VALUING OPTIMIZATION GEOMETRY: INTRINSIC VALUE AND FINITE-STEP TRANSFER GUARANTEES

Anonymous authors
Paper under double-blind review

ABSTRACT

Structured preconditioners can substantially alter local optimization dynamics, but end-to-end optimizer comparisons do not isolate what a fixed geometry is worth at a particular training state. We study this local valuation problem for already-constructed, frozen preconditioning geometries. Under a stochastic quadratic model, optimizing over a global step magnitude yields a scale-invariant intrinsic value and a relative break-even interpretation. We then ask when this model-implied ordering survives a finite update. Under a Lipschitz Hessian, we derive an exact endpoint-wise transfer modulus in a cubic action metric d_3, yielding an attainable ranking frontier and a stochastic action-law transport bound. A preregistered five-trajectory CIFAR-10/ResNet-18 evaluation finds substantial state dependence: 14 of 20 states have resolved ratios, eight have resolved realized parity, and seven are formal reversals. A separate sealed prospective study does not establish unshrunk residual copying as a reliable repair. Together, the results separate intrinsic geometry value, finite-step transfer, and downstream correction, while motivating an uncertainty-aware robust decision rule derived after the prospective study and not evaluated by that protocol.

1 INTRODUCTION

Modern optimizers differ not only in their scalar step schedules but also in the geometry applied to stochastic gradients. Diagonal adaptive methods such as Adam are inexpensive; richer matrix and tensor preconditioners such as K-FAC, Shampoo, and SOAP spend additional computation to reshape the update (Kingma & Ba, 2015; Martens & Grosse, 2015; Gupta et al., 2018; Vyas et al., 2025). Existing work studies how these structures behave, when richer preconditioning helps, and how a preconditioner can be adapted or selected online (Xie et al., 2025; DiPaolo & Gu, 2019; Hu & Simchi-Levi, 2026; Richman & Wüthrich, 2026). We ask a different question that begins *after* candidate construction: given two frozen geometries at the same training state, what is each one locally worth, and when does that ordering survive the finite update that will actually be taken?

The first part is a valuation problem. Conditioning on the current stochastic training state, we optimize out a global step magnitude and assign each frozen geometry a scalar intrinsic value. The resulting quantity is invariant to arbitrary positive rescaling of the preconditioner, so it measures geometry rather than learning-rate units. The second part is a transfer problem. The local quadratic model may favor one geometry while higher-order curvature at a finite radius reverses the realized ordering. Our goal is to identify the information that separates reliable from unreliable transfer.

The empirical picture makes this distinction concrete. In a controlled quadratic benchmark, the Hessian-blind valuation pipeline tracks the constructed-candidate oracle price closely (Spearman 0.9926, log-price MAE 0.0070 over 1,728 calibrated states). We then test point-valued transfer on five independently randomized CIFAR-10/ResNet-18 trajectories at four predeclared checkpoints each. Fourteen states admit a realized ratio; eight have empirically resolved realized parity, and seven satisfy the preregistered formal reversal rule. Exact-model calibration therefore does not imply reliable finite-radius neural ordering.

Our theory characterizes the missing layer. The difference between candidates' higher-order remainders is controlled by a cubic metric on their *actual update actions*. Its exact endpoint modulus can be smaller than a straight-chord or separate-remainder bound because common distortion cancels. It yields an attainable deterministic ranking frontier and a stochastic bound on action laws; simultaneous

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.

finite-action envelopes separately address estimated quantities and data-dependent step selection. A prospective study then tests one operationalization: copying a calibrated residual from a geometry-selected anchor. It fails its directional criterion. A later decision analysis shows why full copying, although minimax for absolute prediction loss under a symmetric interval, is generally not minimax for excess squared loss relative to leaving the raw predictor unchanged.

The paper makes four contributions:

- **Intrinsic valuation.** We define a scale-invariant local value for an already-constructed stochastic preconditioning geometry and retain a Hessian-blind component estimator for the quadratic target.
- **Exact comparative transfer.** A fixed second-order jet does not identify finite-step ordering. With Hessian-Lipschitz information, we derive the endpoint-wise sharp cubic d_3 modulus and a stochastic action-law upper bound.
- **Decision consequences.** The modulus yields an attainable deterministic ranking frontier. Given a valid residual interval, an excess-loss minimax rule characterizes safe corrections, shrinkage, and abstention; it requires uncertainty information not supplied by point estimates.
- **Chronologically separate evidence.** A preregistered five-trajectory study finds interval-qualified reversals. A later sealed prospective study fails to establish full residual copying as a reliable repair; the uncertainty-aware decision rule was derived afterward and remains untested.

