WHEN DOES TRAINING ORDER BUY COMPUTE? EXACT BOUNDARIES AND ROBUST GAIN GUARANTEES

Anonymous authors
Paper under double-blind review

ABSTRACT

When does changing data weights outperform the best fixed mixture, rather than merely a poorly chosen baseline? We give an exact answer for shared-optimum quadratic learning. For two mirror sources at angle ϕ, with orthogonal initial-error mass ε, balanced mixing is optimal over all measurable schedules at budget B exactly when $\varepsilon \geq e^{-2B\cos\phi}$. Below the boundary, two mixed phases are enough to improve performance; sufficiently close to it they attain the global gain ceiling exactly. We then distinguish exactness from robustness. Aligned target anisotropy shifts the exact threshold, but arbitrarily small cross moments or unequal source strengths can destroy exact static optimality. A completed-square bound nevertheless limits the reoptimized gain quadratically in the cross moment, while a uniform perturbation bound covers general changes. Finite-budget plots show where candidate schedules approach the all-schedule ceiling and where a gap remains. Exact Gaussian stochastic gradient descent (SGD) experiments are indexed by distance from the boundary and control source composition through stratified batches and rapid alternation. A uniform SGD-to-flow bound guarantees convergence, not the observed near-boundary gain rate. The results identify both the value and the fragility of training order, without claiming a universal neural-network curriculum.

1 INTRODUCTION

Training on several data sources requires deciding both how much of each source to use and when to use it. Scaling laws inform allocations of model size, data, and training duration (Kaplan et al., 2020; Hoffmann et al., 2022); mixture methods additionally optimize source proportions (Ye et al., 2025; Shukor et al., 2025; Dai and Zheng, 2026). Changing those proportions during training creates another resource-allocation option. A source that is useful early may prepare the model to learn efficiently from a complementary source later. Yet a curriculum that beats uniform sampling may simply use better proportions. To show that changing weights helps, the baseline should be the *best fixed mixture at the same budget*, reoptimized as the budget changes.

This comparison is stronger than matching the curriculum's total source exposure or comparing an order with its reversal. A third static weight can beat all three. Nor does a nonzero local order effect imply that any schedule beats optimized mixing. Dynamic mixture methods and control-based planners already identify useful schedules (Chen et al., 2025; Gu et al., 2025; Mori et al., 2025; Zhao et al., 2026); local geometric methods predict which source should come first (Sweeney, 2026). We instead ask when *no schedule can help*, when a simple one can help, and how much of its apparent benefit survives perturbations or stochastic training.

We isolate these questions in quadratic learning with a shared optimum and a fixed optimizer. The sources disagree only about which error directions they reduce quickly, not about the correct prediction. One phase can reshape the remaining error to make a later phase more effective. The same reshaping can, however, leave more error in a perpendicular direction. Its value therefore depends on the initial error's *second moment*, not only the source Hessians. The schedule is chosen before the initial error is drawn; it cannot inspect each realization and choose a different trajectory.

For two mirror sources separated by an acute angle ϕ, normalize the error mass along their bisector to one and call the perpendicular mass ε. Our first result is an exact boundary: balanced mixing is optimal among *all measurable schedules* at budget B if and only if $\varepsilon \geq e^{-2B\cos\phi}$. Arbitrarily

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.

many costless switches are allowed. Below the threshold, two mixed phases suffice for a strict improvement. Close enough to it, two phases attain the all-schedule loss lower bound exactly, giving gain $(1 + r)/(2\sqrt{r})$ for normalized uncertainty $r = \varepsilon/e^{-2B\cos\phi}$. The excess gain vanishes quadratically as $r \uparrow 1$; a strict mathematical benefit can thus be too small to justify planning.

The exact formula requires the stated structural assumptions. Aligned target weights preserve an exact threshold, with an explicit shift. In contrast, small cross-direction initial error or unequal source strengths can destroy exact equality even inside the nominal no-benefit region. We prove this after reoptimizing the static mixture. We also prove what remains stable: a quadratic upper bound on the fully optimized advantage under a cross moment, and a uniform tolerance for general perturbations. This distinguishes robustness of *small attainable benefit* from robustness of an exact zero-benefit set.

Our experiments test these distinctions directly. Reoptimized perturbed curricula are compared with all-weight static brackets, and finite-budget gains are plotted against a global ceiling that includes the energy bound. Exact expected stochastic gradient descent (SGD) is indexed by normalized uncertainty instead of pooling across opposite sides of the boundary. Stratified batches remove source-count noise; population-gradient updates isolate discretization; exposure-matched rapid alternation controls ordering. These tests include gains that disappear and failures of the fixed preparation rule. The contribution is a quantitative comparison theory with controlled stochastic consequences, not a claim that the resulting switch rule generally improves neural-network training.

2 Related work

Curricula and mixture design. Curriculum learning studies useful presentation orders (Bengio et al., 2009; Pentina et al., 2015). Transfer-based and convex analyses connect difficulty to convergence (Weinshall et al., 2018; Weinshall and Amir, 2020); teacher–student theory separates speed from final-performance effects (Saglietti et al., 2022), and empirical benefits depend on the regime (Hacohen and Weinshall, 2019; Wu et al., 2021). Domain Reweighting with Minimax Optimization (DoReMi), Domain Reweighting with Generalization Estimation (DoGE), Skill-it, and Aioli optimize domain or skill weights (Xie et al., 2023; Fan et al., 2024; Chen et al., 2023; 2025). RegMix predicts static mixtures; RegMix-D and Olmix study dynamic or development-stage mixing (Liu et al., 2025; Zhao et al., 2026; Chen et al., 2026). Mixture scaling models predict or explain performance across proportions (Ye et al., 2025; Shukor et al., 2025; Dai and Zheng, 2026). We give an exact comparison with the best static mixture and show how it changes under perturbations.

Comparison with control approaches. In teacher–student continual learning, Mori et al. (2025) combine exact order-parameter dynamics with backward adjoint and task-selection equations (their Eqs. 5, 8, 10), yielding structured replay protocols. They explicitly note that Pontryagin conditions are necessary, not a global certificate. Theorem 2.1 of Gu et al. (2025) gives necessary adjoint conditions for *time-invariant* example weights in offline language-model data selection, rather than an if-and-only-if comparison between dynamic and static classes. Sweeney (2026) predict local pairwise transfer order with Lie brackets. In our smaller shared-optimum model, a trajectory-wide inequality gives a sufficient global obstruction, and a constructed perturbation proves its necessity. The robustness results further distinguish exact optimality from small optimized regret. These results concern a narrower model than the nonlinear models and empirical methods of those works.

Classical tools and secondary scaling. Switched systems and matrix exponentials have established theories (Liberzon, 2003; Higham, 2008); The Golden–Thompson inequality is due to Golden (1965); Thompson (1965); our use of spectral Jensen follows from scalar Jensen applied to a matrix spectral measure. Solvable SGD, regression, and random-feature models connect spectral tails to learning curves (Bordelon and Pehlevan, 2022; Maloney et al., 2022; Bordelon et al., 2024), while feature learning can change exponents (Bordelon et al., 2025). Our spectral statement is only a rate-envelope corollary and is valid even with independent block schedules. Gaussian concentration in the paid-pilot analysis is also standard (Boucheron et al., 2013). The sharper claims concern the exact curriculum boundary, its attainable gain, and its behavior under specified structural and stochastic changes.

3 MODEL AND THE OPTIMIZED STATIC COMPARATOR

What the scheduler can change. Let $L_i(\theta) = \frac{1}{2}(\theta-\theta^*)^\top H_i(\theta-\theta^*)$, $i = A, B$, with $H_i \succeq 0$ and common minimizer θ^*. Realizable linear regression with source covariance H_i gives this loss after subtracting irreducible label noise. If source i costs $c_i > 0$ per unit of flow time, define $G_i = H_i/c_i$. With normalized training work t, the error obeys

$$\dot{e}(t) = -G(p(t))e(t), \qquad G(p) = pG_A + (1-p)G_B. \tag{1}$$

Only the work fraction $p(t)$ changes. The optimizer, learning-rate convention, model, and target are fixed. Work is the cost-normalized duration of the stated dynamics. Cost normalization allows unequal costs, but the exact theorem requires symmetry of the resulting normalized generators.

A schedule is a measurable $p : [0, B] \to [0, 1]$, chosen using the source matrices and the initial-error second moment, but not the realized error. A finite schedule has transition $\Phi_\pi = e^{-b_M G(p_M)} \cdots e^{-b_1 G(p_1)}$, with $\sum b_j = B$; the rightmost phase acts first. Intermediate fractions are allowed. Switching and planning are free for the main comparison, which favors curricula. With $S = \mathbb{E}[e_0 e_0^\top]$ and evaluation matrix $Q \succeq 0$, risk is

$$\mathcal{R}(\pi) = \tfrac{1}{2}\operatorname{tr}(Q\Phi_\pi S \Phi_\pi^\top). \tag{2}$$

Write $\mathcal{R}^*_{\mathrm{stat}}, \mathcal{R}^*_2, \mathcal{R}^*_{\mathrm{all}}$ for infima over static mixtures, at most two phases, and all measurable schedules. Their ordering is $\mathcal{R}^*_{\mathrm{all}} \le \mathcal{R}^*_2 \le \mathcal{R}^*_{\mathrm{stat}}$. A feasible numerical schedule upper-bounds optimal risk; it does not certify the optimum.

A family with interpretable error directions. Take $v_A = (1, 0)^\top$, $v_B = (\cos\phi, \sin\phi)^\top$, $0 < \phi < \pi/2$. Let u be their unit bisector and v its counterclockwise perpendicular. The reference family is

$$G_A = v_A v_A^\top + \rho I, \quad G_B = v_B v_B^\top + \rho I, \quad S = uu^\top + \varepsilon vv^\top, \quad Q = I, \quad \rho, \varepsilon \ge 0. \tag{3}$$

For instance, $e_0 = u + \sqrt{\varepsilon}\, Zv$ with equiprobable $Z = \pm 1$ has mean u, covariance $\varepsilon vv^\top$, and second moment exactly S. Thus ε is residual variation across initial errors, not a posterior radius automatically reduced by training. At zero the direction is known; at one the second moment is isotropic even if the mean is not zero.

Put $c = \cos\phi$, $a = (1+c)/2$, $d = (1-c)/2$. In the (u, v) basis,

$$G(p) = \begin{pmatrix} \rho + a & z \\ z & \rho + d \end{pmatrix}, \qquad z = (1-2p)\sqrt{ad}. \tag{4}$$

Balanced mixing removes coupling. It learns the bisector at rate $\rho + a$ and the perpendicular direction at the slower rate $\rho + d$. Other weights couple the two coordinates without changing these diagonals.

Proposition 1 (Exact best static mixture)**.** *For every $B \ge 0$, $\varepsilon \ge 0$, balance $p = 1/2$ minimizes static risk, with*

$$\boxed{\mathcal{R}^*_{\mathrm{stat}}(B, \varepsilon) = \tfrac{1}{2}\left[e^{-2(\rho+a)B} + \varepsilon e^{-2(\rho+d)B}\right].} \tag{5}$$

Spectral Jensen gives $u^\top e^{-2BG} u \ge e^{-2Bu^\top Gu}$ and the analogous bound for v; balance attains both. This is a classical inequality applied to the comparator, not a new matrix inequality. Appendix J gives the proof. Gains are static risk divided by candidate risk: one is a tie and values below one are worse. Equal-budget loss ratios are not the same as savings in work needed to reach a common target loss.

4 THE GLOBAL BOUNDARY AND THE SIZE OF THE BENEFIT

Theorem 2 (Exact curriculum–mixture boundary)**.** *For* (3) *and $B > 0$, balanced static mixing is optimal over all measurable schedules if and only if*

$$\boxed{\varepsilon \ge \varepsilon_{\mathrm{crit}}(B) := e^{-2B\cos\phi}.} \tag{6}$$

Below the threshold two mixed phases strictly improve on every static mixture. For fixed $0 < \varepsilon < 1$, static mixing is exactly optimal once $B \ge B_{\mathrm{stop}} = \log(1/\varepsilon)/(2\cos\phi)$.

For example, $\phi = 45^\circ$ and $\varepsilon = 0.01$ give $B_{\mathrm{stop}} \approx 3.26$. More switching cannot help after that budget in this model. Before it, the theorem establishes existence, not the parameters of a global finite-budget optimum.

Why the lower bound covers every schedule. Start from v, write $\Phi(t)v = (x(t), y(t))$, and set $W = y^2 - x^2$. The off-diagonal control cancels:

$$W' = -2(\rho + d)W + 2cx^2, \qquad W(0) = 1.$$

Consequently $\|\Phi v\|^2 \geq b := e^{-2(\rho+d)B}$: switching cannot accelerate the slow initial direction. It can still accelerate the other direction. Let $f = \Phi u$, $g = \Phi v$, $t = \|g\|^2$, $k = f^\top g$. The generator trace is fixed, so $(\det \Phi)^2 = D := e^{-2(1+2\rho)B}$. The Gram determinant identity gives

$$2\mathcal{R}(\pi) = \frac{D + k^2}{t} + \varepsilon t \geq \frac{D}{t} + \varepsilon t, \qquad t \geq b. \tag{7}$$

When $\varepsilon \geq D/b^2 = e^{-2cB}$, the last expression is minimized at b and balance attains it. This argument constrains the entire trajectory and permits measurable controls, not just a specified switch count.

Why two phases suffice below the boundary. Set $z = \gamma r_w$ on the first half and $z = -\gamma$ on the second, where $r_w = [2\cosh(cB/2) - 1]^{-1}$ and $\gamma > 0$ is small. This cancels the first variation of k. Reflection symmetry gives $k = O(\gamma^3)$ and $t = b + \mathcal{K}\gamma^2 + O(\gamma^4)$ for an explicit $\mathcal{K} > 0$. Therefore

$$\mathcal{R}(\pi_\gamma) - \mathcal{R}^*_{\text{stat}} = \tfrac{1}{2}\mathcal{K}(\varepsilon - e^{-2cB})\gamma^2 + O(\gamma^4).$$

The leading term is negative below the boundary. Appendix A proves the claim for arbitrary common diagonals and gives $\mathcal{K}$. This is not a promise that a fixed perturbation amplitude works arbitrarily close to the boundary.

Theorem 3 (An attainable near-boundary gain)**.** *For fixed B, ϕ, ρ, there is $r_0 < 1$ such that, for $r = \varepsilon/\varepsilon_{\text{crit}} \in (r_0, 1]$,*

$$\mathcal{R}^*_{\text{all}} = \mathcal{R}^*_2 = \sqrt{\varepsilon}\, e^{-(1+2\rho)B}, \qquad \frac{\mathcal{R}^*_{\text{stat}}}{\mathcal{R}^*_{\text{all}}} = \frac{1+r}{2\sqrt{r}}. \tag{8}$$

In particular, at $r = 1 - \xi$, the excess optimal gain is $\xi^2/8 + O(\xi^3)$.

To attain the bound, adjust the second phase as a function of the first to make $k = 0$ *exactly*, then choose their amplitude so that $t = \sqrt{D/\varepsilon}$. The implicit function theorem and the positive coefficient $\mathcal{K}$ establish a nonzero feasible interval (Appendix E). Its width is not uniform, but Appendix E.1 certifies every $r \in [7/8, 1]$ at $45°$, $B = 2$, $\rho = 0$ using outward-rounded inequalities, not sampled near-equalities. At $\xi = 0.1$ the leading gain is only 0.13%, so exact optimality need not make scheduling worth its information cost.

Finite-budget gains away from the boundary. For $\varepsilon < \varepsilon_{\text{crit}}$ every schedule also satisfies the energy bound and slow-column bound. We use their maximum,

$$L_{\text{all}} = \max\left\{\sqrt{\varepsilon}e^{-(1+2\rho)B},\ \tfrac{1}{2}(1+\varepsilon)e^{-2(1+\rho)B},\ \tfrac{1}{2}\varepsilon e^{-2(\rho+d)B}\right\}, \tag{9}$$

and $L_{\text{all}} = \mathcal{R}^*_{\text{stat}}$ above the threshold. For any candidate R_{cand}, $\mathcal{R}^*_{\text{stat}}/R_{\text{cand}} \leq \mathcal{R}^*_{\text{stat}}/\mathcal{R}^*_{\text{all}} \leq \mathcal{R}^*_{\text{stat}}/L_{\text{all}}$. Figure 1 shows both sides, rather than presenting an attainable curve alone. The bound-to-candidate factor $R_{\text{cand}}/L_{\text{all}}$ quantifies what remains unknown.

Across 108 angle–budget–uncertainty cases, including 84 below the threshold, the two-phase candidate-to-lower-bound factor has below-boundary median 1.04 and maximum 3.98. At $B = 4$, $45°$, $r = 0.25$, the candidate and ceiling gains both round to 1.25 (the loss gap is 0.15%); at $65°$, $B = 8$, $r = 0$, it is 25.46 versus 101.40. Eight equal-duration phases with multistart local search did not improve these candidates to meaningful precision. That negative search result is not a global certificate. Appendix H gives the full grids and a gap plot.

