \{(z, Kz + e) : z \in \mathbb{R}^d,\ e \in S\}.$$

Consider the continuous linear map

$$f(z, u) := \Delta_A z + \Delta_B u.$$

If it vanishes almost surely, it vanishes everywhere on the support. Indeed, if $f$ were nonzero at a support point, continuity would give a neighborhood on which it remains nonzero; every neighborhood of a support point has positive probability, contradicting almost-sure vanishing. The reverse direction is immediate. This proves the equivalence of statements (i) and (ii).

On a support point $(z, Kz + e)$,

$$f(z, Kz + e) = (\Delta_A + \Delta_B K)z + \Delta_B e. \tag{33}$$

If (33) vanishes for all $z$ and $e \in S$, setting $e = 0$ gives

$$\Delta_A = -\Delta_B K.$$

Setting $z = 0$ gives $\Delta_B e = 0$ for every $e \in S$, which is equivalent to

$$\Delta_B P_S = 0$$

because the range of $P_S$ is $S$. These equations also make (33) vanish for every support point. This proves the equivalence with statement (iii). □

*Proof of the fiber formula and invariants.* Let $(A', B')$ be another affine pair and define

$$\Delta_A := A' - A, \qquad \Delta_B := B' - B.$$

Equality of the two conditional-mean random variables is

$$\Delta_A Z + \Delta_B U = 0 \quad \text{almost surely.}$$

Lemma 14, with $D = \Delta_B$, gives

$$DP_S = 0, \qquad \Delta_A = -DK.$$

Thus every equivalent pair has the form $(A - DK, B + D)$. The converse follows either from the same lemma or by the direct cancellation calculation in Section 5. This proves (10).

For every member of the fiber,

$$(A - DK) + (B + D)K = A + BK, \tag{34}$$

$$(B + D)P_S = BP_S. \tag{35}$$

Conversely, suppose $(A', B')$ satisfies the two invariant equalities. Setting $D = B' - B$, equation (35) gives $DP_S = 0$, while (34) gives $A' = A - DK$. Hence equality of the invariants characterizes membership in the same fiber. □

Every fiber displacement preserves the behavior process itself. Since $DP_S = 0$ and $U = KZ + E$ with $E \in S$ almost surely,

$$(A - DK)Z + (B + D)U = AZ + BU.$$

Adding the same process noise gives the same next state, so the stationary law, predictable covariance and representation margin are preserved.

*Proof of the dimension statement.* Because $P_S$ is an orthogonal projector,

$$DP_S = 0 \quad \Longleftrightarrow \quad D = D(I_p - P_S).$$

Thus $D$ is determined by an arbitrary linear map from $S^\perp$, a space of dimension $p - r$, to $\mathbb{R}^d$. The vector space of such maps has dimension $d(p - r)$.

Equivalently, choose orthogonal matrices in the action space so that the first $r$ basis vectors span $S$. The condition $DP_S = 0$ sets the first $r$ columns of $D$ to zero and leaves $p - r$ arbitrary columns, each in $\mathbb{R}^d$. Finally, the fiber parameterization is injective, since equality of the $B$-components $B + D_1 = B + D_2$ implies $D_1 = D_2$. The affine fiber therefore has the same dimension as its direction space, namely $d(p - r)$. □

**Proposition 15** (Arbitrary unexcited restriction)**.** *For every linear map $T : S^\perp \to \mathbb{R}^d$, the fiber contains a unique displacement $D$ whose restriction to $S$ is zero and whose restriction to $S^\perp$ equals $T - B|_{S^\perp}$. The resulting $B + D$ restricts to $T$ on $S^\perp$.*

*Proof.* Use the orthogonal decomposition $\mathbb{R}^p = S \oplus S^\perp$ and define

$$Ds = 0 \quad (s \in S), \qquad Dw = T(w) - Bw \quad (w \in S^\perp).$$

This defines a unique linear map on $\mathbb{R}^p$. It satisfies $DP_S = 0$, so $(A - DK, B + D)$ lies in the fiber, and its restriction to $S^\perp$ is $T$. Uniqueness follows from the two prescribed restrictions. □

Changing an unexcited column of $B$ changes $A$ by the corresponding feedback contribution. These changes cancel on the behavior support. At $r = p$, the constraint forces $D = 0$; at $r = 0$, every $D$ is admissible and the fiber has dimension $dp$.

Under $Y = QZ$, write $\widetilde{A} = QAQ^\top$, $\widetilde{B} = QB$ and $\widetilde{K} = KQ^\top$. The displacement becomes $\widetilde{D} = QD$, satisfying $\widetilde{D}P_S = 0$ and producing $(\widetilde{A} - \widetilde{D}\widetilde{K}, \widetilde{B} + \widetilde{D})$. Thus the fiber is invariant under the recovered state rotation.

### B.1 Counterfactual disagreement

For a fiber displacement $D$, the difference between the two affine means at an arbitrary pair $(z, u)$ is

$$(-DK)z + Du = D(u - Kz). \tag{36}$$

On behavior data, $u - Kz = e \in S$ and the expression vanishes. At a counterfactual residual $e_\perp \in S^\perp$, the disagreement is $De_\perp$, which may be any vector in the range of $D$. Equation (36) locates the missing information in action residuals outside the explored subspace.

### B.2 An Example with Three Action Coordinates

An example with three actions makes the split concrete. Let $p = 3$ with $S = \mathrm{span}(e_1)$, write $B = [b_1\ b_2\ b_3]$, and let the rows of $K$ be $k_1^\top, k_2^\top, k_3^\top$. Every fiber direction is $D = [0\ d_2\ d_3]$, so the equivalent parameters are $A' = A - d_2k_2^\top - d_3k_3^\top$ and $B' = [b_1\ \ b_2 + d_2\ \ b_3 + d_3]$. The data pin down $b_1$ and $A + b_1k_1^\top + b_2k_2^\top + b_3k_3^\top$, while $b_2$ and $b_3$ move freely inside a $2d$-dimensional fiber. An executed action residual along $e_2$ or $e_3$ breaks the corresponding cancellation.

### B.3 General behavior laws

**Proposition 16** (Distribution-free transition geometry)**.** *The fiber, invariants, dimension, and span criterion do not require Gaussian state or action residuals. They remain exact for any law $U = KZ + E$ satisfying $\mathrm{supp}(Z, E) = \mathbb{R}^d \times S$. If in addition $Z$ and $E$ are centered and uncorrelated, $P = \mathrm{Cov}(Z) \succ 0$, and $\mathrm{Cov}(E) = \Sigma$ with range $S$, then the anchored excess risk is*

$$\mathrm{tr}\{HPH^\top\} + \mathrm{tr}(\Delta_B\Sigma\Delta_B^\top), \qquad H = \Delta_A + \Delta_B K. \tag{37}$$

*Proof of Proposition 16.* The support assumption directly gives

$$\mathrm{supp}(Z, U) = \{(z, Kz + e) : z \in \mathbb{R}^d,\ e \in S\}.$$

The continuity argument in Lemma 14 therefore applies without a Gaussian law and yields the same two annihilator equations. All fiber, invariant, dimension, and augmented-span conclusions are algebraic consequences of those equations. For the risk statement, $\Delta_A Z + \Delta_B U = HZ + \Delta_B E$. Centering and $\mathbb{E}[ZE^\top] = 0$ eliminate the cross term, leaving (37). □

## C Stability and finite sample recovery

### C.1 Prediction risk and distance to the fiber

*Proof of Theorem 4.* Write $Y = AZ + BU + \Xi$. Since $\Xi$ is centered and independent of $(Z, U)$, the Pythagorean identity for conditional expectation gives

$$\mathcal{R}(A', B') - \mathcal{R}(A, B) = \mathbb{E}\|\Delta_A Z + \Delta_B U\|_2^2.$$

Using $U = KZ + E$, the expression inside the norm is $HZ + \Delta_B E$. The two summands are centered and independent, so

$$\mathbb{E}\|HZ + \Delta_B E\|_2^2 = \mathrm{tr}(HH^\top) + \mathrm{tr}(\Delta_B \Sigma \Delta_B^\top),$$

which proves (12). The first inequality in (13) follows immediately. Moreover, $\Sigma \succeq \lambda_+ P_S$, and hence

$$\mathrm{tr}(\Delta_B \Sigma \Delta_B^\top) \geq \lambda_+ \|\Delta_B P_S\|_F^2.$$

Fix $(z, e)$. If $e \in S$, put $G = \Delta_B \Sigma^{1/2}$. Then $\Delta_B e = G\Sigma^{\dagger/2} e$, and the constraint is $\|H\|_F^2 + \|G\|_F^2 \leq \varepsilon$. The evaluation map from the product Frobenius space to $\mathbb{R}^d$ has norm

$$c(z, e) = \sqrt{\|z\|_2^2 + \|\Sigma^{\dagger/2} e\|_2^2}.$$

Cauchy–Schwarz gives the upper bound $\sqrt{\varepsilon} c(z, e)$. If $c(z, e) > 0$, equality is attained by choosing a unit $a \in \mathbb{R}^d$ and the rank-one pair

$$H = \frac{\sqrt{\varepsilon}}{c(z, e)} a z^\top, \qquad G = \frac{\sqrt{\varepsilon}}{c(z, e)} a (\Sigma^{\dagger/2} e)^\top.$$

Set $\Delta_B = G\Sigma^{\dagger/2}$ on $S$, zero on $S^\perp$, and $\Delta_A = H - \Delta_B K$. The case $c(z, e) = 0$ is immediate.

If $e \notin S$, let $w = P_{S^\perp} e \neq 0$ and choose a unit $a \in \mathbb{R}^d$. For $t > 0$ set $\Delta_B = t a w^\top$ and $\Delta_A = -\Delta_B K$. Then $H = 0$, $\Delta_B \Sigma = 0$, and the excess risk is zero, whereas the disagreement is $t a\, w^\top e = t a \|w\|_2^2$. Letting $t \to \infty$ proves the second branch, including when $\varepsilon = 0$. □

**Corollary 17** (Distance to the exact fiber)**.** *Under the premises of Theorem 4, let $D = \Delta_B(I_p - P_S)$. Then $(A - DK, B + D)$ is in the exact fiber through $(A, B)$, and excess risk at most $\varepsilon$ gives*

$$\|B' - (B + D)\|_F \leq \sqrt{\varepsilon / \lambda_+}, \tag{38}$$

$$\|A' - (A - DK)\|_F \leq \sqrt{\varepsilon}\big(1 + \|K\|_{\mathrm{op}} / \sqrt{\lambda_+}\big). \tag{39}$$

*Proof of Corollary 17.* The matrix $D = \Delta_B(I - P_S)$ annihilates $S$, so $(A - DK, B + D)$ belongs to the fiber. Also

$$B' - (B + D) = \Delta_B P_S$$

and, using $H = \Delta_A + \Delta_B K$,

$$A' - (A - DK) = H - \Delta_B P_S K.$$

Apply (13) and the Frobenius/operator product inequality to obtain (38)–(39). □

### C.2 Conditional action coverage

*Proof of Proposition 5.* Since $E \perp Z$ and $Z \sim \mathcal{N}(0, I_d)$, the marginal action covariance is $\Sigma_U = KK^\top + \Sigma$.

For part 1, fix $\Sigma_U \succ 0$ and a subspace $S$ of dimension $r$, and let $V \in \mathbb{R}^{p \times r}$ have orthonormal columns spanning $S$. Because $\Sigma_U^{-1} \succ 0$ and $V$ has full column rank, $V^\top \Sigma_U^{-1} V \succ 0$, so

$$\Sigma := V\big(V^\top \Sigma_U^{-1} V\big)^{-1} V^\top$$

is well defined, positive semidefinite, and has range exactly $S$. Put $T = \Sigma_U^{-1/2}$ and $W = TV$. Then $V^\top \Sigma_U^{-1} V = W^\top W$ and

$$T\Sigma T = W(W^\top W)^{-1} W^\top,$$

which is the orthogonal projection onto $\mathrm{range}(W)$, of rank $r$. Hence

$$\Sigma_U - \Sigma = \Sigma_U^{1/2}\big(I_p - T\Sigma T\big)\Sigma_U^{1/2}$$

is positive semidefinite with rank exactly $p - r$. Since $d \geq p - r$, factor $\Sigma_U - \Sigma = LL^\top$ with $L \in \mathbb{R}^{p \times (p-r)}$ and set $K = [\, L \;\; 0\,] \in \mathbb{R}^{p \times d}$, so that $KK^\top + \Sigma = \Sigma_U$. The pair $(K, \Sigma)$ realizes the prescribed marginal covariance with the prescribed excited subspace. Since $S$ was arbitrary, two

behavior policies with different excited subspaces can share the same action law, so no functional of that law can determine $S$.

For part 2, both summands of $\Sigma_U$ are positive semidefinite, hence $x^\top \Sigma_U x = 0$ if and only if $K^\top x = 0$ and $x^\top \Sigma x = 0$. As $\Sigma$ is symmetric with range $S$, its kernel is $S^\perp$, so $\ker(\Sigma_U) = \ker(K^\top) \cap S^\perp$. Subadditivity of rank gives $\operatorname{rank} \Sigma_U \le \operatorname{rank}(KK^\top) + \operatorname{rank} \Sigma \le d + r < p$, so $\dim \ker(\Sigma_U) \ge p - d - r > 0$. □

### C.3 Exact recovery of the residual span

**Proposition 18** (Genericity in the exact model). *Let $\mathbf{Z} = [Z_1, \dots, Z_n]$ and $\mathbf{U} = [U_1, \dots, U_n]$ be formed from the sampling design of Section 3, in exact state coordinates and with $\operatorname{rank} \Sigma = r$. If $n \ge d + r$ then, almost surely, $\mathbf{Z}\mathbf{Z}^\top \succ 0$ and the ordinary least-squares residual spans the excited subspace,*

$$\operatorname{range}\big(\mathbf{U}(I_n - \Pi)\big) = S, \qquad \Pi = \mathbf{Z}^\top (\mathbf{Z}\mathbf{Z}^\top)^{-1} \mathbf{Z}.$$

*Proof of Proposition 18.* Write $\mathbf{E} = [E_1, \dots, E_n]$, so that $\mathbf{U} = K\mathbf{Z} + \mathbf{E}$. The rows of $\mathbf{Z}$ are $n \ge d$ independent standard Gaussian vectors in $\mathbb{R}^n$, so $\mathbf{Z}\mathbf{Z}^\top \succ 0$ almost surely and $\Pi$ is the orthogonal projection of $\mathbb{R}^n$ onto the row space of $\mathbf{Z}$. Then $\mathbf{Z}(I_n - \Pi) = 0$ and therefore

$$\mathbf{U}(I_n - \Pi) = \mathbf{E}(I_n - \Pi).$$

Factor $\Sigma = V \Lambda V^\top$ with $V \in \mathbb{R}^{p \times r}$ orthonormal and $\Lambda \succ 0$ of size $r$. Since each $E_i$ is a centered Gaussian with covariance $\Sigma$, we may write $\mathbf{E} = V \Lambda^{1/2} \mathbf{G}$ with $\mathbf{G} \in \mathbb{R}^{r \times n}$ having independent standard normal entries; the sampling design makes $\mathbf{G}$ independent of $\mathbf{Z}$. Let $I_n - \Pi = OO^\top$ with $O \in \mathbb{R}^{n \times (n-d)}$ orthonormal. Conditional on $\mathbf{Z}$, rotational invariance gives that $\mathbf{G}O \in \mathbb{R}^{r \times (n-d)}$ again has independent standard normal entries, and $n - d \ge r$ implies $\operatorname{rank}(\mathbf{G}O) = r$ almost surely. Hence

$$\operatorname{range}\big(\mathbf{E}(I_n - \Pi)\big) = V \operatorname{range}\big(\Lambda^{1/2} \mathbf{G} O\big) = V \mathbb{R}^r = S,$$

because $\Lambda^{1/2}$ is invertible. Averaging over $\mathbf{Z}$ preserves the almost-sure conclusion. □

### C.4 Moment estimators and covariance bounds

The identifiable objects themselves can be estimated directly from moments. Let $Y = Z^+$ and set

$$\begin{aligned} M &:= A + BK = \mathbb{E}[Y Z^\top], \\ H_B &:= B\Sigma = \mathbb{E}[Y U^\top] - MK^\top. \end{aligned} \tag{40}$$

For samples, write $\widehat{C}_{VW} = n^{-1} \sum_i V_i W_i^\top$ and define

$$\begin{aligned} \widehat{K} &= \widehat{C}_{UZ}, & \widehat{M} &= \widehat{C}_{YZ}, \\ \widehat{\Sigma} &= \widehat{C}_{UU} - \widehat{K}\widehat{K}^\top, & \widehat{H}_B &= \widehat{C}_{YU} - \widehat{M}\widehat{K}^\top. \end{aligned} \tag{41}$$

For $W = (Z^\top, U^\top, Y^\top)^\top$, set $m = 2d + p$. With $\Gamma = \operatorname{Cov}(W)$ and $L = \|\Gamma\|_{\mathrm{op}}$, the tolerances of Theorem 20 are

$$\begin{aligned} t_n &= m \log 9 + \log(2/\delta), \\ \zeta_n &= 4L\big(\sqrt{t_n/n} + t_n/n\big), \\ \eta_\Sigma &= \zeta_n(1 + 2\|K\|_{\mathrm{op}} + \zeta_n), \\ \eta_H &= \zeta_n(1 + \|K\|_{\mathrm{op}} + \|M\|_{\mathrm{op}} + \zeta_n). \end{aligned} \tag{42}$$

**Lemma 19** (Gaussian covariance and truncated inverse). *Let $W_i \sim \mathcal{N}(0, \Gamma)$ be independent in $\mathbb{R}^m$ and $\widehat{\Gamma} = n^{-1} \sum_i W_i W_i^\top$. With $t = m \log 9 + \log(2/\delta)$,*

$$\|\widehat{\Gamma} - \Gamma\|_{\mathrm{op}} \le 4\|\Gamma\|_{\mathrm{op}} \{\sqrt{t/n} + t/n\}$$

*with probability at least $1 - \delta$. Further, let $A \succeq 0$ have rank $r \ge 1$ and smallest positive eigenvalue $\lambda$, and suppose $\|\widehat{A} - A\|_{\mathrm{op}} \le \eta < \lambda/4$, where $\widehat{A}$ is symmetric. If $\widehat{A}_r$ retains the eigenvalues of $\widehat{A}$ above $2\eta$ and $\widehat{P}$ is its range projector, then*

$$\operatorname{rank}(\widehat{A}_r) = r, \quad \|\widehat{P} - P_A\|_{\mathrm{op}} \le 2\eta/\lambda, \quad \|\widehat{A}_r^\dagger\|_{\mathrm{op}} \le 4/(3\lambda), \quad \|\widehat{A}_r^\dagger - A^\dagger\|_{\mathrm{op}} \le 8\eta/\lambda^2.$$

*Proof.* For fixed unit $x$, $n\,x^\top\widehat{\Gamma}x/(x^\top\Gamma x)$ is $\chi^2_n$ (with the zero-variance case trivial). The standard chi-square Chernoff bound gives, for $s>0$,

$$|x^\top(\widehat{\Gamma}-\Gamma)x| \le 2\|\Gamma\|_{\mathrm{op}}\{\sqrt{s/n}+s/n\}$$

except on a set of probability $2e^{-s}$. A $1/4$-net of the unit sphere has at most $9^m$ points, and for a symmetric matrix its operator norm is at most twice the maximum quadratic form on this net. Union bounding with $s=t$ proves (19).

