GROTHENDIECK MEETS ADAM: ONE PROBE, THREE BUDGETS IN ADAPTIVE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Sign descent and Adam work in the same ℓ_∞ geometry but pay for curvature differently: the standard smoothness that governs normalized descent and the adaptive smoothness that governs adaptive methods can differ by a factor of the dimension. We show that these two constants, and a pointwise constant between them, are one quantity: the largest expected curvature that a random probe can detect. They differ only in how the probe's budget is enforced: surely, on average at each point, or on average over the landscape. For convex objectives in diagonal geometry, adaptive smoothness is a squared 2-summing norm and the ideal preconditioner is its Pietsch measure. Each relaxation is governed by one invariant of the preconditioner algebra. Averaging at a point is a semidefinite relaxation whose gap lies between the largest matrix-block size b and $K_G b$; in diagonal geometry it is at most Grothendieck's constant $K_G < 1.783$, or $\pi/2$ for convex objectives, both asymptotically sharp. Averaging over the landscape costs a curvature rotation number, which is at most the rank of the algebra, the number of directions a preconditioner weighs independently, a bound that convex objectives approach; the rank also replaces the dimension in the known comparison of the two smoothness constants. Rotation is one for quadratics and binary logistic regression and grows linearly for log-sum-exp. In box-constrained ℓ_∞ convex optimization, rotation is the acceleration horizon: up to one logarithm, the minimax risk of deterministic first-order methods over functions of rotation at most k is $1/T$ for $T \le k$ queries and k/T^2 for $k \le T \le d$, and one adaptive method, analysed without choosing a basis, attains both rates in every well-structured geometry. Gradient noise obeys the same three levels.

"The sea advances insensibly in silence, nothing seems to happen, nothing moves, the water is so far off you hardly hear it… yet it finally surrounds the resistant substance." (explained in Appendix J)
—Alexander Grothendieck, *Récoltes et Semailles* (1985–1987, pp. 552–553; trans. McLarty, 2007)

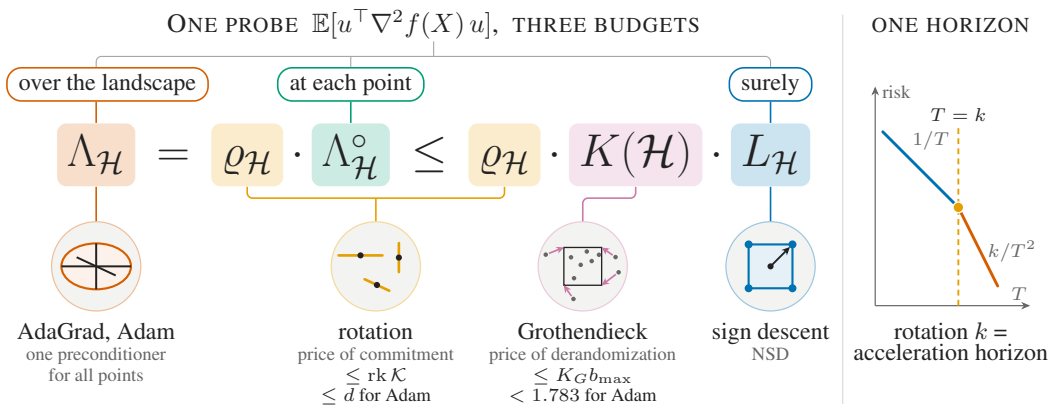

Figure 1: **The paper in one inequality** (eq. (7)). One probe, three budgets (Theorem 2.1); two prices, a Grothendieck constant below 1.783 for Adam (Theorem 3.1) and a rotation number at most the rank (Theorem 4.1); one horizon (Theorem 5.2).

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.

1 INTRODUCTION

Sign descent and Adam work in the same geometry but pay different prices for curvature. Both exploit the ℓ_∞ geometry of the loss (Balles et al., 2020; Bernstein & Newhouse, 2024; Xie et al., 2025a): sign descent is steepest descent for $\|\cdot\|_\infty$, and Adam adapts a diagonal preconditioner to it. Their guarantees depend on how fast the gradient can change in this geometry, but they measure it differently. Normalized steepest descent (NSD) pays the standard smoothness $L_\infty(f) = \sup_x \|\nabla^2 f(x)\|_{\infty\to 1}$. Adaptive methods pay the adaptive smoothness $\Lambda(f)$ of Xie et al. (2025b; 2026), the least trace of a diagonal matrix H with $-H \preceq \nabla^2 f(x) \preceq H$ for every x. The two satisfy $L_\infty \le \Lambda \le d\, L_\infty$, and the factor d separates regimes: Λ supports accelerated rates of order $\Lambda R^2/T^2$ up to logarithms, whereas no deterministic first-order method can guarantee better than order $L_\infty R^2/T$ over a box, up to a logarithm, before the number of queries reaches the dimension (Guzmán & Nemirovski, 2015). What is the factor d made of?

One probe, three budgets. The two constants measure the same thing. Call a random point X together with a random direction u a *probe*; it reads the curvature $\mathbb{E}[u^\top \nabla^2 f(X) u]$. For convex f in the diagonal geometry of Adam, Theorem 2.1 shows

$$\left.\begin{matrix} L_\infty(f) \\ \Lambda^\circ(f) \\ \Lambda(f) \end{matrix}\right\} = \sup \mathbb{E}\big[u^\top \nabla^2 f(X)\, u\big] \quad \text{over probes with} \quad \begin{cases} |u_i| \le 1 & \text{surely,} \\ \mathbb{E}[u_i^2 \mid X] \le 1 & \text{for all } i, \\ \mathbb{E}[u_i^2] \le 1 & \text{for all } i. \end{cases} \tag{1}$$

The middle budget defines a pointwise adaptive smoothness $\Lambda^\circ(f)$. Sign descent charges the probe for leaving the unit box on any outcome. Adam charges it only on average, and on average over the whole landscape, so the probe may inspect different coordinates at different points. The principle holds in every well-structured preconditioner geometry, and for indefinite Hessians with a pair of directions; the budget is then measured by the conditional expectation onto the preconditioner algebra. The factor between standard and adaptive smoothness thus splits into two relaxations, each with its own mathematics.

Surely to pointwise: derandomization. At a fixed point only the probe's second moment matters, so the pointwise constant is a semidefinite relaxation of the standard one. Rounding a random probe back into the unit ball is Grothendieck's problem. For every preconditioner algebra the worst gap lies between the size $b_{\max}$ of its largest matrix block and $K_G b_{\max}$, where $K_G < 1.783$ is Grothendieck's constant. In diagonal geometry it is at most K_G, or $\pi/2$ for convex objectives, and both constants are asymptotically sharp (Theorem 3.1).

Pointwise to global: commitment. One preconditioner serving the whole landscape must pay for curvature wherever it occurs. The ratio $\varrho(f) = \Lambda(f)/\Lambda^\circ(f)$, the *curvature rotation number*, is this price of commitment. It is at most the rank of the algebra, the number of directions a preconditioner can weigh independently, and convex objectives approach this bound: convexity does not help (Theorem 4.1). The same argument replaces d by the rank in $\Lambda \le d\, L$: by d_L for one-sided Shampoo on $d_L \times d_R$ matrices, and by the number of layers for layer-wise scalar step sizes. Rotation is one for quadratics and binary logistic regression, and $d/2$ for log-sum-exp, the multiclass cross-entropy as a function of its logits.

Why Grothendieck, and why only at a point. Write $\nabla^2 f(x) = B_x^\top B_x$ for convex f in diagonal geometry. Then (1) is Pietsch's description of 2-summing norms (Pietsch, 1967): $\Lambda^\circ(f)$ is the largest squared 2-summing norm of the operators $B_x : \ell_\infty^d \to \ell_2$, and the ideal diagonal preconditioner at x is a Pietsch measure. Grothendieck's theorem bounds the 2-summing norm of an operator from ℓ_∞ into a Hilbert space by $\sqrt{\pi/2}$ times its norm (Grothendieck, 1953; Lindenstrauss & Pełczyński, 1968). The adaptive constant $\Lambda(f)$ is the squared 2-summing norm of the whole field $x \mapsto B_x$. That operator takes values in an ℓ_∞-sum of Hilbert spaces, where Grothendieck's theorem fails, and rotation measures the failure.

Consequences. In box-constrained convex optimization with ℓ_∞ geometry, rotation is the acceleration horizon: over functions of rotation at most k, the minimax risk of deterministic first-order methods is LR^2/T up to $T = k$ and kLR^2/T^2 for $k \le T \le d$, up to one logarithm (Theorem 5.2). Both rates are attained by one adaptive method that learns its metric from gradient differences, with a basis-free analysis valid in every well-structured geometry, including one-sided Shampoo (Theorem 5.1; cf. Kovalev, 2026). The hard instances are smoothed maxima of Guzmán & Nemirovski (2015) whose smoothing exponent dials their rotation to the budget. Gradient noise has the same

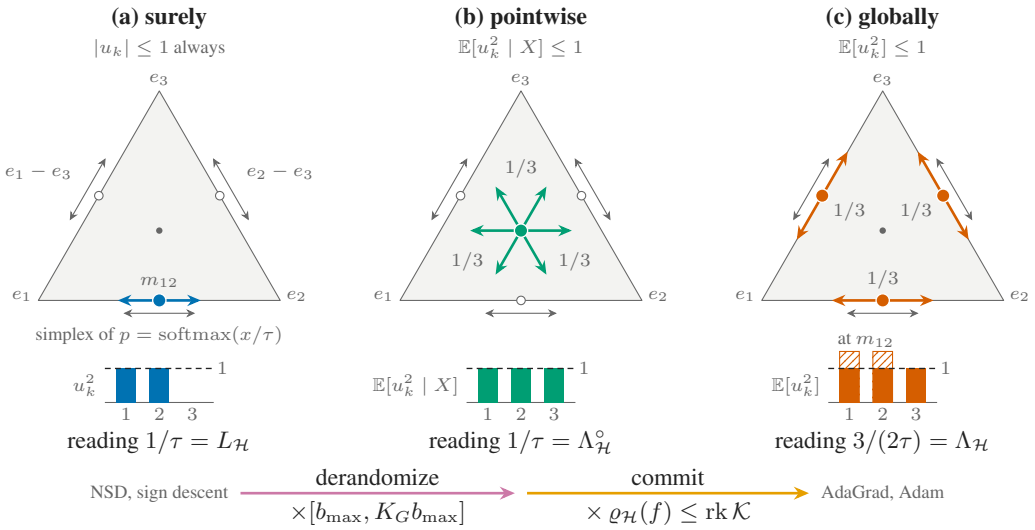

Figure 2: **One probe, three budgets** (Theorem 2.1), on lse_τ in $d = 3$ in diagonal geometry, drawn on the simplex of $p = \mathrm{softmax}(x/\tau)$. Near an edge midpoint $m_{ij} = \frac{1}{2}(e_i + e_j)$ the Hessian tends to $\frac{1}{4\tau}(e_i - e_j)(e_i - e_j)^\top$: curvature runs along the edge (grey). Probe arrows have length $|u|$; bars show the budget spent per coordinate. **(a)** The sure probe $u = e_1 - e_2$ at m_{12} reads $1/\tau = L_{\mathcal{H}}$. **(b)** A probe confined to one point may randomize its direction: at the centroid the three directions $\sqrt{3/2}(e_i - e_j)$, each with probability $\frac{1}{3}$, meet $\mathbb{E}[u_k^2 \mid X] = 1$ and read $1/\tau = \Lambda^{\circ}_{\mathcal{H}}$. **(c)** Under the global budget the same directions are read where each is most curved. The probe overspends at every point (hatched: $\mathbb{E}[u_k^2 \mid X = m_{12}] = \frac{3}{2}$) but not on average, and reads $3/(2\tau) = \Lambda_{\mathcal{H}}$: a rotation of $d/2$ (Proposition 4.2(iii)). Bottom: the two relaxations and their largest factors (Theorems 3.1 and 4.1). Readings at the ties are limits.

three levels, where the preconditioner may depend on the noise realization, on the point, or on nothing. At a single point Gaussian noise costs at most $\pi/2$ times the block size, whereas sparse noise can cost the full rank; noise whose covariance moves with the point can cost the full rank through rotation even when it is Gaussian (Theorem G.1). On Hessians recorded while training five small models with Adam, the pointwise gap stays below 1.13 and the rotation below 1.09, and the whole factor between the two constants stays below 1.15 while d ranges up to 63 (Section 6).

Related work. Balles et al. (2020) showed that separable smoothness, whose constant is the trace of a common diagonal majorant, implies ℓ_∞-smoothness, and related the latter to the diagonal concentration and eigenvalue spread of the Hessian. The preconditioner framework and AdaGrad-type rates are due to Xie et al. (2025b; 2026); our constrained acceleration uses Ene et al. (2021), and our lower bounds build on Guzmán & Nemirovski (2015). The semidefinite inequalities are classical (Grothendieck, 1953; Nesterov, 1998; Alon & Naor, 2006; Pisier, 2012), as are 2-summing norms and Pietsch factorization (Pietsch, 1967; Lindenstrauss & Pełczyński, 1968; Diestel et al., 1995). What is new is the single probing problem, which places these classical objects inside the comparison of optimizers and isolates the landscape quantity they do not control. Sign and spectral geometries supply further context (Bernstein & Newhouse, 2024; Xie et al., 2025a; An et al., 2025; Pethick et al., 2025). One-sided Shampoo is distinguished from two-sided Kronecker preconditioning (Gupta et al., 2018), whose preconditioners do not form the positive cone of an algebra.

2 One probe, three budgets

Preconditioner geometries. $\mathbb{S}_d$ is the space of symmetric $d \times d$ matrices, $\mathbb{S}_d^+$ its positive semidefinite cone, and $\langle A, B\rangle = \mathrm{Tr}(A^\top B)$. Following Xie et al. (2025b; 2026), a preconditioner set is *well-structured* if $\mathcal{H} = \mathbb{S}_d^+ \cap \mathcal{K}$ for a unital subalgebra $\mathcal{K} \subseteq \mathbb{M}_d(\mathbb{R})$, which we may take closed under transposition (replacing $\mathcal{K}$ by $\mathcal{K} \cap \mathcal{K}^\top$ leaves $\mathcal{H}$ unchanged). Let $\pi_{\mathcal{K}}$ be the orthogonal projection onto $\mathcal{K}$. It is a conditional expectation: positive, unital, trace preserving and $\mathcal{K}$-bimodular

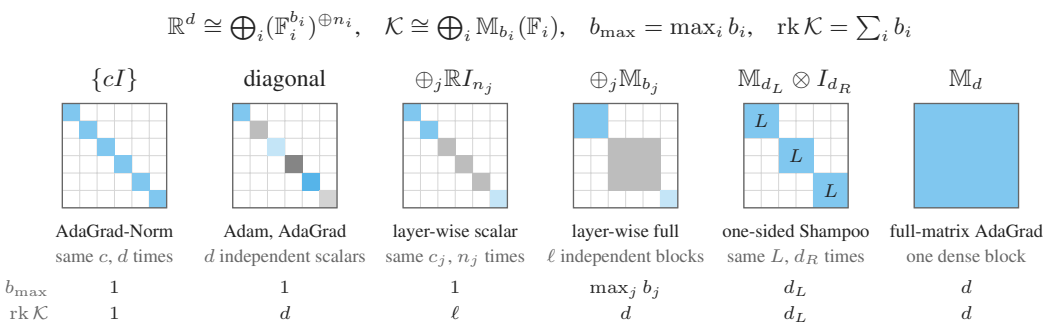

Figure 3: **Well-structured preconditioner algebras as block patterns** ($d = 6$). Each coloured square is a block of free entries; squares of one colour are the same block repeated (tied), and white cells are zero. Each colour is one simple component $\mathbb{M}_{b_i}(\mathbb{R})$ of (3): $b_{\max}$ is the side of the largest square, and $\operatorname{rk}\mathcal{K} = \sum_i b_i$ adds the sides of the distinct colours. These two numbers govern the two relaxations: the worst pointwise gap $K(\mathcal{H})$ lies in $[b_{\max}, K_G b_{\max}]$ (Theorem 3.1), and rotation is at most $\operatorname{rk}\mathcal{K}$ (Theorem 4.1). One-sided Shampoo repeats one $d_L \times d_L$ block d_R times, so $b_{\max} = \operatorname{rk}\mathcal{K} = d_L$. Drawn with layers of sizes $(2, 3, 1)$ and $d_L = 2$, $d_R = 3$.

(Lemma B.1). It measures the geometry (Lemma B.2):

$$\|x\|_{\mathcal{H}} := \sup_{H \in \mathcal{H},\ \operatorname{Tr} H \le 1} \sqrt{x^\top H x} = \|\pi_{\mathcal{K}}(xx^\top)\|_{\mathrm{op}}^{1/2}, \qquad \|g\|_{\mathcal{H},*} = \operatorname{Tr} \pi_{\mathcal{K}}(gg^\top)^{1/2}, \tag{2}$$

and the unit ball is $B_{\mathcal{H}} = \{x : \pi_{\mathcal{K}}(xx^\top) \preceq I\}$. Scalars $\{cI\}$ give AdaGrad-Norm and $\|x\|_2/\sqrt{d}$; diagonal matrices give Adam and AdaGrad and $\|x\|_\infty$; block-diagonal algebras give layer-wise variants; $\mathbb{M}_{d_L} \otimes I_{d_R}$, acting on $\mathbb{R}^{d_L \times d_R}$ by $X \mapsto LX$, gives one-sided Shampoo and ASGO (An et al., 2025) and $\|X\|_{\mathrm{op}}/\sqrt{d_R}$; $\mathbb{M}_d$ gives full-matrix AdaGrad and $\|x\|_2$.

Two invariants. A transpose-closed unital real algebra decomposes orthogonally as

$$\mathbb{R}^d \cong \bigoplus_{i=1}^{m} (\mathbb{F}_i^{b_i})^{\oplus n_i}, \qquad \mathcal{K} \cong \bigoplus_{i=1}^{m} \mathbb{M}_{b_i}(\mathbb{F}_i), \qquad \mathbb{F}_i \in \{\mathbb{R}, \mathbb{C}, \mathbb{H}\}, \tag{3}$$

where the same matrix acts on each of the n_i copies (Lemma B.5, from real Schur's lemma (Rosenberg, 2015, Theorem 1.3)). Two numbers of this decomposition govern everything below: the *block size* $b_{\max} := \max_i b_i$, and the *rank* $\operatorname{rk}\mathcal{K} := \sum_i b_i$, which is the largest number of mutually orthogonal nonzero projections in $\mathcal{K}$ (Lemma B.5(d)). Scalar preconditioning has $b_{\max} = \operatorname{rk}\mathcal{K} = 1$; diagonal preconditioning has $b_{\max} = 1$ and $\operatorname{rk}\mathcal{K} = d$; one-sided Shampoo has $b_{\max} = \operatorname{rk}\mathcal{K} = d_L$; layer-wise scalar step sizes on ℓ layers have $b_{\max} = 1$ and $\operatorname{rk}\mathcal{K} = \ell$. Always $b_{\max} \le \operatorname{rk}\mathcal{K} \le d$.

Three smoothness constants. For $A \in \mathbb{S}_d$ put $\|A\|_{\mathcal{H}} := \sup_{u,v \in B_{\mathcal{H}}} u^\top A v$ and $\nu_{\mathcal{H}}(A) := \min\{\operatorname{Tr} H : H \in \mathcal{H},\ -H \preceq A \preceq H\}$. For $f \in C^2(\mathbb{R}^d)$ define

$$\begin{aligned} L_{\mathcal{H}}(f) &:= \sup_x \|\nabla^2 f(x)\|_{\mathcal{H}}, \qquad \Lambda^\circ_{\mathcal{H}}(f) := \sup_x \nu_{\mathcal{H}}(\nabla^2 f(x)), \\ \Lambda_{\mathcal{H}}(f) &:= \min\left\{ \operatorname{Tr} H :\ H \in \mathcal{H},\ -H \preceq \nabla^2 f(x) \preceq H\ \ \forall x \in \mathbb{R}^d \right\}. \end{aligned} \tag{4}$$

$L_{\mathcal{H}}$ is the standard smoothness, the Lipschitz constant of ∇f from $\|\cdot\|_{\mathcal{H}}$ to $\|\cdot\|_{\mathcal{H},*}$, which governs NSD; $\Lambda_{\mathcal{H}}$ is the adaptive smoothness that governs AdaGrad-type methods (Definitions 2.3 and 2.4 of Xie et al., 2026); $\Lambda^\circ_{\mathcal{H}}$ is new. We assume $L_{\mathcal{H}}(f) < \infty$ throughout. In diagonal geometry $L_{\mathcal{H}}(f) = L_\infty(f)$, $\|A\|_{\mathcal{H}} = \|A\|_{\infty \to 1}$ is NP-hard to compute, and $\nu_{\mathcal{H}}(A)$ is a semidefinite program.

Probes. A *probe* is a finitely supported random triple (X, u, v) in $(\mathbb{R}^d)^3$: a point X of the landscape and two directions. It reads the curvature $\mathbb{E}[u^\top \nabla^2 f(X) v]$. Its directions pay a budget measured by $\pi_{\mathcal{K}}$, enforced at one of three strengths:

$$\textbf{surely: } \pi_{\mathcal{K}}(uu^\top) \preceq I \text{ a.s.;} \quad \textbf{pointwise: } \mathbb{E}[\pi_{\mathcal{K}}(uu^\top) \mid X] \preceq I; \quad \textbf{globally: } \mathbb{E}[\pi_{\mathcal{K}}(uu^\top)] \preceq I, \tag{5}$$

and the same for v. Each budget implies the next. In diagonal geometry $\pi_{\mathcal{K}}(uu^\top) = \operatorname{diag}(u_1^2, \ldots, u_d^2)$, and the three budgets are $|u_i| \le 1$ surely, $\mathbb{E}[u_i^2 \mid X] \le 1$, and $\mathbb{E}[u_i^2] \le 1$.

Theorem 2.1 (the probing principle)**.** Let $\mathcal{H}$ be well-structured and $f \in C^2(\mathbb{R}^d)$ with $L_{\mathcal{H}}(f) < \infty$. The suprema of the readings $\mathbb{E}[u^\top \nabla^2 f(X) v]$ over probes obeying the sure, pointwise and global budgets are $L_{\mathcal{H}}(f)$, $\Lambda^\circ_{\mathcal{H}}(f)$ and $\Lambda_{\mathcal{H}}(f)$, respectively. If f is convex, the same holds for probes with $v = u$.

The sure level restates the definition of $\|\cdot\|_{\mathcal{H}}$. At a fixed point, the semidefinite program $\nu_{\mathcal{H}}(A)$ has the dual $\max\{\langle A, Y - Z\rangle : Y, Z \succeq 0,\ \pi_{\mathcal{K}}(Y + Z) \preceq I\}$ (Lemma D.1), and every dual point is the symmetrized second moment $\mathbb{E}[uv^\top]$ of a probe whose budget is $\pi_{\mathcal{K}}(Y + Z)$. At the global level the dual of the common-majorant program spreads such pairs over many points under one shared budget. A compactness argument reduces the landscape to finitely many Hessians (Appendix C).

The decomposition. Because each budget implies the next, $L_{\mathcal{H}} \le \Lambda^\circ_{\mathcal{H}} \le \Lambda_{\mathcal{H}}$. Define the Grothendieck constants of $\mathcal{H}$ and the rotation number of f by

$$K(\mathcal{H}) := \sup_{0 \neq A \in \mathbb{S}_d} \frac{\nu_{\mathcal{H}}(A)}{\|A\|_{\mathcal{H}}}, \qquad k(\mathcal{H}) := \sup_{0 \neq A \in \mathbb{S}_d^+} \frac{\nu_{\mathcal{H}}(A)}{\|A\|_{\mathcal{H}}}, \qquad \varrho_{\mathcal{H}}(f) := \frac{\Lambda_{\mathcal{H}}(f)}{\Lambda^\circ_{\mathcal{H}}(f)}, \tag{6}$$

with $\varrho_{\mathcal{H}}(f) := 1$ if $\Lambda^\circ_{\mathcal{H}}(f) = 0$. Then

$$L_{\mathcal{H}}(f) \le \Lambda^\circ_{\mathcal{H}}(f) \le \Lambda_{\mathcal{H}}(f) = \varrho_{\mathcal{H}}(f)\, \Lambda^\circ_{\mathcal{H}}(f) \le \varrho_{\mathcal{H}}(f)\, K(\mathcal{H})\, L_{\mathcal{H}}(f), \tag{7}$$

with $k(\mathcal{H})$ in place of $K(\mathcal{H})$ when f is convex. The first relaxation depends on the algebra alone; the second depends on the landscape. Sections 3 and 4 bound them. The rotation number is unrelated to the norm distortion ρ of Xie et al. (2026).

Remark 2.2 (Pietsch measures). For diagonal $\mathcal{H}$ and convex f with $\nabla^2 f(x) = B_x^\top B_x$, the budget $\mathbb{E}[u_i^2] \le 1$ defines weakly 2-summable sequences in ℓ_∞^d, so Theorem 2.1 reads $\Lambda^\circ_{\mathcal{H}}(f) = \sup_x \pi_2(B_x)^2$ and $\Lambda_{\mathcal{H}}(f) = \pi_2(T)^2$ for $T : \ell_\infty^d \to \ell_\infty(\mathbb{R}^d; \ell_2)$, $(Tv)(x) = B_x v$, with optimal preconditioners as Pietsch measures (Theorem A.5). For the full algebra $\nu_{\mathcal{H}}(A) = \operatorname{Tr} A$ on $A \succeq 0$, a squared Hilbert–Schmidt norm; general algebras interpolate, which is why the block size enters Theorem 3.1.

3 Derandomization: a Grothendieck constant for every preconditioner algebra

At a single point a pointwise budget constrains only the probe's second moments, while a sure budget constrains every realization. By Lemma D.1, $\nu_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{H}}$ are the support functions of

$$\mathcal{W}_{\mathcal{H}} = \{Y - Z :\ Y, Z \succeq 0,\ \pi_{\mathcal{K}}(Y + Z) \preceq I\} \quad \text{and} \quad \mathcal{C}_{\mathcal{H}} = \operatorname{conv}\{\tfrac{1}{2}(uv^\top + vu^\top) : u, v \in B_{\mathcal{H}}\},$$

and $K(\mathcal{H})$ is the least K with $\mathcal{W}_{\mathcal{H}} \subseteq K\,\mathcal{C}_{\mathcal{H}}$ (Remark D.2): the integrality gap of a semidefinite relaxation. In diagonal geometry it is Grothendieck's relaxation of the $\infty \to 1$ norm (Alon & Naor, 2006).

Let $K_G^{(n)}$ be the real Grothendieck constant for $n \times n$ matrices. Then $K_G^{(n)} \uparrow K_G < 1.783$, and for positive semidefinite matrices the little Grothendieck inequality has the optimal dimension-free constant $\pi/2$ (Grothendieck, 1953; Krivine, 1979; Braverman et al., 2013; Pisier, 2012).

Theorem 3.1 (Grothendieck constants of preconditioner algebras)**.** Let $\mathcal{H} = \mathbb{S}_d^+ \cap \mathcal{K}$ be well-structured, with block size $b_{\max}$ as in (3). Then

$$b_{\max} \le k(\mathcal{H}) \le K(\mathcal{H}) \le K_G^{(d)}\, b_{\max} < 1.783\, b_{\max}, \qquad k(\mathcal{H}) \le \tfrac{\pi}{2}\, b_{\max}. \tag{8}$$

More precisely:

(a) If $\mathcal{K}$ has one simple component, then $k(\mathcal{H}) = K(\mathcal{H}) = b_1$, for every $\mathbb{F}_1$ and multiplicity n_1.

(b) If $\mathcal{K}$ is commutative, then $K(\mathcal{H}) \le K_G^{(d)}$ and $k(\mathcal{H}) \le \pi/2$.

(c) For the diagonal algebra and $d \ge 2$, $K_G^{(\lfloor d/2 \rfloor)} \le K(\mathcal{D}_d^+) \le K_G^{(d)}$. Moreover $\sup_d K(\mathcal{D}_d^+) = K_G$ and $\sup_d k(\mathcal{D}_d^+) = \pi/2$.

The supremum defining $K(\mathcal{H})$ ranges over all $A \in \mathbb{S}_d$, including matrices outside $\mathcal{K}$.