5 Which parts of the boundary survive perturbation?

Continuity of one useful schedule does not establish stability of an optimized no-benefit region. We therefore change the model, reoptimize the static mixture and curriculum, and distinguish an exact boundary from a tolerance on small gains.

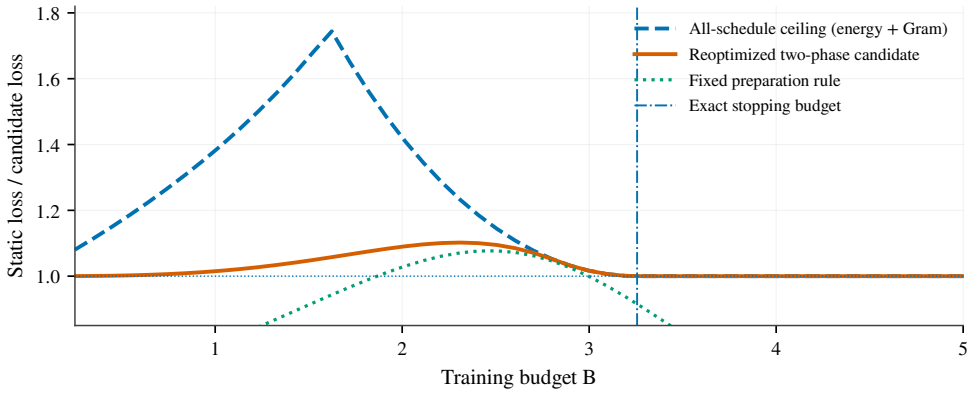


Figure 1: Finite-budget gain at $45°$, $\varepsilon = 0.01$, $\rho = 0$. The all-schedule ceiling includes the energy bound; the reoptimized two-phase curve is an attainable candidate. After B_{stop} the ceiling is exactly one. The fixed preparation rule can be worse than static mixing. The gap between the ceiling and candidate is not claimed closed everywhere.

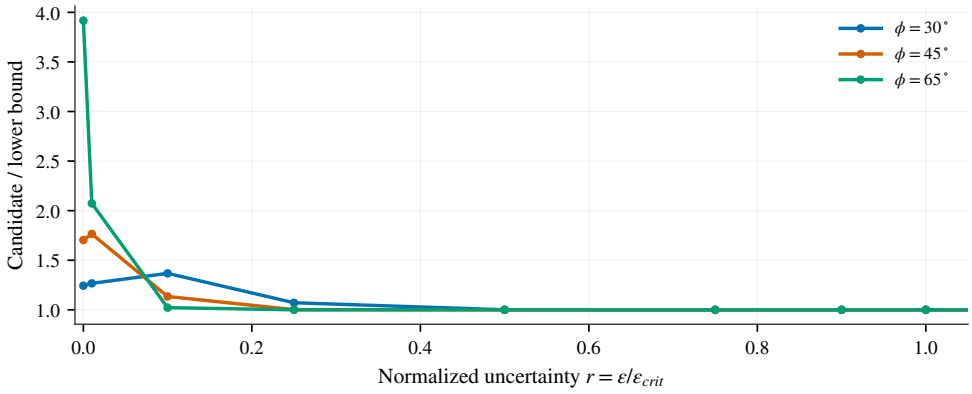


Figure 2: How much finite-budget uncertainty remains? At $B = 4$, the ratio of a two-phase candidate's loss to the all-schedule lower bound is close to one for some nonzero uncertainties but reaches 3.92 at $65°$ and $r = 0$. This is a value bracket, not a claim that the search has certified an optimal schedule.

Aligned targets: a shifted exact boundary. Let $Q = \operatorname{diag}(q_f, q_s)$ in the bisector basis, with positive weights, while retaining diagonal S. Balance is still the best static mixture, and Theorem 7 gives the exact condition

$$\mathcal{R}^*_{\text{all}} = \mathcal{R}^*_{\text{stat}} \quad \Longleftrightarrow \quad \varepsilon \geq \kappa_Q := \frac{q_f}{q_s} e^{-2B\cos\phi}. \tag{10}$$

The proof uses the same slow-column identity in a weighted Gram determinant. Thus identity evaluation is not essential; alignment is. Increasing the target weight on the slow direction makes static mixing preferable at a smaller orthogonal mass.

Cross moments and unequal strengths: exact equality can fail. Fix $e^{-2B\cos\phi} < \varepsilon < 1$. Proposition 9 proves that every sufficiently small nonzero cross moment in S, or a small mismatch between the two source strengths, makes the reoptimized best static mixture strictly improvable by two phases. Their total exposure can equal that of the best static mixture. The proof evaluates the time-dependent first variation at an interior static optimum: a static optimum only makes its time integral zero, whereas a nonconstant variation permits a beneficial early-versus-late perturbation. We do *not* claim that all perturbations merely shift the exact boundary.

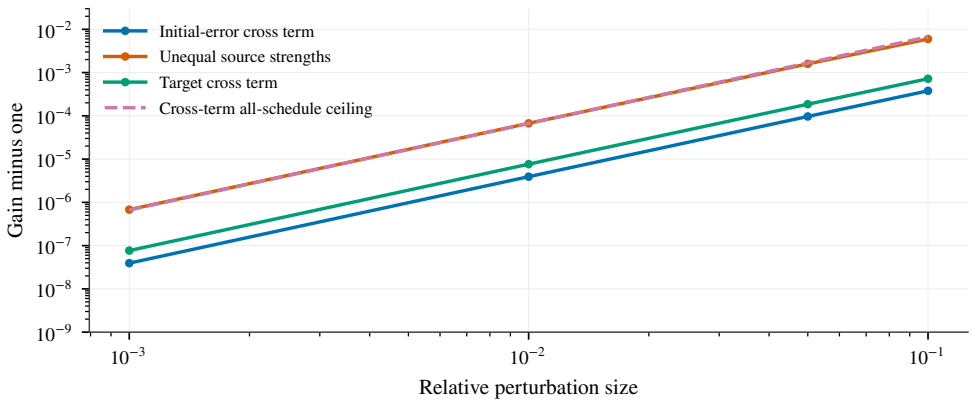


Figure 3: Reoptimized candidates at a nominal no-benefit point: $45°$, $B = 2$, $r = 2$. Initial cross moment is $s_\times = \delta\sqrt{\varepsilon}$; target cross term is δ; source strengths are $1 \pm \delta$. Exact equality is lost, but cross-term gains remain below a quadratic all-schedule ceiling. Each static comparator is bracketed over all weights, not fixed at balance.

What remains robust: a quadratic gain tolerance. Retain the mirror sources and aligned Q, but set $S = \left(\begin{smallmatrix} 1 & s_\times \\ s_\times & \varepsilon \end{smallmatrix}\right) \succeq 0$. The completed-square form of the weighted Gram identity is

$$2\mathcal{R} = \frac{D_Q}{t} + \frac{(k + s_\times t)^2}{t} + (\varepsilon - s_\times^2)t, \quad t \geq q_s e^{-2(\rho+d)B}, \quad D_Q = q_f q_s D.$$

It gives, after *both* classes are reoptimized,

$$\varepsilon - s_\times^2 \geq \kappa_Q \implies 0 \leq \mathcal{R}^*_{\text{stat}} - \mathcal{R}^*_{\text{all}} \leq \tfrac{1}{2} s_\times^2 q_s e^{-2(\rho+d)B}, \qquad \frac{\mathcal{R}^*_{\text{stat}}}{\mathcal{R}^*_{\text{all}}} \leq \frac{\kappa_Q + \varepsilon}{\kappa_Q + \varepsilon - s_\times^2}. \tag{11}$$

Balance is only an admissible upper-risk comparator here, not assumed to remain the best static weight. For $Q = I$ and correlation $s_\times = \delta\sqrt{\varepsilon}$, this bounds the excess gain by $O(\delta^2)$ inside the stated region. A cross term in Q has an analogous bound by reversing the schedule and exchanging Q with S (Appendix D).

An explicit tolerance beyond common diagonals. For general source, moment, or target changes, compare a nominal problem $(G_{i,0}, S_0, Q_0)$ with $(G_{i,1}, S_1, Q_1)$. Let δ_G be the largest source-operator difference, $\delta_S = \|S_1 - S_0\|_1$, and $\delta_Q = \|Q_1 - Q_0\|_{\text{op}}$. If all generators have minimum eigenvalue at least $m \geq 0$, a uniform risk radius is

$$E_B = e^{-2mB}\left[\frac{\delta_Q \operatorname{tr} S_0 + \|Q_1\|_{\text{op}}\delta_S}{2} + \|Q_1\|_{\text{op}} \operatorname{tr} S_0 \min\{2, B\delta_G\}\right]. \tag{12}$$

Proposition 10 proves $|R_1(\pi) - R_0(\pi)| \leq E_B$ for every schedule. Taking infima separately over static and arbitrary schedules therefore gives

$$|\Delta_1 - \Delta_0| \leq 2E_B, \qquad \Delta_j := R^*_{\text{stat},j} - R^*_{\text{all},j}. \tag{13}$$

In a nominal no-benefit region, $\Delta_1 \leq 2E_B$. This answers the stability question without assuming that either optimizer remains unchanged. The bound covers unequal strengths and simultaneous changes, but can be conservative when $B\delta_G$ is appreciable or the target risk is already very small. Absolute and relative robustness are not interchangeable. For a positive nominal static risk $r_0 > E_B$, the perturbed gain is at most $(r_0 + E_B)/(r_0 - E_B)$; otherwise this radius gives no useful relative bound.

A tolerance has an operational interpretation. In the cross-moment case, a proposed improvement larger than the right side of (11) is impossible under the stated model, regardless of how long the scheduler searches. A smaller improvement may exist, but it must still pay for estimating the geometry and constructing the schedule. For generic Hessian changes the broader radius plays the same role, often too conservatively to recommend an action. This is why we report both the exact structural

counterexample and the more informative quadratic special case, rather than treating mere continuity as evidence that the exact boundary persists.

The targeted perturbation grid has 720 cases with source-strength mismatch, initial cross moment, target cross moment, or aligned target anisotropy. Every static bracket has relative width below 2.1×10^{-10}. All cases reoptimize two phases; a 72-case subset also searches eight phases. At the point in Figure 3, a 5% perturbation gives excess gains of $9.70 \times 10^{-3}\%$, 0.16%, and 0.02% for the first three mechanisms, respectively. The cross-moment and target-cross excess-gain ceiling is 0.17%. These small, resolved advantages support fragility of exact equality without suggesting a large practical improvement.

6 Testing the boundary under stochastic training

Control the distance to the boundary. The boundary-indexed grid uses angles $30°, 45°, 65°$, budgets $B = 2, 4$, and $r \in \{0, .01, .05, .1, .25, .5, .75, .9, 1, 1.1, 1.5, 2, 4\}$. For each case set $\varepsilon = re^{-2B\cos\phi}$. With fresh Gaussian examples, zero ridge, and zero label noise, compare steps $\eta = .05, 1/80, 1/320$ and batches $m = 8, 64, 512$. We calculate exact expected SGD risks for per-example source draws, whole-batch source draws, and stratified batches. Population-gradient Euler updates isolate discretization. There are 2,340 conditions.

All methods within a condition receive the same $N = B/\eta$ updates and, when data are sampled, Nm examples. Across step sizes or batch sizes those costs differ; the experiment studies a controlled limit, not a compute-optimal choice of learning rate or batch. The candidate library includes every integer pure-source switch in both orders, 19 weak-steering amplitudes and their reversals, and the best evaluated static weight. It assumes known moments and does not charge search. Stratified fractions are rounded to legal integer source counts. These are existence tests for candidates, not global stochastic optima.

Static lower bounds cover all weights via a curvature envelope. All brackets have relative width below 10^{-3}; 1,332 near-boundary or primary-figure cases are adaptively refined to below 2.1×10^{-10}. For stratification, the lower comparator additionally allows a continuous positive-map relaxation, whereas the upper comparator is implementable. These are float64 evaluations of analytical brackets, not directed-rounding proofs. Figure 4 therefore does not conceal a coarse grid error behind very small gains.

Separate source-count variance from ordering. Let $\Delta H = H_A - H_B$ and T denote the homogeneous second-moment update. With identical expected source fraction p, Proposition 14 gives the exact differences

$$(T_{\text{example}} - T_{\text{strat}})S = \frac{\eta^2}{m} p(1-p)\Delta H S \Delta H, \qquad (T_{\text{batch}} - T_{\text{strat}})S = \eta^2 p(1-p)\Delta H S \Delta H. \tag{14}$$

Both are positive semidefinite. A pure phase removes these terms, so its advantage over random mixing alone need not identify deterministic steering. Increasing batch size removes per-example source-count noise, but not whole-batch source randomness at fixed η. Stratification removes source-count noise directly while retaining within-source sampling noise.

Which discrepancies disappear? At $45°$, $B = 2$, $m = 512$, and $r = 1$, per-batch candidate excess gain falls from 3.89% at $\eta = .05$ to $1.53 \times 10^{-2}\%$ at $\eta = 1/320$. The corresponding population-gradient values are 0.18% and $8.31 \times 10^{-4}\%$; example-source and stratified results approach that smaller discretization effect. Below the threshold, useful gains need not disappear: at $r = .5$ the finest-step stratified candidate still gains about 3.59%. These are resolved candidate comparisons, not assertions of an exact shifted SGD boundary.

A quantitative stochastic tolerance. Proposition 15 supplies a uniform convergence guarantee. Suppose $0 \preceq H_i \preceq LI$, $\eta L \leq 1$, and the homogeneous moment maps are trace-nonexpansive, a condition checked on the entire grid. In parameter dimension D_{par}, define $\omega = \sigma_y^2 D_{\text{par}} L/m$ and $\nu_{\text{strat}} = (D_{\text{par}}+1)L^2/m$. The example-source and batch-source values of ν add $\|\Delta H\|_{\text{op}}^2/(4m)$ and $\|\Delta H\|_{\text{op}}^2/4$, respectively. Population-gradient Euler updates have $\nu = \omega = 0$. Every precommitted,

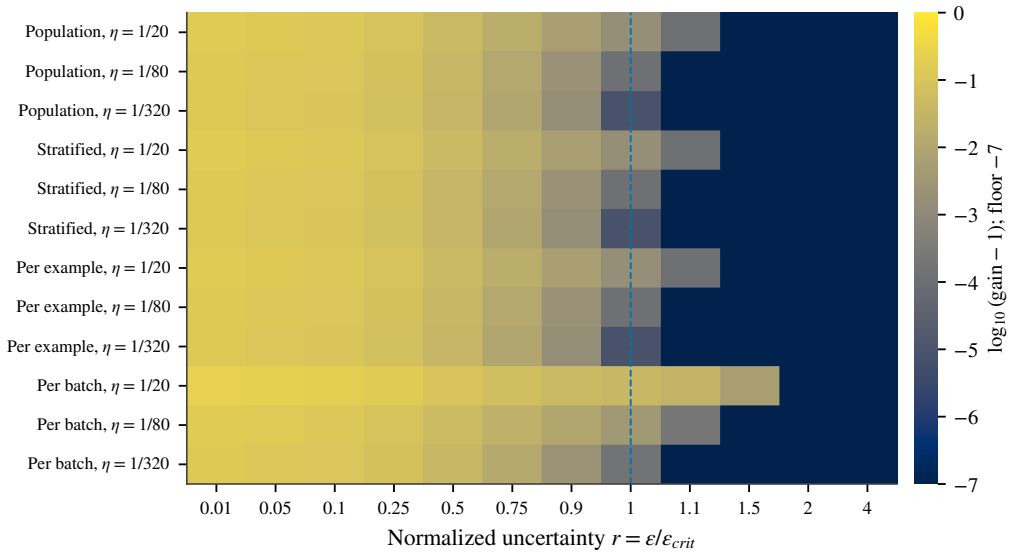


Figure 4: Boundary-indexed exact-risk diagram at $45°$, $B = 2$, $m = 512$. Color shows the logarithm of candidate excess gain, floored at 10^{-7}; it is not a phase classification of globally optimal SGD. The dashed column marks $r = 1$. Above-boundary gains from discretization and whole-batch source randomness diminish as η decreases. Population gradients have no sampling noise.

Table 1: At the nominal boundary $r = 1$, excess candidate gain in percent ($45°$, $B = 2$, $m = 512$). Entries use the same exact-risk calculations as the figures and are checked against refined static brackets; small values use scientific notation with two decimal places. They do not estimate globally optimal stochastic schedules.

Sampling rule	$\eta = 1/20$	$\eta = 1/80$	$\eta = 1/320$
Population gradients	0.18	1.26×10^{-2}	8.31×10^{-4}
Stratified batch	0.18	1.27×10^{-2}	8.25×10^{-4}
Independent per example	0.18	1.26×10^{-2}	8.34×10^{-4}
One source per batch	3.89	0.35	1.53×10^{-2}

grid-aligned schedule satisfies

$$|R_{\mathrm{SGD}}(\pi) - R_{\mathrm{flow}}(\pi)| \le E_{\mathrm{SGD}} := \frac{B\eta}{2}\left[(L^2 + \nu)\{\operatorname{tr} S_0 + B\eta\omega\} + \omega\right]. \tag{15}$$

For $r \ge 1$, the fully optimized stochastic advantage is at most $2E_{\mathrm{SGD}}$. A flow improvement exceeding $2E_{\mathrm{SGD}}$ suffices to survive under the same protocol. The bound follows by comparing a single moment update with its exact-flow step and telescoping over contractive suffixes. It applies to all admitted schedules, not just the library.