The aim is not to introduce another optimizer. It is to provide a valuation and decision layer for comparing frozen optimization geometries at the state where they would be used.

2 Intrinsic local value

Condition on all information available before the evaluation gradient is drawn. Let $w \in \mathbb{R}^d$ be the current parameter vector, $h = \nabla F(w)$, and $H = \nabla^2 F(w)$. A stochastic gradient g satisfies

$$\mathbb{E}[g] = h, \qquad \mathrm{Cov}(g) = \Omega.$$

A frozen preconditioner $P \succ 0$ is measurable with respect to the pre-evaluation information. Writing $Q(s) = -(h^\top s + \frac{1}{2} s^\top H s)$ for the quadratic Taylor gain, the expected predicted gain for update $s = -\eta P g$ is

$$G_P(\eta) = \eta A_P - \frac{\eta^2}{2} B_P, \qquad A_P = h^\top P h, \tag{1}$$

with

$$S_P = h^\top P H P h, \qquad N_P = \mathrm{tr}(PHP\Omega), \qquad B_P = S_P + N_P. \tag{2}$$

The trace form is equivalent to $\mathrm{tr}(HP\Omega P)$ by cyclicity. We allow the neural Hessian to be indefinite. The continuous value below is assigned only in the interior regime $A_P > 0$, $B_P > 0$. Outside it—in particular when $B_P \le 0$—the framework abstains from assigning an unconstrained continuous value; a separately valid predeclared finite action set may still be evaluated.

Proposition 1 (Intrinsic geometry value)**.** *If $A_P > 0$ and $B_P > 0$, then*

$$\eta_P^\star = \frac{A_P}{B_P}, \qquad V_P^{\mathrm{cont}} = \max_{\eta \ge 0} G_P(\eta) = \frac{A_P^2}{2B_P}. \tag{3}$$

For every scalar $c > 0$, $V_{cP}^{\mathrm{cont}} = V_P^{\mathrm{cont}}$, while $\eta_{cP}^\star = \eta_P^\star / c$. Hence the model-optimal actual displacement is unchanged by a global rescaling of P.

For two frozen candidates P_0, P_1, the intrinsic relative value is

$$q_{\mathrm{cont}} = \frac{V_{P_1}^{\mathrm{cont}}}{V_{P_0}^{\mathrm{cont}}}, \tag{4}$$

when both values are positive. If a richer implementation has per-step cost ratio $q_{\mathrm{impl}} = c_1/c_0$, then $q_{\mathrm{cont}} > q_{\mathrm{impl}}$ has the natural interpretation that the model-implied local gain premium exceeds the implementation price. This interpretation is deliberately local: the central question of the paper is when it transfers to finite neural progress.

2.1 HESSIAN-BLIND COMPONENT ESTIMATION

The population Hessian need not be formed explicitly. For the ideal curvature identity, suppose an oracle $Y(v)$ satisfies $\mathbb{E}[Y(v) \mid v] = Hv$ and all statistical roles use independent draws. For independent gradients $g_a, \dots, g_f$, define

$$\widehat{A}_P = g_a^\top P g_b, \tag{5}$$

$$\widehat{S}_P = (Pg_c)^\top Y(Pg_d), \tag{6}$$

$$\widehat{N}_P = (P\zeta)^\top Y(P\zeta), \qquad \zeta = (g_e - g_f)/\sqrt{2}. \tag{7}$$

Then $\mathbb{E}[\widehat{A}_P] = A_P$, $\mathbb{E}[\widehat{S}_P] = S_P$, and $\mathbb{E}[\widehat{N}_P] = N_P$. Components are averaged before forming $\widehat{V} = \bar{A}^2/(2(\bar{S} + \bar{N}))$; neither the nonlinear value nor its ratio is claimed unbiased. The neural implementation uses finite stochastic secants rather than an exact curvature oracle. Under a Lipschitz Hessian, a secant targets a path-averaged Hessian with controlled $O(|\delta|\|v\|^2)$ bias; the supplement states this bridge and reports an exact-HVP diagnostic. The covariance-matched target $N_P = \mathrm{tr}(PHP\Omega)$ differs from isotropic Hutchinson estimation of $\mathrm{tr}(H)$ and from scalable Hessian-diagonal estimation; we claim no generic variance, bias, or compute advantage over those methods (Hutchinson, 1989; Yao et al., 2021). Randomized curvature information and preconditioner diagnostics are well established (Halko et al., 2011; Pilanci & Wainwright, 2017; Frangella et al., 2024; DiPaolo & Gu, 2019); our target is the value-transfer question after a candidate has been frozen.

3 EXACT GEOMETRY OF FINITE-STEP TRANSFER

Let

$$\rho(s) = F(w+s) - F(w) - h^\top s - \frac{1}{2} s^\top H s \tag{8}$$

be the finite-step remainder around the current state.