Two kinds of randomness. Derandomizing a probe has two parts. Its amplitudes can be rounded to signs: after weighting coordinates by their root-mean-square energy, Grothendieck's inequality

Table 1: The two invariants and the resulting constants for common preconditioner sets. Arrows denote suprema over dimension. The last column is the sharp bound on both $\varrho_{\mathcal{H}}(f)$ and $\Lambda_{\mathcal{H}}(f)/L_{\mathcal{H}}(f)$ (Theorem 4.1); Xie et al. (2026) give d.

method	$\mathcal{H}$	$\|x\|_{\mathcal{H}}$	$b_{\max}$	$K(\mathcal{H})$	$k(\mathcal{H})$	$\operatorname{rk}\mathcal{K}$
AdaGrad-Norm	$\{cI\}$	$\|x\|_2/\sqrt{d}$	1	1	1	1
Adam, AdaGrad	diagonal	$\|x\|_\infty$	1	$\to K_G$	$\to \pi/2$	d
layer-wise scalar	$\oplus_j \mathbb{R} I_{n_j}$	$\max_j \|x_j\|_2/\sqrt{n_j}$	1	$\le K_G$	$\le \pi/2$	ℓ
layer-wise full	$\oplus_j \mathbb{M}_{b_j}$	$\max_j \|x_j\|_2$	$\max_j b_j$	$[b_{\max}, K_G b_{\max}]$	$[b_{\max}, \frac{\pi}{2} b_{\max}]$	d
one-sided Shampoo	$\mathbb{M}_{d_L} \otimes I_{d_R}$	$\|X\|_{\mathrm{op}}/\sqrt{d_R}$	d_L	d_L	d_L	d_L
full-matrix AdaGrad	$\mathbb{S}^+_d$	$\|x\|_2$	d	d	d	d

turns a probe with $\mathbb{E}[\pi_{\mathcal{K}}(uu^\top)] \preceq I$ into sign vectors at cost K_G, or $\pi/2$ for positive semidefinite matrices, where the rounding is Gaussian. Its direction inside a matrix block cannot be rounded. In a block $\mathbb{M}_b \otimes I_n$ a pointwise budget asks $\mathbb{E}[U_i U_i^\top] \preceq nI$ of the $b \times n$ matrix U_i of the probe, while a sure budget asks $U_i U_i^\top \preceq nI$ of every realization. A uniformly random direction inside one copy $\mathbb{F}^b$ meets the first budget on average, yet each of its realizations spends b times its share. Formally, the rounded sign vectors have energy $\|u_i\|_2^2 \le b_i r_i$ in component i, with $r_i = n_i \dim_{\mathbb{R}} \mathbb{F}_i$, which places them in $\sqrt{b_{\max}}\, B_{\mathcal{H}}$ (Lemma B.5); the projector onto one irreducible copy in a largest block attains $b_{\max}$. No noncommutative Grothendieck inequality is needed (Appendix D).

In diagonal geometry, therefore, any dimension-growing gap between Adam's and sign descent's smoothness constants must come from the landscape. A matrix preconditioner can already pay its full block size on a single quadratic: for one-sided Shampoo, $f(X) = \frac{1}{2}\|Xe_1\|_2^2$ has standard constant d_R and adaptive constant $d_L d_R$, with $\varrho_{\mathcal{H}}(f) = 1$. Balles et al. (2020) showed that separable smoothness, the adaptive constant of a common diagonal majorant, implies ℓ_∞-smoothness; Theorems 3.1 and 4.1 quantify the converse.

4 Commitment: Curvature Rotation and the Rank of the Algebra

A globally budgeted probe may read different directions at different points. Rotation measures what this buys, and the answer is bounded by how many directions the algebra can weigh independently.

Theorem 4.1 (rotation is bounded by the rank)**.** Let $\mathcal{H}$ be well-structured. For every $f \in C^2(\mathbb{R}^d)$ with $L_{\mathcal{H}}(f) < \infty$,

$$\Lambda_{\mathcal{H}}(f) \;\le\; \operatorname{rk}\mathcal{K} \cdot L_{\mathcal{H}}(f), \qquad \text{hence} \qquad 1 \;\le\; \varrho_{\mathcal{H}}(f) \;\le\; \operatorname{rk}\mathcal{K}.$$

Both upper bounds are sharp, already among convex functions: for every $\delta > 0$ there is a convex $f \in C^\infty(\mathbb{R}^d)$ with bounded Hessian such that $\varrho_{\mathcal{H}}(f) \ge \operatorname{rk}\mathcal{K} - \delta$ and $\Lambda_{\mathcal{H}}(f) \ge (\operatorname{rk}\mathcal{K} - \delta) L_{\mathcal{H}}(f)$.

Proof of the upper bound. In the coordinates of (3), let $P_i \in \mathcal{K}$ project onto component i, put $r_i = n_i \dim_{\mathbb{R}} \mathbb{F}_i$, and let $S = \sum_i \sqrt{r_i} P_i$. The Euclidean ball satisfies $SB_2 \subseteq B_{\mathcal{H}}$ (Lemma B.5), so $\|SAS\|_{\mathrm{op}} \le \|A\|_{\mathcal{H}}$ for every $A \in \mathbb{S}_d$. For a globally budgeted probe, $P_i \in \mathcal{K}$ and the self-adjointness of $\pi_{\mathcal{K}}$ give

$$\mathbb{E}\|S^{-1}u\|_2^2 = \sum_i \frac{\operatorname{Tr}(P_i\, \mathbb{E}[uu^\top])}{r_i} = \sum_i \frac{\operatorname{Tr}(P_i\, \mathbb{E}[\pi_{\mathcal{K}}(uu^\top)])}{r_i} \le \sum_i \frac{\operatorname{Tr} P_i}{r_i} = \sum_i b_i = \operatorname{rk}\mathcal{K},$$

and likewise for v. Hence $\mathbb{E}[u^\top \nabla^2 f(X) v] = \mathbb{E}[(S^{-1}u)^\top S \nabla^2 f(X) S (S^{-1}v)] \le L_{\mathcal{H}}(f)\, \mathbb{E}[\|S^{-1}u\|_2 \|S^{-1}v\|_2] \le \operatorname{rk}\mathcal{K} \cdot L_{\mathcal{H}}(f)$ by Cauchy–Schwarz. Theorem 2.1 turns this into $\Lambda_{\mathcal{H}} \le \operatorname{rk}\mathcal{K} \cdot L_{\mathcal{H}}$, and $\varrho_{\mathcal{H}} \le \Lambda_{\mathcal{H}}/L_{\mathcal{H}}$ because $L_{\mathcal{H}} \le \Lambda^\circ_{\mathcal{H}}$. □

In words, a global budget buys at most $\operatorname{rk}\mathcal{K}$ units of rescaled Euclidean energy, one per direction the algebra can weigh. The bound is approached by convex functions in which each direction bends only where all the others are saturated. In diagonal geometry, let $h(t) = \log(1 + e^t)$, $\tau_k(x) = 1 + \beta \sum_{j \ne k} h(x_j + M)$ and $f(x) = \sum_k \tau_k(x)\, h(x_k/\tau_k(x))$. Each term is a perspective of h, which

is increasing in its scale, evaluated at the convex scale τ_k; hence it is convex. Coordinate k bends with curvature near $\frac{1}{4}$ only where every other coordinate lies below about $-M$, far from its own bending window; elsewhere τ_k dilutes its curvature. For small β and large M, no Hessian has ν much above $\frac{1}{4}$, yet a common majorant must pay $\frac{1}{4}$ in every coordinate (Appendix C, for every algebra). Since $\operatorname{rk}\mathcal{K} = d$ only when every block is real and appears once, Theorem 4.1 sharpens the factor d in $\Lambda_{\mathcal{H}} \le d\, L_{\mathcal{H}}$ (Xie et al., 2026, Proposition 2.5) whenever a block repeats or is complex or quaternionic. For one-sided Shampoo the factor is d_L rather than $d_L d_R$; for layer-wise scalar step sizes it is the number of layers; for AdaGrad-Norm the three constants coincide.

Curvature that does and does not rotate. Rotation is large when different directions are most curved at different points (proofs in Appendix C).

Proposition 4.2 (rotation of basic landscapes)**.** *(i)* For every quadratic $f(x) = \frac{1}{2}x^\top Ax + b^\top x$, $\varrho_{\mathcal{H}}(f) = 1$ and $\Lambda_{\mathcal{H}}(f) = \nu_{\mathcal{H}}(A)$; hence $\sup_{f\ \text{quadratic}} \Lambda_{\mathcal{H}}(f)/L_{\mathcal{H}}(f) = K(\mathcal{H})$, and $= k(\mathcal{H})$ over convex quadratics. *(ii)* Let f be convex. If $\nabla^2 f(x) \preceq \nabla^2 f(x_0) =: B$ for some x_0 and all x, then $\Lambda_{\mathcal{H}}(f) = \Lambda^\circ_{\mathcal{H}}(f) = \nu_{\mathcal{H}}(B)$ and $\varrho_{\mathcal{H}}(f) = 1$ for every well-structured $\mathcal{H}$. This holds for weighted binary logistic regression

$$f(x) = \sum_{j=1}^{n} w_j \log(1 + e^{-y_j a_j^\top x}) + \frac{\lambda}{2}\|x\|_2^2, \qquad w_j \ge 0, \quad y_j \in \{\pm 1\}, \quad \lambda \ge 0, \tag{9}$$

with $x_0 = 0$, because the logistic second derivative peaks at zero. *(iii)* Let $d \ge 2$, $\tau > 0$ and $\operatorname{lse}_\tau(x) = \tau \log \sum_{i=1}^d e^{x_i/\tau}$. In diagonal geometry $L_{\mathcal{H}}(\operatorname{lse}_\tau) = \Lambda^\circ_{\mathcal{H}}(\operatorname{lse}_\tau) = 1/\tau$ and $\Lambda_{\mathcal{H}}(\operatorname{lse}_\tau) = d/(2\tau)$. For the full algebra $\Lambda^\circ_{\mathcal{H}}(\operatorname{lse}_\tau) = (d-1)/(d\tau)$ and $\Lambda_{\mathcal{H}}(\operatorname{lse}_\tau) = (d-1)/(2\tau)$. In both geometries $\varrho_{\mathcal{H}}(\operatorname{lse}_\tau) = d/2 = \operatorname{rk}\mathcal{K}/2$.

In the probe picture, log-sum-exp is the prototypical rotating landscape. Its Hessian $\frac{1}{\tau}(\operatorname{diag}(p) - pp^\top)$, with $p = \operatorname{softmax}(x/\tau)$, is dominated at every point by $\operatorname{diag}(p)/\tau$, of trace $1/\tau$; a probe confined to one point reads at most $1/\tau$. Near $p = \frac{1}{2}(e_i + e_j)$ the curvature concentrates on the pair $\{i, j\}$, and a probe that picks a random pair, moves to the corresponding point and reads along $e_i - e_j$ meets the budget $\mathbb{E}[u_i^2] \le 1$ while reading $d/(2\tau)$. Its curvature always couples two coordinates, which is why it pays half the rank, whereas the saturated construction bends one coordinate at a time and pays nearly all of it. In logistic regression, by contrast, all curvature peaks at the same point, so moving the probe gains nothing. A learned intercept can be included as a feature; with fixed offsets, (ii) still applies if all logits can vanish at a common point.

5 Rotation is an acceleration horizon

We now specialize to the diagonal geometry $\mathcal{H} = \mathcal{D}_d^+$ of Adam and AdaGrad. Theorems 3.1 and 4.1 give

$$\Lambda_{\mathcal{H}}(f) \le K_G\, \varrho_{\mathcal{H}}(f)\, L_{\mathcal{H}}(f), \qquad \Lambda_{\mathcal{H}}(f) \le \tfrac{\pi}{2}\, \varrho_{\mathcal{H}}(f)\, L_{\mathcal{H}}(f) \ \text{ if } f \text{ is convex}, \qquad \varrho_{\mathcal{H}}(f) \le d, \tag{10}$$

with $L_{\mathcal{H}}(f) = L_\infty(f)$. Adaptive methods pay the standard constant times the rotation; can any method avoid that payment? We place an accelerated upper bound and an oracle lower bound in one constrained model.

Acceleration with a feasible output. Let $X_R = \{x : \|x\|_{\mathcal{H}} \le R\}$ for any well-structured $\mathcal{H}$; in diagonal geometry this is the box $\|x\|_\infty \le R$. For C^1 objectives, $\Lambda_{\mathcal{H}}$ is extended through Lipschitz constants (Appendix E). The following method adapts a $\mathcal{K}$-valued metric to gradient differences on any convex compact X with $\|x - y\|_{\mathcal{H}} \le 2R$ on X, such as X_R. Put $a_t = t$, $A_t = t(t+1)/2$, $D = \sqrt{2}R$, $z_0 \in X$ and $G_0 = H_0 = 0$, and for $t = 1, \dots, N$ let

$$\begin{aligned}
&\tilde{x}_t = \tfrac{A_{t-1}\bar{x}_{t-1} + a_t z_{t-1}}{A_t}, \ \ m_t = a_t \nabla f(\tilde{x}_t), \ \ w_t \in \operatorname*{arg\,min}_{w \in X} \big\{\langle m_t, w\rangle + \tfrac{1}{2}\|w - z_{t-1}\|^2_{H_{t-1}}\big\},\\
&\bar{x}_t = \tfrac{A_{t-1}\bar{x}_{t-1} + a_t w_t}{A_t}, \ g_t = a_t \nabla f(\bar{x}_t), \ u_t = g_t - m_t, \ G_t = G_{t-1} + \pi_{\mathcal{K}}(u_t u_t^\top), \ H_t = G_t^{1/2}/D,\\
&z_t \in \operatorname*{arg\,min}_{z \in X} \big\{\langle g_t, z\rangle + \tfrac{1}{2}\|z - z_{t-1}\|^2_{H_{t-1}} + \tfrac{1}{2}\|z - w_t\|^2_{H_t - H_{t-1}}\big\}.
\end{aligned} \tag{11}$$

Theorem 5.1 (basis-free adaptive acceleration)**.** Let f be convex and C^1. After $N \geq 1$ rounds of (11), which make $2N$ gradient queries in X, the output $\bar{x}_N \in X$ satisfies

$$f(\bar{x}_N) - \min_X f \leq \min\left\{\frac{24\sqrt{2}\, L_{\mathcal{H}}(f)R^2}{N+1},\ \frac{32\,\Lambda_{\mathcal{H}}(f)R^2}{N(N+1)}\right\}. \tag{12}$$

The method needs R, but no smoothness constant, rotation number or preconditioner, and it attains both regimes of Theorem 5.2 below. Its proof (Appendix E.1) never chooses a basis: each metric increment is centred at the current play, and the completion of squares $2c \operatorname{Tr} S - \operatorname{Tr}(SB^{+}S) \leq c^2 \operatorname{Tr} B$, for $S \succeq 0$ with $\operatorname{range} S \subseteq \operatorname{range} B$, lets the budget $\operatorname{Tr} B$ of any common majorant pay for the learned metric. On the ball X_R the accelerated rate was also obtained, by a projection-free method, by Kovalev (2026); in diagonal geometry the theorem removes the logarithm of AdaACSA (Ene et al., 2021, Theorem 3.3). For convex C^2 objectives (10) bounds the second term by $16\pi\varrho_{\mathcal{H}}(f)L_\infty(f)R^2/(N(N+1))$, the standard accelerated rate when $\varrho_{\mathcal{H}} = 1$, as for convex quadratics and binary logistic regression.

The horizon is sharp. We measure the risk of deterministic methods with the first-order oracle $(f(x), \nabla f(x))$; our lower bounds also hold for the oracle that returns the entire germ of f at x. Following Guzmán & Nemirovski (2015), a T-step method outputs its final query $x_T \in X_R$, determined by the first $T-1$ answers, and $\mathrm{Risk}_{\mathcal{C},X_R}(T) = \inf_{\text{methods}} \sup_{f\in\mathcal{C}}[f(x_T) - \min_{X_R} f]$.

Theorem 5.2 (rotation is the acceleration horizon)**.** Let $L, R > 0$ and integers $3 \leq T \leq d$ and $k \geq 2$. Define

$$\mathcal{F}_{L,k} = \{f \text{ convex},\ C^1 :\ L_\infty(f) \leq L,\ \Lambda_{\mathcal{D}_d^+}(f) \leq kL_\infty(f)\},$$
$$\mathcal{R}_{L,k} = \{f \text{ convex},\ C^2 :\ L_\infty(f) \leq L,\ \varrho_{\mathcal{H}}(f) \leq k\}.$$

For $\mathcal{C} = \mathcal{F}_{L,k}$ and for $\mathcal{C} = \mathcal{R}_{L,k}$,

$$\frac{c\,LR^2}{\ln T}\min\left\{\frac{1}{T},\ \frac{k}{T^2}\right\} \leq \mathrm{Risk}_{\mathcal{C},X_R}(T) \leq \min\left\{\frac{8LR^2}{T+1},\ \frac{CkLR^2}{T^2}\right\}, \tag{13}$$

where $c, C > 0$ are absolute constants. The lower bound also holds on $\mathcal{F}_{L,k} \cap C^2$.

Up to one logarithm, the minimax risk is LR^2/T while $T \leq k$, the rate of conditional gradient, which needs no preconditioner at all, and kLR^2/T^2 once $T \geq k$, the rate of adaptive acceleration: the curve bends at the rotation budget. The upper bounds are conditional gradient and Theorem 5.1; the lower bound comes from one family whose smoothing exponent dials the rotation.

Proposition 5.3 (the exponent is a rotation dial)**.** Let $m \geq 3$, $r > 2$, $q = r/(r-1)$ and $\epsilon > 0$. The smoothed maximum $F_\epsilon(z) = \max_{\lambda\in\Delta_m}\{\langle z, \lambda\rangle - \frac{\epsilon}{8}\|\lambda\|_q^2\}$ of Guzmán & Nemirovski (2015) is convex and C^2, and in diagonal geometry

$$\frac{2^{2/r}}{\pi}\, m^{1-2/r} \leq \varrho_{\mathcal{H}}(F_\epsilon) \leq \frac{\Lambda_{\mathcal{D}_d^+}(F_\epsilon)}{L_\infty(F_\epsilon)} \leq \frac{m^{1-2/r}}{1-m^{-2}}. \tag{14}$$

Against any deterministic method, a scaled, translated, signed-coordinate copy on $m = T \leq d$ coordinates with $L_\infty \leq L$ has error at least $LR^2/(32(r-1)T^{1+2/r})$ (Proposition F.3). Choosing $T^{1-2/r} \approx \min\{k, T\}$ gives the lower bound of Theorem 5.2.

Near a tie of two coordinates the Hessian is supported on that pair, which forces the lower bound; at the all-way tie it is a multiple of $I - \frac{1}{m}\mathbf{1}\mathbf{1}^\top$, which fixes L_∞ (Appendix F). The exponent trades non-accelerated hardness ($r \approx 3\ln T$, rotation $\approx T$) for accelerated hardness at exactly the exchange rate of the rotation. The statement is about the class: the hard function may depend on the method and the horizon.

6 The two relaxations along training

We record exact Hessians at 22 iterations while training five small models with full-batch Adam: binary logistic regression ($d = 10$), softmax regression on $C \in \{3, 5, 8\}$ digit classes ($d = 6C$), and

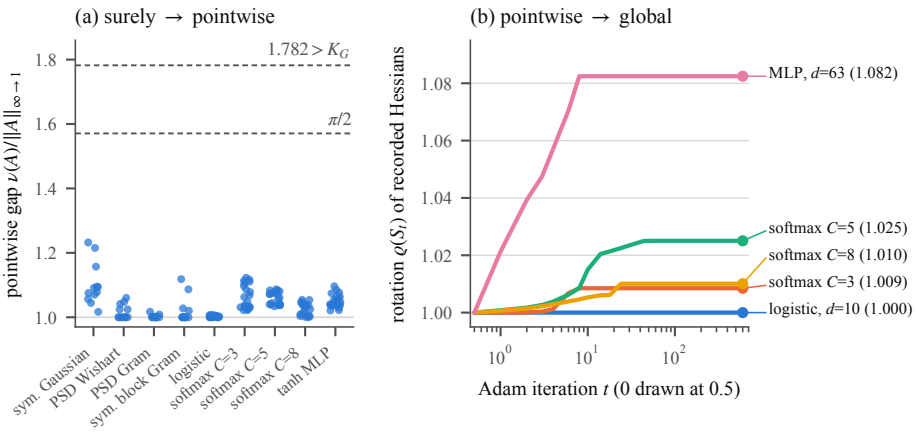

Figure 4: **Both relaxations are small along these training runs. (a)** Pointwise gaps of 48 random matrices ($d = 8$, exact $\|A\|_{\infty\to 1}$) and of the 22 Hessians recorded per model; for $d > 18$ the plotted value is an upper bound (Appendix I). Dashed lines: $\pi/2$ (positive semidefinite) and Krivine's bound $1.782 > K_G$ (Theorem 3.1). **(b)** Rotation of the Hessians recorded up to iteration t; it is one for logistic regression, whose initial Hessian dominates (Proposition 4.2(ii)). Theorems 3.1 and 4.1 bound (a) by K_G ($\pi/2$ if PSD) and (b) by d.

a one-hidden-layer $\tanh$ network ($d = 63$) (Appendix I). In diagonal geometry we compute each pointwise gap $\nu(A)/\|A\|_{\infty\to 1}$ and, for the set S_t of Hessians recorded up to iteration t, the rotation $\varrho(S_t) = \Lambda(S_t)/\max_{A\in S_t}\nu(A)$ of the curvature the run has met at these checkpoints.

Neither relaxation comes close to its worst case on these runs. Every pointwise gap of a recorded Hessian is below 1.13, against the bounds $\pi/2$ and K_G. Rotation, which convex objectives can push to d (Theorem 4.1), is exactly 1 for binary logistic regression, as Proposition 4.2(ii) predicts, between 1.008 and 1.026 for softmax regression, and 1.082 for the network. It accumulates early: for the network it is reached by iteration 8, after which the curvature shrinks inside the majorant already paid for. The whole factor $\Lambda(S_t)/L_\infty(S_t)$ between the two constants stays below 1.15 for every model, while d ranges up to 63. On these runs the optimizer never meets the dimension factor of $\Lambda_{\mathcal{H}} \le d\, L_{\mathcal{H}}$. Large rotations need curvature that moves across coordinates, as in the hard family of Section 5 or in log-sum-exp (Proposition 4.2(iii)), and in diagonal geometry only rotation can grow with the dimension.

7 Discussion

Standard, pointwise and adaptive smoothness are one probing problem read at three levels of information. Conditioning the budget on the cell of X in a partition $\mathcal{P}$ of the landscape gives $\Lambda_{\mathcal{P}}(f) = \sup_{C\in\mathcal{P}} \Lambda_{\mathcal{H}}(f|_C)$, the constant of a preconditioner that may change from cell to cell, between $\Lambda^\circ_{\mathcal{H}}$ and $\Lambda_{\mathcal{H}}$ (Appendix C); it is a spatial analogue of Adam's exponential forgetting, though a rate for Adam requires an analysis of its update rule.

What the decomposition separates. Benchmarks of preconditioners usually conflate two questions. Whether an algebra can represent the curvature at a point is a question about its block size, answered by Theorem 3.1; whether one preconditioner can serve the whole run is a question about rotation, which Theorem 4.1 bounds by the rank and Section 6 measures along training. Algebras with large blocks, such as one-sided Shampoo, can pay the first price in full on a single quadratic. Diagonal algebras pay at most $K_G < 1.783$ at a point, so any dimension-growing gap between Adam and sign descent is rotation, which the runs of Section 6 find small. **Open problems.** (i) For several interacting matrix blocks, $K(\mathcal{H})$ lies between $b_{\max}$ and $K_G b_{\max}$; the exact constant is unknown. (ii) The accelerated rate of Theorem 5.1 holds in every well-structured geometry, and one-sided Shampoo inherits the whole sandwich of Theorem 5.2 for $T \le \min\{d_L, d_R\}$ (Proposition E.5); whether the logarithm in the lower bound is needed is unknown. (iii) Which natural losses rotate: log-sum-exp pays half the rank, saturated functions nearly all of it.

AI USE STATEMENT

The original research ideas and initial manuscript draft were developed by the authors. AI tools were used to assist with checking mathematical details and the correctness of arguments, filling in routine proof steps, drawing TikZ figures, coding assistance and language polishing. The authors take full responsibility for the content and correctness of the final manuscript.

REPRODUCIBILITY STATEMENT

The appendix contains complete proofs, including the landscape duality behind the probing principle, the rank bound and its sharpness, the structural and Grothendieck arguments, the oracle bounds, and the noise calculations. Appendix I specifies the numerical protocols. The anonymous supplementary material contains the scripts, fixed configurations, raw outputs, environment information, and commands that regenerate every figure.

REFERENCES

Noga Alon and Assaf Naor. Approximating the cut-norm via Grothendieck's inequality. *SIAM Journal on Computing*, 35(4):787–803, 2006. doi: 10.1137/S0097539704441629. Conference version: STOC 2004, pp. 72–80, doi:10.1145/1007352.1007371.

Kang An, Yuxing Liu, Rui Pan, Yi Ren, Shiqian Ma, Donald Goldfarb, and Tong Zhang. ASGO: Adaptive structured gradient optimization. *arXiv preprint arXiv:2503.20762*, 2025. URL https://arxiv.org/abs/2503.20762.

Lukas Balles, Fabian Pedregosa, and Nicolas Le Roux. The geometry of sign gradient descent. *arXiv preprint arXiv:2002.08056*, 2020. URL https://arxiv.org/abs/2002.08056.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. doi: 10.1137/080716542.

Aharon Ben-Tal and Arkadi Nemirovski. *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*. MPS-SIAM Series on Optimization. SIAM, Philadelphia, PA, 2001. ISBN 9780898714913. doi: 10.1137/1.9780898718829.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*, 2024. URL https://arxiv.org/abs/2409.20325.

Mark Braverman, Konstantin Makarychev, Yury Makarychev, and Assaf Naor. The Grothendieck constant is strictly smaller than Krivine's bound. *Forum of Mathematics, Pi*, 1:e4, 2013. doi: 10.1017/fmp.2013.4. Conference version: FOCS 2011, pp. 453–462, doi:10.1109/FOCS.2011.77.

Ashok Cutkosky. Anytime online-to-batch, optimism and acceleration. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1446–1454, 2019.

Joe Diestel, Hans Jarchow, and Andrew Tonge. *Absolutely Summing Operators*, volume 43 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1995. ISBN 0521431689.

Alina Ene, Huy L. Nguyen, and Adrian Vladu. Adaptive gradient methods for constrained convex optimization and variational inequalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7314–7321, 2021. doi: 10.1609/aaai.v35i8.16898. URL https://ojs.aaai.org/index.php/AAAI/article/view/16898. Full version: arXiv:2007.08840v3.

Alexander Grothendieck. Résumé de la théorie métrique des produits tensoriels topologiques. *Bol. Soc. Mat. São Paulo*, 8:1–79, 1953. Reprinted in *Resenhas* **2** (1996), no. 4, 401–480.

Alexander Grothendieck. Récoltes et semailles: Réflexions et témoignage sur un passé de mathématicien. Université des Sciences et Techniques du Languedoc, Montpellier, 1985–1987.

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

Cristóbal Guzmán and Arkadi Nemirovski. On lower complexity bounds for large-scale smooth convex optimization. *Journal of Complexity*, 31(1):1–14, 2015. doi: 10.1016/j.jco.2014.08.003.

Steven Heilman. A lower bound for Grothendieck's constant. *arXiv preprint arXiv:2603.22616*, 2026. URL https://arxiv.org/abs/2603.22616.

Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, volume 28 of *Proceedings of Machine Learning Research*, pp. 427–435, Atlanta, Georgia, USA, 2013. PMLR. URL https://proceedings.mlr.press/v28/jaggi13.html.

Chris Jones and Giulio Malavolta. The Grothendieck constant is strictly larger than Davie–Reeds' bound. *arXiv preprint arXiv:2603.30039*, 2026. URL https://arxiv.org/abs/2603.30039.

Pooria Joulani, Anant Raj, András György, and Csaba Szepesvári. A simpler approach to accelerated optimization: Iterative averaging meets optimism. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4984–4993, 2020.

Ali Kavis, Kfir Y. Levy, Francis Bach, and Volkan Cevher. UniXGrad: A universal, adaptive algorithm with optimal guarantees for constrained optimization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Dmitry Kovalev. SGD with adaptive preconditioning: Unified analysis and momentum acceleration. *arXiv preprint arXiv:2506.23803*, 2025.

Dmitry Kovalev. Optimal projection-free adaptive SGD for matrix optimization. *arXiv preprint arXiv:2604.02505*, 2026.

Jean-Louis Krivine. Constantes de Grothendieck et fonctions de type positif sur les sphères. *Advances in Mathematics*, 31(1):16–30, 1979. doi: 10.1016/0001-8708(79)90017-3.

Joram Lindenstrauss and Aleksander Pełczyński. Absolutely summing operators in $\mathcal{L}_p$-spaces and their applications. *Studia Mathematica*, 29(3):275–326, 1968.

Colin McLarty. The rising sea: Grothendieck on simplicity and generality. In Jeremy J. Gray and Karen Hunger Parshall (eds.), *Episodes in the History of Modern Algebra (1800–1950)*, volume 32 of *History of Mathematics*, pp. 301–325. American Mathematical Society, Providence, RI, 2007.