Table 1 isolates one boundary point rather than pooling it with low-uncertainty cases. At large batch size the first three rows nearly agree; whole-batch selection retains additional randomness. At the finest-step population point, $2E_{\mathrm{SGD}} = 6.62 \times 10^{-3}$, or 20.12% of the nominal flow risk, while the observed excess gain is only $8.31 \times 10^{-4}\%$. The radius is about $24{,}259$ times the observed absolute risk improvement. It proves convergence, not the measured displacement of the boundary or its small-step rate; increasing m alone cannot remove Euler error or whole-batch source randomness. Because the weak-steering library and legal stratified fractions are finite, a tie above or below $r = 1$ is not evidence of an exact stochastic optimality region. Positive comparisons are bracketed; absences of improvement remain candidate-specific.

A further 216 comparisons hold pure-source exposure exactly at $50/50$ while reducing the alternation dwell time from two long blocks to one update. For $m = 512$, the maximum relative error of one-update alternation from balanced flow decreases from 0.13 to 9.46×10^{-3} over the tested step

sizes. This control does not change total source use or select a different static proportion. Finally, 16 independent stratified Monte Carlo checks, each with 8,192 trajectories, all lie within three standard errors of the exact moment risks; the largest standardized discrepancy is 2.81. This is an implementation check, not a simultaneous statistical guarantee.

7 COMPUTE INTERPRETATION AND LIMITATIONS

At zero orthogonal mass, the explicit preparation rule aligns the residual with the second source and achieves the globally optimal exponential rate, up to a constant. At 45°, its equal-budget loss gains are 1.89 at $B = 4$ and 6.10 at $B = 8$, yet its asymptotic work-to-target advantage is only 1.17 (Appendix J). Large loss ratios should not be read as equally large compute savings. Additive isotropic process noise contributes at least the balanced schedule's noise risk, forcing optimal gain toward one at large budgets; this is a separate model, not an identification of all SGD with a diffusion (Appendix B).

The spectral result is a comparison corollary. If block scales satisfy $\lambda_j \asymp j^{-\alpha}$, masses $\operatorname{tr} S_j \asymp j^{-1-\beta}$, every source obeys $G_{i,j} \preceq L\lambda_j I$, and one static mixture has $G(p_0)_j \succeq m\lambda_j I$, then both shared and independently chosen block schedules obey $R^*(B, J) \asymp B^{-\beta/\alpha} + J^{-\beta}$, and fixed-capacity work $C = JB$ gives exponent $\beta/(\alpha + 1)$. The lower bound is blockwise; *schedule sharing is not essential*. A common static mixture attains the upper bound. Appendix L proves this comparison. This is an energy-and-tail corollary, not the sharper curriculum-specific contribution of the boundary theorems; changing cross-block budgets, capacity during training, or feature learning is outside it.

AI USE STATEMENT

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

REFERENCES

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. *International Conference on Machine Learning, pp. 41–48*, 2009. https://doi.org/10.1145/1553374.1553380.

Blake Bordelon and Cengiz Pehlevan. Learning curves for SGD on structured features. *International Conference on Learning Representations*, 2022. https://openreview.net/forum?id=WPI2vbkAl3Q.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. *International Conference on Machine Learning, Proceedings of Machine Learning Research (PMLR) 235:4345–4382*, 2024. https://proceedings.mlr.press/v235/bordelon24a.html.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. How feature learning can improve neural scaling laws. *International Conference on Learning Representations*, 2025. https://arxiv.org/abs/2409.17858.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. Concentration Inequalities: A Nonasymptotic Theory of Independence. *Oxford University Press*, 2013. https://academic.oup.com/book/26549.

Mayee F. Chen, Nicholas Roberts, Kush Bhatia, Jue Wang, Ce Zhang, Frederic Sala, and Christopher Ré. Skill-it! A data-driven skills framework for understanding and training language models. *Advances in Neural Information Processing Systems, 36*, 2023. https://papers.nips.cc/paper_files/paper/2023/hash/70b8505ac79e3e131756f793cd80eb8d-Abstract-Conference.html.

Mayee F. Chen, Michael Y. Hu, Nicholas Lourie, Kyunghyun Cho, and Christopher Ré. Aioli: A unified optimization framework for language model data mixing. *International Conference on Learning Representations*, 2025. https://arxiv.org/abs/2411.05735.

Mayee F. Chen, Tyler Murray, David Heineman, Matt Jordan, Hannaneh Hajishirzi, Christopher Ré, Luca Soldaini, and Kyle Lo. Olmix: A framework for data mixing throughout LM development. *arXiv:2602.12237*, 2026. https://arxiv.org/abs/2602.12237.

Rui Dai and Shuran Zheng. Explaining data mixing scaling laws. *International Conference on Machine Learning*, 2026. https://arxiv.org/abs/2606.08167.

Simin Fan, Matteo Pagliardini, and Martin Jaggi. DoGE: Domain reweighting with generalization estimation. *International Conference on Machine Learning, PMLR 235:12895–12915*, 2024. https://proceedings.mlr.press/v235/fan24e.html.

Sidney Golden. Lower bounds for the Helmholtz function. *Physical Review, 137(4B):B1127–B1128*, 1965. https://doi.org/10.1103/PhysRev.137.B1127.

Yuxian Gu, Li Dong, Hongning Wang, Yaru Hao, Qingxiu Dong, Furu Wei, and Minlie Huang. Data selection via optimal control for language models. *International Conference on Learning Representations*, 2025. https://arxiv.org/abs/2410.07064.

Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. *International Conference on Machine Learning, PMLR 97:2535–2544*, 2019. https://proceedings.mlr.press/v97/hacohen19a.html.

Nicholas J. Higham. Functions of Matrices: Theory and Computation. *Society for Industrial and Applied Mathematics*, 2008. https://doi.org/10.1137/1.9780898717778.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. *Advances in Neural Information Processing Systems, 35*, 2022. https://papers.nips.cc/paper_files/paper/2022/hash/c1e2faff6f588870935f114ebe04a3e5-Abstract-Conference.html.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv:2001.08361*, 2020. https://arxiv.org/abs/2001.08361.

Daniel Liberzon. Switching in Systems and Control. *Birkhäuser Boston*, 2003. https://doi.org/10.1007/978-1-4612-0017-8.

Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. RegMix: Data mixture as regression for language model pre-training. *International Conference on Learning Representations*, 2025. https://arxiv.org/abs/2407.01492.

Alexander Maloney, Daniel A. Roberts, and James Sully. A solvable model of neural scaling laws. *arXiv:2210.16859*, 2022. https://arxiv.org/abs/2210.16859.

Francesco Mori, Stefano Sarao Mannelli, and Francesca Mignacco. Optimal protocols for continual learning via statistical physics and control theory. *International Conference on Learning Representations*, 2025. https://arxiv.org/abs/2409.18061.

Anastasia Pentina, Viktoriia Sharmanska, and Christoph H. Lampert. Curriculum learning of multiple tasks. *IEEE Conference on Computer Vision and Pattern Recognition, pp. 5492–5500*, 2015. https://openaccess.thecvf.com/content_cvpr_2015/html/Pentina_Curriculum_Learning_of_2015_CVPR_paper.html.

Luca Saglietti, Stefano Sarao Mannelli, and Andrew Saxe. An analytical theory of curriculum learning in teacher-student networks. *Advances in Neural Information Processing Systems, 35*, 2022. https://proceedings.neurips.cc/paper_files/paper/2022/hash/84bad835faaf48f24d990072bb5b80ee-Abstract-Conference.html.

Mustafa Shukor, Louis Bethune, Dan Busbridge, David Grangier, Enrico Fini, Alaaeldin El-Nouby, and Pierre Ablin. Scaling laws for optimal data mixtures. *Advances in Neural Information Processing Systems, 38*, 2025. https://papers.nips.cc/paper_files/paper/2025/hash/bc1d640f841f752c689aae20b31198c1-Abstract-Conference.html.

John Sweeney. The geometry of sequential learning: Lie-bracket prediction of transfer order. *International Conference on Machine Learning*, 2026. https://arxiv.org/abs/2606.24993.

Colin J. Thompson. Inequality with applications in statistical mechanics. *Journal of Mathematical Physics, 6(11):1812–1813*, 1965. https://doi.org/10.1063/1.1704727.

Daphna Weinshall, Gad Cohen, and Dan Amir. Curriculum learning by transfer learning: Theory and experiments with deep networks. *International Conference on Machine Learning, PMLR 80:5238–5246*, 2018. https://proceedings.mlr.press/v80/weinshall18a.html.

Daphna Weinshall and Dan Amir. Theory of curriculum learning, with convex loss functions. *Journal of Machine Learning Research, 21(222):1–19*, 2020. https://jmlr.org/papers/v21/18-751.html.

Xiaoxia Wu, Ethan Dyer, and Behnam Neyshabur. When do curricula work? *International Conference on Learning Representations*, 2021. https://arxiv.org/abs/2012.03107.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. DoReMi: Optimizing data mixtures speeds up language model pretraining. *Advances in Neural Information Processing Systems, 36*, 2023. https://papers.nips.cc/paper_files/paper/2023/hash/dcba6be91359358c2355cd920da3fcbd-Abstract-Conference.html.

Jiasheng Ye, Peiju Liu, Tianxiang Sun, Jun Zhan, Yunhua Zhou, and Xipeng Qiu. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. *International Conference on Learning Representations*, 2025. https://arxiv.org/abs/2403.16952.

Kaiyan Zhao, Zhongtao Miao, Akiko Aizawa, and Yoshimasa Tsuruoka. RegMix-D: Dynamic data mixing via proxy training trajectories. *arXiv:2606.18663*, 2026. https://arxiv.org/abs/2606.18663.

PROOF AND EXPERIMENT NAVIGATION

The main results are not renumbered when recalled in their proofs.

Claim	Dependencies and location
Static optimum, Proposition 1	Spectral Jensen; proof in Appendix J.
Exact boundary, Theorem 2	Slow-column identity and Gram determinant; Appendix A.
Attainable gain, Theorem 3	Boundary witness and implicit function theorem; Appendix E. Explicit branch certificate in Section E.1.
Structural guarantees	Weighted identity, completed square, switching first variation, uniform Duhamel bound; Appendix D.
SGD bridge	Gaussian moments and composition-noise identity; Appendices F and G.
Structural and boundary experiments	Reoptimized perturbations, finite-budget brackets, normalized uncertainty, stratification and alternation; Appendix H.
Spectral comparison	Per-block energy bounds and a tail sum; Appendix L. Sharing is not needed.

A AN EXACT BOUNDARY FOR ARBITRARY SCHEDULES

This section proves the comparison with arbitrary measurable schedules, without restricting the number of phases. It also distinguishes the threshold for *some useful curriculum* from the more restrictive threshold for one particular preparation rule. All schedules in this section are chosen before drawing the initial error. The schedule may depend on its known second moment, but it must be the same linear evolution for both initial directions.

We prove a slightly more general statement. In an orthonormal basis (u, v), let

$$G(z) = \begin{pmatrix} a & z \\ z & d \end{pmatrix}, \qquad a > d > 0, \quad |z| \le h, \quad 0 < h \le \sqrt{ad}, \qquad S = \begin{pmatrix} 1 & 0 \\ 0 & \varepsilon \end{pmatrix}, \quad Q = I.$$

The two sources have $z = \pm h$ and any intermediate z is a mixture. The normalized family in the main text has diagonals $\rho + a_\phi, \rho + d_\phi$, off-diagonal limit $\sqrt{a_\phi d_\phi}$, and diagonal gap $c = \cos\phi$. In this appendix only, put $c = a - d$ and $s = a + d$.

> **Theorem 2 (general common-diagonal form).** For $B > 0$ and $\varepsilon \ge 0$, balanced static mixing is optimal among all measurable schedules if and only if
>
> $$\varepsilon \ge e^{-2cB}.$$
>
> If $\varepsilon < e^{-2cB}$, a schedule with just two mixed phases strictly improves on *every* static mixture. No assertion that this two-phase witness is globally optimal is needed.

Lemma 4 (A slow column cannot be accelerated)**.** *For every measurable $z : [0, B] \to [-h, h]$, its transition satisfies*

$$\|\Phi(B)v\|^2 \ge e^{-2dB}.$$

Proof. Write the solution from v as $(x(t), y(t))$, initially $(0, 1)$. Define $W(t) = y(t)^2 - x(t)^2$. The off-diagonal terms cancel exactly:

$$W'(t) = -2dW(t) + 2cx(t)^2.$$

Consequently

$$W(B) = e^{-2dB}\left[1 + 2c\int_0^B e^{2dt}x(t)^2\,dt\right] \ge e^{-2dB}. \tag{16}$$

Since $\|(x, y)\|^2 = W + 2x^2$, the result follows. Bounded measurable coefficients give an absolutely continuous solution; the identity and its integration hold almost everywhere. The argument therefore does not depend on a finite number of switches. □

Sufficiency in Theorem 2. Let $f = \Phi u$, $g = \Phi v$, $t = \|g\|^2$, and $k = f^\top g$. The trace of the generator is always s, so Liouville's formula gives $(\det \Phi)^2 = D := e^{-2sB}$. The two-dimensional Gram determinant identity yields

$$2\mathcal{R}(\Phi) = \frac{D + k^2}{t} + \varepsilon t \geq \frac{D}{t} + \varepsilon t, \qquad t \geq b := e^{-2dB}. \tag{17}$$

If $\varepsilon \geq D/b^2 = e^{-2cB}$, the function $D/t + \varepsilon t$ is nondecreasing for $t \geq b$. Thus

$$\mathcal{R}(\Phi) \geq \tfrac{1}{2}(D/b + \varepsilon b) = \tfrac{1}{2}(e^{-2aB} + \varepsilon e^{-2dB}).$$

The balanced schedule has orthogonal columns of these lengths, so it attains the bound. Lemma 16 already shows this is the globally best static risk. □

Necessity in Theorem 2. Consider $z_\eta(t) = \eta\zeta(t)$, where

$$\zeta(t) = \begin{cases} r_w, & 0 \leq t < B/2, \\ -1, & B/2 \leq t \leq B, \end{cases} \qquad r_w = \frac{\sinh(cB/2)}{\sinh(cB) - \sinh(cB/2)} = \frac{1}{2\cosh(cB/2) - 1} \in (0, 1).$$

For $0 < \eta \leq h$ this is feasible. The transition is analytic in η. At zero, $f_0 = (e^{-aB}, 0)$ and $g_0 = (0, e^{-dB})$. Variation of constants gives

$$\partial_\eta \Phi_{12}(0) = -e^{-aB} \int_0^B \zeta(t) e^{ct}\, dt, \qquad \partial_\eta \Phi_{21}(0) = -e^{-dB} \int_0^B \zeta(t) e^{-ct}\, dt.$$

Therefore

$$k'(0) = -\int_0^B \zeta(t) \left(e^{-2aB+ct} + e^{-2dB-ct}\right) dt = 0.$$

Indeed the parenthesized weight is $2e^{-sB}\cosh(c(B-t))$, and r_w is the ratio of its second-half to first-half integral. Conjugation by $\mathrm{diag}(-1, 1)$ changes η to $-\eta$; thus $k(\eta)$ is odd and $t(\eta) = \|g_\eta\|^2$ is even. In particular $k(\eta) = O(\eta^3)$.

For completeness, the increase in the slow column is strictly positive to second order. If $x_1(t) = \partial_\eta x_\eta(t)|_{\eta=0}$, then

$$x_1(t) = -e^{-at} \int_0^t \zeta(r) e^{cr}\, dr.$$

Applying (16) to g_η gives

$$t(\eta) = b + \mathcal{K}\eta^2 + O(\eta^4), \quad \mathcal{K} = 2x_1(B)^2 + 2ce^{-2dB} \int_0^B e^{2dt} x_1(t)^2 dt > 0.$$

Strict positivity holds because $r_w > 0$, so x_1 is nonzero on an initial interval. Substituting in the *exact* Gram expression in (17) gives

$$\mathcal{R}(\Phi_\eta) - \mathcal{R}^*_{\mathrm{stat}}(B) = \tfrac{\mathcal{K}}{2}\left(\varepsilon - e^{-2cB}\right)\eta^2 + O(\eta^4). \tag{18}$$

When $\varepsilon < e^{-2cB}$ the leading coefficient is strictly negative. A sufficiently small positive η therefore gives the claimed two-phase witness. This is an existence result with an explicit family of witnesses, not a claim that a fixed amplitude works uniformly near the boundary. □

Corollary 5 (Finite-budget lower envelope and eventual exact optimality)**.** *In the normalized family, set $\varepsilon_{\mathrm{crit}} = e^{-2cB}$. Every schedule satisfies*

$$\mathcal{R}(\Phi) \geq L(B, \varepsilon) := \begin{cases} \sqrt{\varepsilon}\, e^{-(1+2\rho)B}, & 0 \leq \varepsilon < \varepsilon_{\mathrm{crit}}, \\ \mathcal{R}^*_{\mathrm{stat}}(B, \varepsilon), & \varepsilon \geq \varepsilon_{\mathrm{crit}}. \end{cases}$$

It also satisfies $\mathcal{R}(\Phi) \geq \frac{1}{2}(1+\varepsilon)e^{-2(1+\rho)B}$ and $\mathcal{R}(\Phi) \geq \frac{1}{2}\varepsilon e^{-2(\rho+d)B}$. For $0 < \varepsilon < 1$, the balanced mixture is exactly *optimal for all*

$$B \geq B_{\mathrm{stop}}(\varepsilon) := \frac{\log(1/\varepsilon)}{2\cos\phi}.$$

In particular, the arbitrary-schedule and two-phase loss gains both equal one beyond this budget, a stronger statement than sharing an asymptotic rate.

