WHICH TASK DIFFERENCES MATTER? THE GEOMETRY OF OPTIMAL SHARED ITERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A solver that reuses one learned coefficient schedule across many tasks can underperform because of insufficient training, a restrictive update rule, or missing task information, but its loss cannot identify the cause. For positive-definite quadratics, we separate these sources of excess risk and characterize which task differences change the optimal program, the residual polynomial the schedule implements. Task laws with different spectra and oracle risks can require the same program. Given the population risk of every depth-T program (equivalently, moments through degree $2T$), moment vectors sharing an optimal depth-K program form affine fibers of dimension $\max\{K-2T,0\}$ through each interior point. Across fibers, the leading oracle-relative cost of sharing a program is the squared radius of the smallest ball enclosing the linearized optimal programs in the local risk metric. Under this criterion, the minimax program is optimal for a least-favorable mixture of laws; we prove a sharp worst-case bound, attained by smooth positive densities, on how many laws this mixture needs. A momentum recurrence implements this program while preserving the optimal first T steps shared by all laws in the family.

1 INTRODUCTION

Observed risk can hide three different limitations of a shared solver: insufficient training, a restrictive update rule, and task differences that the solver cannot resolve. Separating them matters because each calls for a different remedy.

We study solvers that apply one learned sequence of scalar coefficients to every task. On a positive-definite quadratic, this *coefficient schedule* determines a terminal residual polynomial, which we call the *program*. Different schedules can implement the same program. Fixed-coefficient unrolled gradient methods have this form (Chen et al., 2022; Takabe and Wadayama, 2022), as does the restricted scalar-attention recurrence examined in Appendix F.5; for looped linear attention with shared weights on in-context linear regression, the population-loss minimizer repeats one preconditioned gradient step (Gatmiry et al., 2024). Adaptive Krylov methods such as CG and MINRES (Hestenes and Stiefel, 1952; Paige and Saunders, 1975; Paquette and Trogdon, 2023) instead choose coefficients per instance through inner products and are far stronger when those are available (Appendix F.1); shared schedules are the relevant class when coefficients must be fixed in advance, as in unrolled networks and iterations that avoid inner products (Gutknecht and Röllin, 2002; Adams et al., 2003). We ask which differences between task laws matter when coefficients must be shared.

Which part of the error can better training remove? A two-task example separates the cost of sharing a program from the cost of restricting its execution. Let $r_0 = e_1$ and

$$A_s = \begin{pmatrix} 3 & 1 & 0 \\ 1 & s & 1 \\ 0 & 1 & 3 \end{pmatrix}, \qquad s \in \{2, 4\}. \tag{1}$$

Both have $A_s r_0 = (3,1,0)^\top$. A first step with stepsize η leaves residual $(I - \eta A_s)r_0$ and has risk $1 - 6\eta + 10\eta^2$, so both choose $\eta = 0.3$. Their depth-two optimal programs differ, with minimum risks $1/35$ for $s = 2$ and $1/131$ for $s = 4$. The best shared program has worst-case relative risk 1.315, reached on both tasks.

The second step may combine the current residual with the previous state through momentum. A reset instead sets its momentum coefficient to zero and restricts the terminal polynomial to

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.

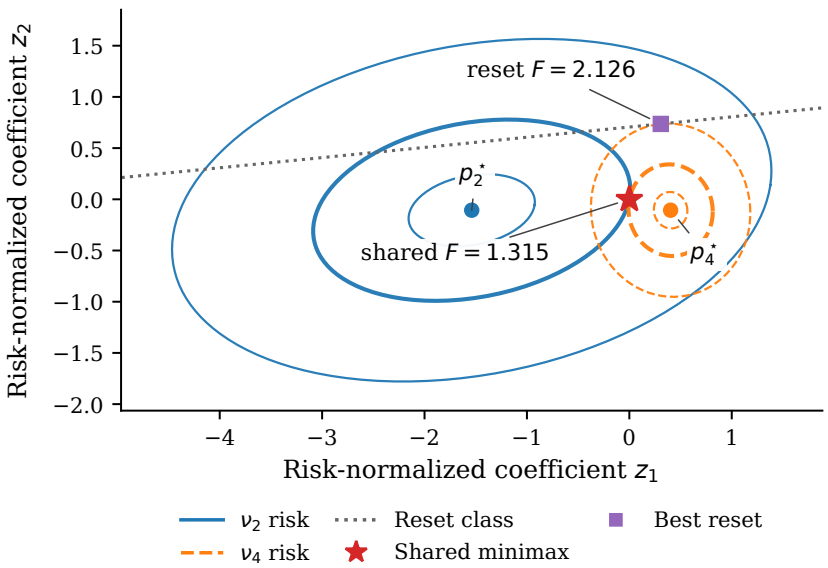


Figure 1: Relative-risk contours of the spectral laws induced by (1), at levels 1.05, 1.315, and 2.126. The centers are the task-specific optimal programs; the star is the best shared program, and the dotted line is the reset class. Coordinates are centered at the shared minimax, whitened by $\bar{H} = \frac{1}{2}\{H_2/R_2^\star(\nu_2) + H_4/R_2^\star(\nu_4)\}$, and rotated to align the oracle centers (Appendix A.7).

$(1 - \eta_1\lambda)(1 - 0.3\lambda)$. The best worst-case relative risk in this restricted class is 2.126 (Figure 1). Relative to the oracle baseline of one, sharing costs 0.315 and the reset restriction adds 0.811. Both programs are already optimal in their respective classes, so more training cannot remove either cost.

Contributions. We characterize exactly which uncertainty changes an optimal shared program. Different spectral laws can share one optimal program; their moment vectors lie on affine fibers whose dimension and transverse rank r we compute (Theorem 1). The same geometry shows that the leading local information price is the squared radius of the smallest enclosing ball. A least-favorable mixture yields a minimax program executable by a momentum recurrence that preserves the common optimal prefix (Theorem 2). At most $r + 1$ laws suffice to certify this program, and this worst-case bound is attained by smooth positive densities, even when constituents may be chosen anywhere inside the family (Theorem 3); to leading order the certificate size is the contact number $k_\star \le r + 1$ of the enclosing ball (Proposition B.1). Under a fixed prior, the value of refining an observation is the expected risk-weighted squared movement of the chosen program, and minimax values are bracketed by the same quantity, so equally accurate observations can have zero or positive value (Section 5).

Numerical scope. Figures 1, 2 and 4 to 6 illustrate these results; Section 6 studies a failure outside the design family under absolute risk, and Appendix F reports the pre-specified audit and a scalar-attention study.

Related work. Spectral methods optimize an iteration for a known law (Pedregosa and Scieur, 2020; Scieur and Pedregosa, 2020; Lacotte and Pilanci, 2020) or a known interval (Golub and Varga, 1961; Fischer, 1996; Agarwal et al., 2021). Related analyses study spectral-tail sensitivity (Cunha et al., 2022) and normal operators (Domingo-Enrich et al., 2021). We ask which differences among plausible laws alter the optimal program when only their low-order moments are known. Optimal-design theory supplies the mixture optimization (Elfving, 1952; Pukelsheim, 2006; Sagnol, 2011), and kernel-polynomial recurrences for a known law supply its execution (Stiefel, 1958; Fischer, 1996; Pedregosa and Scieur, 2020). We add the exact same-program fibers and their rank, a sharp spectral realization of the certificate size, and prefix-preserving attainment under moment uncertainty.

Learned optimizers and unfolded methods train reusable updates (Andrychowicz et al., 2016; Wichrowska et al., 2017; Gupta and Roughgarden, 2017; Chen et al., 2022; Takabe and Wadayama,

2022); linear and looped attention connect learning to iterative computation (Akyürek et al., 2023; Ahn et al., 2023; von Oswald et al., 2023; Zhang et al., 2024; Mahankali et al., 2024; Gatmiry et al., 2024). Here the reusable object is a scalar residual polynomial. Performance estimation and its stepsize designs bound worst-case performance over function classes (Drori and Teboulle, 2014; Taylor et al., 2017; Grimmer, 2024; Altschuler and Parrilo, 2025), and a Wasserstein-robust variant over lifted representations of trajectories is used to certify fixed methods (Park et al., 2025) and to learn stepsizes (Ranjan et al., 2026). We instead isolate the spectral uncertainty that changes the selected program and compare observations by the resulting decision value (Blackwell, 1951; Raiffa and Schlaifer, 1961; Reid and Williamson, 2011).

2 Problem Setup

2.1 Shared Polynomial Iterations

A task consists of $Aw = g$ and an initial point w_0, with $A \succ 0$, $\mathrm{spec}(A) \subset [a, b] \subset (0, \infty)$, and $r_0 = Aw_0 - g \neq 0$. A *regime* is a distribution of tasks, represented for our objective by a residual-weighted spectral probability law ν:

$$R_\nu(p) = \mathbb{E}\frac{\|p(A)r_0\|^2}{\|r_0\|^2} = \int p^2 \, d\nu, \qquad \mathcal{A}_t = \{p : \deg p \le t,\ p(0) = 1\}. \tag{2}$$

In the analysis we call ν a *law*. A coefficient schedule is a sequence of stepsize–momentum pairs (η_t, γ_t), and a *continuation* consists of the steps after an executed prefix. Fixed-coefficient full-history linear methods generate exactly $p(A)r_0$ with $p \in \mathcal{A}_K$ after K steps (Appendix B.2). The *oracle risk* is $R^\star_K(\nu) = \min_{p \in \mathcal{A}_K} R_\nu(p)$; a minimizing polynomial is an *oracle program*.

When the program is chosen, we take as known the population risk of every depth-T program, $R_\nu|_{\mathcal{A}_T}$. This is the objective of a depth-T shared solver on the task population: it fixes the optimal prefix but not its continuation, and it differs from a single task's T-step record, which reveals only that task's moments (Appendix A.2). As a quadratic function of the coefficients, it identifies exactly the moments $m_j = \int \lambda^j \, d\nu$ through degree $2T$ (Appendix A.2). Thus $m_0, \dots, m_{2T}$ are fixed, while the higher moments needed for a target depth $K > T$ remain uncertain. Let $\mathcal{M}$ be a nonempty convex, weakly compact family of laws sharing those known moments. Our results are conditional on this information and on $\mathcal{M}$; estimating either from finite data is a separate statistical problem (Section 7).

2.2 Optimization, Restriction, and Information

Oracle-relative risk compares each law with what it could attain on its own. This prevents the larger absolute baseline from automatically dominating the comparison in the example (1). A positive offset τ limits the influence of nearly zero oracle risks. For $\tau \ge 0$, define

$$\ell_{\nu,\tau}(p) = \frac{R_\nu(p) + \tau}{R^\star_K(\nu) + \tau}, \qquad F_{\mathcal{M}}(p) = \max_{\nu \in \mathcal{M}} \ell_{\nu,\tau}(p), \qquad V_\tau = \min_{p \in \mathcal{A}_K} F_{\mathcal{M}}(p). \tag{3}$$

We write F for $F_{\mathcal{M}}$ when the family is fixed and $V_\tau(\mathcal{M})$ when it must be explicit. The positive-denominator and Gram assumptions are stated in Theorem 2. At $\tau = 0$ they exclude the class of all laws on $[a, b]$ with the known moments: that class contains a law with at most $T + 1$ atoms, given by Gauss quadrature (Golub and Meurant, 2010), whose oracle risk is zero because $T + 1 \le K$. Oracle-relative risk with $\tau = 0$ therefore needs a restricted family, as in our constructions, or $\tau > 0$.

For a feasible subclass $\mathcal{C} \subseteq \mathcal{A}_K$, set $V_{\mathcal{C}} = \inf_{q \in \mathcal{C}} F_{\mathcal{M}}(q)$. The excess of any $p \in \mathcal{C}$ separates into

$$F_{\mathcal{M}}(p) - 1 = \underbrace{F_{\mathcal{M}}(p) - V_{\mathcal{C}}}_{\text{optimization gap}} + \underbrace{V_{\mathcal{C}} - V_\tau}_{\text{restriction price}} + \underbrace{V_\tau - 1}_{\text{information price}}. \tag{4}$$

All three terms use the same family, depth, and risk criterion. Better training reduces the first; a richer update class can reduce the second. For a class outside $\mathcal{A}_K$, such as attention with matrix-valued weights, the middle term becomes a class-change term of either sign (Appendix F.5 compares the two classes under a query-prediction loss). Additional information can reduce the third by allowing the solver to select among programs rather than use one for every law. If p is evaluated on another family $\mathcal{D}$, then $F_{\mathcal{D}}(p) - 1$ equals the right side plus $F_{\mathcal{D}}(p) - F_{\mathcal{M}}(p)$, a term of either sign.

3 Geometry of Optimal Shared Programs

Do different task laws always require different optimal programs? Two positive densities on $[1, 2]$ show that the answer is no. At $T = 1$ and $K = 3$, they share the first two moments but have different higher moments and different oracle risks. Nevertheless, the same cubic $p_\star$ minimizes both risks (Figure 2). The exact polynomial and density construction are given in Appendix A.6.

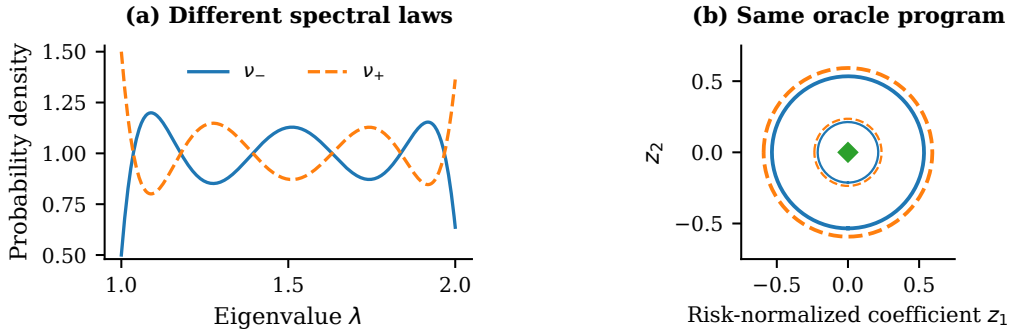


Figure 2: (a) Positive densities ν_- and ν_+ with the same first two moments but different higher moments and oracle risks. (b) Their relative-risk contours at levels 1.05 and 1.315 on a plane through $p_\star$, in coordinates normalized as in Figure 1. The contours share one center because the optimum is common; the Grams nearly coincide while the oracle risks differ by a factor 1.23, so they are nested. In Figure 1 the centers differ. The common optimum is verified exactly in Appendix A.6.

To describe all such coincidences, write $x = (m_{2T+1}, \dots, m_{2K}) \in \mathbb{R}^d$, where $d = 2(K - T)$. For $p_c = 1 + \sum_{j=1}^{K} c_j \lambda^j$, define

$$H(x)_{ij} = m_{i+j}, \quad b(x)_i = m_i, \quad c(x) = -H(x)^{-1} b(x), \quad R^\star(x) = 1 - b(x)^\top H(x)^{-1} b(x), \tag{5}$$

for $1 \le i, j \le K$ wherever $H(x) \succ 0$. Here H is the free-coefficient Gram matrix, $c(x)$ the oracle coefficient vector, and $R^\star(x)$ its risk. Completing the square gives

$$\ell_{x,\tau}(c) - 1 = \frac{(c - c(x))^\top H(x)(c - c(x))}{R^\star(x) + \tau}. \tag{6}$$

For a fixed candidate c, the normal equations $H(x)c + b(x) = 0$ are affine in x. Their feasible solution set is the *same-program fiber* of c. The following result gives both the dimension of this set and the local cost of disagreement between fibers.

Theorem 1 (Same-program fibers and local price)**.** *Let x_0 be an interior point of the feasible unresolved-moment domain, with $H(x) \succ 0$ near x_0 and $R^\star(x_0) + \tau > 0$. Write $H_0 = H(x_0)$ and $c^0 = c(x_0)$. Set $c_0^0 = 1$, extend the coefficients by zero outside degrees $0, \dots, K$, and define $B_{i,h} = c^0_{h-i}$ for $1 \le i \le K$, $2T + 1 \le h \le 2K$.*

(a) Exact fibers and rank. *For feasible x with $H(x) \succ 0$,*

$$\begin{gathered} c(x) = c^0 \iff B(x - x_0) = 0, \qquad Dc(x_0) = -H_0^{-1} B, \\ \operatorname{rank} Dc(x_0) = r := \min\{K, 2(K - T)\}. \end{gathered} \tag{7}$$

(b) Local information price. *Let $\mathsf{S} = -H_0^{-1/2} B / \sqrt{R^\star(x_0) + \tau}$. Let U be a fixed compact convex set; since x_0 is interior, $x_0 + \rho U$ is feasible for small $\rho > 0$. Writing $V_\tau(X)$ for the minimax value over laws whose unresolved moments lie in X, we have*

$$V_\tau(x_0 + \rho U) - 1 = \rho^2 \min_y \max_{u \in U} \|y - \mathsf{S}u\|^2 + O(\rho^3). \tag{8}$$

The underlying risk expansion has a remainder bounded uniformly over $u \in U$ and bounded candidate displacements as $\rho \downarrow 0$.