H. Brendan McMahan. A survey of algorithms and analysis for adaptive online learning. *Journal of Machine Learning Research*, 18(90):1–50, 2017.

Yurii Nesterov. Semidefinite relaxation and nonconvex quadratic optimization. *Optimization Methods and Software*, 9(1-3):141–160, 1998. doi: 10.1080/10556789808805690.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 49069–49104. PMLR, 2025. URL https://proceedings.mlr.press/v267/pethick25a.html.

Albrecht Pietsch. Absolut p-summierende Abbildungen in normierten Räumen. *Studia Mathematica*, 28(3):333–353, 1967.

Gilles Pisier. Grothendieck's theorem, past and present. *Bulletin of the American Mathematical Society*, 49(2):237–323, 2012. doi: 10.1090/S0273-0979-2011-01348-9.

Ronald E. Rietz. A proof of the Grothendieck inequality. *Israel Journal of Mathematics*, 19(3): 271–276, 1974. doi: 10.1007/BF02757725.

Jonathan Rosenberg. Structure and applications of real C^*-algebras, 2015. URL https://arxiv.org/abs/1505.04091. Version 1, Theorem 1.3, p. 2.

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

Shuo Xie, Mohamad Amin Mohamadi, and Zhiyuan Li. Adam exploits ℓ_∞-geometry of loss landscape via coordinate-wise adaptivity. In *International Conference on Learning Representations (ICLR)*, 2025a. URL https://openreview.net/forum?id=PUnD86UEK5. arXiv:2410.08198.

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 (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 68746–68776. PMLR, 2025b. URL https://proceedings.mlr.press/v267/xie25j.html.

Shuo Xie, Tianhao Wang, Beining Wu, and Zhiyuan Li. A tale of two geometries: Adaptive optimizers and non-Euclidean descent. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/7c79d0964c8fc3674dcf1737e8afe89f-Abstract-Conference.html. ICLR conference version; theorem numbering follows this version.

Roadmap. The notation table comes first. Appendix A collects the external results we rely on, with precise statements. Appendix B proves the structural facts about well-structured preconditioner sets (Lemmas B.1–B.5). Appendix C proves the probing principle, the rank bound and its sharpness, and computes the examples. Appendix D proves the duality lemma and Theorem 3.1. Appendix E proves the oracle bounds and accelerated guarantees in Section 5. Appendix F establishes the smoothness, locality, and rotation of the explicit smoothed-max family. Appendix H proves the noise results. Appendix I documents the numerical illustrations. Appendix J relates the quotation under the abstract to the paper. Proofs are broken into numbered steps; displayed equations that are used later are numbered.

NOTATION

Table 2 collects the notation, with the place where each symbol is defined.

Table 2: Notation.

Symbol	Meaning	Defined in
Preconditioner geometry		
$\mathbb{S}_d$, $\mathbb{S}_d^+$	symmetric $d \times d$ matrices; the positive semidefinite cone	Section 2
$\mathcal{K}$, $\mathcal{K}_{\mathrm{s}}$	a unital subalgebra of $\mathbb{M}_d(\mathbb{R})$ closed under transposition; its symmetric part	Section 2
$\mathcal{H} = \mathbb{S}_d^+ \cap \mathcal{K}$	a well-structured preconditioner set	Section 2
$\pi_{\mathcal{K}}$	orthogonal projection onto $\mathcal{K}$, a conditional expectation	Lemma B.1
$\|x\|_{\mathcal{H}}$, $\|g\|_{\mathcal{H},*}$, $B_{\mathcal{H}}$	the norm $\|\pi_{\mathcal{K}}(xx^\top)\|_{\mathrm{op}}^{1/2}$, its dual $\operatorname{Tr} \pi_{\mathcal{K}}(gg^\top)^{1/2}$, the unit ball	(2)
$\mathbb{F}_i$, b_i, n_i	division algebra, block size and multiplicity of the ith component of $\mathcal{K}$	(3)
$b_{\max}$, $\operatorname{rk}\mathcal{K}$	the largest block size $\max_i b_i$; the rank $\sum_i b_i$	(3)
$\mathcal{D}_d^+$	diagonal preconditioners, the geometry of Adam and AdaGrad	Section 2
Smoothness constants and probes		
$\|A\|_{\mathcal{H}}$, $\nu_{\mathcal{H}}(A)$	$\sup_{u,v \in B_{\mathcal{H}}} u^\top A v$; the least trace of $H \in \mathcal{H}$ with $-H \preceq A \preceq H$	Section 2
$L_{\mathcal{H}}$, $\Lambda_{\mathcal{H}}^\circ$, $\Lambda_{\mathcal{H}}$	standard, pointwise adaptive and adaptive smoothness	(4)
L_∞, Λ°, Λ	the same three constants in diagonal geometry	Section 1
(X, u, v)	a probe: a random point and two random directions; its sure, pointwise and global budgets	(5)
$K(\mathcal{H})$, $k(\mathcal{H})$	Grothendieck constants of $\mathcal{H}$: the worst ratio $\nu_{\mathcal{H}}(A)/\|A\|_{\mathcal{H}}$ over symmetric, resp. positive semidefinite A	(6)
$\varrho_{\mathcal{H}}(f)$	the curvature rotation number $\Lambda_{\mathcal{H}}(f)/\Lambda_{\mathcal{H}}^\circ(f)$	(6)
K_G, $K_G^{(n)}$	the real Grothendieck constant and its $n \times n$ version	Theorem A.1
$\pi_2(\cdot)$	the 2-summing norm of an operator	Remark 2.2
lse_τ	log-sum-exp, $\tau \log \sum_i e^{x_i/\tau}$	Proposition 4.2
Oracle complexity and acceleration		
X_R	the ball $\{x : \|x\|_{\mathcal{H}} \le R\}$; the box $\|x\|_\infty \le R$ in diagonal geometry	Section 5
$\mathrm{Risk}_{\mathcal{C},X_R}(T)$	minimax risk of deterministic T-step methods over a class $\mathcal{C}$	Section 5
$\mathcal{F}_{L,k}$, $\mathcal{R}_{L,k}$	convex functions with $L_\infty \le L$ and $\Lambda \le kL_\infty$, resp. $\varrho_{\mathcal{H}} \le k$	Theorem 5.2
F_ϵ, r, q	the smoothed maximum with exponent r, and $q = r/(r-1)$	Proposition 5.3
a_t, A_t, D	averaging weights t and $t(t+1)/2$; the scale $\sqrt{2}R$	(11)
$\tilde{x}_t$, $\bar{x}_t$; w_t, z_t	the two gradient queries of round t; the play and the prox centre	(11)
m_t, g_t, u_t	scaled gradients at $\tilde{x}_t$ and $\bar{x}_t$, and their difference	(11)
G_t, H_t	$\sum_{s \le t} \pi_{\mathcal{K}}(u_s u_s^\top)$ and the learned metric $G_t^{1/2}/D$	(11)
Gradient noise		
$\xi_t(x)$, $\Sigma_t(x)$	gradient noise and its covariance	Section G
σ_{std}, σ°, $\sigma_{\mathcal{H}}$	standard, pointwise and adaptive noise levels	(39)
$\gamma_{\mathcal{H}}$, $\varrho_{\mathcal{H}}^\sigma$	the pointwise noise gap $\sigma^{\circ 2}/\sigma_{\mathrm{std}}^2$ and the noise rotation $\sigma_{\mathcal{H}}^2/\sigma^{\circ 2}$	Section G

A AUXILIARY RESULTS

This section collects, with precise statements and references, the external results that the proofs rely on. Nothing here is new.

A.1 GROTHENDIECK'S INEQUALITIES

Theorem A.1 (Grothendieck's inequality)**.** For every $n \geq 1$ there is a smallest constant $K_G^{(n)}$ such that for every real $n \times n$ matrix $a = (a_{ij})$ and all unit vectors $p_1, \dots, p_n, q_1, \dots, q_n$ of a real Hilbert space,

$$\sum_{i,j=1}^{n} a_{ij} \langle p_i, q_j \rangle \;\leq\; K_G^{(n)} \max_{\epsilon,\delta \in \{\pm 1\}^n} \sum_{i,j=1}^{n} a_{ij} \epsilon_i \delta_j. \tag{15}$$

The sequence $K_G^{(n)}$ is nondecreasing and converges to Grothendieck's constant K_G, and

$$1.676 < K_G < \frac{\pi}{2 \ln(1+\sqrt{2})} = 1.7822\ldots$$

Sources: Grothendieck (1953) (existence of K_G), Krivine (1979) (the upper bound $\pi/(2\ln(1+\sqrt{2}))$), Braverman et al. (2013) (strict inequality), Heilman (2026); Jones & Malavolta (2026) (lower bounds above 1.676); see Pisier (2012) for a survey. Since (15) is invariant under scaling a and is applied to arbitrary unit-vector families, it also holds for vectors of norm at most one (write $p_i = \|p_i\| \hat{p}_i$ and absorb the norms into a, which only decreases the right-hand side).

Theorem A.2 (little Grothendieck inequality)**.** For every $n \geq 1$, every positive semidefinite $a \in \mathbb{S}_n^+$ and all unit vectors $p_1, \dots, p_n$ of a real Hilbert space,

$$\sum_{i,j=1}^{n} a_{ij} \langle p_i, p_j \rangle \;\leq\; \frac{\pi}{2} \max_{\epsilon \in \{\pm 1\}^n} \epsilon^\top a\, \epsilon, \tag{16}$$

and the constant $\pi/2$ cannot be replaced by a smaller one that is valid for all n.

Sources: Grothendieck (1953); Rietz (1974); the SDP form and its use in optimization are due to Nesterov (1998); see Pisier (2012, §5) and Alon & Naor (2006). In Gram form, (16) says that the semidefinite relaxation $\max\{\langle a, Y \rangle : Y \succeq 0,\ Y_{ii} = 1\}$ of $\max_\epsilon \epsilon^\top a \epsilon$ has integrality gap at most $\pi/2$ on PSD matrices. Remark D.3 recalls the standard instance showing that $\pi/2$ is sharp.

A.2 CONVEX ANALYSIS

Theorem A.3 (conic strong duality)**.** Let $\mathcal{C}$ be a closed convex cone in a finite-dimensional Euclidean space V, $\mathcal{C}^* = \{y : \langle y, c \rangle \geq 0\ \forall c \in \mathcal{C}\}$ its dual cone, $\mathcal{P} \subseteq W$ a closed convex cone, $\mathcal{A} : V \to W$ linear, $b \in W$ and $e \in V$. Consider the primal problem $\min\{\langle e, v \rangle :\ v \in \mathcal{C},\ \mathcal{A}v - b \in \mathcal{P}\}$ and its dual $\max\{\langle b, y \rangle :\ y \in \mathcal{P}^*,\ e - \mathcal{A}^* y \in \mathcal{C}^*\}$. If the primal is bounded below and has a strictly feasible point (v_0 in the relative interior of $\mathcal{C}$ with $\mathcal{A}v_0 - b$ in the interior of $\mathcal{P}$), then the two optimal values coincide and the dual optimum is attained.

Source: Ben-Tal & Nemirovski (2001, Theorem 1.4.2 and §4.2).

Theorem A.4 (Sion's minimax theorem)**.** Let $\mathcal{C}$ be a compact convex subset and $\mathcal{P}$ a convex subset of linear topological spaces, and let $\Phi : \mathcal{C} \times \mathcal{P} \to \mathbb{R}$ be upper semicontinuous and quasi-concave in its first argument and lower semicontinuous and quasi-convex in its second. Then $\sup_{c \in \mathcal{C}} \inf_{p \in \mathcal{P}} \Phi(c, p) = \inf_{p \in \mathcal{P}} \sup_{c \in \mathcal{C}} \Phi(c, p)$.

Source: Sion (1958).

Theorem A.5 (Pietsch factorization for ℓ_∞^d)**.** For a linear operator $B : \ell_\infty^d \to E$ into a normed space, the 2-summing norm $\pi_2(B)^2 = \sup\{\sum_k \|Bx_k\|^2 :\ \max_i \sum_k x_{k,i}^2 \leq 1\}$, over finite families (x_k) in $\mathbb{R}^d$, equals $\min\{\sum_i h_i :\ h \geq 0,\ \|Bx\|^2 \leq \sum_i h_i x_i^2\ \forall x\}$.

Sources: Pietsch (1967); Diestel et al. (1995, Chapter 2). The constraint $\max_i \sum_k x_{k,i}^2 \le 1$ is the weak ℓ_2 norm of (x_k), since the extreme points of the unit ball of $(\ell_\infty^d)^* = \ell_1^d$ are $\pm e_i$. For $E = \ell_2$ and $B^\top B = A$ the right-hand side is $\nu_{\mathcal{D}_d^+}(A)$, and for $E = \ell_\infty(\mathbb{R}^d; \ell_2)$ with $(Bv)(x) = B_x v$ it is $\Lambda_{\mathcal{D}_d^+}(f)$; this is Remark 2.2, which Theorem 2.1 also proves directly.

Lemma A.6 (Popoviciu's inequality)**.** If a random variable takes values in an interval $[m, M]$, its variance is at most $(M-m)^2/4$.

Lemma A.7 (Schur complement)**.** For $H \succ 0$ and a vector p, $pp^\top \preceq H$ if and only if $p^\top H^{-1} p \le 1$.

A.3 Oracle complexity of ℓ_∞-smooth convex minimization

We use the black-box model of Guzmán & Nemirovski (2015, §1.3). Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$ with dual norm $\|\cdot\|_*$, $X \subset \mathbb{R}^n$ a convex compact set, and $\mathcal{F}_{\|\cdot\|}(2, L)$ the class of convex C^1 functions $f : \mathbb{R}^n \to \mathbb{R}$ with $\|\nabla f(x) - \nabla f(y)\|_* \le L\|x - y\|$ for all $x, y \in \mathbb{R}^n$. A *local oracle* returns, at a query point x, information that depends only on the restriction of f to a neighbourhood of x; the first-order oracle $(f(x), \nabla f(x))$ is one. A deterministic T-step method queries the oracle at $x_1, \dots, x_T$, each x_t being a deterministic function of the previous answers, and outputs $x_T \in X$. The minimax risk is $\mathrm{Risk}_{\mathcal{F},X}(T) = \inf_{\text{methods}} \sup_{f \in \mathcal{F}} [f(x_T) - \min_X f]$ for a specified oracle. In the main text this is the first-order oracle. The lower bounds below remain valid for the maximal local oracle, which returns the entire germ, and therefore also for the first-order oracle.

Theorem A.8 (Guzmán & Nemirovski, 2015, Corollary 1 with $p = \infty$, $\kappa = 2$, and the scaling remark after their Proposition 2)**.** There is an absolute constant $c_0 > 0$ such that for all $n \ge 3$, $3 \le T \le n$, $L, R > 0$ and every convex compact $X \supseteq \{y \in \mathbb{R}^n : \|y\|_\infty \le R\}$,

$$\mathrm{Risk}_{\mathcal{F}_{\|\cdot\|_\infty}(2,L),X}(T) \;\ge\; c_0 \, \frac{LR^2}{T \ln n}. \tag{17}$$

For $p = \infty$ and $\kappa = 2$ their exponent of T is $\kappa + (\kappa - 1)/p - 1 = 1$ and their smoothing constant is $M = O(\ln n)$; the scaling $f \mapsto f(\cdot/R)$ converts the unit-ball statement into (17). The hard instances are inf-convolution smoothings of functions $\max_{s \le T}[\epsilon_s y_{i_s} - (s-1)\delta]$ whose signs ϵ_s and coordinates i_s are chosen adaptively against the method (a *resisting oracle*).

Theorem A.9 (reducibility under linear surjections; Guzmán & Nemirovski, 2015, §4.2 and Appendix A.3)**.** Let $G : \mathbb{R}^N \to \mathbb{R}^n$ be a linear surjection, $X \subset \mathbb{R}^N$ convex compact, $Y = GX$, and $\mathcal{F}$ a class of convex functions on $\mathbb{R}^n$. Then the minimax risk of the lifted class $\{h \circ G : h \in \mathcal{F}\}$ over X is at least the minimax risk of $\mathcal{F}$ over Y, for deterministic T-step methods using the maximal local oracle in both spaces. A lower bound for this maximal oracle also applies to every less informative local oracle.

The point of the statement is that a local oracle for the lifted function $h \circ G$ could a priori be more informative than one for h; the proof passes to the maximal local oracle (the germ of the function at the query point); because G is surjective, the map (germ of h at Gx) $\mapsto$ (germ of $h \circ G$ at x) is injective, so the two maximal oracles carry the same information, and the two trajectories are linked by $y_t = Gx_t$.

Theorem A.10 (conditional gradient; Jaggi, 2013, Theorem 1)**.** Let X be convex compact and f convex and differentiable on X with curvature constant $C_f := \sup\{\frac{2}{\gamma^2}[f(y) - f(x) - \langle \nabla f(x), y - x\rangle] : x, s \in X,\ \gamma \in (0,1],\ y = x + \gamma(s - x)\}$. The conditional-gradient iteration $x_{t+1} = x_t + \gamma_t(s_t - x_t)$, $s_t \in \arg\min_{s \in X} \langle \nabla f(x_t), s\rangle$, $\gamma_t = 2/(t+2)$, satisfies $f(x_T) - \min_X f \le 2C_f/(T+2)$ and keeps $x_T \in X$. If $\|\nabla f(x) - \nabla f(y)\|_* \le L\|x - y\|$ on X, then $C_f \le L \operatorname{diam}_{\|\cdot\|}(X)^2$.

A.4 Results of Xie et al. (2026) that we use

Throughout, $\mathcal{H} = \mathbb{S}_d^+ \cap \mathcal{K}$ is a well-structured preconditioner set and $\|\cdot\|_{\mathcal{H}}$, $\|\cdot\|_{\mathcal{H},*}$ are as in (2).

Theorem A.11 (Xie et al., 2026, Lemma 2.2 and Proposition 2.5)**.** (a) $\|g\|_{\mathcal{H},*} = \inf\{\|g\|_{H^{-1}} : H \in \mathcal{H},\ H \succ 0,\ \operatorname{Tr} H \le 1\}$. (b) For every differentiable f, $L_{\|\cdot\|_{\mathcal{H}}}(f) \le \Lambda_{\mathcal{H}}(f)$, where $\Lambda_{\mathcal{H}}(f) =$

$\min\{L_{\|\cdot\|_H}(f) : H \in \mathcal{H}, \operatorname{Tr} H \le 1\}$ is the adaptive smoothness in Lipschitz form and $L_{\|\cdot\|}(f) = \sup_{x\neq y} \|\nabla f(x) - \nabla f(y)\|_*/\|x-y\|$.

(Their Proposition 2.5 also states $\Lambda_{\mathcal{H}}(f) \le d\, L_{\|\cdot\|_{\mathcal{H}}}(f)$; we give the two-line proof of this half for the diagonal algebra, the only case we need for C^1 functions, in Lemma E.2(a).) We write $L_{\|\cdot\|_{\mathcal{H}}}(f) = L_{\mathcal{H}}(f)$ and, for the diagonal algebra, $L_{\|\cdot\|_\infty}(f) = L_\infty(f)$.

Theorem A.12 (Xie et al., 2026, Theorem 4.3, deterministic diagonal case)**.** Let $\mathcal{H} = \mathcal{D}_d^+$, let f be convex with a global minimizer x^* and adaptive smoothness $0 < \Lambda_{\mathcal{H}}(f) < \infty$, and run their Algorithm 2 (Nesterov-accelerated AdaGrad with the cumulative preconditioner $V_t = \pi_{\mathcal{K}}\big(\epsilon I + \sum_{s\le t} g_s g_s^\top\big)^{1/2}$, their $\arg\min_{H\in\mathcal{H}}\langle M_t + \epsilon I, H^{-1}\rangle + \operatorname{Tr} H$: projection first, then square root; $\alpha_t = 2/(t+2)$) with learning rate $\eta = D$ and regularization $\epsilon > 0$, on exact gradients, under the assumption $\max_{t<T}\|x_t - x^*\|_\infty \le D$. Then

$$f(\bar{x}_T) - f(x^*) \;\le\; \frac{C}{(T+1)^2}\Big(\Lambda_{\mathcal{H}}(f) D^2 \|S_T\|_{\mathrm{op}} + 2dD\sqrt{\epsilon}\Big), \qquad \|S_T\|_{\mathrm{op}} \le \log\frac{\|V_{T-1}\|_{\mathrm{op}}^2}{\epsilon}, \tag{18}$$

for an absolute constant C, where the bound on $\|S_T\|_{\mathrm{op}}$ is the commutative case of their Lemma 3.3 (whose $(1-\beta)T$ term vanishes for the cumulative variant, $\beta = 1$), the second term of their bound, $\big(\frac{D^2}{2\eta} - \frac{\eta}{2}\big)\mathbb{E}\big[\frac{4}{(T+1)^2}\sum_t g_t^\top V_t^{-1} g_t\big]$, vanishes identically at $\eta = D$, and $\|V_{T-1}\|_{\mathrm{op}}^2 \le \epsilon + \max_i \sum_{t<T} g_{t,i}^2$. Their Theorem E.5 gives the same bound for Algorithm 8 under $\|x^*\|_\infty \le D$ instead of the iterate bound, still with x^* a global minimizer. Its average uses the pre-projection point, so this is not a box-feasible output guarantee. The constrained result used in Section 5 instead comes from Ene et al. (2021).

The term $2dD\sqrt{\epsilon}$ comes from $\operatorname{Tr}(V_{T-1}^{-1}) \le d/\sqrt{\epsilon}$ in the first term of their bound.

Theorem A.13 (Xie et al., 2026, Theorem 3.2; and Section 2.1 there, from Xie et al., 2025a)**.** For $\mathcal{H} = \mathcal{D}_d^+$ and exact gradients, AdaGrad (their Algorithm 1, cumulative variant) with a suitable learning rate satisfies $\frac{1}{T}\sum_{t<T}\|\nabla f(x_t)\|_1 \le \tilde{O}\big(\sqrt{\Delta_0 \Lambda_{\mathcal{H}}(f)/T}\big)$, where $\Delta_0 = f(x_0) - \inf f$ and the $\tilde{O}$ hides a logarithm of $1/\epsilon$ and of the problem constants; sign gradient descent with a suitable step satisfies $\min_{t<T}\|\nabla f(x_t)\|_1 \le O\big(\sqrt{\Delta_0 L_{\|\cdot\|_\infty}(f)/T}\big)$.

By (10), the ratio of the two coefficients, $\sqrt{\Delta_0\Lambda_{\mathcal{H}}(f)/T}$ and $\sqrt{\Delta_0 L_{\mathcal{H}}(f)/T}$, is at most $\sqrt{K_G \varrho_{\mathcal{H}}(f)}$ in diagonal geometry: in the deterministic nonconvex rates, too, adaptivity pays the rotation times a Grothendieck factor.

Theorem A.14 (Xie et al., 2026, Definition 4.1, Proposition B.11 and Theorem 4.5)**.** For stochastic gradients $\nabla f_t(x) = \nabla f(x) + \xi_t(x)$ with $\mathbb{E}\xi_t(x) = 0$, define $\sigma_{\mathrm{std}}^2 = \sup_{t,x}\mathbb{E}\|\xi_t(x)\|_{\mathcal{H},*}^2$ and $\sigma_{\mathcal{H}}^2 = \inf_{H\in\mathcal{H},\, H\succ 0,\, \operatorname{Tr} H\le 1}\sup_{t,x}\mathbb{E}\|\xi_t(x)\|_{H^{-1}}^2$. Then $\sigma_{\mathrm{std}}^2 \le \sigma_{\mathcal{H}}^2 \le d\,\sigma_{\mathrm{std}}^2$. For the algorithmic conclusion assume fresh, independent oracle samples, $L_{\|\cdot\|_{\mathcal{H}}}(f) < \infty$, $\Delta_0 = f(x_0) - \inf f < \infty$, and $\sigma_{\mathcal{H}} < \infty$. For $T \ge 1$, with $\eta > 0$, $\alpha \in (0,1)$ and the initialization $m_0 = \nabla f_0(x_0)$ of their Theorem 4.5, normalized steepest descent with momentum (their Algorithm 3) satisfies $\mathbb{E}\frac{1}{T}\sum_{t<T}\|\nabla f(x_t)\|_{\mathcal{H},*} \le \frac{\Delta_0}{\eta T} + \frac{2\eta}{\alpha}L_{\|\cdot\|_{\mathcal{H}}}(f) + \frac{2\sigma_{\mathcal{H}}}{\alpha T} + 2\sigma_{\mathcal{H}}\sqrt{\alpha}$, which is increasing in $\sigma_{\mathcal{H}}$. Fresh sampling ensures the conditional mean-zero property used to cancel cross terms in their Appendix F.1; pointwise unconditional unbiasedness alone is insufficient for this step.

A.5 Two Gaussian facts

Lemma A.15. If $g \sim N(0,1)$ then $\mathbb{E}|g| = \sqrt{2/\pi}$. If $g_1, \dots, g_d$ are i.i.d. $N(0,1)$ then $\mathbb{E}\big(\sum_i |g_i|\big)^2 = d + d(d-1)\frac{2}{\pi}$.

Proof. The first is a direct integration. For the second, expand the square: the d diagonal terms contribute $\mathbb{E}g_i^2 = 1$ each and the $d(d-1)$ off-diagonal terms contribute $(\mathbb{E}|g|)^2 = 2/\pi$ each. □

B STRUCTURE OF WELL-STRUCTURED PRECONDITIONER SETS

Standing conventions. $\mathcal{K} \subseteq \mathbb{M}_d(\mathbb{R})$ is a unital subalgebra closed under transposition, $\mathcal{H} = \mathbb{S}_d^+ \cap \mathcal{K}$, $\mathcal{K}_{\mathrm{s}} = \mathcal{K} \cap \mathbb{S}_d$, and $\pi_{\mathcal{K}} : \mathbb{M}_d \to \mathcal{K}$ is the orthogonal projection for the inner product $\langle A, B\rangle = \mathrm{Tr}(A^\top B)$. The assumption that $\mathcal{K}$ is closed under transposition is harmless: if $\mathcal{K}'$ is a unital subalgebra that is not, then $\mathbb{S}_d^+ \cap \mathcal{K}' = \mathbb{S}_d^+ \cap (\mathcal{K}' \cap \mathcal{K}'^\top)$ and $\mathcal{K}' \cap \mathcal{K}'^\top$ is a unital subalgebra closed under transposition, so it defines the same $\mathcal{H}$.

Lemma B.1 (structure)**.** (a) *(functional calculus)* If $X \in \mathcal{K}_{\mathrm{s}}$ has Jordan decomposition $X = X_+ - X_-$, then $X_+, X_-, |X| \in \mathcal{H}$; consequently $\operatorname{span}\mathcal{H} = \mathcal{K}_{\mathrm{s}}$. If $X \in \mathcal{K}_{\mathrm{s}}$ is invertible then $X^{-1} \in \mathcal{K}$; if $X \in \mathcal{H}$ then $X^{1/2} \in \mathcal{H}$ and the Moore–Penrose pseudo-inverse X^+ lies in $\mathcal{H}$.
(b) *(self-duality)* If $X \in \mathcal{K}_{\mathrm{s}}$ and $\langle X, H\rangle \geq 0$ for all $H \in \mathcal{H}$, then $X \in \mathcal{H}$.
(c) *(conditional expectation)* $\pi_{\mathcal{K}}(\mathbb{S}_d) \subseteq \mathcal{K}_{\mathrm{s}}$; $\pi_{\mathcal{K}}(I) = I$; $\mathrm{Tr}\,\pi_{\mathcal{K}}(M) = \mathrm{Tr}\,M$; $\pi_{\mathcal{K}}$ is self-adjoint; $\pi_{\mathcal{K}}(\mathbb{S}_d^+) \subseteq \mathcal{H}$; and $\pi_{\mathcal{K}}(AMB) = A\,\pi_{\mathcal{K}}(M)\,B$ for all $A, B \in \mathcal{K}$ and $M \in \mathbb{M}_d$.
(d) *(contraction)* $-\|M\|_{\mathrm{op}} I \preceq \pi_{\mathcal{K}}(M) \preceq \|M\|_{\mathrm{op}} I$ for every $M \in \mathbb{S}_d$.

Proof. (a) Because $\mathcal{K}$ is a finite-dimensional algebra containing I, it contains $p(X)$ for every polynomial p. Let $\sigma(X)$ be the (finite) spectrum of X and choose polynomials $p_+, p_-, p_{|\cdot|}$ that agree on $\sigma(X)$ with $\lambda \mapsto \max(\lambda, 0)$, $\max(-\lambda, 0)$ and $|\lambda|$. Then $X_+ = p_+(X)$, $X_- = p_-(X)$, $|X| = p_{|\cdot|}(X)$ belong to $\mathcal{K}$; being symmetric and PSD, they belong to $\mathcal{H}$. The same argument with $\lambda \mapsto \lambda^{-1}$ (when $0 \notin \sigma(X)$), $\lambda \mapsto \lambda^{1/2}$ (when $X \succeq 0$) and $\lambda \mapsto \lambda^{-1}\mathbf{1}_{\{\lambda \neq 0\}}$ gives $X^{-1} \in \mathcal{K}$, $X^{1/2} \in \mathcal{H}$, $X^+ \in \mathcal{H}$. Finally $X = X_+ - X_-$ shows $\mathcal{K}_{\mathrm{s}} \subseteq \operatorname{span}\mathcal{H}$, and the reverse inclusion is clear.