Proof. Minimize $D/t + \varepsilon t$ on $t \geq b$. The unconstrained minimizer is $t = \sqrt{D/\varepsilon}$; it is feasible precisely below the critical value. At $\varepsilon = 0$, take the infimum as $t \to \infty$, which is a valid, possibly loose lower bound. The other two inequalities follow from the energy derivative and Lemma 4. Rearranging the boundary gives B_{stop}. □

What this theorem does not cover. The main theorem assumes a common, precommitted schedule for the initial-error ensemble. A scheduler that observes each realized error and chooses a different path has a different information model; the same transition cannot then be applied to both columns of S. Appendix D extends the exact result to aligned target anisotropy and gives quantitative tolerances for symmetry breaking. The formula does not automatically apply to state-dependent Hessians, noisy SGD, or a fixed charge for switching. Allowing costless arbitrary switching only strengthens the no-benefit side. Below the boundary the existence of a strictly better schedule does not imply that its small benefit pays for calibration or switching overhead.

B ISOTROPIC PROCESS NOISE: AN UNAVOIDABLE LOSS CONTRIBUTION

Consider the continuous-time model

$$de(t) = -G(p(t))e(t)\,dt + \sigma\,dW(t), \qquad \mathbb{E}[e(0)e(0)^\top] = S_\varepsilon,$$

with a deterministic measurable schedule, standard two-dimensional Brownian motion independent of $e(0)$, and the same normalized symmetric family. This is an explicitly specified additive-noise model, not an assertion that every SGD process has this diffusion.

Proposition 6 (Balanced mixing minimizes the noise contribution)**.** *Define*

$$N(B) = \frac{\sigma^2}{4}\left[\frac{1-e^{-2(\rho+a)B}}{\rho+a} + \frac{1-e^{-2(\rho+d)B}}{\rho+d}\right].$$

Every schedule has expected risk at least $L(B,\varepsilon)+N(B)$, where L may be replaced by the maximum of the lower bounds in Corollary 5. Balanced mixing has risk $\mathcal{R}^*_{\mathrm{stat}}(B,\varepsilon)+N(B)$. Thus it is exactly optimal above the same sufficient uncertainty boundary, and, for every fixed $\sigma > 0$, the best attainable gain tends to one as $B \to \infty$, even when $\varepsilon = 0$.

Proof. The second moment equals

$$\Phi(B,0)S_\varepsilon\Phi(B,0)^\top + \sigma^2\int_0^B \Phi(B,t)\Phi(B,t)^\top dt.$$

For each t, the suffix transition is an admissible schedule of length $B-t$. Theorem 2 at $\varepsilon = 1$ bounds its half squared Frobenius norm below by the balanced value. Integrating this bound gives $N(B)$ exactly. Add the deterministic lower bound. Since $N(B)$ converges to a strictly positive constant while $\mathcal{R}^*_{\mathrm{stat}}(B,\varepsilon) \to 0$, the gain is between one and $(\mathcal{R}^*_{\mathrm{stat}}+N)/(L+N)$ and tends to one. This proof requires the schedule to be independent of the Brownian path. □

C THE GAIN VANISHES QUADRATICALLY AT THE BOUNDARY

The boundary theorem distinguishes a strict improvement from no improvement. The next result quantifies how small the possible gain becomes just below the boundary.

> **Asymptotic consequence of Theorem 3.** Fix $B > 0$, an acute source angle, and a nonnegative ridge. Put $\varepsilon = e^{-2cB}(1-\xi)$ and let $\xi \downarrow 0$. Then
>
> $$\frac{\mathcal{R}^*_{\mathrm{stat}}(B,\varepsilon)}{\mathcal{R}^*_2(B,\varepsilon)} - 1 = \frac{\xi^2}{8} + O(\xi^3), \qquad \frac{\mathcal{R}^*_{\mathrm{stat}}(B,\varepsilon)}{\mathcal{R}^*_{\mathrm{all}}(B,\varepsilon)} - 1 = \frac{\xi^2}{8} + O(\xi^3).$$
>
> The constants are for fixed geometry and budget; this is not uniform as the source gap, off-diagonal strength, or budget tends to zero.

Proof. Use the general-family notation of Appendix A. Let $b = e^{-2dB}$ and $E = e^{-2aB}$. The unconstrained Gram lower envelope is $\sqrt{\varepsilon}e^{-sB} = E\sqrt{1-\xi}$, while the best static risk is $E(1-\xi/2)$. Hence their difference is $E\xi^2/8 + O(\xi^3)$.

For the explicit perturbation used in the necessity proof, choose $\eta^2 = b\xi/(2\mathcal{K})$, with $\mathcal{K} > 0$ as defined there. It is feasible for all sufficiently small ξ. Its slow-column length is $t = b(1+\xi/2) + O(\xi^2)$; its column inner product is $k = O(\xi^{3/2})$. The minimizing length in the Gram lower bound is

$t_* = b/\sqrt{1-\xi} = b(1+\xi/2) + O(\xi^2)$. Since the derivative of $D/t + \varepsilon t$ vanishes at t_*, replacing t_* by t costs $O(\xi^4)$. The additional Gram term $k^2/(2t)$ is $O(\xi^3)$. Thus this two-phase witness has risk at most $E\sqrt{1-\xi} + O(\xi^3)$.

Sandwich the all-schedule optimum and the two-phase optimum between the lower envelope and this feasible witness. Dividing the resulting loss difference by the optimal risk, which tends to $E > 0$, gives the claimed coefficient. A common ridge changes E and $\mathcal{K}$ but cancels from the risk ratios. □

D STRUCTURAL STABILITY: EXACT EXTENSIONS AND APPROXIMATE BOUNDARIES

This appendix separates three questions. An aligned evaluation metric preserves an exact boundary, with a changed threshold. Cross-direction error or unequal source strengths can remove exact equality, even strictly inside the nominal no-benefit region. Nevertheless, the *optimized* advantage remains bounded: a completed-square relaxation bounds the gain quadratically in a cross moment, and a uniform perturbation bound applies to general changes. These statements compare reoptimized competitors, not a fixed nominal preparation rule.

Throughout this appendix the generators in the bisector basis are

$$G(z) = \begin{pmatrix} \alpha & z \\ z & \beta \end{pmatrix}, \quad \alpha > \beta > 0, \quad |z| \le h \le \sqrt{\alpha\beta}, \quad h > 0, \qquad c = \alpha - \beta. \tag{19}$$

The endpoints are the two sources. The normalized main-text family has $\alpha = \rho + a$, $\beta = \rho + d$. Results are for fixed, precommitted schedules of length $B > 0$.

Theorem 7 (An exact boundary for aligned target weights)**.** *Let $Q = \mathrm{diag}(q_f, q_s)$ with $q_f, q_s > 0$, and $S = \mathrm{diag}(1, \varepsilon)$, $\varepsilon \ge 0$. Balanced mixing is the globally best static mixture, with risk*

$$R_{\mathrm{bal}} = \tfrac{1}{2}\left(q_f e^{-2\alpha B} + \varepsilon q_s e^{-2\beta B}\right).$$

It is optimal among all measurable schedules if and only if

$$\varepsilon \ge \kappa_Q := \frac{q_f}{q_s} e^{-2cB}. \tag{20}$$

Below this threshold two mixed phases give a strict improvement over every static mixture.

Proof. For a static generator, write $E = e^{-BG}$, which is symmetric. Direct expansion gives

$$2R = q_f E_{11}^2 + \varepsilon q_s E_{22}^2 + (q_f\varepsilon + q_s) E_{12}^2.$$

Spectral Jensen gives $E_{11} \ge e^{-\alpha B}$ and $E_{22} \ge e^{-\beta B}$. All coefficients are nonnegative, and balance attains the resulting lower bound. This argument does not incorrectly commute Q with E.

For an arbitrary transition put $f = \Phi u$, $g = \Phi v$, $t = g^\top Q g$ and $k = f^\top Q g$. The weighted Gram identity is

$$2R = \frac{D_Q + k^2}{t} + \varepsilon t, \qquad D_Q = q_f q_s e^{-2(\alpha+\beta)B}.$$

For $g = (x, y)$, the identity from Lemma 4 yields

$$t = q_s e^{-2\beta B} + (q_f + q_s) x(B)^2 + 2c q_s e^{-2\beta B} \int_0^B e^{2\beta t} x(t)^2 \, dt \ge b_Q := q_s e^{-2\beta B}. \tag{21}$$

Consequently $D_Q/t + \varepsilon t$ is minimized at b_Q if $\varepsilon \ge D_Q/b_Q^2 = \kappa_Q$. Balance attains equality.

For the converse let $z_\eta(t) = \eta\zeta(t)$, with ζ equal to r_w on the first half and -1 on the second half, where

$$w(t) = q_f e^{-2\alpha B + ct} + q_s e^{-2\beta B - ct}, \qquad r_w = \frac{\int_{B/2}^B w(t)\,dt}{\int_0^{B/2} w(t)\,dt} > 0.$$

Choose η small enough that $|z_\eta| \le h$. Variation of constants gives $k'(0) = -\int \zeta w = 0$. Reflection of the first coordinate changes η to $-\eta$, so $k = O(\eta^3)$. Also $t = b_Q + \mathcal{K}_Q \eta^2 + O(\eta^4)$, where

$$\mathcal{K}_Q = (q_f + q_s) x_1(B)^2 + 2c q_s e^{-2\beta B} \int_0^B e^{2\beta t} x_1(t)^2 \, dt > 0, \quad x_1(t) = -e^{-\alpha t} \int_0^t \zeta(s) e^{cs} \, ds.$$

Strict positivity follows because ζ is nonzero on an initial interval. Substitution into the exact weighted Gram identity gives $R_\eta - R_{\mathrm{bal}} = \frac{1}{2}\mathcal{K}_Q(\varepsilon - \kappa_Q)\eta^2 + O(\eta^4)$, proving the claim. □

Theorem 8 (Quadratic approximate no-benefit region)**.** *Keep* (19) *and diagonal Q from Theorem 7, but allow*

$$S = \begin{pmatrix} 1 & s_\times \\ s_\times & \varepsilon \end{pmatrix} \succeq 0.$$

For every schedule,

$$R \geq L_\times := \tfrac{1}{2} \inf_{t \geq b_Q} \left\{ D_Q/t + (\varepsilon - s_\times^2)t \right\}. \tag{22}$$

In particular, if $\varepsilon - s_\times^2 \geq \kappa_Q$, then, with both *competitor classes reoptimized,*

$$0 \leq \mathcal{R}^*_{\mathrm{stat}} - \mathcal{R}^*_{\mathrm{all}} \leq \tfrac{1}{2} s_\times^2 q_s e^{-2\beta B}, \qquad 1 \leq \frac{\mathcal{R}^*_{\mathrm{stat}}}{\mathcal{R}^*_{\mathrm{all}}} \leq \frac{\kappa_Q + \varepsilon}{\kappa_Q + \varepsilon - s_\times^2}. \tag{23}$$

Here $\mathcal{R}^*_{\mathrm{stat}}$ is not assumed to remain balanced. The lower bound is $L_\times = \sqrt{(\varepsilon - s_\times^2) D_Q}$ when $0 \leq \varepsilon - s_\times^2 < \kappa_Q$; this bound need not be attainable.

Proof. The weighted Gram identity now gives the exact completed square

$$2R = \frac{D_Q + k^2}{t} + 2s_\times k + \varepsilon t = \frac{D_Q}{t} + \frac{(k + s_\times t)^2}{t} + (\varepsilon - s_\times^2)t.$$

Use $t \geq b_Q$ and minimize the remaining one-variable expression. Balanced mixing has $k = 0$ and risk R_{bal}, since its evaluation-weighted cross term vanishes. Thus $\mathcal{R}^*_{\mathrm{stat}} \leq R_{\mathrm{bal}}$, while $\mathcal{R}^*_{\mathrm{all}} \geq R_{\mathrm{bal}} - s_\times^2 b_Q/2$ in the stated region. Subtracting or dividing proves (23). No inference about reoptimization is drawn from the continuity of a single candidate. □

Cross terms in the target metric. Symmetric generators imply that reversing a schedule transposes its transition. The trace identity $R(Q, S, \Phi) = R(S, Q, \Phi^\top)$ therefore preserves the attainable risk sets after exchanging Q and S. For $S = \mathrm{diag}(1, \varepsilon)$, $\varepsilon > 0$, and $Q = \left(\begin{smallmatrix} 1 & t_\times \\ t_\times & 1 \end{smallmatrix}\right) \succ 0$, Theorem 8 gives

$$\varepsilon(1 - t_\times^2) \geq e^{-2cB} \quad \Longrightarrow \quad 0 \leq \mathcal{R}^*_{\mathrm{stat}} - \mathcal{R}^*_{\mathrm{all}} \leq \tfrac{1}{2} t_\times^2 \varepsilon e^{-2\beta B}.$$

This is a tolerance, not a claim of an exact shifted boundary for a rotated target. The square completion is an exact algebraic identity, but dropping its square relaxes attainability. We do not establish a matching coefficient for the reoptimized gain under cross moments.

Proposition 9 (Exact equality is not structurally stable)**.** *Fix the normalized family, $B > 0$, $0 < \phi < \pi/2$, $Q = I$, and $e^{-2B\cos\phi} < \varepsilon < 1$. Either of the following perturbations makes every best static mixture strictly improvable by a two-phase schedule, for all sufficiently small nonzero perturbation parameters:*

1. *$S = \left(\begin{smallmatrix} 1 & s_\times \\ s_\times & \varepsilon \end{smallmatrix}\right)$, with the sources unchanged;*
2. *$G_A = (1 + \delta) v_A v_A^\top + \rho I$ and $G_B = (1 - \delta) v_B v_B^\top + \rho I$, with S unchanged.*

The two phases can preserve the best static mixture's total source exposure. This is a local perturbation statement at a fixed strict interior point, not a classification of all asymmetries.

Proof. For an interior static minimizer p_*, put $G = G(p_*)$ and $D = G_A - G_B$. In its eigenbasis with distinct eigenvalues λ_1, λ_2, the first functional derivative of risk at time t is

$$\begin{aligned} g(t) &= -\operatorname{tr}\left(D e^{-tG} S e^{-(2B-t)G} \right) \\ &= C_0 - 2 D'_{12} S'_{12} e^{-(\lambda_1 + \lambda_2)B} \cosh\left((\lambda_1 - \lambda_2)(B - t) \right). \end{aligned} \tag{24}$$

Primes denote this eigenbasis, not derivatives. If $D'_{12} S'_{12} \neq 0$, the function is strictly monotone on $[0, B)$; its averages over the two halves differ. A perturbation of p_* by $+\eta$ on one half and $-\eta$ on the other, with the sign chosen to reduce risk, is feasible for small η and has strictly negative first variation. It preserves exposure.

At the unperturbed problem balance is the unique static minimizer. Indeed spectral Jensen is strict in the u direction for a nonzero off-diagonal generator. Compactness and uniform continuity imply that every perturbed global static minimizer tends to $1/2$, and hence is interior. The source commutator is nonzero and stays nonzero, so $D'_{12} \neq 0$ at all these minimizers.

It remains to exclude $S'_{12} = 0$, equivalently $[S, G] = 0$. For the cross-moment perturbation, commutation requires $z = cs_\times/(1-\varepsilon) \neq 0$. Since $\operatorname{tr} D = 0$, the diagonal entries of D in the eigenbasis are opposite and nonzero: $\operatorname{tr}(GD) = -4hz \neq 0$. Static stationarity would then require the two eigenvalues of $e^{-BG}Se^{-BG}$ to be equal. At the unperturbed problem they are $e^{-2(\rho+a)B}$ and $\varepsilon e^{-2(\rho+d)B}$, which are distinct by the strict boundary assumption. They remain distinct for sufficiently small perturbations, a contradiction.