3.1 Same-Program Fibers

Theorem 1(a) shows that only r of the $d = 2(K - T)$ unknown moment directions change the oracle program; the remaining $d - r$ span the fiber. Within a fiber, the law changes but its optimal program does not. The oracle risk $R_K^\star = \int p_\star \, d\nu = 1 + \sum_{j=1}^{K} c_j m_j$ is affine along the fiber and generally varies on it, as in Figure 2. Equivalently, a fiber collects the moment vectors of laws for which $\lambda \, d\nu(\lambda)$ has the same K Gauss nodes, the reciprocal steps of a Richardson implementation of the program; the Gauss weights, constrained by the known moments, parametrize it (Remark A.1). When $K \le 2T$, $r = d$: the oracle program and known moments determine every moment through degree $2K$. When $K > 2T$, the fiber through an interior point has dimension $K - 2T$. The pair in Figure 2 realizes this ambiguity with positive densities whose oracle risks differ.

3.2 Local Information Price

Theorem 1(b) gives the leading minimax cost of using one program across competing laws, measured in the local risk metric, in which squared distance is excess relative risk. The map S sends a moment perturbation to the first-order change in the oracle program in this metric. The center of the smallest ball enclosing $\mathsf{S}U$ gives the leading displacement of the optimal shared program; its squared radius gives the leading information price. Rank counts the relevant directions; their images under S determine the cost. For the pair in Figure 2, the moment difference lies in the kernel of S and adds nothing to the price. Figure 3 draws this mechanism.

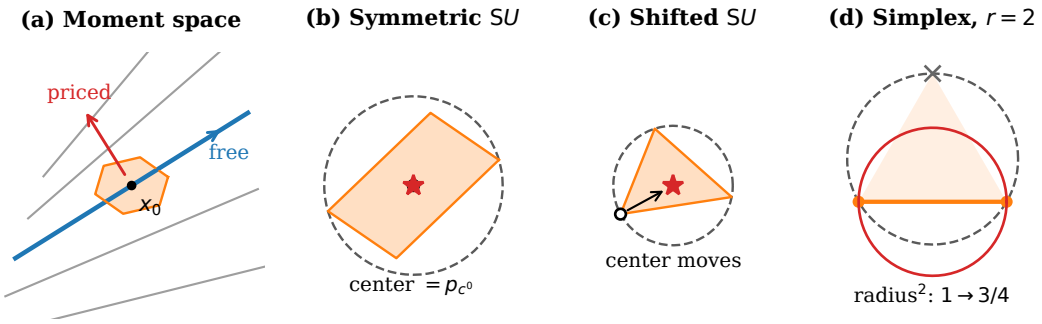


Figure 3: Schematic of Theorem 1. (a) In unresolved-moment space, same-program fibers are straight but not parallel; moving along the fiber through x_0 leaves the program unchanged, while transverse moves change it. (b–d) Program space in the local risk metric: the leading information price is ρ^2 times the squared radius of the smallest ball enclosing $\mathsf{S}U$ (dashed), and the ball's center (star) is the best shared program. (b) If $U = -U$, the center is the reference program p_{c^0}. (c) In this shifted example the center moves, so refitting changes the program at first order; asymmetry alone does not force this, as the centered simplex in (d) shows. (d) For a regular simplex with $r = 2$, deleting a vertex shrinks the squared radius from 1 to $3/4$, the $1/r^2$ loss of Theorem 3.

Remark (Symmetric uncertainty). If $U = -U$, the enclosing ball is centered at zero. The reference program p_{c^0}, optimal at x_0, is already minimax to second order, so refitting improves its worst-case relative risk by at most $O(\rho^3)$. A shift or asymmetry produces a first-order coefficient change when it moves the enclosing center (Figure 3(b,c)).

Figure 4 compares three spectral families on $[1, 9]$ sharing moments through degree eight. At $T = 4$ and $K = 5$, the symmetric family's gain divided by ρ^2 tends to zero as $\rho \downarrow 0$; the shifted and asymmetric triangles, whose risk-metric centers are nonzero, have positive limits. The families are obtained by degree-nine and degree-ten Chebyshev density perturbations (Appendix D).

4 Minimax Synthesis

The optimal shared program can be executed without discarding the common optimal prefix. Some mixture of candidate laws has an oracle program that is minimax for the entire family, and the same mixture certifies that no shared program has lower worst-case risk. For laws $\nu_1, \ldots, \nu_q \in \mathcal{M}$ and

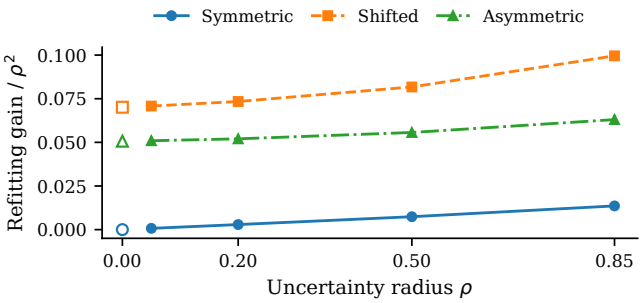


Figure 4: Gain from refitting the reference program, divided by ρ^2, for three uncertainty sets sharing a four-step optimal prefix. Hollow markers at zero are the analytic limits. The symmetric family keeps its leading enclosing-ball center at zero; the other two do not. Oracle-relative risk, $\tau = 0$.

weights $\omega_i \geq 0$, $\sum_i \omega_i = 1$, define

$$J(\omega, \nu_{1:q}) = \frac{R_K^\star(\sum_i \omega_i \nu_i) + \tau}{\sum_i \omega_i R_K^\star(\nu_i) + \tau}. \tag{9}$$

This ratio lower-bounds every shared program's worst-case relative risk (Appendix B.1). The mixture is a *certificate* of this lower bound; its component laws are the *constituents*.

4.1 Attainable Minimax Continuation

Let p_t^{sh} be the common depth-t oracle program for $t \leq T$. Let $\mathcal{C}_K$ contain the programs obtained by continuing from $(p_{T-1}^{\mathrm{sh}}, p_T^{\mathrm{sh}})$ under

$$p_{t+1} = (1 + \gamma_t - \eta_t\lambda)p_t - \gamma_t p_{t-1}, \quad t = T, \ldots, K-1. \tag{10}$$

At $T = 0$, start with $p_{-1} = p_0 = 1$. Write $V_\tau^{\mathrm{cont}} = \inf_{p \in \mathcal{C}_K} F(p)$.

Theorem 2 (Attainable minimax continuation)**.** *Let $\mathcal{M}$ be a nonempty convex family of spectral laws on $[a, b] \subset (0, \infty)$, compact in the weak topology, with common moments through degree $2T$. For $\tau > 0$, assume that at least one member has a positive-definite degree-K Gram; for $\tau = 0$, assume $R_K^\star(\nu) > 0$ for every $\nu \in \mathcal{M}$. Then*

$$V_\tau^{\mathrm{cont}} = V_\tau = \max_{\{(\omega_i, \nu_i)\}} J(\omega, \nu_{1:q}). \tag{11}$$

An attaining mixture can be chosen with at most $r + 1 = \min\{K + 1, 2(K - T) + 1\}$ laws and a positive-definite degree-K Gram. Its depth-t oracle programs satisfy (10) *and coincide with p_t^{sh} for $t \leq T$.*

Once the least-favorable mixture is found, its known-law recurrence implements the minimax program. The mixture problem is a multiresponse c-optimal design (Elfving, 1952; Sagnol, 2011), and the recurrence is the kernel-polynomial construction for a known law (Stiefel, 1958; Pedregosa and Scieur, 2020). In the opening example, the star in Figure 1 is the oracle program of the mixture $0.489\,\nu_2 + 0.511\,\nu_4$, whose certificate value is 1.315. Mixing preserves the common moments, so its first T oracle programs coincide with the family's common optimal prefix. If that prefix has been executed, only the continuation schedule changes (Appendix B.2); for a one-step continuation ($K = T + 1$) from a learned prefix that only approximates it, Proposition C.1 bounds the resulting excess risk. The executed pair matters beyond its terminal program. For the uniform law on $[1, 2]$ with $T = 2$ and $K = 3$, the schedule $(\eta_0, \gamma_0) = (1, 0)$, $(\eta_1, \gamma_1) = (130/291, -25/291)$ reaches the depth-two oracle through $p_1 = 1 - \lambda$ rather than $p_1^{\mathrm{sh}} = 1 - 9\lambda/14$; the best final step from this pair has relative risk 1.89 at $\tau = 0$, whereas the oracle pair attains 1. For $T \geq 1$, one step from a pair $(q_{T-1}, p_T^{\mathrm{sh}})$ reaches only the affine plane $\{(1 + \gamma - \eta\lambda)p_T^{\mathrm{sh}} - \gamma q_{T-1}\}$, which contains the depth-$(T + 1)$ oracle exactly when $q_{T-1} = p_{T-1}^{\mathrm{sh}}$ (Appendix C.3). A finite list of candidate laws is called a *library*. Its optimal mixture is the solution of a convex design problem; Procedure 1 and its stopping rule are in Appendix B.3.

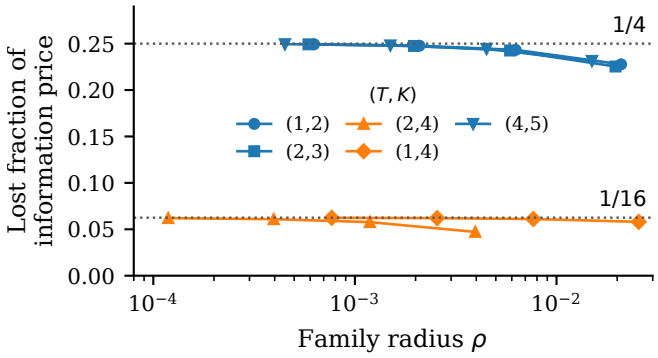


Figure 5: Smooth-density constructions of Theorem 3, labeled by (T, K). After one of $r+1$ vertices is removed and the weights are reoptimized on the remaining face, the lost fraction of the information price approaches $1/r^2$ (dotted) as ρ decreases, with $r = 2$ (blue) or $r = 4$ (orange). The family is $x_0 + \rho U$ with $\mathsf{S}U$ the unit regular simplex, as in the proof of Theorem 3. Oracle-relative risk, $\tau = 0$.

4.2 Certificate Size and Sharpness

Theorem 2 gives a certificate with at most $r+1$ constituents. These laws are the active set a finite-library solver keeps to certify optimality (Appendix B.3), not training tasks or stored programs. The next theorem shows that fewer cannot suffice in general. Let $J_{\le q}(\mathcal{M})$ be the largest value of (9) among mixtures of at most q members of $\mathcal{M}$.

Theorem 3 (Sharp spectral certificate size)**.** *For every $0 \le T < K$, interval $[a, b] \subset (0, \infty)$, and $\tau \ge 0$, there is a family $(\mathcal{M}_\rho)_{0<\rho\le\rho_0}$ of compact convex sets of smooth positive spectral densities with common moments through degree $2T$. The densities approach the uniform density as $\rho \downarrow 0$. For sufficiently small ρ, every maximizing certificate needs at least $r+1$ laws. For $1 \le q \le r$, the best certificate with at most q constituents, which may be any members of $\mathcal{M}_\rho$, satisfies*

$$V_\tau(\mathcal{M}_\rho) - J_{\le q}(\mathcal{M}_\rho) = \rho^2 \frac{r+1-q}{rq} + O(\rho^3),$$
$$V_\tau(\mathcal{M}_\rho) - 1 = \rho^2 + O(\rho^3). \tag{12}$$

The proof realizes a regular simplex of first-order oracle displacements by smooth positive densities. The squared enclosing radius is the largest variance supported on this simplex (Lim and McCann, 2022; Pass, 2023). A distribution on at most $q \le r$ points misses at least $(r+1-q)/(rq)$ of that variance (Appendix B.5); Figure 3(d) draws the case $r = q = 2$. Figure 5 shows the vertex-deletion loss of the smooth-density constructions approaching $1/r^2$ as ρ decreases. More generally, to leading order the smallest certificate has $k_\star$ laws, the fewest points of $\mathsf{S}U$ at the enclosing radius whose convex hull contains the center (Proposition B.1); when that radius is positive, $k_\star = 2$ for centrally symmetric U, and the regular simplex of Theorem 3 attains $k_\star = r+1$.

5 Value of Information

A more accurate label need not have greater decision value. The relevant comparison is between the oracle programs selected with and without the observation.

5.1 Observation Value

Let s index a finite collection of deployment laws ν_s with positive-definite degree-K Grams. An observation Y, with likelihood $f_s(y)$, is obtained before the program is chosen and is independent of the deployment task given s. For a prior π, the Bayes program is the oracle program for

$$\bar{\nu}_{\pi,y} = \sum_s \omega_s(y)\nu_s, \qquad \omega_s(y) \propto \frac{\pi_s f_s(y)}{\zeta_s}, \qquad \zeta_s = R_K^\star(\nu_s) + \tau.$$

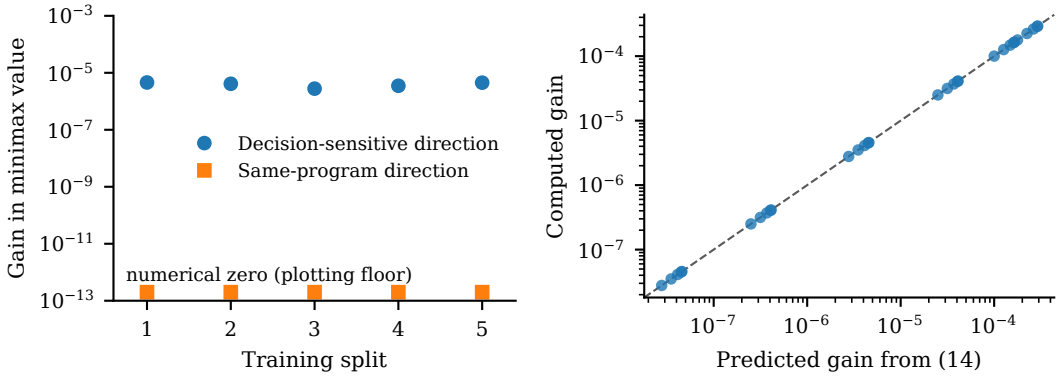


Figure 6: Two channels with 80% regime-label accuracy. Left: gains at $\rho = 0.1$ along decision-sensitive and same-program directions; the fiber channel changes the value by at most 1.2×10^{-13}. Right: computed gains for the decision-sensitive channels against (14) (dashed: equality), one point per split and radius $\rho \in \{0.01, 0.03, 0.1, 0.3, 0.6, 0.8\}$. Oracle-relative risk, $\tau = 0$.

The factor $1/\zeta_s$ accounts for oracle-relative risk; a least-favorable prior makes the rule minimax.

Let Z be a function of Y, hence a coarser observation. Write c_Y, c_Z for the Bayes coefficient vectors and $h_Y(\pi), h_Z(\pi)$ for their expected relative risks under the same prior. Then

$$h_Z(\pi) - h_Y(\pi) = \mathbb{E}_\pi[(c_Z - c_Y)^\top W_Y (c_Z - c_Y)], \quad W_Y = \mathbb{E}_\pi[H_s/\zeta_s \mid Y], \tag{13}$$

where H_s is the free-coefficient Gram of ν_s. This is the squared-loss case of the comparison of experiments (Blackwell, 1951; Raiffa and Schlaifer, 1961): under a fixed prior, the value is the expected squared change in the chosen program, weighted by its conditional risk matrix. The minimax gain lies between the same quantity evaluated at the two least-favorable priors (Appendix C).

For two regimes with moment vectors $x_0 \pm \rho u$, consider a binary channel that reports the wrong one with probability $e \in (0, 1/2)$. Let V_e and V_{no} be the minimax values with and without the channel. Equal prior weights are least favorable in the leading symmetric quadratic problem, giving

$$V_{\text{no}} - V_e = (1 - 2e)^2 \rho^2 \|\mathsf{S}u\|^2 + O(\rho^3). \tag{14}$$

Along an exact same-program fiber, even perfect identification has zero decision value. Across fibers, revealing which regime holds can help even when the best unobserved program stays at the center to first order.

5.2 Equal Accuracy, Unequal Decision Value

Figure 6 compares two equally accurate channels. In a controlled construction, for each of five base laws pooled from diabetes ridge training splits, moment-preserving perturbations generate one pair along an exact fiber and another along a decision-sensitive direction (Appendix D.2). Both channels report the regime correctly with probability 0.8. Only the decision-sensitive channel improves the minimax value, in agreement with (14). Law identification and program selection need not be the same problem.

6 Spectral Misspecification

An optimal program need not remain reliable on tasks outside its design family. The empirical family constrains risk at its spectral atoms, so a program can be small there and grow rapidly elsewhere. A diabetes ridge example shows the consequence (Figure 7). For tasks formed from 64 sampled data rows (Efron et al., 2004),

$$A = X^\top X/64 + \lambda_{\text{reg}} I, \qquad g = X^\top y/64, \qquad w_0 = 0, \quad r_0 = -g.$$

The minimax design minimizes the largest regime-mean risk R_ν from (2), an absolute rather than oracle-relative criterion; mean training minimizes their average. The preprocessing and complete protocol are in Appendix F.1.

