THE STATE BETWEEN STAGES: IDENTIFYING SUFFICIENT STATISTICS FROM SECOND DERIVATIVES

Anonymous authors
Paper under double-blind review

ABSTRACT

A deterministic differentiable program often lets one block of its inputs interact with the rest only through a few numbers, its *interface*, whose dimension is the state a staged solver must carry. Representation learning estimates such a summary from data; we compute it from second derivatives, even where no variable holds it. The blocks' cross-Hessian factors through the interface, so its rank at any set of second-block values bounds from below the dimension of every linear interface and every smooth one with a full-rank Jacobian. Under analyticity, constant rank, a full-rank summary map, and values drawn from a density, the rank almost surely equals the minimal dimension of an analytic interface, which can lie below that of any linear one. On a control rollout it reads 2 where the loop carries 4 numbers, recovering the zero-effort miss (where the mass would land if it coasted), which no line computes. Solving in stages through it reaches the joint optimum. On a published robot-arm benchmark it reads 2 against 3 carried numbers. Given only the objective, it recovers the interface of 200 of 200 random objectives and of 6 Feynman formulas, and two ratio interfaces as payoff differences.

1 INTRODUCTION

Many differentiable programs let one block of their inputs interact with the rest through far fewer numbers than either block has. A control rollout that scores only where a point mass ends up lets its first thrust phase interact with the second only through where the mass would land if it coasted from the handoff (its zero-effort miss), two numbers where the loop carries four. The distance of a robot arm's end from its shoulder depends on the first two segments only through where the shoulder lies in the frame of the second, two numbers where the loop carries three. In the gravitational potential difference $Gm_1m_2(1/r_2-1/r_1)$, the two masses reach the separation r_2 only through their product. We call the smallest such summary the program's *interface*. When the blocks are decisions taken in sequence, the interface is the state a staged solver must carry, and the problem solves exactly through it, one stage at a time. Which state to carry is a question in dynamic programming (Bellman, 1957), in control and trajectory optimization, in planning and reinforcement learning, and in influence diagrams (Howard & Matheson, 1984) and the Bayesian-network methods that evaluate them (Shachter, 1986). The interface's dimension is also the number of inputs a value function or planner for the second block needs and the width an encoder between the blocks needs. Programs do not mark these summaries; the rollout never even computes its own. We know of no automatic tool for matrix block structure or state abstraction that looks for them in a program's source.

Representation learning has an answer to a version of this question. Bisimulation metrics (Ferns et al., 2011) compare states through a known transition kernel, whereas their deep-reinforcement-learning descendants learn a compressed state from sampled interaction with an environment and certify it approximately, in expectation, under a learned dynamics model (Zhang et al., 2021). Causal state abstraction keeps the state variables that affect dynamics or reward and discards the rest (Wang et al., 2024). Our setting is the complementary one, where the objective is a program we can read and differentiate. Because the objective's cross-Hessian, its mixed second derivative between one block of inputs (the context) and the rest (the probe), factors through whatever interface the objective has, its rank and its singular directions are statements about that interface, computed rather than estimated. Read as representation learning, the cross-Hessian rank is the exact counterpart of train-

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.

ing a k-dimensional state encoder and choosing k by validation, with k identified from the objective itself.

Where the interface is a variable the program already computes, a syntactic search finds it. A *dominator* of the traced computation graph, in the sense of compiler flow analysis (Lengauer & Tarjan, 1979), separates the blocks, and an exact two-stage rewrite follows, in the form that Theorem A.1 of Appendix A states and proves. The exact abstraction methods for influence diagrams and factor graphs that Section 5 reviews share the requirement that the summary be realized by a node. Neither the rollout's zero-effort miss nor the ratio $m_2/(m_1 + m_2)$, through which the center of mass $(m_1 r_1 + m_2 r_2)/(m_1 + m_2)$ depends on the second position, is realized by a node. This paper removes that node-realization hypothesis, reading the interface from the objective's derivatives rather than from its syntax. Section 6 names the hypotheses it holds fixed.

Our instrument, the cross-Hessian rank, bounds from below the dimension of every linear interface and every smooth one with a full-rank Jacobian, identifies that dimension exactly under analyticity and constant rank, recovers the interface's direction where no node carries it, and composes along a chain of blocks. Section 3 reads it on a control rollout (which we then solve in stages), a published robot-arm benchmark, and a synthetic objective built so that the reading at any single probe value undercounts. Section 4 then turns it into a discovery procedure handed nothing but the objective, a box of context values, and random probe values, and reports what it returns on Feynman formulas as a symbolic-regression benchmark prints them and on random objectives whose interfaces stayed sealed until scoring, against a node-gradient baseline and a learned one.

2 Identification from the Cross-Hessian

Some objectives have a low-dimensional interface with no corresponding node anywhere in the trace, because the modeler wrote the expression in a form that never materializes it. By contrast, the syntactic search of Section 1 can return only a variable some equation computes. We give an identification result that detects this case and recovers the interface's direction where that search cannot.

Write the objective as

$$f(x, y) \;=\; g(x) \;+\; h\big(\phi(x),\, y\big), \tag{1}$$

where the vector x, ranging over an open set $U \subseteq \mathbb{R}^n$, collects the block of inputs whose interface we want, the *context*, and the vector y, ranging over an open set $V \subseteq \mathbb{R}^p$, collects the remaining inputs, which we treat as a *probe* the analyst may move freely. The map $\phi : U \to \mathbb{R}^k$ is the candidate interface, a k-dimensional summary that need not correspond to any traced variable; h is the part of the objective that couples the summary to the probe; and g is an arbitrary offset depending on x alone. We impose no regularity on g, because every quantity below starts with a derivative in y, which removes it. We call an interface that allows such an offset *offset-refined*. The condition is weak separability of the coupling (Gorman, 1959; Goldman & Uzawa, 1964), relaxed to allow an additive term in x alone that the summary need not carry.

Fix a base point x_0 in U and write $\phi_0 := \phi(x_0)$ for the interface value there. Let N be an open neighborhood of ϕ_0 in $\mathbb{R}^k$ on which h is defined, and assume $h \in C^2(N \times V)$. The *cross-derivative matrix*

$$W_\phi(y) \;:=\; \left[\frac{\partial^2 h}{\partial \phi_a \, \partial y_b}(\phi, y)\right]_{a \le k,\; b \le p} \;\in\; \mathbb{R}^{k \times p} \tag{2}$$

collects the second derivatives that differentiate once in a coordinate of the interface and once in a coordinate of the probe. We abbreviate $W(y) := W_{\phi_0}(y)$ and write $c_b(\phi, y) \in \mathbb{R}^k$ for the b-th column of $W_\phi(y)$, that is, the vector of derivatives of $\partial h/\partial y_b$ in each interface coordinate. Two dimensions are then in play. The first is the dimension $r^\star$ of the *first-order span at the base point*,

$$L(\phi) \;:=\; \operatorname{span}\{\, c_b(\phi, y) \;:\; y \in V,\; 1 \le b \le p \,\} \;\subseteq\; \mathbb{R}^k, \qquad r^\star \;:=\; \dim L(\phi_0), \tag{3}$$

the span of every cross-derivative column that the probe variable can produce while the interface is held at ϕ_0. The second is the *minimal linear offset-refined dimension* $\underline{d}$, defined as the least rank of a linear map $\pi : \mathbb{R}^k \to \mathbb{R}^{k'}$ for which there exist functions A on a neighborhood of ϕ_0 and H on the image of π times V with $h(\phi, y) = A(\phi) + H(\pi(\phi), y)$ there. Here and in every decomposition below, A and H are *arbitrary functions*, with no continuity or differentiability imposed, since the statements below must hold against competing interfaces that may be nowhere differentiable.

An experimenter measures neither of those. Write J for the Jacobian $D\phi(x_0)$ of the summary map at the base point, of size $k \times n$. For probes $y_1, \ldots, y_m \in V$ let $S(y_1, \ldots, y_m)$ be the $(mp) \times n$ matrix obtained by stacking the mixed second derivatives of f itself vertically, one $p \times n$ block per probe,

$$S(y_1, \ldots, y_m) \;=\; \begin{bmatrix} D_x D_y f(x_0, y_1) \\ \vdots \\ D_x D_y f(x_0, y_m) \end{bmatrix} \;=\; M(y_1, \ldots, y_m)\, J, \qquad M \;:=\; \begin{bmatrix} W(y_1)^\top \\ \vdots \\ W(y_m)^\top \end{bmatrix}, \tag{4}$$

so that the row of M indexed by the pair (i, b) is $c_b(\phi_0, y_i)^\top$. Every entry of S is a second derivative of the objective, obtainable directly, but W, L and ϕ itself are not observable. We write $\hat{r} := \operatorname{rank} S(y_1, \ldots, y_m)$ for the observed stacked rank. The factorization $S = MJ$ is a chain rule whose inner dimension is k, so $\hat{r} \le k$ whatever the probes. We fix this orientation, probes indexing rows and coordinates of x indexing columns, throughout; the interface direction is then a *right*-singular vector of S. Our implementation differentiates the reverse-mode context gradient forward along each probe coordinate, one Hessian-vector product per coordinate at a small multiple of a gradient's cost, so a stack of m probes costs about mp gradient evaluations however large n is; we stack three probes.

The half of the result that needs only C^2 is the lower bound, one-sided in the direction a discovery method wants.

Proposition 2.1 (One-sided rank bound)**.** *Assume $h \in C^2(N \times V)$, and let $y_1, \ldots, y_m$ be any probes whatsoever in V, with observed stacked rank $\hat{r} = \operatorname{rank} S(y_1, \ldots, y_m)$. Then:*

1. ***(Linear interfaces.)*** *If $\pi : \mathbb{R}^k \to \mathbb{R}^{k'}$ is linear and there are a neighborhood $N_1 \subseteq N$ of ϕ_0 and functions A, H, of no assumed regularity, with $h(\phi, y) = A(\phi) + H(\pi(\phi), y)$ for all $(\phi, y) \in N_1 \times V$, then $\hat{r} \le r^\star \le \operatorname{rank} \pi$. In particular $\hat{r} \le r^\star \le \underline{d}$.*
2. ***(Nonlinear interfaces.)*** *If $\rho : N_1 \to \mathbb{R}^s$ is real-analytic and is a submersion at ϕ_0, meaning that its Jacobian $D\rho(\phi_0)$ has full rank s, and if there are functions A, H, again of no assumed regularity, with $h(\phi, y) = A(\phi) + H(\rho(\phi), y)$ for all $(\phi, y) \in N_1 \times V$, then $\hat{r} \le r^\star \le s$.*

Consequently an observed stacked rank of r shows that no offset-refined interface of dimension below r exists at ϕ_0, whether linear or an analytic submersion. The bound is deterministic and holds for the tuple actually used, whatever the rank of J.