(b) Test the hypothesis with $H = X_- \in \mathcal{H}$. Since $X_+ X_- = 0$,

$$\langle X, X_-\rangle = \langle X_+, X_-\rangle - \langle X_-, X_-\rangle = -\|X_-\|_F^2.$$

Hence $\langle X, X_-\rangle \geq 0$ forces $X_- = 0$, i.e. $X = X_+ \in \mathcal{H}$.

(c)

Step 1: $\pi_{\mathcal{K}}$ preserves symmetry. Transposition is a Frobenius isometry mapping $\mathcal{K}$ onto itself, so it commutes with the orthogonal projection onto $\mathcal{K}$: $\pi_{\mathcal{K}}(M^\top) = \pi_{\mathcal{K}}(M)^\top$. For $M \in \mathbb{S}_d$ this gives $\pi_{\mathcal{K}}(M) \in \mathcal{K}_{\mathrm{s}}$.

Step 2: unitality and trace preservation. $I \in \mathcal{K}$ gives $\pi_{\mathcal{K}}(I) = I$; and since an orthogonal projection is self-adjoint, $\mathrm{Tr}\,\pi_{\mathcal{K}}(M) = \langle \pi_{\mathcal{K}}(M), I\rangle = \langle M, \pi_{\mathcal{K}}(I)\rangle = \mathrm{Tr}\,M$.

Step 3: positivity. Let $M \succeq 0$. For every $H \in \mathcal{H} \subseteq \mathcal{K}$, $\langle \pi_{\mathcal{K}}(M), H\rangle = \langle M, \pi_{\mathcal{K}}(H)\rangle = \langle M, H\rangle \geq 0$. Since $\pi_{\mathcal{K}}(M) \in \mathcal{K}_{\mathrm{s}}$ by Step 1, (b) yields $\pi_{\mathcal{K}}(M) \in \mathcal{H}$.

Step 4: bimodule property. Let $A, B \in \mathcal{K}$. Both $\pi_{\mathcal{K}}(AMB)$ and $A\pi_{\mathcal{K}}(M)B$ lie in $\mathcal{K}$, so it suffices to test them against an arbitrary $C \in \mathcal{K}$. Using $A^\top C B^\top \in \mathcal{K}$,

$$\langle \pi_{\mathcal{K}}(AMB), C\rangle = \langle AMB, C\rangle = \langle M, A^\top C B^\top\rangle = \langle \pi_{\mathcal{K}}(M), A^\top C B^\top\rangle = \langle A\,\pi_{\mathcal{K}}(M)\,B, C\rangle.$$

(d) Apply Step 3 and Step 2 to the two PSD matrices $\|M\|_{\mathrm{op}} I \pm M$. □

Lemma B.2 (the two norms)**.** With $\|\cdot\|_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{H},*}$ as in (2):
(a) $\|x\|_{\mathcal{H}}^2 = \|\pi_{\mathcal{K}}(xx^\top)\|_{\mathrm{op}}$ and $B_{\mathcal{H}} = \{x : \pi_{\mathcal{K}}(xx^\top) \preceq I\}$.
(b) $\|x\|_2/\sqrt{d} \leq \|x\|_{\mathcal{H}} \leq \|x\|_2$ and $\|g\|_2 \leq \|g\|_{\mathcal{H},*} \leq \sqrt{d}\,\|g\|_2$.
(c) $\|g\|_{\mathcal{H},*} = \mathrm{Tr}\left(\pi_{\mathcal{K}}(gg^\top)^{1/2}\right) = \inf\{\|g\|_{H^{-1}} : H \in \mathcal{H},\ H \succ 0,\ \mathrm{Tr}\,H \leq 1\}$.
(d) For every $\Sigma \in \mathbb{S}_d^+$,

$$\inf\left\{\langle \Sigma, H^{-1}\rangle :\ H \in \mathcal{H},\ H \succ 0,\ \mathrm{Tr}\,H \leq 1\right\} = \left(\mathrm{Tr}\ \pi_{\mathcal{K}}(\Sigma)^{1/2}\right)^2. \tag{19}$$

Proof. (a) By self-adjointness of $\pi_{\mathcal{K}}$ and $H \in \mathcal{K}$,

$$\|x\|_{\mathcal{H}}^2 = \sup_{H \in \mathcal{H},\ \mathrm{Tr}\,H \leq 1} \langle H, xx^\top\rangle = \sup_{H \in \mathcal{H},\ \mathrm{Tr}\,H \leq 1} \langle H, P\rangle, \qquad P := \pi_{\mathcal{K}}(xx^\top) \in \mathcal{H}.$$

Upper bound: $H, P \succeq 0$ give $\langle H, P\rangle \le \mathrm{Tr}(H)\|P\|_{\mathrm{op}} \le \|P\|_{\mathrm{op}}$.

Lower bound: if $P \neq 0$, let Π be the spectral projector of P onto the eigenspace of $\lambda_{\max}(P)$. It is a polynomial in P, hence $\Pi \in \mathcal{H}$, and $H := \Pi/\operatorname{Tr}\Pi$ has $\operatorname{Tr} H = 1$ and $\langle H, P\rangle = \lambda_{\max}(P) = \|P\|_{\mathrm{op}}$. The description of $B_{\mathcal{H}}$ follows.

(b) $H = I/d$ gives $\|x\|_{\mathcal{H}} \ge \|x\|_2/\sqrt{d}$; Lemma B.1(d) gives $\pi_{\mathcal{K}}(xx^\top) \preceq \|x\|_2^2 I$, hence $\|x\|_{\mathcal{H}} \le \|x\|_2$. The dual inequalities follow by duality of norms.

(c) and (d). Let $P := \pi_{\mathcal{K}}(\Sigma) \in \mathcal{H}$ (for (c), $\Sigma = gg^\top$). We prove (19) first.

Step 1: reduction to P. For $H \in \mathcal{H}$ with $H \succ 0$, $H^{-1} \in \mathcal{K}$ by Lemma B.1(a), so $\langle \Sigma, H^{-1}\rangle = \langle P, H^{-1}\rangle$.

Step 2: lower bound. By Cauchy–Schwarz for the Frobenius inner product,

$$\mathrm{Tr}(P^{1/2}) = \left\langle P^{1/2}H^{-1/2},\, H^{1/2}\right\rangle \le \|P^{1/2}H^{-1/2}\|_F\, \|H^{1/2}\|_F = \langle P, H^{-1}\rangle^{1/2}\,(\operatorname{Tr} H)^{1/2},$$

so $\langle P, H^{-1}\rangle \ge (\operatorname{Tr} P^{1/2})^2$ whenever $\operatorname{Tr} H \le 1$.

Step 3: upper bound. For $\delta > 0$ put $H_\delta := (P+\delta I)^{1/2}/\operatorname{Tr}(P+\delta I)^{1/2}$, an element of $\mathcal{H}$ with $H_\delta \succ 0$ and $\operatorname{Tr} H_\delta = 1$. Then

$$\langle P, H_\delta^{-1}\rangle = \operatorname{Tr}\left((P+\delta I)^{1/2}\right)\left\langle P, (P+\delta I)^{-1/2}\right\rangle \le \left(\operatorname{Tr}(P+\delta I)^{1/2}\right)^2 \xrightarrow[\delta\downarrow 0]{} (\operatorname{Tr} P^{1/2})^2.$$

This proves (19) and the second equality in (c).

Step 4: first equality in (c), upper bound. For $x \in B_{\mathcal{H}}$ and $H \in \mathcal{H}$, $H \succ 0$,

$$\langle g, x\rangle \le \|g\|_{H^{-1}}\|x\|_H, \qquad \|x\|_H^2 = \operatorname{Tr}(H)\, x^\top \tfrac{H}{\operatorname{Tr} H} x \le \operatorname{Tr}(H)\|x\|_{\mathcal{H}}^2 \le \operatorname{Tr} H,$$

so $\|g\|_{\mathcal{H},*} \le \inf_{\operatorname{Tr} H \le 1}\|g\|_{H^{-1}} = \operatorname{Tr} P^{1/2}$ by Steps 2–3.

Step 5: first equality in (c), lower bound. Let $Q := (P^{1/2})^+ \in \mathcal{H}$ (Lemma B.1(a)) and $x := Qg$. Since $P^{1/2}Q$ is the orthogonal projector onto $\mathrm{range}(P)$,

$$\langle g, x\rangle = \langle gg^\top, Q\rangle = \langle P, Q\rangle = \operatorname{Tr}\left(P^{1/2}\cdot P^{1/2}Q\right) = \mathrm{Tr}(P^{1/2}).$$

By the bimodule property, $\pi_{\mathcal{K}}(xx^\top) = Q\,\pi_{\mathcal{K}}(gg^\top)\,Q = QPQ$, which is the same projector; hence $\|x\|_{\mathcal{H}}^2 = \|QPQ\|_{\mathrm{op}} \le 1$ by (a), i.e. $x \in B_{\mathcal{H}}$. Therefore $\|g\|_{\mathcal{H},*} \ge \langle g, x\rangle = \operatorname{Tr} P^{1/2}$. □

Remark B.3. Lemma B.2(c) is Lemma B.8 of Xie et al. (2026) in different notation: their $P_{\mathcal{H}}(M)$ equals $\pi_{\mathcal{K}}(M)^{1/2}$ for $M \succeq 0$ by their Lemmas B.2 and B.4, and the second equality is their Lemma 2.2 (Theorem A.11(a)). The proof above keeps the paper self-contained.

Lemma B.4 (commutative algebras)**.** If $\mathcal{K}$ is commutative, then $\mathcal{K}_{\mathrm{s}}$ is itself a commutative unital algebra of symmetric matrices,

$$\mathcal{K}_{\mathrm{s}} = U\left\{\,\mathrm{diag}(c_1 I_{n_1}, \ldots, c_m I_{n_m}) :\ c \in \mathbb{R}^m\right\}U^\top \quad \text{for some orthogonal } U,$$

and $\mathcal{H} = \mathbb{S}_d^+ \cap \mathcal{K}_{\mathrm{s}}$. All objects of Section 2 depend on $\mathcal{K}$ only through $\mathcal{K}_{\mathrm{s}}$, because $\pi_{\mathcal{K}}|_{\mathbb{S}_d}$ is the orthogonal projection onto $\mathcal{K}_{\mathrm{s}}$. In particular a commutative $\mathcal{H}$ is split with all $b_i = 1$.

Proof. Products of commuting symmetric matrices are symmetric, so $\mathcal{K}_{\mathrm{s}}$ is closed under products. A commuting family of symmetric matrices is simultaneously orthogonally diagonalizable, and a unital algebra of diagonal matrices consists of the functions that are constant on the blocks of a partition of the coordinates, which is the displayed form. Finally, $\pi_{\mathcal{K}}(M) \in \mathcal{K}_{\mathrm{s}}$ for $M \in \mathbb{S}_d$ (Lemma B.1(c)) and $\pi_{\mathcal{K}}(M)$ is the closest point of $\mathcal{K} \supseteq \mathcal{K}_{\mathrm{s}}$ to M, hence also the closest point of $\mathcal{K}_{\mathrm{s}}$. □

B.1 Real, complex, and quaternionic matrix components

Lemma B.5 (real block representations)**.** Let $\mathcal{K} \subseteq \mathbb{M}_d(\mathbb{R})$ be unital and closed under transposition. There is an orthogonal decomposition

$$V := \mathbb{R}^d = \bigoplus_{i=1}^{m} V_i, \qquad V_i \cong (\mathbb{F}_i^{b_i})^{\oplus n_i}, \qquad \mathbb{F}_i \in \{\mathbb{R}, \mathbb{C}, \mathbb{H}\},$$

in which $\mathcal{K}$ consists of independent matrices $L_i \in \mathbb{M}_{b_i}(\mathbb{F}_i)$, each acting by left multiplication on all n_i copies. Write

$$q_i := \dim_{\mathbb{R}} \mathbb{F}_i, \qquad r_i := q_i n_i, \qquad d_i := b_i r_i,$$

and let P_i be the orthogonal projector onto V_i. Then:

(a) $P_i \in \mathcal{K}$. For every $H \in \mathcal{H}$, its restriction H_i to V_i satisfies $\|H_i\|_{\mathrm{op}} \le \mathrm{Tr}(H_i)/r_i$. All traces here are real matrix traces.

(b) $\{x = \bigoplus_i x_i : \|x_i\|_2 \le \sqrt{r_i}\ \forall i\} \subseteq B_{\mathcal{H}}$.

(c) If W is one irreducible copy $\mathbb{F}_i^{b_i}$ inside V_i and E_W its real orthogonal projector, then

$$\pi_{\mathcal{K}}(E_W) = P_i/n_i, \qquad E_W B_{\mathcal{H}} = \sqrt{r_i}\, B_2(W).$$

In particular, $\|g\|_{\mathcal{H},*} = \sqrt{r_i}\|g\|_2$ for $g \in W$.

(d) The rank $\mathrm{rk}\,\mathcal{K} = \sum_i b_i$ is the largest number of mutually orthogonal nonzero projections in $\mathcal{K}$.

Proof. We include the finite-dimensional representation argument to fix the real trace normalization. The only classification input is real Schur's lemma: the commutant of an irreducible real C^*-representation is $\mathbb{R}$, $\mathbb{C}$, or $\mathbb{H}$ (Rosenberg, 2015, Theorem 1.3, p. 2).

Step 1: a finite-dimensional double-commutant argument. For a unital transpose-closed algebra $\mathcal{A}$ on a real Euclidean space E, denote its commutant by $\mathcal{A}'$. We show $\mathcal{A}'' = \mathcal{A}$. Choose an orthonormal basis $e_1, \dots, e_s$ of E, and in E^s consider

$$\mathcal{U} = \{(Ae_1, \dots, Ae_s) : A \in \mathcal{A}\}.$$

This subspace is invariant under the diagonal action of $\mathcal{A}$ and its transpose. Its orthogonal projector Q therefore commutes with that action, so every block of Q belongs to $\mathcal{A}'$. If $T \in \mathcal{A}''$, the diagonal action of T commutes with Q. Since $(e_1, \dots, e_s) \in \mathcal{U}$, it follows that $(Te_1, \dots, Te_s) \in \mathcal{U}$, which means $T = A$ for some $A \in \mathcal{A}$.

Step 2: irreducible components. Every invariant subspace has an invariant orthogonal complement, so induction decomposes V orthogonally into irreducible subspaces. On an irreducible subspace W, real Schur's lemma identifies the commutant with a division $*$-algebra $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}, \mathbb{H}\}$. Regard its action as right scalar multiplication, using the opposite algebra if necessary; each of these division algebras is isomorphic to its opposite. An orthonormal $\mathbb{F}$-basis identifies W isometrically with $\mathbb{F}^b$, with real inner product $\mathrm{Re}(u^* v)$. Such a basis can be constructed by successively taking unit vectors in the orthogonal complements of their $\mathbb{F}$-spans, because unit-norm scalars act orthogonally and the commutant is closed under adjoints. For the first claim, let T lie in the commutant $\mathcal{D}$, which is transpose closed because $\mathcal{A}$ is. Then $T^\top T$ is a symmetric element of the division algebra $\mathcal{D}$; the subalgebra it generates is a field, so its minimal polynomial is irreducible, and having only real roots it is linear. Hence $T^\top T = N(T) I$ with $N(T) \ge 0$. The form N is multiplicative, so by Hurwitz's theorem it is the standard squared norm of $\mathbb{R}$, $\mathbb{C}$ or $\mathbb{H}$; unit scalars therefore act orthogonally, and polarization gives $\langle w\alpha, w\beta\rangle = \mathrm{Re}(\bar{\alpha}\beta)\|w\|^2$. Step 1 now identifies the algebra on W with the full algebra of right-$\mathbb{F}$-linear maps, namely left multiplication by $\mathbb{M}_b(\mathbb{F})$.

Step 3: multiplicities and orthogonality. Group equivalent irreducible subspaces. Their identifications may be chosen orthogonal: if S is an invertible intertwiner, then $S^\top S$ commutes with the representation, and the polar factor $S(S^\top S)^{-1/2}$ is an orthogonal intertwiner. There are no nonzero intertwiners between inequivalent groups. Within a group the commutant consists of arbitrary matrices of scalar intertwiners between the copies. Taking its commutant, using Step 1, gives exactly one independent $L_i \in \mathbb{M}_{b_i}(\mathbb{F}_i)$ repeated on n_i copies. The projector onto an entire group commutes with this commutant, so $P_i \in \mathcal{K}$.

Step 4: trace multiplicity and the product ball. A self-adjoint L_i is a real symmetric operator commuting with right-$\mathbb{F}_i$ multiplication. Each of its real eigenspaces is therefore an $\mathbb{F}_i$-subspace, so every eigenvalue has real multiplicity divisible by q_i on one copy and by $r_i = q_i n_i$ on V_i. For

$H_i \succeq 0$, this proves (a). Consequently

$$x^\top H x \le \sum_i \frac{\|x_i\|_2^2}{r_i} \operatorname{Tr}(H_i) \le \Big(\max_i \frac{\|x_i\|_2^2}{r_i} \Big) \operatorname{Tr} H,$$

which proves (b) by the definition of $\|\cdot\|_{\mathcal{H}}$.

Step 5: projection onto one copy. For every $M \in \mathcal{K}$, its action is identical on the n_i copies, so $\langle E_W, M\rangle = \operatorname{Tr}(M|_W) = \operatorname{Tr}(M|_{V_i})/n_i = \langle P_i/n_i, M\rangle$. This proves $\pi_{\mathcal{K}}(E_W) = P_i/n_i$. To prove the ball identity, let $x \in B_{\mathcal{H}}$ and $y = E_W x \neq 0$. The $\mathbb{F}_i$-linear orthogonal projection onto the field line spanned by y, repeated on all n_i copies and zero on other components, is an algebra projection Q_y of real rank r_i. Thus Q_y/r_i is an admissible preconditioner, and

$$1 \ge x^\top (Q_y/r_i) x \ge \|y\|_2^2/r_i.$$

This gives one inclusion; the reverse follows from (b) by using vectors supported in W. Taking the support function gives the asserted dual norm on W.

Step 6: the rank. A projection in $\mathcal{K}$ acts on component i as an $\mathbb{F}_i$-linear orthogonal projection Q_i of $\mathbb{F}_i^{b_i}$, repeated on the n_i copies. For mutually orthogonal nonzero projections $Q^{(1)}, \dots, Q^{(N)}$, the $\mathbb{F}_i$-ranks of the $Q_i^{(k)}$ add up to at most b_i for each i, while each $Q^{(k)}$ has positive $\mathbb{F}_i$-rank in some component. Hence $N \le \sum_i b_i$. The $\sum_i b_i$ projections onto the coordinate $\mathbb{F}_i$-lines of the components attain the bound. □

C Proofs for Sections 2 and 4

Lemma C.1. For every $A \in \mathbb{S}_d$,

$$\|A\|_{\mathrm{op}} \;\le\; \|A\|_{\mathcal{H}} \;\le\; \nu_{\mathcal{H}}(A) \;\le\; d\,\|A\|_{\mathrm{op}}. \tag{20}$$

If $A \succeq 0$ then $\|A\|_{\mathcal{H}} = \sup_{u \in B_{\mathcal{H}}} u^\top A u$. Moreover $\nu_{\mathcal{H}}$ is a norm on $\mathbb{S}_d$ and the minimum defining it is attained.

Proof. Left inequality. $B_2 \subseteq B_{\mathcal{H}}$ by Lemma B.2(b), so the supremum over $B_{\mathcal{H}} \times B_{\mathcal{H}}$ dominates the supremum over unit vectors.

Right inequality. $H = \|A\|_{\mathrm{op}} I \in \mathcal{H}$ satisfies $-H \preceq A \preceq H$ and has trace $d\|A\|_{\mathrm{op}}$.

Middle inequality. Let $H \in \mathcal{H}$ satisfy $-H \preceq A \preceq H$, and let $u, v \in B_{\mathcal{H}}$.

Step 1: a block-matrix argument. The congruence

$$\frac{1}{\sqrt{2}}\begin{pmatrix} I & I \\ I & -I \end{pmatrix} \begin{pmatrix} H & A \\ A & H \end{pmatrix} \frac{1}{\sqrt{2}}\begin{pmatrix} I & I \\ I & -I \end{pmatrix} = \begin{pmatrix} H+A & 0 \\ 0 & H-A \end{pmatrix} \succeq 0$$

shows $\left(\begin{smallmatrix} H & A \\ A & H \end{smallmatrix}\right) \succeq 0$. Compressing with $\left(\begin{smallmatrix} u & 0 \\ 0 & v \end{smallmatrix}\right) \in \mathbb{R}^{2d \times 2}$ gives $\left(\begin{smallmatrix} u^\top H u & u^\top A v \\ u^\top A v & v^\top H v \end{smallmatrix}\right) \succeq 0$, i.e.

$$(u^\top A v)^2 \le (u^\top H u)(v^\top H v). \tag{21}$$

Step 2: passing to the induced norm. If $H \neq 0$, $H/\operatorname{Tr} H$ is admissible in the definition of $\|\cdot\|_{\mathcal{H}}$, so $u^\top H u = \operatorname{Tr}(H)\, u^\top \frac{H}{\operatorname{Tr} H} u \le \operatorname{Tr}(H)\|u\|_{\mathcal{H}}^2 \le \operatorname{Tr} H$, and likewise for v. (If $H = 0$ then $A = 0$.) With (21), $u^\top A v \le \operatorname{Tr} H$. Taking the supremum over $u, v \in B_{\mathcal{H}}$ and the minimizing H gives $\|A\|_{\mathcal{H}} \le \nu_{\mathcal{H}}(A)$.

PSD case. For $A \succeq 0$, $u^\top A v \le \frac{1}{2}(u^\top A u + v^\top A v) \le \sup_{w \in B_{\mathcal{H}}} w^\top A w$.

Norm axioms and attainment. If $-H_i \preceq A_i \preceq H_i$ for $i = 1, 2$ then $-(H_1 + H_2) \preceq A_1 + A_2 \preceq H_1 + H_2$ with $H_1 + H_2 \in \mathcal{H}$, giving the triangle inequality; homogeneity and $\nu_{\mathcal{H}}(-A) = \nu_{\mathcal{H}}(A)$ are clear; $\nu_{\mathcal{H}}(A) = 0$ forces $H = 0$ and hence $A = 0$. The feasible set is closed and Tr is coercive on $\mathbb{S}_d^+$, so the minimum is attained. □

C.1 THE PROBING PRINCIPLE

Throughout this subsection $A(x) := \nabla^2 f(x)$. By Lemma C.1, $\|A(x)\|_{\mathrm{op}} \le L_{\mathcal{H}}(f)$ for every x. For a finite set $F \subset \mathbb{R}^d$ put

$$\Lambda_{\mathcal{H}}(F) := \min\{\operatorname{Tr} H : \ H \in \mathcal{H},\ -H \preceq A(x) \preceq H \ \ \forall x \in F\}.$$

Write $\operatorname{sym}(uv^\top) = \frac{1}{2}(uv^\top + vu^\top)$. All probes are finitely supported, so all conditional expectations below are elementary.

Lemma C.2 (moment representation). *(a)* For random vectors u, v, the matrices $Y := \frac{1}{4}\mathbb{E}[(u+v)(u+v)^\top]$ and $Z := \frac{1}{4}\mathbb{E}[(u-v)(u-v)^\top]$ are positive semidefinite, $Y - Z = \mathbb{E}[\operatorname{sym}(uv^\top)]$, and $Y + Z = \frac{1}{2}\mathbb{E}[uu^\top + vv^\top]$. *(b)* Conversely, for $Y, Z \in \mathbb{S}^+_d$ there is a finitely supported pair (u, v) with $\mathbb{E}[uv^\top] = Y - Z$ and $\mathbb{E}[uu^\top] = \mathbb{E}[vv^\top] = Y + Z$. If $Z = 0$ one can take $v = u$.

Proof. (a) is a direct expansion. For (b), write $Y = \sum_{k=1}^m y_k y_k^\top$ and $Z = \sum_{l=1}^n z_l z_l^\top$ by spectral decomposition (if $Y = Z = 0$ take $u = v = 0$). Let J be uniform on $\{1, \dots, m+n\}$ and put $(u, v) = \sqrt{m+n}\,(y_J, y_J)$ if $J \le m$ and $(u, v) = \sqrt{m+n}\,(z_{J-m}, -z_{J-m})$ otherwise. Then $\mathbb{E}[uv^\top] = \sum_k y_k y_k^\top - \sum_l z_l z_l^\top$ and $\mathbb{E}[uu^\top] = \mathbb{E}[vv^\top] = Y + Z$. □

Lemma C.3 (finite reduction and landscape duality). *(a)* $\Lambda_{\mathcal{H}}(f) = \sup_F \Lambda_{\mathcal{H}}(F)$, the supremum over finite sets $F \subset \mathbb{R}^d$. *(b)* For every finite F,

$$\Lambda_{\mathcal{H}}(F) = \max\Big\{ \sum_{x \in F} \langle A(x), Y_x - Z_x \rangle : \ Y_x, Z_x \succeq 0,\ \pi_{\mathcal{K}}\Big(\sum_{x \in F} (Y_x + Z_x) \Big) \preceq I \Big\},$$

and if every $A(x)$, $x \in F$, is positive semidefinite, a maximizer with all $Z_x = 0$ exists.

Proof. (a) Clearly $\Lambda_{\mathcal{H}}(F) \le \Lambda_{\mathcal{H}}(f)$. Let $s := \sup_F \Lambda_{\mathcal{H}}(F)$, which is finite because $H = L_{\mathcal{H}}(f) I$ is feasible for every F. The sets $C_F := \{H \in \mathcal{H} : \operatorname{Tr} H \le s,\ -H \preceq A(x) \preceq H \ \forall x \in F\}$ are nonempty and compact, and $C_{F \cup F'} \subseteq C_F \cap C_{F'}$. By the finite intersection property some H lies in every C_F, in particular in every $C_{\{x\}}$; it is feasible for $\Lambda_{\mathcal{H}}(f)$ with $\operatorname{Tr} H \le s$.

(b) This is Lemma D.1 with one pair of multipliers per constraint. The primal is $\min\{\operatorname{Tr} H : H \in \mathcal{H},\ H - A(x) \in \mathbb{S}^+_d,\ H + A(x) \in \mathbb{S}^+_d \ (x \in F)\}$. For $H \in \mathcal{H}$ its Lagrangian with multipliers $Y_x, Z_x \in \mathbb{S}^+_d$ is

$$\begin{aligned} &\operatorname{Tr} H - \sum_x \langle Y_x, H - A(x) \rangle - \sum_x \langle Z_x, H + A(x) \rangle \\ &\quad = \sum_x \langle A(x), Y_x - Z_x \rangle + \Big\langle H,\ I - \pi_{\mathcal{K}}\Big(\sum_x (Y_x + Z_x) \Big) \Big\rangle, \end{aligned}$$

using $\langle H, M \rangle = \langle H, \pi_{\mathcal{K}}(M) \rangle$ for $H \in \mathcal{K}$. Since $\mathcal{H}$ is self-dual in $\mathcal{K}_{\mathrm{s}}$ (Lemma B.1(b)), the infimum over $H \in \mathcal{H}$ is 0 if $\pi_{\mathcal{K}}(\sum_x (Y_x + Z_x)) \preceq I$ and $-\infty$ otherwise. The point $(2 \max_{x \in F} \|A(x)\|_{\mathrm{op}} + 1) I$ is strictly feasible, so Theorem A.3 gives equality and dual attainment. If all $A(x) \succeq 0$, replacing (Y_x, Z_x) by $(Y_x, 0)$ keeps feasibility, because $\pi_{\mathcal{K}}$ is positive, and does not decrease the objective. □

Proof of Theorem 2.1. Sure budget. $\pi_{\mathcal{K}}(uu^\top) \preceq I$ means $u \in B_{\mathcal{H}}$ (Lemma B.2(a)). Hence every sure reading is an average of values $u^\top A(x) v$ with $u, v \in B_{\mathcal{H}}$ and is at most $L_{\mathcal{H}}(f)$; deterministic probes come arbitrarily close.