For unequal strengths, S has distinct eigenvalues, so commutation requires G to be diagonal in the bisector basis. Its off-diagonal entry is $h(1-\delta-2p)$, forcing $p = (1-\delta)/2$. There the diagonal entries are $\lambda_1 = \rho + (1-\delta^2)a$ and $\lambda_2 = \rho + (1-\delta^2)d$, whereas D has diagonal entries $2\delta a, 2\delta d$. The static derivative equals

$$-2\delta B\left(ae^{-2\lambda_1 B} + \varepsilon de^{-2\lambda_2 B}\right) \neq 0.$$

This again contradicts stationarity. Equation (24) now supplies the improving two-phase perturbation in both cases. □

Proposition 10 (Uniform tolerance for general perturbations)**.** *Let two problems have positive semidefinite generators $G_{i,j} \succeq mI$, second moments $S_j \succeq 0$, and targets $Q_j \succeq 0$, for $j = 0, 1$, $i = A, B$, $m \geq 0$. Write*

$$\delta_G = \max_i \|G_{i,1} - G_{i,0}\|_{\mathrm{op}}, \quad \delta_S = \|S_1 - S_0\|_1, \quad \delta_Q = \|Q_1 - Q_0\|_{\mathrm{op}},$$

where $\|\cdot\|_1$ is the nuclear norm. Define

$$E_B = e^{-2mB}\left[\tfrac{1}{2}\delta_Q \operatorname{tr} S_0 + \tfrac{1}{2}\|Q_1\|_{\mathrm{op}}\delta_S + \|Q_1\|_{\mathrm{op}} \operatorname{tr} S_0 \min\{2, B\delta_G\}\right]. \tag{25}$$

For every common schedule, $|R_1(\pi) - R_0(\pi)| \leq E_B$. Consequently, for $\Delta_j = R^*_{\mathrm{stat},j} - R^*_{\mathrm{all},j}$,

$$|\Delta_1 - \Delta_0| \leq 2E_B.$$

In a nominal no-benefit region, $0 \leq \Delta_1 \leq 2E_B$. If additionally $E_B < R^*_{\mathrm{stat},0} = r_0$, the perturbed gain is at most $(r_0 + E_B)/(r_0 - E_B)$.

Proof. Contractivity and Duhamel's formula give $\|\Phi_j(B,t)\| \leq e^{-m(B-t)}$ and $\|\Phi_1 - \Phi_0\| \leq e^{-mB}\min\{2, B\delta_G\}$, uniformly over measurable schedules. Add and subtract risks with the same target, second moment, and transition in turn. Trace duality bounds the target and second-moment terms by the first two summands in (25). Factoring the difference of the two propagated second moments bounds the remaining term by its third summand. Taking an infimum over any common competitor class changes its value by at most E_B; apply this separately to static and arbitrary schedules. This infimum argument is why the guarantee remains valid after reoptimizing both classes. The ratio bound follows by bounding the two perturbed optimum risks above and below by $r_0 \pm E_B$. □

The generic radius is intentionally conservative and grows with the perturbation horizon $B\delta_G$. It does not prove that a small relative gain remains useful at arbitrarily large budgets. The stronger quadratic cross-moment bound is used wherever its more specific assumptions apply.

E An attainable interval of the finite-budget gain ceiling

The gain ceiling in Corollary 5 is not globally tight at every uncertainty. Nevertheless, the near-boundary expansion can be strengthened to exact attainment on an interval, using two equal-duration mixed phases. This does not identify how far that interval extends at every budget.

Proposition 11 (Exact attainment sufficiently close to the boundary)**.** *For fixed $B > 0$ and geometry* (19)*, with $Q = I$, there is a number $r_0 \in [0, 1)$ such that, for $\varepsilon = re^{-2cB}$ and $r \in (r_0, 1]$,*

$$\mathcal{R}^*_{\text{all}}(B, \varepsilon) = \mathcal{R}^*_2(B, \varepsilon) = \sqrt{\varepsilon}\, e^{-(\alpha+\beta)B}, \qquad \frac{\mathcal{R}^*_{\text{stat}}}{\mathcal{R}^*_{\text{all}}} = \frac{1+r}{2\sqrt{r}}. \tag{26}$$

An analogous statement holds for positive diagonal Q, replacing the threshold by κ_Q and the optimum by $\sqrt{\varepsilon D_Q}$. The guaranteed interval depends on the fixed geometry and budget; no uniform lower bound on its width is asserted.

Proof. Consider two phases of length $B/2$, with off-diagonals z_1, z_2. Let $k(z_1, z_2) = f^\top Q g$ and $t(z_1, z_2) = g^\top Q g$. Both are analytic. At $(0, 0)$, $k = 0$, and its derivatives are the negatives of the strictly positive first-half and second-half integrals of $w(t)$ from Theorem 7. The implicit function theorem therefore gives an analytic $z_2 = \psi(z_1)$ near zero with $k(z_1, \psi(z_1)) = 0$, $\psi(0) = 0$, and nonzero negative derivative. Reflection of the first coordinate implies $\psi(-z_1) = -\psi(z_1)$.

On this curve, (21) yields $t(z_1, \psi(z_1)) = b_Q + \widetilde{\mathcal{K}}_Q z_1^2 + O(z_1^4)$ with $\widetilde{\mathcal{K}}_Q > 0$. The nonzero first-phase perturbation ensures strict positivity exactly as in the weighted boundary proof. In the coordinate $u_1 = z_1^2$, this is an analytic function with positive derivative at zero. Its image contains $[b_Q, b_Q + \delta_t)$ for some $\delta_t > 0$, while both controls remain in $[-h, h]$. For $\varepsilon < \kappa_Q$ sufficiently close to it, $t_* = \sqrt{D_Q/\varepsilon}$ lies in this image. At that control, $k = 0$ and the Gram identity attains the global lower bound $R = \sqrt{\varepsilon D_Q}$. At the threshold balance attains the same value. Substitution into the exact static formula gives (26). □

The gain plots evaluate the full energy-plus-Gram lower bound at all shown budgets. Outside the proven local interval, closeness of an optimized candidate to that bound is numerical evidence, not a claim that its parameters solve global optimal control. Conversely, a large bound-to-candidate ratio remains an explicitly unresolved finite-budget gap.

E.1 A USABLE INTERVAL WITH AN EXPLICIT ARITHMETIC CERTIFICATE

The local existence proof does not specify how far its branch can be continued. The following criterion makes that question finite and testable; it does not require solving the control problem globally. All interval checks below use explicit formulas for a two-phase transition.

Lemma 12 (A sufficient branch-continuation test)**.** *Fix the common-diagonal family with $Q = I$ and two phases of length $B/2$. Write $z_1 = x$, $z_2 = -\upsilon x$, $X = x^2$, and let $k = f^\top g$, $t = g^\top g$. Reflection implies that $F(X, \upsilon) := k(x, -\upsilon x)/x$ has an analytic extension to $X = 0$. Suppose $0 < v_- < v_+$ and the rectangle $0 \le X \le X_*$, $v_- \le \upsilon \le v_+$ satisfies*

$$F(X, v_-) < 0 < F(X, v_+), \qquad \partial_\upsilon F(X, \upsilon) > 0, \qquad \sqrt{X_*}\max\{1, v_+\} \le h.$$

At $X = X_*$, suppose an interval $[v_*^-, v_*^+] \subseteq [v_-, v_+]$ brackets a root of F, and every value of $t(X_*, \upsilon)$ on that interval is at least $b/\sqrt{r_*}$, for $0 < r_* < 1$ and $b = e^{-2\beta B}$. Then two feasible phases attain the all-schedule lower bound for every $r \in [r_*, 1]$*, where $\varepsilon = re^{-2(\alpha-\beta)B}$.*

Proof. The face signs and positive derivative give a unique root $\upsilon(X)$ in the rectangle for each X. This root varies continuously, including at zero. On the resulting feasible path, $k = 0$, $t(0) = b$, and $t(X_*) \ge b/\sqrt{r_*}$. The intermediate value theorem covers every length $b/\sqrt{r}$ between these endpoints. The Gram identity then attains its minimum. Monotonicity of t along the path is *not* required. This criterion is sufficient; failing to verify it does not exclude another feasible branch. □

Multiplying F and t by known positive factors does not change the root test, provided b is scaled with t. That normalization simplifies the concrete certificate.

Corollary 13 (Certified interval at a representative geometry)**.** *At $\phi = 45°$, $B = 2$, $\rho = 0$, $Q = I$, and $S = \operatorname{diag}(1, re^{-2\sqrt{2}})$, the conclusion of Theorem 3 holds for every*

$$\boxed{r \in [7/8, 1].}$$

This is a conservative certified interval, not the largest possible attainment interval. In particular, $\mathcal{R}^*_{\text{stat}}/\mathcal{R}^*_{\text{all}} = (1+r)/(2\sqrt{r})$ throughout it.

Finite certificate and analytic implication. Put $\delta = \sqrt{2}/4$, so $h = \delta$, and remove the scalar trace factor from the transition by defining $P = e\Phi$. Define the entire functions

$$C(q) = \cosh\sqrt{q} = \sum_{n\geq 0} \frac{q^n}{(2n)!}, \qquad H(q) = \frac{\sinh\sqrt{q}}{\sqrt{q}} = \sum_{n\geq 0} \frac{q^n}{(2n+1)!}.$$

Set $q_1 = 1/8 + X$, $q_2 = 1/8 + \upsilon^2 X$, and $C_i = C(q_i)$, $H_i = H(q_i)$. Direct multiplication of the two symmetric exponentials gives

$$\begin{aligned} P_{11} &= (C_2 - \delta H_2)(C_1 - \delta H_1) - \upsilon X H_1 H_2, \\ P_{22} &= (C_2 + \delta H_2)(C_1 + \delta H_1) - \upsilon X H_1 H_2, \\ P_{12}/x &= -(C_2 - \delta H_2)H_1 + \upsilon H_2(C_1 + \delta H_1) =: A, \\ P_{21}/x &= \upsilon H_2(C_1 - \delta H_1) - (C_2 + \delta H_2)H_1 =: J. \end{aligned}$$

The nonsingular quantities to verify are therefore

$$\widetilde{F} = P_{11}A + JP_{22}, \qquad \widetilde{t} = XA^2 + P_{22}^2, \qquad \widetilde{b} = e^{\sqrt{2}}.$$

Use $X \in [0, 1/25]$ and $\upsilon \in [37/25, 8/5]$. The interval calculation subdivides these intervals into 100 and eight equal rational cells. Outward-rounded enclosures on the faces and all 800 derivative boxes give

$$\widetilde{F}(X, 37/25) < -\tfrac{872}{10^4}, \quad \widetilde{F}(X, 8/5) > \tfrac{868}{10^4}, \quad \partial_\upsilon \widetilde{F} > \tfrac{21047}{10^4}. \tag{27}$$

Feasibility is strict since $|z_1| \leq 1/5$ and $|z_2| \leq 8/25 < \sqrt{2}/4$. At $X = 1/25$, the calculation additionally brackets the unique root by $\upsilon \in [\frac{155786246}{10^8}, \frac{155786247}{10^8}]$ and bounds

$$\tfrac{44042691103}{10^{10}} < \widetilde{t} < \tfrac{44042691186}{10^{10}}, \qquad \tfrac{41132503787}{10^{10}} < \widetilde{b} < \tfrac{41132503789}{10^{10}}.$$

The outward-rounded lower bound on $(7/8)\widetilde{t}^2 - \widetilde{b}^2$ exceeds 0.05. Thus $\widetilde{t} > \widetilde{b}/\sqrt{7/8}$ at the endpoint. Lemma 12 proves the all-r claim; checking only a finite list of r values would not do so. □

Arithmetic, not a root-finder tolerance. The interval calculation uses integer endpoints on a 10^{-40} grid. Addition, multiplication, division, and rational square roots are enclosed outwards using integer floor, ceiling, and square-root operations. It sums the positive series for C and H through degree 24. Each omitted tail is bounded above by its first omitted term divided by one minus an upper bound on all subsequent term ratios. Here $0 \leq q_i < 1$, so those ratios are strictly below one. The derivatives use $C' = H/2$ and $H' = (C - H)/(2q)$, with $q \geq 1/8$. An 80-term positive exponential series similarly bounds $\widetilde{b}$. These operations enclose entire cells and their analytic derivatives; they are not high-precision point evaluations interpreted as intervals. An unresolved strict inequality is treated as a failed certificate.

For orientation only, solving for the controls numerically at $r = 7/8$ gives $p_1 \approx 0.22$, $p_2 \approx 0.94$, with one unit of work per phase and optimal excess gain 0.22%. The printed, rounded controls do not themselves attain exact equality. The proof concerns the enclosed branch; the separate 80-digit control table is an illustration. The rectangle certificate implies an endpoint ratio near 0.87, but we deliberately state the simpler, conservative interval $[7/8, 1]$ and make no claim of maximality or uniformity across geometries.

F Exact Gaussian SGD and a global static bracket

The continuous-time theorem does not describe all sources of SGD noise. We therefore use an exact discrete-time calculation rather than fitting a diffusion to the training runs. The moment closure is a standard Gaussian calculation; related solvable SGD dynamics are studied by Bordelon and Pehlevan (2022). Our use of it is to compare a candidate schedule with a global bracket on all static mixture weights.

F.1 One source

Consider linear regression with a shared target, fresh independent examples $x \sim N(0, H)$, and independent zero-mean label noise of variance σ_y^2. A minibatch of size m takes one gradient step

of size η. The source choices and data at each update are independent of the current error. Let $S = \mathbb{E}[ee^\top]$. Then

$$S^+ = \mathcal{T}_H(S) + q_H, \qquad q_H = \eta^2\sigma_y^2 H/m,$$
$$\mathcal{T}_H(S) = S - \eta(HS + SH) + \eta^2[(1 + m^{-1})HSH + m^{-1}\operatorname{tr}(HS)H]. \tag{28}$$

To see this, write the update as $e^+ = (I - \eta\widehat{H})e + \eta\widehat{n}$. Gaussian fourth moments give $\mathbb{E}[xx^\top Sxx^\top] = 2HSH + \operatorname{tr}(HS)H$. The $m(m-1)$ cross-example terms contribute HSH, while the m same-example terms contribute the preceding fourth moment. The label-noise cross term vanishes, and $\mathbb{E}[\widehat{n}\widehat{n}^\top] = \sigma_y^2 H/m$. This proves (28). The distribution of e need not be Gaussian; its second moment and independence from the fresh batch suffice.

F.2 Two different meanings of a mixed minibatch

In the *batch-source* convention, a single domain is sampled with probability p, and all m examples come from it. Thus $T(p) = p\mathcal{T}_A + (1-p)\mathcal{T}_B$ and $q(p) = pq_A + (1-p)q_B$.

In our primary *example-source* convention, the domain is sampled independently for each example. Let $H_p = pH_A + (1-p)H_B$ and $M_i(S) = 2H_iSH_i + \operatorname{tr}(H_iS)H_i$. The correct map is

$$T(p)S = S - \eta(H_pS + SH_p) + \eta^2\left[(1 - m^{-1})H_pSH_p + m^{-1}\{pM_A(S) + (1-p)M_B(S)\}\right]. \tag{29}$$

The noise injection remains $q(p)$. This follows by separating the same-example and cross-example terms as above. In particular,

$$T(p) = T_0 + pT_1 + p^2T_2, \qquad T_0 = \mathcal{T}_B, \quad T_2(S) = \eta^2(1 - m^{-1})\Delta HS\Delta H, \quad T_1 = \mathcal{T}_A - \mathcal{T}_B - T_2,$$

where $\Delta H = H_A - H_B$. For batch-source sampling $T_2 = 0$. The conventions agree at $m = 1$ or a pure source, but not in general. The exact risk of any prescribed schedule is computed by composing these affine maps. An extra constant coordinate turns them into ordinary 4×4 matrices in a three-dimensional basis of symmetric 2×2 matrices.

F.3 A lower bound between evaluated static weights

A dense grid alone does not prove that a static comparator is globally optimal. Define its exact expected risk after N updates by

$$f(p) = \tfrac{1}{2}\operatorname{tr}\left[T(p)^N S_0 + \sum_{j=0}^{N-1} T(p)^j q(p)\right].$$

We bound every cell between grid points without assuming that f is convex. On symmetric matrices, let $\|\cdot\|_1$ denote the nuclear norm. Each $T(p)$ is positive because it is an expectation of maps $S \mapsto MSM^\top$. Its induced nuclear norm is $\|T(p)^*(I)\|_{\mathrm{op}}$. The upper bound follows by decomposing an input into positive and negative parts; equality follows by choosing a rank-one input in a top eigendirection of $T(p)^*(I)$.

For a cell $[l, r]$, let v upper-bound this map norm throughout the cell. The larger endpoint eigenvalue suffices: in the batch-source case $T(p)^*(I)$ is affine; in the example-source case it is matrix convex, since its quadratic coefficient is the positive semidefinite matrix $\eta^2(1 - m^{-1})(\Delta H)^2$. Let $\Delta_1 \geq \sup_{[l,r]} \|T'(p)\|_{1\to 1}$, $\Delta_2 \geq \|T''(p)\|_{1\to 1}$, $Q \geq \sup_{[l,r]} \|q(p)\|_1$, and $\Delta_q = \|q_A - q_B\|_1$. Differentiating each ordered product and using submultiplicativity gives

$$|f''(p)| \leq M := \tfrac{1}{2}\operatorname{tr}(S_0)\left[N(N-1)\Delta_1^2 v^{N-2} + N\Delta_2 v^{N-1}\right]$$
$$+ \tfrac{1}{2}Q\left[\Delta_1^2 \sum_{j=2}^{N-1} j(j-1)v^{j-2} + \Delta_2 \sum_{j=1}^{N-1} jv^{j-1}\right] + \Delta_1\Delta_q \sum_{j=1}^{N-1} jv^{j-1}. \tag{30}$$

Terms whose counting coefficient is zero are omitted. The Δ_1^2 terms count two differentiated factor positions; the Δ_2 terms count one twice-differentiated position. Differentiating the affine noise injection gives the last term. These arguments do not require the factors to commute.