Two features of Proposition 2.1 govern how it is used. A *high* observed rank establishes irreducibility, whereas a low one is never by itself evidence of compressibility, since an unlucky probe tuple can depress the rank arbitrarily. The bound also holds without full rank of J (a deficient J can pull the observed rank below $r^\star$, weakening the bound without invalidating it), without analyticity, and without connectedness of N or V, though the decomposition must hold on all of $N_1 \times V$, since $L(\phi_0)$ is a span over every $y \in V$. Part 1 is the mixed-block form of the classical bound that a function of k ridge directions has Hessian rank at most k (Pinkus, 2015), read across the block boundary rather than within one block.

Reading the observed rank as a measurement rather than only as a bound requires the reverse inequality. The next theorem obtains it under extra hypotheses.

Theorem 2.2 (Generic attainment of the first-order span)**.** *Assume h is real-analytic on $N \times V$, that V is open and connected, that $J = D\phi(x_0)$ has full rank k, and that the number of probes satisfies $m \ge r^\star$ (for which $m \ge k$ is always sufficient). Then*

$$Z_m \;:=\; \{\,(y_1,\dots,y_m)\in V^m \;:\; \operatorname{rank} S(y_1,\dots,y_m) < r^\star\,\} \tag{5}$$

is closed in V^m, has empty interior, and has Lebesgue measure zero; indeed Z_m is contained in the zero set of a single real-analytic function on V^m that is not identically zero. Consequently, if the probe tuple is drawn from any probability law on V^m that is absolutely continuous with respect to Lebesgue measure, then $\hat{r} = r^\star$ with probability one.

As the next example shows, $r^\star$ can lie strictly below the linear dimension $\underline{d}$.

Example 2.3 (A linear interface dimension no single base point reveals)**.** *Take $n = k = 2$, $p = 1$, $\phi = \mathrm{id}$, and $h(\phi, y) = y(\phi_1 + \frac{1}{2}\phi_2^2)$. Then $r^\star = 1$ at every base point while the minimal linear offset-refined dimension is $\underline{d} = 2$ (Appendix A), so the single-base-point reading of 1 is a sound lower bound on $\underline{d}$ and no estimate of it.*

Theorem 2.4 (Exactness against nonlinear interfaces)**.** *Assume h is real-analytic on $N \times V$ and that V is open and connected. Suppose the first-order span has locally constant dimension at the base point, that is, there is a neighborhood $N_1 \subseteq N$ of ϕ_0 on which $\dim L(\phi) = r^\star$ for every $\phi \in N_1$. Then there exist a neighborhood $N_2 \subseteq N_1$ of ϕ_0, a real-analytic submersion $\rho : N_2 \to \mathbb{R}^{r^\star}$, and real-analytic functions A on N_2 and H on $\rho(N_2) \times V$ such that*

$$h(\phi, y) \;=\; A(\phi) \;+\; H\big(\rho(\phi),\, y\big) \qquad \textit{for all } (\phi, y) \in N_2 \times V. \tag{6}$$

Combined with part 2 of Proposition 2.1, the number $r^\star$ is therefore exactly the minimal dimension of an offset-refined interface that is an analytic submersion at ϕ_0, and when J has full rank, by Theorem 2.2, the observed stacked rank equals it almost surely.

Theorem 2.4 turns the observed rank from a bound into a measurement, naming the quantity measured: the number of scalar functions of the interface, linear or not, through which the coupling can be routed. In Example 2.3, $\rho(\phi) = \phi_1 + \phi_2^2/2$ is a one-dimensional analytic interface whose gradient $(1, \phi_2)$ spans $L(\phi)$ everywhere, so the report of 1 is exact and the excess of $\underline{d} = 2$ is what insisting on a linear interface adds; Appendix A records two routes to $\underline{d}$ and what each costs.

The constant-rank hypothesis of Theorem 2.4 is the one hypothesis whose failure makes the reported number wrong rather than merely conservative. It is also not removable, since Example A.3 in the appendix exhibits a coupling whose first-order span drops rank exactly at the base point, where the cross-Hessian reads 0 for every probe tuple while the minimal analytic interface dimension nearby is 1.

Because the equality $\hat{r} = r^\star$ of Theorem 2.2 holds only almost surely, the readings of Section 3 do not use it. A reading that equals the dimension k of a known interface is exact for the tuple used, since $\hat{r} \le r^\star \le k$. The theorem does cover the runs of Section 4, because they draw their probes at random. Proofs are in Appendix A.

A program whose inputs form a chain of blocks $d_1, \dots, d_T$, as the phases of a rollout do, is handled one split at a time: at split j, take x to be the parameters together with $d_1, \dots, d_j$ and let the whole downstream block $(d_{j+1}, \dots, d_T)$ play the probe. When the objective factors along the chain, meaning that at every split j it is a sum of terms free of the downstream block plus a term reading that block only through a split-j interface, each subproblem is an instance of (1) and Proposition 2.1 applies unchanged. At every split the one-sidedness that makes the bound safe survives, with no hypothesis beyond C^2, and the number of probes needed is set by $r^\star$ at that split rather than by T.

The natural way of running that loop, freezing the upstream blocks as backward induction does, is unsound, since freezing deletes columns from the composite Jacobian in $S = MJ$ and so can only lower the rank. Appendix B exhibits a two-dimensional split-2 interface that reads 1 at every frozen value, undetectable by any sweep, whereas probing the upstream blocks reads 2, and reports the probed loop on three- and four-block chains of inlined residuals absent from the trace, with its constructed negatives.

3 READING THE CROSS-HESSIAN ON DIFFERENTIABLE PROGRAMS

We knew the interfaces of the three objectives below before reading them, so they test the instrument, and Section 4 turns to discovery. They are a control rollout that never computes its interface, a published benchmark function whose interface no line computes, and the case the stacking exists for, an interface that each probe alone undercounts.

Our first objective is a differentiable control rollout written the usual way, one loop over the horizon that carries a planar point mass's position $\mathbf{q}$ and velocity $\dot{\mathbf{q}}$ forward under damped, saturated thrust and reads both phases' thrusts from one array. The context x is the initial state and the first phase's thrusts, 10 numbers, and the probe y is the second phase's 10. The loss is the effort accumulated step by step plus a pseudo-Huber penalty on the terminal miss. Because the damping is linear, the second phase shifts the terminal position by an amount that does not depend on the first. Hence the first phase affects the loss only through the zero-effort miss $\mathbf{q} + \tau\dot{\mathbf{q}} - \mathbf{q}^*$, where the mass would end relative to the target $\mathbf{q}^*$ if it coasted from the handoff, with $\tau = 0.452$ set by the damping and the length of the second phase. No line of the loop computes the miss. Where the loop carries 4 numbers across the handoff, the cross-Hessian reads rank 2 and its top right-singular vectors span the miss's gradients to within an angle of 1.5×10^{-7} radians. That bound holds for every cross-Hessian alignment we report, whose principal cosines all equal 1 up to float64 rounding; the run itself asserts only an angle below 1.5×10^{-5}. A planner or value function for the second phase therefore needs 2 inputs where the state offers 4.

We solve the rollout in stages through the miss. With the initial state fixed, we tabulate the second phase's optimal cost once on a 25×25 grid of misses, 625 entries where the handoff state at that resolution would need 390,625, and solve the first phase against the table. The staged optimum matches the joint one over all 16 thrust coordinates to within 10^{-9}, whereas the same table indexed by the handoff position alone misses it by 6.6×10^{-4}. Scoring the terminal velocity as well makes the whole handoff state the interface, raising the reading to 4. Averaging the loss over 8 scenarios of second-phase noise, random targets and gusts, leaves the reading at 2, because additive noise shifts the terminal position by a constant in each scenario. Because a random second-phase damping rate changes the coast time by scenario, it widens the reading to the handoff state's 4.

The second objective, the robot-arm function of An & Owen (2001), is a published test function for surrogate models. The function returns the distance from the shoulder to the end of an arm of four segments, segment i with length L_i and angle θ_i relative to the segment before, so that the end is at $\sum_i L_i(\cos A_i, \sin A_i)$ with $A_i = \theta_1 + \cdots + \theta_i$. We take the first two segments as the context, 4 numbers, and the last two as the probe. Across the split the loop carries 3 numbers, the two partial sums and the angle A_2. However, the distance does not change when the whole arm turns about the shoulder, so the last two segments see the first two only through the shoulder's two coordinates in the frame of the second segment, which the function never forms. Stacked over probes, the cross-Hessian reads rank 2 and spans the gradients of those coordinates to within that angle. One probe alone reads 1, because the distance's gradient with respect to the shoulder's position is a unit vector, which a change of probe can turn but not stretch. Extended to twenty segments split after ten, the same loop reads rank 2 from 20 context numbers, again to within that angle. With the end's vertical offset weighted, a score that a rotation changes, the cross-Hessian reads all 3 numbers the loop carries.