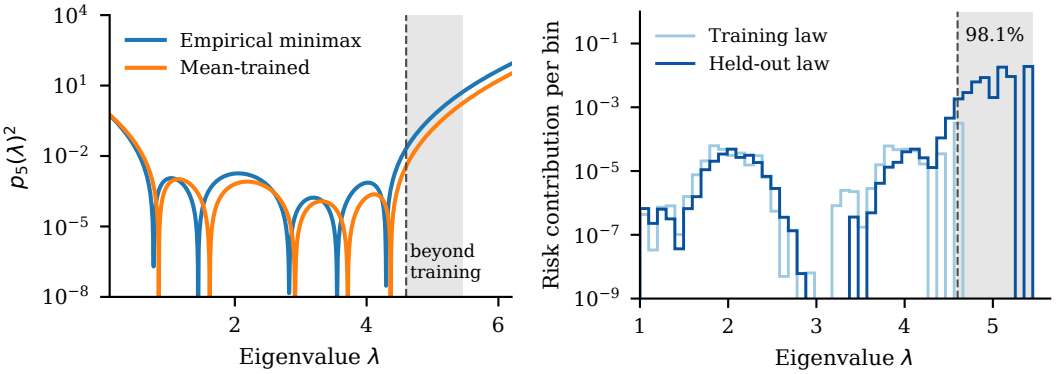


Figure 7: Diabetes ridge development split 101 at $K = 5$ (Appendix F.1), the split with the largest minimax/mean ratio among the five development splits. Left: squares of the residual polynomials. Right: the minimax polynomial's contributions to the worst regime's mean risk R_ν under its training and held-out laws. The dashed line is the training maximum 4.60, and the shaded band extends to the largest test eigenvalue 5.45; 98.1% of minimax test risk lies in this band. Across the 95 primary audit splits, the median share is 0.594 for held-out tasks and 0.403 for new tasks from the training rows (Table 6).

Held-out tasks place substantial spectral weight where the minimax polynomial is large. The mean-trained polynomial grows less rapidly beyond the training maximum. Because the empirical family has no atoms there, even exact optimization leaves this region unconstrained; the broader audit documents the associated transfer losses (Appendix F). Held-out tasks bootstrap 64 of 132 rows rather than 64 of 310. In an order-dependent, descriptive decomposition (Appendix F.3), the median log-risk increments for new tasks from the training rows, the smaller pool, and new rows are 0.251, 0.615, and 0.006 for the shared minimax (0.016, 0.286, and 0.058 for mean training), so at the median most of the increase appears before new rows enter; new rows give the widest split-to-split spread. Probability mass does not measure distance in the risk metric of Theorem 1(b). With $\mathcal{H}_\rho = H_0^{-1/2} H_\rho H_0^{-1/2}$ comparing a mixture Gram H_ρ with the base Gram H_0, a 0.1% mixture extending the support gives $\|\mathcal{H}_\rho - I\|_2 = 837$, whereas weight 10^{-8} gives 8.38×10^{-3}, where the local prediction is within 2% of the exact excess (Appendix D.3, Table 1). Proposition E.1 bounds the price by this deviation, the variation of the oracle-risk denominators, which can be large even along a fiber, and the linearized radius.

7 Discussion

The opening example and the same-program pair illustrate opposite possibilities. Identical available information can conceal different optimal programs, while different laws can require exactly the same program. The fibers identify the distinctions that leave the program unchanged. The risk metric measures disagreement between programs, and a least-favorable mixture yields an optimal shared program executable from the common optimal prefix.

Our results concern scalar-coefficient iterations on positive-definite quadratics with a specified family of spectral laws. The fiber identity and minimax attainment are exact; the enclosing-ball formula describes the leading local price. Within this setting, better training closes the optimization gap, a richer update rule can reduce the restriction price, and observations are valuable when they resolve disagreement about the program to run. In a restricted scalar-attention model, paired coefficient movements follow the query oracle along decision-sensitive directions, with median Gram-metric cosine at least 0.9998 over initializations; along its first-order null directions the query oracle barely changes, and training does not resolve the smallest such movements (Appendix F.5).

Choosing the family from finite data remains a statistical problem (Kong and Valiant, 2017). A close fit to the observed spectrum can miss regions where the chosen program amplifies residuals. The geometry tells us which task distinctions matter once the family is specified. What must be learned is the part of the task law that changes the computation.

AI USE STATEMENT

In this work, we used generative AI tools to write the proofs, design the numerical constructions and the pre-specified audit, generate the synthetic spectral laws, preprocess the diabetes dataset, implement all methods and experiments, and interpret the results. We have not used generative AI tools to develop the theoretical framework, formulate the mathematical claims, provide the critical ingredients of their proofs, or propose and refine hypotheses; translation and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to search, identify, and summarize the literature, suggest experimental parameters, create the figures, propose the title and the structure of the paper, draft the text, format the references, perform numerical checks, and edit the paper for readability. We have reviewed all AI-assisted work: the authors checked every proof line by line and reviewed and tested all code. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is primarily theoretical. Its empirical audit uses only the publicly available, de-identified diabetes dataset of Efron et al. (2004) (442 patients, distributed with scikit-learn); we collected no new data and conducted no experiments involving human subjects. We are not aware of conflicts of interest or sponsorship relevant to this work. The authors have read and adhere to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

Appendices A to C and E contain complete proofs of all results. Appendix D specifies the controlled spectral constructions, and Appendix F specifies the ridge audit, including data preprocessing, task generation, seeds, and policies.

REFERENCES

Mark Adams, Marian Brezina, Jonathan Hu, and Ray Tuminaro. Parallel multigrid smoothing: Polynomial versus Gauss–Seidel. *Journal of Computational Physics*, 188(2):593–610, 2003. https://doi.org/10.1016/S0021-9991(03)00194-3.

Naman Agarwal, Surbhi Goel, and Cyril Zhang. Acceleration via fractal learning rate schedules. In *ICML*, PMLR 139:87–99, 2021. https://proceedings.mlr.press/v139/agarwal21a.html.

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. In *NeurIPS 36*, pages 45614–45650, 2023. https://proceedings.neurips.cc/paper_files/paper/2023/hash/8ed3d610ea4b68e7afb30ea7d01422c6-Abstract-Conference.html.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? Investigations with linear models. In *ICLR*, 2023. https://openreview.net/forum?id=0g0X4H8yN4I.

Jason M. Altschuler and Pablo A. Parrilo. Acceleration by stepsize hedging: Silver Stepsize Schedule for smooth convex optimization. *Mathematical Programming*, 213(1–2):1105–1118, 2025. https://doi.org/10.1007/s10107-024-02164-2.

Marcin Andrychowicz, Misha Denil, Sergio Gómez Colmenarejo, Matthew W. Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando de Freitas. Learning to learn by gradient descent by gradient descent. In *NeurIPS 29*, pages 3981–3989, 2016. https://proceedings.neurips.cc/paper/2016/hash/fb87582825f9d28a8d42c5e5e5e8b23d-Abstract.html.

David Blackwell. Comparison of experiments. In Jerzy Neyman, editor, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 93–102. University of California Press, 1951. https://projecteuclid.org/euclid.bsmsp/1200500222.

Tianlong Chen, Xiaohan Chen, Wuyang Chen, Howard Heaton, Jialin Liu, Zhangyang Wang, and Wotao Yin. Learning to optimize: A primer and a benchmark. *Journal of Machine Learning Research*, 23(189):1–59, 2022. https://jmlr.org/papers/v23/21-0308.html.

Leonardo Cunha, Gauthier Gidel, Fabian Pedregosa, Damien Scieur, and Courtney Paquette. Only tails matter: Average-case universality and robustness in the convex regime. In *ICML*, PMLR 162:4474–4491, 2022. https://proceedings.mlr.press/v162/cunha22a.html.

Yohann De Castro, Fabrice Gamboa, Didier Henrion, Roxana Hess, and Jean-Bernard Lasserre. Approximate optimal designs for multivariate polynomial regression. *The Annals of Statistics*, 47(1):127–155, 2019. https://doi.org/10.1214/18-AOS1683.

Holger Dette and William J. Studden. *The Theory of Canonical Moments with Applications in Statistics, Probability, and Analysis*. Wiley, 1997.

Carles Domingo-Enrich, Fabian Pedregosa, and Damien Scieur. Average-case acceleration for bilinear games and normal matrices. In *ICLR*, 2021. https://openreview.net/forum?id=H0syOoy3Ash.

Yoel Drori and Marc Teboulle. Performance of first-order methods for smooth convex minimization: A novel approach. *Mathematical Programming*, 145(1–2):451–482, 2014. https://doi.org/10.1007/s10107-013-0653-0.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004. https://doi.org/10.1214/009053604000000067.

Gustav Elfving. Optimum allocation in linear regression theory. *The Annals of Mathematical Statistics*, 23(2):255–262, 1952. https://doi.org/10.1214/aoms/1177729442.

Bernd Fischer. *Polynomial Based Iteration Methods for Symmetric Linear Systems*. Wiley–Teubner Series Advances in Numerical Mathematics. Wiley and B. G. Teubner, 1996. https://doi.org/10.1007/978-3-663-11108-5.

Khashayar Gatmiry, Nikunj Saunshi, Sashank J. Reddi, Stefanie Jegelka, and Sanjiv Kumar. Can looped Transformers learn to implement multi-step gradient descent for in-context learning? In *ICML*, PMLR 235:15130–15152, 2024. https://proceedings.mlr.press/v235/gatmiry24b.html.

Gene H. Golub and Gérard Meurant. *Matrices, Moments and Quadrature with Applications*. Princeton University Press, 2010.

Gene H. Golub and Richard S. Varga. Chebyshev semi-iterative methods, successive overrelaxation iterative methods, and second order Richardson iterative methods. Part I. *Numerische Mathematik*, 3:147–156, 1961. https://doi.org/10.1007/BF01386013.

Benjamin Grimmer. Provably faster gradient descent via long steps. *SIAM Journal on Optimization*, 34(3):2588–2608, 2024. https://doi.org/10.1137/23M1588408.

Rishi Gupta and Tim Roughgarden. A PAC approach to application-specific algorithm selection. *SIAM Journal on Computing*, 46(3):992–1017, 2017. https://doi.org/10.1137/15M1050276.

Martin H. Gutknecht and Stefan Röllin. The Chebyshev iteration revisited. *Parallel Computing*, 28(2):263–283, 2002. https://doi.org/10.1016/S0167-8191(01)00139-9.

Magnus R. Hestenes and Eduard L. Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49(6):409–436, 1952. https://doi.org/10.6028/jres.049.044.

Ronald A. Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26, 1966. https://doi.org/10.1109/TSSC.1966.300074.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964. https://doi.org/10.1214/aoms/1177703732.

Samuel Karlin and William J. Studden. *Tchebycheff Systems: With Applications in Analysis and Statistics*. Interscience, 1966.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Weihao Kong and Gregory Valiant. Spectrum estimation from samples. *The Annals of Statistics*, 45(5):2218–2247, 2017. https://doi.org/10.1214/16-AOS1525.

M. G. Kreĭn and A. A. Nudel'man. *The Markov Moment Problem and Extremal Problems*. Volume 50 of *Translations of Mathematical Monographs*. American Mathematical Society, 1977.

Jonathan Lacotte and Mert Pilanci. Optimal randomized first-order methods for least-squares problems. In *ICML*, PMLR 119:5587–5597, 2020. https://proceedings.mlr.press/v119/lacotte20a.html.

Tongseok Lim and Robert J. McCann. Geometrical bounds for variance and recentered moments. *Mathematics of Operations Research*, 47(1):286–296, 2022. https://doi.org/10.1287/moor.2021.1125.

Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45:503–528, 1989. https://doi.org/10.1007/BF01589116.

Arvind Mahankali, Tatsunori Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. In *ICLR*, 2024. https://proceedings.iclr.cc/paper_files/paper/2024/hash/87f476af4053961667c2c08e9f4b850e-Abstract-Conference.html.

Christopher C. Paige and Michael A. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM Journal on Numerical Analysis*, 12(4):617–629, 1975. https://doi.org/10.1137/0712047.

Elliot Paquette and Thomas Trogdon. Universality for the conjugate gradient and MINRES algorithms on sample covariance matrices. *Communications on Pure and Applied Mathematics*, 76:1085–1136, 2023. https://doi.org/10.1002/cpa.22081.

Jisun Park, Vinit Ranjan, and Bartolomeo Stellato. Data-driven analysis of first-order methods via distributionally robust optimization. arXiv:2511.17834, 2025.

Brendan Pass. Generalized barycenters and variance maximization on metric spaces. *Journal of Fixed Point Theory and Applications*, 25, Article 5, 2023. https://doi.org/10.1007/s11784-022-01015-x.

Fabian Pedregosa and Damien Scieur. Acceleration through spectral density estimation. In *ICML*, PMLR 119:7553–7562, 2020. https://proceedings.mlr.press/v119/pedregosa20a.html.

Friedrich Pukelsheim. *Optimal Design of Experiments*. Volume 50 of *Classics in Applied Mathematics*. SIAM, 2006. https://doi.org/10.1137/1.9780898719109.

Howard Raiffa and Robert Schlaifer. *Applied Statistical Decision Theory*. Harvard University, 1961; Wiley Classics Library reprint, 2000.

Vinit Ranjan, Jisun Park, and Bartolomeo Stellato. Distributionally-robust learning to optimize. arXiv:2605.06585, 2026.

Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12(22):731–817, 2011. https://jmlr.org/papers/v12/reid11a.html.

Guillaume Sagnol. Computing optimal designs of multiresponse experiments reduces to second-order cone programming. *Journal of Statistical Planning and Inference*, 141(5):1684–1708, 2011. https://doi.org/10.1016/j.jspi.2010.11.031.

Damien Scieur and Fabian Pedregosa. Universal average-case optimality of Polyak momentum. In *ICML*, PMLR 119:8565–8572, 2020. https://proceedings.mlr.press/v119/scieur20a.html.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958. https://doi.org/10.2140/pjm.1958.8.171.

Eduard L. Stiefel. Kernel polynomials in linear algebra and their numerical applications. In *Further Contributions to the Solution of Simultaneous Linear Equations and the Determination of Eigenvalues*, volume 49 of *National Bureau of Standards Applied Mathematics Series*, pages 1–22, 1958.

Satoshi Takabe and Tadashi Wadayama. Convergence acceleration via Chebyshev step: Plausible interpretation of deep-unfolded gradient descent. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E105-A(8):1110–1120, 2022. https://doi.org/10.1587/transfun.2021EAP1139.

Adrien B. Taylor, Julien M. Hendrickx, and François Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161(1–2):307–345, 2017. https://doi.org/10.1007/s10107-016-1009-3.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *ICML*, PMLR 202:35151–35174, 2023. https://proceedings.mlr.press/v202/von-oswald23a.html.

Olga Wichrowska, Niru Maheswaranathan, Matthew W. Hoffman, Sergio Gómez Colmenarejo, Misha Denil, Nando de Freitas, and Jascha Sohl-Dickstein. Learned optimizers that scale and generalize. In *ICML*, PMLR 70:3751–3760, 2017. https://proceedings.mlr.press/v70/wichrowska17a.html.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained Transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024. https://jmlr.org/papers/v25/23-1042.html.

A PROOFS: POLYNOMIAL RISK AND SAME-PROGRAM GEOMETRY

Main result	Proof location
Polynomial risk, moment information, regularity	Appendices A.1 to A.3
Exact same-program sets and local price	Appendices A.4 and A.5
Executable minimax certificate	Appendices B.1 and B.2
Sharp smooth-density construction	Appendix B.5
Noisy observations and decision value	Appendix C
Learned pair and perturbation bound	Appendix C.3

A.1 QUADRATIC ORACLE AND CONTINUITY

For a degree-ordered basis $\psi = (\psi_0, \dots, \psi_K)^\top$, let $z = \psi(0)$ and $G_\nu = \int \psi\psi^\top d\nu$. If $G_\nu \succ 0$, constrained quadratic minimization gives

$$\widetilde{c}_\nu = \frac{G_\nu^{-1} z}{z^\top G_\nu^{-1} z}, \qquad R_K^\star(\nu) = \frac{1}{z^\top G_\nu^{-1} z}. \tag{A.1.1}$$

Indeed, the stationarity equation is $2G_\nu\widetilde{c}+\mu z = 0$ and normalization is $z^\top\widetilde{c} = 1$. Positive definiteness gives uniqueness. In the free monomial coordinates of Equation (5), expanding and completing the square gives

$$R_x(c) = 1 + 2b(x)^\top c + c^\top H(x) c, \qquad c(x) = -H(x)^{-1} b(x),$$
$$R_x(c) - R^\star(x) = (c - c(x))^\top H(x)(c - c(x)).$$

These are the usual Christoffel extremal formulas (De Castro et al., 2019).