3.1 SECOND-ORDER INFORMATION IS NOT ENOUGH

A local quadratic valuation depends only on the second-order jet $(F(w), h, H)$. That information alone cannot identify finite-step ranking.

Proposition 2 (No free finite-step transfer)**.** *Fix a one-dimensional quadratic $q(t)$ and distinct nonzero points t_0, t_1. For arbitrary target deviations $y_0, y_1 \in \mathbb{R}$, there exists a polynomial $r(t) = t^3(a + bt)$ such that $r(0) = r'(0) = r''(0) = 0$ and $r(t_i) = y_i$ for $i \in \{0, 1\}$. Consequently $F = q + r$ has the same value, gradient, and Hessian as q at the expansion point, while its two finite-step values can be changed arbitrarily.*

The construction is explicit:

$$b = \frac{y_1/t_1^3 - y_0/t_0^3}{t_1 - t_0}, \qquad a = \frac{y_0}{t_0^3} - b t_0.$$

Thus an unconditional mapping from a local second-order value to a finite-step ordering is not identifiable without additional regularity information.

3.2 EXACT COMPARATIVE TRANSFER GEOMETRY

Assume that $\nabla^2 F$ is L_H-Lipschitz on $w + B(0, R)$, where $R \geq \max\{\|s_0\|, \|s_1\|\}$:

$$\|\nabla^2 F(x) - \nabla^2 F(y)\|_{\mathrm{op}} \leq L_H \|x - y\|. \tag{9}$$

The absolute Taylor bound $|\rho(s)| \leq L_H \|s\|^3/6$ does not use the relation between two actions. Define the cubic action distance

$$d_3(x, y) = 3 \inf_{\gamma: x \to y} \int_0^1 \|\gamma(t)\|^2 \|\gamma'(t)\| \, dt. \tag{10}$$

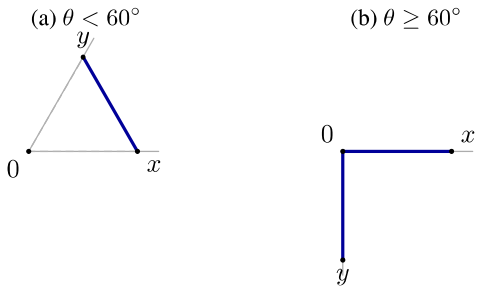

Figure 1: Cubic action geometry in cone coordinates $u = \|s\|^3$ with angular separation 3θ. The shortest weighted path (dark) is direct below 60° and goes through the apex at or above 60°. The gray dashed route in (a) is a longer separate-action path.

For $a = \|x\|$, $b = \|y\|$ and angle $\theta \in [0, \pi]$ between nonzero endpoints, it has the closed form

$$d_3(x, y) = \left(a^6 + b^6 - 2a^3b^3 \cos(\min\{3\theta, \pi\})\right)^{1/2}. \tag{11}$$

If either endpoint is zero, $d_3(x, y) = a^3 + b^3$. For $\theta < \pi/3$ the minimizing cone path is direct; for $\theta \geq \pi/3$ it passes through the origin. In particular, $d_3(x, y) = 2a^3 \sin(3\theta/2)$ for equal nonzero radii only when $\theta \leq \pi/3$.

Theorem 1 (Exact endpoint transfer modulus)**.** *Under* (9)*,*

$$|\rho(s_1) - \rho(s_0)| \leq \frac{L_H}{6} d_3(s_0, s_1). \tag{12}$$

For every endpoint pair, the right side is the exact supremum over functions with the stated fixed second-order jet and Hessian-Lipschitz constant at most L_H. For distinct endpoints without a Hessian-Lipschitz bound, the supremum is infinite.

The proof in the supplement integrates $\|\nabla\rho(s)\| \leq L_H\|s\|^2/2$ along a shortest weighted path and constructs equality-attaining cubics. In contrast, the earlier straight-chord radius replaces d_3 by $(\|s_0\|^2 + \langle s_0, s_1\rangle + \|s_1\|^2)\|s_1 - s_0\|$; separate absolute control uses $\|s_0\|^3 + \|s_1\|^3$. Both are valid upper bounds on d_3, but neither is the exact endpoint modulus. The geometry is on actions, not directly on preconditioner matrices.