We built the third objective so that no single probe counts its interface, as on the arm, and no node gradient at a single probe is an interface direction: a three-dimensional context x, a scalar probe y, the coupling $h(\phi, y) = \phi_1 y + \phi_2 y^2/2$ with the nonlinear two-dimensional interface $\phi_1 = x_1x_2 + x_3$ and $\phi_2 = x_1 - x_2^2x_3$, written flat with the products distributed and the sums interleaved,

$$f(x,y) \;=\; y\,x_1x_2 \;+\; \tfrac{1}{2}y^2x_1 \;+\; y\,x_3 \;-\; \tfrac{1}{2}y^2x_2^2x_3 \;+\; g(x), \qquad g(x) \;=\; \log(1+x_1) - \tfrac{1}{2}|x|^2, \tag{7}$$

so that no node of the trace equals ϕ_1 or ϕ_2 and no partial sum has an x-gradient that is a fixed interface direction. The dominator search of Section 1 reports no cut (the node-gradient search is measured in Section 4). Because y is scalar, each single-probe cross-Hessian is a 1×3 row, of rank at most 1 by shape and of rank exactly 1 at each of the probes $y \in \{0.3, 0.9, 1.6\}$ taken alone, since the row vanishes at none of them. Two probes stacked read rank 2, which the third leaves unchanged. The stacked singular values are $(2.86, 0.94)$ followed by a null one, with the top two

right-singular vectors spanning $\{\nabla\phi_1, \nabla\phi_2\}$ to within the same angle. We checked by hand that the coupling is polynomial and that $\nabla\phi_1$, $\nabla\phi_2$ are independent at the base point $(0.7, 0.4, 1.1)$, hence nearby, so Theorem 2.4 makes 2 the minimal analytic offset-refined dimension there. A collinear twin with $\phi_2 = 2\phi_1$ reads rank 1 however many probes are stacked.

Appendix B reports readings of the rank-drop coupling of Example A.3 and of a rescaled undercount context. At scale, an objective whose context of dimension 1000 reaches a scalar probe through 3 quadratic coordinates, coupled non-quadratically so that the probe count binds, reads rank 3 from 3 probes, that is, 3 Hessian-vector products, aligned with the planted tangent space to within that angle, and rank 2 from two probes.

4 Discovery Without a Planted Answer

Up to here we read the cross-Hessian at one base point against an answer we knew. Handed nothing but the objective, a box of context points, and probes drawn at random, the instrument becomes a discovery procedure in three steps. First, the run reads the cross-Hessian at 40 random points of the box and refuses unless the rank agrees at every one of them, which checks the constant-rank hypothesis of Theorem 2.4 on the sample. Second, it looks first for a closed form, r polynomial coordinates $\hat{\phi}$ whose gradients lie in the identified row space at every sampled point as the null vectors of a generalized eigenproblem (the fraction of a candidate's gradient energy leaving the plane, summed over the points) at the smallest degree where r independent null coordinates appear, then row-reduces over the monomials to the sparsest coordinates with the same joint level sets. Where no degree fits, it takes as $\hat{\phi}$ the r payoff differences $f(x, y_i) - f(x, y_*)$ at swept probes, which Corollary A.1.1 of Appendix A shows are interface coordinates near each point where their gradients are independent, a condition the run checks at every sampled point. Third, it tests the recovered statistic against the objective with no reference to the cross-Hessian: two context points on one level set of $\hat{\phi}$ must give the same payoff differences $f(x, y) - f(x, y_*)$ at every probe, a condition the run scores by the worst relative mismatch over random pairs. Where the context is itself optimized, the run then solves the split in two stages, tabulating the probe's optimum once as $\overline{u}(\eta) = \max_y [f(\sigma(\eta), y) - f(\sigma(\eta), y_*)]$ on a grid in the statistic's value η through a section σ of $\hat{\phi}$ and maximizing $f(x, y_*) + \overline{u}(\hat{\phi}(x))$ in stage 1.

On the undercount objective the procedure returns, at degree 3, the pair $x_1 - x_2^2 x_3$ and $x_3 + x_1 x_2$, the interface with its coordinates swapped, after refusing degree two because only one coordinate fits there, and on the collinear twin the single coordinate $x_3 + x_1 x_2$.

To test objectives we did not write, we take 9 formulas from the Feynman database as the symbolic-regression benchmark SRSD (Matsubara et al., 2022) codes them, following Udrescu & Tegmark (2020), with the physical constants ε_0, G and k_B fixed, one variable fixed in each of two formulas, and every other variable in $[1, 3]$, including the speed of light, which is the Larmor formula's probe. Each splits its variables into a context block and a scalar probe. On the 6 formulas whose first block enters through a product (the Coulomb force, a gravitational potential difference, two-source interference, the ideal gas, drift velocity and Larmor power), the procedure returns the product itself at degree 2: $q_1 q_2$, $m_1 m_2$, $I_1 I_2$, nT, μq and qa. These are easy cases, since the trace computes most of the products as nodes (see the node-gradient search below). Two of the products are offset-refined: the term $-Gm_1 m_2/r_1$ of the potential difference and the term $I_1 + I_2$ of the interference $I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\delta$ are offsets in the context alone, so a reading that did not allow an offset would need two dimensions. On two formulas the interface is a ratio: the driven oscillator $qE_f/(m(\omega_0^2 - \omega^2))$ reaches the frequency only through qE_f/m; the center of mass reaches the second position only through $m_2/(m_1 + m_2)$, again with an offset. The cross-Hessian reads rank 1 on both and recovers the ratio's gradient to within an angle of 1.5×10^{-7} radians. Because no polynomial has a ratio's level sets, the polynomial fit refuses both; the procedure instead returns one payoff difference for each, a statistic the program evaluates but the run does not reduce to a closed form. The shifted Gaussian $\exp(-((\theta - \theta_1)/\sigma)^2/2)/(\sqrt{2\pi}\,\sigma)$ reads rank 2, its whole context, and is refused as irreducible. Every sufficiency residual is below 10^{-9}. Solved in two stages along the automatic split, m_1 in stage 1 and r_2 in stage 2, the potential difference comes within 10^{-7} of its joint optimum, and the same split with m_1 alone as the statistic misses it by 1.6. Appendix B sweeps the rank-drop coupling as a positive control.

To run the procedure on interfaces we did not know, we sealed a generator: for each seed it draws a context dimension between 3 and 5, a rank of zero, one, or two, degree-two polynomial interface coordinates, couplings cubic in a scalar probe, and an offset, expands the whole objective into monomials, and sums them in a seeded random order. It rejects any draw in which a node of the trace computes an interface coordinate or an affine combination of them. The protocol commits a digest of each truth to the repository before discovery runs, refuses a problem whose digest differs from the committed one, hands discovery nothing but the flat evaluation, and reads the truth only when the result is scored. Over 200 seeds, 56 separable, 64 of rank one and 80 of rank two, the procedure returned the sealed rank on all 200, refusing the separable seeds as rank zero and matching the sealed tangent space at every sampled point on the coupled ones, with every sufficiency residual below 10^{-9}. The generator is ours and its interfaces are polynomial, inside the class the fit searches. Each seed's interface was nonetheless unknown to us until scoring, including which seeds were separable.

The node-gradient search, run on the same objectives, stacks each traced node's x-Jacobian over the probes and tests whether its rows recover the interface tangent space. On the undercount objective 1 of 26 x-dependent nodes does, the accumulated coupling node; on the product formulas between 1 and 4 do, the product itself and its downstream transforms; on the rollout 5 of 85, the terminal position and the penalty built on it, every one computed after the second phase's thrusts; and on the robot arm 2 of 48, the squared distance and its root, both computed after the last two segments. On the oscillator 1 of 9 does, the output itself, and on the center of mass 0 of 15 do, since its ratio is computed nowhere and every node downstream of it also carries the offset. The search reaches the direction wherever the trace materializes the interface, or a function of it and the probe, but does not choose that node among the rest or return a dimension.

A learned abstraction is the representation-learning route to the same decomposition. We fit $f = g_\theta(x) + h_\theta(\phi_\theta(x), y)$ with small networks and an encoder of width k to 20,000 samples of the center-of-mass formula, whose ratio the polynomial fit refuses. The validation loss falls by a factor of 4.1×10^3 from $k = 0$ to $k = 1$ and then drifts, within a factor of 1.24, to its minimum at $k = 3$. We obtain the width 1 only by setting by hand a tolerance of a factor of 2 on the best loss. At $k = 1$ the encoder's gradient aligns with the ratio's to a mean cosine of 0.999. The learned route reaches an interface the polynomial fit cannot express, approximately and with a width chosen by that tolerance, where the cross-Hessian identifies the width and the direction. The cross-Hessian reading has a threshold too, but on the two ratio formulas its cutoff of 10^{-10} times the leading singular value lies more than 5 orders of magnitude above the largest one it discards.

5 Related Work: Baselines and Positioning

Closest in mechanism are exact state-abstraction methods for decision problems presented as influence diagrams (Howard & Matheson, 1984; Shachter, 1986), which search the diagram for a summary through which a decision's influence must pass, certify it graphically, and solve the problem in stages with no loss of optimality. The sufficient information set of Yuan et al. (2010) is a graphical Markov state for each decision, found as a minimum separator, that may add information arcs to the diagram; the s-reachability criterion of Koller & Milch (2003) decides which decisions are strategically relevant; and the Nielsen-Jensen line (Nielsen & Jensen, 1999; Nielsen, 2001; Jensen & Nielsen, 2007) restricts a decision's relevant past, which limited-memory influence diagrams (Lauritzen & Nilsson, 2001) make a first-class object. The summary they return must be a node of the diagram they are handed, which makes them exact and confines them to such summaries. Recent work on influence diagrams (Lee et al., 2021; Everitt et al., 2021; van Merwijk et al., 2022; Hansen et al., 2022; Oliehoek et al., 2021) has not left the declared diagram. Our method evaluates second derivatives instead, giving up the graphical guarantees to reach the case the diagram-side search must refuse, such as the center-of-mass ratio of Section 4, which is not a vertex, so a search confined to the diagram's vertices cannot return it. Two-stage stochastic programming names the linear case: the first-stage decision x reaches the expected recourse cost only through the tender, its image under a technology matrix the modeler supplies (Nazareth & Wets, 1986), so the tender's dimension is that matrix's rank. In our setting the cross-Hessian reads a nonlinear tender that nobody supplied. The same line divides this paper from factored dynamic programming (Boutilier et al., 2000), nonserial dynamic programming, bucket elimination, and the junction-tree literature (Bertelè & Brioschi, 1972; Dechter, 1999; Koller & Friedman, 2009; Wainwright & Jordan, 2008), whose separator is syntactic, read off the sparsity pattern of a factor graph it takes as given, where ours is semantic,

read off the numerical rank of a mixed derivative. Because functional-form structure invisible to the sparsity pattern can compress further, a syntactic separator is only an upper bound on the interface dimension.