For the second claim, Weyl's inequality places the $r$ signal eigenvalues above $\lambda-\eta>3\lambda/4>2\eta$. Every other eigenvalue of $\widehat{A}$ lies in $[-\eta,\eta]$. Thus thresholding selects exactly $r$ positive eigenvalues and $\|\widehat{A}_r^\dagger\|_{\mathrm{op}}\le 4/(3\lambda)$. The Davis–Kahan sin-theta inequality, applied across the gap from zero to $\lambda$, gives

$$s := \|\widehat{P}-P_A\|_{\mathrm{op}} \le \frac{\eta}{\lambda-\eta} \le \frac{2\eta}{\lambda}. \tag{43}$$

We now derive the inverse constant without identifying the two signal subspaces. Put $\widetilde{A}=\widehat{A}_r=\widehat{P}\widehat{A}\widehat{P}$. The spectral projector $\widehat{P}$ commutes with $\widehat{A}$, and every discarded eigenvalue has absolute value at most $\eta$. Consequently,

$$\|\widetilde{A}-A\|_{\mathrm{op}} \le \|\widetilde{A}-\widehat{A}\|_{\mathrm{op}} + \|\widehat{A}-A\|_{\mathrm{op}} \le 2\eta.$$

For two symmetric matrices of the same rank, direct multiplication verifies the Moore–Penrose identity

$$\begin{aligned}\widetilde{A}^\dagger - A^\dagger = &-\widetilde{A}^\dagger(\widetilde{A}-A)A^\dagger \\ &+\widetilde{A}^\dagger(I-P_A)-(I-\widehat{P})A^\dagger.\end{aligned}$$

The last two terms contain only cross-subspace projectors because $\widetilde{A}^\dagger(I-P_A)=\widetilde{A}^\dagger\widehat{P}(I-P_A)$ and $(I-\widehat{P})A^\dagger=(I-\widehat{P})P_AA^\dagger$. Using (43),

$$\begin{aligned}\|\widetilde{A}^\dagger-A^\dagger\|_{\mathrm{op}} &\le \frac{4}{3\lambda}(2\eta)\frac{1}{\lambda}+\frac{4}{3\lambda}s+\frac{1}{\lambda}s \\ &\le \left(\frac{8}{3}+\frac{8}{3}+2\right)\frac{\eta}{\lambda^2}=\frac{22}{3}\frac{\eta}{\lambda^2}<\frac{8\eta}{\lambda^2}.\end{aligned}$$

This proves every constant in (19) and never inverts a null-space direction. □

### C.5 Recovery of rank, subspace and invariants

**Theorem 20** (Finite-sample recovery of rank, subspace, and invariants)**.** *Let $W=(Z^\top,U^\top,Y^\top)^\top\in\mathbb{R}^m$, where $m=2d+p$, let $L=\|\operatorname{Cov}(W)\|_{\mathrm{op}}$, and let $\zeta_n,\eta_\Sigma,\eta_H$ be the explicit tolerances* (42) *of Appendix C.4, each of order $L\sqrt{(m+\log(1/\delta))/n}$ up to factors in $\|K\|_{\mathrm{op}}$ and $\|M\|_{\mathrm{op}}$. With probability at least $1-\delta$,*

$$\|\widehat{K}-K\|_{\mathrm{op}}\vee\|\widehat{M}-M\|_{\mathrm{op}}\le\zeta_n,\quad \|\widehat{\Sigma}-\Sigma\|_{\mathrm{op}}\le\eta_\Sigma,\quad \|\widehat{H}_B-H_B\|_{\mathrm{op}}\le\eta_H. \tag{44}$$

*If $r\ge 1$ and $\lambda_+>4\eta_\Sigma$, let $\widehat{P}$ project onto the eigenvectors of $\widehat{\Sigma}$ with eigenvalue greater than $2\eta_\Sigma$, and let $\widehat{\Sigma}_r^\dagger$ invert exactly those eigenvalues. Then on the same event $\operatorname{rank}(\widehat{P})=r$ and*

$$\|\widehat{P}-P_S\|_{\mathrm{op}}\le 2\eta_\Sigma/\lambda_+, \tag{45}$$

$$\|\widehat{H}_B\widehat{\Sigma}_r^\dagger-BP_S\|_{\mathrm{op}}\le\frac{4\eta_H}{3\lambda_+}+\frac{8\|H_B\|_{\mathrm{op}}\eta_\Sigma}{\lambda_+^2}. \tag{46}$$

*The thresholding rule is thus sample-computable, and the beta-min condition says when its returned rank equals the structural rank.*

For a regularized target that requires no spectral gap, fix $\alpha>\eta_\Sigma$ and define

$$G_\alpha=H_B(\Sigma+\alpha I_p)^{-1},\qquad \widehat{G}_\alpha=\widehat{H}_B(\widehat{\Sigma}+\alpha I_p)^{-1}.$$

The same event gives

$$\|\widehat{G}_\alpha - G_\alpha\|_{\mathrm{op}} \le \frac{\eta_H}{\alpha - \eta_\Sigma} + \frac{\|H_B\|_{\mathrm{op}}\eta_\Sigma}{\alpha(\alpha - \eta_\Sigma)}. \tag{47}$$

On a covariance eigenvector of value $\lambda$, $G_\alpha$ equals $\lambda/(\lambda+\alpha)$ times the corresponding action effect. The target and its error therefore vary continuously as the covariance approaches singularity.

*Proof of Theorem 20.* The vector $W = (Z, U, Y)$ is centered jointly Gaussian. Lemma 19 gives $\|\widehat{\Gamma} - \Gamma\|_{\mathrm{op}} \le \zeta_n$. Every displayed raw cross-moment is a block of this matrix, so its error is at most $\zeta_n$. The identities in (40) follow from $Y = MZ + BE + \Xi$, independence, and centering. In particular,

$$\mathbb{E}[YZ^\top] = M, \qquad \mathbb{E}[YU^\top] = MK^\top + B\Sigma.$$

Consequently,

$$\begin{aligned}
\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}} &\le \zeta_n + \|(\widehat{K} - K)\widehat{K}^\top + K(\widehat{K} - K)^\top\|_{\mathrm{op}} \\
&\le \zeta_n(1 + 2\|K\|_{\mathrm{op}} + \zeta_n) = \eta_\Sigma, \\
\|\widehat{H}_B - H_B\|_{\mathrm{op}} &\le \zeta_n + \|(\widehat{M} - M)\widehat{K}^\top + M(\widehat{K} - K)^\top\|_{\mathrm{op}} \\
&\le \zeta_n(1 + \|K\|_{\mathrm{op}} + \|M\|_{\mathrm{op}} + \zeta_n) = \eta_H.
\end{aligned}$$

This proves (44).

Apply the second half of Lemma 19 with $A = \Sigma$. Since $H_B\Sigma^\dagger = BP_S$,

$$\|\widehat{H}_B\widehat{\Sigma}_r^\dagger - BP_S\|_{\mathrm{op}} \le \eta_H\|\widehat{\Sigma}_r^\dagger\|_{\mathrm{op}} + \|H_B\|_{\mathrm{op}}\|\widehat{\Sigma}_r^\dagger - \Sigma^\dagger\|_{\mathrm{op}},$$

which is (46). Finally, $\lambda_{\min}(\widehat{\Sigma} + \alpha I) \ge \alpha - \eta_\Sigma > 0$. The resolvent identity and the same product decomposition give

$$\|\widehat{G}_\alpha - G_\alpha\|_{\mathrm{op}} \le \frac{\eta_H}{\alpha - \eta_\Sigma} + \frac{\|H_B\|_{\mathrm{op}}\eta_\Sigma}{\alpha(\alpha - \eta_\Sigma)}.$$

□

### C.6 Thresholds computed from data

We can compute the threshold from the sample. Set $m = 2d + p$, $t_n = m\log 9 + \log(2/\delta)$, and let

$$a_n = 4\{\sqrt{t_n/n} + t_n/n\}, \qquad \widehat{\Gamma} = n^{-1}\sum_{i=1}^{n} W_i W_i^\top.$$

When $a_n < 1$, define

$$\begin{aligned}
\bar{\zeta}_n &= \frac{a_n}{1 - a_n}\|\widehat{\Gamma}\|_{\mathrm{op}}, \\
\bar{\eta}_\Sigma &= \bar{\zeta}_n(1 + 2\|\widehat{K}\|_{\mathrm{op}} + \bar{\zeta}_n), \\
\bar{\eta}_H &= \bar{\zeta}_n(1 + \|\widehat{K}\|_{\mathrm{op}} + \|\widehat{M}\|_{\mathrm{op}} + \bar{\zeta}_n).
\end{aligned} \tag{48}$$

**Corollary 21** (Data-driven rank threshold under beta-min)**.** *On the probability-$1 - \delta$ event of Theorem 20, all four bounds in* (44) *hold with the barred, sample-computable tolerances in* (48)*. If $\lambda_+ > 4\bar{\eta}_\Sigma$, thresholding the eigenvalues of $\widehat{\Sigma}$ at $2\bar{\eta}_\Sigma$ returns rank $r$, and* (45)–(46) *hold with barred tolerances. Thus the estimator and its cutoff use only the sample; the usual beta-min condition states exactly when the returned rank equals structural rank.*

*Proof of Corollary 21.* On (19), $\|\widehat{\Gamma}\|_{\mathrm{op}} \ge (1 - a_n)L$, hence $\zeta_n = a_n L \le \bar{\zeta}_n$. Also $\|K\|_{\mathrm{op}} \le \|\widehat{K}\|_{\mathrm{op}} + \bar{\zeta}_n$ and similarly for $M$. Repeating the two algebraic expansions in the preceding proof gives exactly the barred tolerances. Lemma 19 then applies with $\eta = \bar{\eta}_\Sigma$. □

### C.7 Propagation of representation error

We next join representation stability to moment recovery. This also records the rotations of every estimand. Put

$$T_U := \mathbb{E}\|U\|_2^2 = \|K\|_F^2 + \operatorname{tr}\Sigma, \qquad \Delta_h := 2\varepsilon/\gamma_{\mathrm{rep}},$$

and define

$$\begin{aligned} b_K &= \sqrt{T_U \Delta_h}, \qquad b_M = 2\sqrt{d\Delta_h} + \Delta_h, \\ b_\Sigma &= b_K(2\|K\|_{\mathrm{op}} + b_K), \\ b_H &= \sqrt{T_U \Delta_h} + b_M(\|K\|_{\mathrm{op}} + b_K) \\ &\quad + \|M\|_{\mathrm{op}} b_K. \end{aligned} \tag{49}$$

**Theorem 22** (Learned-encoder downstream recovery)**.** *Let an admissible $h$ satisfy* (29)*, and let $Q$ be supplied by Theorem 11. On an independent transition sample set define $\widetilde{Z}_i = h(g(Z_i))$, $\widetilde{Y}_i = h(g(Y_i))$ and form the estimators in* (41) *with $(Z, Y)$ replaced by $(\widetilde{Z}, \widetilde{Y})$. Their population plug-in versions obey*

$$\begin{aligned} \|\widetilde{K} - KQ^\top\|_{\mathrm{op}} &\le b_K, & \|\widetilde{M} - QMQ^\top\|_{\mathrm{op}} &\le b_M, \\ \|\widetilde{\Sigma} - \Sigma\|_{\mathrm{op}} &\le b_\Sigma, & \|\widetilde{H} - QH_B\|_{\mathrm{op}} &\le b_H. \end{aligned} \tag{50}$$

*For the empirical bounds, assume the independently fixed encoder satisfies*

$$\kappa := \sup_{\|a\|_2=1} \|a^\top(\widetilde{Z}^\top, U^\top, \widetilde{Y}^\top)^\top\|_{\psi_2} < \infty.$$

*Set $m = 2d + p$, $s_n = m + \log(8/\delta)$, and*

$$\chi_n = 96\kappa^2\{\sqrt{s_n/n} + s_n/n\}.$$

*With probability at least $1 - \delta$, the corresponding empirical errors relative to the rotated true targets are at most*

$$\begin{aligned} \beta_K &= b_K + \chi_n, \qquad \beta_M = b_M + \chi_n, \\ \beta_\Sigma &= b_\Sigma + \chi_n[1 + 2(\|K\|_{\mathrm{op}} + b_K) + \chi_n], \\ \beta_H &= b_H + \chi_n[1 + \|K\|_{\mathrm{op}} + b_K + \|M\|_{\mathrm{op}} \\ &\quad + b_M + \chi_n]. \end{aligned} \tag{51}$$

*If $\lambda_+ > 4\beta_\Sigma$, thresholding $\widehat{\Sigma}$ at $2\beta_\Sigma$ recovers $r$ and*

$$\begin{aligned} \|\widehat{P} - P_S\|_{\mathrm{op}} &\le 2\beta_\Sigma/\lambda_+, \\ \|\widehat{H}\widehat{\Sigma}_r^\dagger - QBP_S\|_{\mathrm{op}} &\le \frac{4\beta_H}{3\lambda_+} + \frac{8\|H_B\|_{\mathrm{op}}\beta_\Sigma}{\lambda_+^2}. \end{aligned} \tag{52}$$

*Hence encoder optimization and approximation error, covariance sampling, subspace recovery, and affine-effect recovery appear in one bound. Cross-fit folds may replace the independent transition split.*

**Lemma 23** (Sub-Gaussian covariance)**.** *For independent centered vectors $V_i \in \mathbb{R}^m$ satisfying $\sup_{\|a\|=1} \|a^\top V_i\|_{\psi_2} \le \kappa$,*

$$\left\| n^{-1} \sum_i \left( V_i V_i^\top - \mathbb{E}[V_i V_i^\top] \right) \right\|_{\mathrm{op}} \le 96\kappa^2\{\sqrt{(m + \log(8/\delta))/n} + (m + \log(8/\delta))/n\}$$

*with probability at least $1 - \delta$.*

*Proof.* For fixed unit $a$, the centered square $(a^\top V_i)^2 - \mathbb{E}(a^\top V_i)^2$ has sub-exponential norm at most $4\kappa^2$. Bernstein's Chernoff argument bounds its empirical mean by $16\kappa^2\{\sqrt{s/n} + s/n\}$ with failure probability at most $2e^{-s}$. Use a $1/4$-net of size at most $9^m$, the factor-two quadratic-form net inequality, and $s = m \log 9 + \log(2/\delta)$. Put $s_n = m + \log(8/\delta)$. Since $s \le 3s_n$, the resulting bound is

$$32\sqrt{3}\,\kappa^2\sqrt{s_n/n} + 96\kappa^2 s_n/n \le 96\kappa^2\{\sqrt{s_n/n} + s_n/n\}.$$

□

*Proof of Theorem 22.* Write $e_Z = \widetilde{Z} - QZ$ and $e_Y = \widetilde{Y} - QY$. Stationarity and Theorem 11 give $\mathbb{E}\|e_Z\|^2, \mathbb{E}\|e_Y\|^2 \le \Delta_h$. Cauchy–Schwarz yields

$$\|\mathbb{E}[Ue_Z^\top]\|_{\mathrm{op}} \le \sqrt{T_U \Delta_h}$$

and, after expanding $\mathbb{E}[(QY + e_Y)(QZ + e_Z)^\top] - QMQ^\top$, yields $2\sqrt{d\Delta_h} + \Delta_h$. These are $b_K, b_M$. Expanding the two plug-in products gives

$$\|\widetilde{\Sigma} - \Sigma\|_{\mathrm{op}} \le b_K(2\|K\|_{\mathrm{op}} + b_K)$$

and

$$\|\widetilde{H} - QH_B\|_{\mathrm{op}} \le \sqrt{T_U \Delta_h} + b_M(\|K\|_{\mathrm{op}} + b_K) + \|M\|_{\mathrm{op}} b_K.$$

This proves (50).

Apply Lemma 23 to $V = (\widetilde{Z}, U, \widetilde{Y})$. Every raw-moment block is within $\chi_n$ of its population value. Repeating the product expansions, now around the population plug-ins, gives the sampling terms in (51). Adding the population biases proves those bounds. The truncated-inverse part of Lemma 19, with $\eta = \beta_\Sigma$, proves (52). The aligned targets follow from $K_Q = KQ^\top$, $M_Q = QMQ^\top$, and $H_Q = QH_B$. □

### C.8 One stationary trajectory

The i.i.d. theorem extends to dependent transitions by spacing the moments at the mixing time. For the stationary triple process $T_t = (Z_t, U_t, Z_{t+1})$, use the convention

$$\beta_T(k) = \sup_t \beta\{\sigma(T_j : j \le t), \sigma(T_j : j \ge t + k)\}.$$

Thus the fact that adjacent triples share $Z_{t+1}$ is already included in $\beta_T(1)$; spacing is measured between triple indices.

**Corollary 24** (Geometrically mixing trajectory)**.** *Assume $\beta_T(k) \le c_\beta e^{-k/\tau}$. Given $N$ transitions $T_1, \dots, T_N$ (and hence $N + 1$ states), set*

$$\ell = \max\{1, \lceil \tau \log(2c_\beta N/\delta) \rceil\}, \qquad n_{\mathrm{eff}} = 1 + \lfloor (N - 1)/\ell \rfloor.$$

*For $t_j = 1 + (j - 1)\ell$, compute the moment estimators from $T_{t_1}, \dots, T_{t_{n_{\mathrm{eff}}}}$. Every conclusion of Theorem 20 holds with $n = n_{\mathrm{eff}}$ and confidence parameter $\delta/2$, hence with overall probability at least $1 - \delta$ after coupling. The same replacement applies to Theorem 22 for an encoder fixed independently of the evaluated trajectory.*

*Proof of Corollary 24.* Write $t_j = 1 + (j - 1)\ell$. Iterative Berbee coupling replaces $T_{t_j}$, in order, by independent variables $T^\star_{t_j}$ having the same stationary triple marginal. At step $j$, the previous selected indices belong to $\sigma(T_k : k \le t_{j-1})$ and the next index is $t_j = t_{j-1} + \ell$; under the displayed convention the mismatch probability is at most $\beta_T(\ell)$. A union bound therefore gives total coupling failure at most

$$(n_{\mathrm{eff}} - 1)\beta_T(\ell) \le N c_\beta e^{-\ell/\tau} \le \delta/2.$$

The coupling applies to spaced transition triples, with dependence measured by $\beta_T$. On the coupling event the selected triples equal an i.i.d. sample. Apply the relevant i.i.d. theorem with confidence parameter $\delta/2$ and union bound the sampling and coupling failures. □

**Corollary 25** (Nonstationary start with burn-in)**.** *Suppose the controlled state process is Markov with stationary law $\pi$, the stationary triple process satisfies the preceding mixing bound, and*

$$\|\mathcal{L}(Z_b) - \pi\|_{\mathrm{TV}} \le c_0 e^{-b/\tau_0}.$$

*Choose $b \ge \lceil \tau_0 \log(2c_0/\delta) \rceil$, discard the first $b$ transitions, and apply Corollary 24 to the remaining $N - b$ transitions with confidence input $\delta/2$. Then all of its moment, subspace, and invariant bounds hold with probability at least $1 - \delta$.*

*Proof of Corollary 25.* Maximal coupling matches $Z_b$ to an independent draw from $\pi$ with failure probability at most $c_0 e^{-b/\tau_0} \le \delta/2$. On a match, use the same future action and process innovations for the two Markov chains; their entire post-burn-in paths then coincide, and the coupled path is stationary. Corollary 24, invoked with confidence input $\delta/2$, fails with probability at most $\delta/2$. A union bound proves the claim. □

## D Intervention Directions and Joint Identification

Consider augmenting the residual action by independent scalar Gaussian perturbations. Let

$$E_k = E + \sum_{i=1}^{k} v_i \eta_i, \qquad \eta_i \sim \mathcal{N}(0, \tau_i^2), \quad \tau_i^2 > 0, \tag{53}$$

where the $\eta_i$ are mutually independent and independent of $(Z, E)$. Define $U_k = KZ + E_k$ and

$$S_k := \operatorname{span}(S, v_1, \ldots, v_k).$$

**Lemma 26** (Range of augmented Gaussian excitation)**.** *Under* (53)*,*

$$\operatorname{range}(\operatorname{Cov}(E_k)) = S_k = \operatorname{span}(S, v_1, \ldots, v_k).$$

*Moreover, the vector space* $\{D : DP_{S_k} = 0\}$ *has dimension* $d(p - \dim S_k)$ *and equals* $\{0\}$ *precisely when* $S_k = \mathbb{R}^p$*.*

*Proof.* Independence and centering give

$$\operatorname{Cov}(E_k) = \Sigma_k := \Sigma + \sum_{i=1}^{k} \tau_i^2 v_i v_i^\top.$$

For every $x \in \mathbb{R}^p$,

$$x^\top \Sigma_k x = x^\top \Sigma x + \sum_{i=1}^{k} \tau_i^2 (x^\top v_i)^2. \tag{54}$$

Every term is nonnegative. Hence the left side vanishes exactly when $x \in \ker \Sigma = S^\perp$ and $x \perp v_i$ for every $i$. Therefore

$$\ker \Sigma_k = S_k^\perp.$$

For a symmetric matrix, the range is the orthogonal complement of the kernel, proving $\operatorname{range}(\Sigma_k) = S_k$.