For fixed deterministic actions let $Q_j = Q(s_j)$, $R_j = F(w) - F(w + s_j)$, $\Delta_Q = Q_1 - Q_0$, and $\Delta_R = R_1 - R_0$. With $r_{01} = L_H d_3(s_0, s_1)/6$, the sharp frontier is

$$\Delta_R \in [\Delta_Q - r_{01}, \Delta_Q + r_{01}]. \tag{13}$$

Thus $\Delta_Q > r_{01}$ guarantees $R_1 > R_0$, and $\Delta_Q < -r_{01}$ guarantees the reverse. Equality at either boundary allows a tie; when $|\Delta_Q| < r_{01}$ both strict rankings are attainable (unless $r_{01} = 0$). This is a conditional frontier, not a neural certificate: the experiments do not supply L_H.

For stochastic action laws μ_0, μ_1 with finite third moments, assume global Hessian Lipschitzness or common bounded support inside a ball where it holds. Writing $S_j \sim \mu_j$, $R_j = \mathbb{E}[F(w) - F(w + S_j)]$ and $Q_j = \mathbb{E}Q(S_j)$, Theorem 1 gives

$$|(\Delta_R - \Delta_Q)| \leq \frac{L_H}{6} W_{1,d_3}(\mu_0, \mu_1). \tag{14}$$

This is an upper bound, not an exact distribution-wise minimax characterization. A common-random-number (CRN) pairing gives a computable *coupling cost* above W_{1,d_3}, not the Wasserstein distance unless optimality is shown. W_{1,d_3} is a pseudometric on parameterized geometries and a metric only after identifying geometries with the same action law.

Estimated finite-action choices require a separate statistical layer: if component intervals and transfer envelopes hold *simultaneously* across a fixed action grid, they remain valid at a data-dependent selected radius. Unbiased point estimators alone do not provide such bands. The selection and cost-adjusted utility details are in the supplement; no neural interval is presented as certified here.

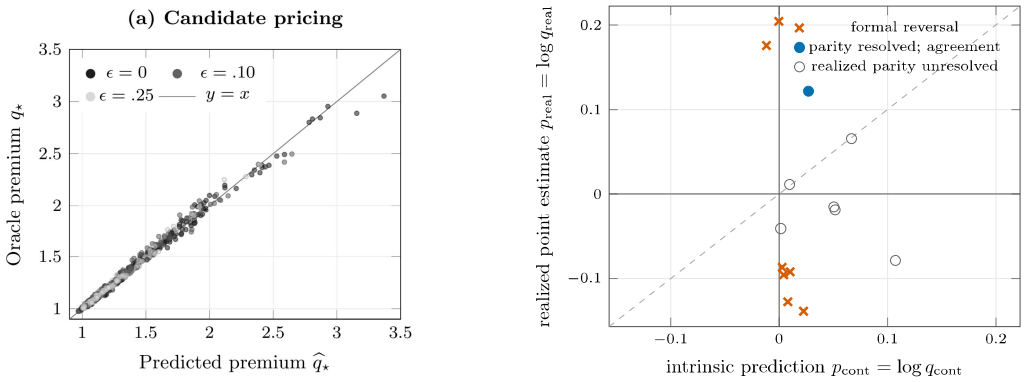

Figure 2: Value versus finite-step transfer. (a) Predicted and oracle break-even premiums in the exact-quadratic synthetic benchmark (marker gray level denotes mismatch ϵ). (b) Intrinsic and realized log ratios in the preregistered neural study: crosses mark seven interval-qualified reversals, the filled circle an agreement, and open circles unresolved parity. The diagonal is equality, not a certification boundary. Exact-model valuation recovery does not ensure neural finite-step ordering.

4 Experiments

We use two complementary evaluations. The synthetic study tests the valuation estimator in a setting where the quadratic model is exact. The neural study tests whether that intrinsic value transfers to finite progress along real training trajectories. Full protocols, per-state tables, audit hashes, and additional diagnostics are in the supplement.

4.1 Exact-model calibration

The calibrated synthetic family uses $d = 64$, ranks $r \in \{4, 8\}$, paired curvature amplitudes $a \in \{.2, .5, .8, .9\}$, gradient-energy fractions $\gamma \in \{.15, .40, .65, .90\}$, mismatch $\epsilon \in \{0, .10, .25\}$, three noise levels, and six deterministic seeds, yielding 1,728 base states. Candidate construction and valuation use disjoint sample streams, and the population Hessian is unavailable to the estimator. In the formal high-probe run, predicted constructed-candidate prices achieve Spearman 0.99258, log-price MAE 0.00704, and zero invalid valuations. At the benchmark's all-in effective compute price $q_{\text{eff}} = 2$ (which may amortize construction overhead, unlike the local per-step q_{impl}), the value rule attains precision 1.0, recall 0.838, and mean decision regret 3.20×10^{-5}, compared with 0.182 regret for a gain-only purchase rule. These results establish recovery of the intended intrinsic quantity when the quadratic model is exact, not neural transfer.