In reinforcement learning, spectral option discovery (Machado et al., 2017) reads a task's structure from the Laplacian of its transition graph, estimated from sampled transitions, where the cross-Hessian reads a cross-derivative between decision blocks exactly. Exact model minimization under equivalences that preserve kernel and reward (Givan et al., 2003), the bisimulation metrics of Ferns et al. (2011), the encoders trained to reproduce them (Zhang et al., 2021), and the causal state abstractions of Wang et al. (2024) all certify a state equivalence that preserves optimal value under a stochastic transition kernel, computed from a known model in the first two and learned from interaction in the last two. MDP homomorphisms (Ravindran & Barto, 2003) give the algebraic form of the exact criterion, while the taxonomy of Li et al. (2006) orders these abstractions by what they preserve; in its terms the offset-refined interface preserves stage-2 value differences up to a stage-1 constant, a Q^*-irrelevance with an additive term allowed. Here the objective is a fully observed deterministic program, so the identification is an exact statement about second derivatives, at the cost of saying nothing about a stochastic kernel, the gap Section 6 states as open.

Active subspace methods (Constantine et al., 2014; Zahm et al., 2020) and certified dimension reduction for Bayesian inverse problems (Zahm et al., 2022) build a low-rank object from one function's own gradient, an outer product of ∇f with itself averaged over a sampling measure, and ridge recovery (Cohen et al., 2012; Fornasier et al., 2012; Tyagi & Cevher, 2014) reads the same object from point queries. Ours is a *mixed*, block-versus-rest cross-Hessian, whose rank measures coupling *between* one block of inputs and the rest of the objective. Their subspaces are also linear and so, by Example 2.3, can overcount the interface dimension. In logarithmic coordinates a linear subspace is a set of power products, which is how Constantine et al. (2017) recover the dimensionless groups of a physical law, and a ratio such as $m_2/(m_1 + m_2)$ is not a power product. Sufficient dimension reduction estimates the reduction from a sample, by inverse regression (Li, 1991; Cook & Li, 2002) or from the regression function's Hessian (Li, 1992), whose leading eigenvectors estimate our directions from a sample, and Lee et al. (2013) define the nonlinear reduction that Theorem 2.4 measures. Vanishing mixed partials are an older test. The Leontief-Sono criterion for functional separability (Leontief, 1947; Sono, 1961) is the one-dimensional case of the cross-Hessian's own rank condition, a condition on ratios of first derivatives for a group of variables to enter through a single function; Gorman (1959) gives its global form. The distinction of Blackorby et al. (1977) between separability and functional structure is ours between an interface and an offset-refined one. Interaction detection in machine learning (Friedman & Popescu, 2008; Tsang et al., 2018; Janizek et al., 2021) asks whether a fitted model's mixed partial between two features is nonzero, and does not read a rank across a block. In symbolic regression, the generalized-symmetry test of Udrescu et al. (2020) detects $f = g(h(x'), x'')$ for a scalar h by checking that the normalized gradient $\nabla_{x'} f$ of a neural-network fit does not vary with x''. That is our rank-one case without an offset, estimated from a fit. An additive term in x' alone, or a second interface coordinate, rotates that gradient and fails the test, whereas the mixed derivative removes the offset and its rank counts the coordinates. The additive-split tests of partially separable optimization (Griewank & Toint, 1982) and symbolic regression (Udrescu & Tegmark, 2020) detect a factorization $f = \sum_i f_i$, where the cross-Hessian reads the rank of a coupling that does not factor. Coupled input-output reduction (Chen et al., 2025) reduces a forward map's input and output jointly, by linear maps.

6 Limitations and Scope

The identification itself has two limitations beyond the hypotheses attached to each statement in Section 2.

One is the gap between what the cross-Hessian reports, the minimal dimension of an *analytic* interface, and what an analyst committed to a linear summary wants; Example 2.3 shows the first can lie strictly below the second, and Appendix A records what the two routes to $\underline{d}$ cost.

The other is that the instrument checks its hypotheses only in part. The sweep of Section 4 tests constant rank on its sample and the sufficiency test checks the recovered statistic against the objective, so an analyst pointing the procedure at an unfamiliar objective learns whether the returned coordinates are sufficient on the box. Analyticity, which the run does not check, enters only the

almost-sure attainment of Theorem 2.2, since the decomposition survives with twice differentiability and constant rank (Proposition A.2). Between sample points, a rank drop can escape the sweep. The numerical rank carries a conditional guarantee: by Weyl's inequality the reading stays a lower bound on the rank of S while the error in evaluating S is below 10^{-10} of its leading singular value, an error we do not bound. For the linear case, ridge recovery (Fornasier et al., 2012) carries query and noise bounds. Estimated from samples by Stein's identity, the device behind principal Hessian directions (Li, 1992), the rows of a two-block network read rank 11 where the exact stack reads 6 (Appendix B).

Three hypotheses bound the work as a whole. First, we relax the requirement that the interface be a named quantity but retain a modeler who has already chosen the split into blocks and named the coordinates the statistic may depend on. Second, because the program is deterministic, the interface summarizes one block's influence on a fixed function of the other, which covers a sample average over noise entering after the interface, as on the rollout, but not a value function defined by an expectation over a transition kernel, for which the approximate information state (Subramanian et al., 2022) is the nearest sufficient-statistic notion. We have not attempted that extension, on which any policy in a stochastic environment depends. Third, the coupling is twice differentiable on open sets, so an interface among discrete choices, such as the parity of two binary variables that a diagram draws at no vertex, lies outside the identification and remains open.

The evidence lacks one case, a decision program at scale that someone else wrote, such as a simulator or a planning model whose interface combines quantities the trace never names. We wrote the rollout and the sealed generator, whose polynomials of degree two we knew by family if not by seed. Both the Feynman formulas and the robot arm are someone else's, coded as their authors wrote them, but they are small functions with no decisions in them. On most of the formulas the product the procedure finds is also a node of the trace; the arm's interface follows from a rotation symmetry that a kinematicist would write down by hand. A trained network is a harder case for a different reason, since its low rank is soft and the reading then depends on the cutoff, as a saturated unit already shows (Appendix B).

Our demonstrated refusals cover a separable objective, a second coupling channel, a frozen chain, a collinear interface, the separable seeds, an irreducible formula, and the rank-drop sweep, which catches the rank drop only because a sample point coincides with it. At the rank-drop point itself the single-point reading undercounts, reporting a smaller interface than the coupling has nearby. For an interface outside the polynomial class, as on the two ratio formulas, the procedure returns payoff differences, which are sufficient on the box but cost evaluations of the program and do not give a closed form.

7 Conclusion

The cross-Hessian identifies the direction of the interface through which one block of a differentiable program interacts with the rest, bounds its dimension from below against every linear interface and every smooth one with a full-rank Jacobian, and, under analyticity and constant rank, measures that dimension exactly. The state a staged solver must carry, which a modeler otherwise picks by hand or by validating encoders of several widths, can therefore be computed from the program itself, as it is for a control rollout that never computes its zero-effort miss, a published robot arm, and sealed random objectives. Still open are a decision program at scale that someone else wrote and, beyond it, the stochastic and discrete boundaries.

AI use statement

Under our direction, and with all of their output reviewed by us, we used large language model assistants to formulate mathematical claims, provide critical ingredients for proofs and write them, design experiments, implement methods, and interpret results; to draft and edit text; and to search the literature. We did not use them to develop the conceptual framework, to propose or refine hypotheses, or to generate synthetic data. Translation, dataset cleaning, and qualitative data analysis are not applicable to this work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Reproducibility statement

Appendix A proves every result of Section 2 under the hypotheses stated with it, and Appendix B gives the chain constructions. The supplementary material contains the code for every experiment, the seeded generator and the digest manifest of the sealed protocol in Section 4, one script that runs every experiment the paper reports, and a test suite that checks each reported rank, cosine, residual, and gap against its tolerance.

References

Jian An and Art Owen. Quasi-Regression. *Journal of Complexity*, 17(4):588–607, 2001. doi: 10.1006/jcom.2001.0588.

Richard Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, 1957.

Umberto Bertelè and Francesco Brioschi. *Nonserial Dynamic Programming*, volume 91 of *Mathematics in Science and Engineering*. Academic Press, New York, 1972.

Charles Blackorby, Daniel Primont, and R. Robert Russell. Separability vs Functional Structure: A Characterization of Their Differences. *Journal of Economic Theory*, 15(1):135–144, 1977. doi: 10.1016/0022-0531(77)90072-2.

Craig Boutilier, Richard Dearden, and Moisés Goldszmidt. Stochastic Dynamic Programming with Factored Representations. *Artificial Intelligence*, 121(1–2):49–107, 2000. doi: 10.1016/S0004-3702(00)00033-3.

Qiao Chen, Elise Arnaud, Ricardo Baptista, and Olivier Zahm. Coupled Input-Output Dimension Reduction: Application to Goal-Oriented Bayesian Experimental Design and Global Sensitivity Analysis. *SIAM Journal on Scientific Computing*, 47:A2403–A2430, 2025. doi: 10.1137/24m1670421.

Albert Cohen, Ingrid Daubechies, Ronald DeVore, Gerard Kerkyacharian, and Dominique Picard. Capturing Ridge Functions in High Dimensions from Point Queries. *Constructive Approximation*, 35(2):225–243, 2012. doi: 10.1007/s00365-011-9147-6.

Paul G. Constantine, Eric Dow, and Qiqi Wang. Active Subspace Methods in Theory and Practice: Applications to Kriging Surfaces. *SIAM Journal on Scientific Computing*, 36(4):A1500–A1524, 2014. doi: 10.1137/130916138.

Paul G. Constantine, Zachary del Rosario, and Gianluca Iaccarino. Data-Driven Dimensional Analysis: Algorithms for Unique and Relevant Dimensionless Groups. *arXiv preprint arXiv:1708.04303*, 2017. doi: 10.48550/arXiv.1708.04303.

R. Dennis Cook and Bing Li. Dimension Reduction for Conditional Mean in Regression. *The Annals of Statistics*, 30(2):455–474, 2002. doi: 10.1214/aos/1021379861.

Rina Dechter. Bucket Elimination: A Unifying Framework for Reasoning. *Artificial Intelligence*, 113(1–2):41–85, 1999. doi: 10.1016/s0004-3702(99)00059-4.

Tom Everitt, Ryan Carey, Eric D Langlois, Pedro A Ortega, and Shane Legg. Agent Incentives: A Causal Perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11487–11495, 2021. doi: 10.1609/aaai.v35i13.17368.