We also need continuity when a Gram can be singular. On a positive interval, minimizing polynomials can be chosen from the compact root family

$$\left\{ \prod_{j=1}^{K} (1 - \lambda/\xi_j) : \xi_j \in [a, b] \right\}.$$

For a full-rank law, the minimizer is the normalized degree-K polynomial orthogonal for $\lambda d\nu$. To see why it has K simple interior zeros, multiply it by the product of its interior sign-change factors if there are fewer than K. That product has degree below K, but its product with the oracle has constant nonzero sign on the support, contradicting orthogonality. Positive definiteness rules out vanishing on the whole support. For a law on at most K points, use its normalized annihilator and pad the remaining roots inside $[a, b]$. The risk of this compact family is jointly continuous in the roots and the weakly varying law, so its minimum $R_K^\star(\nu)$ is continuous, including at singular Grams. It is concave in ν, being the infimum of linear risk functionals.

A.2 WHAT A SHALLOW COMPUTATION IDENTIFIES

For $p = 1 + \sum_{j=1}^{T} c_j \lambda^j$, its complete population risk is a quadratic polynomial in c. Its linear coefficients give $m_1, \dots, m_T$, and its quadratic coefficients give m_{i+j}, $1 \le i, j \le T$. Together with $m_0 = 1$, these identify exactly the moments through $2T$. No higher moment enters the objective.

For an individual task, a nondegenerate T-step Krylov record recovers $r_0, Ar_0, \dots, A^T r_0$. Its inner products give

$$m_{i+j} = \frac{(A^i r_0)^\top A^j r_0}{\|r_0\|^2}, \qquad 0 \le i, j \le T.$$

Thus the record determines that task's residual-weighted moments through $2T$, whereas the population-information model specifies the risk function over the task distribution.

A.3 THE ZERO-RESOLUTION DOMAIN

For a probability law on $[a, b] \subset (0, \infty)$, the following are equivalent:

$$R_K^\star(\nu) > 0; \qquad G_K(\nu) \succ 0; \qquad |\operatorname{supp} \nu| \ge K + 1.$$

A singular Gram gives a nonzero degree-at-most-K polynomial vanishing on the support, so the support has at most K points. Conversely, for a support of size $j \le K$, $\prod_{i=1}^{j}(1-\lambda/\lambda_i)$ has zero risk and equals one at zero. With at least $K+1$ points, the Gram is positive definite, and Equation (A.1.1) gives a positive value. On a compact family the Gram's least eigenvalue is continuous and positive, hence uniformly positive. The same is then true of $R_K^\star$. Thus at $\tau = 0$ positivity of every denominator already implies the required uniform regularity.

A.4 The affine set and its exact dimension

For $c^0 = c(x_0)$, holding this vector fixed in the normal equations gives

$$b(x) + H(x)c^0 = B(x - x_0), \qquad B_{i,h} = c^0_{h-i}.$$

Since $H(x) \succ 0$, the right side vanishes if and only if c^0 is the unique oracle at x. Differentiating at x_0 yields $Dc(x_0) = -H_0^{-1}B$.

The coefficient c^0_K is nonzero. If p_{c^0} had degree below K, $p_{c^0} + t\lambda p_{c^0}$ would be feasible and stationarity would imply $\int \lambda p^2_{c^0} d\nu = 0$. This contradicts $a > 0$, positive definiteness, and $p_{c^0}(0) = 1$.

If $d \le K$, take rows $i = K - d + j$ and columns $h = 2K - d + \ell$, $j, \ell = 1, \dots, d$. Their entries are $c^0_{K+\ell-j}$, a lower-triangular matrix with nonzero diagonal c^0_K. If $d \ge K$, take all rows and columns $h = K+1, \dots, 2K$, which are among the unresolved moments; the same triangular argument gives rank K. Therefore the rank is $\min(K, d)$. At an interior feasible point the affine section has dimension $d - \min(K, d) = \max(K - 2T, 0)$.

When $K \le 2T$, the unresolved moments can be reconstructed successively from the preferred program:

$$m_h = -\frac{1}{c^0_K} \sum_{j=0}^{K-1} c^0_j m_{h-K+j}, \qquad h = 2T+1, \dots, 2K.$$

Every moment on the right is already known or was recovered earlier.

Remark A.1 (Gauss-node view). The oracle program is the degree-K orthogonal polynomial of $\lambda\, d\nu(\lambda)$ normalized by $p(0) = 1$ (Appendix A.1). Its zeros are therefore the K nodes $\xi_1, \dots, \xi_K$ of Gauss quadrature for $\lambda\, d\nu$, and $p_\star = \prod_j (1 - \lambda/\xi_j)$. Quadrature exactness gives $m_h = \sum_j w_j \xi_j^{h-1}$ for $h = 1, \dots, 2K$ with positive weights w_j. Two interior moment vectors share the oracle program exactly when they share the nodes, and with the nodes fixed the moments are a rank-K Vandermonde image of the weights. The known moments $m_1, \dots, m_{2T}$ impose $\min\{2T, K\}$ independent linear constraints, which leave $\max\{K - 2T, 0\}$ free weights and recover the fiber dimension (Karlin and Studden, 1966; Kreĭn and Nudel'man, 1977; Dette and Studden, 1997). Writing $f = f(0) + \lambda g$ for $\deg f \le 2K$ and applying the rule to g gives $\int f\, d\nu = f(0)\big(1 - \sum_j w_j/\xi_j\big) + \sum_j (w_j/\xi_j) f(\xi_j)$: up to degree $2K$, ν agrees with the quadrature rule that adds a prescribed node at zero. Taking $f = p_\star^2$ gives $R_K^\star = 1 - \sum_j w_j/\xi_j$, the weight at the prescribed node, which is the Christoffel function of ν at zero (Golub and Meurant, 2010). The oracle risk is therefore affine in the free weights along a fiber; it is positive at interior points and approaches zero only as $\sum_j w_j/\xi_j \to 1$ at the boundary of the moment space.

A.5 Risk curvature and the enclosing ball

Let $D_0 = R^\star(x_0) + \tau$. The oracle map and risk are smooth on a fixed positive-definite neighborhood. Implicit differentiation of the normal equations and differentiation of the minimized quadratic give

$$\nabla^2 R^\star(x_0) = -2B^\top H_0^{-1} B, \qquad \mathsf{S} = -H_0^{-1/2} B / \sqrt{D_0}.$$

For $x = x_0 + \rho u$, $u \in U$, write a candidate as $c = c^0 + \rho\sqrt{D_0} H_0^{-1/2} y$. Uniform Taylor expansion of the exact completed square gives

$$\rho^{-2}\big(\ell_{x_0+\rho u,\tau}(c) - 1\big) = \|y - \mathsf{S}u\|^2 + O(\rho), \tag{A.5.1}$$

uniformly for bounded y and $u \in U$. It suffices to consider bounded y: the candidate c^0 has worst excess $O(\rho^2)$, and a uniform positive eigenvalue bound in the exact risk forces any minimizing c to

be $O(\rho)$ from the uniformly $O(\rho)$ oracle centers. Taking the maximum over u and minimum over y in Equation (A.5.1) proves Equation (8) with a uniform $O(\rho^3)$ remainder.

The equivalent mixture expansion is useful for sharpness. For any probability measure ω on U, Taylor-expand numerator and denominator around x_0:

$$\frac{R^\star(x_0+\rho\mathbb{E}_\omega u)+\tau}{\mathbb{E}_\omega R^\star(x_0+\rho u)+\tau} = 1+\rho^2\,\mathrm{tr}(\mathsf{S}^\top\mathsf{S}\ \mathrm{Cov}_\omega\, u)+O(\rho^3). \tag{A.5.2}$$

Lemma A.2 (Uniform mixture remainder)**.** *On the compact perturbation set U, the cubic remainder in the local risk expansion is bounded by $C\rho^3$ uniformly over every probability measure on U, including measures whose support and weights depend on ρ.*

Proof. Bounded third derivatives give a pointwise remainder at most $M_3(\sup_U \|u\|)^3\rho^3/6$. Integrating cannot enlarge this bound, and the denominator is uniformly separated from zero. □

For a compact set Z, the maximum variance of a distribution on Z equals the squared radius of its smallest enclosing ball (Lim and McCann, 2022; Pass, 2023). One direct proof is that every variance is at most $\max_{z\in Z}\|z-y\|^2$ for every y. At an enclosing-ball center y_*, the active squared-distance gradients have zero in their convex hull. Thus some probability weights on contact points have mean y_* and variance equal to the enclosing radius squared. Apply this to $Z=\mathsf{S}U$.

If $U=-U$, then the minimum ball of $\mathsf{S}U$ is centered at zero. The reference program c^0 therefore has the optimal leading ρ^2 price; its refitting gain is $O(\rho^3)$. In a shifted family the center can be nonzero, producing an $O(\rho)$ change in the optimal coefficients and a quadratic improvement.

A.6 An exact nontrivial same-program construction

The common cubic program in Figure 2 is

$$p_\star = 1-\frac{825}{398}\lambda+\frac{835}{597}\lambda^2-\frac{245}{796}\lambda^3. \tag{A.6.1}$$

Let P_j denote a Legendre polynomial. Define on $[1,2]$

$$h(\lambda)=\frac{1}{8078520514}\big[-26471025P_3(2\lambda-3)+173412225P_4(2\lambda-3) \\ -1077504120P_5(2\lambda-3)+6801133144P_6(2\lambda-3)\big].$$

The numerator's absolute coefficients sum to its denominator, so $|h|\le 1$. The densities $d\nu_\pm = (1\pm h/2)d\lambda$ are positive and normalized. They are obtained by solving the three homogeneous linear conditions $\int \lambda^i p_\star h\,d\lambda=0$, $i=1,2,3$, in the four-dimensional shifted-Legendre span. Orthogonality already imposes $\int\lambda^j h\,d\lambda=0$ for $j\le 2$; the nullspace supplies the direction, and normalization supplies positivity.

For $p_\star$ in Equation (A.6.1), exact integration gives

$$\int\lambda^i p_\star\,d\lambda=\int\lambda^i p_\star h\,d\lambda=0,\quad i=1,2,3,\qquad \int\lambda^3 h\,d\lambda=-\frac{756315}{32314082056}\neq 0.$$

Thus the higher moments and degree-three Grams differ while both strictly convex normal equations have the same solution. The oracle risks are

$$R_3^\star(\nu_-)=\frac{14489366453}{462996167698368},\qquad R_3^\star(\nu_+)=\frac{17824715603}{462996167698368}.$$

For $p_\alpha=p_\star+10^{-3}\alpha\lambda^3$ the plotted normalized risks equal

$$1+\frac{3675117131226057}{6339097823187500}\alpha^2,\qquad 1+\frac{734989406197107}{1559662615262500}\alpha^2.$$

A distribution label has no decision value in this family even though the oracle risk itself is distribution-dependent.

A.7 Coordinates and conditioning

For Figure 1, let H_s be the free degree-two Gram of ν_s and set $\bar{H} = \frac{1}{2}\{H_2/R_2^\star(\nu_2) + H_4/R_2^\star(\nu_4)\}$. The plotted coordinates are $z = Q\bar{H}^{1/2}(c - c^{\mathrm{mm}})$, where c^{mm} is the shared minimax and the orthogonal matrix Q places the two oracle centers on a horizontal line. The same transform is applied to both risks and the reset line.

Under $x = x_0 + N\xi$ with N invertible, the same physical perturbation set becomes $N^{-1}U$ and the decision map becomes $\mathsf{S}N$. Their image remains $(\mathsf{S}N)(N^{-1}U) = \mathsf{S}U$, so the local price is invariant. Defining a fresh unit ball after changing coordinates instead changes the uncertainty family. With shape matrix Σ, $\mathsf{S}\Sigma^{1/2}$ describes the quantitatively relevant directions. Rank identifies the directions that matter; the singular values of $\mathsf{S}\Sigma^{1/2}$ quantify how much.

For a smooth physical parameterization with fixed lower moments, $x(\theta_0 + \rho u) = x_0 + \rho Ju + O(\rho^2)$, the leading image is $\mathsf{S}JU$. Central symmetry of parameter perturbations gives a centrally symmetric leading image, while higher-order curvature can remain. Chebyshev density perturbations are useful controlled examples because orthogonality fixes the lower moments and $|T_j| \le 1$ gives a transparent positivity bound.

For an ellipsoidal family $U = \Sigma^{1/2}\{v : \|v\| \le 1\}$, $\mathsf{S}U$ is a centered ellipsoid, so the leading price is $\rho^2\sigma_{\max}(\mathsf{S}\Sigma^{1/2})^2$ and two laws suffice to leading order (Proposition B.1).

B Proofs: minimax continuation and sharp certificates

B.1 Proof of the certificate identity

We prove Theorem 2. Put $\zeta_\nu = R_K^\star(\nu) + \tau$. For $\tau > 0$ these are bounded below by τ; for $\tau = 0$ use Appendix A.3. One positive-definite law makes F coercive on the finite-dimensional free coefficient space. Continuity and compactness of $\mathcal{M}$ give an attained minimum c_*. The subdifferential rule for a maximum of differentiable convex functions gives

$$0 \in \mathrm{conv}\{\nabla \ell_\nu(c_*) : \ell_\nu(c_*) = V_\tau\}.$$

Carathéodory's theorem supplies at most $K + 1$ active laws and weights π_i such that

$$\sum_i \frac{\pi_i}{\zeta_i}(H_i c_* + b_i) = 0.$$

Define $C = \sum_i \pi_i/\zeta_i$ and $\omega_i = (\pi_i/\zeta_i)/C$. This is the normal equation for $\bar{\nu} = \sum_i \omega_i \nu_i$, so c_* minimizes $R_{\bar{\nu}}$, even if that quadratic is singular.

For any mixture and any feasible p,

$$\max_i \frac{R_{\nu_i}(p) + \tau}{\zeta_i} \ge \frac{\sum_i \omega_i (R_{\nu_i}(p) + \tau)}{\sum_i \omega_i \zeta_i} \ge \frac{R_K^\star(\bar{\nu}) + \tau}{\sum_i \omega_i R_K^\star(\nu_i) + \tau}.$$

For the active mixture, equality holds at c_* because each active normalized loss equals V_τ. This proves both inequalities in Equation (11).

If $V_\tau > 1$, an attaining barycenter cannot be singular: a singular barycenter has oracle risk zero and certificate value at most one. If $V_\tau = 1$, every minimax polynomial minimizes every constituent risk; select a singleton positive-definite law, whose unique oracle is that polynomial. Thus a full-rank attaining mixture always exists under the stated assumptions.

To obtain the second support bound, write $x \in \mathbb{R}^d$, $d = 2(K - T)$, and take the compact convex hull of points $(x, R^\star(x) + \tau)$. A mixture is represented by $(\bar{x}, \bar{\zeta})$ and has value $(R^\star(\bar{x}) + \tau)/\bar{\zeta}$. If this convex hull has affine dimension at most d, Carathéodory already gives $d + 1$ constituents. Otherwise it is full-dimensional in $\mathbb{R}^{d+1}$. A maximizer lies on its lower boundary: decreasing $\bar{\zeta}$ at fixed $\bar{x}$ increases its positive ratio. A supporting hyperplane exposes a face of dimension at most d. Every positive-weight constituent in a representation of the boundary point lies on that face, so Carathéodory gives $d + 1$ constituents. Compactness supplies a finite representation, and preserving $(\bar{x}, \bar{\zeta})$ preserves the ratio and the barycenter's moment Gram. When $V_\tau = 1$, the positive-definite singleton chosen above already satisfies either bound. Combining this with the $K + 1$ bound gives $\min\{K + 1, d + 1\}$.

B.2 WHY THE CERTIFICATE SUPPLIES A CONTINUATION

The degree-t oracle p_t obeys $\int \lambda^j p_t(\lambda)\, d\nu = 0, 1 \le j \le t$. Equivalently, it is the degree-t normalized orthogonal polynomial for the positive measure $\lambda d\nu$. The three-term recurrence for these polynomials, normalized by $p_t(0) = 1$, is

$$p_{t+1} = (1 + \gamma_t - \eta_t\lambda)p_t - \gamma_t p_{t-1}.$$

The corresponding vector update is $w_{t+1} = w_t - \eta_t(Aw_t - g) + \gamma_t(w_t - w_{t-1})$. This is the kernel-polynomial construction for a known law (Stiefel, 1958; Pedregosa and Scieur, 2020). The degree-t oracle is uniquely determined by the moments through $2t$, so any mixture of laws sharing moments through $2T$ has the identical oracle prefix for all $t \le T$. Applying the recurrence to the full-rank attaining mixture proves prefix preservation and attainment.

Fixed-coefficient full-history methods cannot produce more than degree-t residual polynomials after t updates, by induction. Conversely, choosing scalar GD prefixes gives a degree-ordered basis of residual polynomials; an affine combination of their residuals produces any degree-K polynomial with constant one. Hence the unrestricted final class is exactly $\mathcal{A}_K$. The three-term recurrence (10) started from $p_{-1} = p_0 = 1$ also generates all of $\mathcal{A}_K$ (Appendix F.5 gives an explicit construction); what Theorem 2 adds is attainment when the first T steps are fixed.