Figure 2(a) shows that the estimator can recover the intended intrinsic price when the quadratic model is exact. The neural experiment changes the transfer question: whether a locally valued ordering survives the finite action actually evaluated at a training state.

4.2 Preregistered neural transfer study

We next evaluate CIFAR-10 with ResNet-18. A deterministic split assigns 44,000 examples to the training trajectory and 2,000 each to pricing, oracle-update, and held-out oracle-evaluation roles. Five independent 30-epoch AdamW trajectories (seeds 1–5) use the same cosine learning-rate schedule. Four checkpoints are predeclared at approximately 5%, 20%, 50%, and 80% of the trajectory. Seed 0 is a pilot and excluded from formal inference. This controlled scale permits independently randomized training trajectories, disjoint measurement roles, repeated paired held-out gains, and auditable frozen counterfactual geometries. It does not establish prevalence in larger models.

The two frozen geometries are deliberately simple to interpret. P_0 is a diagonal second-moment operator built only from historical gradients,

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2, \qquad P_0 g = g / \sqrt{v_t + \epsilon_D}.$$

Table 1: Preregistered neural transfer study. A ratio is resolved when both candidate held-out gain intervals are strictly positive. Realized parity is resolved only when the paired log-ratio interval excludes zero. A formal reversal additionally requires the unrounded intrinsic prediction to lie strictly on the opposite side.

Seed	Ratio / 4	Parity / 4	Formal reversals	Math-valid / 4
1	3	2	2	4
2	3	1	1	3
3	3	1	0	4
4	2	1	1	4
5	3	3	3	4
All	14 / 20	8 / 20	7	19 / 20

P_1 is a block-Shampoo-style shadow operator. For a matrix-shaped gradient G it maintains

$$L_t = \beta L_{t-1} + (1-\beta)GG^\top, \qquad R_t = \beta R_{t-1} + (1-\beta)G^\top G,$$

and applies $(L+\lambda_L I)^{-1/4} G (R+\lambda_R I)^{-1/4}$. Convolutional tensors are flattened by output channel, matrix blocks are capped at dimension 128, and one-dimensional parameters use diagonal treatment. Neither geometry updates the common AdamW trajectory; both are frozen counterfactual candidates at each checkpoint.

The measurement protocol uses the same stochastic augmentation law as the training trajectory. Pricing uses eight paired probe replicates at primary secant scale $\rho = 3 \times 10^{-4}$. Oracle tuning uses eight independent replicates over the predeclared radius grid, and held-out evaluation uses sixteen fresh paired replicates. The data roles, checkpoint hashes, measurement law, and five formal seeds were sealed before the five-run study. This design isolates state-local candidate comparison from trajectory effects: both geometries are evaluated from the same checkpoint rather than being allowed to generate diverging training paths.

At trajectory level, the intrinsic point predictor outperforms a neutral parity predictor on two of five independently randomized runs and underperforms it on three (full values in the supplement). We therefore do not interpret the original point ratio as a generally reliable predictor. Its role is the model-side valuation target; the transfer layer determines whether that valuation is deployable at a particular finite action.

Figure 2(b) shows the central neural result: among 20 predeclared states, 14 have resolved ratios, eight of those have resolved realized parity, and seven meet the formal reversal rule. The other six ratio-resolved states are not counted as reversals, even though four point estimates cross parity. Nineteen of twenty states remain mathematically valid; weak-signal cases are retained rather than clipped or assigned an artificial ratio. A predeclared exact-HVP diagnostic at the 50% pilot checkpoint leaves the qualitative finite-radius mismatch intact, so finite-secant approximation is not a complete explanation.

The five-trajectory study was preregistered before the transfer theory in Section 3; it tests unconditional point transfer, not that later theorem. A historical post-theory proxy study on untouched seeds 6–8 found within-trajectory Spearman correlations $(0.4, 1.0, 0.8)$ for a paired action-scale proxy and $(0.2, 1.0, 0.8)$ for a separate proxy. The paired advantage was not consistent; selected-action ranking was correct in three states, incorrect in three, and unresolved in six. These pre-d_3 proxy measurements are mixed and do not validate the exact d_3 refinement.

4.3 Prospective residual copying

We separately sealed a prospective five-trajectory study (seeds 9–13) before observing outcomes. At four checkpoints, the gauge-fixed realizable family $P_\lambda = (1-\lambda)P_0 + \lambda P_1$ used anchors $\lambda \in \{0, 1\}$ and unseen targets $\lambda \in \{.2, .4, .6, .8\}$. Model/pricing, anchor calibration, and held-out evaluation had disjoint roles. The primary predictor added the selected anchor's full calibrated residual to the target's intrinsic prediction; selection used CRN-averaged d_3 distance on pricing-role actions.