Norm Ferns, Prakash Panangaden, and Doina Precup. Bisimulation Metrics for Continuous Markov Decision Processes. *SIAM Journal on Computing*, 40(6):1662–1714, 2011. doi: 10.1137/10080484x.

Massimo Fornasier, Karin Schnass, and Jan Vybiral. Learning Functions of Few Arbitrary Linear Parameters in High Dimensions. *Foundations of Computational Mathematics*, 12(2):229–262, 2012. doi: 10.1007/s10208-012-9115-y.

Jerome H. Friedman and Bogdan E. Popescu. Predictive Learning via Rule Ensembles. *The Annals of Applied Statistics*, 2(3):916–954, 2008. doi: 10.1214/07-AOAS148.

Robert Givan, Thomas Dean, and Matthew Greig. Equivalence Notions and Model Minimization in Markov Decision Processes. *Artificial Intelligence*, 147(1–2):163–223, 2003. doi: 10.1016/s0004-3702(02)00376-4.

S. M. Goldman and H. Uzawa. A Note on Separability in Demand Analysis. *Econometrica*, 32(3): 387–398, 1964. doi: 10.2307/1913043.

William M Gorman. Separable Utility and Aggregation. *Econometrica*, 27(3):469–481, 1959. doi: 10.2307/1909472.

Andreas Griewank and Philippe L. Toint. Partitioned Variable Metric Updates for Large Structured Optimization Problems. *Numerische Mathematik*, 39:119–137, 1982. doi: 10.1007/bf01399316.

Eric A. Hansen, Jinchuan Shi, and James Kastrantas. Strategy Graphs for Influence Diagrams. *Journal of Artificial Intelligence Research*, 75:1177–1221, 2022. doi: 10.1613/jair.1.13865.

Ronald A. Howard and James E. Matheson. Influence Diagrams. In Ronald A. Howard and James E. Matheson (eds.), *Readings on the Principles and Applications of Decision Analysis*, volume II, pp. 719–762. Strategic Decisions Group, Menlo Park, CA, 1984. doi: 10.1287/deca.1050.0020. Reprinted in Decision Analysis 2(3):127–143, 2005.

Joseph D. Janizek, Pascal Sturmfels, and Su-In Lee. Explaining Explanations: Axiomatic Feature Interactions for Deep Networks. *Journal of Machine Learning Research*, 22(104):1–54, 2021. doi: 10.48550/arXiv.2002.04138.

Finn V. Jensen and Thomas D. Nielsen. *Bayesian Networks and Decision Graphs*. Information Science and Statistics. Springer, 2 edition, 2007. doi: 10.1007/978-0-387-68282-2.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

Daphne Koller and Brian Milch. Multi-Agent Influence Diagrams for Representing and Solving Games. *Games and Economic Behavior*, 45(1):181–221, 2003. doi: 10.1016/s0899-8256(02)00544-4.

Steffen L. Lauritzen and Dennis Nilsson. Representing and Solving Decision Problems with Limited Information. *Management Science*, 47(9):1235–1251, 2001. doi: 10.1287/mnsc.47.9.1235.9779.

Junkyu Lee, Radu Marinescu, and Rina Dechter. Submodel Decomposition Bounds for Influence Diagrams. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 12147–12157, 2021. doi: 10.1609/aaai.v35i13.17442.

Kuang-Yao Lee, Bing Li, and Francesca Chiaromonte. A General Theory for Nonlinear Sufficient Dimension Reduction: Formulation and Estimation. *The Annals of Statistics*, 41(1):221–249, 2013. doi: 10.1214/12-AOS1071.

Thomas Lengauer and Robert Endre Tarjan. A Fast Algorithm for Finding Dominators in a Flowgraph. *ACM Transactions on Programming Languages and Systems*, 1(1):121–141, 1979. doi: 10.1145/357062.357071.

Wassily Leontief. Introduction to a Theory of the Internal Structure of Functional Relationships. *Econometrica*, 15(4):361–373, 1947. doi: 10.2307/1905335.

Ker-Chau Li. Sliced Inverse Regression for Dimension Reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991. doi: 10.1080/01621459.1991.10475035.

Ker-Chau Li. On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein's Lemma. *Journal of the American Statistical Association*, 87(420): 1025–1039, 1992. doi: 10.1080/01621459.1992.10476258.

Lihong Li, Thomas J. Walsh, and Michael L. Littman. Towards a Unified Theory of State Abstraction for MDPs. In *Proceedings of the Ninth International Symposium on Artificial Intelligence and Mathematics (ISAIM)*, 2006.

Marlos C. Machado, Marc G. Bellemare, and Michael Bowling. A Laplacian Framework for Option Discovery in Reinforcement Learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2295–2304, 2017. doi: 10.48550/arxiv.1703.00956.

Yoshitomo Matsubara, Naoya Chiba, Ryo Igarashi, and Yoshitaka Ushiku. Rethinking Symbolic Regression Datasets and Benchmarks for Scientific Discovery. *arXiv preprint arXiv:2206.10540*, 2022. doi: 10.48550/arXiv.2206.10540.

J. L. Nazareth and R. J.-B. Wets. *Algorithms for Stochastic Programs: The Case of Nonstochastic Tenders*, volume 28 of *Mathematical Programming Studies*, pp. 1–28. Springer Berlin Heidelberg, 1986. doi: 10.1007/BFb0121123.

Thomas D. Nielsen. Decomposition of Influence Diagrams. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU), Lecture Notes in Computer Science 2143*, pp. 144–155. Springer, 2001. doi: 10.1007/3-540-44652-4_14.

Thomas D. Nielsen and Finn V. Jensen. Welldefined Decision Scenarios. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 502–511. Morgan Kaufmann, 1999.

Frans A Oliehoek, Stefan Witwicki, and Leslie P Kaelbling. A Sufficient Statistic for Influence in Structured Multiagent Environments. *Journal of Artificial Intelligence Research*, 70:789–870, 2021. doi: 10.1613/jair.1.12136.

Allan Pinkus. *Ridge Functions*. Cambridge University Press, 2015. doi: 10.1017/CBO9781316408124.

Balaraman Ravindran and Andrew G. Barto. Smdp Homomorphisms: An Algebraic Approach to Abstraction in Semi-Markov Decision Processes. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1011–1016, 2003.

Ross D Shachter. Evaluating Influence Diagrams. *Operations Research*, 34(6):871–882, 1986. doi: 10.1287/opre.34.6.871.

Masazo Sono. The Effect of Price Changes on the Demand and Supply of Separable Goods. *International Economic Review*, 2(3):239–271, 1961. doi: 10.2307/2525430.

Jayakumar Subramanian, Amit Sinha, Raihan Seraj, and Aditya Mahajan. Approximate Information State for Approximate Planning and Reinforcement Learning in Partially Observed Systems. *Journal of Machine Learning Research*, 23(12):1–83, 2022. doi: 10.48550/arxiv.2010.08843.

Michael Tsang, Dehua Cheng, and Yan Liu. Detecting Statistical Interactions from Neural Network Weights. In *International Conference on Learning Representations (ICLR)*, 2018. doi: 10.48550/arXiv.1705.04977.

Hemant Tyagi and Volkan Cevher. Learning Non-Parametric Basis Independent Models from Point Queries via Low-Rank Methods. *Applied and Computational Harmonic Analysis*, 37(3):389–412, 2014. doi: 10.1016/j.acha.2014.01.002.

Silviu-Marian Udrescu and Max Tegmark. Ai Feynman: A Physics-Inspired Method for Symbolic Regression. *Science Advances*, 6(16), 2020. doi: 10.1126/sciadv.aay2631.

Silviu-Marian Udrescu, Andrew Tan, Jiahai Feng, Orisvaldo Neto, Tailin Wu, and Max Tegmark. Ai Feynman 2.0: Pareto-Optimal Symbolic Regression Exploiting Graph Modularity. In *Advances in Neural Information Processing Systems*, volume 33, 2020. doi: 10.48550/arXiv.2006.10782.

Chris van Merwijk, Ryan Carey, and Tom Everitt. A Complete Criterion for Value of Information in Soluble Influence Diagrams. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36 (9):10034–10041, 2022. doi: 10.1609/aaai.v36i9.21242.

Martin J Wainwright and Michael I Jordan. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008. doi: 10.1561/2200000001.

Zizhao Wang, Caroline Wang, Xuesu Xiao, Yuke Zhu, and Peter Stone. Building Minimal and Reusable Causal State Abstractions for Reinforcement Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2024. doi: 10.48550/arXiv.2401.12497.

Changhe Yuan, Xiaojian Wu, and Eric A. Hansen. Solving Multistage Influence Diagrams using Branch-and-Bound Search. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 691–700. AUAI Press, 2010.

Olivier Zahm, Paul G. Constantine, Clémentine Prieur, and Youssef M. Marzouk. Gradient-Based Dimension Reduction of Multivariate Vector-Valued Functions. *SIAM Journal on Scientific Computing*, 42(1):A534–A558, 2020. doi: 10.1137/18M1221837.

Olivier Zahm, Tiangang Cui, Kody J. H. Law, Alessio Spantini, and Youssef Marzouk. Certified Dimension Reduction in Nonlinear Bayesian Inverse Problems. *Mathematics of Computation*, 91 (336):1789–1835, 2022. doi: 10.1090/mcom/3737. arXiv:1807.03712 (2018).

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning Invariant Representations for Reinforcement Learning without Reconstruction. In *International Conference on Learning Representations*, 2021. doi: 10.48550/arXiv.2006.10742.

A Proofs for Identification from the Cross-Hessian

A.1 Two-Block Dominator Exactness

The claim taken as given in Section 1 is that a dominator justifies an exact two-stage rewrite. We state and prove it here, in the two-block form the paper uses.