For cell width $w = r - l$, the scalar interpolation remainder yields

$$\min\{f(l), f(r)\} - Mw^2/8 \le \inf_{p\in[l,r]} f(p).$$

The minimum of these lower bounds, truncated at zero, is a global lower bracket; the best evaluated point is an upper bracket. A candidate below the lower bracket beats every static mixture in exact real arithmetic.

Implementation and limitations. An orthonormal Frobenius basis represents each map by a 3×3 matrix. Since $\|S\|_F \le \|S\|_1 \le \sqrt{2}\|S\|_F$, $\sqrt{2}$ times its spectral norm bounds its induced nuclear norm. The derivative is affine in p, so endpoint norms bound Δ_1; Δ_2 is constant. The experiments use trace-nonexpansive maps ($v \le 1$). Each finite derivative sum is upper-bounded by the smaller of its value at $v = 1$ and its infinite geometric-series derivative. These are conservative upper bounds, not truncated approximations. Grid sizes increase from 1,025 to 4,097, 16,385, or 65,537 when needed to target a relative bracket width below 10^{-3}. All bounds are evaluated in float64, not directed-rounding arithmetic. The paper reports them as numerical evaluations of an analytical bracket, not formal computer-verified certificates. Monte Carlo checks and dense-grid tests independently test the implementation.

G Separating steering, source composition, and discretization

The deterministic theorem is not claimed to be an exact SGD threshold. We instead identify the noise terms that vary with source composition, and give a uniform approach to the deterministic comparison. The experiments are indexed by $r = \varepsilon/\varepsilon_{\mathrm{crit}}(B)$; a separate value of ε is used at each angle and budget to hold r fixed.

G.1 Stratified batches remove source-count noise

Let $H_p = pH_A + (1-p)H_B$ and $\Delta H = H_A - H_B$. Within each update use exactly mp independent Gaussian examples from source A and $m(1-p)$ from B, requiring mp to be an integer. Put

$$V_i(S) = H_i S H_i + \mathrm{tr}(H_i S) H_i.$$

Independence of examples and Gaussian fourth moments give the exact stratified second-moment map

$$T_{\mathrm{strat}}(p)S = (I - \eta H_p)S(I - \eta H_p) + \frac{\eta^2}{m}\{pV_A(S) + (1-p)V_B(S)\}. \tag{31}$$

The common label-noise injection is $q(p) = \eta^2\sigma_y^2 H_p/m$. Indeed, the covariance of the batch covariance is the sum of the independent within-source covariances divided by m^2. There is no random source count.

Proposition 14 (Exact composition-noise differences)**.** *For every $S \succeq 0$ and legal stratified fraction p, the example-source and batch-source maps from Appendix F satisfy*

$$[T_{\mathrm{example}}(p) - T_{\mathrm{strat}}(p)]S = \frac{\eta^2}{m}p(1-p)\Delta H S \Delta H, \tag{32}$$

$$[T_{\mathrm{batch}}(p) - T_{\mathrm{strat}}(p)]S = \eta^2 p(1-p)\Delta H S \Delta H. \tag{33}$$

Both differences are positive semidefinite. For the same legal precommitted schedule and initial moment, stratification has no larger expected risk for any positive semidefinite target metric.

Proof. Subtract (31) from the explicit maps in Appendix F. The identity $pH_A S H_A + (1-p)H_B S H_B - H_p S H_p = p(1-p)\Delta H S \Delta H$ gives both expressions. Induct on the number of updates: each map preserves positive semidefinite order, the injections agree, and the difference of maps applied to a positive semidefinite moment is positive semidefinite. Trace against a positive semidefinite target preserves the ordering. □

Thus increasing m removes example-source count noise, but not whole-batch source randomness at fixed step size. Switching to pure phases eliminates the count term in either convention, so

its benefit alone does not establish deterministic steering. Our stratified and population-gradient comparisons remove this confound directly. The continuous-p formula in (31) is a positive-map relaxation for noninteger mp, not an implementable fixed-size stratified batch. Static upper bounds and candidate schedules use legal integer counts; allowing every p for a lower bound only strengthens the comparator.

G.2 A UNIFORM FINITE-STEP BRIDGE TO THE DETERMINISTIC THEOREM

Use dimension D_{par}, generators $0 \preceq H_i \preceq LI$, and $N\eta = B$ with $\eta L \le 1$. All schedules here are precommitted and constant on update intervals. For the following bound assume also that the homogeneous moment maps are trace-nonexpansive for all admitted fractions, that is, $T(p)^*(I) \preceq I$. This condition is checked directly for the reported grids, rather than inferred from mean-gradient stability alone. Define

$$\nu_{\mathrm{strat}} = \frac{(D_{\mathrm{par}}+1)L^2}{m}, \qquad \omega = \frac{\sigma_y^2 D_{\mathrm{par}} L}{m},$$
$$\nu_{\mathrm{example}} = \nu_{\mathrm{strat}} + \frac{\|\Delta H\|_{\mathrm{op}}^2}{4m}, \quad \nu_{\mathrm{batch}} = \nu_{\mathrm{strat}} + \frac{\|\Delta H\|_{\mathrm{op}}^2}{4}.$$

For population-gradient Euler updates set $\nu = \omega = 0$. This control has neither sampled gradients nor labels and is included only to isolate discretization.

Proposition 15 (Uniform SGD-to-flow risk radius)**.** *Under these conditions, for every admitted schedule π and $Q = I$,*

$$|R_{\mathrm{SGD}}(\pi) - R_{\mathrm{flow}}(\pi)| \le E_{\mathrm{SGD}} := \frac{B\eta}{2}\left[(L^2+\nu)\{\operatorname{tr} S_0 + B\eta\omega\} + \omega\right]. \tag{34}$$

In the symmetric family, if $r \ge 1$ and balanced mixing is admitted, then

$$0 \le R^*_{\mathrm{stat,SGD}} - R^*_{\mathrm{schedule,SGD}} \le 2E_{\mathrm{SGD}}. \tag{35}$$

The schedule class may include every grid-aligned sequence of admitted fractions. Below the deterministic boundary, a flow candidate's gain transfers whenever its absolute *improvement exceeds $2E_{\mathrm{SGD}}$. No exact finite-step SGD boundary is asserted.*

Proof. For $K = I - \eta H_p$ and $F = e^{-\eta H_p}$, scalar Taylor remainder on each eigenvalue gives $\|F - K\|_{\mathrm{op}} \le \eta^2 L^2/2$, while $\|F\|_{\mathrm{op}}, \|K\|_{\mathrm{op}} \le 1$. Therefore

$$\|FSF^\top - KSK^\top\|_1 \le \eta^2 L^2 \operatorname{tr} S \quad (S \succeq 0).$$

The Gaussian within-source covariance has trace at most $\eta^2(D_{\mathrm{par}}+1)L^2 \operatorname{tr} S/m$; Proposition 14 bounds the composition terms. The label-noise injection has trace at most $\eta^2\omega$. The one-step error is consequently at most $\eta^2[(L^2+\nu)\operatorname{tr} S + \omega]$ in nuclear norm.

Trace nonexpansiveness implies $\operatorname{tr} S_k \le \operatorname{tr} S_0 + N\eta^2\omega = \operatorname{tr} S_0 + B\eta\omega$. Telescope the difference between the iterated stochastic moment maps and the exact flow maps, using the same interval fractions. Exact-flow suffix maps contract the nuclear norm. Summing the local errors and dividing by two gives (34). Taking infima preserves a uniform error radius. The deterministic infimum equals balanced static risk when $r \ge 1$, even over the larger class of all measurable schedules; applying the radius to the two stochastic competitor classes proves (35). The below-boundary statement follows by subtracting the two risk errors for the candidate and its best static comparator. □

At fixed B this bound vanishes as $\eta \to 0$, for any fixed batch size, provided the assumptions hold along that limit. It is deliberately conservative: it provides a uniform tolerance over schedules, not a fitted description of the observed boundary displacement. Increasing batch size alone removes within-source and per-example composition variance, but leaves Euler error and, for whole-batch selection, source-choice variance. This separates the three limits that the phase diagram tests.

H PROTOCOLS FOR THE STRUCTURAL AND BOUNDARY-FOCUSED EXPERIMENTS

The structural and boundary experiments isolate three questions that aggregate schedule wins cannot answer: whether a reoptimized advantage survives a perturbation, whether a candidate nearly attains an all-schedule bound, and whether stochastic deviations from the deterministic boundary disappear in a controlled limit. These experiments use exact matrix or moment dynamics except for the explicitly identified Monte Carlo validation. Their grids are model configurations, not samples from an application population.

H.1 PERTURBATIONS WITH BOTH COMPETITORS REOPTIMIZED

In the bisector basis use angles $30°, 45°, 65°$, $B = 2, 4$, $\rho = 0$, and $r = \varepsilon/e^{-2B\cos\phi} \in \{.25, .75, 1, 1.25, 2, 4\}$. At each point apply $\delta \in \{0, 10^{-3}, .01, .05, .1\}$ in four ways:

Initial cross moment:	$S = \left(\begin{smallmatrix} 1 & \delta\sqrt{\varepsilon} \\ \delta\sqrt{\varepsilon} & \varepsilon \end{smallmatrix}\right),\ Q = I;$
Unequal strengths:	$G_A = (1+\delta)G_{A,0},\ G_B = (1-\delta)G_{B,0};$
Target cross term:	$Q = \left(\begin{smallmatrix} 1 & \delta \\ \delta & 1 \end{smallmatrix}\right),\ S = \mathrm{diag}(1, \varepsilon);$
Aligned target:	$Q = \mathrm{diag}(1+\delta, 1-\delta),\ S = \mathrm{diag}(1, \varepsilon).$

Unmentioned matrices remain nominal. All second moments and targets are positive definite on this grid. There are 720 records, including repeated nominal controls at $\delta = 0$ across perturbation types. For each record, static risk is reoptimized and globally bracketed using the method below. A free-duration two-phase candidate is then searched with the limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm with bound constraints (L-BFGS-B) and analytical matrix-exponential derivatives. Its seven starts are the best static fraction, increasing and decreasing weights, pure A then B with duration fractions .15, .3, .7, and one seeded random start. The best static schedule is always included. The random seed is the fixed record index.

A 72-case subset ($45°$, $B = 2$, nonzero sizes .01, .05, .1) additionally searches eight equal-duration phases, using four starts. The at-most-eight candidate set includes the free-duration two-phase solution, since eight equal-duration phases alone do not contain every such solution. No random geometry or initial-error draw is used for these population risks.

All-weight static brackets. For $E(p) = e^{-BG(p)}$, Duhamel differentiation bounds $\|E'\| \le B\|G_A - G_B\|e^{-mB}$ and $\|E''\| \le B^2\|G_A - G_B\|^2 e^{-mB}$ on a cell whose generators have eigenvalues at least m. It follows that

$$|R''(p)| \le 2\|Q\|_{\mathrm{op}}\,\mathrm{tr}(S)B^2\|G_A - G_B\|_{\mathrm{op}}^2 e^{-2mB}.$$

Concavity of the minimum eigenvalue makes the smaller endpoint eigenvalue a valid cellwise m. Scalar interpolation bounds the cell minimum below by the smaller endpoint risk minus this curvature times cell width squared divided by eight. An adaptive priority queue refines only cells that could improve the best evaluated risk. Local optimization supplies an upper bound only. All 720 brackets reach relative width below 2.1×10^{-10}. Small floating-point padding is a guard, not outward-rounded interval arithmetic.

H.2 FINITE-BUDGET CANDIDATES AND THE GLOBAL CEILING

The ceiling grid uses the same three angles, $B = 1, 2, 4, 8$, and $r = 0, .01, .1, .25, .5, .75, .9, 1, 2$, totaling 108 cases. Two-phase search uses nine starts; eight equal-duration phases use five. Every gain ceiling includes the volume, energy, and slow-column bounds, and is set to one in the exact no-benefit region. The following slice illustrates both tight and loose regimes:

Angle, budget	r	Two-phase gain	All-schedule ceiling	Candidate/lower risk
$45°, 4$	0	1.89	3.23	1.70
$45°, 4$	.01	1.85	3.26	1.76
$45°, 4$	.10	1.53	1.74	1.13
$45°, 4$	.25	1.25	1.25	1.00
$45°, 4$	.50	1.06	1.06	1.00
$65°, 8$	0	25.46	101.40	3.98

Equality after rounding is not proof that a displayed numerical schedule is exactly optimal. It supplies a small value bracket, while Theorem 3 proves exact attainment on a possibly smaller local interval. Among the 84 below-boundary configurations, the median candidate/lower-risk ratio is 1.04. The maximum over the grid is 3.98. Eight-phase local searches did not improve the two-phase candidates by 0.1% in any of the 108 cases; this does not rule out an undiscovered better schedule.

Figure 2 in the main text shows the same quantity at fixed budget across all three angles.

The main fixed-ε curve uses 45°, $\varepsilon = .01$, and 80 evenly spaced budgets in $[.25, 8]$. It records the preparation rule, a reoptimized two-phase candidate, the combined ceiling, and the volume-only ceiling. The energy bound is particularly important at low uncertainty and short budgets; omitting it would inflate the stated uncertainty about attainable gain.

H.3 Exact SGD phase grid and adaptive comparator accuracy

The 2,340-condition grid is specified in Section 6. There are 702 cases for each sampling convention and 234 population-gradient controls. Every source shares the same linear target, and initial moments in the bisector basis are $S_0 = \mathrm{diag}(1, re^{-2B\cos\phi})$. Source strengths are one, the ridge is zero, and label noise is zero in this boundary grid. The normalized work is $B = N\eta$; N and batch size are not silently counted as equal compute across different protocols.

The dimensionless amplitude γ is a fraction of the maximum off-diagonal magnitude $h = \sqrt{ad}$, not a mixture fraction or an unnormalized coupling. The exact grid is

$$\Gamma = \{50^{j/6}/500 : j = 0, \dots, 6\} \cup \{3/20 + 17j/220 : j = 0, \dots, 11\}.$$

For $r_w = [2\cosh(cB/2) - 1]^{-1}$ the first-order cancellation schedule is

$$z_1 = h\gamma r_w, \quad z_2 = -h\gamma, \qquad p_1 = (1 - \gamma r_w)/2, \quad p_2 = (1 + \gamma)/2. \tag{36}$$

Each weight is used for $N/2$ updates, and (p_2, p_1) is a second candidate. The twelve-point linear grid includes $5/22$. For stratification we replace each p_j separately by $\mathrm{round}(mp_j)/m$, using ties-to-even rounding; this need not preserve the unrounded exposure exactly. The pure-phase library uses every integer switch time from zero to N, in both source orders. The minimum risk is selected in enumeration order (first occurrence for exact ties). The best static point replaces it only if strictly better. The mixed library is not a continuous optimization over all two-phase schedules; it is intentionally reproducible and is never called the optimal SGD curriculum.

The moment static bracket uses the analytical curvature bound in Appendix F. Initial grids have 2,049 or 8,193 points. For stratified batches its upper risk is minimized over all $m + 1$ legal fractions; its lower bracket also allows the algebraic positive-map relaxation at noninteger mp. We adaptively refine every record with $r \geq .75$, and all primary-figure records (45°, $B = 2$, $m = 512$), for 1,332 refined records in total. Their relative widths are below 2.1×10^{-10}; all remaining brackets have width below 10^{-3}. This matters because some above-boundary gains are much smaller than 0.1%. Positive tiny gains discussed in the main text exceed the refined lower bracket, not merely a coarse best grid point. All underlying moment maps are checked to be trace-nonexpansive.

We also evaluate each selected SGD schedule under exact flow with the same fractions and duration, and compare it with an exposure-matched static schedule. Thus a favorable noisy-training comparison can be distinguished from the same trajectory's deterministic effect. Per-example and whole-batch sampling retain different source-count noise, as quantified in Proposition 14. The stratified comparison removes that term; population-gradient controls additionally remove within-source randomness.

H.4 Exposure-matched alternation and independent sampled training

The alternation grid holds $\phi = 45^\circ$, $B = 2, 4$, $r = .1, .5, 1, 2$, and the same three step sizes and batch sizes. Every pure-source schedule uses exactly $N/2$ updates from each source. Dwell times are $N/2$, $N/8$, or one update, giving 216 records. Their risks are compared with balanced per-example and stratified mixtures and with the corresponding flow trajectories. Alternation changes order without changing total source use; decreasing its physical dwell time tests convergence to balanced flow.