B.3 ABSOLUTE LOSS, FINITE LIBRARIES, AND FEASIBLE NUMERICAL OUTPUTS

The same argument with $\zeta_i = 1$ gives

$$\min_{p\in\mathcal{A}_K} \max_{\nu\in\mathcal{M}} R_\nu(p) = \max_{\bar\nu\in\mathcal{M}} R^\star_K(\bar\nu).$$

The attaining law's oracle has the same recurrence and preserves a common prefix when one exists. At $T = 0$ no common moments beyond mass are needed, which is the setting of the empirical ridge benchmark. For relative risk on a convex library hull, vertex checking suffices: the numerator is affine in a mixture and the oracle denominator is concave, so any mixture ratio is bounded by the largest vertex ratio.

In a finite library, let $\psi = (\psi_0, \dots, \psi_K)^\top$ be a degree-ordered polynomial basis, $z = \psi(0)$, and $G_i = \int \psi\psi^\top d\nu_i$. With $\zeta_i = R^\star_K(\nu_i) + \tau$, the finite-library dual is

$$V_\tau = \left[\min_{\pi\in\Delta:\, \widetilde{M}_\pi \succ 0} z^\top \left(\sum_i \pi_i \frac{G_i + \tau zz^\top}{\zeta_i}\right)^{-1} z\right]^{-1},$$

where $\widetilde{M}_\pi = \sum_i \pi_i(G_i + \tau zz^\top)/\zeta_i$. The inverse is used only on the displayed positive-definite domain; an attaining prior in this domain exists by the full-rank mixture argument above. Equivalently, singular boundary priors can be assigned their constrained quadratic Bayes value, without writing an ordinary inverse. This is the multiresponse c-optimal criterion with $c = z$ (Sagnol, 2011).

Procedure 1: finite-library synthesis. The following computation uses the Grams and normalizers just defined.

1. Form $\widetilde{G}_i = (G_i + \tau zz^\top)/\zeta_i$ and choose a probability vector π with $\widetilde{M}_\pi = \sum_i \pi_i \widetilde{G}_i \succ 0$.
2. Minimize $z^\top \widetilde{M}_\pi^{-1} z$ over the probability simplex. This is the c-optimal-design criterion with $c = z$, which has an SOCP formulation (Sagnol, 2011).
3. At any feasible prior, form $p_\pi(\lambda) = \psi(\lambda)^\top \widetilde{M}_\pi^{-1} z / (z^\top \widetilde{M}_\pi^{-1} z)$ and compute

   $$\underline{V} = (z^\top \widetilde{M}_\pi^{-1} z)^{-1}, \qquad \overline{V} = \max_i \ell_{\nu_i,\tau}(p_\pi).$$

   Weak duality makes $\underline{V}$ a lower bound; evaluating a feasible program makes $\overline{V}$ an upper bound. Stop when $\overline{V} - \underline{V} \le \varepsilon$.
4. Convert the normalized-risk prior to law weights $\omega_i = (\pi_i/\zeta_i)/\sum_j \pi_j/\zeta_j$ and use the oracle recurrence of $\sum_i \omega_i\nu_i$ to obtain the continuation.

The prior weights π_i and spectral mixture weights ω_i are generally different; the reweighting is required by relative risk. For a continuous family, an exchange solve repeatedly optimizes the current finite library and adds a law maximizing the current program's relative risk over the entire family. That maximization supplies the upper bound used by the same stopping rule. For absolute risk, set $\zeta_i = 1$ and omit $\tau z z^\top$.

An arbitrary feasible coefficient vector need not itself be a mixture oracle. It can nevertheless be replaced by one without increasing any candidate risk. Minimize

$$\min_{p \in \mathcal{A}_K} \max_i \{\ell_i(p) - \ell_i(p_0)\}.$$

The optimum is nonpositive since p_0 is feasible. Its active-gradient weights identify its optimizer as a mixture oracle, by the preceding normal-equation argument. This replacement preserves depth and every common optimal prefix. An approximate replacement with per-law excess at most ε adds at most ε to the reported upper bound.

B.4 A POSITIVE-DENSITY NEIGHBORHOOD WITH ARBITRARY UNRESOLVED-MOMENT DIRECTIONS

Let $d\nu_0 = d\lambda/(b-a)$ and let $G^{(2K)} = (\int \lambda^{i+j} d\nu_0)_{i,j=0}^{2K}$, a positive-definite matrix. Define dual polynomials

$$g_h(\lambda) = \sum_{j=0}^{2K} (G^{(2K)})^{-1}_{hj} \lambda^j, \quad \text{so that} \quad \int \lambda^i g_h \, d\nu_0 = \mathbf{1}_{i=h}.$$

For an unresolved vector u, set $f_u = \sum_{h=2T+1}^{2K} u_h g_h$. The density

$$d\nu_{\rho,u} = (1 + \rho f_u) d\nu_0$$

preserves all moments through $2T$ and has unresolved moment vector $x_0 + \rho u$. On compact U, $C_U = \max_{u \in U, \lambda \in [a,b]} |f_u(\lambda)| < \infty$. Taking $\rho C_U < 1/2$ gives smooth densities bounded below by half the uniform density. Every associated Gram is then uniformly positive definite. Hence every $x_0 + \rho u$, $u \in U$, is realized by a smooth positive spectral density.

B.5 PROOF OF SHARPNESS

The image of S has dimension r. Choose a unit regular simplex $v_0, \ldots, v_r$ in this image:

$$\|v_i\| = 1, \qquad v_i^\top v_j = -1/r \ (i \neq j), \qquad \sum_i v_i = 0.$$

Choose preimages u_i with $\mathsf{S} u_i = v_i$ and let $U = \mathrm{conv}\{u_i\}$. The preceding density construction makes $x_0 + \rho U$ a valid compact convex spectral family at small ρ.

The maximum variance on this simplex is one, attained by uniform vertex weights. If a distribution has at most q support points in the simplex, its variance is at most

$$\frac{(r+1)(q-1)}{rq} = 1 - \frac{r+1-q}{rq}.$$

Here support points may lie in the interior. To prove the bound, fix their weights. Variance, as a function of any one support point with the others fixed, is convex, with quadratic coefficient $\omega_i(1-\omega_i) \geq 0$. Its maximum over the simplex can therefore be achieved at a vertex. Repeat for all support points, then merge repeated vertices. For weights $\beta_1, \ldots, \beta_q$ on distinct vertices,

$$\mathrm{Var}(v) = 1 - \left\| \sum_i \beta_i v_i \right\|^2 = \frac{r+1}{r} \left(1 - \sum_i \beta_i^2 \right),$$

maximized by equal weights. This proves the bound, with equality for a uniform q-vertex sub-simplex.

Lemma A.2 permits maximizing Equation (A.5.2) over all certificates and, separately, over certificates of size at most q, yielding Equation (12). For $q \leq r$, its positive quadratic coefficient dominates the uniform cubic remainder at sufficiently small ρ. Thus no such certificate attains the exact optimum. The support upper bound in Theorem 2 supplies $r+1$ constituents, for every (T, K).

Proposition B.1 (Contact number). *Under the assumptions of Theorem 1(b), let $\mathcal{M}_\rho$ contain the laws on $[a,b]$ with the known moments and unresolved moments in $x_0+\rho U$. Let z and R be the center and radius of the smallest ball containing $\mathsf{S}U$, and let $k_\star$ be the least number of points of $\mathsf{S}U$ at distance R from z whose convex hull contains z. For sufficiently small ρ, every maximizing certificate of $\mathcal{M}_\rho$ has at least $k_\star$ constituents, and some certificate with $k_\star$ constituents has value $V_\tau(\mathcal{M}_\rho)-O(\rho^3)$. If $R>0$, then $2\le k_\star\le \dim\operatorname{aff}\mathsf{S}U+1\le r+1$; $k_\star=2$ when U is centrally symmetric, and $k_\star=r+1$ for the regular simplex of Theorem 3.*

Proof. By (A.5.2) and Lemma A.2, a certificate whose constituents have unresolved moments $x_0+\rho u_i$ with weights ω_i has value $1+\rho^2\operatorname{Var}_\omega(\mathsf{S}u)+O(\rho^3)$, uniformly. For any distribution of Y on $\mathsf{S}U$ with mean $\bar y$, $\operatorname{Var}(Y)=\mathbb{E}\|Y-\bar y\|^2\le\mathbb{E}\|Y-z\|^2\le R^2$, with equality exactly when $\bar y=z$ and $\|Y-z\|=R$ almost surely. The center of the smallest enclosing ball lies in the convex hull of the contact points, so by Carathéodory's theorem some $k_\star\le\dim\operatorname{aff}\mathsf{S}U+1$ of them carry a distribution with mean z and variance R^2. Hence $V_\tau(\mathcal{M}_\rho)=1+\rho^2R^2+O(\rho^3)$, as in Theorem 1(b), and a $k_\star$-point certificate attains it up to $O(\rho^3)$. Distributions on at most $q<k_\star$ points form a compact set on which the variance is continuous and, by the equality case, strictly below R^2; its maximum $v_q<R^2$ gives $V_\tau(\mathcal{M}_\rho)-J_{\le q}(\mathcal{M}_\rho)\ge\rho^2(R^2-v_q)-C\rho^3>0$ for small ρ. If $R>0$, a single point has variance zero, so $k_\star\ge2$. For centrally symmetric U with $R>0$, a farthest point y of $\mathsf{S}U$ and its reflection $2z-y$ with weights $\frac12$ give $k_\star=2$. For the regular simplex the contact points are the $r+1$ vertices, and $z=0$ lies in the convex hull of no proper subset, so $k_\star=r+1$. □

C Proofs: observations, retained computation, and finite data

C.1 The optimal rule after a noisy record

Let $s=1,\dots,m$ index laws with positive-definite degree-K Grams and let $f_s>0$ be record densities with respect to a common measure. Calibration and deployment are independent given s. Put $\zeta_s=R_K^\star(\nu_s)+\tau$. For a prior π, define

$$a_\pi(y)=\sum_s\frac{\pi_sf_s(y)}{\zeta_s},\qquad \bar\nu_{\pi,y}=\frac{\sum_s\pi_sf_s(y)\nu_s/\zeta_s}{a_\pi(y)}.$$

Conditional minimization separates by record, so

$$h(\pi)=\inf_{p_\bullet}\sum_s\pi_s\mathbb{E}_s\ell_s(p_Y)=\int a_\pi(y)[R_K^\star(\bar\nu_{\pi,y})+\tau]\,dy. \tag{C.1.1}$$

Its unique action at each positive-probability record is the mixture oracle.

Here minimax interchange and attainment can be justified without assuming a finite observation alphabet. First restrict coefficients to a sufficiently large common ball. In free coordinates each per-law risk is a positive-definite quadratic. Along a ray outside a large ball its radial derivative is positive simultaneously for all laws, so radial clipping decreases every conditional loss. With reference measure $m^{-1}\sum_s f_s(y)dy$, bounded measurable actions form a closed convex bounded subset of a Hilbert L^2 space. It is weakly compact; the per-law risk functionals are convex, continuous, and weakly lower semicontinuous. The prior simplex is compact. Convex minimax therefore gives $V(f)=\max_\pi h(\pi)$, with an attained action (Sion, 1958).

The unique conditional minimizers lie in the same bounded ball. Differentiation of Equation (C.1.1) along the simplex is dominated by a constant multiple of $\sum_s f_s$, so its derivative in direction $e_s-\pi$ equals the posterior rule's risk under law s minus $h(\pi)$. At a maximizing prior every such derivative is nonpositive. Its rule consequently has all per-law risks at most $h(\pi)$ and is minimax, including at a boundary prior. Reweighting of the ordinary posterior by $1/\zeta_s$ is essential for relative risk; absolute risk uses $\zeta_s=1$. Common lower moments make every posterior mixture preserve the common optimal prefix.

C.2 Value of a refined observation

For a fixed prior, the conditional quadratic loss after observing Y has Hessian $2W_Y$ and minimum at c_Y. If Z is a coarsening, c_Z is feasible after seeing Y. Conditional completion of the square and the

tower property give

$$h_Z(\pi) - h_Y(\pi) = \mathbb{E}_\pi[(c_Z - c_Y)^\top W_Y (c_Z - c_Y)].$$

This is the squared-loss specialization of the decision-theoretic comparison of experiments (Blackwell, 1951; Raiffa and Schlaifer, 1961; Howard, 1966; Reid and Williamson, 2011). It refines the ordering $h_Y \le h_Z$ into the amount saved by moving the decision, not by recovering a label. For a stochastic garbling, adjoin its independent randomization; convexity removes unnecessary decision randomization.

Write this quantity as $I_\pi(Y;Z)$. If π_Y^*, π_Z^* maximize the two Bayes objectives, evaluating each objective at the other's maximizing prior gives

$$I_{\pi_Y^*}(Y;Z) \le V(Z) - V(Y) \le I_{\pi_Z^*}(Y;Z).$$

Equality of minimax values holds precisely when some coarse least-favorable prior has zero decision movement: the forward implication uses a maximizer of h_Y, and the reverse uses a zero-movement maximizer of h_Z.

C.3 A LEARNED PAIR AND ITS OPTIMAL FINAL STEP

Let (q_{T-1}, q_T) be the residual polynomials produced by a trained solver after $T-1$ and T steps.

Proposition C.1 (Optimal final step from a learned pair)**.** *Optimizing the retained pair over (η, γ) is a convex quadratic minimax problem. At $K = T+1$, with the same distributions and loss normalization,*

$$V_\tau \le V_{\mathrm{pair}} \le V_{\mathrm{reset}},$$

where reset fixes $\gamma = 0$. If the pair is within $L^2(\nu)$ errors δ_T, δ_{T-1} of an attaining reference sequence uniformly in ν, then

$$V_{\mathrm{pair}} \le \left(\sqrt{V_\tau} + e_*/\sqrt{\zeta_{\min}}\right)^2, \quad e_* = \max_{\lambda \in [a,b]} |1 + \gamma_* - \eta_* \lambda|\, \delta_T + |\gamma_*| \delta_{T-1}. \tag{C.3.1}$$

Here $\zeta_{\min} = \inf_\nu (R_K^\star(\nu) + \tau)$; for absolute risk use $\zeta_{\min} = 1$.

Fix $q_T(0) = q_{T-1}(0) = 1$. With

$$v(\lambda) = (-\lambda q_T,\ q_T - q_{T-1})^\top, \quad \vartheta = (\eta, \gamma)^\top,$$

we have

$$R_\nu(q_T + v^\top \vartheta) = a_\nu + 2 b_\nu^\top \vartheta + \vartheta^\top H_\nu \vartheta, \quad H_\nu = \int v v^\top d\nu \succeq 0.$$

The maximum of the normalized quadratics is convex. If the direction polynomials are independent and the coefficient Grams are uniformly positive definite, this objective is coercive and strictly convex; otherwise work on the quotient by the common null directions. Reset is the subfamily $\gamma = 0$ and unrestricted $\mathcal{A}_{T+1}$ is a relaxation, proving $V_\tau \le V_{\mathrm{pair}} \le V_{\mathrm{reset}}$.

For Equation (C.3.1), let (p_{T-1}^*, p_T^*) be prefixes of any attaining reference program, with final coefficients (η_*, γ_*). The difference produced by those coefficients from the actual pair is

$$(1 + \gamma_* - \eta_* \lambda)(q_T - p_T^*) - \gamma_* (q_{T-1} - p_{T-1}^*).$$

Its $L^2(\nu)$ norm is at most e_*. Minkowski, applied to the residual and a constant coordinate $\sqrt{\tau}$, bounds the normalized risk by $(\sqrt{V_\tau} + e_*/\sqrt{\zeta_{\min}})^2$. Reoptimizing the pair only improves this feasible action. For absolute loss the denominator is one; for an evaluated reference upper bound $\overline{V}$ replace V_τ by $\overline{V}$. The bound needs only an attaining reference sequence, so it also applies to the ridge audit without a common optimal prefix.

A sharper computable error keeps cancellation between the two state errors:

$$e_{\mathrm{direct}} = \sup_{\nu \in \mathcal{M}} \left\| (1 + \gamma_* - \eta_* \lambda)(q_T - p_T^*) - \gamma_* (q_{T-1} - p_{T-1}^*) \right\|_{L^2(\nu)}.$$

The same bound holds with e_* replaced by $e_{\mathrm{direct}} \le e_*$. Finite-library Grams evaluate this quantity without separately maximizing the two error terms.

For an exact common optimal prefix, the next mixture-oracle polynomial is not divisible by the current one: consecutive orthogonal polynomials have no common zero. Thus a reset class $q_T \mathcal{A}_1$ can have strictly larger optimum.