Pointwise budget, upper bound. Let x be a value of X. Conditional on $X = x$, Lemma C.2(a) gives $Y_x, Z_x \succeq 0$ with $\mathbb{E}[u^\top A(x) v \mid X = x] = \langle A(x), Y_x - Z_x \rangle$ and $\pi_{\mathcal{K}}(Y_x + Z_x) = \frac{1}{2}\mathbb{E}[\pi_{\mathcal{K}}(uu^\top) + \pi_{\mathcal{K}}(vv^\top) \mid X = x] \preceq I$, because $\pi_{\mathcal{K}}$ is linear. Weak duality (Step 1 of the proof of Lemma D.1) bounds this by $\nu_{\mathcal{H}}(A(x)) \le \Lambda^\circ_{\mathcal{H}}(f)$. Averaging over x proves the claim.

Pointwise budget, attainment. For each x the dual maximum in Lemma D.1 is attained; Lemma C.2(b) realizes it by a probe with $X \equiv x$ and budget $\pi_{\mathcal{K}}(Y + Z) \preceq I$. Taking the supremum over x gives $\Lambda^\circ_{\mathcal{H}}(f)$.

Global budget, upper bound. Let $\mu_x = \Pr(X = x)$ and, by Lemma C.2(a) applied conditionally, $Y_x := \frac{\mu_x}{4}\mathbb{E}[(u+v)(u+v)^\top \mid X = x]$ and $Z_x := \frac{\mu_x}{4}\mathbb{E}[(u-v)(u-v)^\top \mid X = x]$. Then $\mathbb{E}[u^\top A(X)v] = \sum_x \langle A(x), Y_x - Z_x\rangle$ and $\sum_x \pi_{\mathcal{K}}(Y_x + Z_x) = \frac{1}{2}\mathbb{E}[\pi_{\mathcal{K}}(uu^\top) + \pi_{\mathcal{K}}(vv^\top)] \preceq I$. For every H feasible for $\Lambda_{\mathcal{H}}(f)$, using $\langle A(x), Y_x\rangle \le \langle H, Y_x\rangle$, $-\langle A(x), Z_x\rangle \le \langle H, Z_x\rangle$ and $H \in \mathcal{K}$,

$$\mathbb{E}[u^\top A(X)v] \le \Big\langle H, \sum_x (Y_x + Z_x)\Big\rangle = \Big\langle H, \pi_{\mathcal{K}}\Big(\sum_x (Y_x + Z_x)\Big)\Big\rangle \le \operatorname{Tr} H.$$

Minimizing over H gives the bound $\Lambda_{\mathcal{H}}(f)$.

Global budget, attainment. Fix a finite F and a maximizer $(Y_x, Z_x)_{x\in F}$ in Lemma C.3(b). Decompose every Y_x and Z_x spectrally into rank-one terms, and list all resulting triples (x, w, s), where w is a rank-one factor and $s = +1$ for factors of Y_x, $s = -1$ for factors of Z_x. If the list is empty then $\Lambda_{\mathcal{H}}(F) = 0$. Otherwise let N be its length, choose a triple uniformly at random and set $X = x$, $u = \sqrt{N}\, w$, $v = s\sqrt{N}\, w$. Then $\mathbb{E}[u^\top A(X)v] = \sum_x \langle A(x), Y_x - Z_x\rangle = \Lambda_{\mathcal{H}}(F)$ and $\mathbb{E}[uu^\top] = \mathbb{E}[vv^\top] = \sum_x (Y_x + Z_x)$, so the global budget holds. Lemma C.3(a) completes the proof.

Convex objectives. If every $A(x) \succeq 0$, the upper bounds above apply in particular to probes with $v = u$. For attainment, the pointwise dual optimum can be taken with $Z = 0$ (Lemma D.1) and the global one with all $Z_x = 0$ (Lemma C.3(b)); the constructions then have $v = u$. At the sure level, $\sup_{u,v\in B_{\mathcal{H}}} u^\top Av = \sup_{u\in B_{\mathcal{H}}} u^\top Au$ for $A \succeq 0$ (Lemma C.1). □

The decomposition (7). The sure budget implies the pointwise one, which implies the global one, so Theorem 2.1 gives $L_{\mathcal{H}} \le \Lambda^\circ_{\mathcal{H}} \le \Lambda_{\mathcal{H}}$. The identity $\Lambda_{\mathcal{H}} = \varrho_{\mathcal{H}}\Lambda^\circ_{\mathcal{H}}$ is the definition of $\varrho_{\mathcal{H}}$, and $\Lambda^\circ_{\mathcal{H}} \le K(\mathcal{H})L_{\mathcal{H}}$ (resp. $k(\mathcal{H})L_{\mathcal{H}}$ for convex f, whose Hessians are positive semidefinite) follows by applying the definition of $K(\mathcal{H})$ (resp. $k(\mathcal{H})$) at every point. The primal reading of the sure level for $A \succeq 0$ follows from $\min\{\operatorname{Tr} H : H \in \mathcal{H},\ u^\top Hu \ge u^\top Au\} = u^\top Au/\|u\|^2_{\mathcal{H}}$, by homogeneity and the definition of $\|\cdot\|_{\mathcal{H}}$.

Cellwise constants. For a set $C \subseteq \mathbb{R}^d$ write $\Lambda_{\mathcal{H}}(f|_C) := \min\{\operatorname{Tr} H : H \in \mathcal{H},\ -H \preceq \nabla^2 f(x) \preceq H\ \forall x \in C\}$. For a partition $\mathcal{P}$ of $\mathbb{R}^d$ into cells, the budget $\mathbb{E}[\pi_{\mathcal{K}}(uu^\top) \mid C(X)] \preceq I$ and the same for v, where $C(X)$ is the cell containing X, yields readings at most $\sup_C \Lambda_{\mathcal{H}}(f|_C)$ by the global upper bound applied conditionally on each cell, and the attainment argument applied inside one cell shows equality. This justifies the formula for $\Lambda_{\mathcal{P}}$ in Section 7.

What the preconditioner may know. By duality, the budgets are information patterns for a dominating preconditioner: in $\Lambda^\circ_{\mathcal{H}}$ it may depend on the point, and in $\Lambda_{\mathcal{H}}$ it may not. For convex f the sure level lets it depend on the direction as well, since $\|A\|_{\mathcal{H}} = \sup_u \min\{\operatorname{Tr} H : H \in \mathcal{H},\ u^\top Hu \ge u^\top Au\}$ for $A \succeq 0$. The methods follow the same pattern. The NSD step for $\|\cdot\|_{\mathcal{H}}$ is a preconditioned step with $H \propto \pi_{\mathcal{K}}(gg^\top)^{1/2}$ (Lemma B.2), which depends on the current gradient g and hence on the direction of the step itself; AdaGrad accumulates $\pi_{\mathcal{K}}(\sum_{s\le t} g_s g_s^\top)^{1/2}$ over the whole run (Xie et al., 2025b); Adam's exponential average lies in between. The constant that governs a method reflects what its preconditioner is allowed to know.

C.2 The rank bound and its sharpness

We use the notation of Lemma B.5: $V = \bigoplus_i V_i$, $P_i \in \mathcal{K}$ the projector onto V_i, $q_i = \dim_{\mathbb{R}} \mathbb{F}_i$, $r_i = q_i n_i$, and $d_i = b_i r_i = \operatorname{Tr} P_i$. Put $S := \sum_i \sqrt{r_i} P_i$.

Lemma C.4 (the rescaled Euclidean metric). *(a)* $SB_2 \subseteq B_{\mathcal{H}}$, hence $\|SAS\|_{\mathrm{op}} \le \|A\|_{\mathcal{H}}$ for all $A \in \mathbb{S}_d$ and $\|Sg\|_2 \le \|g\|_{\mathcal{H},*}$ for all g. *(b)* If $M \succeq 0$ and $\pi_{\mathcal{K}}(M) \preceq I$, then $\operatorname{Tr}(S^{-1}MS^{-1}) \le \operatorname{rk}\mathcal{K}$.

Proof. (a) For $y \in B_2$ and $x = Sy$, $\|x_i\|_2 = \sqrt{r_i}\|y_i\|_2 \le \sqrt{r_i}$, so $x \in B_{\mathcal{H}}$ by Lemma B.5(b). Then $\|SAS\|_{\mathrm{op}} = \sup_{y,y'\in B_2}(Sy)^\top A(Sy') \le \|A\|_{\mathcal{H}}$ and $\|Sg\|_2 = \sup_{y\in B_2}\langle g, Sy\rangle \le \|g\|_{\mathcal{H},*}$. (b) Since $P_i \in \mathcal{K}$ and $\pi_{\mathcal{K}}$ is self-adjoint, $\operatorname{Tr}(P_i M) = \langle P_i, \pi_{\mathcal{K}}(M)\rangle \le \operatorname{Tr} P_i$. Hence $\operatorname{Tr}(S^{-1}MS^{-1}) = \sum_i \operatorname{Tr}(P_i M)/r_i \le \sum_i d_i/r_i = \sum_i b_i$. □

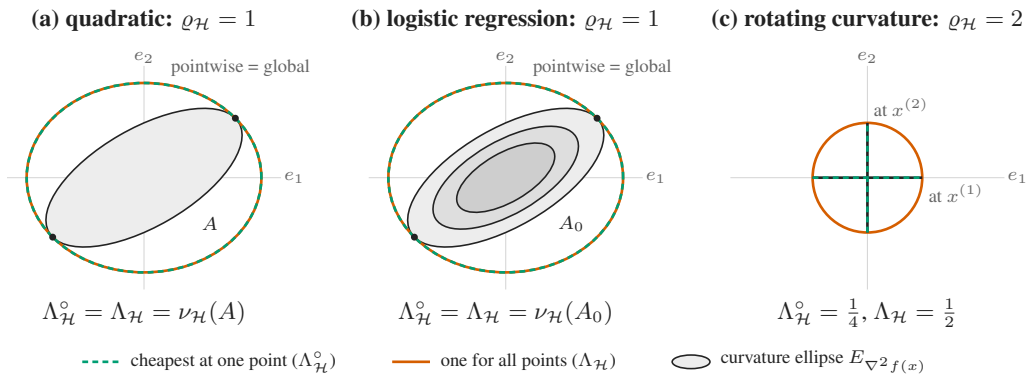

Figure 5: **Commitment in pictures** ($d = 2$, diagonal geometry). For $A \succeq 0$, $A \preceq H$ exactly when the ellipse $E_A = A^{1/2}B_2$ lies inside E_H, and a diagonal H is an axis-aligned ellipse whose trace is the sum of its squared semi-axes. $\Lambda^\circ_{\mathcal{H}}$ pays for the cheapest such ellipse at each point (dashed), $\Lambda_{\mathcal{H}}$ for one ellipse containing the curvature of every point (solid); the cheapest diagonal majorant of A is $\operatorname{diag}(A_{11}+|A_{12}|, A_{22}+|A_{12}|)$, touching E_A at the dots. **(a)** A quadratic has one Hessian, and **(b)** in logistic regression every Hessian lies inside the one at the origin (drawn: sA_0, $s = 1, 0.55, 0.25$), so $\varrho_{\mathcal{H}} = 1$ (Proposition 4.2). **(c)** Curvature that bends one coordinate at a time costs $\frac{1}{4}$ at each point but $\frac{1}{2}$ in total: $\varrho_{\mathcal{H}} = 2 = \operatorname{rk}\mathcal{K}$. The needles are the limiting Hessians $\frac{1}{4}e_k e_k^\top$ of the saturated construction of Theorem 4.1.

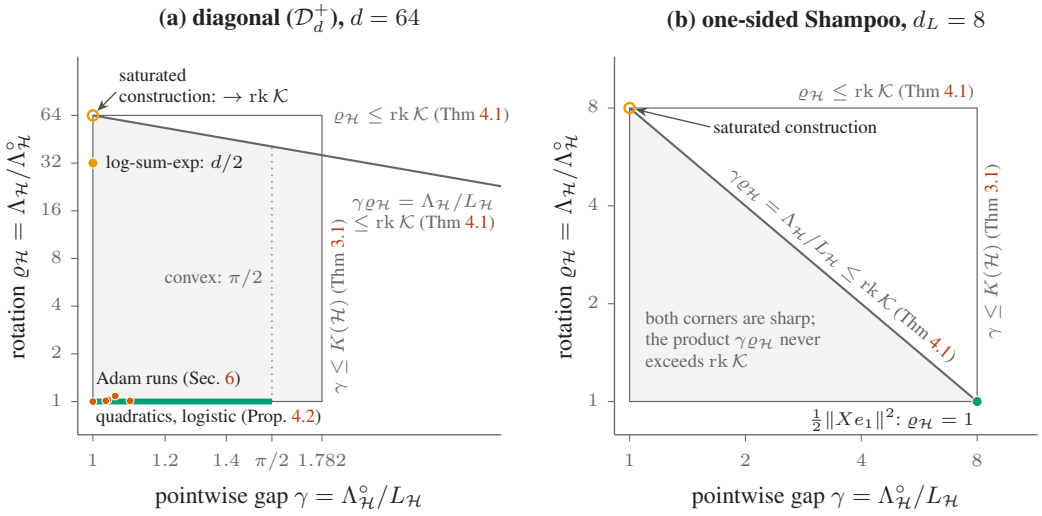

Figure 6: **The two relaxations share one budget.** With $\gamma := \Lambda^\circ_{\mathcal{H}}/L_{\mathcal{H}}$ the pointwise gap, Theorem 3.1 gives $\gamma \le K(\mathcal{H})$, and Theorem 4.1 bounds both the rotation $\varrho_{\mathcal{H}}$ and the whole factor $\gamma\varrho_{\mathcal{H}} = \Lambda_{\mathcal{H}}/L_{\mathcal{H}}$ by $\operatorname{rk}\mathcal{K}$. Every landscape therefore lies in the shaded region below the hyperbola $\gamma\varrho_{\mathcal{H}} = \operatorname{rk}\mathcal{K}$, a line on these log–log axes; the box shows the two bounds taken separately. **(a)** Diagonal geometry, $d = 64$: the pointwise gap is below $K(\mathcal{D}^+_d) < K_G$ (line at Krivine's bound 1.782), and below $\pi/2$ for convex f, so only rotation can grow with d. Convex quadratics and binary logistic regression have $\varrho_{\mathcal{H}} = 1$, log-sum-exp pays half the rank (Proposition 4.2), and the saturated construction approaches the full rank (open marker). Dots: the Adam runs of Section 6, with pointwise gap at most 1.10 and rotation at most 1.09. **(b)** One-sided Shampoo, $d_L = 8$: here $K(\mathcal{H}) = \operatorname{rk}\mathcal{K} = d_L$ and only the product bound binds; the quadratic $\frac{1}{2}\|Xe_1\|^2$ attains the corner $(d_L, 1)$, and the saturated construction approaches $(1, d_L)$.

Proof of the upper bound in Theorem 4.1. Let (X, u, v) be a globally budgeted probe. Lemma C.4(b) with $M = \mathbb{E}[uu^\top]$, whose projection is $\mathbb{E}[\pi_{\mathcal{K}}(uu^\top)] \preceq I$, gives $\mathbb{E}\|S^{-1}u\|_2^2 \le \operatorname{rk}\mathcal{K}$,

and likewise for v. By Lemma C.4(a),

$$\mathbb{E}[u^\top A(X)v] = \mathbb{E}\big[(S^{-1}u)^\top SA(X)S\,(S^{-1}v)\big] \le L_{\mathcal{H}}(f)\,\big(\mathbb{E}\|S^{-1}u\|_2^2\,\mathbb{E}\|S^{-1}v\|_2^2\big)^{1/2} \le \operatorname{rk}\mathcal{K}\cdot L_{\mathcal{H}}(f).$$

Theorem 2.1 gives $\Lambda_{\mathcal{H}}(f) \le \operatorname{rk}\mathcal{K}\cdot L_{\mathcal{H}}(f)$, and $\varrho_{\mathcal{H}}(f) = \Lambda_{\mathcal{H}}/\Lambda^\circ_{\mathcal{H}} \le \Lambda_{\mathcal{H}}/L_{\mathcal{H}} \le \operatorname{rk}\mathcal{K}$. Also $\varrho_{\mathcal{H}} \ge 1$ because $\Lambda^\circ_{\mathcal{H}} \le \Lambda_{\mathcal{H}}$. □

For the sharpness we fix, for each component i and $j \le b_i$, the real unit vector $e_{ij} \in V_i$ that corresponds to the jth standard basis vector of the first copy of $\mathbb{F}_i^{b_i}$. There are $\operatorname{rk}\mathcal{K}$ such vectors, and they are orthonormal.

Lemma C.5 (an extremal family)**.** Let $A_{ij} := r_i^{-1}e_{ij}e_{ij}^\top$. Then $\|A_{ij}\|_{\mathcal{H}} = \nu_{\mathcal{H}}(A_{ij}) = 1$. Moreover, if $\eta \ge 0$ and $H \in \mathcal{H}$ satisfies $H \succeq A_{ij} - \eta I$ for all (i,j), then $\operatorname{Tr} H \ge \operatorname{rk}\mathcal{K} - \eta d$.

Proof. By Lemma B.5(c), $\sup_{x\in B_{\mathcal{H}}}\langle e_{ij}, x\rangle = \sqrt{r_i}$, so $\|A_{ij}\|_{\mathcal{H}} = 1$ (PSD formula of Lemma C.1). Let $Q_{ij} \in \mathcal{K}$ be the projector onto the $\mathbb{F}_i$-line through the jth basis vector, acting on every copy of $\mathbb{F}_i^{b_i}$ in V_i. It has real rank r_i and $Q_{ij} \succeq e_{ij}e_{ij}^\top$, so $r_i^{-1}Q_{ij}$ is feasible for $\nu_{\mathcal{H}}(A_{ij})$ with trace 1; with $\|A_{ij}\|_{\mathcal{H}} \le \nu_{\mathcal{H}}(A_{ij})$ this gives $\nu_{\mathcal{H}}(A_{ij}) = 1$.

For the trace bound, recall from the proof of Lemma B.5 that $\mathcal{K}$ acts on each copy $W \cong \mathbb{F}_i^{b_i}$ by left matrix multiplication, commuting with right multiplication by scalars, and that right multiplication by a unit scalar is orthogonal for the real inner product $\operatorname{Re}(w^*w')$. Let $s_1, \dots, s_{q_i}$ be a real orthonormal basis of $\mathbb{F}_i$. The vectors $e_{ij}s_t$ form a real orthonormal basis of the first copy W_i, and

$$(e_{ij}s_t)^\top H(e_{ij}s_t) = e_{ij}^\top He_{ij} \;\ge\; r_i^{-1} - \eta.$$

Hence $\operatorname{Tr}(H|_{W_i}) \ge q_ib_i(r_i^{-1}-\eta)$. Since H acts identically on the n_i orthogonal copies, $\operatorname{Tr}(P_iH) \ge n_iq_ib_i(r_i^{-1}-\eta) = b_i - \eta d_i$, and summing over i gives $\operatorname{Tr} H \ge \operatorname{rk}\mathcal{K} - \eta d$. □

Proof of sharpness in Theorem 4.1: a convex construction. If $\operatorname{rk}\mathcal{K} = 1$ there is nothing to prove, so let $\operatorname{rk}\mathcal{K} \ge 2$. Enumerate the vectors of Lemma C.5 as $e_1, \dots, e_{\operatorname{rk}\mathcal{K}}$, let $r_{(k)}$ be the value of r_i for e_k, and write $y_k = \langle e_k, x\rangle$. Let $h(t) = \log(1+e^t)$, so that $h' = \sigma$ is the logistic function, $0 < h'' = \sigma(1-\sigma) \le h''(0) = \frac14$ and $h''(t) \le e^{-|t|}$; put $q(t) = h(t) - th'(t)$, so that $q'(t) = -th''(t)$ and $0 < q \le q(0) = \log 2$. For $\beta \in (0,1]$ and $M > 0$ define

$$\tau_k(x) = 1 + \beta\sum_{j\ne k} h(y_j + M), \qquad g_k(x) = \tau_k(x)\,h\big(y_k/\tau_k(x)\big), \qquad f = \sum_{k=1}^{\operatorname{rk}\mathcal{K}} \frac{g_k}{r_{(k)}}.$$

Step 1: convexity. The perspective $(s,\tau) \mapsto \tau h(s/\tau)$ is jointly convex on $\mathbb{R}\times(0,\infty)$, and it is nondecreasing in τ because its τ-derivative is $q(s/\tau) > 0$. Since y_k is linear and τ_k is convex with $\tau_k \ge 1$, each g_k is convex and C^∞, and so is f.

Step 2: the Hessian. With $t_k = y_k/\tau_k$, a direct computation gives

$$\nabla^2 g_k = \frac{h''(t_k)}{\tau_k}\,w_kw_k^\top + q(t_k)\,\nabla^2\tau_k, \qquad w_k = e_k - t_k\nabla\tau_k,$$

where $\nabla\tau_k = \beta\sum_{j\ne k}\sigma(y_j+M)e_j$ and $\nabla^2\tau_k = \beta\sum_{j\ne k}h''(y_j+M)e_je_j^\top$. By Lemma B.5(c), $\|e_j\|_{\mathcal{H},*} = \sqrt{r_{(j)}}$, and $\nu_{\mathcal{H}}(ww^\top) = \|w\|_{\mathcal{H},*}^2$ for every vector w (Lemma B.2(c) and the Schur complement). Hence, with $B := \beta\,\operatorname{rk}\mathcal{K}\sqrt{d}$,

$$\nu_{\mathcal{H}}(w_kw_k^\top) \le \Big(\sqrt{r_{(k)}} + |t_k|\beta\sum_{j\ne k}\sqrt{r_{(j)}}\Big)^2 \le r_{(k)}\big(1+|t_k|B\big)^2, \quad \nu_{\mathcal{H}}(\nabla^2\tau_k) \le \frac{\beta}{4}\sum_j r_{(j)} \le \frac{\beta d}{4}.$$

Since $\nu_{\mathcal{H}}$ is a norm (Lemma C.1), for every x

$$\nu_{\mathcal{H}}\big(\nabla^2 f(x)\big) \le \sum_k u_k(x) + \operatorname{rk}\mathcal{K}\,\frac{\beta d\log 2}{4}, \qquad u_k := \frac{h''(t_k)}{\tau_k}\big(1+|t_k|B\big)^2. \tag{22}$$

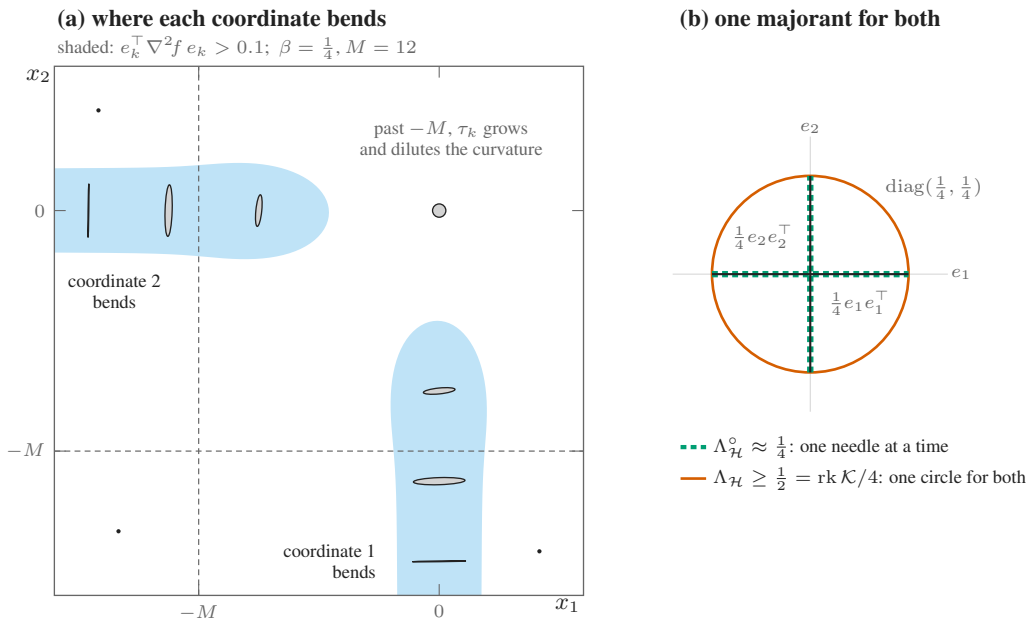

Figure 7: **The construction behind the sharpness of Theorem 4.1** ($d = 2$, diagonal geometry, $\mathrm{rk}\,\mathcal{K} = 2$; display parameters $\beta = \frac{1}{4}$, $M = 12$). **(a)** Shaded: where the curvature $e_k^\top \nabla^2 f\, e_k$ along a coordinate exceeds 0.1. Glyphs: $\nabla^2 f(x)$ at sample points, with semi-axes proportional to its eigenvalues; dots mark curvature below 0.02. Coordinate k bends, with curvature near $\frac{1}{4}$, only near $x_k = 0$ and only while the other coordinate lies below $-M$; past $-M$ its scale $\tau_k = 1 + \beta h(x_j + M)$ grows and dilutes the curvature. The two strips are disjoint, so at every point at most one needle is long. **(b)** Deep in the strips the Hessians tend to $\frac{1}{4} e_1 e_1^\top$ and $\frac{1}{4} e_2 e_2^\top$ (drawn as $A^{1/2} B_2$, as in Figure 5). Each needle is its own cheapest diagonal majorant, but a single diagonal majorant must contain both, so $\Lambda_{\mathcal{H}} \geq \frac{1}{2} = \mathrm{rk}\,\mathcal{K}/4$ while $\Lambda^\circ_{\mathcal{H}} \to \frac{1}{4}$ as $\beta \to 0$ and $M \to \infty$. For the drawn parameters $\Lambda^\circ_{\mathcal{H}} = L_{\mathcal{H}} = 0.30$ and $\Lambda_{\mathcal{H}} = 0.51$.

In particular the Hessian of f is bounded.

Step 3: at most one coordinate bends at a time. Let $c_1 = \sup_t h''(t)(1 + |t|B)^2$, which is finite because $h''(t) \leq e^{-|t|}$, and for $\epsilon \in (0, c_1)$ let $T_\epsilon = \sup\{|t| : h''(t)(1 + |t|B)^2 \geq \epsilon\} < \infty$. Suppose $u_k(x) > \epsilon$ and $u_l(x) > \epsilon$ for some $k \neq l$. Since $u_k \leq c_1/\tau_k$, both τ_k and τ_l are below c_1/ϵ. From $\tau_k \geq 1 + \beta h(y_l + M) \geq 1 + \beta(y_l + M)$ we get $y_l < -M + c_1/(\epsilon\beta)$. From $u_l > \epsilon$ and $\tau_l \geq 1$ we get $|t_l| \leq T_\epsilon$, hence $|y_l| = |t_l|\tau_l < T_\epsilon c_1/\epsilon$. If

$$M \geq \frac{c_1}{\epsilon\beta} + \frac{T_\epsilon c_1}{\epsilon},$$

these two bounds are incompatible, so at most one $u_k(x)$ exceeds ϵ, and (22) gives

$$\Lambda^\circ_{\mathcal{H}}(f) \leq c_1 + (\mathrm{rk}\,\mathcal{K} - 1)\epsilon + \mathrm{rk}\,\mathcal{K}\,\frac{\beta d \log 2}{4}.$$

Step 4: every coordinate bends somewhere. Fix k and let $x_n = -n \sum_{j \neq k} e_j$, so that $y_k(x_n) = 0$ and $y_j(x_n) = -n$ for $j \neq k$. As $n \to \infty$, $\tau_k(x_n) \to 1$, $\nabla\tau_k(x_n) \to 0$ and $\nabla^2\tau_k(x_n) \to 0$, so $\nabla^2 g_k(x_n) \to h''(0) e_k e_k^\top = \frac{1}{4} e_k e_k^\top$. For $j \neq k$, $\tau_j(x_n) \to 1 + \beta h(M)$ and $t_j = -n/\tau_j \to -\infty$; since $h''(t)(1 + |t|B)^2 \to 0$ and $q(t) \to 0$ as $t \to -\infty$, while $\nabla\tau_j$ and $\nabla^2\tau_j$ stay bounded, $\nabla^2 g_j(x_n) \to 0$. Hence $\nabla^2 f(x_n) \to \frac{1}{4} e_k e_k^\top / r_{(k)} = \frac{1}{4} A_k$ in the notation of Lemma C.5. Every H feasible for $\Lambda_{\mathcal{H}}(f)$ satisfies $H \succeq \nabla^2 f(x_n)$ for all n, hence $H \succeq \frac{1}{4} A_k$ for every k, and Lemma C.5 gives $\mathrm{Tr}\, H \geq \mathrm{rk}\,\mathcal{K}/4$. Thus $\Lambda_{\mathcal{H}}(f) \geq \mathrm{rk}\,\mathcal{K}/4$.