In an orthonormal basis adapted to $S_k$, a matrix that annihilates $S_k$ has $\dim S_k$ zero columns and $p - \dim S_k$ arbitrary columns in $\mathbb{R}^d$. This gives the dimension formula. The direction space reduces to the zero matrix exactly when $p - \dim S_k = 0$. □

*Proof of Theorem 6.* The augmented action retains the feedback-plus-residual form

$$U_k = KZ + E_k.$$

Lemma 26 identifies the residual support as $S_k$. Applying Lemma 14 with $S_k$ in place of $S$ gives

$$\mathcal{I}_k(A, B) = \{(A - DK, B + D) : DP_{S_k} = 0\}.$$

The same lemma gives its dimension $d(p - \dim S_k)$. The parameterization is injective through its $B$-component, so the fiber is a singleton precisely when its direction space is zero, equivalently when $S_k = \mathbb{R}^p$.

It remains to establish the scalar count. For any $k$ directions,

$$\dim S_k \le r + k.$$

If $k < p - r$, then $\dim S_k < p$, so $S_k^\perp$ contains a nonzero vector $w$. Choose any nonzero $a \in \mathbb{R}^d$ and set

$$D = aw^\top. \tag{55}$$

For $s \in S_k$, $w^\top s = 0$, hence $Ds = 0$ and $DP_{S_k} = 0$. Since $a, w \neq 0$, the matrix $D$ is nonzero and produces a distinct pair in the fiber. Thus every design with fewer than $p - r$ scalar directions leaves ambiguity.

For the converse count, choose $p - r$ vectors forming a basis of $S^\perp$. Their span together with $S$ is $\mathbb{R}^p$, so the fiber is a singleton. Combining the lower and upper bounds proves the minimum count in (16). □

**Corollary 27** (Dimension removed by a perturbation batch)**.** *Let*

$$q := \dim\big(\mathrm{span}(P_{S^\perp} v_1, \dots, P_{S^\perp} v_k)\big).$$

*Then the augmented fiber has dimension $d(p-r-q)$. Thus each newly independent quotient direction removes exactly $d$ affine degrees of freedom.*

*Proof.* The quotient contribution of the added directions has dimension $q$, so $\dim S_k = r + q$. Substitute this identity into the fiber dimension $d(p - \dim S_k)$ of Theorem 6. □

For example, when $S = \mathrm{span}(e_1) \subset \mathbb{R}^3$, adding $e_2$ leaves one unexcited direction. A further perturbation along $e_1 + e_2$ is redundant, while one along $e_1 + e_3$ completes the span.

### D.1 Stationary exploration and simultaneous retraining

The physical intervention changes state covariance but not the closed-loop matrix $F = A + BK$. This permits an end-to-end statement under the stationary law created by the intervention.

**Theorem 28** (Simultaneous representation and transition identification)**.** *Let $F = A + BK$ be Schur stable, let $Q_\xi = \mathrm{Cov}(\Xi)$, and augment the residual covariance to*

$$\Sigma_\Omega = \Sigma + V\Omega V^\top, \qquad \Omega \succ 0,$$

*where the columns of $V$ span $S^\perp$. The unique stationary covariance is*

$$P_\Omega = \sum_{j=0}^{\infty} F^j (B\Sigma_\Omega B^\top + Q_\xi)(F^j)^\top. \tag{56}$$

*Assume $P_\Omega \succ 0$ and define the whitened predictable covariance*

$$R_\Omega = P_\Omega^{-1/2}\,\mathrm{Cov}(FZ + BE_\Omega) P_\Omega^{-1/2}.$$

*If*

$$\lambda_{\min}(R_\Omega) > \lambda_{\max}(R_\Omega)^2, \tag{57}$$

*then population retraining under the augmented stationary law recovers $P_\Omega^{-1/2} Z$ up to an orthogonal transform and point identifies the entire affine transition in those coordinates. Thus the same intervention simultaneously supplies representation recovery and full action-effect identification, while the stationary covariance change is absorbed by whitening.*

*Proof of Theorem 28.* Under the augmented policy the dynamics are

$$Z^+ = FZ + BE_\Omega + \Xi.$$

Schur stability makes the series in (56) converge in operator norm and gives the unique solution of the discrete Lyapunov equation. Gaussian innovations then give the stationary law $\mathcal{N}(0, P_\Omega)$. With $\bar{Z} = P_\Omega^{-1/2} Z$, the process becomes

$$\bar{Z}^+ = P_\Omega^{-1/2} A P_\Omega^{1/2} \bar{Z} + P_\Omega^{-1/2} BU + P_\Omega^{-1/2}\Xi, \quad U = K P_\Omega^{1/2} \bar{Z} + E_\Omega.$$

Both $\bar{Z}$ and $\bar{Z}^+$ are standard Gaussian, and their predictable covariance is exactly $R_\Omega$. Theorem 1 therefore gives orthogonal recovery of $\bar{Z}$. Because $\Omega \succ 0$ on $S^\perp$, Lemma 26 gives $\mathrm{range}(\Sigma_\Omega) = S \oplus S^\perp = \mathbb{R}^p$. The full-rank endpoint of the fiber theorem forces $D = 0$, proving point identification. □

There is an explicit energy range in which an existing margin persists. Suppose the unperturbed stationary covariance is $I_d$, put $R = I_d - Q_\xi$, and define

$$\kappa_F = \sum_{j=0}^{\infty} \|F^j BV\|_{\mathrm{op}}^2,$$

$$h_\Omega = \kappa_F \|\Omega\|_{\mathrm{op}}, \qquad s_\Omega = 1 - (1 + h_\Omega)^{-1/2},$$

and

$$\delta_\Omega = \|Q_\xi\|_{\mathrm{op}}(2s_\Omega + s_\Omega^2).$$

**Corollary 29** (Explicit margin-preserving energy range)**.** *Under the preceding definitions,*

$$\lambda_{\min}(R_\Omega) - \lambda_{\max}(R_\Omega)^2 \geq \gamma_{\mathrm{rep}} - (1+2\rho)\delta_\Omega - \delta_\Omega^2.$$

*Consequently every positive-definite $\Omega$ for which the right side is positive satisfies the simultaneous conclusion of Theorem 28. The closed-loop stability condition is independent of $\Omega$, and a weighted constraint $\mathrm{tr}(W\Omega) \leq P$ can be imposed simultaneously with Theorem 30.*

*Proof of Corollary 29.* Subtract the baseline and augmented Lyapunov series to obtain

$$\Delta P := P_\Omega - I_d = \sum_{j=0}^{\infty} F^j B V \Omega V^\top B^\top (F^j)^\top \succeq 0, \qquad \|\Delta P\|_{\mathrm{op}} \leq h_\Omega.$$

Functional calculus on $I + \Delta P$ gives

$$\|P_\Omega^{-1/2} - I\|_{\mathrm{op}} \leq 1 - (1+h_\Omega)^{-1/2} = s_\Omega.$$

Stationarity yields

$$R_\Omega = I - P_\Omega^{-1/2} Q_\xi P_\Omega^{-1/2}, \qquad R = I - Q_\xi.$$

Writing $P_\Omega^{-1/2} = I + H$ and expanding the two products shows $\|R_\Omega - R\|_{\mathrm{op}} \leq \|Q_\xi\|_{\mathrm{op}}(2\|H\|_{\mathrm{op}} + \|H\|_{\mathrm{op}}^2) \leq \delta_\Omega$. Weyl's inequality then gives

$$\lambda_{\min}(R_\Omega) \geq \lambda_{\min}(R) - \delta_\Omega, \qquad \lambda_{\max}(R_\Omega) \leq \rho + \delta_\Omega.$$

Subtracting the square of the second bound from the first proves (29). Schur stability concerns only $F$, which is unchanged by residual-noise covariance. □

# E Action Effect Estimation and Probe Energy Allocation

The span theorem decides whether the missing effect can be learned. We first solve an exact regression benchmark and then propagate all quantities needed to construct its residual from learned coordinates. Let $q = p - r$, let $V \in \mathbb{R}^{p\times q}$ have orthonormal columns spanning $S^\perp$, and write

$$T = BV, \qquad M = A + BK, \qquad G_S = BP_S.$$

During an intervention draw $x_i \sim \mathcal{N}(0, \Omega)$ independently and use

$$U_i = KZ_i + E_i + Vx_i.$$

Without observing $E_i$, we can form the residual

$$\begin{aligned} R_i &:= Y_i - MZ_i - G_S(U_i - KZ_i) \\ &= Tx_i + \xi_i, \qquad \xi_i \sim \mathcal{N}(0, Q_\xi). \end{aligned} \tag{58}$$

The noise is independent of the design and $\nu := \mathrm{tr}\, Q_\xi > 0$. Equation (58) is the oracle benchmark in which the latent coordinate and nuisance matrices are known; Theorem 33 below replaces every one of them by its first-stage estimate.

For squared Frobenius loss and the unrestricted parameter space $T \in \mathbb{R}^{d\times q}$, define

$$\mathfrak{R}_n(\Omega) := \inf_{\widehat{T}} \sup_{T} \mathbb{E}_T \|\widehat{T} - T\|_F^2.$$

## E.1 Minimax regression risk

**Theorem 30** (Exact minimax quotient-regression risk)**.** *If $\Omega \succ 0$ and $n > q+1$, then*

$$\mathfrak{R}_n(\Omega) = \frac{\nu}{n-q-1}\,\mathrm{tr}(\Omega^{-1}), \tag{59}$$

*attained by ordinary least squares. If $\mathrm{rank}(\Omega) < q$ or $n \leq q+1$, then $\mathfrak{R}_n(\Omega) = +\infty$. Under a positive-definite action cost $W \succ 0$ and total expected energy $n\,\mathrm{tr}(W\Omega) \leq \mathcal{E}$, the unique optimal covariance and its risk are*

$$\Omega_\star = \frac{\mathcal{E}}{n\,\mathrm{tr}(W^{1/2})} W^{-1/2}, \qquad \inf_\Omega \mathfrak{R}_n(\Omega) = \frac{\nu n \{\mathrm{tr}(W^{1/2})\}^2}{\mathcal{E}(n-q-1)}. \tag{60}$$

*Proof of Theorem 30.* Let $X = [x_1, \dots, x_n] \in \mathbb{R}^{q\times n}$ and $R = [R_1, \dots, R_n]$. Conditional on a full-row-rank $X$, ordinary least squares is

$$\widehat{T} = RX^\top (XX^\top)^{-1} = T + \Xi X^\top (XX^\top)^{-1}.$$

Using independence of the columns of $\Xi$,

$$\mathbb{E}[\|\widehat{T} - T\|_F^2 \mid X] = \operatorname{tr}(Q_\xi)\operatorname{tr}((XX^\top)^{-1}) = \nu \operatorname{tr}((XX^\top)^{-1}). \tag{61}$$

We verify that this is the conditional minimax value even for correlated or singular output noise. Choose an orthogonal $O$ with $OQ_\xi O^\top = \operatorname{diag}(\sigma_1^2, \dots, \sigma_d^2)$ and rotate both the response and $T$ by $O$. Frobenius loss is unchanged, and the rotated rows are independent scalar regressions. For every $\sigma_j^2 > 0$, put the prior $\mathcal{N}(0, \tau^2 I_q)$ on row $j$ of $OT$. Its posterior covariance is

$$\{\tau^{-2} I_q + \sigma_j^{-2} XX^\top\}^{-1}.$$

The integrated Bayes risk increases as $\tau \to \infty$ to $\sigma_j^2 \operatorname{tr}\{(XX^\top)^{-1}\}$. A row with $\sigma_j^2 = 0$ is recovered exactly because $X$ has full row rank and contributes zero. Summing the independent row risks gives the right side of (61). Bayes risk lower-bounds minimax risk, while ordinary least squares gives the matching upper bound.

For Gaussian design, $XX^\top \sim \mathsf{Wishart}_q(\Omega, n)$. When $n > q + 1$ and $\Omega \succ 0$,

$$\mathbb{E}[(XX^\top)^{-1}] = \frac{\Omega^{-1}}{n - q - 1},$$

which proves (59). If $n < q$, every design has a null direction. For $n = q$ or $q + 1$, the diffuse-prior Bayes risks increase to a quantity whose inverse-Wishart expectation is infinite; monotone convergence therefore makes the unconditional minimax risk infinite. If $\Omega$ is singular, a deterministic $v \in \ker \Omega$ obeys $v^\top x_i = 0$ almost surely, so $T$ and $T + av^\top$ induce the same law for every $a \in \mathbb{R}^d$ and unrestricted risk is infinite.

It remains to optimize $\Omega$. Hilbert–Schmidt Cauchy–Schwarz gives

$$\operatorname{tr}(\Omega^{-1})\operatorname{tr}(W\Omega) \geq \{\operatorname{tr}(W^{1/2})\}^2. \tag{62}$$

Equality holds precisely when $\Omega$ is proportional to $W^{-1/2}$. The budget is active because scaling $\Omega$ upward reduces the risk. Substituting $\operatorname{tr}(W\Omega) = \mathcal{E}/n$ gives (60); setting $W = I_q$ and $\nu = d\sigma^2$ gives its isotropic specialization. The equality condition also proves uniqueness. □

The random-design degrees-of-freedom factor is exact for this Gaussian design experiment. The following comparison uses the same total energy $\mathcal{E}$, so the only ratio is the inverse-Wishart factor.

**Proposition 31** (Deterministic and amplitude-constrained designs)**.** *For fixed $x_1, \dots, x_n$ with Gram matrix $G_X = \sum_i x_i x_i^\top$,*

$$\inf_{\widehat{T}} \sup_{T} \mathbb{E}_T \|\widehat{T} - T\|_F^2 = \begin{cases} \nu \operatorname{tr}(G_X^{-1}), & G_X \succ 0, \\ +\infty, & \operatorname{rank}(G_X) < q. \end{cases}$$

*Under the same weighted total budget $\operatorname{tr}(WG_X) \leq \mathcal{E}$, the unique optimal Gram matrix and risk are*

$$G_{X,\star} = \frac{\mathcal{E}}{\operatorname{tr}(W^{1/2})} W^{-1/2}, \qquad \frac{\nu\{\operatorname{tr}(W^{1/2})\}^2}{\mathcal{E}}. \tag{63}$$

*Thus optimal random Gaussian design has exactly $n/(n - q - 1)$ times the optimal deterministic risk. For $W = I_q$, if $\|x_i\|_2 \leq a$ and $n$ is a multiple of $q$, cycling equally through an orthonormal basis with $\|x_i\|_2 = \sqrt{\mathcal{E}/n}$ attains the optimum whenever $\mathcal{E} \leq na^2$.*

*Proof of Proposition 31.* The conditional calculation and rowwise diffuse-prior lower bound in (61) apply verbatim to fixed $X$. If $G_X$ is singular, shifting $T$ along a fixed vector in $\ker G_X$ leaves the experiment unchanged and makes the supremum risk infinite. If $G_X \succ 0$, apply (62) with $\Omega = G_X$. Under $\operatorname{tr}(WG_X) \leq \mathcal{E}$ this yields

$$\operatorname{tr}(G_X^{-1}) \geq \frac{\{\operatorname{tr}(W^{1/2})\}^2}{\mathcal{E}},$$

with equality uniquely at the matrix in (63). For $W = I_q$, the cyclic construction has $n/q$ copies per basis direction, each squared norm $\mathcal{E}/n$, so its Gram is $(\mathcal{E}/q)I_q$ and its amplitude is admissible under the stated condition. Dividing the optimal random risk by this deterministic value gives $n/(n-q-1)$ under the same total budget. □

The design remains quantitative when $S^\perp$ is estimated. Let $\widehat{V}$ have orthonormal columns spanning $\widehat{S}^\perp$ and inject $\widehat{V}x$.

**Corollary 32** (Oracle-aligned cost of subspace error)**.** *Suppose $\widehat{S}$ and $S$ have dimension $r$ and $s = \|\widehat{P} - P_S\|_{\mathrm{op}} < 1$. Set $C = V^\top \widehat{V}$. After removing the identified component $BP_S\widehat{V}x$, the response coefficient is $TC$. If this alignment is known for the purpose of the oracle benchmark and $x \sim \mathcal{N}(0, \mathcal{E}(nq)^{-1}I_q)$, then*

$$\mathfrak{R}_n^{\mathrm{eff}} \le \frac{\nu q^2 n}{\mathcal{E}(n-q-1)(1-s^2)}. \tag{64}$$

*The executable estimator below does not use $C$ or the true $V$; it estimates the missing action operator in ambient action coordinates.*

*Proof of Corollary 32.* For orthonormal bases $V, \widehat{V}$ of the two orthogonal complements, principal-angle identities give

$$\sigma_{\min}(C)^2 = \sigma_{\min}(V^\top \widehat{V})^2 = 1 - \|\widehat{P} - P_S\|_{\mathrm{op}}^2 = 1 - s^2.$$

After subtracting the identified action component,

$$B(I - P_S)\widehat{V}x = BV(V^\top \widehat{V})x = TCx.$$

The effective covariance in true missing coordinates is $\mathcal{E}(nq)^{-1}CC^\top$. Theorem 30 and $\mathrm{tr}\{(CC^\top)^{-1}\} \le q/(1-s^2)$ prove (64). □

### E.2 Estimation in learned coordinates

The joint bound uses a first-stage sample to estimate the behavior geometry and an independent intervention batch. This sample split makes the selected intervention subspace fixed during analysis of the second stage; cross-fitting gives the same conclusion foldwise.