Which pairs reach the next oracle. Let $T \geq 1$ and let q have degree at most $T-1$, and let $p^{\star}_{T+1} = (1+\gamma_T - \eta_T\lambda)p^{\mathrm{sh}}_T - \gamma_T p^{\mathrm{sh}}_{T-1}$ be the three-term relation of a law with the common moments. If also $p^{\star}_{T+1} = (1+\gamma-\eta\lambda)p^{\mathrm{sh}}_T - \gamma q$, subtracting gives $[(\gamma-\gamma_T)-(\eta-\eta_T)\lambda]\,p^{\mathrm{sh}}_T = \gamma q - \gamma_T p^{\mathrm{sh}}_{T-1}$. The right side has degree at most $T-1$, so both sides vanish: $\gamma=\gamma_T$, $\eta=\eta_T$, and $\gamma_T(q-p^{\mathrm{sh}}_{T-1})=0$. Consecutive orthogonal polynomials have no common zero, so $\gamma_T \neq 0$ and $q = p^{\mathrm{sh}}_{T-1}$.

D Controlled spectral checks

These checks use exact constructions or finite spectral inputs, with $\tau = 0$ for oracle-relative risks.

D.1 Smooth realizations of sharpness and uncertainty shape

The primary shape family is supported on $[1,9]$ with $\lambda = 5+4x$ and Chebyshev reference measure ν_0, with T_j denoting the degree-j Chebyshev polynomial. Its density ratio is

$$1 + \theta_9 \mathsf{T}_9(x) + \theta_{10}\mathsf{T}_{10}(x), \qquad \theta = \rho v, \quad \rho \in \{0.05, 0.20, 0.50, 0.85\}.$$

Write v for a parameter vector. The three uncertainty sets are the convex hulls of

$$\begin{aligned} V_{\mathrm{sym}} &= \{(-0.5,-0.5),(-0.5,0.5),(0.5,-0.5),(0.5,0.5)\},\\ V_{\mathrm{shifted}} &= \{(0,0),(1,0),(0,0.45)\},\\ V_{\mathrm{asym}} &= \{(-0.2,0),(0.85,0.15),(0.15,0.35)\}. \end{aligned}$$

The last two have nonzero centers in the risk metric. All three preserve moments through degree eight. Their density is at least $1-\rho > 0$. At depth $K=5$ and prefix depth $T=4$, relative risk with $\tau = 0$ is evaluated over the full convex hull. The five sharpness constructions use $(T,K) = (1,2),(2,3),(2,4),(1,4),(4,5)$ and check the vertex-deleted value against the $1/r^2$ prediction of Theorem 3.

Figures 4 and 5 show the retained finite-construction values and their respective limiting predictions.

D.2 Moment-preserving observation controls

For each of the five diabetes training splits used for this construction (seeds 101–105), pool the three training regimes into a base measure ν_0. A weighted QR factorization of the Chebyshev polynomials $\mathsf{T}_0,\ldots,\mathsf{T}_{2K}$ produces orthonormal density coordinates f_j. In a free polynomial basis ϕ with $\phi(0)=0$, define

$$H = \int \phi\phi^{\top} d\nu_0 = LL^{\top}, \qquad D_{ij} = \int \phi_i p_0 f_j \, d\nu_0, \qquad S_{\mathrm{QR}} = -L^{-1}D/\sqrt{R^{\star}_K(\nu_0)}.$$

Only the coordinates $f_{2T+1},\ldots,f_{2K}$ enter D and S_{QR}, so every perturbation preserves the moments through degree $2T$. At $T=1, K=3$, choose a leading right-singular direction and a right-null direction; both have unit $L^2(\nu_0)$ norm and are divided by a common factor that makes the larger of their sup-norms on the support of ν_0 equal to one. Form $d\nu_s = (1+s\rho f)d\nu_0$. The density ratio is then at least $1-\rho$, which is 0.2 at $\rho = 0.8$, and the largest lower-moment error over all 60 records is 3.5×10^{-18}. Both binary channels flip the regime label with probability 0.2. The decision-sensitive gain agrees with $0.36\rho^2\|S_{\mathrm{QR}}v\|^2$ to relative error at most 1.7×10^{-5} at $\rho=0.1$; null-direction gains are at most 1.2×10^{-13}. All 25 tested ranks equal $\min\{K, 2(K-T)\}$.

D.3 Finite-shift identity and the local range

For two laws with training Gram H_0, target Gram H_1, and target forcing g, completing the square gives the exact refitting gain $\Delta = g^{\top}H_1^{-1}g$ and the frozen-metric prediction $\widehat{\Delta} = g^{\top}H_0^{-1}g$. Writing $\mathcal{H} = H_0^{-1/2}H_1H_0^{-1/2}$ yields the bracket (E.5.1). On three Gaussian paths and 30 points per path, all $90/90$ finite-law evaluations satisfy that bracket.

We independently recomputed the retained mixture path using weighted-QR coordinates and 80-digit monomial moments. Let ρ mix the base law with the fixed top-doubled endpoint law. The local

Table 1: Independent mixture-path recomputation. The first row is the original grid point; the smaller radii, added after the original grid, show the approach to the local regime.

Mixture weight ρ	Excess $V_0 - 1$	Quadratic prediction	$\|\mathcal{H}_\rho - I\|_2$
9.99×10^{-4}	1.31	2.56×10^{5}	837
1.00×10^{-8}	2.52×10^{-5}	2.57×10^{-5}	0.00838
1.00×10^{-9}	2.56×10^{-7}	2.57×10^{-7}	8.38×10^{-4}
1.00×10^{-10}	2.57×10^{-9}	2.57×10^{-9}	8.38×10^{-5}
1.00×10^{-11}	2.57×10^{-11}	2.57×10^{-11}	8.38×10^{-6}
1.00×10^{-12}	2.57×10^{-13}	2.57×10^{-13}	8.38×10^{-7}

prediction is $\rho^2 g'^\top H_0^{-1} g'/(4R_0^\star)$. At the original first point, $\rho = 0.001/1.001$, the minimax excess is 1.31 while the prediction is 2.56×10^5. The QR and high-precision prediction agree to relative error 2.06×10^{-15}: the large value is not a monomial-conditioning artifact for these inputs.

The base free Gram has condition number 2.48×10^9 in monomial coordinates, 2.95×10^3 in the scaled-Chebyshev basis, and numerical identity after QR whitening. Yet the first mixture point already has $\|\mathcal{H}_\rho - I\|_2 = 837$. A small probability mass placed far outside the base support therefore makes a large risk-metric perturbation (Table 1). The endpoint support reaches 9.19 versus the base maximum 4.89. This is the finite-law counterpart of the above-support risk mechanism in Section 6: mass size alone does not control the metric needed by the local theorem.

Across all 30 original mixture points, high-precision two-law prior optimization closes the numerical primal–dual gap below 10^{-50}. At $\rho = 10^{-12}$ the prediction/price ratio is 1.00000187, recovering the local expansion. Regenerating the retained endpoints from their Gaussian draws changes individual atoms by at most 1.25×10^{-14}.

E Objective choice, observation value, and spectral regularization

E.1 Relative versus absolute risk

The oracle map $c(x)$, its affine fibers, and its derivative rank do not depend on how different regimes are normalized. The minimax price does. In particular, the symmetry and sharp certificate statements for oracle-relative risk cannot be transferred to raw absolute risk by rescaling S.

For the *additive oracle regret* objective

$$A(\rho) = \min_c \max_{u \in U} \{R_{x_0+\rho u}(c) - R^\star(x_0 + \rho u)\},$$

the same completed-square proof gives

$$A(\rho) = \rho^2 \operatorname{rad}(\mathsf{S}_{\mathrm{add}} U)^2 + O(\rho^3), \qquad \mathsf{S}_{\mathrm{add}} = -H_0^{-1/2} B.$$

The corresponding mixture certificate is the Jensen gap $R^\star(\sum_i \omega_i x_i) - \sum_i \omega_i R^\star(x_i)$. Raw absolute risk retains the variation of the oracle baseline $R^\star(x)$, including its first-order term.

A symmetric absolute-risk counterexample. Let $T = 1, K = 2$, let ν_0 be uniform on $[1, 2]$, and let $h(\lambda) = P_4(2\lambda - 3)$, where P_4 is the Legendre polynomial. The densities $1 \pm \rho h$, $0 < \rho < 1$, are positive and have equal moments through degree three, hence the required common prefix. Orthogonality and direct integration give

$$\int_1^2 \lambda^j h(\lambda) d\lambda = 0 \ (0 \le j < 4), \qquad \int_1^2 \lambda^4 h(\lambda) d\lambda = \frac{1}{630}.$$

For $p = 1 + c_1\lambda + c_2\lambda^2$,

$$R_{\nu_+}(p) - R_{\nu_-}(p) = \frac{\rho c_2^2}{315} \ge 0.$$

Thus the absolute minimax polynomial on the symmetric family $\mathrm{conv}\{\nu_-, \nu_+\}$ is the oracle of ν_+. Its coefficients move at first order, because

$$\left.\frac{dc_+}{d\rho}\right|_0 = -H_0^{-1}\begin{pmatrix}0\\ c_2(0)/630\end{pmatrix} \neq 0.$$

The leading oracle coefficient is nonzero by the argument in Appendix A. Therefore central symmetry does not eliminate the leading absolute-risk refitting gain.

There is also a certificate distinction. For raw absolute risk on a compact convex family,

$$\min_{p\in\mathcal{A}_K}\max_{\nu\in\mathcal{M}} R_\nu(p) = \max_{\bar\nu\in\mathcal{M}} R_K^\star(\bar\nu).$$

A single maximizing law $\bar\nu$ is already a certificate when interior laws are admissible constituents. Representing that law by *library vertices* may require a mixture, but the relative-risk lower bound of $r+1$ arbitrary constituents is not an absolute-risk lower bound.

E.2 A BINARY OBSERVATION'S LOCAL VALUE

Consider two regimes with moments $x_s = x_0 + s\rho u$, $s \in \{-1, +1\}$, and a binary symmetric channel with label-flip probability $e \in (0, 1/2)$. Put $z = \mathsf{S}u$. Completing the square uniformly over the two regimes reduces the local relative risk to

$$\ell_s(c^0 + \rho\sqrt{R_0^\star + \tau}H_0^{-1/2}y) = 1 + \rho^2\|y - sz\|^2 + O(\rho^3).$$

For prior $\pi = \Pr(s = +1)$, the Bayes-optimal local action after a record is the posterior mean of sz. Its squared-error risk is the posterior variance times $\|z\|^2$. If $q = e + (1-2e)\pi$ is the probability of the positive record, then

$$\mathbb{E}\,\mathrm{Var}(s \mid Y) = \frac{4\pi(1-\pi)e(1-e)}{q(1-q)}.$$

Writing $t = \pi(1-\pi)$ gives $q(1-q) = e(1-e) + (1-2e)^2 t$, so this ratio is maximized at $t = 1/4$, or $\pi = 1/2$. Consequently

$$V_e = 1 + 4e(1-e)\rho^2\|z\|^2 + O(\rho^3), \qquad V_{\mathrm{no}} = 1 + \rho^2\|z\|^2 + O(\rho^3),$$

which proves Equation (14). The remainders are uniform over priors: each record has probability at least e, the posterior weighted Grams have a common positive-definite bound, and the same bounded-action argument as in Appendix A applies. Hence maximizing the expansion over π is valid. At $e = 0$ the risk is exactly one; at $e = 1/2$ the record has no value.

If $Bu = 0$ and the family remains in the same affine fiber, one polynomial is optimal in both regimes. It achieves relative loss one with or without observations, so every observation has exactly zero value, independently of the local expansion. Conversely, a symmetric family can have $Bu \neq 0$: the unobserved compromise may stay at the center to first order even though revealing the sign is useful.

E.3 UNIFORM BACKGROUND MASS AND WHAT IT CONTROLS

Let U be the probability measure uniform on $[a, b]$. In its orthonormal Legendre basis, $p(\lambda) = \sum_{j=0}^K \alpha_j\sqrt{2j+1}P_j((2\lambda - a - b)/(b-a))$. Since $|P_j(t)| \le 1$ on $[-1, 1]$, Cauchy–Schwarz yields

$$|p(\lambda)|^2 \le \Big(\sum_{j=0}^K \alpha_j^2\Big)\Big(\sum_{j=0}^K (2j+1)\Big) = (K+1)^2 R_U(p).$$

Thus the uniform component penalizes large values anywhere in the known interval, including regions with no training atoms. For the dimensionless penalty

$$\max_s R_{\widehat{v}_s}(p)/\widehat{V} + \alpha R_U(p)/R_U(p_0),$$

if its value is at most v and $\alpha > 0$, every probability law ξ on that interval satisfies $R_\xi(p) \le (K+1)^2 v R_U(p_0)/\alpha$. The large value of $R_U(p_0)$ makes this generic bound loose in the ridge study; the direct interval guard in Appendix E.6 avoids that conversion.

Uniform ϵ-smoothing regularizes the family; Huber-type contamination (Huber, 1964) instead adds the worst-law term $\epsilon\|p\|_\infty^2$. Both that penalty and the max-form interval guard defined below admit a worst-atom oracle.

E.4 Relations to spectral Wasserstein balls

Here W_1 denotes Wasserstein distance between *one-dimensional spectral laws*, with transport cost $|\lambda - \lambda'|$. This is distinct from the lifted trajectory distance used by Park et al. (2025) and Ranjan et al. (2026). For laws on the common positive interval $[a, b]$, every coupling and the mean-value theorem give

$$|m_j(\nu) - m_j(\mu)| \le j b^{j-1} W_1(\nu, \mu), \qquad j \ge 1.$$

A spectral W_1 ball of radius δ is therefore contained in these moment *slabs*. Except in degenerate cases it is not contained in a slice imposing exact moment equalities. Conversely, if $W_1(\nu_i, \mu) \le \delta$ for every library vertex, mixing their couplings proves

$$\operatorname{conv}\{\nu_i\} \subseteq \{\nu : W_1(\nu, \mu) \le \delta\}.$$

A finite set of exact moment equalities alone does not imply membership in an arbitrarily small prescribed Wasserstein ball: distinct positive densities obtained by an orthogonal high-degree perturbation share those moments but have positive W_1 distance. Quantitative containment therefore requires radii or additional assumptions.

E.5 A finite-shift diagnostic for the local oracle map

Let ν_0, ν_1 have positive-definite free-coefficient Grams H_0, H_1, and let $p_0 = 1 + \phi^\top c_0$ minimize R_{ν_0}. For the coordinate decomposition, assume also that the degree-$2K$ moment Gram of ν_0 is positive definite, so weighted orthogonalization through degree $2K$ is defined. No common-moment condition is imposed in this subsection. Fix a training-orthonormal polynomial basis $f_0 = 1, f_1, \dots, f_{2K}$, with $\int f_j \, d\nu_0 = 0$ for $j \ge 1$, and define

$$\delta_j = \int f_j \, d(\nu_1 - \nu_0), \qquad D_{ij} = \int \phi_i p_0 f_j \, d\nu_0, \qquad g = \int \phi p_0 \, d\nu_1.$$

Normal-equation orthogonality gives $g = D\delta$. If $H_0 = L_0 L_0^\top$, set $S_{\mathrm{all}} = -L_0^{-1} D$ and write $\delta = \delta_\parallel + \delta_\perp$ with $\delta_\perp \in \ker S_{\mathrm{all}}$. The frozen local predictor is

$$\widehat{\Delta} = \|S_{\mathrm{all}} \delta_\parallel\|^2 = g^\top H_0^{-1} g.$$

The map D has rank K. Indeed, $v^\top D = 0$ makes $(v^\top \phi) p_0$ orthogonal to $f_1, \dots, f_{2K}$; its mean is zero by the normal equations. Positive definiteness of the degree-$2K$ moment Gram then makes it the zero polynomial. Since $p_0 \ne 0$ and the free basis is independent, $v = 0$. Thus $\ker S_{\mathrm{all}}$ has dimension K in $\mathbb{R}^{2K}$. If U is uniform on the unit sphere of the $L^2(\nu_0)$-orthonormal coordinates $f_1, \dots, f_{2K}$ and $P_\perp$ is the orthogonal null projector, then

$$\mathbb{E}\|P_\perp U\|^2 = \operatorname{tr}(P_\perp \mathbb{E}[UU^\top]) = \frac{K}{2K} = \frac{1}{2}.$$

If an entire feasible shift lies in the nullspace, p_0 remains exactly optimal; in a mixed shift, the null component leaves g unchanged but may alter H_1, hence the final coefficients and gain.

Completing the target-law square gives $c_1 = c_0 - H_1^{-1} g$ and

$$\Delta = R_{\nu_1}(p_0) - R_K^\star(\nu_1) = g^\top H_1^{-1} g.$$

Writing $\mathcal{H} = L_0^{-1} H_1 L_0^{-\top}$ gives

$$\frac{\widehat{\Delta}}{\lambda_{\max}(\mathcal{H})} \le \Delta \le \frac{\widehat{\Delta}}{\lambda_{\min}(\mathcal{H})}, \tag{E.5.1}$$

since its eigenvalues agree with the symmetrically whitened matrix. For $\Delta > 0$, the same eigenvalue bounds give, without a small-shift condition,

$$\lambda_{\min}(\mathcal{H}) \le \widehat{\Delta}/\Delta \le \lambda_{\max}(\mathcal{H}), \qquad |\widehat{\Delta}/\Delta - 1| \le \|\mathcal{H} - I\|_2.$$

If $g = 0$, both gains are zero and the ratio is left undefined. The additional condition $\|\mathcal{H} - I\| \le \varepsilon < 1$ gives

$$\frac{\widehat{\Delta}}{1+\varepsilon} \le \Delta \le \frac{\widehat{\Delta}}{1-\varepsilon}.$$

For a local shift of size ρ, $g = O(\rho)$ and $H_1 - H_0 = O(\rho)$, recovering the additive $O(\rho^3)$ expansion. Equation (E.5.1) measures the finite-shift range of the local metric; evaluating the exact identity with H_1 has the complexity of a target normal-equation solve.