Table 2: Sealed prospective residual-copying result. D_s is the trajectory-level mean paired squared-loss difference, corrected minus raw; negative favors copying. Coverage counts joint-valid target–checkpoint cases out of 16 per seed. The five trajectories, not the 76 cases, are the independent units.

Seed	Coverage	$D_s \times 10^6$
9	16/16	$+6.913789698$
10	12/16	$+6.014190657$
11	16/16	-41.313655212
12	16/16	$+39.152631359$
13	16/16	$+8.293447950$

The trajectory effect D_s is the mean paired squared-loss difference, corrected minus raw; *negative* favors correction. Trajectories were independently randomized but evaluated under one common predeclared measurement design; checkpoints and targets are repeated observations.

The sealed study had 76/80 joint-valid target–checkpoint cases; Table 2 reports all five trajectory effects and coverage. The equal-weight mean was $+3.812080890 \times 10^{-6}$ and the median $+6.913789698 \times 10^{-6}$. Only one of five trajectories improved, so coverage met the preregistered adequacy gate but the directional success criterion **failed**. The d_3-versus-Euclidean nearest-anchor comparison was mixed. An integrity audit classified the pricing-role action geometry as a *surrogate law*: its gradients came from the pricing pool, whereas calibrated and evaluated updates used the oracle-update pool. The mismatch magnitude is unavailable, and no neural L_H certificate is claimed.

A post-hoc case decomposition finds 50 beneficial corrections, 16 wrong-sign failures, and 10 same-sign overshoots among the 76 valid cases. This is consistent with the squared-loss identity $D = c(c - 2t)$ for correction c and target residual t, but does not identify a unique cause. Fixed-anchor, Euclidean, chord, and d_3 copying all had adverse equal-seed means and helped only one trajectory. The prospective result tests full residual copying, *not* the uncertainty-aware correction of Proposition 3.

5 From transfer intervals to robust corrections

The geometry certifies an interval; the downstream loss determines how to act on it. Suppose the *operational* residual $t = R - \widehat{Q}$ is known to lie in a valid interval $[L, U]$. A correction c changes squared loss relative to the raw predictor by

$$(t - c)^2 - t^2 = c^2 - 2ct, \quad \Phi(c; L, U) = \begin{cases} c^2 - 2cL, & c \geq 0, \\ c^2 - 2cU, & c \leq 0, \end{cases}$$

where Φ is worst-case excess loss over the interval.

Proposition 3 (Robust excess-loss correction)**.** *For $L \leq U$, the unique minimizer of Φ is $c^\star = \mathrm{Proj}_{[L,U]}(0)$, with value $-\operatorname{dist}(0, [L, U])^2$. The complete set of uniformly non-worsening corrections is*

$$\mathcal{C}_{\mathrm{safe}}(L, U) = \begin{cases} [0, 2L], & L > 0, \\ \{0\}, & L \leq 0 \leq U, \\ [2U, 0], & U < 0. \end{cases}$$

The piecewise quadratics give the proof: on either half-axis the worst endpoint is fixed, and each branch has a unique quadratic minimum; solving $\Phi(c) \leq 0$ gives the displayed safe set. For a symmetric interval $[\mu - r, \mu + r]$, $r \geq 0$, the minimizer is $\operatorname{sign}(\mu)(|\mu| - r)_+$. It abstains for $|\mu| \leq r$. When $r < |\mu| < 2r$, a nonzero robust correction exists but full copying is not uniformly safe. For a *nonzero* full-copy correction $c = \mu$, uniform non-degradation holds iff $|\mu| \geq 2r$, and strict worst-case improvement iff $|\mu| > 2r$; $\mu = 0$ is trivially identical to the raw predictor. Even when full copying is safe, the minimax correction remains shrunk by r whenever $r > 0$.

Full copying is the minimax center for *absolute* worst-case prediction loss, $\min_c \sup_{t \in [\mu - r, \mu + r]}(t - c)^2 = r^2$, but not generally for excess loss relative to abstention. The prospective experiment used an unshrunk full-copy residual rule; its failure therefore does not falsify the transfer theorem.

This decision consequence was derived after the sealed prospective study and was not evaluated by that protocol. A true d_3 residual-difference bound can help form an interval, but operational prediction additionally involves differential model/pricing error and calibration noise (supplement). No calibrated neural interval was available here.