**Theorem 33** (Two-stage learned-coordinate exploration)**.** *Let $Q$ align a learned encoder with the latent state and define*

$$K_0 = KQ^\top, \qquad M_0 = QMQ^\top, \qquad G_0 = QBP_S.$$

*Suppose first-stage estimates satisfy*

$$\begin{aligned} \|\widehat{K} - K_0\|_{\mathrm{op}} &\le a_K, & \|\widehat{M} - M_0\|_{\mathrm{op}} &\le a_M, \\ \|\widehat{G} - G_0\|_{\mathrm{op}} &\le a_G, & \|\widehat{M} - M_0\|_F &\le a_M^F, \\ \|\widehat{G} - G_0\|_F &\le a_G^F, & \|\widehat{P} - P_S\|_{\mathrm{op}} &= s < 1, \end{aligned}$$

*where $\mathrm{rank}(\widehat{P}) = r$. Let $\widehat{V}$ be any orthonormal basis of $\mathrm{range}(I_p - \widehat{P})$. On an independent batch choose design vectors $x_i$ independently of the process noise and collect*

$$\begin{aligned} U_i &= KZ_i + E_i + \widehat{V}x_i, \\ Y_i &= MZ_i + B(E_i + \widehat{V}x_i) + \xi_i. \end{aligned}$$

*Write $\widetilde{Z}_i = h(g(Z_i))$, $\widetilde{Y}_i = h(g(Y_i))$, and form the entirely observable residual and least-squares estimators*

$$\begin{aligned} \widehat{R}_i &= \widetilde{Y}_i - \widehat{M}\widetilde{Z}_i - \widehat{G}(U_i - \widehat{K}\widetilde{Z}_i), \\ G_X &= \sum_i x_i x_i^\top, \qquad \widehat{L} = [\widehat{R}_i]X^\top G_X^{-1}, \\ \widehat{B} &= \widehat{G} + \widehat{L}\widehat{V}^\top, \qquad \widehat{A} = \widehat{M} - \widehat{B}\widehat{K}, \end{aligned} \tag{65}$$

*where $X = [x_i]$ and assume $G_X \succ 0$.*

*For the realized batch let*

$$Z_0 = [QZ_i], \qquad W_0 = [U_i - K_0 QZ_i],$$
$$E_Z = [\widetilde{Z}_i - QZ_i], \qquad E_Y = [\widetilde{Y}_i - QY_i],$$

*and define the joint nuisance/representation remainder*

$$\mathcal{D}_n = \|E_Y\|_F + \|\widehat{G}\widehat{K} - \widehat{M}\|_{\mathrm{op}}\|E_Z\|_F + (a_M + \|\widehat{G}\|_{\mathrm{op}} a_K)\|Z_0\|_F + a_G\|W_0\|_F. \tag{66}$$

*Put $B_{\perp,0} = QB(I_p - P_S)$ and*

$$\mathcal{N}_\delta(X) = \sqrt{\nu \operatorname{tr}(G_X^{-1})} + \sqrt{\frac{2\|Q_\xi\|_{\mathrm{op}} \log(1/\delta)}{\lambda_{\min}(G_X)}}. \tag{67}$$

*Conditional on the first stage and $X$, with probability at least $1 - \delta$ over the Gaussian process noise,*

$$\|\widehat{B} - QB\|_F \le a_G^F + \mathcal{N}_\delta(X) + \frac{\mathcal{D}_n}{\sqrt{\lambda_{\min}(G_X)}} + s\|B_{\perp,0}\|_F =: \mathcal{B}_n, \tag{68}$$

$$\|\widehat{A} - QAQ^\top\|_F \le a_M^F + \|\widehat{K}\|_{\mathrm{op}}\mathcal{B}_n + \|QB\|_F a_K. \tag{69}$$

*All baseline-subtraction, behavior-gain, identified-effect, subspace, encoder, and intervention-regression errors therefore enter one finite-sample bound. The quantities $a_K, a_M, a_G, s$ may be set to the bounds of Theorem 22; operator bounds imply $a_M^F \le \sqrt{d}\, a_M$ and $a_G^F \le \sqrt{d}\, a_G$.*

*If $h(g(z)) = Qz$ exactly, then $E_Z = 0$; also $E_Y = 0$ for every exploration next-state law absolutely continuous with respect to the standard Gaussian law. For approximate encoders,* (66) *records the exact additional errors under the intervention law, accounting for the change from the passive distribution.*

*Proof of Theorem 33.* Let

$$L_0 = B_{\perp,0}\widehat{V} = QB(I - P_S)\widehat{V}.$$

The exact learned-coordinate residual satisfies

$$R_i^0 := QY_i - M_0 QZ_i - G_0(U_i - K_0 QZ_i) = QB(I - P_S)(E_i + \widehat{V}x_i) + Q\xi_i = L_0 x_i + Q\xi_i,$$

because $(I - P_S)E_i = 0$. Thus no observation of $E_i$ is needed.

Put $D = [\widehat{R}_i - R_i^0]$. Direct expansion of (65) gives the exact identity

$$D = E_Y + (\widehat{G}\widehat{K} - \widehat{M})E_Z + \{\widehat{G}(\widehat{K} - K_0) - (\widehat{M} - M_0)\}Z_0 - (\widehat{G} - G_0)W_0. \tag{70}$$

The triangle and product inequalities imply $\|D\|_F \le \mathcal{D}_n$. Least squares therefore obeys

$$\widehat{L} - L_0 = Q\Xi X^\top G_X^{-1} + DX^\top G_X^{-1}, \qquad \|X^\top G_X^{-1}\|_{\mathrm{op}} = \lambda_{\min}(G_X)^{-1/2}. \tag{71}$$

Conditional on $X$, the first matrix in (71) is centered Gaussian after vectorization, with covariance

$$G_X^{-1} \otimes (QQ_\xi Q^\top).$$

Its expected squared Frobenius norm is $\nu \operatorname{tr}(G_X^{-1})$, while the operator norm of this covariance is $\|Q_\xi\|_{\mathrm{op}}/\lambda_{\min}(G_X)$. Gaussian concentration for the Euclidean norm, together with Jensen's inequality for its mean, yields

$$\|Q\Xi X^\top G_X^{-1}\|_F \le \mathcal{N}_\delta(X)$$

with conditional probability at least $1-\delta$. Combining this event with (71) gives

$$\|\widehat{L} - L_0\|_F \le \mathcal{N}_\delta(X) + \mathcal{D}_n/\sqrt{\lambda_{\min}(G_X)}.$$

Since $\widehat{V}\widehat{V}^\top = I_p - \widehat{P}$,

$$\|L_0\widehat{V}^\top - B_{\perp,0}\|_F = \|B_{\perp,0}\widehat{P}\|_F \le s\|B_{\perp,0}\|_F.$$

Using $\widehat{B} = \widehat{G} + \widehat{L}\widehat{V}^\top$ proves (68). Finally, $QAQ^\top = M_0 - QBK_0$, and hence

$$\widehat{A} - QAQ^\top = (\widehat{M} - M_0) - (\widehat{B} - QB)\widehat{K} - QB(\widehat{K} - K_0).$$

The Frobenius/operator product inequality proves (69). □

Independent stationary resets make errors in the fitted behavior dynamics part of a fixed output covariance. In exact coordinates, the resulting regression gives an exact risk decomposition even when the first stage estimates are inaccurate.

**Proposition 34** (Conditional risk in exact coordinates)**.** *Condition on the first stage $\mathcal{F}_1$, which fixes the orthogonal alignment $Q$, the estimates $\widehat{K}, \widehat{M}, \widehat{G}, \widehat{P}, \widehat{V}$, and $X \in \mathbb{R}^{q\times n}$. Suppose $q \ge 1$, $n \ge q$, $G_X = XX^\top \succ 0$, and $\widehat{V}^\top\widehat{V} = I_q$ with $\widehat{P} = I_p - \widehat{V}\widehat{V}^\top$. The first stage estimates may have arbitrary finite errors. Use independent stationary resets with $Z_i \sim \mathcal{N}(0, I_d)$, $E_i \sim \mathcal{N}(0, \Sigma)$, and $\xi_i \sim \mathcal{N}(0, Q_\xi)$, mutually independent across variables and samples, and independent of $\mathcal{F}_1$. Assume the encoder returns $QZ_i$ and $QY_i$ exactly for each intervention transition, with $U_i = KZ_i + E_i + \widehat{V}x_i$. Define*

$$C = QB - \widehat{G}, \qquad H = M_0 - \widehat{M} + \widehat{G}(\widehat{K} - K_0),$$
$$Q_{\text{eff}} = HH^\top + C\Sigma C^\top + QQ_\xi Q^\top, \qquad \nu_{\text{eff}} = \operatorname{tr}(Q_{\text{eff}}).$$

*The conditional squared error of* (65) *is*

$$\mathbb{E}[\|\widehat{B} - QB\|_F^2 \mid \mathcal{F}_1] = \|(\widehat{G} - QB)\widehat{P}\|_F^2 + \nu_{\text{eff}} \operatorname{tr}(G_X^{-1}). \tag{72}$$

*Let $W_a \succ 0$ measure action cost, set $W_q = \widehat{V}^\top W_a \widehat{V}$, and suppose $\mathcal{E} > 0$ and $\nu = \operatorname{tr}(Q_\xi) > 0$. Among fixed probe designs satisfying $\operatorname{tr}(W_q G_X) \le \mathcal{E}$, the unique Gram matrix minimizing this conditional risk is*

$$G_{X,\star} = \frac{\mathcal{E}}{\operatorname{tr}(W_q^{1/2})} W_q^{-1/2}. \tag{73}$$

*The minimum equals*

$$\|(\widehat{G} - QB)\widehat{P}\|_F^2 + \nu_{\text{eff}}\{\operatorname{tr}(W_q^{1/2})\}^2/\mathcal{E}.$$

*Proof.* Substituting the transition and action equations into the observed residual gives

$$\widehat{R}_i = C\widehat{V}x_i + \eta_i, \qquad \eta_i = HQZ_i + CE_i + Q\xi_i.$$

Conditional on $\mathcal{F}_1$, the $\eta_i$ are independent centered Gaussian vectors with covariance $Q_{\text{eff}}$. Writing $N = [\eta_i]$ and $\Delta_L = NX^\top G_X^{-1}$ gives

$$\widehat{L} = C\widehat{V} + \Delta_L, \qquad \widehat{B} - QB = (\widehat{G} - QB)\widehat{P} + \Delta_L\widehat{V}^\top.$$

The two error terms are orthogonal in Frobenius inner product for every realization because $\widehat{P}\widehat{V} = 0$. Also, $\|\Delta_L\widehat{V}^\top\|_F = \|\Delta_L\|_F$ and $\operatorname{Cov}(\operatorname{vec}\Delta_L \mid \mathcal{F}_1) = G_X^{-1} \otimes Q_{\text{eff}}$. Taking expectations proves (72).

For fixed first stage estimates, the projection error and $\nu_{\text{eff}}$ are constant across probe designs, with $\nu_{\text{eff}} \ge \nu > 0$. The trace inequality and equality condition in Proposition 31 therefore give (73) and its minimum risk. For $n \ge q$, the design $X = G_{X,\star}^{1/2}[I_q \;\; 0_{q\times(n-q)}]$ realizes this Gram matrix. □

Equation (72) separates the error retained on $\operatorname{range}(\widehat{P})$ from the variance reduced by probing its orthogonal complement. Errors in the nuisance matrices enter the effective noise covariance, preserving the optimal allocation shape within the selected subspace. With $\widehat{K} = K_0$, $\widehat{M} = M_0$, $\widehat{G} = G_0$, and $\widehat{P} = P_S$, the projection error vanishes and $\nu_{\text{eff}} = \nu$, recovering the exact regression risk. For a fixed first stage, the optimal variance decreases inversely with total energy, while the projection error remains constant. This identifies the part of the estimation error that additional probes can reduce after the intervention subspace has been selected.

### E.3 Representation error under interventions

For representation transfer, use independent stationary resets $(Z_i, E_i)$ from the behavior law, independently of the first stage. Conditional on that stage, choose a deterministic design or Gaussian draws $x_i \sim \mathcal{N}(0, \Omega)$ independent of all $(Z_i, E_i, \xi_i)$. Each $Z_i$ then retains the passive law, while stationarity gives

$$\mathrm{Cov}(Y_i) = I_d + \Theta, \qquad \Theta = B\widehat{V}\Omega\widehat{V}^\top B^\top, \tag{74}$$

for a Gaussian design with $\mathrm{Cov}(x_i) = \Omega$. A deterministic design leaves the covariance at $I_d$ and shifts the mean by $B\widehat{V}x_i$. These are the next-state laws after one intervention from a stationary reset. Section D.1 studies the stationary covariance $P_\Omega$ reached under sustained exploration. Write $e(z) = h(g(z)) - Qz$ for the aligned encoder error and $\Delta_h = \mathbb{E}\|e(Z)\|_2^2$ for its passive risk. The next proposition transfers this risk to the intervention law.

**Proposition 35** (Passive-to-intervention transfer)**.** *Under this reset protocol, suppose $\mathbb{E}\|e(Z)\|_2^4 \le C_4^2\Delta_h^2$ for a finite $C_4 \ge 1$.*

1. *For a Gaussian design with $\lambda_{\max}(\Theta) < 1$, and $\theta_1, \dots, \theta_d$ the eigenvalues of $\Theta$,*

$$\mathbb{E}\|e(Y_i)\|_2^2 \le C_4\,\kappa(\Omega)\,\Delta_h, \qquad \kappa(\Omega) = \prod_{j=1}^{d}\big(1-\theta_j^2\big)^{-1/4}.$$

   *The factor equals one at $\Theta = 0$, increases in each $\theta_j$, and diverges as $\lambda_{\max}(\Theta) \uparrow 1$.*
2. *For a deterministic design,*

$$\mathbb{E}\|e(Y_i)\|_2^2 \le C_4\Delta_h \exp\big(\tfrac{1}{2}\|B\widehat{V}x_i\|_2^2\big).$$

*Writing $c_i$ for the corresponding factor, Markov's inequality gives, for any $\delta' \in (0,1)$, with probability at least $1-\delta'$,*

$$\|E_Y\|_F \le \Big(\frac{C_4\Delta_h}{\delta'}\sum_i c_i\Big)^{1/2}. \tag{75}$$

*Proof of Proposition 35.* Under the reset protocol $Z_i \sim \mathcal{N}(0, I_d)$ and $Y_i = MZ_i + BE_i + B\widehat{V}x_i + \xi_i$. Stationarity of (2) gives $\mathrm{Cov}(MZ_i + BE_i + \xi_i) = I_d$, and the design is independent of $(Z_i, E_i, \xi_i)$, so $Y_i \sim \mathcal{N}(0, I_d + \Theta)$ for a Gaussian design with $\mathrm{Cov}(x_i) = \Omega$, and $Y_i \sim \mathcal{N}(B\widehat{V}x_i, I_d)$ for a deterministic design. This proves (74).

Let $P = \mathcal{N}(0, I_d)$ be the passive law, under which $\Delta_h$ is defined, and let $P_i$ be the law of $Y_i$. For nonnegative $f$, Cauchy–Schwarz gives

$$\mathbb{E}_{P_i} f = \mathbb{E}_P\Big[f\,\frac{dP_i}{dP}\Big] \le \big(\mathbb{E}_P f^2\big)^{1/2}\big(1+\chi^2(P_i\,\|\,P)\big)^{1/2}.$$

For centered Gaussians with covariances $\Sigma_1$ and $\Sigma_0$, completing the square gives

$$1+\chi^2 = |\Sigma_0|^{1/2}\,|\Sigma_1|^{-1}\,\big|2\Sigma_1^{-1} - \Sigma_0^{-1}\big|^{-1/2},$$

finite exactly when $2\Sigma_1^{-1} \succ \Sigma_0^{-1}$. With $\Sigma_0 = I_d$ and $\Sigma_1 = I_d + \Theta$, whose eigenvalues are $1+\theta_j$, this is finite exactly when $\theta_j < 1$ for all $j$, and

$$\begin{aligned} 1+\chi^2 &= \prod_{j=1}^{d}(1+\theta_j)^{-1}\Big(\frac{2}{1+\theta_j} - 1\Big)^{-1/2} \\ &= \prod_{j=1}^{d}\big((1+\theta_j)(1-\theta_j)\big)^{-1/2} = \prod_{j=1}^{d}\big(1-\theta_j^2\big)^{-1/2}. \end{aligned}$$

For a mean shift, the same computation with $P_i = \mathcal{N}(m_i, I_d)$ and $m_i = B\widehat{V}x_i$ gives $1+\chi^2 = \exp(\|m_i\|_2^2)$, which is finite for every design. Taking $f = \|e\|_2^2$ and applying the kurtosis bound,

$$\mathbb{E}\|e(Y_i)\|_2^2 \le \big(\mathbb{E}_P\|e\|_2^4\big)^{1/2}(1+\chi^2)^{1/2} \le C_4\Delta_h\,c_i,$$

with $c_i = \kappa(\Omega)$ in the Gaussian case and $c_i = \exp(\|m_i\|_2^2/2)$ in the deterministic case. Each Gaussian factor equals one at $\theta_j = 0$, is increasing in $\theta_j$ on $[0, 1)$, and the product diverges as any $\theta_j \uparrow 1$.

Finally $\mathbb{E}\|E_Y\|_F^2 = \sum_i \mathbb{E}\|e(Y_i)\|_2^2 \le C_4\Delta_h \sum_i c_i$, so Markov's inequality applied to $\|E_Y\|_F^2$ gives (75). □

Combining (75) with (68) gives a joint probability of at least $1 - \delta - \delta'$. For fixed designs the following corollary controls all batch terms in (66).

**Corollary 36** (Joint bound under stationary resets)**.** *Condition on first-stage estimates satisfying Theorem 33. Use the reset protocol and fourth-moment condition of Proposition 35, with a fixed design $X = [x_i]$ satisfying $G_X \succ 0$. For $\delta \in (0, 1)$ set $\eta = \delta/5$ and*

$$\begin{aligned} r_Z &= \sqrt{nd} + \sqrt{2\log(1/\eta)}, \\ r_E &= \sqrt{n\,\mathrm{tr}\,\Sigma} + \sqrt{2\|\Sigma\|_{\mathrm{op}}\log(1/\eta)}, \\ r_{E_Z} &= \sqrt{n\Delta_h/\eta}, \\ r_{E_Y} &= \left\{ \frac{C_4\Delta_h}{\eta} \sum_i \exp\!\big(\tfrac{1}{2}\|B\widehat{V}x_i\|_2^2\big) \right\}^{1/2}. \end{aligned}$$

*Define*

$$\begin{aligned} \overline{\mathcal{D}}_n &= r_{E_Y} + \|\widehat{G}\widehat{K} - \widehat{M}\|_{\mathrm{op}} r_{E_Z} \\ &\quad + (a_M + \|\widehat{G}\|_{\mathrm{op}} a_K) r_Z + a_G(r_E + \|X\|_F), \\ \overline{\mathcal{B}}_n &= a_G^F + \mathcal{N}_\eta(X) + \frac{\overline{\mathcal{D}}_n}{\sqrt{\lambda_{\min}(G_X)}} + s\|B_{\perp,0}\|_F. \end{aligned} \tag{76}$$

*With conditional probability at least $1 - \delta$ over the second-stage batch,*

$$\begin{aligned} \|\widehat{B} - QB\|_F &\le \overline{\mathcal{B}}_n, \\ \|\widehat{A} - QAQ^\top\|_F &\le a_M^F + \|\widehat{K}\|_{\mathrm{op}}\overline{\mathcal{B}}_n + \|QB\|_F a_K. \end{aligned} \tag{77}$$

*If the first-stage conditions hold with probability at least $1 - \delta_0$, the joint probability is at least $1 - \delta_0 - \delta$.*

*Proof of Corollary 36.* Condition on the first stage and the fixed design. The vectorizations of $Z_0$ and $\mathsf{E} = [E_i]$ are centered Gaussian, with covariance matrices $I_{nd}$ and $I_n \otimes \Sigma$, respectively. Gaussian concentration for the Euclidean norm and Jensen's inequality give

$$\|Z_0\|_F \le r_Z, \qquad \|\mathsf{E}\|_F \le r_E,$$

each with failure probability at most $\eta$. Stationarity gives $\mathbb{E}\|E_Z\|_F^2 = n\Delta_h$, so Markov's inequality gives $\|E_Z\|_F \le r_{E_Z}$ with the same failure probability. The deterministic case of Proposition 35 gives $\|E_Y\|_F \le r_{E_Y}$ with failure probability at most $\eta$. Since $W_0 = \mathsf{E} + \widehat{V}X$ and $\widehat{V}^\top\widehat{V} = I_q$,

$$\|W_0\|_F \le r_E + \|X\|_F.$$

These four events imply $\mathcal{D}_n \le \overline{\mathcal{D}}_n$. Intersect them with the Gaussian noise event in the proof of Theorem 33, using confidence input $\eta$. A union bound gives total failure probability at most $5\eta = \delta$, and the same pointwise least-squares decomposition proves (77). Averaging over the first stage and adding its failure probability proves the last assertion. □

### E.4 Exploration energy and representation transfer

Scaling the design strengthens the regression signal and shifts the next-state distribution. For $X_t = \sqrt{t}\,X_1$, the noise radius (67) decreases as $t^{-1/2}$. The transfer factor grows exponentially in $t\|B\widehat{V}x_i\|_2^2/2$ for a fixed probe. For Gaussian probes, $\kappa(t\Omega_0)$ grows with the eigenvalues of $\Theta$ and diverges at the spectral threshold. The resulting error bound therefore balances estimation noise against representation error under the shifted law.