Step 5: conclusion. Since $h''(t)\,|t| \leq 1/e$ and $h''(t)\,t^2 \leq 4/e^2$, we have $c_1 \leq \frac{1}{4} + \frac{2B}{e} + \frac{4B^2}{e^2}$. Given $\delta \in (0, 1)$, choose β so small that $c_1 \leq \frac{1}{4} + \frac{\delta}{3}$ and $\mathrm{rk}\,\mathcal{K}\,\beta d \log 2/4 \leq \frac{\delta}{3}$, then $\epsilon = \delta/(3\,\mathrm{rk}\,\mathcal{K})$, which

is below $\frac{1}{6} < c_1$, and M as in Step 3. Then $L_{\mathcal{H}}(f) \le \Lambda^{\circ}_{\mathcal{H}}(f) \le \frac{1}{4} + \delta$ and $\Lambda_{\mathcal{H}}(f) \ge \operatorname{rk}\mathcal{K}/4$, so both $\varrho_{\mathcal{H}}(f)$ and $\Lambda_{\mathcal{H}}(f)/L_{\mathcal{H}}(f)$ are at least $\operatorname{rk}\mathcal{K}/(1+4\delta)$. Applying this with $\delta' = \min\{\frac{1}{2}, \delta/(4\operatorname{rk}\mathcal{K})\}$ in place of δ gives $\operatorname{rk}\mathcal{K}/(1+4\delta') \ge \operatorname{rk}\mathcal{K}(1-4\delta') \ge \operatorname{rk}\mathcal{K} - \delta$, as claimed. Moreover $L_{\mathcal{H}}(f) \ge \|\frac{1}{4}A_k\|_{\mathcal{H}} = \frac{1}{4}$, so the pointwise gap of f is at most $1+4\delta$: essentially the whole factor is rotation. □

C.3 Proof of Proposition 4.2

(i) Quadratics. The Hessian is the constant matrix A, so $\Lambda^{\circ}_{\mathcal{H}} = \Lambda_{\mathcal{H}} = \nu_{\mathcal{H}}(A)$ and $L_{\mathcal{H}} = \|A\|_{\mathcal{H}}$; the suprema over A are the definitions of $K(\mathcal{H})$ and $k(\mathcal{H})$.

(ii) A dominant Hessian. A minimum-trace majorant H of B satisfies $-H \preceq 0 \preceq \nabla^2 f(x) \preceq B \preceq H$ for every x, so $\Lambda_{\mathcal{H}} \le \nu_{\mathcal{H}}(B)$. Evaluation at x_0 gives $\Lambda^{\circ}_{\mathcal{H}} \ge \nu_{\mathcal{H}}(B)$, and $\Lambda^{\circ}_{\mathcal{H}} \le \Lambda_{\mathcal{H}}$ gives equality. For (9), the scalar logistic second derivative $\sigma(t)(1-\sigma(t))$ is at most $1/4$, attained at $t=0$, hence $\nabla^2 f(x) \preceq \frac{1}{4}\sum_j w_j a_j a_j^{\top} + \lambda I = \nabla^2 f(0)$.

(iii) Log-sum-exp. Assume $d \ge 2$ (for $d = 1$ the function is affine). Write

$$p = p(x) = \operatorname{softmax}(x/\tau), \qquad M_p := \operatorname{diag}(p) - pp^{\top}, \qquad \nabla^2 \operatorname{lse}_\tau(x) = \tfrac{1}{\tau} M_p.$$

As x ranges over $\mathbb{R}^d$, p ranges over the open simplex Δ°. For $u, v \in \mathbb{R}^d$ viewed as functions on $[d]$ with law p,

$$u^{\top} M_p v = \operatorname{Cov}_p(u, v), \qquad v^{\top} M_p v = \operatorname{Var}_p(v) \ge 0. \tag{23}$$

Step 1: the diagonal algebra, $L_{\mathcal{H}}(\operatorname{lse}_\tau) = 1/\tau$. For $u, v \in [-1,1]^d$, (23) and Cauchy–Schwarz give $u^{\top} M_p v \le \sqrt{\operatorname{Var}_p(u)\operatorname{Var}_p(v)} \le 1$, since a variable with values in $[-1,1]$ has variance at most 1. Equality is approached as $p \to \frac{1}{2}(e_1 + e_2)$ with $u = v = e_1 - e_2$.

Step 2: the diagonal algebra, $\Lambda^{\circ}_{\mathcal{H}}(\operatorname{lse}_\tau) = 1/\tau$. The diagonal matrix $\operatorname{diag}(p)$ majorizes M_p on both sides: $M_p \preceq \operatorname{diag}(p)$ is immediate, and

$$-\operatorname{diag}(p) \preceq M_p \iff pp^{\top} \preceq 2\operatorname{diag}(p) \iff p^{\top}(2\operatorname{diag}(p))^{-1}p = \tfrac{1}{2}\textstyle\sum_i p_i \le 1$$

by the Schur complement (Lemma A.7). Thus $\nu_{\mathcal{H}}(M_p) \le \operatorname{Tr}\operatorname{diag}(p) = 1$ for every p, so $\Lambda^{\circ}_{\mathcal{H}} \le 1/\tau$; Step 1 and $L_{\mathcal{H}} \le \Lambda^{\circ}_{\mathcal{H}}$ give equality.

Step 3: the diagonal algebra, $\Lambda_{\mathcal{H}}(\operatorname{lse}_\tau) = d/(2\tau)$. Upper bound. For a unit vector v, Popoviciu's inequality (Lemma A.6) and $\max_i v_i - \min_i v_i \le \sqrt{2}\|v\|_2$ give

$$\lambda_{\max}(M_p) = \sup_{\|v\|_2=1} \operatorname{Var}_p(v) \le \sup_{\|v\|_2=1} \tfrac{1}{4}(\max_i v_i - \min_i v_i)^2 \le \tfrac{1}{2}, \tag{24}$$

so $H = \frac{1}{2\tau} I$ satisfies $-H \preceq 0 \preceq \nabla^2 \operatorname{lse}_\tau(x) \preceq H$ for all x, and $\Lambda_{\mathcal{H}} \le d/(2\tau)$. *Lower bound.* Let $H = \operatorname{diag}(h)$ satisfy $H \succeq M_p/\tau$ for all $p \in \Delta^{\circ}$. Letting $p \to \frac{1}{2}(e_i + e_j)$ gives $M_p \to \frac{1}{4}(e_i - e_j)(e_i - e_j)^{\top}$, whence $h_i + h_j \ge 1/\tau$ for all $i \ne j$. Summing over the $\binom{d}{2}$ pairs, $(d-1)\operatorname{Tr} H \ge \binom{d}{2}/\tau$, i.e. $\operatorname{Tr} H \ge d/(2\tau)$.

Step 4: the full algebra, $\Lambda^{\circ}_{\mathcal{H}}(\operatorname{lse}_\tau) = (d-1)/(d\tau)$. For $\mathcal{H} = \mathbb{S}^+_d$, $\nu_{\mathbb{S}^+_d}(M_p) = \|M_p\|_{\mathrm{nuc}} = \operatorname{Tr} M_p = 1 - \|p\|_2^2$, and $\sup_{p \in \Delta^{\circ}}(1 - \|p\|_2^2) = 1 - 1/d$.

Step 5: the full algebra, $\Lambda_{\mathcal{H}}(\operatorname{lse}_\tau) = (d-1)/(2\tau)$. We compute the least trace of $H \succeq 0$ with $H \succeq M_p$ for all p; $\Lambda_{\mathcal{H}}(\operatorname{lse}_\tau)$ is $1/\tau$ times that number. The feasible set is convex and invariant under conjugation by permutation matrices, because the family $\{M_p\}$ is. Averaging a minimizer over all permutations produces a feasible minimizer $H = aI + bJ$ with $J = \mathbf{1}\mathbf{1}^{\top}$. Since $M_p\mathbf{1} = 0$, both M_p and H leave $\mathbf{1}^{\perp}$ invariant, and $H - M_p \succeq 0$ is equivalent to $a + bd \ge 0$ and $a \ge \lambda_{\max}(M_p)$. By (24) and its approximate equality case, $\sup_p \lambda_{\max}(M_p) = \frac{1}{2}$. Hence $a \ge \frac{1}{2}$, $b \ge -a/d$, and $\operatorname{Tr} H = d(a+b) \ge a(d-1) \ge \frac{d-1}{2}$, with equality for $H^* = \frac{1}{2}(I - J/d)$. Therefore $\Lambda_{\mathcal{H}}(\operatorname{lse}_\tau) = (d-1)/(2\tau)$ and $\varrho_{\mathcal{H}}(\operatorname{lse}_\tau) = d/2$.

For the scalar algebra, $\operatorname{rk}\mathcal{K} = 1$, and Theorem 4.1 gives $L_{\mathcal{H}} = \Lambda^{\circ}_{\mathcal{H}} = \Lambda_{\mathcal{H}}$ for every f.

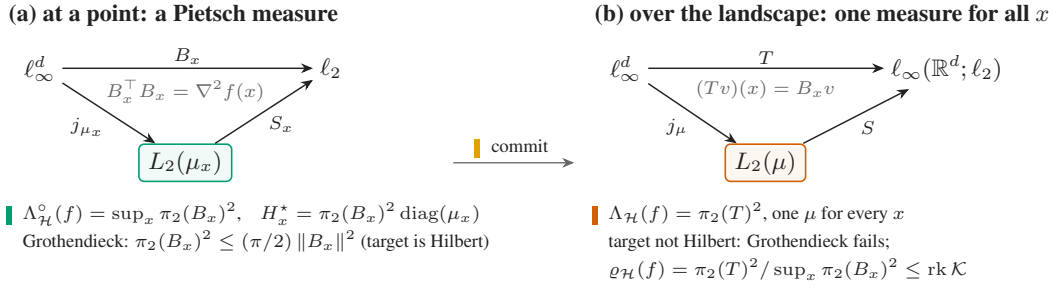

Figure 8: **Why Grothendieck, and why only at a point** (convex f, diagonal geometry, $\nabla^2 f(x) = B_x^\top B_x$). **(a)** At a point, $B_x : \ell_\infty^d \to \ell_2$ factors through $L_2(\mu_x)$ for a probability measure μ_x on the coordinates (Pietsch). The least $\|S_x\|$ is $\pi_2(B_x)$, the optimal μ_x is the normalized diagonal of the ideal preconditioner $H_x^\star = \pi_2(B_x)^2 \operatorname{diag}(\mu_x)$, and $\Lambda^\circ_{\mathcal{H}}(f) = \sup_x \pi_2(B_x)^2$ (Remark 2.2). The target is a Hilbert space, so Grothendieck's theorem gives $\pi_2(B_x)^2 \le \frac{\pi}{2}\|\nabla^2 f(x)\|_{\infty\to 1}$: the pointwise gap of Theorem 3.1. **(b)** A committed preconditioner is one measure μ for the whole field $Tv = (B_x v)_x$, with $\Lambda_{\mathcal{H}}(f) = \pi_2(T)^2$. Its target $\ell_\infty(\mathbb{R}^d;\ell_2)$ is not a Hilbert space, Grothendieck's theorem fails there, and the rotation $\varrho_{\mathcal{H}}(f) = \pi_2(T)^2/\sup_x \pi_2(B_x)^2 \le \operatorname{rk}\mathcal{K}$ (Theorem 4.1) measures the failure.

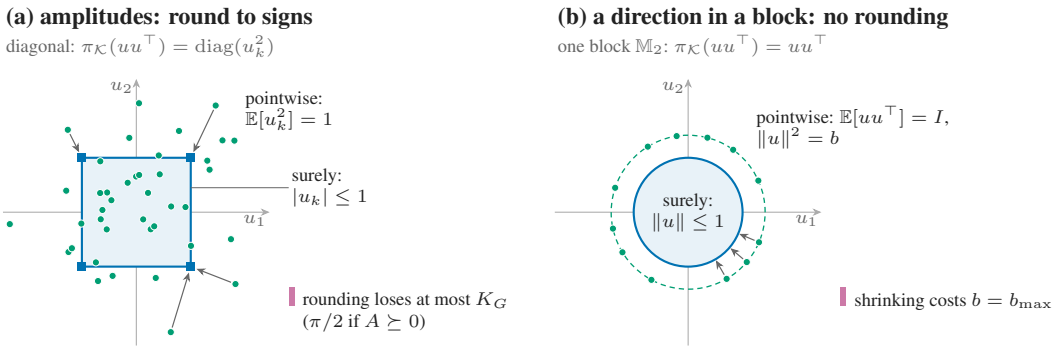

Figure 9: **Two kinds of randomness in a pointwise probe** (Section 3; $d = 2$, both panels on one scale). **(a)** Amplitudes can be rounded. In diagonal geometry the pointwise budget $\mathbb{E}[u_k^2] \le 1$ admits Gaussian amplitudes $u \sim N(0, I_2)$ (dots: 40 draws), many of them outside the sure budget $|u_k| \le 1$ (square). Grothendieck's inequality rounds such probes to sign vectors (the corners) at a cost of at most K_G in the reading; for positive semidefinite Hessians the signs of a Gaussian probe (arrows) cost at most $\pi/2$. Both are dimension-free worst cases; the drawing is schematic. **(b)** A direction inside a matrix block cannot be rounded. In one block $\mathbb{M}_2$ ($b = 2$), a direction uniform on the circle of radius $\sqrt{b}$ meets $\mathbb{E}[uu^\top] = I$, yet every realization has $\|u\|^2 = b$ and spends b times its share of the sure budget $\|u\| \le 1$ (disk). Shrinking onto the disk (arrows) divides the reading of $A = I$ by b, and no sure probe reads more than 1: the factor $b = b_{\max}$ is attained (Theorem 3.1(a)).

D GROTHENDIECK CONSTANTS

D.1 DUALITY

Lemma D.1 (duality)**.** For every well-structured $\mathcal{H}$ and $A \in \mathbb{S}_d$,

$$\nu_{\mathcal{H}}(A) = \max\big\{\langle A, Y - Z\rangle : Y, Z \succeq 0,\ \pi_{\mathcal{K}}(Y + Z) \preceq I\big\}. \tag{25}$$

If $A \succeq 0$, one may take $Z = 0$.

Proof of Lemma D.1. Step 1: weak duality. Let $H \in \mathcal{H}$ satisfy $-H \preceq A \preceq H$ and let $Y, Z \succeq 0$ satisfy $\pi_{\mathcal{K}}(Y + Z) \preceq I$. Then $\langle A, Y\rangle \le \langle H, Y\rangle$ and $-\langle A, Z\rangle \le \langle H, Z\rangle$, and since $H \in \mathcal{K}$ and $\pi_{\mathcal{K}}$

is self-adjoint,

$$\langle A, Y - Z\rangle \;\le\; \langle H, Y + Z\rangle = \langle H, \pi_{\mathcal{K}}(Y + Z)\rangle \;\le\; \langle H, I\rangle = \operatorname{Tr} H,$$

the last inequality because $H \succeq 0$ and $I - \pi_{\mathcal{K}}(Y + Z) \succeq 0$. Minimizing over H and maximizing over (Y, Z) gives "$\ge$" in (25).

Step 2: the conic formulation. The primal is $\min\{\operatorname{Tr} H : \; H \in \mathcal{H},\; H - A \in \mathbb{S}^+_d,\; H + A \in \mathbb{S}^+_d\}$: a linear objective over the closed convex cone $\mathcal{H}$ (inside the space $\mathcal{K}_{\mathrm{s}}$) with two conic constraints in $\mathbb{S}^+_d$. Its Lagrangian with multipliers $Y, Z \in \mathbb{S}^+_d$ is, for $H \in \mathcal{H}$,

$$\mathcal{L}(H; Y, Z) = \operatorname{Tr} H - \langle Y, H - A\rangle - \langle Z, H + A\rangle = \langle A, Y - Z\rangle + \big\langle H,\; I - \pi_{\mathcal{K}}(Y + Z)\big\rangle,$$

where we used $\langle H, Y + Z\rangle = \langle H, \pi_{\mathcal{K}}(Y + Z)\rangle$ for $H \in \mathcal{K}$.

Step 3: the dual function. The dual cone of $\mathcal{H}$ inside $\mathcal{K}_{\mathrm{s}}$ is $\mathcal{H}$ itself (Lemma B.1(b)), and $I - \pi_{\mathcal{K}}(Y + Z) \in \mathcal{K}_{\mathrm{s}}$. Hence $\inf_{H \in \mathcal{H}} \langle H, I - \pi_{\mathcal{K}}(Y + Z)\rangle$ equals 0 if $I - \pi_{\mathcal{K}}(Y + Z) \in \mathcal{H}$, i.e. if $\pi_{\mathcal{K}}(Y + Z) \preceq I$, and equals $-\infty$ otherwise. The dual problem is therefore exactly the right-hand side of (25).

Step 4: strong duality. The point $H_0 = (2\|A\|_{\mathrm{op}} + 1)I$ is positive definite, hence in the relative interior of $\mathcal{H}$, and $H_0 \pm A \succ 0$; the primal is bounded below by 0. Theorem A.3 gives equality of the optimal values and attainment of the dual maximum.

Step 5: the PSD case. If $A \succeq 0$ then $\langle A, Z\rangle \ge 0$, and $\pi_{\mathcal{K}}(Y) \preceq \pi_{\mathcal{K}}(Y + Z)$ by positivity of $\pi_{\mathcal{K}}$; so replacing (Y, Z) by $(Y, 0)$ keeps feasibility and does not decrease $\langle A, Y - Z\rangle$. □

Remark D.2 (the integrality-gap picture). For $u, v \in B_{\mathcal{H}}$ write $\frac{1}{2}(uv^\top + vu^\top) = Y - Z$ with $Y = \frac{1}{4}(u + v)(u + v)^\top$ and $Z = \frac{1}{4}(u - v)(u - v)^\top$; then $\pi_{\mathcal{K}}(Y + Z) = \frac{1}{2}\pi_{\mathcal{K}}(uu^\top + vv^\top) \preceq I$ by Lemma B.2(a). Hence $\mathcal{C}_{\mathcal{H}} \subseteq \mathcal{W}_{\mathcal{H}}$, both $\|\cdot\|_{\mathcal{H}}$ and $\nu_{\mathcal{H}}(\cdot)$ are support functions of these sets, and $K(\mathcal{H}) = \min\{K : \; \mathcal{W}_{\mathcal{H}} \subseteq K\,\mathcal{C}_{\mathcal{H}}\}$. For the diagonal algebra, $\mathcal{C}_{\mathcal{H}}$ is the unit ball of the norm dual to $\|\cdot\|_{\infty\to 1}$ and $\mathcal{W}_{\mathcal{H}} = \{Y - Z : \; Y, Z \succeq 0,\; \operatorname{diag}(Y + Z) \le \mathbf{1}\}$ is the feasible set of the Grothendieck SDP; computing $\|A\|_{\infty\to 1}$ is NP-hard (Alon & Naor, 2006).

D.2 Proof of Theorem 3.1

Coordinates and real traces. Use Lemma B.5, and put $q_i = \dim_{\mathbb{R}} \mathbb{F}_i$, $r_i = q_i n_i$, $d_i = b_i r_i$. Orthogonal conjugation preserves both matrix norms: $\nu_{U\mathcal{H}U^\top}(A) = \nu_{\mathcal{H}}(U^\top A U)$ and $\|A\|_{U\mathcal{H}U^\top} = \|U^\top A U\|_{\mathcal{H}}$. We may therefore work in a real orthonormal basis adapted to the components V_i. Nothing in the following proof assumes $A \in \mathcal{K}$.

For the split special case, identify the i-th component with a real $b_i \times n_i$ matrix X_i, using row-major vectorization. Direct computation of the orthogonal projection gives

$$\pi_{\mathcal{K}}(M) = \bigoplus_i \tfrac{1}{n_i} \operatorname{Tr}_2(M_{ii}) \otimes I_{n_i}, \qquad B_{\mathcal{H}} = \{x : \|X_i\|_{\mathrm{op}} \le \sqrt{n_i}\;\forall i\}. \tag{26}$$

Indeed, $\operatorname{Tr}_2(x_i x_i^\top) = X_i X_i^\top$, and $\pi_{\mathcal{K}}(xx^\top) \preceq I$ is equivalent to $X_i X_i^\top \preceq n_i I_{b_i}$. The general proof needs only the product-ball inclusion in Lemma B.5(b).

Proof of the upper bounds in (8). Fix $A \in \mathbb{S}_d$ and a dual-feasible pair $Y = PP^\top$, $Z = QQ^\top$, with rows $p_\alpha, q_\alpha \in \mathbb{R}^N$. Pad the factors with zero columns if necessary so their column dimensions agree.

Step 1: energy in each component. Since $P_i \in \mathcal{K}$ and $\pi_{\mathcal{K}}(Y + Z) \preceq I$,

$$\sum_{\alpha \in V_i} (\|p_\alpha\|^2 + \|q_\alpha\|^2) = \operatorname{Tr}(P_i(Y + Z)) = \operatorname{Tr}(P_i \pi_{\mathcal{K}}(Y + Z)) \le \operatorname{Tr} P_i = b_i r_i. \tag{27}$$

Here $\alpha \in V_i$ means that the coordinate belongs to the chosen basis of V_i.

Step 2: a scalar Grothendieck sum. Set $r'_\alpha = (p_\alpha, q_\alpha)$ and $s'_\alpha = (p_\alpha, -q_\alpha)$. These vectors have the same norm w_α. Normalize them to unit vectors $\hat r_\alpha, \hat s_\alpha$ (arbitrarily if $w_\alpha = 0$), and put $D = \operatorname{diag}(w)$, $\widetilde A = DAD$. Then

$$\langle A, Y - Z\rangle = \sum_{\alpha,\beta} \widetilde A_{\alpha\beta} \langle \hat r_\alpha, \hat s_\beta\rangle \le K_G^{(d)} \max_{\epsilon,\delta \in \{\pm 1\}^d} (D\epsilon)^\top A (D\delta), \tag{28}$$

by Theorem A.1.

Step 3: feasibility after rounding. For $u = D\epsilon$, (27) gives $\|u_i\|_2^2 \le b_i r_i$. Hence

$$\|u_i/\sqrt{b_{\max}}\|_2^2 \le r_i \qquad \forall i,$$

so $u \in \sqrt{b_{\max}}\, B_{\mathcal{H}}$ by Lemma B.5(b); the same holds for $D\delta$. Thus every rounded bilinear form in (28) is at most $b_{\max}\|A\|_{\mathcal{H}}$. Maximizing over dual-feasible pairs proves $K(\mathcal{H}) \le K_G^{(d)} b_{\max}$. The inequality $k(\mathcal{H}) \le K(\mathcal{H})$ follows directly from their definitions.

Step 4: PSD matrices. If $A \succeq 0$, use $Z = 0$ in Lemma D.1. Now $\widetilde{A} = DAD \succeq 0$, so the little Grothendieck inequality gives

$$\langle A, Y\rangle \le \tfrac{\pi}{2} \max_{\epsilon} (D\epsilon)^\top A (D\epsilon) \le \tfrac{\pi}{2} b_{\max}\|A\|_{\mathcal{H}}.$$

The last inequality uses Step 3 and the PSD formula in Lemma C.1. This proves $k(\mathcal{H}) \le (\pi/2) b_{\max}$. □

Proof of the lower bound $k(\mathcal{H}) \ge b_{\max}$. Choose a component i with $b_i = b_{\max}$, and let $A = E_W$ be the real orthogonal projector onto one irreducible copy $W \cong \mathbb{F}_i^{b_i}$ in V_i. It is PSD, but need not belong to $\mathcal{K}$.

If $H \in \mathcal{H}$ satisfies $H \succeq E_W$, its restriction to W is at least the identity. Since the same matrix acts on all equivalent copies, $H|_{V_i} \succeq I_{V_i}$, whence $\operatorname{Tr} H \ge d_i$. Conversely the central projector P_i is feasible and has trace d_i, so $\nu_{\mathcal{H}}(E_W) = d_i$. By Lemma B.5(c) and the PSD norm formula,

$$\|E_W\|_{\mathcal{H}} = \sup_{x \in B_{\mathcal{H}}} \|E_W x\|_2^2 = r_i.$$

Their ratio is $d_i/r_i = b_i = b_{\max}$, as claimed. In the split coordinates (26), this is the original column projector $I_{b_i} \otimes e_1 e_1^\top$. □

Proof of (a): one simple component. Write $d = br$, where $r = n \dim_{\mathbb{R}} \mathbb{F}$. The scalar preconditioner gives $\nu_{\mathcal{H}}(A) \le d\|A\|_{\mathrm{op}}$ for every $A \in \mathbb{S}_d$. Lemma B.5(b) gives $\sqrt{r} B_2 \subseteq B_{\mathcal{H}}$, so $\|A\|_{\mathcal{H}} \ge r\|A\|_{\mathrm{op}}$. Therefore $K(\mathcal{H}) \le d/r = b$, matching the PSD lower bound above. □

Proof of (b): commutative algebras. Lemma B.4 gives a split generating algebra with every $b_i = 1$ and the same $\mathcal{H}$. Applying (8) proves the claim. □

Proof of (c): the diagonal algebra. Step 1: the lower bound $K(\mathcal{D}_{2n}^+) \ge K_G^{(n)}$. Let $a \in \mathbb{R}^{n\times n}$ and unit vectors $p_i, q_j \in \mathbb{R}^N$ be arbitrary (any instance of Grothendieck's inequality can be realized in a finite-dimensional space), and set

$$d = 2n, \qquad A = \begin{pmatrix} 0 & a \\ a^\top & 0 \end{pmatrix} \in \mathbb{S}_d.$$

For $u = (u_1, u_2)$, $v = (v_1, v_2) \in \{\pm 1\}^{2n}$ one has $u^\top A v = u_1^\top a v_2 + u_2^\top a^\top v_1$, and the two terms depend on disjoint sets of sign variables; hence $\|A\|_{\infty\to 1} = 2\|a\|_{\infty\to 1}$. Let $P, Q \in \mathbb{R}^{n\times N}$ have rows p_i, q_j, and define

$$Y = \tfrac{1}{2}\begin{pmatrix} P \\ Q \end{pmatrix}\begin{pmatrix} P \\ Q \end{pmatrix}^\top, \qquad Z = \tfrac{1}{2}\begin{pmatrix} P \\ -Q \end{pmatrix}\begin{pmatrix} P \\ -Q \end{pmatrix}^\top.$$

Both are PSD; $\operatorname{diag}(Y + Z) = \mathbf{1}$ because the i-th diagonal entry is $\frac{1}{2}\|p_i\|^2 + \frac{1}{2}\|p_i\|^2 = 1$ (and similarly for the q_j); and

$$Y - Z = \begin{pmatrix} 0 & PQ^\top \\ QP^\top & 0 \end{pmatrix}, \qquad \text{so} \qquad \langle A, Y - Z\rangle = 2\sum_{i,j} a_{ij}\langle p_i, q_j\rangle.$$

By weak duality (Step 1 of the proof of Lemma D.1; strong duality is not needed here), $\nu_{\mathcal{D}^+}(A) \ge 2\sum_{ij} a_{ij}\langle p_i, q_j\rangle$, hence

$$\frac{\nu_{\mathcal{D}^+}(A)}{\|A\|_{\infty\to 1}} \ge \frac{\sum_{ij} a_{ij}\langle p_i, q_j\rangle}{\|a\|_{\infty\to 1}}.$$

Taking the supremum over all instances gives $K(\mathcal{D}^+_{2n}) \geq K_G^{(n)}$. For odd d, a matrix supported on the first $d-1$ coordinates has the same ν and the same $\|\cdot\|_{\infty\to 1}$ in $\mathbb{R}^d$, so $K(\mathcal{D}^+_d) \geq K(\mathcal{D}^+_{d-1})$.

Step 2: the upper bound is (8) with $b_{\max} = 1$. Since $K_G^{(n)} \uparrow K_G$, $\sup_d K(\mathcal{D}^+_d) = K_G$.

Step 3: the little constant. For $A \succeq 0$, Lemma D.1 gives $\nu_{\mathcal{D}^+_d}(A) = \max\{\langle A, Y\rangle : Y \succeq 0,\ \operatorname{diag} Y \leq \mathbf{1}\}$, and the constraint may be taken with equality, because adding a nonnegative diagonal to Y keeps it PSD and does not decrease $\langle A, Y\rangle$. In Gram form this is $\max \sum_{ij} A_{ij}\langle p_i, p_j\rangle$ over unit vectors, while $\|A\|_{\infty\to 1} = \max_\epsilon \epsilon^\top A\epsilon$ by Lemma C.1. Thus $k(\mathcal{D}^+_d)$ is the integrality gap of the little Grothendieck SDP restricted to $d \times d$ PSD matrices, and the statement that $\pi/2$ is the optimal constant in Theorem A.2 is exactly $\sup_d k(\mathcal{D}^+_d) = \pi/2$. □

Remark D.3 (sharpness of $\pi/2$). That $\pi/2$ cannot be improved in Theorem A.2 is classical (see Pisier, 2012, §5); the standard witness is the Gram matrix $A = PP^\top$ of n independent uniformly random unit vectors $p_i \in \mathbb{R}^m$ with $1 \ll m \ll n$. The Gram point $Y = PP^\top$ is feasible for the SDP with $\langle A, Y\rangle = \sum_{ij}\langle p_i, p_j\rangle^2 \approx n^2/m$, whereas

$$\max_\epsilon \epsilon^\top A\epsilon = \max_\epsilon \Big\|\sum_i \epsilon_i p_i\Big\|_2^2 \approx n^2 \sup_{\|u\|_2=1} \big(\mathbb{E}|\langle p, u\rangle|\big)^2 \approx n^2 \cdot \frac{2}{\pi m},$$

so the ratio approaches $\pi/2$. In the language of Proposition 4.2(i) this is the convex quadratic $f(x) = \frac{1}{2}x^\top PP^\top x$, for which $\Lambda_{\mathcal{H}}/L_{\mathcal{H}} \to \pi/2$.