6 Related work and limitations

Structured adaptive optimization spans diagonal methods, Kronecker-factored curvature, tensor preconditioning, and recent matrix-geometry methods (Kingma & Ba, 2015; Duchi et al., 2011; Martens & Grosse, 2015; Gupta et al., 2018; Yao et al., 2021; Liu et al., 2024; Vyas et al., 2025). Recent analyses increasingly separate the effect of preconditioner structure from optimizer heuristics and show that a richer structure is not automatically superior (Xie et al., 2025; Eschenhagen et al., 2025). Our question begins after candidate construction: we ask for a state-local value of a frozen geometry and for conditions under which its ordering transfers to a finite update.

Preconditioner selection and optimizer switching are also related. Randomized diagnostics can compare preconditioners without forming all spectral information (DiPaolo & Gu, 2019); online formulations can adapt or optimize the preconditioner itself (Hu & Simchi-Levi, 2026); repeated-resampling approaches study adaptive optimizer selection along a single training path (Richman & Wüthrich, 2026). Recent geometric work develops certificates for structural reachability, projection, and condition-number properties of restricted preconditioner families (Li, 2026b;a). Those certificates concern which geometries can be expressed or reached; ours concerns whether the local value assigned to two already-constructed candidates transfers to their realized finite-step progress.

Recent work also clarifies the boundary of local quadratic reasoning. Meterez et al. (2026) show that quadratic models can be predictive over substantial windows of LLM pretraining. Our claim is compatible: an informative local model does not by itself guarantee a pairwise finite-action ordering. Concurrently, Guo et al. (2026) analyze finite-step adequacy along a proposed looped-Transformer direction using pathwise curvature variation. We do not claim novelty for generic finite-step failure. Our object is comparative transfer between two frozen geometries that induce coupled stochastic actions. The exact cubic ground metric refines the earlier straight-chord action bound; metric-cone geometry, Taylor remainder bounds, and Wasserstein transport are not claimed as new in isolation.

Finally, randomized curvature sketches and stochastic second-order methods motivate Hessian-vector or secant information (Halko et al., 2011; Pilanci & Wainwright, 2017; Frangella et al., 2024). Hutchinson probing estimates traces under an isotropic probe law, and AdaHessian uses related randomized products for scalable diagonal curvature estimation (Hutchinson, 1989; Yao et al., 2021). Our covariance-matched term $\mathrm{tr}(PHP\Omega)$ is a different target; no general efficiency or accuracy dominance is asserted. Common random numbers are also established in stochastic trust-region methods (Ha et al., 2025). Neither curvature probing nor paired randomness is claimed as new here; the theorem controls the higher-order difference under a shared coupling.

7 Discussion and conclusion

A frozen optimization geometry has two distinct notions of value. The first is intrinsic: what the local stochastic quadratic model predicts after optimizing over a global step magnitude. The second is operational: what the corresponding finite update actually delivers. Treating these as the same quantity hides the transfer problem.

The exact conditional theory gives this transfer problem a cubic geometry on actions. It identifies the endpoint-wise worst-case differential remainder and the deterministic ranking frontier under a Hessian-Lipschitz assumption. The stochastic transport inequality bounds expected residual differences for true action laws, but the cost of a measured CRN coupling is only an upper bound on the corresponding Wasserstein distance. Finite-action envelopes remain conditional on simultaneous component and transfer bounds.

The neural evidence has a narrower meaning. The first preregistered study shows that unconditional point transfer can reverse; the historical proxy comparison is mixed. The separate prospective study did not meet its criterion for full residual copying. In that study, geometry probes and evaluated updates used different sampling pools, so the true-law transfer radius also contains unmeasured

cross-law terms. The exact neural quadratic gain was unavailable, leaving differential pricing/model error unidentified. Calibration sampling error is descriptive, not an established cause of failure. ReLU networks are nonsmooth at activation boundaries, and no neural L_H upper bound or conventional five-sample significance inference is claimed.

The robust excess-loss rule shows that knowing a valid residual interval would support shrinkage or abstention rather than unconditional copying. Such intervals require calibrated uncertainty scales and were not available prospectively. A geometry-only nearest-anchor rule can ignore a common positive L_H multiplier in *distance ranking*; the signal-aware rule $\arg\max_A(|a_A| - r_A)_+$ generally cannot. The scope remains local CIFAR-10/ResNet-18 counterfactual updates from frozen shadow geometries on AdamW trajectories, not optimizer switching, global convergence, or speedup. A matched-law, independently calibrated test of robust correction is future work.

AI USE STATEMENT

Generative AI tools assisted with literature discovery and retrieval; brainstorming and refining research questions; theoretical formulation and proof checking; experimental protocol design and review; implementation and debugging; artifact auditing; result interpretation; scientific figure preparation and refinement; and language polishing. The authors reviewed all AI-assisted output and checked citations, proofs, algebra, numerical outputs, experimental artifacts, assumptions, code outputs, figures, and final claims. Generative AI did not generate the synthetic or neural datasets. The authors take responsibility for the final paper, including all AI-assisted text, claims, proofs, code, and artifacts.