**Corollary 37** (Interior optimum in exploration energy)**.** *Fix a design $X_1$ with $G_{X_1} \succ 0$ and scale it as $X_t = \sqrt{t}\, X_1$ for $t > 0$, holding the first stage and the realized $(Z_i, E_i)$ fixed. Assume $B\widehat{V} X_1 \neq 0$ and a nondegenerate representation, $\Delta_h > 0$. Then the bound obtained by substituting* (75) *for $\|E_Y\|_F$ in* (68) *is continuous in $t$ and diverges both as $t \downarrow 0$ and as $t \to \infty$, so it attains its minimum at a finite $t$. For a Gaussian design family $\Omega = t\Omega_0$, $\Omega_0 \succ 0$, couple the designs by scaling the same initial draws. The corresponding bound using the Gaussian transfer factor attains its minimum before $\lambda_{\max}(\Theta) \uparrow 1$.*

*Proof of Corollary 37.* Write $X_t = \sqrt{t}\, X_1$, so $G_{X_t} = tG_{X_1}$ and, by (67), $\mathcal{N}_\delta(X_t) = t^{-1/2}\mathcal{N}_\delta(X_1)$ and $\lambda_{\min}(G_{X_t})^{-1/2} = t^{-1/2}\lambda_{\min}(G_{X_1})^{-1/2}$.

Two contributions vary with $t$. The second-stage action residual is $W_0(t) = \mathsf{E} + \widehat{V}\sqrt{t}\, X_1$ with $\mathsf{E} = [E_i]$, so its normalized term is

$$\Phi(t) := \frac{a_G \|W_0(t)\|_F}{\lambda_{\min}(G_{X_t})^{1/2}} = \frac{a_G \big\|t^{-1/2}\mathsf{E} + \widehat{V} X_1\big\|_F}{\lambda_{\min}(G_{X_1})^{1/2}},$$

which is not proportional to $t^{-1/2}$. Writing $D_0 \geq 0$ for the remaining terms of $\mathcal{D}_n$, which the hypotheses do fix, and bounding $\|E_Y\|_F$ by (75),

$$\begin{aligned}\mathcal{B}_n(t) &\leq a_G^F + s\|B_{\perp,0}\|_F + \Phi(t) \\ &\quad + t^{-1/2}\Big[\mathcal{N}_\delta(X_1) + \frac{D_0 + R(t)}{\lambda_{\min}(G_{X_1})^{1/2}}\Big], \\ R(t) &= \Big(\frac{C_4\Delta_h}{\delta'}\sum_i c_i(t)\Big)^{1/2}.\end{aligned}$$

All terms are nonnegative and continuous on $(0, \infty)$.

As $t \downarrow 0$ the bracket is at least $\mathcal{N}_\delta(X_1) > 0$, so the last term diverges, and the bound with it. As $t \to \infty$, $\Phi(t)$ converges to the finite limit $a_G\|\widehat{V}X_1\|_F\lambda_{\min}(G_{X_1})^{-1/2}$, while for a deterministic design $c_i(t) = \exp(t\|B\widehat{V}x_i^{(1)}\|_2^2/2)$ and $B\widehat{V}X_1 \neq 0$ makes some exponent positive, so $R(t)$ grows like $e^{t\theta/4}$ for some $\theta > 0$. Since $\Delta_h > 0$ gives $R(t) > 0$, the product $t^{-1/2}R(t)$ diverges and the bound diverges. A continuous function on $(0, \infty)$ that diverges at both ends attains its minimum in the interior.

For the Gaussian family $\Omega = t\Omega_0$ the same argument applies on $(0, t_\star)$, where $t_\star = \sup\{t : \lambda_{\max}(tB\widehat{V}\Omega_0\widehat{V}^\top B^\top) < 1\}$ is finite because $B\widehat{V} \neq 0$; there $\Phi$ stays bounded and $\kappa \to \infty$ as $t \uparrow t_\star$.

If $\Delta_h = 0$ then $R \equiv 0$. If in addition the first stage is exact, so that $a_G = 0$ and $D_0 = 0$, the bound reduces to $a_G^F + s\|B_{\perp,0}\|_F + t^{-1/2}\mathcal{N}_\delta(X_1)$, which is strictly decreasing, and no interior minimum exists. □

The transfer condition depends on $B\widehat{V}$. The fiber permits arbitrary action effects on $S^\perp$, so a positive probe covariance alone does not bound the induced change in the next state distribution. Intervention samples estimate the excess covariance $\Theta$ in (74), up to encoder error; an additional bound on the missing action effect also controls this quantity. Each probability bound evaluates a design fixed independently of its evaluation batch.

Corollary 37 optimizes the energy scale of the stated upper bound. With exact representation and first stage estimates, this bound decreases as $t^{-1/2}$. Approximate first stage estimates retain nuisance terms. Theorem 30 optimizes allocation across directions at a fixed energy.

**Corollary 38** (Oracle reduction lower bound)**.** *Any joint procedure whose model class contains the exact-encoder, known-nuisance Gaussian subexperiment* (58) *has, under the weighted expected-energy budget of Theorem 30, minimax missing-effect risk at least*

$$\frac{\nu n\{\mathrm{tr}(W^{1/2})\}^2}{\mathcal{E}(n - q - 1)}$$

*when restricted to nonadaptive Gaussian covariance designs and $n > q + 1$. Thus the noise/energy term in* (68) *reduces to the exact minimax rate when the first-stage and representation errors vanish;*

*the additional terms are perturbations created by learning the coordinate, nuisances, and quotient subspace.*

*Proof of Corollary 38.* Restrict the joint model to instances with exact encoder, known $K_0, M_0, G_0, S$, and independent Gaussian interventions. Giving these quantities to an estimator can only decrease minimax risk. On this restricted experiment the sufficient residual is exactly (58); optimizing its exact conditional minimax risk under the total-energy constraint is Theorem 30. The displayed value is therefore a lower bound for every procedure on the larger joint problem. □

### E.5 Allocation with partial passive information

**Proposition 39** (Water-filling design under partial passive information)**.** *Let $\Lambda \succeq 0$ and $W \succ 0$ be $k \times k$ matrices that commute, with a common orthonormal eigenbasis and respective eigenvalues $\lambda_j, w_j$, and let $\mathcal{E} > 0$. Among $J \succeq 0$ with $\Lambda + J \succ 0$ and $\operatorname{tr}(WJ) \le \mathcal{E}$, the unique minimizer of $\operatorname{tr}[(\Lambda + J)^{-1}]$ is diagonal in that basis with*

$$j_j^\star = \left(c\, w_j^{-1/2} - \lambda_j\right)_+, \tag{78}$$

*where $c > 0$ solves $\sum_j w_j j_j^\star = \mathcal{E}$. If $k = q$ and $\Lambda = 0$, then $J_\star = \mathcal{E} W^{-1/2} / \operatorname{tr}(W^{1/2})$ and $J_\star/n$ is the covariance of Theorem 30.*

*Proof of Proposition 39.* Work in the common orthonormal eigenbasis of $W$ and $\Lambda$. Conjugating a feasible $J$ by any diagonal sign matrix preserves feasibility, cost and objective. Averaging over all such sign matrices gives $\operatorname{diag}(J)$, and convexity of the inverse trace implies that this averaging cannot increase the objective. Thus an optimum is diagonal. Writing $j_j$ for its entries, the problem is

$$\min_{j_j \ge 0} \sum_j \frac{1}{\lambda_j + j_j} \quad \text{subject to} \quad \sum_j w_j j_j \le \mathcal{E},$$

with $\lambda_j + j_j > 0$. A strictly positive feasible diagonal design exists because $\mathcal{E} > 0$. The objective diverges at a singular total-information matrix and is strictly convex on positive-definite total information, so its minimizer exists and is unique. Each coordinate objective decreases with $j_j$, hence the budget is active. With multipliers $\mu > 0$ and $\eta_j \ge 0$, the KKT conditions are

$$-\frac{1}{(\lambda_j + j_j)^2} + \mu w_j - \eta_j = 0, \qquad \eta_j j_j = 0.$$

On active coordinates, $\lambda_j + j_j = c\, w_j^{-1/2}$ with $c = \mu^{-1/2}$; on inactive coordinates, $\lambda_j \ge c\, w_j^{-1/2}$. Together these are (78), and $c$ is fixed by the active constraint. The sum $\sum_j w_j (c\, w_j^{-1/2} - \lambda_j)_+$ is continuous and strictly increasing in $c$ on the region where it is positive, so its value equals $\mathcal{E}$ at a unique $c > 0$.

When $\Lambda = 0$ on the entire $q$-dimensional design space, every coordinate is active, giving $c = \mathcal{E} / \operatorname{tr}(W^{1/2})$ and $J_\star = \mathcal{E} W^{-1/2} / \operatorname{tr}(W^{1/2})$. Dividing by $n$ gives the per-sample covariance in Theorem 30. For a random Gram matrix $G_X$ with $\mathbb{E} G_X = n\Omega$, operator convexity also gives

$$\mathbb{E} \operatorname{tr}[(\Lambda + G_X)^{-1}] \ge \operatorname{tr}[(\Lambda + n\Omega)^{-1}],$$

whenever the inverses and expectation are defined. The design objective above uses the expected accumulated information $J = n\Omega$. □

### E.6 Task-weighted quotient design

**Proposition 40** (When the quotient design is task-optimal)**.** *Let $W, H_{\text{task}} \in \mathbb{R}^{q \times q}$ be positive definite and $\mathcal{E} > 0$. The unique minimizer of $\operatorname{tr}(H_{\text{task}} J^{-1})$ over $J \succ 0$ with $\operatorname{tr}(WJ) \le \mathcal{E}$ is*

$$J_\star^H = \frac{\mathcal{E}\, W^{-1/2} N^{1/2} W^{-1/2}}{\operatorname{tr}(N^{1/2})}, \qquad N = W^{1/2} H_{\text{task}} W^{1/2}.$$

*Its per-sample covariance $J_\star^H / n$ coincides with the design of Theorem 30 exactly when $H_{\text{task}} = \alpha I_q$ for some $\alpha > 0$.*

*Proof of Proposition 40.* Write $J = W^{-1/2}\Psi W^{-1/2}$ with $\Psi \succ 0$. Then $\mathrm{tr}(H_{\mathrm{task}}J^{-1}) = \mathrm{tr}(N\Psi^{-1})$, where $N = W^{1/2}H_{\mathrm{task}}W^{1/2} \succ 0$, and $\mathrm{tr}(WJ) = \mathrm{tr}\,\Psi$. The Frobenius Cauchy–Schwarz inequality gives

$$\{\mathrm{tr}(N^{1/2})\}^2 = \langle \Psi^{-1/2}N^{1/2}, \Psi^{1/2}\rangle_F^2 \le \mathrm{tr}(N\Psi^{-1})\,\mathrm{tr}\,\Psi \le \mathcal{E}\,\mathrm{tr}(N\Psi^{-1}).$$

Equality holds precisely at $\Psi_\star = \mathcal{E}N^{1/2}/\,\mathrm{tr}(N^{1/2})$, which is positive definite. This proves the formula and uniqueness. For $H_{\mathrm{task}} = I_q$, $N = W$, so $J_\star = \mathcal{E}W^{-1/2}/\,\mathrm{tr}(W^{1/2})$. The general design agrees with this one exactly when $N^{1/2}$ is a positive scalar multiple of $W^{1/2}$; squaring and conjugating by $W^{-1/2}$ gives $H_{\mathrm{task}} = \alpha I_q$ for some $\alpha > 0$. □

## F Global feature geometry and local nonlinear identification

The affine fiber extends to controlled means that are linear in a known feature map. Let $\psi : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}^k$ be a known feature map with finite second moments, including a constant coordinate when an intercept is present, and consider

$$\mu_\Theta(z,u) = \Theta\psi(z,u), \qquad \Theta \in \mathbb{R}^{d\times k}.$$

The uncentered Gram matrix

$$G_\psi = \mathbb{E}[\psi(Z,U)\psi(Z,U)^\top], \qquad T_\psi = \mathrm{range}(G_\psi)$$

retains the constant term and its information about the basepoint.

**Theorem 41** (Global nonlinear feature-span identification)**.** *The complete behavior-law identified set in the feature-linear class is*

$$\mathcal{I}_\psi(\Theta) = \{\Theta + D : DP_{T_\psi} = 0\}. \tag{79}$$

*It has dimension $d(k - \mathrm{rank}\,G_\psi)$ and is a singleton iff $G_\psi \succ 0$. For the anchored squared prediction risk,*

$$\mathcal{R}_\psi(\Theta') - \mathcal{R}_\psi(\Theta) = \mathrm{tr}\{(\Theta'-\Theta)G_\psi(\Theta'-\Theta)^\top\}.$$

*At a query $v = (z,u)$, write $\ell_\psi(v) := \psi(v)^\top G_\psi^\dagger \psi(v)$. Its exact modulus is*

$$\begin{aligned}&\sup_{\substack{\mathcal{R}_\psi(\Theta')-\mathcal{R}_\psi(\Theta)\\ \le\varepsilon}} \|\mu_{\Theta'}(v) - \mu_\Theta(v)\|_2 \\ &\quad= \begin{cases}\sqrt{\varepsilon\,\ell_\psi(v)}, & \psi(v)\in T_\psi,\\ +\infty, & \psi(v)\notin T_\psi.\end{cases}\end{aligned} \tag{80}$$

*For finitely many exploration laws with positive sampling weights $\pi_j$, the pooled fiber is obtained by replacing $T_\psi$ with $\mathrm{span}_j\,\mathrm{range}(G_{\psi,j})$; point identification holds iff that span is $\mathbb{R}^k$.*

*Proof of Theorem 41.* For $D = \Theta' - \Theta$,

$$\mathbb{E}\|D\psi(V)\|_2^2 = \mathrm{tr}(DG_\psi D^\top) = \|DG_\psi^{1/2}\|_F^2.$$

This vanishes iff $D$ annihilates $\mathrm{range}(G_\psi^{1/2}) = \mathrm{range}(G_\psi) = T_\psi$, equivalently $DP_{T_\psi} = 0$. This proves the complete fiber. In a basis adapted to $T_\psi$, exactly $k - \mathrm{rank}\,G_\psi$ columns of $D$ are free, giving the dimension and singleton claims. Orthogonality of conditional-mean residual noise gives the anchored excess-risk identity.

For a query whose feature $x = \psi(v)$ lies in $T_\psi$, put $H = DG_\psi^{1/2}$. Then $Dx = HG_\psi^{\dagger/2}x$. The Frobenius ball $\|H\|_F^2 \le \varepsilon$ has evaluation norm $\sqrt{\varepsilon}\|G_\psi^{\dagger/2}x\|_2$. It is attained by the same rank-one construction as in the proof of Theorem 4. If $x \notin T_\psi$, choose $w = P_{T_\psi^\perp}x \neq 0$ and $D = taw^\top$. This has zero excess risk and query disagreement $ta\|w\|^2$, proving the infinite branch.

Under several laws, the pooled Gram is $G_{\mathrm{pool}} = \sum_j \pi_j G_{\psi,j}$. Since every term is positive semidefinite,

$$\ker G_{\mathrm{pool}} = \bigcap_j \ker G_{\psi,j}, \qquad \mathrm{range}\,G_{\mathrm{pool}} = \mathrm{span}_j\,\mathrm{range}\,G_{\psi,j}.$$

Applying the single-law result proves the final assertion. □

This result is global in $(z, u)$ and includes polynomial, Fourier, kernel dictionary, and features from the final layer of a neural network. For nonlinear parameterizations, the tangent Gram characterizes regular local identification.

Let $\theta \in \Theta \subset \mathbb{R}^k$ and let $\mu_\theta : \mathcal{V} \to \mathbb{R}^d$ be differentiable as a map into $L^2(P; \mathbb{R}^d)$. Write $J_\theta(v) \in \mathbb{R}^{d\times k}$ for its derivative and

$$G_\theta = \mathbb{E}[J_\theta(V)^\top J_\theta(V)].$$

**Theorem 42** (Nonlinear tangent-Gram theorem and converse)**.** *For every parameter direction $a$,*

$$\lim_{t\to 0} \frac{\mathbb{E}\|\mu_{\theta+ta}(V) - \mu_\theta(V)\|_2^2}{t^2} = a^\top G_\theta a. \tag{81}$$

*Thus the derivative of the behavior-law mean map is injective iff $G_\theta \succ 0$; its tangent non-identification space is exactly* $\ker G_\theta$.

*Suppose, on a ball $\|h\|_2 \le r_0$, the $L^2$ remainder obeys*

$$\|\mu_{\theta+h} - \mu_\theta - J_\theta h\|_{L^2(P)} \le (L/2)\|h\|_2^2. \tag{82}$$

*If $\lambda = \lambda_{\min}(G_\theta) > 0$, then*

$$\|\mu_{\theta+h} - \mu_\theta\|_{L^2(P)} \ge \{\sqrt{\lambda} - (L/2)\|h\|_2\}\|h\|_2. \tag{83}$$

*Hence the parameter is locally point identified on $0 < \|h\|_2 < \min(r_0, 2\sqrt{\lambda}/L)$ and is locally $\sqrt{\lambda}/2$-stable on the smaller ball $\|h\|_2 \le \min(r_0, \sqrt{\lambda}/L)$. Conversely, for every unit $a \in \ker G_\theta$,*

$$\|\mu_{\theta+ta} - \mu_\theta\|_{L^2(P)} \le (L/2)t^2, \tag{84}$$

*so no positive first-order lower-Lipschitz identification constant exists.*

*Across exploration laws $P_j$ with weights $\pi_j > 0$, the pooled tangent Gram is $\sum_j \pi_j G_{\theta,j}$ and regular local identification holds iff* $\operatorname{span}_j \operatorname{range}(G_{\theta,j}) = \mathbb{R}^k$.

*Proof of Theorem 42.* Fréchet differentiability in $L^2(P; \mathbb{R}^d)$ gives

$$\mu_{\theta+ta} - \mu_\theta = tJ_\theta a + o_{L^2}(t).$$

After division by $t$, continuity of the squared $L^2$ norm proves (81). Moreover, $a^\top G_\theta a = \|J_\theta a\|_{L^2(P)}^2$, so the derivative kernel is exactly $\ker G_\theta$, and it is trivial precisely when $G_\theta \succ 0$.

Under (82), the reverse triangle inequality gives

$$\|\mu_{\theta+h} - \mu_\theta\|_{L^2(P)} \ge \|J_\theta h\|_{L^2(P)} - (L/2)\|h\|^2 \ge \sqrt{\lambda}\|h\| - (L/2)\|h\|^2.$$

This is (83) and proves both stated radii. If $a \in \ker G_\theta$, then $J_\theta a = 0$ in $L^2(P)$; substituting $h = ta$ in the remainder bound proves (84).