E Constrained acceleration and oracle complexity

This appendix proves the accelerated guarantee of Theorem 5.1 for every well-structured geometry, compares it with the global-minimizer guarantee of Xie et al. (2026), and then assembles the oracle lower bounds, using the explicit C^2 resisting-oracle construction of Appendix F.

E.1 A basis-free accelerated method

Throughout this subsection $\mathcal{H} = \mathbb{S}^+_d \cap \mathcal{K}$ is well-structured, X is convex and compact with $\|x - y\|_{\mathcal{H}} \leq 2R$ for $x, y \in X$, and $\|v\|^2_M = v^\top Mv$ for $M \succeq 0$; M^+ is the Moore–Penrose pseudoinverse and $M^{+/2} = (M^{1/2})^+$. For C^1 functions, $\Lambda_{\mathcal{H}}(f) = \inf_{H\in\mathcal{H},\,\operatorname{Tr} H\leq 1} L_{\|\cdot\|_H}(f)$ with the extended dual seminorms of the next subsection; by Lemma E.2(c), which holds for every $H \succeq 0$, this agrees with (4) for convex C^2 functions, and scaling $B = cH$ shows $\Lambda_{\mathcal{H}}(f) = \inf\{\operatorname{Tr} B :\ B \in \mathcal{H},\ L_{\|\cdot\|_B}(f) \leq 1\}$. The minimizers in (11) exist because X is compact; when they are not unique, any choice works below.

Lemma E.1 (tools)**.** Let $M, G, U \in \mathcal{H}$, $B \in \mathcal{H}$, and $x, u \in \mathbb{R}^d$.
(a) $u^\top Mu = \operatorname{Tr}(M\pi_{\mathcal{K}}(uu^\top)) \leq \operatorname{Tr}(M)\,\|u\|^2_{\mathcal{H}}$.
(b) $u \in \operatorname{range}\pi_{\mathcal{K}}(uu^\top)$.
(c) If $G' = G + U$, then $\operatorname{Tr}(G'^{+/2}U) \leq 2\operatorname{Tr}(G'^{1/2} - G^{1/2}) \leq 2\operatorname{Tr} U^{1/2}$.
(d) If f is convex and C^1 with $L_{\|\cdot\|_B}(f) \leq 1$, then for all x, y the difference $\delta = \nabla f(x) - \nabla f(y)$ lies in $\operatorname{range} B$ and $\langle\delta, x - y\rangle \geq \|\delta\|^2_{B^+}$.
(e) If $S \succeq 0$ has $\operatorname{range} S \subseteq \operatorname{range} B$ and $c \geq 0$, then $2c\operatorname{Tr} S - \operatorname{Tr}(SB^+S) \leq c^2\operatorname{Tr} B$.

Proof. *(a)* Since $M \in \mathcal{K}$ and $\pi_{\mathcal{K}}$ is a self-adjoint projection onto $\mathcal{K}$, $\langle M, uu^\top\rangle = \langle\pi_{\mathcal{K}}(M), uu^\top\rangle = \langle M, \pi_{\mathcal{K}}(uu^\top)\rangle$. The inequality is $\operatorname{Tr}(MP) \leq \operatorname{Tr}(M)\|P\|_{\mathrm{op}}$ for $M, P \succeq 0$ together with Lemma B.2(a).

(b) The orthogonal projector Q onto $\ker\pi_{\mathcal{K}}(uu^\top)$ is a polynomial in $\pi_{\mathcal{K}}(uu^\top) \in \mathcal{K}_{\mathrm{s}}$, hence lies in $\mathcal{K}$. By the bimodule property $\pi_{\mathcal{K}}(Quu^\top Q) = Q\pi_{\mathcal{K}}(uu^\top)Q = 0$, and trace preservation gives $\|Qu\|_2^2 = \operatorname{Tr}\pi_{\mathcal{K}}(Quu^\top Q) = 0$.

(c) For $\epsilon > 0$ the function $\Phi(Y) = \operatorname{Tr} Y^{1/2}$ is concave on $\mathbb{S}^+_d$ with gradient $\frac{1}{2}Y^{-1/2}$ at $Y \succ 0$, so $\Phi(G + \epsilon I) \leq \Phi(G' + \epsilon I) - \frac{1}{2}\operatorname{Tr}((G' + \epsilon I)^{-1/2}U)$. Since $0 \preceq U \preceq G'$, U vanishes on

$\ker G'$, and letting $\epsilon \downarrow 0$ gives the first inequality. For the second, $W = [G^{1/2}, U^{1/2}] \in \mathbb{R}^{d\times 2d}$ has $WW^\top = G'$, so $\operatorname{Tr} G'^{1/2} = \|W\|_{\mathrm{nuc}} \le \|[G^{1/2}, 0]\|_{\mathrm{nuc}} + \|[0, U^{1/2}]\|_{\mathrm{nuc}} = \operatorname{Tr} G^{1/2} + \operatorname{Tr} U^{1/2}$.

(d) The extended seminorm $\|\delta\|_{B,*}$ is finite, so $\delta \in \operatorname{range} B$ and $\|\delta\|_{B,*} = \|\delta\|_{B^+}$. Integrating the Lipschitz bound along a segment gives $f(z) \le f(y) + \langle \nabla f(y), z-y\rangle + \frac{1}{2}\|z-y\|_B^2$ for all y, z. The convex function $\phi(z) = f(z) - \langle \nabla f(x), z\rangle$ is minimized at x and obeys the same bound, so with $s = B^+\nabla\phi(y)$, $\phi(x) \le \phi(y-s) \le \phi(y) - \frac{1}{2}\|\nabla\phi(y)\|_{B^+}^2$, that is, $f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2}\|\delta\|_{B^+}^2$. Adding the same inequality with x, y exchanged proves (d).

(e) With $P = B^{1/2}B^{+/2}$ the projector onto $\operatorname{range} B$, $PS = S$, and $0 \le \operatorname{Tr}\big((cB^{1/2} - SB^{+/2})(cB^{1/2} - SB^{+/2})^\top\big) = c^2 \operatorname{Tr} B - 2c \operatorname{Tr} S + \operatorname{Tr}(SB^+S)$. □

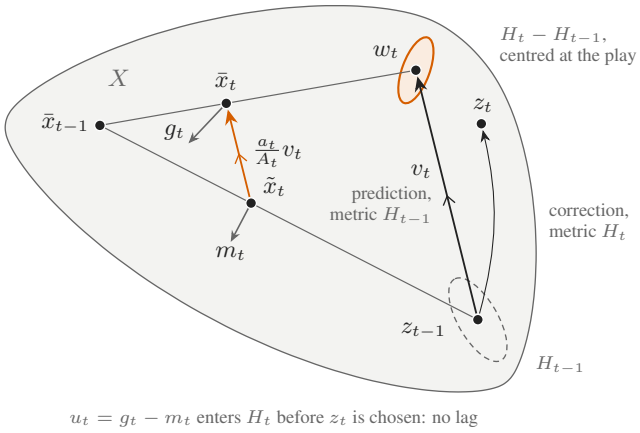

Figure 10: **The gradient difference meets a metric that already contains it.** One round of (11) at $t = 4$, computed exactly for a convex quadratic in the plane with full-matrix metrics. The queries $\tilde{x}_t$ and $\bar{x}_t$ are the same convex combination, with weight $a_t/A_t = 0.4$, of $\bar{x}_{t-1}$ with z_{t-1} and with the play w_t; hence $\bar{x}_t - \tilde{x}_t = (a_t/A_t)v_t$ is a scaled copy of the prediction step $v_t = w_t - z_{t-1}$ (chevrons mark the parallel sides), which links $u_t = g_t - m_t$ to v_t. The ellipses are level sets of the two proximal terms of the correction: H_{t-1} centred at z_{t-1} and the increment $H_t - H_{t-1}$ centred at w_t (shapes exact, sizes not to scale). Because the increment is centred at the play, u_t is measured in H_t.

Proof of Theorem 5.1. Write $v_t = w_t - z_{t-1}$ and $U_t = \pi_{\mathcal{K}}(u_t u_t^\top)$, so $G_t = \sum_{s\le t} U_s \in \mathcal{H}$ and $H_t = G_t^{1/2}/D \in \mathcal{H}$ by Lemma B.1(a),(c). The square root is operator monotone, so $H_t - H_{t-1} \in \mathcal{H}$.

Step 1: feasibility and online-to-batch. The point $\bar{x}_t = A_t^{-1}\sum_{s\le t} a_s w_s$ is a convex combination of points of X, and so is $\tilde{x}_t$; hence every query and the output lie in X. Fix $y \in X$. From $A_t\bar{x}_t = A_{t-1}\bar{x}_{t-1} + a_t w_t$ we get $A_{t-1}(\bar{x}_{t-1} - \bar{x}_t) = a_t(\bar{x}_t - w_t)$, and convexity at $\bar{x}_t$, applied to $\bar{x}_{t-1}$ with weight A_{t-1} and to y with weight a_t, gives $A_t f(\bar{x}_t) - A_{t-1}f(\bar{x}_{t-1}) \le a_t f(y) + \langle g_t, w_t - y\rangle$. Summing, with $A_0 = 0$,

$$A_N\big(f(\bar{x}_N) - f(y)\big) \le \sum_{t=1}^{N} \langle g_t, w_t - y\rangle. \tag{29}$$

Step 2: one round. The quadratic $\psi_t(z) = \frac{1}{2}\|z - z_{t-1}\|_{H_{t-1}}^2 + \frac{1}{2}\|z - w_t\|_{H_t - H_{t-1}}^2$ has Hessian H_t, so optimality of z_t gives, for $y \in X$, $\langle g_t, z_t - y\rangle \le \psi_t(y) - \psi_t(z_t) - \frac{1}{2}\|y - z_t\|_{H_t}^2$. Optimality of w_t, tested at z_t, gives $\langle m_t, w_t - z_t\rangle \le \frac{1}{2}\|z_t - z_{t-1}\|_{H_{t-1}}^2 - \frac{1}{2}\|z_t - w_t\|_{H_{t-1}}^2 - \frac{1}{2}\|v_t\|_{H_{t-1}}^2$. Add these to $\langle g_t, w_t - y\rangle = \langle g_t, z_t - y\rangle + \langle m_t, w_t - z_t\rangle + \langle u_t, w_t - z_t\rangle$. The terms in $z_t - z_{t-1}$ cancel, and the two terms in $z_t - w_t$ combine into $-\frac{1}{2}\|z_t - w_t\|_{H_t}^2$. Because the increment is centred at the play w_t, the gradient difference meets the metric H_t that already contains it: $u_t \in \operatorname{range} U_t \subseteq \operatorname{range} H_t$

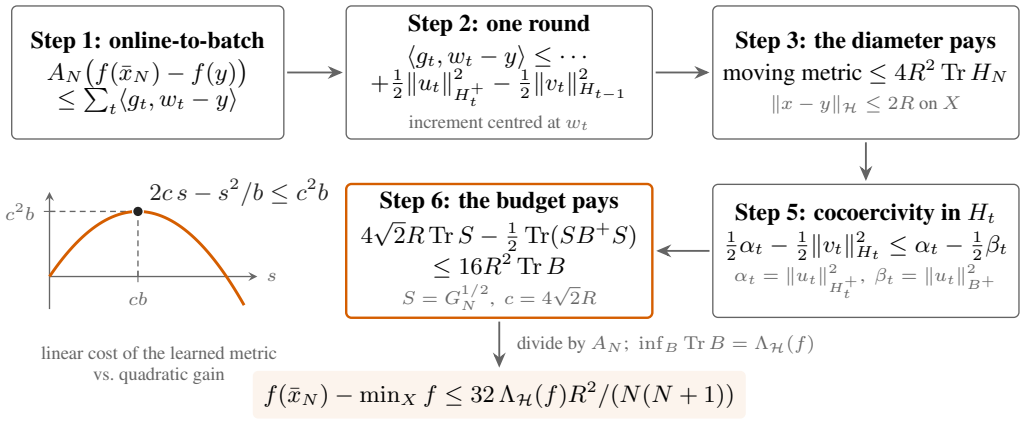

Figure 11: **The ledger of the proof of Theorem 5.1: the global budget pays for the learned metric.** The diameter pays for the moving metric (Step 3). Cocoercivity, read in the learned metric H_t, turns the negative prediction term into the gain $-\frac{1}{2}\beta_t$ measured by a majorant B (Step 5). The completion of squares $2c \operatorname{Tr} S - \operatorname{Tr}(SB^+S) \le c^2 \operatorname{Tr} B$ of Lemma E.1(e), applied with $c = 4\sqrt{2}R$ (inset: its scalar form, drawn for $c = b = 1$), lets the budget $\operatorname{Tr} B$ of any common majorant pay for $S = G_N^{1/2}$; the infimum over B is $\Lambda_{\mathcal{H}}(f)$. No step compares H_t with B in the Loewner order. Step 4, the non-accelerated bound, is omitted.

by (b), so $\langle u_t, e\rangle - \frac{1}{2}\|e\|^2_{H_t} \le \frac{1}{2}\|u_t\|^2_{H_t^+}$ for every e. Therefore

$$\langle g_t, w_t - y\rangle \le \tfrac{1}{2}\|y - z_{t-1}\|^2_{H_{t-1}} - \tfrac{1}{2}\|y - z_t\|^2_{H_t} + \tfrac{1}{2}\|y - w_t\|^2_{H_t - H_{t-1}} + \tfrac{1}{2}\|u_t\|^2_{H_t^+} - \tfrac{1}{2}\|v_t\|^2_{H_{t-1}}. \quad (30)$$

Step 3: the diameter pays for the moving metric. The points y, w_t, z_{t-1} lie in X, so $\|y - w_t\|_{\mathcal{H}}, \|v_t\|_{\mathcal{H}} \le 2R$, and (a) bounds $\frac{1}{2}\|y - w_t\|^2_{H_t - H_{t-1}}$ and $\frac{1}{2}\|v_t\|^2_{H_t - H_{t-1}}$ by $2R^2 \operatorname{Tr}(H_t - H_{t-1})$. Writing $-\frac{1}{2}\|v_t\|^2_{H_{t-1}} = -\frac{1}{2}\|v_t\|^2_{H_t} + \frac{1}{2}\|v_t\|^2_{H_t - H_{t-1}}$, summing (30), telescoping and using $H_0 = 0$,

$$\sum_{t=1}^N \langle g_t, w_t - y\rangle \le 4R^2 \operatorname{Tr} H_N + \sum_{t=1}^N \Big(\tfrac{1}{2}\alpha_t - \tfrac{1}{2}\|v_t\|^2_{H_t}\Big), \qquad \alpha_t := \|u_t\|^2_{H_t^+}. \quad (31)$$

By (a), $\alpha_t = \operatorname{Tr}(H_t^+ U_t) = D \operatorname{Tr}(G_t^{+/2} U_t)$, so (c) gives $\sum_t \alpha_t \le 2D \operatorname{Tr} G_N^{1/2}$; also $\operatorname{Tr} H_N = \operatorname{Tr} G_N^{1/2}/D$.

Step 4: the non-accelerated bound. Drop the negative terms in (31). With $D = \sqrt{2}R$, the right-hand side is at most $(4R^2/D + D) \operatorname{Tr} G_N^{1/2} = 3\sqrt{2}R \operatorname{Tr} G_N^{1/2} \le 3\sqrt{2}R \sum_t \operatorname{Tr} U_t^{1/2}$ by the second inequality of (c), and $\operatorname{Tr} U_t^{1/2} = \|u_t\|_{\mathcal{H},*}$ by Lemma B.2(c). Since $\bar{x}_t - \tilde{x}_t = (a_t/A_t)v_t$ and $a_t^2/A_t = 2t/(t+1) < 2$, $\|u_t\|_{\mathcal{H},*} \le a_t L_{\mathcal{H}}(f)\|\bar{x}_t - \tilde{x}_t\|_{\mathcal{H}} \le 2L_{\mathcal{H}}(f)\|v_t\|_{\mathcal{H}} \le 4L_{\mathcal{H}}(f)R$. Hence the regret is at most $12\sqrt{2}L_{\mathcal{H}}(f)R^2 N$, and (29) with $A_N = N(N+1)/2$ gives the first bound.

Step 5: cocoercivity read in the learned metric. Let $B \in \mathcal{H}$ with $L_{\|\cdot\|_B}(f) \le 1$. By (d), $\delta_t = \nabla f(\bar{x}_t) - \nabla f(\tilde{x}_t) \in \operatorname{range} B$ and $\langle u_t, v_t\rangle = A_t\langle \delta_t, \bar{x}_t - \tilde{x}_t\rangle \ge A_t\|\delta_t\|^2_{B^+} = (A_t/a_t^2)\|u_t\|^2_{B^+} \ge \frac{1}{2}\beta_t$, where $\beta_t := \|u_t\|^2_{B^+} = \operatorname{Tr}(B^+U_t)$ by (a). Cauchy–Schwarz in the metric H_t gives $\langle u_t, v_t\rangle \le \sqrt{\alpha_t}\,\|v_t\|_{H_t}$. If $\alpha_t > 0$, then $\|v_t\|^2_{H_t} \ge \beta_t^2/(4\alpha_t)$ and, since $\beta_t^2/\alpha_t \ge 4\beta_t - 4\alpha_t$,

$$\tfrac{1}{2}\alpha_t - \tfrac{1}{2}\|v_t\|^2_{H_t} \le \tfrac{1}{2}\alpha_t - \tfrac{\beta_t^2}{8\alpha_t} \le \alpha_t - \tfrac{1}{2}\beta_t;$$

if $\alpha_t = 0$, then $u_t = 0$ and the same bound holds. No step compares H_t with B.

Step 6: the budget pays for the learned metric. Insert Step 5 into (31) and use Step 3 and $4R^2/D + 2D = 4\sqrt{2}R$:

$$\sum_{t=1}^N \langle g_t, w_t - y\rangle \le 4\sqrt{2}R \operatorname{Tr} S - \tfrac{1}{2}\operatorname{Tr}(SB^+S), \qquad S = G_N^{1/2},$$

since $\operatorname{Tr}(SB^+S) = \operatorname{Tr}(B^+G_N) = \sum_t \beta_t$. Each u_t lies in $\operatorname{range} B$, and the projector P onto $\operatorname{range} B$ lies in $\mathcal{K}$, so $U_t = PU_tP$ by the bimodule property and $\operatorname{range} S \subseteq \operatorname{range} B$. Lemma E.1(e) with $c = 4\sqrt{2}R$ bounds the right-hand side by $16R^2 \operatorname{Tr} B$. By (29), $f(\bar{x}_N) - f(y) \le 32R^2 \operatorname{Tr} B/(N(N+1))$. The run does not depend on B, so taking y a minimizer over X and the infimum over B proves the second bound. □

The proof combines four known ingredients: anytime online-to-batch conversion (Cutkosky, 2019), optimistic extragradient steps as in Kavis et al. (2019) and Joulani et al. (2020), AdaGrad metrics built from gradient differences, and proximal increments centred at the play (McMahan, 2017). None of its steps chooses a basis: the learned metric is compared with the unknown majorant only through the trace in Step 6, where the global budget $\operatorname{Tr} B$ pays for it. For the ball X_R, Kovalev (2026, Theorem 5) proves the accelerated rate for preconditioner spaces satisfying Assumption 1 there, which include diagonal, full-matrix and one-sided Shampoo geometries, by a projection-free variant of one-sided Shampoo with Nesterov acceleration, with an additional term of order $\delta R \dim(X)/K^2$ from an initialization $\delta > 0$. The argument above allows any compact convex domain, has no additional term and yields the non-accelerated bound in the same run; Kovalev (2025) accelerates AdaGrad-type methods when the smoothness operator commutes with the preconditioners. In diagonal geometry the coordinate-wise analysis of AdaACSA (Ene et al., 2021, Theorem 3.3) gives the same rate with an extra factor $\ln(2\beta_i)$ per coordinate. The natural matrix analogue of its multiplicative update, an exponential of accumulated movements, need not increase in the Loewner order, which that analysis uses.

E.2 Dimension reduction and the two function classes

For a positive semidefinite H, the dual of $\|x\|_H = \sqrt{x^\top H x}$ is understood as an extended seminorm: it equals $\|g\|_{H^+}$ on $\operatorname{range} H$ and $+\infty$ off that range. We define $L_{\|\cdot\|_H}(f)$ as the least A for which $\|\nabla f(x) - \nabla f(y)\|_{H,*} \le A\|x - y\|_H$ for every pair x, y. This formulation includes null directions without a $0/0$ quotient.

Lemma E.2. (a) For every C^1 function f on $\mathbb{R}^d$, $\Lambda^+_{\mathcal{D}^+_d}(f) \le dL_\infty(f)$. (b) If $h : \mathbb{R}^m \to \mathbb{R}$ is C^1 and $f = h \circ \Pi_m$, where $\Pi_m(x) = (x_1, \dots, x_m)$, then $L_\infty(f) = L_\infty(h)$ and $\Lambda^+_{\mathcal{D}^+_d}(f) = \Lambda^+_{\mathcal{D}^+_m}(h)$. For C^2 functions the same equality holds for pointwise adaptive smoothness, and hence for rotation. (c) If f is C^2, $H \succeq 0$ and $c \ge 0$, then $L_{\|\cdot\|_H}(f) \le c$ if and only if $-cH \preceq \nabla^2 f(x) \preceq cH$ for every x. Consequently the Lipschitz form of $\Lambda^+_{\mathcal{D}^+_d}$ agrees with (4).

Proof. *(a)* Choose $H = I/d$. The inequalities $\|g\|_2 \le \|g\|_1$ and $\|v\|_\infty \le \|v\|_2$ give

$$\Lambda^+_{\mathcal{D}^+_d}(f) \le dL_{\|\cdot\|_2}(f) \le dL_\infty(f).$$

(b) The gradient of f is $(\nabla h(\Pi_m x), 0)$, which proves the standard-smoothness equality by restriction to the first m coordinates. Extending a feasible diagonal weight for h by zeros proves the "$\le$" direction for adaptive smoothness; restricting a weight for f to its leading $m \times m$ block proves the reverse direction, without increasing its trace. For a C^2 function, the Hessian is $\operatorname{diag}(\nabla^2 h, 0)$ and the identical extension/restriction argument applies pointwise to its semidefinite majorant.

(c) If $-cH \preceq \nabla^2 f \preceq cH$ everywhere, the block-matrix argument of Lemma C.1 gives $|w^\top \nabla^2 f(z)(y - x)| \le c\|w\|_H \|y - x\|_H$, and integrating along the segment from x to y gives $\langle \nabla f(y) - \nabla f(x), w\rangle \le c\|w\|_H \|y - x\|_H$ for every w. For $w \in \ker H$ this forces $\langle \nabla f(y) - \nabla f(x), w\rangle = 0$, so the extended dual seminorm of the gradient difference is at most $c\|y - x\|_H$. Conversely, a Lipschitz bound gives $v^\top \nabla^2 f(x) v = \lim_{t\to 0} \langle \nabla f(x + tv) - \nabla f(x), v\rangle / t \in [-c\, v^\top H v,\ c\, v^\top H v]$. Finally, $H' = cH$ turns $\min\{c : L_{\|\cdot\|_H}(f) \le c\}$ over $\operatorname{Tr} H \le 1$ into the least trace of a two-sided majorant. □

Proof of Theorem 5.2. Lower bound: tune the exponent to the budget. Put $m = T$,

$$\kappa = \min\left\{k(1 - T^{-2}),\ Te^{-2/3}\right\}, \qquad r = \frac{2 \ln T}{\ln(T/\kappa)}.$$

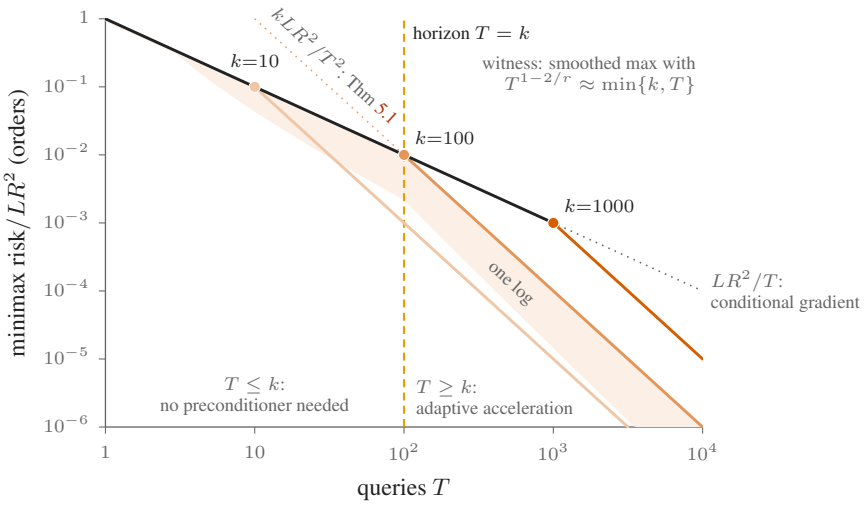

Figure 12: **Rotation is the acceleration horizon** (Theorem 5.2). For convex f with $L_\infty \le L$ and rotation at most k on the box of radius R, up to constants and one logarithm the minimax risk is LR^2/T until $T = k$ and kLR^2/T^2 afterwards; shown are the orders $\min\{1/T, k/T^2\}$ for $k = 10, 100, 1000$. Before the horizon conditional gradient, which uses no preconditioner, is optimal (dotted: LR^2/T continued); after it the method of Theorem 5.1 is (dotted: kLR^2/T^2 before the kink, $k = 100$). The band is the one-log gap between the two bounds of Theorem 5.2 for $k = 100$. The hard instances are smoothed maxima whose exponent r sets $T^{1-2/r} \approx \min\{k, T\}$. In one-sided Shampoo geometry the horizon is $\min\{k, d_L, d_R\}$ (Proposition E.5).

Since $k \ge 2$ and $T \ge 3$, we have $1 < \kappa \le Te^{-2/3}$, so $r \in (2, 3\ln T]$ and $T^{1-2/r} = \kappa$. Proposition F.3, proved in Appendix F, gives for each deterministic method at horizon T a convex C^2 function with $L_\infty(f) \le L$ and

$$\frac{\Lambda_{\mathcal{D}_d^+}(f)}{L_\infty(f)} \le \frac{\kappa}{1 - T^{-2}} \le k, \qquad \varrho_{\mathcal{H}}(f) \le \frac{\Lambda_{\mathcal{D}_d^+}(f)}{L_\infty(f)} \le k,$$

where the second inequality uses $L_\infty \le \Lambda^\circ_{\mathcal{H}}$. Hence f belongs to $\mathcal{F}_{L,k} \cap C^2$ and to $\mathcal{R}_{L,k}$. Its error is at least $LR^2\kappa/(32(r-1)T^2)$. Now $\kappa \ge \min\{\frac{8}{9}k,\, e^{-2/3}T\} \ge e^{-2/3}\min\{k, T\}$ and $r - 1 \le 3\ln T$, so

$$f(x_T) - \min_{X_R} f \;\ge\; \frac{LR^2}{96e^{2/3}\ln T}\,\min\left\{\frac{1}{T},\, \frac{k}{T^2}\right\}.$$

When $k \ge T$ the choice gives $r = 3\ln T$, the construction of Guzmán & Nemirovski (2015) with active dimension T; for the original C^1 class alone, Theorems A.8 and A.9 with Lemma E.2 give that case as well. When $k(1 - T^{-2}) < Te^{-2/3}$ the smaller exponent lowers the rotation to $\kappa \le k$ and makes the instance hard at the accelerated rate.

Conditional-gradient upper bound. For either class, the curvature constant on X_R is at most $L_\infty(f)\,\mathrm{diam}_\infty(X_R)^2 \le 4LR^2$. Run $N = T - 1$ conditional-gradient updates. Theorem A.10 yields an error at most $8LR^2/(N+2) = 8LR^2/(T+1)$. The resulting feasible point is a permitted final query x_T.