Monte Carlo validation uses 8,192 independent trajectories in each of 16 conditions: step .05 or $1/80$, batch 8 or 32, orthogonal mass .01 or .2, and either balanced sampling or two mixed half-duration

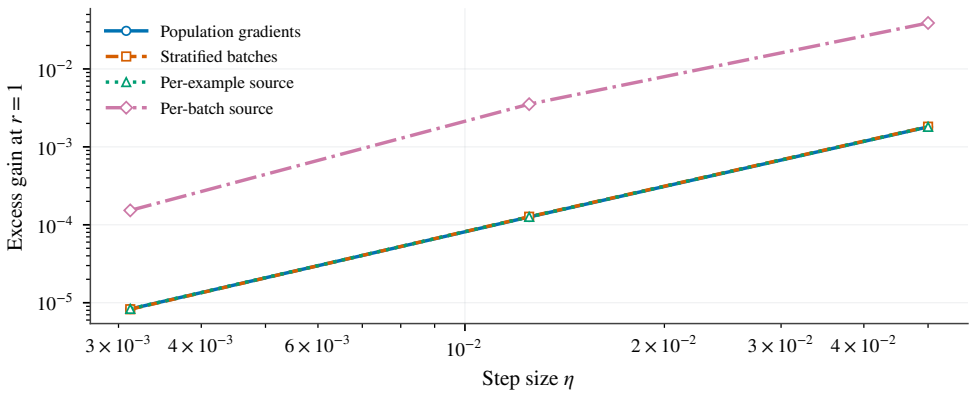


Figure 5: Excess candidate gain at the deterministic boundary, $r = 1$, for $45°$, $B = 2$, $m = 512$. Whole-batch source choice creates a larger discrepancy that persists when batch size alone increases. Reducing the step size removes it; the other conventions approach the population-gradient discretization effect. The plotted values are resolved by the refined static brackets.

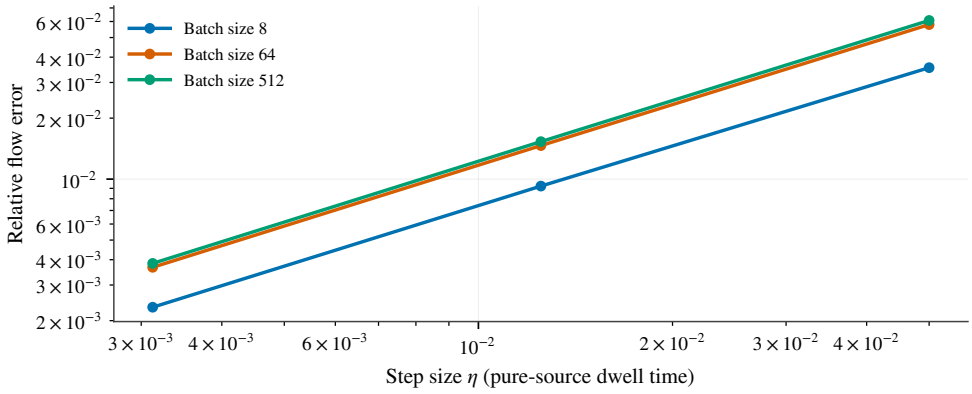


Figure 6: One-update pure-source alternation approaches balanced flow at fixed $50/50$ exposure. The slice has $45°$, $B = 2$, $r = .5$. This checks ordering at fixed exposure, separately from changing random batch composition. Across the full $m = 512$ grid, the maximum relative error falls from 0.13 to 9.46×10^{-3} as the step decreases.

phases with fractions $.25, .75$. Here $B = 2$, source angle is $45°$, and label standard deviation is $.1$. Actual fixed-composition Gaussian minibatches are generated independently across updates. The initial error is $(1, \sqrt{\varepsilon}Z)$ with Z standard normal, independent of training data. Its second moment equals that used in the exact recursion. The largest absolute standardized discrepancy between a sample mean and exact risk is 2.81; all 16 are below three standard errors. This tests the stratified map, not just an algebraic comparison between implementations of that map.

Numerical scope. Ordinary float64 arithmetic and seeded local searches are used. Reported risks are exact-model expectations or explicitly labeled Monte Carlo estimates. Tests cover matrix derivatives, target-weighted boundaries, cross-moment lower bounds, optimized perturbation comparisons, composition identities, exact local attainment, and moment-to-flow bounds. They support implementation checks, not independent verification of the mathematical proofs.

H.5 Consistency of reported risks and an independent precision check

Table 1 and its quoted values use the same exact-risk calculations as Figures 4 and 5. At the finest step, the excess gains are $8.31 \times 10^{-4}\%$ (population), $8.25 \times 10^{-4}\%$ (stratified), and $8.34 \times 10^{-4}\%$

(per example). The batch-source value is $1.53 \times 10^{-2}\%$. All comparisons use the full numerical precision before display rounding.

A direct check requiring no static-weight search. For population-gradient Euler updates, let $M(p) = I - \eta G(p)$. When $\eta(1+\rho) \le 1$, the scalar function $f(\lambda) = (1-\eta\lambda)^{2N}$ is convex on the stable eigenvalue interval. Spectral Jensen and the common diagonals therefore prove that $p = 1/2$ is an exact best static comparator, with

$$R_{\mathrm{stat},\eta} = \tfrac{1}{2}\left[(1-\eta(\rho+a))^{2N} + \varepsilon(1-\eta(\rho+d))^{2N}\right].$$

For $\phi = 45°$, $B = 2$, $\rho = 0$, $r = 1$, and $\eta = 1/320$, we have $N = 640$. The winning mixed amplitude is $\gamma = 5/22$ and

$$p_1 = (1-(5/22)r_w)/2, \qquad p_2 = 27/44.$$

With 320 updates in each phase, direct 80-digit powers of $M(p_2)^{320}M(p_1)^{320}$ give

$$\begin{aligned} R_{\mathrm{stat},\eta} &\approx 3.28 \times 10^{-2}, \\ R_{\pi} &\approx 3.28 \times 10^{-2}, \\ 100(R_{\mathrm{stat},\eta}/R_{\pi} - 1) &\approx 8.31 \times 10^{-4}\ \%. \end{aligned}$$

The two risks agree at the displayed precision; their difference is 2.73×10^{-7}. For stratification, the selected fractions are $218/512$ and $314/512$. Per-example sampling selects the same unrounded fractions as population gradients. Whole-batch sampling selects $\gamma = 101/220$, with second fraction $321/440$. Exact fractions are evaluated before any displayed rounding.

Independent numerical basis and scope. The independent precision checker rebuilds the second-moment maps in coordinates (S_{11}, S_{12}, S_{22}), rather than the orthonormal coordinates used in the float64 calculation. It uses the distributional fourth-moment identities directly, without reusing the float64 map coefficients. At 80 digits it checks every pure switch and all 38 mixed candidates for all twelve Table 1 conditions, and compares the selected schedules and risks with the float64 evaluations of the same conditions. The population winner is additionally checked by direct 2×2 matrix powers. These checks do not remove the finite-candidate limitation or turn the stochastic static brackets into directed-rounding certificates. The interval certificate in Section E.1 is a separate computation with separate arithmetic guarantees.

Calibrating the stochastic radius in risk units. At that population-gradient point the nominal flow static risk is 3.29×10^{-2}. Equation (15) gives $2E_{\mathrm{SGD}} \approx 6.62 \times 10^{-3}$, or 20.12% of that risk. The actual candidate improvement relative to best static Euler training is only 2.73×10^{-7}, approximately $24{,}259$ times smaller than the bound. The uniform $O(\eta)$ radius is a sufficient convergence statement over all admitted schedules. It does not quantitatively explain the observed boundary displacement or imply the apparently faster decay of this finite library's excess gain. Sampling noise, discretization, and candidate selection remain distinct sources of that behavior.

I BASIC COMPARISON PROOFS

Notation. All matrix norms without a subscript are Euclidean operator norms, except $\|\cdot\|_F$ and $\|\cdot\|_*$. Risk is one half of the stated quadratic form. The budget variable is B, while $d = (1-\cos\phi)/2$ is the slow eigenvalue of the balanced, unregularized mixture. A second moment need not be a centered covariance.

Commuting, isotropic, and condition-number comparisons. If G_A and G_B commute, all their convex combinations commute. A product of exponentials is the exponential of their sum. The same identity holds for measurable schedules by simultaneous diagonalization and solving each scalar equation. Its effective mixture weight is $B^{-1}\int_0^B p(t)\,dt$ for $B > 0$; the zero-budget statement is immediate.

For two phases let $X = b_1G_1$, $Y = b_2G_2$, $\Phi = e^{-Y}e^{-X}$, and $\Psi = e^{-(X+Y)}$. Cyclicity of trace gives

$$\|\Phi\|_F^2 = \mathrm{tr}(e^{-2X}e^{-2Y}).$$

Golden–Thompson, applied to the symmetric matrices $-2X$ and $-2Y$, gives $\|\Phi\|_F^2 \geq \operatorname{tr}(e^{-2(X+Y)}) = \|\Psi\|_F^2$. The exponent $X+Y$ equals $BG(\bar{p})$, with $\bar{p} = (b_1 p_1 + b_2 p_2)/B$. Thus Ψ is a feasible static comparator. Since static schedules belong to the two-phase class, their optimized risks coincide in the isotropic setting.

For positive definite S, Q, denote their smallest and largest eigenvalues by s_-, s_+, q_-, q_+. For every matrix M,

$$\tfrac{1}{2} q_- s_- \|M\|_F^2 \leq \tfrac{1}{2} \operatorname{tr}(QMSM^\top) \leq \tfrac{1}{2} q_+ s_+ \|M\|_F^2 .$$

Applying the upper bound to the exposure-matched static transition and the lower bound to the curriculum gives $\mathcal{R}(\Psi) \leq \kappa(Q)\kappa(S)\mathcal{R}(\Phi)$. The optimal static risk is at most $\mathcal{R}(\Psi)$ for each curriculum. Taking the infimum over two-phase curricula proves the ratio bound. No product inequality with three or more independent exponents is used. □

J Exact static optimum and preparation formula

Lemma 16 (Common-diagonal static benchmark)**.** *Suppose symmetric generators G_i have the same diagonal $(\gamma_1, \dots, \gamma_D)$ in an orthonormal basis (u_j). Suppose their convex hull contains $G_0 = \sum_j \gamma_j u_j u_j^\top$, and $S = \sum_j s_j u_j u_j^\top$ with $s_j \geq 0$. For $Q = I$, G_0 globally minimizes static risk at every $B \geq 0$, with value $\frac{1}{2}\sum_j s_j e^{-2B\gamma_j}$.*

Proof. For a static generator G with eigenpairs (λ_k, z_k),

$$u_j^\top e^{-2BG} u_j = \sum_k (u_j^\top z_k)^2 e^{-2B\lambda_k} \geq \exp\left(-2B \sum_k (u_j^\top z_k)^2 \lambda_k\right) = e^{-2B\gamma_j}.$$

The weights are nonnegative and sum to one. The inequality is scalar Jensen for the convex function $x \mapsto e^{-2Bx}$. Summing against $s_j/2$ gives the lower bound. The diagonal generator G_0 attains it. This is a direct spectral Jensen consequence; operator convexity of the exponential is neither assumed nor needed. □

Proof of Proposition 1. In the (u, v) basis, $G(p) - \rho I$ has the form

$$\begin{pmatrix} a & z(p) \\ z(p) & d \end{pmatrix}, \qquad z(p) = (1-2p)\sqrt{ad}. \tag{37}$$

Indeed $u^\top v_A = u^\top v_B = \sqrt{a}$, whereas $v^\top v_A = -\sqrt{d}$ and $v^\top v_B = \sqrt{d}$. The diagonal is independent of p, and the off-diagonal vanishes at $p = 1/2$. Apply Lemma 16 with masses $1, \varepsilon$. □

Preparation notation. For the family in (3), put $\tau = \log((1+c)/c)$, $A_\phi = a/c^2$, $D_\phi = d(1+2c)^2/c^2$, and $F_\phi = c^2/a$. Write $S_\varepsilon = uu^\top + \varepsilon vv^\top$. For $B \geq \tau$, the preparation schedule uses source A for τ and source B for $B - \tau$; its risk is $\mathcal{R}_{\mathrm{prep}}$.

Derivation of the preparation risk. The common ridge commutes with all terms, so every transition is multiplied by $e^{-\rho B}$. It suffices to compute at $\rho = 0$. In the original source basis, $u = (\sqrt{a}, \sqrt{d})^\top$ and $v = (-\sqrt{d}, \sqrt{a})^\top$. Since $e^{-\tau} = c/(1+c) = c/(2a)$,

$$e^{-\tau v_A v_A^\top} u = \begin{pmatrix} c/(2\sqrt{a}) \\ \sqrt{d} \end{pmatrix} = \frac{v_B}{2\sqrt{a}}.$$

The identity uses $\sin\phi = 2\sqrt{ad}$. Thus the squared norm after the second phase is

$$\left\| e^{-(B-\tau) v_B v_B^\top} e^{-\tau v_A v_A^\top} u \right\|^2 = \frac{e^{-2(B-\tau)}}{4a} = \frac{a}{c^2} e^{-2B} = A_\phi e^{-2B}.$$

For the orthogonal vector, define $w = e^{-\tau v_A v_A^\top} v = (-c\sqrt{d}/(2a), \sqrt{a})^\top$ and $v_B^\perp = (-\sin\phi, \cos\phi)^\top$. Direct inner products give

$$(v_B^\top w)^2 = \frac{d(1+2c)^2}{(1+c)^2}, \qquad ((v_B^\perp)^\top w)^2 = \frac{c^2}{a} = F_\phi.$$

The parallel component is multiplied by $e^{-2(B-\tau)}$ and the perpendicular component is unchanged. Therefore its final squared norm is $D_\phi e^{-2B} + F_\phi$. Weighting the two initial components by 1 and ε, dividing by two, and restoring $e^{-2\rho B}$ gives

$$\mathcal{R}_{\mathrm{prep}}(B,\varepsilon) = \tfrac{1}{2} e^{-2\rho B}\left[(A_\phi + \varepsilon D_\phi)e^{-2B} + \varepsilon F_\phi\right]. \tag{38}$$

□

Proof of the rate and work claims. At $\varepsilon = 0$, dividing (5) by (38) gives $F_\phi e^{2(1-a)B} = F_\phi e^{2dB}$. Strict improvement is equivalent to $2dB > \log A_\phi$ whenever the preparation schedule is feasible.

For every admissible allocation, $G(p(t)) \preceq (1+\rho)I$. The scalar energy derivative satisfies

$$\frac{d}{dt}\|e(t)\|^2 = -2e(t)^\top G(p(t))e(t) \geq -2(1+\rho)\|e(t)\|^2.$$

Integration gives $\|e(B)\|^2 \geq e^{-2(1+\rho)B}\|e(0)\|^2$. The initial vector u has unit norm. This proves the lower bound for arbitrary measurable schedules; the explicit schedule supplies the upper bound. The bounds imply the same exponential rate for $\mathcal{R}_2^*$ and $\mathcal{R}_{\mathrm{all}}^*$.