Finally, the derivative into the weighted direct sum of the $L^2(P_j; \mathbb{R}^d)$ spaces has squared norm $a^\top(\sum_j \pi_j G_{\theta,j})a$. The positive-semidefinite kernel/range identity used in the feature proof establishes the pooled span criterion. □

# G Controlled systems and robot manipulation

## G.1 Controlled systems and acquisition

**Systems and observations.** Eight physical systems have four state and four action coordinates. In a common action basis, two residual standard deviations are $0.18$; the other two are $0$, $0.02$, $0.06$ or $0.18$, defining missing, weak, anisotropic and isotropic coverage. All conditions retain the same dynamics, feedback, process noise and control objective in coordinates whitened under the stationary law with missing coverage. Observations apply $y_i = z_i + 0.16z_i^3$, optionally followed by $y_i \leftarrow y_i + 0.22\tanh(y_{i-1})$ for $i > 1$, and orthogonal mixing. Two training seeds and two observation maps give 32 cases per condition.

**Representation and screening.** Each case uses 3,072 training, 4,096 validation and 4,096 geometry transitions. The predictive encoder has two identity-initialized tanh residual blocks of width 32, amplitude $0.8$ and unit spectral-norm bounds. Training uses empirical whitening and 800 full-batch Adam steps at learning rate $0.002$. Checkpoint selection every 50 steps, including step zero, minimizes passive validation prediction risk normalized by current-state covariance.

ConditionalGap fits an affine behavior predictor on one data half and computes action residual covariance on the other. It selects the lower side of the largest adjacent eigenvalue gap when the ratio is at least two and the gap is at least $0.25$ times the largest eigenvalue. The two halves must select the same dimension with projector distance at most $0.25$; otherwise acquisition uses the full action space. GlobalGap applies this rule to raw actions. GlobalMatchedDim uses the smallest raw-action eigenvectors with the dimension selected by ConditionalGap. Other policies are FullBalanced, RandomFull and NoNewData.

**Acquisition and estimation.** Methods share reset anchors, passive data and random draws. Antipodal probes have norm $0.8$ and independent process noise within each pair. Budgets of 24 and 96 transitions match perturbation and command energy across methods. PairedTwoStage first fits action effects in the span of paired action differences, then holds them fixed while fitting the complement, state dynamics and intercept on passive data. Ridge regularization is $0.001$. EqualMassJoint fits all coefficients jointly with equal total weights for passive and intervention data. We also compare a linear encoder of observations.

**Control and aggregation.** A six-step quadratic controller runs for 32 steps on 32 common evaluation episodes, with action norm capped at four. Cost is mean squared physical tracking error plus $0.05$ times squared action norm. Paired relative cost reductions are averaged over seeds and maps within each system, then equally over the eight systems. The primary comparison uses the predictive encoder, PairedTwoStage and 24 transitions. Its 98.75% intervals use 20,000 bootstrap resamples of whole systems, with Bonferroni adjustment across four conditions.

Table 2: Acquisition policies across coverage conditions, under the shared predictive encoder and PairedTwoStage at 24 transitions. Entries are mean system-level percentage reductions in closed-loop cost relative to balanced full-space acquisition, so positive values are better and exact zeros denote policies that fall back to the same full-space acquisition. The Directed intervals at $98.75\%$ are $[26.55, 78.33]$, $[26.61, 78.26]$ and $[36.27, 77.85]$ for missing, weak and anisotropic coverage.

| | Coverage deficient | | | Complete |
|---|---|---|---|---|
| Policy | Unexcited | Weak full | Anisotropic | Isotropic |
| NoNewData | $-121.05$ | $-204.36$ | $-419.91$ | $+55.29$ |
| RandomFull | $-60.59$ | $-60.68$ | $-61.43$ | $-61.86$ |
| GlobalGap | 0.00 | 0.00 | 0.00 | $+25.58$ |
| GlobalMatchedDim | $-136.51$ | $-230.56$ | $-400.23$ | 0.00 |
| ConditionalGap | $\mathbf{+53.34}$ | $\mathbf{+53.33}$ | $\mathbf{+57.19}$ | 0.00 |

Conditional residuals recover the missing and weak subspaces in all 32 cases per condition, with mean projector error $0.011$. The raw-action screen selects the full space in these conditions. GlobalMatchedDim gives higher cost than FullBalanced despite using the same number of probe directions as Directed, supporting the importance of their orientation. EqualMassJoint also gives positive direction effects of $40.61\%$, $40.65\%$ and $44.47\%$ under missing, weak and anisotropic coverage. Across the two encoders, fitters and budgets, mean reductions in these conditions range from $9.27\%$ to $57.19\%$.

### G.2 Replication on separate physical systems

Eight additional systems, with two training seeds and two observation maps, test missing coverage using the same predictive encoder, screen, paired probes and controller. System means receive equal weight; $95\%$ intervals resample the eight systems with 20,000 bootstrap draws. Directed reduces cost by $38.46\%$ at 24 transitions, with interval $[25.88, 51.87]$, and $17.81\%$ at 96, with interval $[6.38, 31.64]$. The primary comparison uses 24 transitions. EqualMassJoint yields respective mean reductions of $34.75\%$ and $17.58\%$, supporting the benefit across both fitting rules.

### G.3 Evaluating energy allocation

We compare uniform and weighted allocation within a common subspace $\widehat{V}$, using the residual estimator in (65). For action cost $W_a \succ 0$, write $W_q = \widehat{V}^\top W_a \widehat{V}$ and let $\mathcal{E}$ denote total weighted probe energy. The two designs have Gram matrices

$$G_{\text{uniform}} = \frac{\mathcal{E}}{\operatorname{tr}(W_q)} I_q, \qquad G_{\text{weighted}} = \frac{\mathcal{E}}{\operatorname{tr}(W_q^{1/2})} W_q^{-1/2}.$$

Deterministic sign frames realize each Gram matrix with weighted energy $\mathcal{E}/n$ per transition. Both methods share first-stage estimates, reset states, behavior residuals and process noise. We measure squared Frobenius error of the fitted action effect, $\|\widehat{B} - QB\|_F^2$. The alignment $Q$ is used for evaluation in learned coordinates.

**Exact coordinates.** Sixteen new physical systems have four state and action coordinates and two missing action directions. Each system supplies two independent first-stage fits from 768 passive transitions. Ordinary least squares estimates $K, M, G$, and the bottom two residual eigenvectors define $\widehat{V}$. An additional reference uses the true first-stage parameters and subspace. We compare equal costs $W_a = I_4$ with $W_a = I_4 + V \operatorname{diag}(-0.8, 0.8) V^\top$, where $V$ spans the true missing subspace. Each configuration uses 2,048 independent acquisition repetitions.

With estimated first-stage parameters, $n = 96$ and $\mathcal{E} = 15.36$, weighted allocation lowers mean error from $0.881113$ to $0.751681$, with a mean paired reduction of $14.80\%$ (Table 3). Probe variance decreases by $19.86\%$, compared with the $20\%$ reduction predicted by Proposition 34. Equal costs produce identical designs and estimates for both methods. Errors are averaged over acquisition repetitions and first-stage fits within each system. Percentage reductions average $1 - R_{\text{weighted}}/R_{\text{uniform}}$ over systems. Tables report nominal $95\%$ intervals from 20,000 bootstrap resamples of whole systems.

Table 3: Allocation in exact coordinates at $n = 96$ and $\mathcal{E} = 15.36$. Uniform and Weighted report mean squared action-effect error. Reduction is the mean paired percentage over 16 systems, with a $95\%$ system bootstrap interval. First stage denotes known parameters or estimates from passive data.

| First stage | Costs | Uniform | Weighted | Reduction (%) |
|---|---|---|---|---|
| Known | Equal | 0.645798 | 0.645798 | 0.00 [0.00, 0.00] |
| Known | Unequal | 0.645798 | 0.517373 | 19.89 [19.76, 20.00] |
| Estimated | Equal | 0.881113 | 0.881113 | 0.00 [0.00, 0.00] |
| Estimated | Unequal | 0.881113 | 0.751681 | 14.80 [14.18, 15.49] |

**Separating energy and transition count.** At $n = 96$, we vary $\mathcal{E}$ from 3.84 to 245.76 with independent acquisition streams across energies. The slopes of log probe variance against log energy are $-1.0007$ for uniform and $-1.0009$ for weighted allocation. At the largest energy, weighted allocation has total error $0.261945$, comprising projection error $0.229502$ and probe variance $0.032443$. At fixed $\mathcal{E} = 15.36$, the same Gram matrix is realized using 24, 48, 96 or 192 transitions. Weighted error ranges from $0.749626$ to $0.751681$ (Table 4). These measurements follow the decomposition into a fixed projection term and a variance term controlled by the probe Gram matrix.

### G.4 Evaluating allocation with learned representations

We use 64 frozen predictive encoders from 16 physical systems, two observation maps and two training seeds under missing coverage. The observation maps use $z + 0.16 \tanh(z)$, with optional bounded triangular coupling and orthogonal mixing. We cross exact and learned coordinates with true and estimated subspaces. Exact and learned coordinates use the same physical transitions for each probe design. The estimated subspace contains the bottom two eigenvectors of the covariance of action residuals from the fitted MLP behavior predictor. Each coordinate choice fits $K, M, G$ on the same 3,072 passive transitions; these estimates are shared across subspaces and designs. Least squares uses a relative singular value cutoff of $10^{-7}$.

The grid combines $n \in \{24, 96\}$, $\mathcal{E} \in \{15.36, 61.44\}$, both allocations, and costs $I_4$ or $\operatorname{diag}(0.2, 0.2, 1.8, 1.8)$ in ambient action coordinates. Each configuration uses 512 acquisition

Table 4: Separate energy and transition sweeps in exact coordinates with estimated first-stage parameters and unequal costs. Uniform and Weighted are total mean squared errors. Weighted variance is the probe contribution; the common projection term is $0.229502$.

| $\mathcal{E}$ | $n$ | Uniform | Weighted | Weighted variance |
|---|---|---|---|---|
| Fixed transition count $n = 96$ | | | | |
| 3.84 | 96 | 2.831365 | 2.312603 | 2.083102 |
| 7.68 | 96 | 1.532006 | 1.270988 | 1.041486 |
| 15.36 | 96 | 0.881113 | 0.751681 | 0.522179 |
| 30.72 | 96 | 0.554649 | 0.489670 | 0.260169 |
| 61.44 | 96 | 0.391890 | 0.359401 | 0.129900 |
| 122.88 | 96 | 0.310822 | 0.294516 | 0.065014 |
| 245.76 | 96 | 0.270053 | 0.261945 | 0.032443 |
| Fixed energy $\mathcal{E} = 15.36$ and Gram matrix | | | | |
| 15.36 | 24 | 0.879961 | 0.750394 | 0.520893 |
| 15.36 | 48 | 0.878804 | 0.749626 | 0.520124 |
| 15.36 | 96 | 0.881113 | 0.751681 | 0.522179 |
| 15.36 | 192 | 0.881910 | 0.750983 | 0.521481 |

repetitions. At $n = 24$ and $\mathcal{E} = 61.44$, weighted allocation with learned coordinates and estimated directions reduces error by $7.37\%$ and probe variance by $9.83\%$ (Table 5). Quadrupling energy at the same transition count gives a probe variance ratio of $0.2501$. The four coordinate and subspace combinations have weighted errors between $0.1915$ and $0.1933$. The mean projector distance between estimated and true subspaces is $0.003892$. For each encoder, paired reductions compare risks averaged over acquisition repetitions. We then average over observation maps and training seeds within each physical system. Bootstrap intervals resample the $16$ systems.

Table 5: Allocation with unequal ambient action costs across $64$ frozen encoders from $16$ physical systems. Uniform and Weighted report mean squared action-effect error. Reductions are paired within each encoder and averaged within systems, with $95\%$ system bootstrap intervals. The lower panel crosses coordinate and subspace choices on the same systems.

| $\mathcal{E}$ | $n$ | Uniform | Weighted | Reduction (%) |
|---|---|---|---|---|
| Learned coordinates and estimated subspace | | | | |
| 15.36 | 24 | 0.685334 | 0.625852 | 9.13 [6.38, 11.94] |
| 15.36 | 96 | 0.685651 | 0.625395 | 9.27 [6.44, 12.17] |
| 61.44 | 24 | 0.207587 | 0.192946 | 7.37 [5.10, 9.72] |
| 61.44 | 96 | 0.206313 | 0.191567 | 7.48 [5.15, 9.90] |

| Weighted allocation at $n = 24$, $\mathcal{E} = 61.44$ | | |
|---|---|---|
| Coordinates | Subspace | Mean squared error |
| Exact | True | 0.191534 |
| Exact | Estimated | 0.191523 |
| Learned | True | 0.193246 |
| Learned | Estimated | 0.192946 |

### G.5 Testing spectral screening across dimensions

We vary the residual spectrum and weak direction count at action dimensions $p \in \{4, 8, 16\}$, with state dimension $d = p$. Eight development systems and 16 separate evaluation systems each cover 12 configurations. At fixed system and dimension, dynamics, feedback and process noise are shared across spectra. Stationary reset covariances are recomputed for each behavior law. Independent geometry samples of size $256$ or $1{,}024$ provide cross-fitted linear action residuals in exact coordinates. The ratio screen selects the largest eligible adjacent eigenvalue ratio, requiring a ratio of at least two, a gap of at least $0.25$ times the largest eigenvalue, agreement of split dimensions, and projector distance at most $0.25$. ConditionalGap provides the comparison based on the largest absolute gap.

All acquisition methods share first-stage fits from $768$ passive transitions and use the residual estimator in (65). Each method receives $96$ transitions and energy $61.44$, realized with probe norm $0.8$. Each evaluation system uses two first-stage fits and $512$ acquisition repetitions per configuration. Clear

spectra have weak to strong variance ratio $1 : 100$; the exact case has zero weak variance. Boundary spectra have ratio $1 : 2$, decaying spectra have geometrically spaced eigenvalues from $0.1$ to $1$, and isotropic spectra have equal eigenvalues. All spectra are scaled by $0.18^2$.

Both screens recover the weak dimension in every clear and exact configuration at both geometry sample sizes. With $1{,}024$ geometry samples and half the action directions weak, mean error reductions against FullBalanced are $44.41\%$, $59.03\%$ and $67.15\%$ at dimensions $4$, $8$ and $16$ (Table 6). At dimension eight, the reductions are $71.36\%$ for one weak direction and $20.00\%$ for seven. Concentrating energy yields larger savings when fewer directions need additional excitation. Boundary, decaying and isotropic spectra activate full-space acquisition at both sample sizes, reproducing the FullBalanced design and risk. Both screens select the same dimensions throughout this grid. The numerical supplement contains both geometry sample sizes, all allocation grids, and the complete coordinate and subspace comparisons.

Table 6: Spectral screening on $16$ evaluation systems using $1{,}024$ geometry samples. Configurations give action dimension $p$ and weak dimension $q$; $\cdot$ denotes a spectrum without a designated weak dimension. Screen and Full report mean squared action-effect error at $n = 96$, $\mathcal{E} = 61.44$. Reduction compares Screen with FullBalanced, with $95\%$ system bootstrap intervals. Both screens give the reported results.

| Spectrum $(p, q)$ | Fallback (%) | Screen | Full | Reduction (%) |
|---|---|---|---|---|
| Clear $(4, 2)$ | 0 | 0.361 | 0.651 | 44.41 [38.81, 49.93] |
| Clear $(8, 4)$ | 0 | 2.155 | 5.259 | 59.03 [57.92, 60.05] |
| Clear $(16, 8)$ | 0 | 14.156 | 43.091 | 67.15 [66.89, 67.40] |
| Clear $(8, 1)$ | 0 | 1.506 | 5.259 | 71.36 [70.15, 72.65] |
| Clear $(8, 7)$ | 0 | 4.207 | 5.259 | 20.00 [19.21, 20.72] |
| Exact $(8, 4)$ | 0 | 2.133 | 5.232 | 59.24 [58.25, 60.19] |
| Boundary $(8, 4)$ | 100 | 5.259 | 5.259 | 0.00 [0.00, 0.00] |
| Boundary $(16, 8)$ | 100 | 43.091 | 43.091 | 0.00 [0.00, 0.00] |
| Decay $(8, \cdot)$ | 100 | 5.259 | 5.259 | 0.00 [0.00, 0.00] |
| Decay $(16, \cdot)$ | 100 | 43.091 | 43.091 | 0.00 [0.00, 0.00] |
| Isotropic $(8, \cdot)$ | 100 | 5.259 | 5.259 | 0.00 [0.00, 0.00] |
| Isotropic $(16, \cdot)$ | 100 | 43.091 | 43.091 | 0.00 [0.00, 0.00] |

### G.6 Robot manipulation

**Data and models.** ManiSkill3 `PickCube-v1` uses state observations and `pd_ee_delta_pos` control. The three translation coordinates form the action space for acquisition. Common training data combine behavior transitions with paired horizontal probes, leaving vertical motion unexcited. The dynamics input excludes the cached flag for grasp contact, which is used for selecting evaluation starts. Models share normalizers, validation data and three hidden layers of width 128. Training uses 180 epochs, batch size 256 and learning rate $10^{-3}$. The behavior estimator uses 15 nearest neighbors and translation limit $0.7$, with 280 training, 60 validation and 60 test episodes.

**Acquisition rules.** The five policies are RandomFull, GlobalMean, EnsembleDisagreement, Directed and the oracle vertical line. GlobalMean selects the smallest covariance direction after subtracting the global action mean. Directed uses conditional residuals from the behavior estimator. EnsembleDisagreement selects the line with largest normalized signed action-effect variance among five models on a grid of 263 lines. Five acquisition repetitions share anchors and opposite perturbations of norm $0.20$. Each supplies 128 transitions; control is evaluated at prefixes of 16, 32 and 64 transitions using three model seeds.

**Evaluation and uncertainty.** From 400 held-out successful demonstrations, we select the earliest state with four consecutive grasped frames, object height at least $0.04$ m and remaining goal distance at least $0.06$ m. All 330 eligible episodes enter evaluation, comprising 243 starts dominated by vertical motion and 87 by horizontal motion. MPC compares the behavior action with positive and negative Cartesian perturbations, predicts four steps, executes one action and replans for at most 16 steps. We average budgets within each acquisition repetition, model seed and episode, then independently bootstrap these three axes. The four success and distance contrasts against RandomFull and EnsembleDisagreement use 50,000 draws and 98.75% Bonferroni intervals.

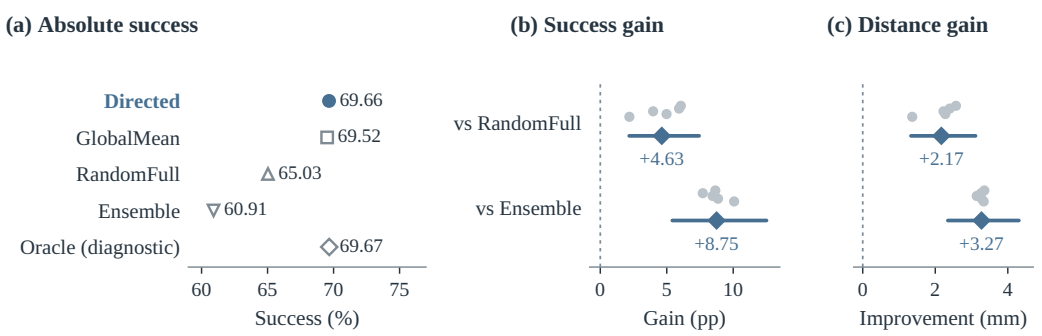



Figure 5: PickCube control over 330 starts. (a) Absolute success. (b,c) Paired effects show acquisition repetitions in gray and pooled means with 98.75% intervals in blue. Bootstrap crosses acquisition, model seed and episode after budget aggregation.