Proposition E.1 (Non-asymptotic price bracket)**.** *Let X be a set of unresolved-moment vectors on which $H(x) \succ 0$ and $\zeta(x) = R_K^\star(x) + \tau > 0$, and let $\zeta_0 = \zeta(x_0)$. Suppose $(1-\varepsilon)H_0 \preceq H(x) \preceq (1+\varepsilon)H_0$ and $|\zeta(x)/\zeta_0 - 1| \le \varepsilon_\zeta$ on X, with $\varepsilon, \varepsilon_\zeta < 1$. With S as in Theorem 1(b), let R_{S} be the radius of the smallest ball containing $\{\mathsf{S}(x - x_0) : x \in X\}$ and $M = \sup_{x\in X} \|\mathsf{S}(x - x_0)\|$. Then*

$$\sqrt{\tfrac{1-\varepsilon}{1+\varepsilon_\zeta}}\Big(R_{\mathsf{S}} - \tfrac{\varepsilon M}{1-\varepsilon}\Big)_+ \le \sqrt{V_\tau(X) - 1} \le \sqrt{\tfrac{1+\varepsilon}{1-\varepsilon_\zeta}}\Big(R_{\mathsf{S}} + \tfrac{\varepsilon M}{1-\varepsilon}\Big).$$

Proof. Write $y = H_0^{1/2}(c - c_0)/\sqrt{\zeta_0}$ and $s(x) = H_0^{1/2}(c(x) - c_0)/\sqrt{\zeta_0}$. By (6), the relative risk of p_c at x exceeds one by $\|c - c(x)\|^2_{H(x)}/\zeta(x)$, which lies between $(1-\varepsilon)/(1+\varepsilon_\zeta)$ and $(1+\varepsilon)/(1-\varepsilon_\zeta)$ times $\|y - s(x)\|^2$. Taking the supremum over x and the infimum over y bounds $V_\tau(X) - 1$ by the same factors times the squared radius of $\{s(x)\}$. The identity $c(x) - c_0 = -H(x)^{-1}B(x - x_0)$ of Appendix A.4 gives $s(x) = \mathcal{H}(x)^{-1}\mathsf{S}(x - x_0)$ with $\mathcal{H}(x) = H_0^{-1/2}H(x)H_0^{-1/2}$ and $\|\mathcal{H}(x)^{-1} - I\|_2 \le \varepsilon/(1-\varepsilon)$, so the two sets are within Hausdorff distance $\varepsilon M/(1-\varepsilon)$. The radius of the smallest enclosing ball is 1-Lipschitz in Hausdorff distance. □

For a polytope X both constants follow from vertex data: $\mathcal{H}(x) - I$ is affine in x, and $R_K^\star$ is concave with gradient $\partial R_K^\star/\partial m_h(x_0) = [p_{c_0}^2]_h$, the coefficient of λ^h in $p_{c_0}^2$, so ζ is bounded below at the vertices and above by its tangent plane at x_0. As X shrinks to x_0, $\varepsilon, \varepsilon_\zeta$ and M are $O(\rho)$ and the bracket recovers (8). The denominator term is the binding one when $R_K^\star(x_0) + \tau$ is small, since $\varepsilon_\zeta \approx |\nabla R_K^\star \cdot (x - x_0)|/(R_K^\star(x_0) + \tau)$; a larger offset τ shrinks it. A family parallel to the fiber through x_0, lying in $x' + \ker B$ with $B(x' - x_0) \neq 0$, has $R_{\mathsf{S}} = 0$ but generally a positive price, at most of order $(\varepsilon M)^2$; both (8) and the bracket depend on the base point, and expanding at a point of the family is sharper.

E.6 An interval guard and its computable lower bound

For $\kappa \ge 1$, define the convex interval-guard problem

$$\min_{p,t} t \quad \text{subject to} \quad p(0) = 1, \quad R_{\widehat{\nu}_s}(p) \le t \;\forall s, \quad p(\lambda)^2 \le \kappa t \;\forall \lambda \in [a, b].$$

The dimensionless parameter κ bounds interval amplification relative to the optimized epigraph level t. At $\kappa = 1$, every empirical constraint is dominated by the interval constraint, giving classical interval-Chebyshev design. Let p_0 be an empirical minimizer with value $\widehat{V} > 0$. For any $\kappa \ge \|p_0\|_\infty^2/\widehat{V}$, it is feasible at $t = \widehat{V}$, so the guard recovers an empirical minimizer exactly. For any feasible pair, every probability law supported on the interval has risk at most κt.

In a polynomial basis ψ, interval points contribute rank-one matrices $\psi(\lambda)\psi(\lambda)^\top/\kappa$ alongside empirical Grams. Any probability mixture M of these matrices with $M \succ 0$ gives the lower bound $(z^\top M^{-1} z)^{-1}$, where $z = \psi(0)$. The proof is simply minimization of that mixture's quadratic under $z^\top c = 1$. A feasible polynomial supplies the upper bound, evaluated at both endpoints and every real derivative root. This combines a finite-mixture numerical lower check with a continuum primal check.

E.7 Absolute-risk information and execution budgets

For operational regimes $s = 1, \dots, m$, perfect regime revelation permits a different degree-K polynomial in each regime and has minimax value $O_K = \max_s R_K^\star(\nu_s)$. Without revelation, convex duality gives $W_K = \max_{\pi\in\Delta_m} R_K^\star(\sum_s \pi_s \nu_s)$. Consequently $O_K \le W_K$, and the fixed-depth information improvement is $W_K - O_K$. A redesigned shared degree-$(K+1)$ program has value W_{K+1}. Since multiplying a degree-K minimizer by $1 - 2\lambda/(a+b)$ is feasible at degree $K+1$, the scalar interval bound with $\chi = (b-a)/(b+a)$ proves

$$W_{K+1} \le \max_s R_{\nu_s}((1 - 2\lambda/(a+b))p_K^{\mathrm{mm}}) \le \chi^2 W_K.$$

Thus the achieved one-product comparison is W_{K+1} versus O_K. In the absolute-risk analogue of Equation (4), the reference baseline is O_K and the information term is $W_K - O_K$.

Amortized execution cost. If a record is acquired once and reused over M solves, let C_{obs} be the total incremental acquisition and prefix cost after canceling costs common to the policies. The record policy costs $MK + C_{\mathrm{obs}}$ and one extra shared update on every solve costs $M(K+1)$. The latter is no more expensive exactly when $C_{\mathrm{obs}} \geq M$. Combined with $W_{K+1} < O_K$, this gives charged-cost dominance; replacing $<$ by $\leq$ gives weak risk dominance.

The operational regime set is part of the decision problem. If every law in the convex hull is itself a revealable regime, the perfect-information value becomes $\max_{\nu \in \mathrm{conv}\{\nu_s\}} R^\star_K(\nu) = W_K$, eliminating the absolute information gap.

F Empirical audit of learned schedules and transfer

The pre-specified audit (PA) fixed its policies and evaluation seeds 106–200 after development on seeds 101–105. Seeds index overlapping splits of the same 442 rows; the row bootstrap in Appendix F.2 refits the entire audit. For a policy on a split, L is the natural-log ratio of its risk to that of the mean-trained shared schedule. Log-ratio magnitudes below 10^{-5} print as zero.

F.1 Pre-specified policy comparison

PA loads the diabetes rows with `load_diabetes(scaled=False)` and uses 310 training and 132 held-out rows. Training-only standardization is followed by radial clipping $x \leftarrow x \min\{1, \sqrt{10}/\|x\|\}$. Each task bootstraps 64 rows; the regime parameters are $0.1, 0.3, 1$. The training/test banks have 64/256 tasks per regime. Since $\|x\|^2 \leq 10$, $\lambda_{\max}(X^\top X/64) \leq 10$, yielding the pre-specified interval $[0.1, 11]$ without test spectra. Responses are centered and scaled using the training rows. For each task, $g = X^\top y/64$, $w_0 = 0$, and $r_0 = -g$; spectral weights are $|u_i^\top r_0|^2/\|r_0\|^2$. Absolute-risk fits have no offset τ. All log ratios are natural logarithms paired within a split against mean-trained shared schedules. We regenerated PA seeds 101–105 from the public rows and retained per-purpose RNG seeds. Recomputed minimax test risks agree with the stored PA values to relative error at most 1.42×10^{-13}. Figure 7 uses split 101 of the development implementation, which preceded PA and draws different rows: for the same seed identifier, the two 310-row training sets share only 211–220 rows.

Table 2: Longer Adam optimization (Kingma and Ba, 2015) on PA seeds 101–200. Gaps are relative to the empirical worst-regime optimum; all 100 fits at each depth meet the 1% threshold after 20,000 updates.

K	Adam 2,500 median gap	Adam 20,000 median gap	Below 1% at 20,000
3	2.59%	1.40×10^{-7}	100/100
5	4.12%	2.10×10^{-7}	100/100
7	4.89%	2.70×10^{-7}	100/100

The fixed-interval $\kappa = 10^4$ guard has P90 risk 0.00449 and maximum split mean 0.00501, versus 0.00661 and 0.126 for mean training. Its median log ratio is 0.605, a tail–median tradeoff.

Closing the training optimization gap worsens held-out risk here. The worst-regime objective after 2,500 updates (ID 2) has median log ratio 0.131 and P90 0.0189; its converged optimum (ID 17) has 0.435 and 0.0249, while the median training gap falls from 4.12% to 2.1×10^{-7} (seeds 101–200, Table 2). Since the gap shrinks while held-out risk rises, the family-shift term of the absolute-risk analogue of (4) (Appendix E.7) dominates this comparison; in the sequential decomposition of Appendix F.3, most of that shift appears at the median before new rows enter. The converged pooled-mean optimum (IDs 4, 16, and 18) has P90 0.00690, slightly above the 0.00661 of the 2,500-update reference used for L.

F.2 Row-level bootstrap of the overlapping audit

We draw 200 bootstrap samples of all 442 source rows with replacement. For each draw we rebuild all 95 row splits using the original seed schedule and refit standardization, clipping, task banks, mean-Adam schedules, and empirical minimax. Thus each outer replicate contains the complete

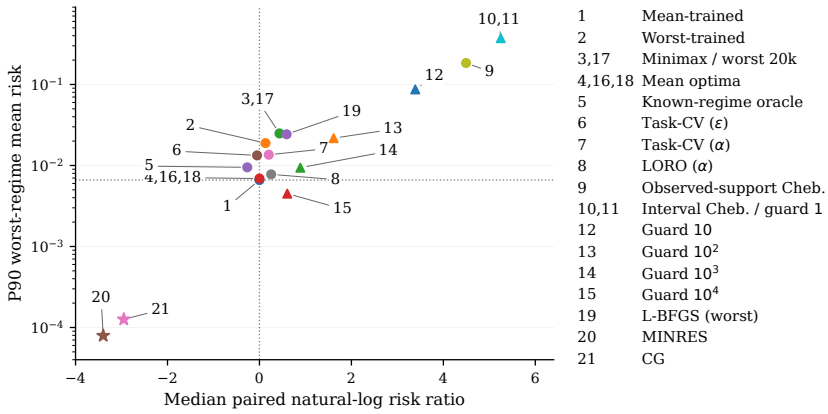


Figure 8: Median natural-log risk ratio and P90 for the 21 policies keyed in Table 4. Within the shared-schedule class, central risk and tail protection favor different policies. MINRES and CG are adaptive references. Grouped IDs overlap at plotting resolution; full distributions are in Figure 9.

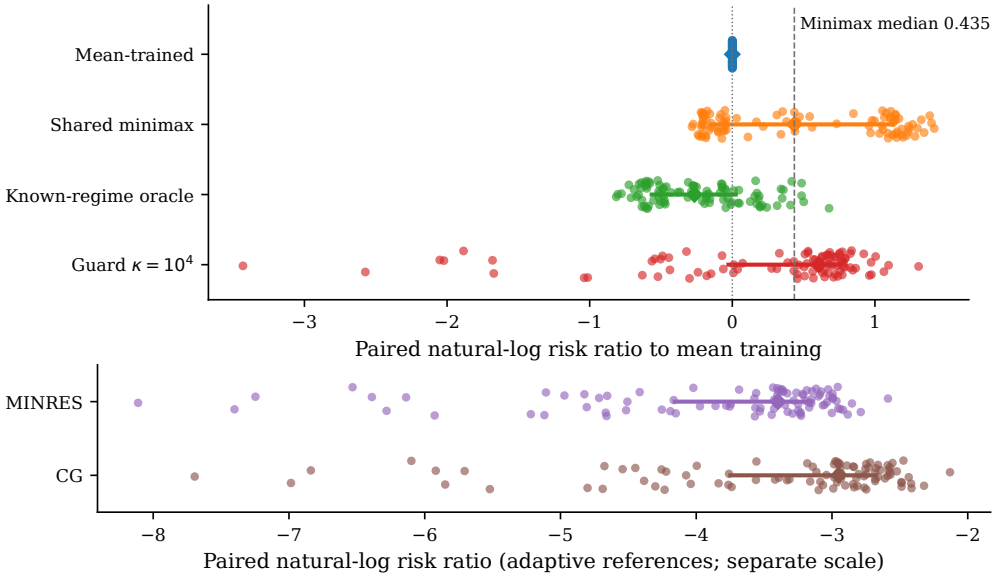


Figure 9: Each panel shows all 95 unchanged log ratios per policy. The upper panel isolates shared and regime-specific policies; adaptive references use a separate horizontal scale below. The upper dashed line is the minimax median 0.435.

Table 3: Depth-five policies on the 95 audit splits: median of e^L and P90 of worst-regime mean risk.

Policy	Median e^L	P90
Mean-trained	1	0.00661
Shared minimax	1.54	0.0249
Known-regime oracle	0.767	0.00949
Guard $\kappa = 10^4$	1.83	0.00449
MINRES	0.0334	7.92×10^{-5}
CG	0.0523	0.000126

Table 4: All 21 policies on PA seeds 106–200, $K = 5$. IDs match Figure 8. Log-ratio magnitudes below 10^{-5} are displayed as zero. P90 and Max summarize split-level worst-regime *mean* risk. Task-CV and LORO choose the indicated hyperparameter by task-level cross-validation and by leave-one-regime-out validation; L-BFGS is the limited-memory quasi-Newton method of Liu and Nocedal (1989).

ID	Policy	Median log	IQR	P90	Max
1	Mean-trained shared schedule	0	$[0, 0]$	0.00661	0.126
2	Worst-trained shared schedule	0.131	$[-0.127, 0.869]$	0.0189	0.323
3	Shared empirical minimax	0.435	$[-0.148, 1.13]$	0.0249	0.360
4	Pooled-mean oracle	0	$[0, 0]$	0.00690	0.126
5	Known-regime oracle	-0.266	$[-0.565, 0.0241]$	0.00949	0.205
6	Task-CV (ϵ)	-0.0519	$[-0.155, 0.408]$	0.0133	0.360
7	Task-CV (α)	0.203	$[-0.0172, 0.645]$	0.0136	0.257
8	LORO (α)	0.255	$[0.0747, 0.500]$	0.00777	0.0408
9	Observed-support Chebyshev	4.49	$[3.84, 4.75]$	0.184	0.197
10	Fixed-support Chebyshev	5.25	$[4.74, 5.53]$	0.373	0.398
11	Fixed-support guard $\kappa = 1$	5.25	$[4.74, 5.53]$	0.373	0.398
12	Fixed-support guard $\kappa = 10$	3.39	$[2.73, 3.78]$	0.0870	0.137
13	Fixed-support guard $\kappa = 10^2$	1.61	$[1.20, 1.85]$	0.0218	0.0692
14	Fixed-support guard $\kappa = 10^3$	0.889	$[0.657, 1.06]$	0.00940	0.0223
15	Fixed-support guard $\kappa = 10^4$	0.605	$[-0.0274, 0.750]$	0.00449	0.00501
16	Mean-trained (20,000 updates)	0	$[0, 0]$	0.00690	0.126
17	Worst-trained (20,000 updates)	0.435	$[-0.148, 1.13]$	0.0249	0.360
18	L-BFGS (mean)	0	$[0, 0]$	0.00690	0.126
19	L-BFGS (worst)	0.591	$[-0.00789, 1.27]$	0.0243	0.421
20	MINRES	-3.40	$[-4.16, -3.15]$	7.92×10^{-5}	1.33×10^{-4}
21	CG	-2.95	$[-3.75, -2.68]$	1.26×10^{-4}	2.12×10^{-4}

overlapping audit; no split-level independence is imposed. Percentiles at 2.5% and 97.5% give empirical row-bootstrap intervals. Duplicate source rows can occur in both partitions of an outer draw, as in the ordinary nonparametric bootstrap; intervals measure sensitivity under that empirical row distribution.