Theorem A.1 (Two-block dominator exactness)**.** *Consider an objective $\Phi = c_0 + \sum_{j\in\mathcal{J}} \Phi_j(\mathbf{v}_{\Lambda_j})$ over a deterministic dataflow DAG, each term Φ_j reading the variables in its scope Λ_j, with decisions partitioned as $\mathbf{D} = \{d\} \sqcup \mathbf{D}_2$ and every exogenous input held fixed. Throughout, "reachable" counts the length-zero path, so a vertex is reachable from itself. Call a term coupled if its scope contains a variable reachable from d and a variable reachable from some member of $\mathbf{D}_2$, and write $\mathcal{C}(d)$ for the union of the coupled scopes. Let $\Phi^{(2)}$ be the sum of the terms whose scope contains a variable reachable from $\mathbf{D}_2$, and $\Phi^{(1)}$ the sum of c_0 and the remaining terms. Let ν be a vertex such that every directed path from d to a member of $\mathcal{C}(d)$ passes through ν, and such that no member of $\mathbf{D}_2$ is an ancestor of ν, so that ν is a function $\nu(d)$ of d alone. Then*

$$\max_{d,\,\mathbf{d}_2} \Phi \;=\; \max_{d} \Big[\Phi^{(1)}(d) + \max_{\mathbf{d}_2} \Phi^{(2)}(\nu(d), \mathbf{d}_2) \Big], \tag{8}$$

with equality rather than inequality, provided the stage-2 rule may be chosen separately at each realized value of ν.

Proof. Each stage-1 term reads no variable reachable from $\mathbf{D}_2$, so $\Phi^{(1)}$ is a function of d alone. The content of the claim is that every stage-2 term reads d only through ν. A stage-2 term that is not coupled reads no variable reachable from d, so it does not read d at all. If $\nu = d$ the coupled terms read d directly, which is reading it through ν, and there is nothing more to show. Assume henceforth $\nu \neq d$. Let Q be the set of vertices reachable from d by a directed path avoiding ν; the length-zero convention puts $d \in Q$, the construction keeps $\nu \notin Q$, and the dominator hypothesis gives $Q \cap \mathcal{C}(d) = \varnothing$. Run the induction over vertices outside $Q \cup \{\nu\}$, not merely outside Q: for such a v, a parent in Q would extend a ν-avoiding path to v and place $v \in Q$, so every parent is ν or lies again outside $Q \cup \{\nu\}$, and in particular d is not a parent of v. An induction along a topological order writes each such v as a function of ν and the stage-2 choices with no residual dependence on d. The restriction to that range is necessary, because the step fails at $v = \nu$, where d may be a parent. Every variable in a coupled scope is therefore ν or lies outside $Q \cup \{\nu\}$, so each coupled term reads d only through ν, and $\Phi^{(2)}$ is a function of $(\nu(d), \mathbf{d}_2)$ as displayed. The interchange of the two maxima is an equality, not an inequality, because the inner maximizer may be chosen separately at each realized value of ν; this is the standard interchange of maxima over a product with no cross-dependence.

□

The argument covers the two-block case only. We do not use a generalization to three or more blocks, which does not follow from it.

A.2 Proofs of the Identification Results

These proofs keep the notation of Section 2: the model (1), the cross-derivative matrix (2), the first-order span (3), and the observable stack (4).

The verification of Example 2.3 is short. With $h(\phi, y) = y(\phi_1 + \frac{1}{2}\phi_2^2)$ and $J = I_2$, the cross-derivative column is $W_\phi(y) = (1,\ \phi_2)^\top$ for every y, so $L(\phi) = \operatorname{span}\{(1, \phi_2)\}$ and $r^\star = 1$ at every base point, with no base point at which the single-base-point probe reads 2. The minimal linear offset-refined dimension is nonetheless $\underline{d} = 2$: a nonzero a with $a_1 + a_2\phi_2 = 0$ for all ϕ in an open set forces $a = 0$, so no linear functional of the interface absorbs the coupling.

We begin with Proposition 2.1, the one-sided bound.

Proof. Write $r^\star = \dim L(\phi_0)$ as in (3).

Step 1: $\hat{r} \le r^\star$ in all cases. By (4) we have $S = MJ$, so $\operatorname{rank} S \le \operatorname{rank} M$. Every row of M is some $c_b(\phi_0, y_i)^\top$ and so lies in $L(\phi_0)$, so the row space of M is a subspace of $L(\phi_0)$ and $\operatorname{rank} M \le r^\star$. This step uses only that $S = MJ$, which asks only C^2 regularity of h and differentiability of ϕ at x_0.

Step 2: part 1, the linear case. Let $a \in \ker \pi$ and $\phi \in N_1$. Since N_1 is open there is $\varepsilon > 0$ with $\phi + ta \in N_1$ for $|t| < \varepsilon$, and since $a \in \ker \pi$ we have $\pi(\phi + ta) = \pi(\phi)$. The two instances of the decomposition at $\phi + ta$ and at ϕ therefore share the same H term for every y, and subtracting them gives

$$h(\phi + ta,\, y) - h(\phi,\, y) \;=\; A(\phi + ta) - A(\phi) \qquad \text{for all } y \in V,\ |t| < \varepsilon. \tag{9}$$

The right-hand side does not depend on y. Applying $\partial/\partial y_b$, an operation that touches only h and is therefore legitimate under $h \in C^2$ regardless of how badly behaved A and H are, gives $\partial h/\partial y_b(\phi + ta, y) = \partial h/\partial y_b(\phi, y)$ for all $|t| < \varepsilon$. The map $t \mapsto \partial h/\partial y_b(\phi + ta, y)$ is then constant, and differentiable in t because $h \in C^2$, with derivative $a^\top c_b(\phi, y)$ at $t = 0$. Hence $a^\top c_b(\phi, y) = 0$. Taking $\phi = \phi_0$ and ranging over $y \in V$ and b yields $L(\phi_0) \subseteq (\ker \pi)^\perp = \operatorname{rowspace}(\pi)$, so $r^\star \le \operatorname{rank} \pi$. Minimizing over admissible π gives $r^\star \le \underline{d}$.

Step 3: part 2, the nonlinear case. Because ρ is a submersion at ϕ_0, its level set $\Sigma := \rho^{-1}(\rho(\phi_0)) \cap N_1$ is, near ϕ_0, an embedded analytic submanifold of dimension $k - s$ whose tangent space at ϕ_0 is $\ker D\rho(\phi_0)$. Fix $a \in \ker D\rho(\phi_0)$ and choose a C^1 curve ζ in Σ with $\zeta(0) = \phi_0$ and $\zeta'(0) = a$. Since $\rho(\zeta(t)) = \rho(\phi_0)$ for all small t, subtracting instances of the decomposition along the curve gives $h(\zeta(t), y) - h(\phi_0, y) = A(\zeta(t)) - A(\phi_0)$, again independent of y. Differentiating in y_b and then in t at $t = 0$, exactly as in Step 2, gives $a^\top c_b(\phi_0, y) = 0$. Hence $L(\phi_0) \subseteq (\ker D\rho(\phi_0))^\perp = \operatorname{rowspace} D\rho(\phi_0)$, a subspace of dimension s, so $r^\star \le s$.

□

The difference-quotient device in Steps 2 and 3 is required because A and H are arbitrary functions and need not be differentiable, so the direct argument, which would differentiate the decomposition to get $\nabla_\phi h = \nabla A + (D\rho)^\top \nabla_u H$, is unavailable. Restricting to a level set of the interface map removes H from the picture entirely and leaves a difference in A that is constant in y and is therefore annihilated by the very first derivative taken. The same device is what allows g in (1) to be arbitrary.

Without the submersion hypothesis in part 2 the level set through ϕ_0 may be a singular analytic set whose tangent cone $\mathcal{T}_{\phi_0}\Sigma$ spans a proper subspace of $\ker D\rho(\phi_0)$. Since curves in Σ then reach only the directions of that cone, the argument delivers the weaker bound $r^\star \le k - \dim \operatorname{span} \mathcal{T}_{\phi_0}\Sigma$. The submersion hypothesis is satisfied automatically by the interfaces that Theorem 2.4 constructs, which is the comparison the rank bound is used for.

We turn to Theorem 2.2, the reverse inequality.

Proof. If $r^\star = 0$ then $W(y) = 0$ for every y, so $\operatorname{rank} S = 0 = r^\star$ for every tuple and $Z_m = \emptyset$; the conclusion is trivial. Assume $r^\star \geq 1$.

Reduction to M. Since $J : \mathbb{R}^n \to \mathbb{R}^k$ has rank k it is surjective, so $\operatorname{im}(MJ) = M(J(\mathbb{R}^n)) = M(\mathbb{R}^k) = \operatorname{im} M$, so $\operatorname{rank} S(y) = \operatorname{rank} M(y)$ for every tuple. This is the only use of full rank. Then $Z_m = \{y : \operatorname{rank} M(y) < r^\star\}$, which is the common zero set of all $r^\star \times r^\star$ minors of M, each continuous in the tuple, so Z_m is closed in V^m.

A nonvanishing minor. Since $\dim L(\phi_0) = r^\star$, choose probes $z_1, \ldots, z_{r^\star} \in V$ and column indices $b_1, \ldots, b_{r^\star} \in \{1, \ldots, p\}$ such that the vectors $w_j := c_{b_j}(\phi_0, z_j)$ form a basis of $L(\phi_0)$. Because $m \geq r^\star$, the row indices $(1, b_1), (2, b_2), \ldots, (r^\star, b_{r^\star})$ are pairwise distinct rows of M, their first coordinates being distinct. Fix any $v \in V$ and set

$$y^* \;:=\; (z_1, \ldots, z_{r^\star},\, v, \ldots, v) \in V^m. \tag{10}$$

The rows of $M(y^*)$ at those indices are $w_1^\top, \ldots, w_{r^\star}^\top$, which are linearly independent, so that $r^\star \times k$ submatrix has rank $r^\star$ and some set Γ of $r^\star$ column indices makes the corresponding square submatrix nonsingular. Having fixed such a Γ, define $\Delta : V^m \to \mathbb{R}$ as the minor of $M(y)$ on those rows and the columns Γ. The order matters, because Γ is chosen from the tuple y^* and the row set before Δ is defined from it.

Each entry of $M(y)$ is a partial derivative of the real-analytic h evaluated at (ϕ_0, y_i) and is real-analytic in y_i, so Δ is real-analytic on V^m, and $\Delta(y^*) \neq 0$ by construction.

Conclusion. Since V is open and connected, so is V^m in $\mathbb{R}^{mp}$. The zero set of a real-analytic function on a connected open set that is not identically zero is closed, has empty interior by the identity theorem, and has Lebesgue measure zero. Since Z_m lies in the zero set of Δ, the set Z_m, closed by the reduction step, has empty interior and measure zero. Finally, a probability law absolutely continuous with respect to Lebesgue measure assigns Z_m probability zero, and $\operatorname{rank} S \leq r^\star$ always, so $\hat{r} = r^\star$ almost surely. Independence across probes is not needed, only absolute continuity of the joint law.