Table 7: PickCube control results over 330 episodes. The last column is the distance advantage of Directed over that policy, with its 98.75% interval from the crossed bootstrap. Oracle is diagnostic only, and Figure 5 shows the matching success contrasts.

| Acquisition | Success | Distance (mm) | Gain (mm) |
|---|---|---|---|
| RandomFull | 65.03% | 30.163 | 2.168 [1.330, 3.111] |
| GlobalMean | 69.52% | 28.033 | 0.038 [−0.067, 0.169] |
| EnsembleDisagreement | 60.91% | 31.269 | 3.275 [2.348, 4.305] |
| **Directed** | **69.66%** | **27.995** | NA |
| Oracle | 69.67% | 27.992 | −0.003 |

The improvement is concentrated in starts dominated by vertical motion, where Directed gains 5.81 success percentage points with interval [2.79, 9.33] and 2.867 mm in final distance with interval [1.750, 4.081]. The corresponding estimates for horizontal motion are 1.33 points and 0.215 mm.

**Direction estimation and alignment.** On common held-out anchors, conditional behavior subtraction gives projector error 0.01428, compared with 0.09380 for global mean subtraction and zero when the recorded action is subtracted. The conditional and global regressions have respective behavior prediction RMSEs of 0.05492 and 0.09152. Figure 3b varies the probe line's projector distance from the vertical under the same paired acquisition and budgets. Each point averages three model seeds and three budgets within 186 eligible episodes dominated by vertical motion, then weights episodes equally. The curve connects accurate alignment to improved success in this task.

# H Offline behavior coverage and counterfactual responses

## H.1 Datasets, inputs and simulator restoration

We evaluate medium and expert policies for HalfCheetah, Hopper and Walker2d from the Minari Gymnasium v5 collections (Younis et al., 2024), and human demonstrations for Adroit Door and Pen from D4RL v2 (Fu et al., 2020). Locomotion states are restored from recorded positions and velocities. For Hopper and Walker2d, evaluation uses observations strictly inside the velocity clipping bounds $[-10, 10]$, as summarized in Table 8. Recorded actions reproduce the next observations to tolerance $10^{-6}$ on these states. Adroit states are restored from the stored initial state and action history. Because its observations omit velocities, we compare current observations with a history input for both screening and dynamics prediction.

## H.2 Conditional Action Coverage and Response Error

An MLP with three hidden layers of width 256 predicts actions, with five episode folds and an inner episode split for early stopping. Covariance eigenvalues come from held-out action residuals.

We compare the bottom and top $k = \min(3, \lfloor p/2 \rfloor)$ conditional eigenvectors with the bottom $k$ marginal eigenvectors and three random directions. Table 8 reports locomotion spectra. Their smallest eigenvalues are 8.1% to 54.2% of the largest, placing the comparison in a regime of uneven full coverage.

Table 8: Locomotion behavior coverage. The residual covariance uses MLP predictions on held-out episodes. All eigenvalues exceed 1% of the largest; $\theta_{\max}$ is the largest principal angle (degrees) between the bottom-$k$ conditional and marginal subspaces, $k = \min(3, \lfloor p/2 \rfloor)$. Eligible is the fraction of unclipped states available for exact simulator restoration.

| Dataset | $p$ | $\lambda_{\min}/\lambda_{\max}$ | $\theta_{\max}$ | Eligible (%) |
|---|---|---|---|---|
| halfcheetah-medium | 6 | 0.2529 | 85.0 | 100.00 |
| halfcheetah-expert | 6 | 0.0811 | 83.3 | 100.00 |
| hopper-medium | 3 | 0.5421 | 82.6 | 87.60 |
| hopper-expert | 3 | 0.3116 | 58.1 | 71.72 |
| walker2d-medium | 6 | 0.1478 | 83.9 | 62.70 |
| walker2d-expert | 6 | 0.1459 | 78.4 | 56.76 |

Locomotion models use an 80/20 episode split. Adroit uses five outer episode folds, each supplying 200 evaluation states, with behavior regressions, directions and dynamics fitted only on outer training episodes. Each dynamics ensemble contains five MLPs with four hidden layers of width 256. Training uses batch size 1,024, learning rate $10^{-3}$, weight decay $10^{-5}$ and at most 60,000 updates. Early stopping uses 10% of training episodes and patience eight. Architecture selection uses one-step error on the inner validation split on Hopper-medium; the architecture is shared across datasets. At 1,000 eligible held-out states per dataset, simulator branches use $a^{\pm} = \mathrm{clip}(a \pm 0.3v, [-1, 1]^p)$, where $a$ is the recorded action and $v$ a unit direction. Models receive the executed actions. Predicted and true differences in next observations are standardized by the training standard deviations of observation increments. For a direction set $D$, the relative response error and directional ratio are

$$e(D) = \left\{ \frac{\sum_{i,v \in D} \|\widehat{\Delta}_{iv} - \Delta_{iv}\|_2^2}{\sum_{i,v \in D} \|\Delta_{iv}\|_2^2} \right\}^{1/2}, \qquad R = \frac{e(D_{\mathrm{bottom}})}{e(D_{\mathrm{top}})}.$$

Here $\widehat{\Delta}$ uses the ensemble mean and $\Delta$ the simulator branches. Values $R > 1$ indicate larger relative error in less excited directions. A predictor assigning zero action response has $e(D) = 1$.

Intervals use 2,000 bootstrap resamples of evaluation episodes and have nominal 95% coverage separately for each dataset. Locomotion intervals condition on the fitted models and episode split. Adroit aggregation over optimization seeds is specified below. We also evaluate marginal and random directions and sensitivity to action clipping.

### H.3 Locomotion results and response scale

Five of the six locomotion datasets have larger relative response errors along less excited directions, with the largest contrasts on Hopper (Table 9). Halfcheetah-expert has the reverse ordering. The association between excitation and fitted response error depends on both prediction quality and response amplitude.

To separate these factors, let $R_{\mathrm{abs}}$ be the bottom/top ratio of absolute response RMSE in the same standardized coordinates, and $Q$ the corresponding RMS ratio of true responses. The identity $R = R_{\mathrm{abs}}/Q$ expresses how response amplitude enters the reported metric. Table 9 gives this descriptive decomposition. Halfcheetah-expert has similar absolute errors across the two direction groups ($R_{\mathrm{abs}} = 1.03$), while its true responses are larger in the less excited directions ($Q = 1.33$), giving $R = 0.77$.

Excluding clipped direction pairs preserves the error ordering for Hopper and Walker2d, while Halfcheetah-expert moves to $R = 1.041$ (Table 10). This sensitivity uses the subsets where both branch actions remain within their bounds.

Table 9: Locomotion response errors and their amplitude decomposition. The relative RMSE ratio is $R = e_{\text{bottom}}/e_{\text{top}} = R_{\text{abs}}/Q$, where $R_{\text{abs}}$ compares absolute response RMSE and $Q$ compares true response RMS, using the same training coordinate standardization. Brackets give nominal 95% episode bootstrap intervals conditional on fitted models; $N_{\text{ep}}$ counts evaluation episodes.

| Dataset | $N_{\text{ep}}$ | $e_{\text{bottom}}$ | $e_{\text{top}}$ | $R$ | $R_{\text{abs}}$ | $Q$ |
|---|---|---|---|---|---|---|
| Hopper-medium | 255 | 0.0798 | 0.0189 | 4.215 [3.567, 5.006] | 1.886 [1.600, 2.238] | 0.448 |
| Hopper-expert | 211 | 0.1256 | 0.0451 | 2.785 [1.934, 4.197] | 1.077 [0.741, 1.634] | 0.387 |
| Walker2d-medium | 200 | 0.0911 | 0.0768 | 1.186 [1.071, 1.326] | 1.418 [1.278, 1.584] | 1.195 |
| Walker2d-expert | 196 | 0.0914 | 0.0680 | 1.344 [1.230, 1.483] | 1.553 [1.426, 1.711] | 1.155 |
| Halfcheetah-medium | 199 | 0.3120 | 0.2876 | 1.085 [1.014, 1.161] | 1.332 [1.238, 1.429] | 1.228 |
| Halfcheetah-expert | 199 | 0.2138 | 0.2768 | 0.773 [0.663, 0.905] | 1.030 [0.879, 1.210] | 1.332 |

Table 10: Directional error ratios and sensitivity to action clipping. $R_{\text{marg}} = e_{\text{bottom}}/e_{\text{marginal}}$ and $R_{\text{rand}} = e_{\text{bottom}}/e_{\text{random}}$ have nominal 95% episode-bootstrap intervals. $R_{\text{unclipped}}$ excludes direction pairs for which either action clips; its bottom and top denominators can involve different state subsets. The last column counts retained bottom/top branch pairs; intervals resample episodes.

| Dataset | $R_{\text{marg}}$ | $R_{\text{rand}}$ | $R_{\text{unclipped}}$ | Pairs |
|---|---|---|---|---|
| Hopper-medium | 3.222 [2.524, 4.311] | 2.382 [1.990, 2.749] | 3.662 | 625/694 |
| Hopper-expert | 1.912 [1.614, 2.312] | 2.381 [1.993, 2.758] | 2.253 | 467/655 |
| Walker2d-medium | 1.150 [1.067, 1.246] | 0.975 [0.921, 1.036] | 1.222 | 1316/1092 |
| Walker2d-expert | 1.108 [1.031, 1.213] | 0.970 [0.907, 1.039] | 1.587 | 1608/1239 |
| Halfcheetah-medium | 1.031 [0.960, 1.108] | 1.000 [0.915, 1.103] | 1.071 | 408/658 |
| Halfcheetah-expert | 1.121 [1.015, 1.233] | 1.064 [0.969, 1.158] | 1.041 | 777/1010 |

### H.4 Observation information and directional error in Adroit

History input is $(o_t, o_{t-1}, o_{t-2}, a_{t-1}, a_{t-2})$ for both the conditional behavior regressor and the dynamics ensemble. Missing history at episode starts repeats the initial observation and uses zero past actions. Models predict the current observation increment. Both input modes share five outer episode folds, 1,000 evaluation states and the dynamics architecture. Directions are estimated separately from each mode's training data. Each input mode uses three optimization seeds.

The energies of response errors and true responses are pooled over seeds before computing $R$. Bootstrap draws resample episodes with shared weights across seeds and input modes. The independent sampling unit is the episode; all optimization seeds use the same 25 episodes per task. Table 11 reports pooled estimates; Figure 6 shows the individual seeds.

Table 11: Observation and history inputs on the same Adroit episodes. Response energies are pooled across three optimization seeds before computing $R$. Intervals resample the 25 episodes with shared weights across modes and seeds. Bottom and top columns report relative response RMSE.

| Task | Input | $R$ [95% interval] | Bottom | Top | One-step MSE |
|---|---|---|---|---|---|
| Door | Observation | 0.734 [0.703, 0.764] | 0.998 | 1.361 | 0.431 |
| Door | History | 1.234 [1.191, 1.287] | 0.997 | 0.808 | 0.348 |
| Pen | Observation | 1.010 [0.991, 1.028] | 0.998 | 0.988 | 0.184 |
| Pen | History | 1.027 [1.021, 1.034] | 0.997 | 0.971 | 0.160 |

History improves prediction on both tasks. On Door, the change in the directional ratio comes primarily from better responses in the more excited directions, while the less excited directions retain error near one. Pen shows a smaller change in directional errors. These comparisons separate the contribution of observation information from the remaining difficulty of estimating weakly excited action effects.

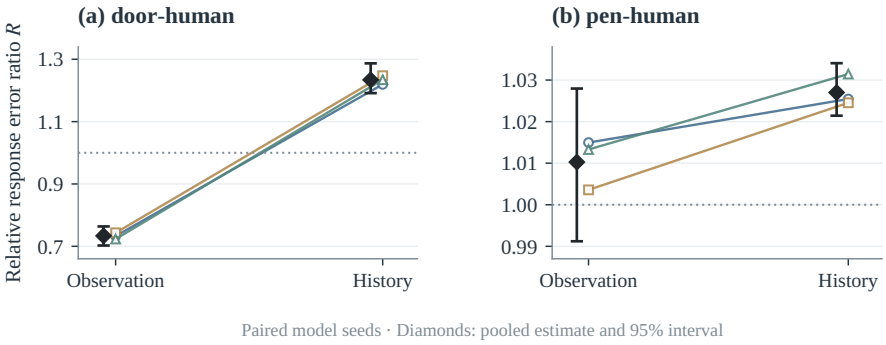



Figure 6: Observation and history inputs on Adroit. Colored lines connect the same optimization seed across input modes. Black diamonds pool response energies over seeds; bars are nominal 95% episode bootstrap intervals. The dotted line marks equal relative response error in the two direction groups.

# I CONTROLLED DEPARTURES FROM THE MODELING ASSUMPTIONS

## I.1 OBSERVATION MODEL AND DEPARTURES

We test how departures from the controlled assumptions affect directed acquisition and how estimation error reaches the controller. Observations use the invertible scalar map $z + 0.16\tanh(z)$, optional bounded triangular coupling and orthogonal mixing. The map has linear tail growth, giving finite observation covariances under the Gaussian and Student noise laws used here. Student-$t_3$ noise has a finite second moment and an infinite fourth moment.

The study uses 16 physical systems, two training seeds, two observation maps, and missing and weak coverage. The Gaussian reference and two levels of each of five departures give 22 condition and coverage groups and 1,408 encoder cases. Each case uses the controlled study's passive splits, perturbation norm 0.8, and budgets of 24 and 96 transitions. Table 12 defines the departures.

Table 12: Fixed departures from the controlled linear Gaussian setup. Percentages measure the nonlinear conditional-mean variance left after its best affine predictor. V4 preserves unconditional residual covariance.

| | Departure | Mild | Strong |
|---|---|---|---|
| V1 | Nonlinear latent dynamics | 1% | 4% |
| V2 | Nonlinear behavior mean | 2% | 8% |
| V3 | Student-$t$ noise with matching covariance | 5 degrees of freedom | 3 degrees of freedom |
| V4 | State-dependent residual availability | Region probability 0.5 | Region probability 0.1 |
| V5 | Acquisition without paired resets | Segments of 4 steps | One 96-step segment |

V1 adds $\alpha O\tanh(2Gz)$ to the latent dynamics, with $G$ of unit norm and $O = -G^\top$. V2 adds $\beta O\tanh(2Hz)$ to the behavior mean, with random $H$ and output rotation. Amplitudes set the nonlinear variance fractions in Table 12. Calibration uses 100,000 independent chains and 60 burn-in steps. Gain caps are 1.5 and 1.2, respectively, and sampled local closed-loop spectral radii are below one.

V3 changes both process noise and action residual noise to Student distributions with matching covariances. V4 gates the residual vector inside a state half-space and rescales its nonzero values to preserve unconditional covariance. This changes information availability within a fixed deficient subspace. V5 collects running segments initialized at shared simulator anchors. Its strong condition is one 96-transition segment, with budget 24 using its prefix. The fitter uses residual regression from (17); the other conditions use paired responses from the same reset state.

### I.2 Acquisition, control and statistical comparison

Directed uses RatioGapMLP. A two-fold MLP with two tanh hidden layers of width 64 estimates the conditional behavior mean, using 1,500 updates at learning rate 0.003. The screen selects the largest adjacent eigenvalue ratio among gaps with magnitude at least $0.25$ times the largest eigenvalue and ratio at least two. The two splits must select the same dimension and have projector distance at most $0.25$; otherwise it uses full-space acquisition. Comparators use the MLP screen selecting by gap magnitude, the linear conditional screen, FullBalanced, and GlobalGap.

The primary controller is the six-step quadratic MPC of the controlled study, denoted Base MPC. Three controller comparisons use the behavior policy. One uses fixed behavior feedback and an ensemble objective. The other two recompute a constant residual command from the current encoded state, using a point model or an ensemble. The ensemble uses 128 residual wild bootstrap draws with leverage correction. In V5, signs are shared within blocks of four transitions ending at reset boundaries. NoNewData executes the fitted behavior policy with zero added transitions.

Policies share passive data, physical systems, evaluation starts, targets, noise streams and the appropriate fitter. Perturbation energies are 15.36 and 61.44 at the two budgets. Running acquisition allows policies to visit different states, so total action energy can differ. The primary comparison is Directed against FullBalanced with Base MPC at 24 transitions.

For each case, the paired effect is $1 - C_{\mathrm{method}}/C_{\mathrm{FullBalanced}}$, using the same controller for both costs. We average training seeds and observation maps within each physical system, then give the 16 systems equal weight. Nominal 95% intervals use 20,000 bootstrap resamples of whole systems, separately for each comparison. Controller comparisons also report absolute costs.

### I.3 Effects across departures and controllers

At 24 transitions, mean gains and uncertainty vary across departures. At 96 transitions, Directed lowers mean cost in all 22 conditions. Table 13 reports every condition at both budgets. Table 14 compares acquisition rules, and Table 15 summarizes coverage diagnostics. Strong nonlinear behavior increases error of the linear screen; Student-$t_3$ noise also increases fallback rates for the nonlinear screens.

Table 13: Assumption departures with Base MPC. Entries are mean paired cost reductions (%) with nominal 95% intervals over 16 physical systems. Each acquisition is compared with FullBalanced using the same controller. Training seeds and observation maps are averaged within each system. Budget 24 is primary.

| Condition | Coverage | 24 transitions | 96 transitions |
|---|---|---|---|
| Gaussian | Missing | 40.64 [25.94, 55.11] | 16.35 [9.88, 23.89] |
| Gaussian | Weak | 16.61 [−21.77, 47.89] | 15.71 [9.84, 22.58] |
| V1 mild | Missing | 39.09 [26.70, 52.22] | 14.32 [8.63, 21.39] |
| V1 mild | Weak | 37.85 [27.80, 49.19] | 14.29 [8.60, 21.45] |
| V1 strong | Missing | 27.57 [12.87, 41.43] | 15.51 [8.42, 24.09] |
| V1 strong | Weak | 34.31 [22.00, 45.90] | 14.19 [8.55, 21.81] |
| V2 mild | Missing | 12.78 [−28.65, 45.53] | 16.28 [10.02, 23.65] |
| V2 mild | Weak | 31.88 [18.56, 46.22] | 16.50 [10.26, 23.85] |
| V2 strong | Missing | 20.11 [−23.74, 50.01] | 17.04 [10.77, 24.36] |
| V2 strong | Weak | 39.11 [22.51, 55.13] | 16.98 [10.69, 24.31] |
| V3 mild | Missing | −13.43 [−92.86, 44.50] | 12.89 [7.81, 19.57] |
| V3 mild | Weak | 7.09 [−34.88, 41.15] | 12.84 [7.67, 19.59] |
| V3 strong | Missing | −4.72 [−84.67, 41.02] | 7.99 [5.28, 11.39] |
| V3 strong | Weak | −5.61 [−85.56, 40.37] | 7.87 [5.14, 11.24] |
| V4 mild | Missing | 22.81 [−20.67, 50.84] | 16.25 [10.38, 23.15] |
| V4 mild | Weak | 28.35 [−8.81, 53.09] | 16.31 [10.11, 23.51] |
| V4 strong | Missing | 33.82 [19.89, 48.13] | 16.64 [10.69, 23.62] |
| V4 strong | Weak | 22.14 [−19.69, 50.14] | 16.95 [10.68, 24.28] |
| V5 mild | Missing | 17.06 [−50.77, 58.19] | 12.45 [8.54, 17.96] |
| V5 mild | Weak | 49.89 [38.17, 61.43] | 12.35 [8.28, 18.22] |
| V5 strong | Missing | 18.22 [−48.37, 58.83] | 12.58 [8.47, 18.41] |
| V5 strong | Weak | 49.96 [37.51, 61.95] | 12.40 [8.42, 18.23] |

Table 14: Mean absolute cost over the 1,408 cases for every acquisition rule and controller. Each of the 22 conditions has equal weight. NoNewData uses the behavior policy with zero added transitions and has mean cost 5.155. Ratio MLP denotes RatioGapMLP; Gap MLP denotes ConditionalGapMLP; Linear denotes ConditionalGap. Paired contrasts for each condition are retained in the accompanying numerical supplement.