REPRODUCIBILITY STATEMENT

The supplementary material records the frozen data-role split, training and measurement configurations, checkpoint and artifact hashes, all five original trajectory summaries, the historical proxy study, and exact synthetic result sources. Candidate construction and valuation use separated statistical roles. Complete proofs, state-level tables, the finite-secant bridge, and protocol details are provided there.

REFERENCES

Conner DiPaolo and Weiqing Gu. A randomized algorithm for preconditioner selection. *arXiv preprint arXiv:1908.00633*, 2019. URL https://arxiv.org/abs/1908.00633.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

Runa Eschenhagen, Aaron Defazio, Tsung-Hsien Lee, Richard E. Turner, and Hao-Jun Shi. Purifying Shampoo: Investigating Shampoo's heuristics by decomposing its preconditioner. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5507. URL https://papers.nips.cc/paper_files/paper/2025/hash/f14f4eda29a74c02c803699a09529bb9-Abstract-Conference.html.

Zachary Frangella, Pratik Rathore, Shipu Zhao, and Madeleine Udell. SketchySGD: Reliable stochastic optimization via randomized curvature estimates. *SIAM Journal on Mathematics of Data Science*, 6(4):1173–1204, 2024. doi: 10.1137/23M1575330.

Zhihao Guo, Zonghan Wu, Haizhou Du, Huan Huo, Yilei Shao, Athanasios V. Vasilakos, and Qingsong Wen. Right Direction, Wrong Step: Geometric analysis of finite-step failure in looped transformers, 2026. URL https://arxiv.org/abs/2609.16665.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1842–1850. PMLR, 2018. URL https://proceedings.mlr.press/v80/gupta18a.html.

Yunsoo Ha, Sara Shashaani, and Raghu Pasupathy. Complexity of zeroth- and first-order stochastic trust-region algorithms. *SIAM Journal on Optimization*, 35(3):2098–2127, 2025. doi: 10.1137/24M1664484.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53 (2):217–288, 2011. doi: 10.1137/090771806.

Haichen Hu and David Simchi-Levi. Optimizing the preconditioner: A black-box online-to-nonconvex conversion with static regret minimization oracles. *arXiv preprint arXiv:2607.17607*, 2026. URL https://arxiv.org/abs/2607.17607.

Michael F. Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics—Simulation and Computation*, 18(3):1059–1076, 1989. doi: 10.1080/03610918908812806.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Zavier Li. Structured preconditioning in affine-invariant geometry: Projection, certificates, and kronecker separation. *arXiv preprint arXiv:2607.12286*, 2026a. URL https://arxiv.org/abs/2607.12286.

Zavier Li. Restricted dynamic geometric complexity: Certificates for structured preconditioning. *arXiv preprint arXiv:2607.07204*, 2026b. URL https://arxiv.org/abs/2607.07204.

Hong Liu, Zhiyuan Li, David Hall, Percy Liang, and Tengyu Ma. Sophia: A scalable stochastic second-order optimizer for language model pre-training. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3xHDeA8I9i.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2408–2417. PMLR, 2015. URL https://proceedings.mlr.press/v37/martens15.html.

Alexandru Meterez, Pranav Ajit Nair, Depen Morwani, Cengiz Pehlevan, Sham Kakade, and Alex Damian. A defense of the quadratic model, 2026. URL https://arxiv.org/abs/2607.21716.

Mert Pilanci and Martin J. Wainwright. Newton Sketch: A near linear-time optimization algorithm with linear-quadratic convergence. *SIAM Journal on Optimization*, 27(1):205–245, 2017. doi: 10.1137/15M1021106.

Ronald Richman and Mario V. Wüthrich. Many optimizers but only one training path: Repeated resampling for adaptive optimizer selection. *arXiv preprint arXiv:2608.18810*, 2026. URL https://arxiv.org/abs/2608.18810.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham M. Kakade. SOAP: Improving and stabilizing Shampoo using Adam for language modeling. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=IDxZhXrpNf.

Shuo Xie, Tianhao Wang, Sashank J. Reddi, Sanjiv Kumar, and Zhiyuan Li. Structured preconditioners in adaptive optimization: A unified analysis. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 68746–68776. PMLR, 2025. URL https://proceedings.mlr.press/v267/xie25j.html.

Zhewei Yao, Amir Gholami, Sheng Shen, Mustafa Mustafa, Kurt Keutzer, and Michael W. Mahoney. ADAHESSIAN: An adaptive second order optimizer for machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10665–10673, 2021. doi: 10.1609/aaai.v35i12.17275.