Table 5: Post-hoc whole-row bootstrap, 200 outer draws and 95 refitted splits per draw, $K = 5$. Intervals use the numerically stabilized refit described in the text. The high-loss fraction is the share of splits whose minimax/mean log ratio L exceeds the post-hoc cutoff 0.5.

Estimand	Frozen audit	Unresampled refit	95% row-bootstrap interval
High-loss fraction	0.463	0.463	$[0.358, 0.537]$
Overall median e^L	1.54	1.54	$[0.920, 2.02]$
Minimax P90 mean risk	0.0249	0.0249	$[0.0131, 0.0481]$

The fast refit evaluates the same quadratic training loss and analytic coefficient Jacobian through a QR factor, using the original 2,500-step Adam schedule and initializations. On unresampled rows, minimax risks match to relative error 3.48×10^{-8} and the overall median ratio remains 1.54. Mean-trained risks differ from the original fits by a median absolute log difference of 2.90×10^{-5} and at most 31.6% (seed 107); the mean-reference P90 is 0.00690 under the refit versus 0.00661

originally. Intervals refer to this QR refit. The largest minimax relative gap is 8.06×10^{-8} across 19,000 fits.

F.3 Spectral mass above the training maximum

For iid continuous task maxima, 256 test tasks exceed the maximum of 64 training tasks $256/65 \approx 3.94$ times in expectation. The exact no-exceedance probability is $64/(64+256) = 0.200$. Across PA seeds 101–200 the mean exceedance counts are 3.98 for 256 further training-pool tasks and 15.6 for the 256 held-out tasks; on primary seeds 106–200 they are 4.05 and 15.9, with medians 3 and 5. The same-pool count and share (Table 6) isolate finite sampling of the training rows. Tasks drawn from 132-row subsets of the training rows exceed the training maximum 7.1 to 9.6 times on average (seeds 106–200, five subsets), between the same-pool 4.05 and the held-out 15.9, so pool size and row identity both raise the counts. A sequential decomposition, which is descriptive and depends on the order of its steps, splits the log-risk change from the 64 training tasks to the 256 held-out tasks into a task-count term (256 new tasks from the 310 training rows), a pool-size term (tasks from a 132-row subset of those rows), and a row-identity term (tasks from the 132 held-out rows); no 310-row held-out pool exists, so the pool-size step is taken within the training rows. For the shared minimax the three medians are 0.251, 0.615, and 0.00595 (mean-trained: 0.0159, 0.286, and 0.0584); the pool-size step is the largest at the median for both, and the row-identity term has the widest interquartile range, $[-0.580, 1.09]$ for the shared minimax.

Table 6: PA median fraction of worst-regime risk from atoms above the training maximum.

K	PA seeds	n	Same-pool share	Held-out share
3	101–200	100	0.150	0.250
3	106–200	95	0.149	0.232
5	101–200	100	0.404	0.632
5	106–200	95	0.403	0.594
7	101–200	100	0.653	0.852
7	106–200	95	0.651	0.851

F.4 Retained versus reset continuation

The median retained/reset test-risk ratio is 0.939 for the learned prefix and 0.986 for a perturbed one, with retention lower on 66/100 and 57/100 seeds; these compare the retained and reset classes of Proposition C.1 at $T = 4$.

Table 7: PA held-out continuation at $K = 5$ from a four-step prefix, seeds 101–200.

Four-step prefix	Median retained/reset test risk	Retention lower
Learned	0.939	66/100
Perturbed	0.986	57/100

F.5 Restricted attention and optimization resolution

The prompt dimension is ten and the context length is $m = 64$. The base covariance is derived from the training rows of PA development split 101. Each of three depths $K \in \{3, 5, 7\}$ has seven covariance conditions and five optimizer initializations. Both scalar and full attention use fixed 20,000-update Adam training without architecture-specific learning-rate tuning or validation-checkpoint selection. This study uses the query prediction loss below, not the residual spectral risk of Section 2: its Gram H_q lacks the Hankel moment structure behind Theorem 1, and its null directions come from a $K \times 55$ Jacobian, whose null space has dimension at least $55 - K$ for any smooth map. It therefore tracks program movement in one trained model and does not test the exact fibers or rank of Theorem 1.

Model and loss. Prompt i has context rows $X_i \in \mathbb{R}^{64\times 10}$ and responses $y_i \in \mathbb{R}^{64}$, with $A_i = X_i^\top X_i/64 + 0.3I$ and $g_i = X_i^\top y_i/64$. The query target is $y_{q,i} = x_{q,i}^\top \beta_i + 0.1\,\epsilon_{q,i}$ with $\beta_i \sim$

$N(0, I_{10}/10)$ and $\epsilon_{q,i} \sim N(0,1)$. Models are trained on $N = 4096$ prompts, with 1024 validation and 1024 test prompts, by minimizing $J_N(\theta) = N^{-1}\sum_i(\widehat{y}_{\theta,i} - y_{q,i})^2$; the value 0.3 enters only through A_i, not as a parameter penalty. With the unnormalized convention $\mathrm{Att}(q; M_k, M_v) = qM_k^\top M_v$, one read of the prompt rows gives $\mathrm{Att}(z^\top; X, X/m)^\top = X^\top Xz/m$, and adding the ridge term gives Az. Scalar attention fixes the attention projections so that each layer reads $Aw_t - g$ from the prompt, and trains K unconstrained real step sizes and $K-1$ unconstrained momentum coefficients. With $w_{-1} = w_0 = 0$ and $\gamma_0 = 0$, layer t computes $w_{t+1} = w_t - \eta_t(Aw_t - g) + \gamma_t(w_t - w_{t-1})$, so $p_{t+1} = (1 + \gamma_t - \eta_t\lambda)p_t - \gamma_t p_{t-1}$ as in (10). This architecture realizes every $p \in \mathcal{A}_K$: a normalized linear factor $1 + \alpha_1\lambda$ needs one layer ($\eta = -\alpha_1$, $\gamma = 0$), and a normalized quadratic factor $1 + \alpha_1\lambda + \alpha_2\lambda^2$ needs two ($\eta = 1, \gamma = 0$, then $\eta = \alpha_2, \gamma = -\alpha_1 - \alpha_2 - 1$). The query oracle below is therefore attainable by the architecture in exact arithmetic, although nonconvex training need not attain it. Full attention replaces the scalars by matrix-valued weights shared across prompts. In the fixed basis $\phi_j(\lambda) = (\lambda/b)\,\mathsf{T}_{j-1}\big((2\lambda - a - b)/(b-a)\big)$, with $a = 0.3$ and b the largest eigenvalue over the base training prompts, held fixed across conditions, the program $p_c = 1 + \sum_{j=1}^K c_j\phi_j$ predicts

$$\widehat{y}_i(c) = x_{q,i}^\top A_i^{-1}[I - p_c(A_i)]g_i = F_i^\top c, \qquad F_{ij} = -b^{-1}x_{q,i}^\top \mathsf{T}_{j-1}\big((2A_i - (a+b)I)/(b-a)\big)g_i.$$

Hence $J_N(c) = c^\top H_q c - 2b_q^\top c + \|y_q\|_2^2/N$ with $H_q = F^\top F/N$ and $b_q = F^\top y_q/N$, and the query oracle $c_q^\star$ is the least-squares solution of $Fc \approx y_q$.

Covariance perturbations. With base covariance $\Sigma_0 = L_0L_0^\top$, the perturbed covariances are $\Sigma_\rho = L_0(I + \rho Q)L_0^\top$ for $\rho \in \{10^{-3}, 10^{-2}, 10^{-1}\}$, where Q is symmetric with $\|Q\|_F = 1$, so every Σ_ρ remains positive definite. For each depth, Q comes from the derivative of the base training query oracle. With a Frobenius-orthonormal basis $B_1, \ldots, B_{55}$ of symmetric 10×10 matrices, column k of D is the central difference of $c_q^\star\big(L_0(I \pm hB_k)L_0^\top\big)$ with $h = 5 \times 10^{-5}$ (a check at $2h$ is also computed). With $C^\top C = H_{q,0}$ and $CD = U_D\Sigma_D V_D^\top$, the sensitive direction is the first right singular vector and the null direction is the last right-null vector; for the chosen vector ϑ, $Q = \sum_k \vartheta_k B_k$. All conditions reuse the same Gaussian draws, teacher vectors, and noise, and regenerate features from the Cholesky factor of each covariance. The null direction has a numerically zero first derivative; the oracle is not invariant at finite ρ.

Scalar-polynomial check. On the first 32 test designs, let $r_i(g) = g - A_i\widehat{w}_i(g)$ be the residual response to input g, read from the predictions for queries $e_1, \ldots, e_{10}$, let $R_i = [r_i(e_1), \ldots, r_i(e_{10})]$, and let $s_i = \max(\|R_i\|_F, 10^{-300})$. A model passes if all four conditions hold: $\max_i \|r_i(0)\|_\infty \le 10^{-10}$; $\max_{i,\, g\in\mathcal{P}} \|r_i(g) - R_i g\|_2/s_i \le 10^{-8}$ for $\mathcal{P} = \{\xi_1, \xi_2, \xi_1+\xi_2, -2\xi_1\}$ with $\xi_1 = \mathbf{1}/\sqrt{10}$ and $(\xi_2)_j = (-1)^j/\sqrt{10}$; $\max_i \|A_iR_i - R_iA_i\|_F/(\|A_i\|_F s_i) \le 10^{-8}$; and $\max_i \|R_i - p_{\widehat{c}}(A_i)\|_F/s_i \le 10^{-6}$ for a single coefficient vector $\widehat{c}$ fitted by least squares across all 32 designs. The separate optimization criterion is $(J_{\text{learned}} - J_{\text{oracle}})/J_{\text{oracle}} \le 0.01$, required for at least four of five initializations per depth and condition.

Coefficient movement. Movement is measured in the fixed base-query Gram norm $\|\Delta c\|_{H_{q,0}} = (\Delta c^\top H_{q,0}\Delta c)^{1/2} = \|F_0\Delta c\|_2/\sqrt{4096}$, with basis and metric fixed at the base condition. Because the Hessian of J_N is $2H_q$, this is the half-Hessian norm; the Hessian-induced norm is larger by $\sqrt{2}$. Learned movement compares the same initialization, $\Delta c = c_{\rho,\text{seed}} - c_{0,\text{seed}}$, and oracle movement is $c_q^\star(\Sigma_\rho) - c_q^\star(\Sigma_0)$; neither is normalized by the oracle risk. Scalar-model coefficients are computed algebraically from the learned step and momentum recurrence and transformed into the basis ϕ_j; the 32-design polynomial fit only gates the scalar-polynomial check.

At $\rho = 10^{-3}$, depth-five and depth-seven null-direction oracle movements are 1.03×10^{-8} and 1.86×10^{-8}, versus learned 9.06×10^{-8} and 1.85×10^{-7}. At $\rho = 10^{-2}$, the learned and oracle null-movement norms differ by 46% and 49%.

Norms alone do not show direction, so Table 9 compares the movement vectors $u = \Delta\widehat{c}$ and $v = \Delta c^\star$ through $E = \|u - v\|_{H_{q,0}}$, $E/\|v\|_{H_{q,0}}$, and the Gram-metric cosine, as medians over the five initializations. Across the nine sensitive-direction conditions, the median relative vector error is at most 1.57% and the median cosine at least 0.99988. Of the 45 sensitive-direction runs, 41 have relative vector error below 4%; the other four come from one initialization, with errors up to

Table 8: Learned median / query-oracle coefficient movement in the base-query Gram norm. All 105 scalar runs meet the 1% objective-gap criterion.

Depth	direction	$\rho = 0.001$	$\rho = 0.01$	$\rho = 0.1$
3	null	$(4.59/2.97)\times10^{-8}$	$(2.98/2.96)\times10^{-6}$	$(2.84/2.83)\times10^{-4}$
3	sensitive	$(7.12/7.12)\times10^{-4}$	$(6.94/6.94)\times10^{-3}$	$(5.61/5.6)\times10^{-2}$
5	null	$(9.06/1.03)\times10^{-8}$	$(1.52/1.04)\times10^{-6}$	$(1.16/1.13)\times10^{-4}$
5	sensitive	$(5.34/5.34)\times10^{-4}$	$(5.12/5.12)\times10^{-3}$	$(3.75/3.75)\times10^{-2}$
7	null	$(18.5/1.86)\times10^{-8}$	$(2.75/1.85)\times10^{-6}$	$(1.81/1.79)\times10^{-4}$
7	sensitive	$(4.06/4.07)\times10^{-4}$	$(3.77/3.78)\times10^{-3}$	$(2.38/2.38)\times10^{-2}$

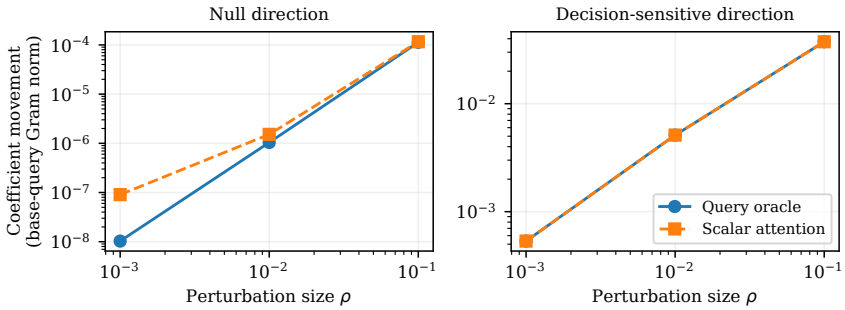


Figure 10: Learned and query-oracle coefficient movement at $K = 5$, measured in the base-query Gram norm.

50.1% at depth seven. Along null directions at $\rho = 10^{-3}$, median absolute errors of 3.35×10^{-8} to 1.82×10^{-7} are comparable to or larger than the oracle displacements, and median cosines are 0.38 to 0.69; we do not claim directional tracking of these near-null movements. The 1% objective gap does not by itself resolve such small movements: base coefficient-to-oracle errors in the same norm range from 3.2×10^{-5} to 5.5×10^{-3}, and paired movement errors are smaller because errors at the two conditions cancel.

Table 9: Movement vectors of learned and oracle coefficients in the base-query Gram norm. Medians over five initializations, except the last column.

K	Direction	ρ	Oracle norm	Median E	Median rel. E	Median cosine	Max E
3	sensitive	0.001	7.12×10^{-4}	5.72×10^{-7}	0.000803	0.9999997	6.84×10^{-7}
3	sensitive	0.01	6.94×10^{-3}	5.71×10^{-6}	0.000823	0.9999997	6.77×10^{-6}
3	sensitive	0.1	5.60×10^{-2}	5.01×10^{-5}	0.000894	0.9999996	5.65×10^{-5}
3	null	0.001	2.97×10^{-8}	3.35×10^{-8}	1.13	0.6850083	3.74×10^{-8}
3	null	0.01	2.96×10^{-6}	3.40×10^{-7}	0.115	0.9935472	3.80×10^{-7}
3	null	0.1	2.83×10^{-4}	3.90×10^{-6}	0.0137	0.9999058	4.28×10^{-6}
5	sensitive	0.001	5.34×10^{-4}	3.20×10^{-6}	0.006	0.9999821	3.42×10^{-5}
5	sensitive	0.01	5.12×10^{-3}	2.98×10^{-5}	0.00582	0.9999831	9.66×10^{-5}
5	sensitive	0.1	3.75×10^{-2}	1.56×10^{-4}	0.00416	0.9999916	1.75×10^{-4}
5	null	0.001	1.03×10^{-8}	8.72×10^{-8}	8.46	0.3832459	1.01×10^{-7}
5	null	0.01	1.04×10^{-6}	8.63×10^{-7}	0.829	0.8388965	9.99×10^{-7}
5	null	0.1	1.13×10^{-4}	7.69×10^{-6}	0.0678	0.9979287	9.05×10^{-6}
7	sensitive	0.001	4.07×10^{-4}	6.39×10^{-6}	0.0157	0.9998883	2.04×10^{-4}
7	sensitive	0.01	3.78×10^{-3}	4.21×10^{-5}	0.0111	0.9999442	1.61×10^{-3}
7	sensitive	0.1	2.38×10^{-2}	1.05×10^{-4}	0.00441	0.9999903	3.10×10^{-3}
7	null	0.001	1.86×10^{-8}	1.82×10^{-7}	9.78	0.3753440	1.10×10^{-6}
7	null	0.01	1.85×10^{-6}	1.79×10^{-6}	0.966	0.7646322	8.40×10^{-6}
7	null	0.1	1.79×10^{-4}	1.68×10^{-5}	0.0944	0.9957191	2.16×10^{-4}

Full attention passes the scalar-polynomial check in $0/105$ runs, so its maps lie outside $\mathcal{A}_K$. It beats scalar attention on $35/35$, $27/35$, and $0/35$ held-out comparisons at depths three, five, and seven; its class optimum is not computed, so these comparisons combine restriction and optimization effects.