□

If J has rank $k' < k$, a weaker statement holds. One has $\operatorname{rank} S(y) = \dim(J^\top L(\phi_0))$ at its maximum, which equals $\dim L - \dim(L \cap (\operatorname{im} J)^\perp)$. The minor argument applies with two changes: the probe count must satisfy $m \geq \dim(J^\top L)$, and the probes and column indices must be chosen so that the images $J^\top c_{b_j}(\phi_0, z_j)$ are linearly independent. Choosing a basis of L and then applying $J^\top$ does not suffice, since $J^\top$ may collapse it.

Read as a whole, the argument is two moves. Full rank of J, used nowhere else, makes the map $M \mapsto MJ$ rank-preserving, so the observable S sees exactly the span $L(\phi_0)$ rather than its image under some projection. Rank preservation converts a statement about the parameterization ϕ into one about h. Then one exhibits a single $r^\star \times r^\star$ minor of M that is real-analytic in the probes and nonzero at one explicitly constructed tuple, and appeals to closedness and nullity of the zero set of a nontrivial real-analytic function on a connected open set. Both connectedness of V and analyticity in y are needed here, since a probe domain in two pieces, or a coupling that is flat on part of the probe range, produces a deficiency set with nonempty interior.

The probe-count hypothesis $m \geq r^\star$ is sharp and cannot be lowered to $\lceil k/p \rceil$: taking $h(\phi, y) = \phi_1 y_1 + \phi_2 y_1^2/2$ generates all of L through a single probe coordinate, so one needs $r^\star$ separate probes no matter how large p is. A large probe dimension does not substitute for probe count when the cross-derivative columns vary rather than accumulate. Since $r^\star \leq k$ always, the requirement falls on $r^\star$, and $m \geq k$ is always a safe threshold.

Last comes Theorem 2.4, which says what the measured number is.

Proof. Set $s := r^\star$, fix a reference probe $y_* \in V$, and put $u_y(\phi) := h(\phi, y) - h(\phi, y_*)$.

The span is generated by gradients. We claim $L(\phi) = \operatorname{span}\{\nabla_\phi u_y(\phi) : y \in V\}$ for $\phi \in N_1$. For the inclusion of the right side in the left, join y_* to y by a smooth path ζ in V, possible because V is open and connected. Differentiating $u_y(\phi) = \int_0^1 \frac{d}{dt} h(\phi, \zeta(t))\, dt$ in ϕ gives $\nabla_\phi u_y(\phi) = \int_0^1 \sum_b c_b(\phi, \zeta(t))\, \zeta_b'(t)\, dt$, an integral of an integrand taking values in the closed subspace $L(\phi)$. For the reverse inclusion, $c_b(\phi, y) = \lim_{\varepsilon \to 0} \varepsilon^{-1}(\nabla_\phi u_{y+\varepsilon e_b}(\phi) - \nabla_\phi u_y(\phi))$ lies in the closed span of those gradients.

An involutive distribution. Let $\mathcal{D}(\phi) := L(\phi)^\perp$, of constant dimension $k - s$ on N_1 by hypothesis. Choose probes and column indices whose cross-derivative columns are independent at ϕ_0. As functions of ϕ, these columns remain independent on a possibly smaller neighborhood, which we again call N_1, and by constancy of $\dim L(\phi)$ they span $L(\phi)$ there, so $\mathcal{D}$ is the kernel of an analytic surjective bundle map and is an analytic distribution. It is involutive: for vector fields X, Y with values in $\mathcal{D}$ and any y, the one-form $\omega := du_y$ is exact, hence closed, so $0 = d\omega(X, Y) = X(\omega(Y)) - Y(\omega(X)) - \omega([X, Y]) = -\omega([X, Y])$, and since $\mathcal{D} = \bigcap_y \ker(du_y)$ by the previous paragraph, the bracket $[X, Y]$ again takes values in $\mathcal{D}$.

Straightening. By the Frobenius theorem in the analytic category there are a neighborhood N_2 of ϕ_0 and analytic coordinates $(w, z) \in \mathbb{R}^s \times \mathbb{R}^{k-s}$ on N_2, with connected slices, in which $\mathcal{D} = \operatorname{span}\{\partial/\partial z\}$. Put $\rho(\phi) := w(\phi)$, an analytic submersion onto an open subset of $\mathbb{R}^s$. Each u_y has vanishing z-derivatives and is constant on the connected slices, so $u_y(\phi) = \tilde{u}_y(\rho(\phi))$. Setting $A(\phi) := h(\phi, y_*)$ and $H(w, y) := u_y(\sigma(w))$ for a local analytic section σ of ρ gives $h(\phi, y) = A(\phi) + H(\rho(\phi), y)$ on $N_2 \times V$ with H jointly analytic. Minimality is part 2 of Proposition 2.1, whose submersion hypothesis is satisfied by this ρ by construction.

□

The straightening step is, in substance, the classical theorem on functional dependence under constant rank, applied to the whole family $\{u_y\}_{y \in V}$ at once. It also shows that the interface need not be searched for in any function class, because the program's own payoff differences realize it.

Corollary A.1.1 (Payoff differences realize the interface)**.** *Under the hypotheses of Theorem 2.4, fix $y_* \in V$ and write $u_y(\phi) := h(\phi, y) - h(\phi, y_*)$. There exist probes $y_1, \dots, y_{r^\star} \in V$ whose gradients $\nabla_\phi u_{y_i}(\phi_0)$ are linearly independent. For any such probes the map $\rho' := (u_{y_1}, \dots, u_{y_{r^\star}})$ can serve as the submersion ρ of Theorem 2.4 on a possibly smaller neighborhood of ϕ_0. In the coordinates of the program, $u_{y_i}(\phi(x)) = f(x, y_i) - f(x, y_*)$, because the offset g cancels, so $r^\star + 1$ evaluations of f compute an interface.*

Proof. The first step of the preceding proof gives $L(\phi_0) = \operatorname{span}\{\nabla_\phi u_y(\phi_0) : y \in V\}$, a space of dimension $r^\star$, so $r^\star$ of those gradients are linearly independent. Let ρ, A, H and N_2 be as in the theorem. Subtracting the decomposition at y_* from the one at y gives $u_y(\phi) = H(\rho(\phi), y) - H(\rho(\phi), y_*) =: \tilde{u}_y(\rho(\phi))$ on N_2, with $\tilde{u}_y$ real-analytic on $\rho(N_2)$. Put $\tilde{u} := (\tilde{u}_{y_1}, \dots, \tilde{u}_{y_{r^\star}})$, so that $\rho' = \tilde{u} \circ \rho$ and $D\rho'(\phi_0) = D\tilde{u}(\rho(\phi_0))\, D\rho(\phi_0)$. The left side has rank $r^\star$ by the choice of probes, so the $r^\star \times r^\star$ matrix $D\tilde{u}(\rho(\phi_0))$ is invertible, and by the analytic inverse function theorem $\tilde{u}$ has a real-analytic inverse near $\rho(\phi_0)$. On the corresponding neighborhood $N_3 \subseteq N_2$ of ϕ_0 we have $\rho = \tilde{u}^{-1} \circ \rho'$, hence $h(\phi, y) = A(\phi) + H(\tilde{u}^{-1}(\rho'(\phi)), y)$ there, with $(w, y) \mapsto H(\tilde{u}^{-1}(w), y)$ real-analytic. Since $D\rho'(\phi_0)$ has full rank $r^\star$ and full rank is an open condition, ρ' is a real-analytic submersion after shrinking N_3.

□

Analyticity is not what makes the decomposition exist. It enters Theorem 2.2, where it makes the deficient probe tuples a null set, and it makes ρ, A and H analytic. With twice differentiability alone the payoff differences still realize the interface.

Proposition A.2 (Exactness under twice differentiability)**.** *Assume $h \in C^2(N \times V)$ with V open and connected, and suppose $\dim L(\phi) = r^\star$ for every ϕ in a neighborhood $N_1 \subseteq N$ of ϕ_0. Fix $y_* \in V$, write $u_y(\phi) := h(\phi, y) - h(\phi, y_*)$, and choose probes $y_1, \dots, y_{r^\star}$ whose gradients $\nabla_\phi u_{y_i}(\phi_0)$ are linearly independent. Then $F := (u_{y_1}, \dots, u_{y_{r^\star}})$ is a C^2 submersion on a neighborhood $N_2 \subseteq N_1$ of ϕ_0, and there is a function H on $F(N_2) \times V$, C^2 in its first argument, with*

$$h(\phi, y) \;=\; h(\phi, y_*) \;+\; H\big(F(\phi),\, y\big) \qquad \textit{for all } (\phi, y) \in N_2 \times V. \tag{11}$$

Part 2 of Proposition 2.1 holds for every C^1 map ρ that is a submersion at ϕ_0, since its proof uses only that the level set through ϕ_0 is, by the C^1 implicit function theorem, an embedded C^1 submanifold with tangent space $\ker D\rho(\phi_0)$*, and a C^1 curve in it. Hence $r^\star$ is the minimal dimension of an offset-refined interface that is a C^1 submersion at ϕ_0.*

Proof. The first step of the proof of Theorem 2.4 differentiates under an integral along a path in V, which a piecewise-C^1 polygonal path in the open connected V supplies, and takes a difference quotient in y, both legitimate for $h \in C^2$, so $L(\phi) = \operatorname{span}\{\nabla_\phi u_y(\phi) : y \in V\}$ on N_1 and the probes exist. Each u_{y_i} is C^2 in ϕ, and the rows $\nabla_\phi u_{y_i}(\phi)$ of $DF(\phi)$ are independent at ϕ_0, hence on a neighborhood contained in N_1, where F is a submersion. Those rows lie in $L(\phi)$, which has dimension $r^\star$, so they span it, and $L(\phi) = (\ker DF(\phi))^\perp$ there. Completing F with $k - r^\star$ linear functionals whose restriction to $\ker DF(\phi_0)$ is an isomorphism gives a map with invertible differential at ϕ_0, and the inverse function theorem then supplies C^2 coordinates $(w, z) \in \mathbb{R}^{r^\star} \times \mathbb{R}^{k-r^\star}$ on a box N_2 in which $F(\phi) = w$, so each level set of F in N_2 is a connected slice whose tangent spaces are $\ker DF$. Fix $y \in V$. Along any C^1 curve γ in a slice, $\frac{d}{dt} u_y(\gamma(t)) = \nabla_\phi u_y(\gamma(t))^\top \gamma'(t) = 0$, because $\nabla_\phi u_y \in L = (\ker DF)^\perp$ and $\gamma' \in \ker DF$. Hence u_y is constant on each slice, and $H(w, y) := u_y(\sigma(w))$, for the section $\sigma(w)$ with coordinates (w, z_0), satisfies $u_y(\phi) = H(F(\phi), y)$ on N_2 for every $y \in V$ at once.