Solving the two explicit risk formulas for their first crossing of a sufficiently small ℓ gives the required works $\log(1/(2\ell))/(2(\rho+a))$ for static mixing and $\log(A_\phi/(2\ell))/(2(\rho+1))$ for preparation. The preparation crossing eventually exceeds τ. For the optimal arbitrary schedule, the lower and upper rate bounds bracket the required work between $\log(1/(2\ell))/(2(1+\rho))$ and $\log(A_\phi/(2\ell))/(2(1+\rho))$. Their difference is constant in ℓ, so they give the same limiting static-to-curriculum work ratio. □

K Preparation-rule uncertainty and its peak law

Derivation of the preparation uncertainty threshold. Subtracting the two exact formulas yields

$$2e^{2\rho B}(\mathcal{R}_{\mathrm{stat}}^* - \mathcal{R}_{\mathrm{prep}}) = N(B) - \varepsilon K(B), \qquad K(B) = F_\phi + D_\phi e^{-2B} - e^{-2dB}.$$

At $\varepsilon = 1$, the isotropic comparison in Appendix I implies $N(B) - K(B) \leq 0$, since the preparation curriculum is a two-phase schedule and the balanced static mixture is optimal. If $N(B) > 0$, it follows that $K(B) \geq N(B) > 0$, and strict improvement is exactly $\varepsilon < N(B)/K(B)$. This threshold is at most one. As $B \to \infty$, $N(B) \sim e^{-2aB}$ and $K(B) \to F_\phi > 0$, proving the asymptotic expression. If $N(B) \leq 0$, the difference is nonpositive both at $\varepsilon = 0$ and at $\varepsilon = 1$. Affinity in ε proves nonpositivity throughout $[0,1]$. No assertion for $\varepsilon > 1$ is required for this threshold statement. □

Lemma 17 (Slow-eigenvector cone)**.** *For all $p \in [0,1]$, the smaller eigenvalue of $G(p)$ in* (3) *is at most $\rho + d$. Its unit slow eigenvector can be chosen to make angle at most $\phi/2$ with v, and slow eigenvectors for any two mixtures have squared inner product at least c^2.*

Proof. In (37), $a - d = c > 0$ and $|z(p)| \leq \sqrt{ad} = \sin\phi/2$. The smaller eigenvalue is

$$\mu_-(p) = \rho + \frac{1 - \sqrt{c^2 + 4z(p)^2}}{2} \leq \rho + \frac{1-c}{2} = \rho + d.$$

Diagonalization rotates the principal axes by an angle ψ satisfying $|\tan(2\psi)| = 2|z|/c \leq \tan\phi$, with $|2\psi| < \pi/2$. Hence $|\psi| \leq \phi/2$. The slow axes lie in an angular cone of width $\phi < \pi/2$, and consistently oriented unit vectors have inner product at least $\cos\phi = c$. □

Proposition 18 (Arbitrary-stage volume obstruction)**.** *For* (3)*, $\varepsilon > 0$, and every measurable schedule,*

$$\mathcal{R}(\pi) \geq \sqrt{\varepsilon}\, e^{-(1+2\rho)B}. \tag{39}$$

Proof. The trace of every generator is $1 + 2\rho$. Liouville's determinant formula gives $\det \Phi_\pi = e^{-(1+2\rho)B}$. The two eigenvalues of $\Phi_\pi S_\varepsilon \Phi_\pi^\top$ are positive, with product $\varepsilon e^{-2(1+2\rho)B}$. Arithmetic–geometric mean bounds their sum below by twice the square root of this product. Dividing by two proves the result. □

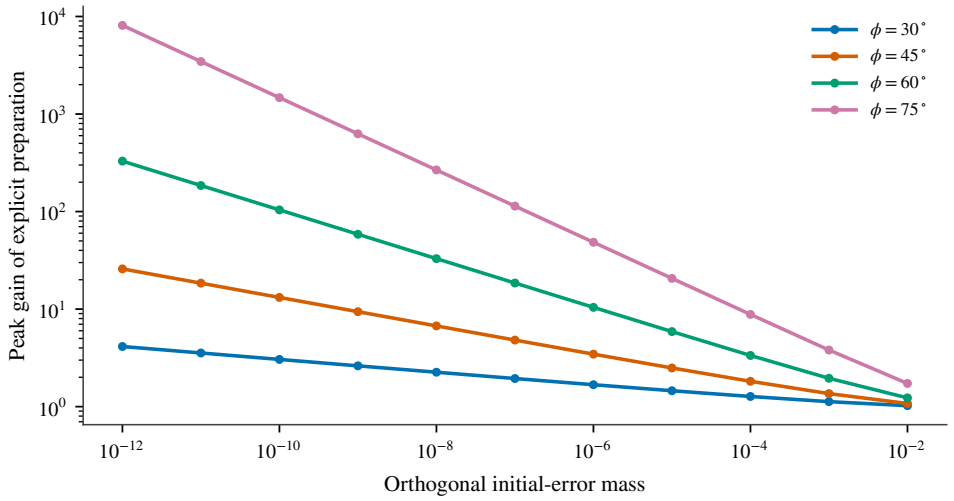


Figure 7: The peak preparation gain grows as orthogonal uncertainty decreases. The exponent is $d = (1 - \cos\phi)/2$, so the gain can increase slowly when the source directions are close. The plotted maxima are one-dimensional numerical evaluations of the exact formulas; the peak-gain asymptotic above is derived analytically.

Derivation of the peak preparation gain. Put $x = e^{-2B}$, $A_\varepsilon = A_\phi + \varepsilon D_\phi$, and suppress the fixed angle subscript on F. The ratio is

$$\frac{x^a + \varepsilon x^d}{A_\varepsilon x + \varepsilon F}, \qquad 0 < x \le e^{-2\tau}, \quad a + d = 1.$$

For $f(x) = x^a/(A_\varepsilon x + \varepsilon F)$, differentiation shows its unique unrestricted maximum is at

$$x_* = \frac{a\varepsilon F}{dA_\varepsilon}, \qquad f(x_*) = a^a d^d A_\varepsilon^{-a} F^{-d} \varepsilon^{-d}.$$

For sufficiently small ε, x_* lies in the admissible interval. The second summand $g(x) = \varepsilon x^d/(A_\varepsilon x + \varepsilon F)$ is uniformly $O(\varepsilon^d)$: substitute $x = \varepsilon t$ and use boundedness of $t^d/(A_\varepsilon t + F)$ on $t > 0$, uniformly as $\varepsilon \downarrow 0$. Consequently the maximum of $f + g$ differs from the maximum of f by at most $O(\varepsilon^d)$. Since $F = A_\phi^{-1}$ and $a - d = c$, the leading constant converges to $a^a d^d A_\phi^{-a+d} = a^a d^d A_\phi^{-c}$. □

L SPECTRAL AND CAPACITY-WORK PROOF

Lemma 19 (Power-law exponential tail sum)**.** *For fixed $\alpha, \beta, k > 0$,*

$$\sum_{j \ge 1} j^{-1-\beta} e^{-kBj^{-\alpha}} \asymp B^{-\beta/\alpha} \quad (B \ge 1).$$

The same conclusion holds when the weights and scales are comparable, with fixed positive constants, to $j^{-1-\beta}$ and $j^{-\alpha}$.

Proof. Write $J_B = \lceil B^{1/\alpha} \rceil$. For $j \ge J_B$, the exponential is bounded below by a positive constant depending only on k, α, and the tail sum is comparable to $J_B^{-\beta}$. This proves the lower bound. For the upper bound, the same tail is at most a constant times $J_B^{-\beta}$. Partition $1 \le j < J_B$ into dyadic bands with j between $2^{-l-1}J_B$ and $2^{-l}J_B$, omitting empty bands. The total weight of a band is at most a constant times $J_B^{-\beta} 2^{(l+1)\beta}$, while its exponential is at most $\exp(-k'2^{l\alpha})$ for a fixed $k' > 0$. The series over bands converges independently of B, because exponential decay in $2^{l\alpha}$ dominates the geometric factor. Comparability changes only constants. □

Proof of the spectral comparison in Section 7. Use the assumptions stated there: $w_j = \operatorname{tr} S_j \asymp j^{-1-\beta}$, $\lambda_j \asymp j^{-\alpha}$, every source satisfies $0 \preceq G_{i,j} \preceq L\lambda_j I$, and one common static mixture satisfies $G(p_0)_j \succeq m\lambda_j I$, with $\alpha, \beta, L, m > 0$ and identity evaluation. Capacity J trains the first J blocks and leaves the remaining mass unlearned. The notation $\mathcal{R}^*$ below covers either shared or independent block schedules. For every trained block and its own arbitrary measurable schedule $p_j(t)$, the energy differential inequality gives

$$\|e_j(B)\|^2 \geq e^{-2L\lambda_j B} \|e_j(0)\|^2 .$$

Expectation gives the same bound with initial mass w_j. The proposed static mixture gives the upper bound $\mathbb{E}\|e_j(B)\|^2 \leq e^{-2m\lambda_j B} w_j$. Hence

$$\tfrac{1}{2}\sum_{j\leq J} w_j e^{-2L\lambda_j B} + \tfrac{1}{2}\sum_{j>J} w_j \leq \mathcal{R}^*(B,J) \leq \tfrac{1}{2}\sum_{j\leq J} w_j e^{-2m\lambda_j B} + \tfrac{1}{2}\sum_{j>J} w_j.$$

The lower expression is at least the full exponential sum, because each untrained tail term is at least its exponentially damped version. It is also at least the untrained mass. Lemma 19 and $\sum_{j>J} w_j \asymp J^{-\beta}$ therefore give a constant times $B^{-\beta/\alpha} + J^{-\beta}$ as a lower bound, using $\max(x,y) \geq (x+y)/2$. The upper expression is at most the full exponential sum plus the tail, proving the corresponding upper bound. The lower inequality therefore also holds for independently chosen block schedules, while the common static mixture supplies the upper inequality for the more restrictive shared class. No simultaneous realization of separate optimal block schedules is needed.

For total work $JB \leq C$, the lower bound gives, up to constants,

$$B^{-\beta/\alpha} + J^{-\beta} \geq B^{-\beta/\alpha} + (B/C)^{\beta}.$$

The two terms balance at $B \asymp C^{\alpha/(\alpha+1)}$. Either side of that scale is bounded below by a constant times $C^{-\beta/(\alpha+1)}$. Taking J to be an integer within a constant factor of $C^{1/(\alpha+1)}$ and using the common static mixture gives the matching upper bound. The cost counts a fixed capacity during training; it is not a result about dynamically activating features. □

M Perturbation, discretization, and statistical bounds

Lemma 20 (Uniform transition perturbation)**.** *Suppose two time-dependent symmetric positive semidefinite generators differ in operator norm by at most δ_G almost everywhere on $[0,T]$. Their transitions satisfy $\left\|\Phi(T) - \widehat{\Phi}(T)\right\| \leq \min\{2, T\delta_G\}$.*

Proof. All subinterval transitions are contractions. Variation of constants gives

$$\Phi(T) - \widehat{\Phi}(T) = -\int_0^T \Phi(T,s)(G(s) - \widehat{G}(s))\widehat{\Phi}(s,0)\, ds.$$

The integral norm is at most $T\delta_G$. The triangle inequality and contraction bounds also give two. For source estimates with per-source errors bounded by δ_G, every common convex allocation has an error at most δ_G. Thus this result is simultaneous over schedules. □

Uniform risk and oracle bounds. Let $S, \widehat{S} \succeq 0$ satisfy $\left\|S - \widehat{S}\right\|_* \leq \delta_S$, where $\|\cdot\|_*$ is the nuclear norm, and let the true and estimated generators satisfy the assumptions of Lemma 20. For $Q = I$, split the risk difference by first replacing S by $\widehat{S}$ and then replacing Φ by $\widehat{\Phi}$. The covariance term has absolute value at most $\delta_S/2$, since $\left\|\Phi^\top \Phi\right\| \leq 1$. If $\eta_\Phi = \left\|\Phi - \widehat{\Phi}\right\|$, then

$$\left\|\Phi^\top\Phi - \widehat{\Phi}^\top\widehat{\Phi}\right\| \leq (\|\Phi\| + \left\|\widehat{\Phi}\right\|)\eta_\Phi \leq 2\eta_\Phi.$$

Because $\widehat{S} \succeq 0$, the remaining risk difference is at most $\operatorname{tr}(\widehat{S})\eta_\Phi$. Apply Lemma 20 to obtain $|\mathcal{R}(\pi) - \widehat{\mathcal{R}}(\pi)| \leq r_T := \delta_S/2 + \operatorname{tr}(\widehat{S}) \min\{2, T\delta_G\}$.

If an empirical library minimizer has tolerance η, compare its true risk to its estimated risk, use empirical optimality, and compare the selected reference's estimated risk to its true risk. The two estimation differences contribute at most $2r_T$, proving the oracle inequality. An infimum can replace an attained optimum by an arbitrarily small additional tolerance. All terms use the same post-pilot budget T. □

Robustness of a strict comparison. For a nominal schedule with gap $\Delta = \mathcal{R}^*_{\mathrm{stat},0}(B) - \mathcal{R}_0(\pi) > 0$, a uniform radius r_B for both schedule and static risks gives $\mathcal{R}^*_{\mathrm{stat}}(B) - \mathcal{R}(\pi) \geq \Delta - 2r_B$. Thus every positive nominal gap has a nonempty neighborhood of source/second-moment perturbations in which the comparison persists. This is an absolute-risk statement. Since the gap can be exponentially small, it is not a uniform large-budget robustness guarantee.

Lemma 21 (Euler-to-flow bound at a specified schedule)**.** *Suppose every step uses a symmetric positive semidefinite generator with norm at most L, the common step size is η with $\eta L \leq 1$, and there are N steps, with total work $B = N\eta$. The Euler transition product and the exact piecewise-constant flow with the same stepwise allocation differ by at most $B\eta L^2/2$ in operator norm.*

Proof. For $x \in [0,1]$, $0 \leq e^{-x} - (1-x) \leq x^2/2$. Spectral calculus therefore bounds each one-step discrepancy by $\eta^2 L^2/2$. Both the Euler factor and the exponential factor have norm at most one. A telescoping product identity bounds the total discrepancy by the sum of one-step discrepancies, namely $N\eta^2 L^2/2$. The bound compares flow and Euler with identical stepwise allocations; changing a phase boundary is a separate schedule perturbation. □

Gaussian observation event. For n independent $N(e_j, \sigma^2)$ observations of each coordinate of an unknown deterministic residual $e \in \mathbb{R}^D$, let $\widehat{e}$ be the vector of sample means. Then $\widehat{e} - e$ has law $(\sigma/\sqrt{n})Z$ with $Z \sim N(0, I_D)$. Gaussian concentration for the 1-Lipschitz Euclidean norm gives $\Pr(\|Z\| \geq \mathbb{E}\|Z\| + t) \leq e^{-t^2/2}$; Jensen gives $\mathbb{E}\|Z\| \leq \sqrt{D}$. Choosing $t = \sqrt{2\log(1/\delta)}$ yields $\|e - \widehat{e}\| \leq r := (\sigma/\sqrt{n})(\sqrt{D} + \sqrt{2\log(1/\delta)})$ with probability at least $1 - \delta$. These are classical concentration tools (Boucheron et al., 2013). Candidate selection does not require a union bound when all candidate risks are bounded on this same state-ball event.

N Uniform static envelopes and the paid gate

This section derives a global lower comparator for a schedule selected from coordinate observations. It is not needed for the exact-family optimum. Consider known generators, fixed full budget B, and $E(p) = e^{-BG(p)}$. Partition $[0,1]$ into cells centered at p_q with half-width h_q. Let m_q be the minimum of the smallest eigenvalues of the two endpoint generators. Concavity of the smallest eigenvalue implies $G(p) \succeq m_q I$ throughout the cell. Duhamel's formula then gives

$$\|E(p) - E(p_q)\| \leq \eta_q := B\,\|G_A - G_B\|\, h_q e^{-Bm_q}. \tag{40}$$

Both generators along the interpolation have the stated lower bound, so the two exponential factors in the integral contribute e^{-Bm_q}. If true and estimated source generators differ by at most δ_G, one can add $\min\{2, B\delta_G\}$ to η_q by a triangle inequality.

Write $N = \|\widehat{e}\|$, $y_q = E(p_q)\widehat{e}$, $n_q = \|E(p_q)\|$, and $n_q^+ = \min\{1, n_q + \eta_q\}$. On the state ball $\|e - \widehat{e}\| \leq r$, the triangle inequality supplies the cell-wise lower bound

$$L_q^{\mathrm{amp}} = \tfrac{1}{2}\left[\|y_q\| - n_q r - \eta_q(N + r)\right]_+^2. \tag{41}$$

A usually sharper bound follows from convexity of $f_p(e) = \|E(p)e\|^2/2$:

$$f_p(e) \geq f_p(\widehat{e}) - r\left\|E(p)^2\widehat{e}\right\|.$$

The first term is bounded below by $\frac{1}{2}[\|y_q\| - \eta_q N]_+^2$. To bound the gradient norm, use

$$(E(p)^2 - E(p_q)^2)\widehat{e} = E(p)(E(p) - E(p_q))\widehat{e} + (E(p) - E(p_q))y_q.$$

Hence a second cell-wise lower bound is

$$L_q^{\mathrm{dir}} = \tfrac{1}{2}[\|y_q\| - \eta_q N]_+^2 - r\left(\|E(p_q)y_q\| + \eta_q(\|y_q\| + n_q^+ N)\right). \tag{42}$$

It follows that

$$L_s(B) = \max\left\{0, \min_q \max\{L_q^{\mathrm{amp}}, L_q^{\mathrm{dir}}\}\right\} \leq \inf_{p\in[0,1]} \tfrac{1}{2}\|E(p)e\|^2. \tag{43}$$

The bounds cover the whole cell rather than only its center, so optimizing on a grid does not invalidate the comparator.

Candidate upper bound. For $e = \widehat{e} + z$ with $\|z\| \le r$, expand

$$\tfrac{1}{2}\|\Phi e\|^2 = \tfrac{1}{2}\|\Phi\widehat{e}\|^2 + z^\top\Phi^\top\Phi\widehat{e} + \tfrac{1}{2}z^\top\Phi^\top\Phi z.$$

Cauchy–Schwarz gives the upper bound $U_\Phi = \frac{1}{2}\|\Phi\widehat{e}\|^2 + r\left\|\Phi^\top\Phi\widehat{e}\right\| + \frac{1}{2}r^2\|\Phi\|_{\mathrm{op}}^2$. The amplitude bound $\frac{1}{2}(\|\Phi\widehat{e}\| + r\|\Phi\|)^2$ is also valid, so the smaller of the two bounds can be used. If an estimated transition has error η_Φ, its directional bound can be enlarged by $\frac{1}{2}\eta_\Phi(2\left\|\widehat{\Phi}\right\| + \eta_\Phi)(N + r)^2$.

Decision and budget. If each coordinate observation costs c_{query} and planning costs h, the remaining training budget is $T = B - Dnc_{\mathrm{query}} - h$. Each candidate may have a different remaining training budget if its additional cost differs. A valid upper bound uses that actual remaining budget. The baseline lower bound always uses the full original B. Taking the smallest candidate upper bound is valid on the common confidence event. If it is below (43), the accepted candidate beats every full-budget static mixture. A static fallback after rejection has only T training work, not B, so a claim of unconditional no-regression would be false.

Arithmetic scope. These acceptance bounds are real-arithmetic guarantees. A numerical acceptance decision requires controlled rounding error; directed-rounding linear algebra or a proved floating-point error bound can supply it.