| Controller | Budget | Ratio MLP | Gap MLP | Linear | FullBalanced | GlobalGap |
|---|---|---|---|---|---|---|
| Base MPC | 24 | 5.645 | 5.578 | 5.895 | 14.240 | 14.297 |
| Base MPC | 96 | 2.888 | 2.889 | 3.010 | 4.773 | 4.992 |
| Fixed behavior, ensemble | 24 | 4.357 | 4.356 | 4.380 | 4.789 | 4.780 |
| Fixed behavior, ensemble | 96 | 4.131 | 4.130 | 4.152 | 4.262 | 4.271 |
| Current state, point | 24 | 4.410 | 4.395 | 4.473 | 6.755 | 6.672 |
| Current state, point | 96 | 3.611 | 3.612 | 3.641 | 4.363 | 4.397 |
| Current state, ensemble | 24 | 3.861 | 3.861 | 3.896 | 4.680 | 4.653 |
| Current state, ensemble | 96 | 3.560 | 3.560 | 3.590 | 3.984 | 3.981 |

Table 15: Coverage diagnostics across departures. Each entry gives the maximum fallback percentage and maximum mean projector error across the missing and weak coverage groups, each containing 64 cases. Projector means include fallback cases; the imposed deficient subspace has dimension two. Ratio MLP, Gap MLP and Linear denote the three conditional screens.

| Condition | Ratio MLP | Gap MLP | Linear | GlobalGap |
|---|---|---|---|---|
| Gaussian | 0.0 / 0.006 | 0.0 / 0.006 | 0.0 / 0.006 | 100.0 / 1.000 |
| V1 mild | 0.0 / 0.006 | 0.0 / 0.006 | 0.0 / 0.007 | 100.0 / 1.000 |
| V1 strong | 0.0 / 0.005 | 0.0 / 0.005 | 0.0 / 0.006 | 100.0 / 1.000 |
| V2 mild | 0.0 / 0.006 | 0.0 / 0.006 | 0.0 / 0.021 | 81.2 / 0.998 |
| V2 strong | 0.0 / 0.007 | 0.0 / 0.007 | 7.8 / 0.249 | 59.4 / 0.993 |
| V3 mild | 0.0 / 0.010 | 0.0 / 0.010 | 6.2 / 0.067 | 100.0 / 1.000 |
| V3 strong | 12.5 / 0.154 | 9.4 / 0.215 | 26.6 / 0.379 | 100.0 / 1.000 |
| V4 mild | 0.0 / 0.007 | 0.0 / 0.007 | 0.0 / 0.006 | 100.0 / 1.000 |
| V4 strong | 0.0 / 0.013 | 0.0 / 0.013 | 0.0 / 0.012 | 100.0 / 1.000 |
| V5 mild | 0.0 / 0.005 | 0.0 / 0.005 | 0.0 / 0.005 | 100.0 / 1.000 |
| V5 strong | 0.0 / 0.005 | 0.0 / 0.005 | 0.0 / 0.005 | 100.0 / 1.000 |

Controller choice changes the cost of response estimation error. Ensemble control reduces extreme costs, while Base MPC has lower cost in most individual cases. Table 16 compares the four controllers on the same acquisition data at both budgets.

Table 16: Controller comparisons within RatioGapMLP on the 1,408 cases. The column for higher cost counts cases exceeding Base MPC at the same budget. Radius counts use the true local closed-loop Jacobian at the origin before action clipping.

| Controller | Budget | Mean cost | Max cost | Higher cost | Radius $\geq 1$ |
|---|---|---|---|---|---|
| Base MPC | 24 | 5.645 | 54.640 | 0 | 76 |
| Fixed behavior, ensemble | 24 | 4.357 | 10.215 | 1277 | 0 |
| Current state, point | 24 | 4.410 | 18.764 | 1282 | 0 |
| Current state, ensemble | 24 | 3.861 | 9.454 | 1257 | 0 |
| Base MPC | 96 | 2.888 | 8.531 | 0 | 0 |
| Fixed behavior, ensemble | 96 | 4.131 | 10.148 | 1408 | 0 |
| Current state, point | 96 | 3.611 | 9.410 | 1408 | 0 |
| Current state, ensemble | 96 | 3.560 | 9.419 | 1405 | 0 |

**Control gains over the behavior policy.** We additionally compare Directed with NoNewData on the same evaluation starts, targets and noise streams. For case costs $C$ and $C_0$, the paired gain is $1 - C/C_0$. We average within each physical system over training seeds, observation maps and the 22 condition and coverage groups, then weight the 16 systems equally. Nominal 95% intervals use 20,000 paired bootstrap resamples of whole systems. Gains compare evaluation control costs at the stated acquisition budgets. Table 17 reports the mean gain, the 95th percentile of case average costs, and the number of cases with $C < C_0$. The percentile describes variation across cases, each averaged over evaluation episodes.

At 24 transitions, Base MPC improves 1,285 of 1,408 cases, but high costs in a small fraction of cases give a mean gain of $-8.98\%$. Residual ensemble control lowers the cost percentile from 34.386 to 4.494 and gives a mean gain of $25.05\%$. At 96 transitions, Base MPC improves every case in this grid and has lower mean and percentile costs than residual ensemble control. The distribution across cases explains why lower cost in most cases can coexist with higher mean cost.

The independent systems in Appendix I.4 reproduce this change in mean cost ordering. Residual ensemble control has lower mean cost at 24 transitions, while Base MPC has lower mean cost from 48 transitions onward. Response interpolation in Table 18 connects these comparisons to the controller's sensitivity to acquisition response error.

Table 17: Directed compared with the behavior policy on 1,408 cases. Gain is the mean paired relative cost reduction with nominal 95% system bootstrap intervals. Q95 is the 95th percentile of case average costs. The behavior policy has mean cost 5.155 and Q95 5.742. Residual ensemble denotes current state ensemble control.

| Controller | Budget | Mean cost | Gain (%) [95% CI] | Q95 | Improved cases |
|---|---|---|---|---|---|
| Base MPC | 24 | 5.645 | $-8.98\,[-31.34,\ 12.33]$ | 34.386 | 1285/1408 |
| Residual ensemble | 24 | 3.861 | $25.05\,[24.07,\ 25.98]$ | 4.494 | 1402/1408 |
| Base MPC | 96 | 2.888 | $43.80\,[42.80,\ 44.75]$ | 3.175 | 1408/1408 |
| Residual ensemble | 96 | 3.560 | $30.87\,[29.90,\ 31.82]$ | 3.848 | 1408/1408 |

### I.4 Response estimation and budget sensitivity

Response interpolation measures the controller's sensitivity to acquisition error. For 192 cases with missing coverage, we interpolate encoded responses toward noiseless encoded pairs at the same design and refit both estimation stages. Evaluation noise and initial states are fixed. Halving response error reduces the count at budget 24 of local closed-loop radii at least one from 11 to zero (Table 18). The noiseless reference isolates acquisition error in this diagnostic experiment.

Table 18: Response interpolation on 192 cases with missing coverage, Base MPC and 24 added transitions. The weight multiplies the encoded response difference from the noiseless acquisition reference. Both estimation stages are refitted at every weight; control evaluation noise and initial states are held fixed.

| Error weight | Mean cost | Maximum cost | Local radius $\geq 1$ |
|---|---|---|---|
| 0.00 | 2.844 | 8.416 | 0 |
| 0.25 | 2.868 | 8.440 | 0 |
| 0.50 | 2.943 | 8.481 | 0 |
| 0.75 | 3.358 | 51.766 | 1 |
| 1.00 | 5.960 | 54.640 | 11 |

Budget sensitivity is evaluated on 16 additional physical systems, two training seeds, two observation maps, and Gaussian, Student-$t_5$ and Student-$t_3$ noise under both coverage conditions. Each of the 384 encoder cases receives four repetitions of acquisition noise. Directed and FullBalanced share each repetition and use Base MPC, residual point control and residual ensemble control at 24, 32, 48, 64 and 96 transitions. The primary comparison in this study uses 48 transitions. Increasing the budget raises both transition count and perturbation energy.

Bootstrap intervals resample physical systems after averaging maps, training seeds and acquisition repetitions within systems. With Base MPC, Directed reduces mean cost in all six conditions at 24 transitions. At 48 transitions, reductions range from $17.15\%$ to $31.73\%$. Directed also lowers mean cost in all six conditions at budgets 64 and 96. Figure 7 reports all five budgets for each noise law, coverage condition and controller.

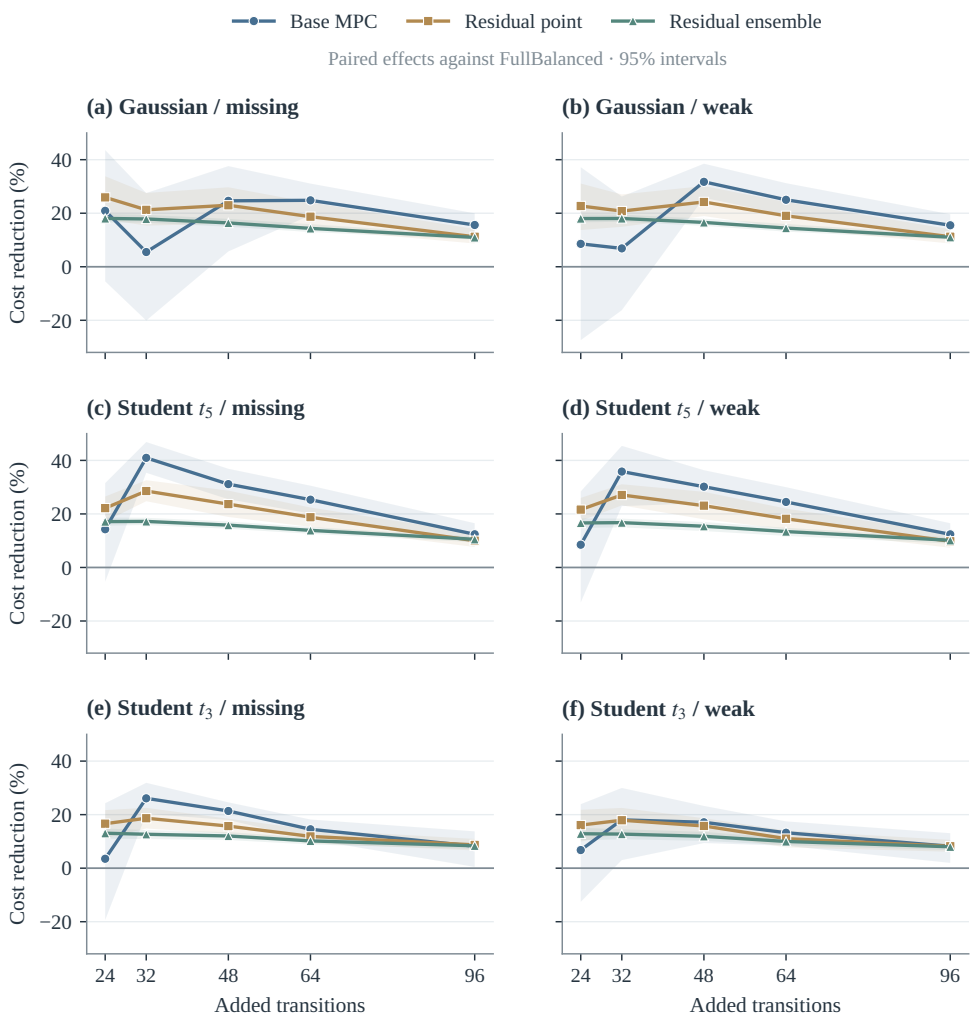



Figure 7: Acquisition budget and control cost on 16 physical systems. Lines show mean paired cost reductions against FullBalanced with the same controller; bands are nominal 95% system bootstrap intervals. Each panel contains 256 acquisition repetitions grouped within the 16 systems. Perturbation energy is $0.64n$ at $n$ added transitions.

# J Pixel world model with a shared learned representation

## J.1 Behavior data and the coverage diagnostic

We use `cheetah-run` from the DeepMind Control Suite (Tassa et al., 2018), with action repeat two, $64 \times 64$ RGB observations, stacks of three frames and 500 agent steps per episode. A state-based SAC behavior policy (Haarnoja et al., 2018) is trained once for 200,000 environment steps. For each of three model seeds, a random subspace of dimension three $S \subset \mathbb{R}^6$ receives independent Gaussian behavior noise of standard deviation 0.3, while the commanded noise is zero on $S^{\perp}$. Executed actions are clipped to $[-1, 1]^6$. The learned world model and Directed screen use pixels, actions and rewards. The behavior policy uses physics state.

Each seed supplies 100 passive and ten held-out episodes. Action clipping introduces variation on $S^{\perp}$; we measure its executed coverage with passive clipping rates of $1.77\%$ to $2.00\%$. The conditional screen regresses executed actions on passive encoder latents using five episode folds and an ELU MLP with two hidden layers of width 256, trained for 60 epochs with AdamW, learning rate $10^{-3}$ and

weight decay $10^{-4}$. It selects the low-variance side of the largest spectral gap when the upper/lower eigenvalue ratio is at least two, the gap is at least 0.25 times the largest eigenvalue, and the distance between split projectors is at most 0.25. Otherwise acquisition falls back to the full action space. GlobalGap applies the same rule to raw executed actions. Projector error is measured against the imposed $S^{\perp}$; a dimension mismatch has error one.

### J.2 World model and acquisition

The passive model has a convolutional encoder with 64 latent coordinates, five dynamics heads with two hidden layers of width 512, reward and behavior policy heads of width 256, and a pixel decoder. The training loss combines latent consistency with a target encoder updated by exponential moving averages, reward prediction, pixel reconstruction and behavior cloning, each with weight one. Rollout horizon is five, temporal weight is 0.5, batch size is 256, and training uses 30,000 updates of AdamW at learning rate $3 \times 10^{-4}$ and weight decay $10^{-6}$, with 1,000 warmup updates, gradient clipping at 20, random image shifts with padding four, and target update coefficient 0.01. The model has deterministic latent dynamics.

The policies are Directed, FullBalanced, GlobalGap, EnsembleDisagreement, the true $S^{\perp}$ as a diagnostic Oracle, and NoNewData. Each active policy collects 30 additional behavior episodes; budgets ten and 30 are episode prefixes. Starts and exogenous noise streams are shared across policies, with no resets within acquisition episodes. Directed and the two subspace comparators draw random fixed-norm directions in their respective spans; FullBalanced draws them in the full action space. EnsembleDisagreement uses the leading direction of disagreement among the five passive dynamics heads' action Jacobians, with random sign. The commanded perturbation norm is 0.5 at each step, so commanded perturbation energy is exactly 1,250 or 3,750 for every active policy. Clipping changes the executed perturbations. Both commanded and executed perturbation energies are reported in Table 19.

Within each seed, all policies share the frozen passive encoder and its target encoding. Identically initialized dynamics, reward, decoder and policy heads are retrained using the same schedule on passive data plus each policy's acquisition prefix. NoNewData uses passive data alone. The shared encoder makes the latent response target common to all acquisition policies.

### J.3 Response and Control Evaluation Methods

At 40 states from each held-out episode, we restore physics state and branch under the recorded action sequence plus and minus $0.3v$ for three successive agent steps, clipping each action to its bounds. Directions comprise three basis vectors each of $S^{\perp}$ and $S$, and three random directions. The simulator branches are encoded by the shared frozen encoder. At horizon $k$, normalized squared response error is

$$E_k(D) = \frac{\sum_{i,v \in D} \|\widehat{\Delta}_{iv}^{(k)} - \Delta_{iv}^{(k)}\|_2^2}{\sum_{i,v \in D} \|\Delta_{iv}^{(k)}\|_2^2}.$$

The two primary diagnostics are projector error of the conditional screen and Directed versus FullBalanced $E_3(S^{\perp})$ at budget ten. The acquisition effect is $1 - E_3^{\mathrm{Directed}}/E_3^{\mathrm{FullBalanced}}$ within a seed, followed by a mean with equal weight per seed. The model seed is the independent unit for comparing acquisition methods. Table 19 reports both budgets and all direction groups.

Control uses MPPI constrained around a common behavior cloning prior. Within each model seed, every acquisition policy uses the policy head of NoNewData and the same frozen encoder. The planner optimizes residual actions with norm at most $0.25$, retains the candidate with zero residual, and evaluates the final proposed sequence. It uses horizon six, 128 samples, three iterations, 32 elites, temperature $0.5$ and discount $0.99$. Settings are selected on two separate tuning seeds. Every fitted model uses ten evaluation starts shared across acquisition methods. Control return is reported alongside the primary comparison of response errors.

### J.4 Response Prediction and Control Results

The conditional screen selects three directions in all seeds, with projector errors from $0.026$ to $0.075$. GlobalGap selects five directions and has projector error one. A reference screen using physics

states selects three directions, with errors from $0.015$ to $0.038$. Directed reduces $E_3(S^{\perp})$ against FullBalanced by a mean of $12.04\%$ at ten episodes and $10.89\%$ at 30. EnsembleDisagreement has lower mean response error at the smaller budget, while Directed has lower error at the larger budget (Figure 4).

Directed improves return over FullBalanced in each seed at ten episodes; the comparison varies across seeds at 30. GlobalGap has the highest mean return at the smaller budget. The differing rankings by response accuracy and return connect control performance to how the planner uses the model. Table 19 summarizes every policy at both budgets, including the executed energy and clipping rates.

Table 19: Mean action response errors, control return and acquisition cost over the three model seeds, with equal seed weights. Figure 4 shows individual seeds for $E_3(S^{\perp})$ and return. Errors are normalized squared errors; $\perp$ denotes the imposed $S^{\perp}$. Each seed's return averages ten common starts. Ensemble denotes EnsembleDisagreement and BC the common behavior policy. Energy measures executed perturbations relative to the clipped behavior action; commanded energies at 10 and 30 episodes are 1,250 and 3,750. Clip is the fraction of clipped action coordinates.

| Policy | Episodes | $E_1(\perp)$ | $E_3(\perp)$ | $E_3(S)$ | $E_3(\mathrm{rand})$ | Return | Energy | Clip (%) |
|---|---|---|---|---|---|---|---|---|
| NoNewData | 0 | 1.2846 | 1.0004 | 0.4644 | 0.6679 | 311.44 | 0.00 | N/A |
| Directed | 10 | 0.8595 | 0.6662 | 0.4650 | 0.5360 | 331.42 | 1195.56 | 5.07 |
| Ensemble | 10 | 0.8371 | 0.6518 | 0.4654 | 0.5309 | 314.82 | 1168.62 | 6.01 |
| FullBalanced | 10 | 0.9752 | 0.7571 | 0.4652 | 0.5753 | 317.86 | 1181.60 | 4.91 |
| GlobalGap | 10 | 0.9896 | 0.7563 | 0.4658 | 0.5779 | 338.88 | 1182.52 | 4.84 |
| Oracle | 10 | 0.8719 | 0.6745 | 0.4643 | 0.5396 | 332.17 | 1197.51 | 5.00 |
| Directed | 30 | 0.7323 | 0.5503 | 0.4597 | 0.4894 | 336.54 | 3589.89 | 5.08 |
| Ensemble | 30 | 0.7582 | 0.5775 | 0.4627 | 0.5062 | 324.14 | 3503.64 | 6.09 |
| FullBalanced | 30 | 0.8169 | 0.6168 | 0.4567 | 0.5182 | 331.91 | 3540.30 | 5.00 |
| GlobalGap | 30 | 0.8479 | 0.6360 | 0.4594 | 0.5311 | 310.37 | 3547.06 | 4.87 |
| Oracle | 30 | 0.7349 | 0.5461 | 0.4595 | 0.4893 | 311.12 | 3588.63 | 5.10 |
| Common BC | 0 | N/A | N/A | N/A | N/A | 312.42 | 0.00 | N/A |