□

The independence of the gradients $\nabla_\phi u_{y_i}$ follows from that of the observable gradients $\nabla_x[f(x, y_i) - f(x, y_*)] = J^\top \nabla_\phi u_{y_i}$, which is the condition the discovery run checks.

The constant-rank hypothesis of Theorem 2.4 is not removable, as a counterexample small enough to state in full shows.

Example A.3 (Constant rank is not removable)**.** *Take $n = k = p = 1$, $\phi = \mathrm{id}$, $\phi_0 = 0$, and $h(\phi, y) = \phi^2 y/2$ on $\mathbb{R} \times \mathbb{R}$. Then $W_\phi(y) = \phi$, so $L(\phi) = \operatorname{span}\{\phi\}$ and $\dim L(\phi)$ jumps from 0 at the base point to 1 at every other point. The first-order span thus drops rank exactly at ϕ_0, where the hypothesis of Theorem 2.4 fails.*

The cross-Hessian reads $r^\star = \dim L(0) = 0$, and reads it for every probe tuple. But the minimal analytic interface dimension near ϕ_0 is 1, not 0: an interface of dimension 0 would mean $h(\phi, y) = A(\phi) + H(y)$, which forces the cross-derivative $\partial^2 h / \partial\phi\, \partial y$ to vanish identically near 0, whereas it equals ϕ. Without constant rank the observed rank strictly undershoots even the nonlinear dimension, which falsifies the conclusion of Theorem 2.4.

Example 2.3 and Example A.3 are not interchangeable. In Example 2.3 the constant-rank hypothesis holds, the cross-Hessian is correct about the nonlinear dimension, and the discrepancy is with the linear target. In Example A.3 the constant-rank hypothesis fails and the cross-Hessian is simply wrong, with no change of target able to repair it. Only the second is a limitation of the method.

Remark A.4 (Two routes to the linear dimension and their costs)**.** *Example 2.3 shows that the minimal linear offset-refined dimension $\underline{d}$ is not identified from cross-Hessians at a single base point. The two routes below recover it and target the linear quantity rather than the nonlinear one that Theorem 2.4 identifies.*

Multi-base-point probing. Drop the restriction to a fixed base point and stack blocks $W_{\phi_i}(y_i)^\top$ at probe pairs $(\phi_i, y_i) \in N \times V$. The proof of Theorem 2.2 applies verbatim with $L(\phi_0)$ replaced by the span of all cross-derivative columns over $N \times V$ and with V^m replaced by $(N \times V)^m$, which is open and connected when N and V are. The stacked rank then equals $\underline{d}$ almost surely under any absolutely continuous probe law. This leaves the hypotheses on h unchanged but has a cost: the block observable at a point x_i is $W_{\phi(x_i)}(y_i)^\top D\phi(x_i)$, not $W_{\phi(x_i)}(y_i)^\top$, so recovering the probe blocks requires knowing $D\phi(x_i)$ and applying the pseudoinverse frame correction $D\phi(x_i)^+ = D\phi(x_i)^\top (D\phi(x_i) D\phi(x_i)^\top)^{-1}$, available near x_0 because full rank is an open condition. Without that correction, stacking the raw blocks across different base points mixes row spaces expressed in different frames, which can break even the elementary bound of Proposition 2.1.

Taylor-augmented probing. Alternatively, stay at the single base point and differentiate further in the interface variable. With R_q the span of all $\partial_\phi^\alpha W_{\phi_0}(y)e_j$ with $|\alpha| \le q$, over $y \in V$ and $j \le p$, the dimensions $r_q := \dim R_q$ are nondecreasing in q, bounded by k, and converge to $\underline{d}$ when h is analytic in ϕ and N is connected, with $r_0 = r^\star$. Each r_q is a valid lower bound. The route identifies $\underline{d}$ only in the limit and does not supply an effective termination criterion, since stabilization $r_q = r_{q+1}$ at one step does not prove that the limit has been reached and no a priori bound on the required order is available for general analytic h.

Both routes answer the linear question, so both can overshoot the quantity of interest. In Example 2.3 each returns 2, correctly, while a one-dimensional analytic sufficient statistic exists and the single-base-point cross-Hessian reads exactly 1. A discovery method that asks through how few numbers the coupling can be routed wants the nonlinear answer, whereas a method that "repairs" itself by varying the base point will report the linear one and will report it correctly, having answered a different question.

B Chain Composition and Further Readings

Backward induction fixes the upstream decisions before treating the next split, so a cross-Hessian taken on the frozen problem reports *its* interface rather than the chain's. With $\phi_2(d_1, d_2) = (d_1 + d_2, d_1 d_2)$ and $h_2(\eta, d_3) = \eta_1 d_3 + \eta_2 d_3^2$ the split-2 interface is two-dimensional, yet fixing d_1 at any value makes both coordinates functions of d_2 alone and the reading is 1 at every frozen value. Neither the rank nor the recovered direction varies with d_1, so no sweep detects it. Probing the upstream decisions instead of freezing them costs one wider column block and reads the interface the chain actually has, 2 on this example.

We ran the probed loop on a three-decision chain with inlined residual interfaces $R_1 = s_0 - \psi_1(d_1)$ and $R_2 = R_1 - \psi_2(d_2)$, each ψ_j a quadratic penalty about its own anchor, which the trace does not compute at any node. Both splits read rank 1 and recover ∇R_1 and ∇R_2 to within the angle stated in Section 3. A four-decision extension reads rank 1 at each of its three splits. Two further readings exercise the hypotheses directly. The rank-drop coupling of Example A.3, $h = x^2 y/2$, reads rank 0 at the base point $x = 0$ for every probe tuple and rank 1 at $x = 0.3$, the misread that example predicts. Since a random sweep meets the rank-drop point with probability zero, the discovery procedure's positive control sweeps a grid of 41 points placed through it, where one reading of rank zero against rank one elsewhere makes the procedure refuse the coupling. Rescaling the undercount context into units that stretch x_1 by 10^3 and shrink x_3 by 10^3, and the reverse, leaves the rank at 2 and the plane aligned with the rescaled interface gradients to within the same angle, the rows of S transforming as covectors under the diagonal change of units, $\nabla_z \phi = \lambda \odot \nabla_x \phi$ for $z = x/\lambda$. Three constructed negatives guard the chain readings. With the coupling removed altogether the stack reads rank 0; with d_1 reaching downstream through a second channel, a penalty anchored on d_1 as well as the residual, it reads rank 2; and on the collapse example above the frozen reading stays at 1 while probing d_1 reads 2.

A two-block network has a body $z_n = \tanh(B\xi_n + \beta)$ with 4 units on inputs $\xi_n \in \mathbb{R}^3$, a linear head $\kappa^\top z_n + \kappa_0$, and the squared loss over 2 fixed inputs. The context is the body's 16 parameters and the probe the head's 5, so the interface is the activation matrix, of dimension at most 2×4. With the last unit saturated by a bias of -14, three random heads read rank 3 on one input and 6 on two, inputs times the 3 units that carry signal, aligned with the live activation gradients to within an angle of 1.5×10^{-7} radians, and with no unit saturated the same network reads 8. From derivatives of the loss alone, without knowledge of the architecture, the cross-Hessian rules out every summary of the body narrower than 6 numbers, linear or an analytic submersion. Two scopes govern the reading. The activations are nodes of the trace, so the node-gradient search of Section 4 reaches this interface too, through 4 of 8 nodes, the head's output and the loss built on it, and the cross-Hessian adds the count. Since the saturated channel is present in exact arithmetic, where the analytic interface dimension is 8, the live width is also a threshold reading: 6 counts the channels above 10^{-10} times the strongest.

The network has the loss

$$f\big((B,\beta),(\kappa,\kappa_0)\big) \;=\; \tfrac{1}{2}\textstyle\sum_n \big(\kappa^\top z_n + \kappa_0 - t_n\big)^2 \tag{12}$$

with targets t_n, context $x = (B, \beta)$ and probe $y = (\kappa, \kappa_0)$. The head reaches the body only through the activation matrix $Z = (z_{n,j})$, so in (1) the interface is $\phi(x) = Z$. Differentiating in κ_j gives $\sum_n (\kappa^\top z_n + \kappa_0 - t_n)\, z_{n,j}$, whose x-gradient lies in the span of the activation gradients $\nabla_x z_{n,j} = (1 - z_{n,j}^2)(\xi_n, 1)$ on unit j's block of parameters. The saturating bias scales the last unit's gradients by $\operatorname{sech}^2(-14) \approx 2.8 \times 10^{-12}$. The saturated channel falls below the cutoff but remains measurable: on one input its singular value is 2.5×10^{-12} of the leading one, as the sech^2 factor predicts, far above the float64 floor near 10^{-16}. Saturation that sets in only near the base point is the rank drop Example A.3 warns about.

Two noisy readings of the same network show the value of the exact reading. Estimating each row by Stein's identity from 10^5 Gaussian draws of width 0.05 about the body reads rank 11 where the exact stack reads 6, because the design's curvature makes the sampling noise rank-increasing. Its first extra singular value, 1.8×10^{-4} of the leading one, falls only as the width over the square root of the sample size (1.8×10^{-3} at 10^3 draws), far above the cutoff at any feasible sample. Averaging the loss over input noise of scale 0.05, discretized on 5 quadrature nodes, behaves the same way. Each node carries its own activations, the extra channels enter at 2.9×10^{-4} of the leading singular value and scale as the square of the noise, and the default cutoff reads rank 9 while a cutoff stated at 10^{-2} reads 6.