Accelerated upper bound. Round t of (11) needs the answers at $\tilde{x}_1, \bar{x}_1, \ldots, \tilde{x}_t$ to form $\bar{x}_t$, so $N = \lfloor T/2 \rfloor$ rounds use at most $T - 1$ answers; for odd T one query is repeated, ties in the minimizations are broken by taking the minimizer of least Euclidean norm, which makes the method deterministic, and $x_T = \bar{x}_N$. On $\mathcal{F}_{L,k}$, $\Lambda \le kL$; on $\mathcal{R}_{L,k}$, $\Lambda \le (\pi/2)kL$. For $T \ge 3$, $N(N+1) \ge (T^2-1)/4 \ge 2T^2/9$, so Theorem 5.1 gives an error at most $144\Lambda R^2/T^2 \le 72\pi kLR^2/T^2$. The infimum defining minimax risk is no larger than either method's worst-case bound, giving their minimum. Neither upper-bound argument requires $T \le k$. □

Corollary E.3 (commitment is the right uniform constant)**.** Let $2 \le k \le T \le d$ and $T \ge 3$. If a deterministic local-oracle method guarantees $f(x_T) - \min_{X_R} f \le g(f)R^2/T^2$ for its Tth query x_T and every $f \in \mathcal{F}_{L,k}$, then $\sup_{\mathcal{F}_{L,k}} g \ge ckL/\ln T \ge (c/\ln T)\sup_{\mathcal{F}_{L,k}} \Lambda_{\mathcal{D}_d^+}$. The conclusion also holds on the C^2 subclass.

No accelerated guarantee improves on the adaptive constant uniformly over the class by more than a logarithm, at any horizon $k \le T \le d$.

Proof of Corollary E.3. For $k \le T \le d$, Theorem 5.2 implies

$$\sup_{\mathcal{F}_{L,k}} g \cdot \frac{R^2}{T^2} \ge \mathrm{Risk}_{\mathcal{F}_{L,k},X_R}(T) \ge \frac{ckLR^2}{T^2 \ln T}.$$

Thus $\sup g \ge ckL/\ln T$, while $\sup_{\mathcal{F}_{L,k}} \Lambda_{\mathcal{D}_d^+} \le kL$ by definition. The same proof applies to $\mathcal{F}_{L,k} \cap C^2$ because its lower bound has already been established. This compares suprema; it is not a lower bound for each individual $g(f)$. □

E.3 The separate accelerated AdaGrad guarantee

Proposition E.4 (rotation in the cited AdaGrad bound)**.** Let f be convex and C^2, $0 < \Lambda_{\mathcal{H}}(f) < \infty$, and let x^* be a global minimizer. Under the hypotheses of Theorem A.12, take $D > 0$, $\eta = D$ and $\epsilon = (\Lambda_{\mathcal{H}}(f)D/d)^2$. After N updates,

$$f(\bar{x}_N) - f(x^*) \le \frac{C\Lambda_{\mathcal{H}}(f)D^2}{(N+1)^2} \ln\left(2 + \frac{dG}{\Lambda_{\mathcal{H}}(f)D}\right) \le \frac{C\pi\varrho_{\mathcal{H}}(f)L_{\mathcal{H}}(f)D^2}{2(N+1)^2} \ln\left(2 + \frac{dG}{\Lambda_{\mathcal{H}}(f)D}\right),$$

where $G^2 = \max_i \sum_{t<N} g_{t,i}^2$. Algorithm 2 of Xie et al. (2026) requires its iterate-distance bound, and their Algorithm 8 requires $\|x^*\|_\infty \le D$; neither statement replaces x^* by an arbitrary constrained minimizer. If $\Lambda_{\mathcal{H}}(f) = 0$ and a global minimizer exists, f is constant and the error is zero.

Proof. Theorem A.12 bounds the error by

$$\frac{C}{(N+1)^2}\left(\Lambda_{\mathcal{H}}(f)D^2 \ln \frac{\|V_{N-1}\|_{\mathrm{op}}^2}{\epsilon} + 2dD\sqrt{\epsilon}\right).$$

The chosen ϵ makes the second term $2\Lambda_{\mathcal{H}}(f)D^2$, and

$$\ln \frac{\|V_{N-1}\|_{\mathrm{op}}^2}{\epsilon} \le \ln\left(1 + \frac{d^2G^2}{(\Lambda_{\mathcal{H}}(f)D)^2}\right) \le 2\ln\left(2 + \frac{dG}{\Lambda_{\mathcal{H}}(f)D}\right).$$

Absorb the constant term into this positive logarithm and insert (10). Algorithm 8 averages its unprojected intermediate points, so feasibility of $\bar{x}_N$ is not asserted here. This distinction is why the constrained comparison uses Theorem 5.1. □

E.4 One-sided Shampoo geometry

Let $\mathcal{H} = \{L \otimes I_{d_R} : L \in \mathbb{S}_{d_L}^+\}$ act on $X \in \mathbb{R}^{d_L \times d_R}$ by $X \mapsto LX$, so that $\|X\|_{\mathcal{H}} = \|X\|_{\mathrm{op}}/\sqrt{d_R}$, and let $X_R = \{X : \|X\|_{\mathcal{H}} \le R\}$. Define $\mathcal{F}_{L,k}^{\mathrm{Sh}}$ and $\mathcal{R}_{L,k}^{\mathrm{Sh}}$ as in Theorem 5.2, with $L_{\mathcal{H}}$, $\Lambda_{\mathcal{H}}$ and $\varrho_{\mathcal{H}}$ of this geometry in place of L_∞, $\Lambda_{\mathcal{D}_d^+}$ and the diagonal rotation, and let Risk refer to X_R.

Proposition E.5 (one-sided Shampoo)**.** For integers $3 \le T \le \min\{d_L, d_R\}$ and $k \ge 2$, and for both classes,

$$\frac{c\,LR^2}{\ln T} \min\left\{\frac{1}{T}, \frac{k}{T^2}\right\} \le \mathrm{Risk}(T) \le \frac{8LR^2}{T+1},$$

with the constant c of Theorem 5.2. On $\mathcal{F}_{L,k}^{\mathrm{Sh}}$, moreover, $\mathrm{Risk}(T) \le 144kLR^2/T^2$ for every $T \ge 3$, and on both classes $\mathrm{Risk}(T) \le 8\min\{d_L, d_R\}\,LR^2/T^2$ for every $T \ge 2$.

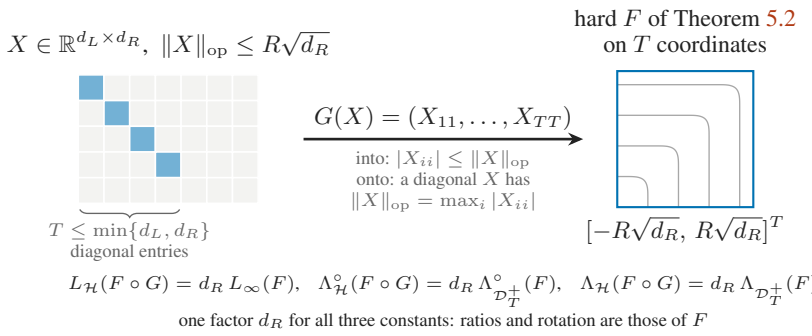

Figure 13: **The lift behind Proposition E.5.** The first $T \le \min\{d_L, d_R\}$ diagonal entries map the ball $\{\|X\|_{\mathrm{op}} \le R\sqrt{d_R}\}$ of one-sided Shampoo geometry onto the box $[-R\sqrt{d_R}, R\sqrt{d_R}]^T$, because $|X_{ii}| \le \|X\|_{\mathrm{op}}$ and a diagonal matrix has operator norm equal to its largest entry. Composing with this map multiplies all three constants by d_R, so ratios and rotation are those of F in diagonal geometry, and the rotation-tuned witness of Theorem 5.2, run with L/d_R and $R\sqrt{d_R}$ in place of L and R, transfers with the same error. Drawn with $d_L = 5$, $d_R = 7$, $T = 4$; the box is a two-coordinate stand-in with level sets of the smoothed maximum ($r = 3$).

Proof. Step 1: diagonal entries. Let $G(X) = (X_{11}, \dots, X_{TT})$. Since $|X_{ii}| \le \|X\|_{\mathrm{op}}$ and diagonal matrices have operator norm equal to their largest absolute entry, G maps X_R onto the box $[-R\sqrt{d_R}, R\sqrt{d_R}]^T$.

Step 2: the three constants lift exactly. Let $F \in C^2(\mathbb{R}^T)$, $f = F \circ G$ and $A = \nabla^2 F(G(X))$, so that the Hessian of f is the quadratic form $V \mapsto \sum_{i,j \le T} A_{ij} V_{ii} V_{jj}$. The first T diagonal entries $G(U)$ of a matrix $U \in B_{\mathcal{H}} = \{\|U\|_{\mathrm{op}} \le \sqrt{d_R}\}$ range exactly over $[-\sqrt{d_R}, \sqrt{d_R}]^T$, so $\|\nabla^2 f(X)\|_{\mathcal{H}} = d_R \|A\|_{\infty \to 1}$. If $L \otimes I \succeq \pm \nabla^2 f(X)$, testing on diagonal matrices V gives $\mathrm{diag}(L_{11}, \dots, L_{TT}) \succeq \pm A$, hence $\mathrm{Tr}(L \otimes I) = d_R \,\mathrm{Tr}\, L \ge d_R\, \nu_{\mathcal{D}_T^+}(A)$. Conversely, if $\mathrm{diag}(h) \succeq \pm A$, then $L = \mathrm{diag}(h, 0, \dots, 0)$ satisfies $\langle V, (L \otimes I)V \rangle = \sum_{i \le T} h_i \sum_j V_{ij}^2 \ge \sum_i h_i V_{ii}^2 \ge |\sum_{ij} A_{ij} V_{ii} V_{jj}|$. Applying these facts at one point and uniformly over all points gives

$$L_{\mathcal{H}}(f) = d_R L_\infty(F), \qquad \Lambda^\circ_{\mathcal{H}}(f) = d_R \Lambda^\circ_{\mathcal{D}_T^+}(F), \qquad \Lambda_{\mathcal{H}}(f) = d_R \Lambda_{\mathcal{D}_T^+}(F),$$

so $\Lambda_{\mathcal{H}}/L_{\mathcal{H}}$ and $\varrho_{\mathcal{H}}$ are those of F in diagonal geometry.

Step 3: lower bound. Run the construction in the proof of Theorem 5.2 on the coordinates $G(X_t)$ of the queries, with L/d_R and $R\sqrt{d_R}$ in place of L and R, and answer with the local oracle of $f^t = F^t \circ G$. Germs of f^t at X_t are determined by germs of F^t at $G(X_t)$, so the transcript is consistent with the final $f = F \circ G$, and by Step 1, $f(X_T) - \min_{X_R} f = F(G(X_T)) - \min F$ over the box. By Step 2 and Lemma E.2(c), which holds for every $H \succeq 0$, f belongs to both classes, and its error is at least $c\,(L/d_R)(R\sqrt{d_R})^2 \min\{1/T, k/T^2\}/\ln T$.

Step 4: upper bounds. Conditional gradient on X_R gives $8LR^2/(T+1)$ exactly as in the diagonal case, because Theorem A.10 holds in any norm and $\mathrm{diam}_{\mathcal{H}}(X_R) = 2R$. For the second bound, $\|U\|_{\mathcal{H}} \le \|U\|_F/\sqrt{d_R}$ and $\|g\|_{\mathcal{H},*} = \sqrt{d_R}\|g\|_{\mathrm{nuc}} \ge \sqrt{d_R}\|g\|_F$ show that ∇f is $(L_{\mathcal{H}}(f)/d_R)$-Lipschitz in the Frobenius norm, and the Frobenius diameter of X_R is at most $2R\sqrt{d_R \min\{d_L, d_R\}}$. Euclidean projected acceleration with $T-1$ gradients (Beck & Teboulle, 2009), whose extrapolated queries may leave X_R as the model permits and whose output is feasible, has error at most $2(L_{\mathcal{H}}/d_R)\,\mathrm{diam}_F(X_R)^2/T^2$. Finally, $\Lambda_{\mathcal{H}} \le kL$ on $\mathcal{F}^{\mathrm{Sh}}_{L,k}$, and Theorem 5.1 holds in every well-structured geometry, so the accelerated upper bound in the proof of Theorem 5.2 gives $144kLR^2/T^2$. □

Proposition E.5 gives on $\mathcal{F}^{\mathrm{Sh}}_{L,k}$, for $T \le \min\{d_L, d_R\}$, the sandwich of Theorem 5.2: up to one logarithm the risk is LR^2/T before the horizon k and kLR^2/T^2 after it. In this geometry the horizon is also capped by $\min\{d_L, d_R\}$, beyond which Euclidean acceleration gives $8\min\{d_L, d_R\}LR^2/T^2$ whatever k is. On $\mathcal{R}^{\mathrm{Sh}}_{L,k}$ the accelerated bound can lose up to the factor $K(\mathcal{H}) = d_L$ of Theorem 3.1(a): the quadratic $\frac{L}{2d_R}\|Xe_1\|_2^2$ has $\varrho_{\mathcal{H}} = 1$, $L_{\mathcal{H}} = L$ and $\Lambda_{\mathcal{H}} = d_L L$.

F SMOOTH HARD INSTANCES WITH QUANTITATIVELY LARGE ROTATION

We give an explicit version of the smoothed-max construction of Guzmán & Nemirovski (2015). Their smoothing argument uses a kernel with an exponent slightly larger than two to ensure that the kernel itself is C^2 at the origin. Here the kernel is exactly $2\|h\|_r^2$. We prove directly that the resulting smoothed maximum is C^2, and establish its locality and rotation without a limiting argument.

F.1 REGULARITY, SMOOTHNESS AND LOCALITY OF THE EXACT DUAL

Lemma F.1. Let $n \geq 1$, $r > 2$, $q = r/(r-1)$ and $\epsilon > 0$. Define

$$F_\epsilon(z) = \max_{\lambda \in \Delta_n} \left\{ \langle z, \lambda \rangle - \frac{\epsilon}{8}\|\lambda\|_q^2 \right\}. \tag{32}$$

Then F_ϵ is convex and C^2 on $\mathbb{R}^n$, with

$$L_\infty(F_\epsilon) \leq \frac{4(r-1)n^{2/r}}{\epsilon}. \tag{33}$$

If $z_j < \max_i z_i - \epsilon/4$, its maximizing vector has $\lambda_j = 0$. Consequently, removing any collection of terms satisfying this strict inequality leaves F_ϵ unchanged on a neighbourhood of z; the exponent q is kept fixed when terms are removed.

Proof. Step 1: solve the optimality conditions. Write $c = \epsilon/8$. Strict convexity of $\|\lambda\|_q^2$ gives a unique maximizer, and compactness of Δ_n gives its existence. Let μ be the multiplier of $\sum_i \lambda_i = 1$. With $S = \sum_i \lambda_i^q$, the optimality conditions are

$$z_i - \mu = 2cS^{2/q-1}\lambda_i^{q-1} \quad \text{if } \lambda_i > 0, \qquad z_i \leq \mu \quad \text{if } \lambda_i = 0. \tag{34}$$

Put $s_i = (z_i - \mu)_+$, $Z = \sum_i s_i^{r-1}$ and $W = \sum_i s_i^r$. Then

$$\lambda_i = \frac{s_i^{r-1}}{Z}, \qquad \frac{Z}{W^{(r-2)/r}} = 2c, \qquad \mu < \max_i z_i. \tag{35}$$

Conversely, a solution of (35) satisfies (34), so it produces the maximizer.

Step 2: regularity across active-set boundaries. For fixed z, the left-hand side of the scalar equation in (35) decreases continuously from $+\infty$ to 0 as μ increases from $-\infty$ to $\max_i z_i$. To verify strict decrease, set $A = \sum_i s_i^{r-2}$. Cauchy–Schwarz gives $Z^2 \leq AW$, and hence

$$\frac{\partial}{\partial \mu} \ln\left(\frac{Z}{W^{(r-2)/r}}\right) = -(r-1)\frac{A}{Z} + (r-2)\frac{Z}{W} \leq -\frac{A}{Z} < 0.$$

Thus there is a unique μ. Since $r > 2$, the maps $s \mapsto (s_+)^{r-1}$ and $s \mapsto (s_+)^r$ are C^1, including at zero. At every solution, $Z, W > 0$ and the derivative in μ is nonzero. The implicit-function theorem therefore makes $\mu(z)$ a C^1 function locally at every z, and uniqueness joins these local functions globally. Equation (35) makes $\lambda(z)$ C^1. The envelope formula $\nabla F_\epsilon(z) = \lambda(z)$ proves that F_ϵ is C^2. For $n = 1$ the same conclusion is also immediate from $F_\epsilon(z) = z - \epsilon/8$.

Step 3: the global smoothness bound. Let $\phi(h) = 2\|h\|_r^2$. Its convex conjugate is $\phi^*(w) = \|w\|_q^2/8$, as follows by optimizing first over direction using Hölder and then over the scalar norm. Convex duality gives

$$F_\epsilon(z) = \min_{h \in \mathbb{R}^n} \left\{ \max_i (z_i + h_i) + \epsilon\phi(h/\epsilon) \right\}. \tag{36}$$

The minimum exists because the quadratic kernel is coercive; there is no duality gap because both primal summands are finite and continuous. At its minimizer h, the saddle-point conditions give $-\nabla\phi(h/\epsilon) = \lambda(z)$.

For $h \neq 0$, direct differentiation and Hölder yield

$$\begin{aligned} v^\top \nabla^2 \phi(h) v &\leq 4(r-1)\|h\|_r^{2-r} \sum_i |h_i|^{r-2} v_i^2 \\ &\leq 4(r-1)\|v\|_r^2 \leq M_n \|v\|_\infty^2, \qquad M_n = 4(r-1)n^{2/r}. \end{aligned}$$

The term omitted in the first inequality is negative semidefinite because $r > 2$. Although ϕ need not be C^2 at zero, its gradient is continuous there. Integrating the displayed bound along a line segment, splitting at the origin if needed, proves the global upper-quadratic inequality with coefficient M_n.

Use $h - v$ as a candidate in (36) at $z + v$. The max term remains unchanged, so

$$F_\epsilon(z+v) \le F_\epsilon(z) + \langle \lambda(z), v\rangle + \frac{M_n}{2\epsilon}\|v\|_\infty^2.$$

Since F_ϵ is convex and C^2, its Hessian is positive semidefinite and satisfies $v^\top \nabla^2 F_\epsilon(z) v \le (M_n/\epsilon)\|v\|_\infty^2$. For a positive semidefinite matrix A, Cauchy–Schwarz in the A-seminorm shows $\|A\|_{\infty\to 1} = \max_{\|v\|_\infty \le 1} v^\top A v$. This proves (33).

Step 4: locality as equality of germs. Because $\|\lambda\|_q \le \|\lambda\|_1 = 1$, equation (34) gives

$$0 \le z_i - \mu = \frac{\epsilon}{4}\|\lambda\|_q^{2-q}\lambda_i^{q-1} \le \frac{\epsilon}{4} \quad \text{for every active coordinate.}$$

A largest coordinate is active, so $\max_i z_i - \mu \le \epsilon/4$. Therefore $z_j < \max_i z_i - \epsilon/4$ implies $z_j < \mu$, and hence $\lambda_j = 0$. A strict gap persists on a neighbourhood of z. On that neighbourhood the full maximizer is feasible for the simplex with those coordinates removed, and the two optimal values coincide. Thus the two functions have the same germ, not merely the same value and gradient at z. □

F.2 The exponent is a rotation dial

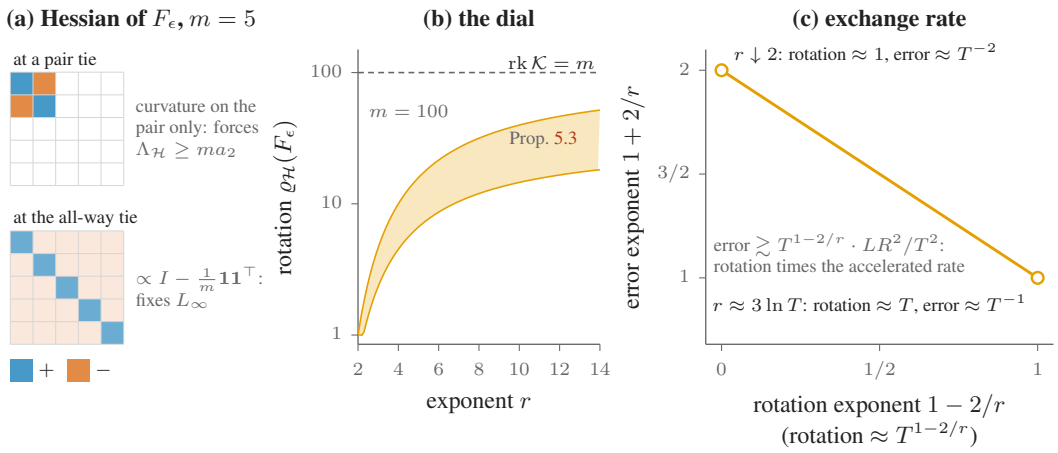

Figure 14: **The smoothed maximum as a rotation dial** (Proposition 5.3). **(a)** For $m = 5$: at a tie of two coordinates the Hessian is $\frac{a_2}{2}(e_i - e_j)(e_i - e_j)^\top$ with $a_2 = 4(r-1)2^{2/r-1}/\epsilon$; it lives on that pair, so every common diagonal majorant pays for every coordinate, $\Lambda_{\mathcal{H}} \ge ma_2$. At the all-way tie it is $\frac{4(r-1)}{\epsilon} m^{2/r-1}(I - \frac{1}{m}\mathbf{1}\mathbf{1}^\top)$, which fixes L_∞ (Lemma F.2). Each matrix is normalized to largest entry 1. **(b)** For $m = 100$ the exponent moves the rotation between $\max\{1, (2^{2/r}/\pi)m^{1-2/r}\}$ and $m^{1-2/r}/(1-m^{-2})$, always below $\mathrm{rk}\,\mathcal{K} = m$. **(c)** The resisting-oracle error is at least $T^{1-2/r} \cdot LR^2/(32(r-1)T^2)$ (Proposition F.3). Each unit of rotation exponent costs one unit of error exponent: from rotation ≈ 1 and error $\approx T^{-2}$ ($r \downarrow 2$) to rotation $\approx T$ and error $\approx T^{-1}$ up to a logarithm ($r \approx 3 \ln T$). Tuning $T^{1-2/r} \approx \min\{k, T\}$ gives the lower bound of Theorem 5.2.

Lemma F.2 (curvature of the smoothed maximum)**.** Let $m \ge 2$, $r > 2$, $q = r/(r-1)$, $\epsilon > 0$, and let F_ϵ be (32) with $n = m$. In the diagonal geometry of $\mathbb{R}^m$:

(a) $\nabla^2 F_\epsilon(z) \preceq \frac{4(r-1)}{\epsilon} I$ for every z; hence $\Lambda_{\mathcal{D}_m^+}(F_\epsilon) \le 4(r-1)m/\epsilon$.

(b) $\nabla^2 F_\epsilon(0) = \frac{4(r-1)}{\epsilon}\, m^{2/r-1}\big(I - \frac{1}{m}\mathbf{1}\mathbf{1}^\top\big)$; hence $L_\infty(F_\epsilon) \ge \frac{4(r-1)}{\epsilon}\, m^{2/r}\big(1 - m^{-2}\big)$.

(c) Consequently $\Lambda_{\mathcal{D}_m^+}(F_\epsilon) \le \dfrac{m^{1-2/r}}{1-m^{-2}}\, L_\infty(F_\epsilon)$.

Proof. (a) In Step 3 of the proof of Lemma F.1, the negative semidefinite term was dropped from $\nabla^2\phi(h)$, leaving $\nabla^2\phi(h) \preceq 4(r-1)\|h\|_r^{2-r}\,\mathrm{diag}(|h_i|^{r-2}) \preceq 4(r-1)I$, because $|h_i| \le \|h\|_r$. The same candidate $h-v$ as there gives $F_\epsilon(z+v) \le F_\epsilon(z) + \langle \lambda(z), v\rangle + \frac{2(r-1)}{\epsilon}\|v\|_2^2$, and for the convex C^2 function F_ϵ this is the claimed Hessian bound. Taking $H = \frac{4(r-1)}{\epsilon} I$ in the definition of Λ gives the trace bound.

(b) Near $z = 0$ all coordinates are active and, with $\psi(\lambda) = \frac{\epsilon}{8}\|\lambda\|_q^2$, the maximizer satisfies $z - \mu\mathbf{1} = \nabla\psi(\lambda)$ and $\mathbf{1}^\top\lambda = 1$ (equation (34)). At $z = 0$, symmetry gives $\lambda = \mathbf{1}/m$, where $\nabla^2\psi(\lambda) = aI + b\mathbf{1}\mathbf{1}^\top$ with $a = \frac{\epsilon}{4}(q-1)m^{(2-q)/q}$ and some b. Differentiating, $dz - d\mu\,\mathbf{1} = (aI + b\mathbf{1}\mathbf{1}^\top)d\lambda$ with $\mathbf{1}^\top d\lambda = 0$, so $d\lambda = a^{-1}(I - \frac{1}{m}\mathbf{1}\mathbf{1}^\top)dz$. Since $\nabla F_\epsilon = \lambda$ and $q - 1 = 1/(r-1)$, $(2-q)/q = 1 - 2/r$, this is the stated Hessian. For the lower bound on L_∞, evaluate $u^\top\nabla^2 F_\epsilon(0)u$ at a sign vector with $|\mathbf{1}^\top u| \le 1$, for which $u^\top(I - \frac{1}{m}\mathbf{1}\mathbf{1}^\top)u \ge m - \frac{1}{m}$, and use the PSD formula of Lemma C.1.

(c) Divide (a) by (b). □

Proof of the geometric statement in Proposition 5.3. Set $n = m$. Regularity follows from Lemma F.1, and the upper bound on $\Lambda_{\mathcal{H}}/L_\infty$ is Lemma F.2(c); since $L_\infty \le \Lambda^\circ_{\mathcal{H}}$, it also bounds $\varrho_{\mathcal{H}}$. It remains to prove the lower bound on $\varrho_{\mathcal{H}}$. We use the diagonal geometry in dimension m throughout this proof.

Step 1: realize all two-coordinate Hessians. Fix $i \ne j$. Choose $z_i = z_j$ and the other coordinates sufficiently smaller that only i, j are active on a neighbourhood. Symmetry gives $\lambda_i = \lambda_j = 1/2$. On this simplex edge, the Hessian of $c\|\lambda\|_q^2$ in the tangent direction has eigenvalue $2c(q-1)2^{2/q-1}$. Indeed, its Hessian is

$$2c(q-1)S^{2/q-1}\,\mathrm{diag}(\lambda_i^{q-2}) + 2c(2-q)S^{2/q-2}(\lambda_i^{q-1})_i(\lambda_i^{q-1})_i^\top,$$

and at the uniform two-coordinate vector the rank-one term vanishes on the edge's tangent space. Differentiating the edge optimality condition therefore gives

$$\nabla^2 F_\epsilon(z) = \frac{a_2}{2}(e_i - e_j)(e_i - e_j)^\top, \qquad a_2 = \frac{4(r-1)}{\epsilon}\,2^{-1+2/r}. \tag{37}$$

The other Hessian entries vanish because the function is locally independent of the inactive coordinates.

Step 2: every uniform majorant pays for every coordinate. If $\mathrm{diag}(h)$ majorizes the Hessian everywhere, testing (37) on $e_i - e_j$ gives $h_i + h_j \ge 2a_2$. Summing over all unordered pairs yields $(m-1)\sum_i h_i \ge m(m-1)a_2$, so

$$\Lambda_{\mathcal{D}_m^+}(F_\epsilon) \ge m a_2. \tag{38}$$

Step 3: divide by pointwise smoothness. The convex diagonal case of Theorem 3.1 and Lemma F.1 give

$$\Lambda^\circ_{\mathcal{D}_m^+}(F_\epsilon) \le \frac{\pi}{2}L_\infty(F_\epsilon) \le \frac{\pi}{2}\frac{4(r-1)m^{2/r}}{\epsilon}.$$

Combining this with (38) proves

$$\varrho_{\mathcal{H}}(F_\epsilon) \ge \frac{m}{\pi}\left(\frac{2}{m}\right)^{2/r} = \frac{2^{2/r}}{\pi}\,m^{1-2/r}.$$

Translation of the input, signed coordinate permutations, and positive scaling of the function preserve $\varrho_{\mathcal{H}}$, $\Lambda_{\mathcal{H}}/L_\infty$ and $\Lambda^\circ_{\mathcal{H}}/L_\infty$; lifting to inactive coordinates preserves them by Lemma E.2. □

F.3 A SINGLE FINAL FUNCTION CONSISTENT WITH EVERY ORACLE ANSWER

Proposition F.3 (an explicit C^2 resisting-oracle witness)**.** Let $L, R > 0$, $3 \le m \le d$, $r > 2$, and let a deterministic local-oracle method output its mth query $x_m \in X_R$, based on the first $m-1$

answers. There is a convex C^2 function f such that

$$L_\infty(f) \le L, \qquad \Lambda^+_{\mathcal{D}_d}(f) \le \frac{m^{1-2/r}}{1-m^{-2}}\, L_\infty(f), \qquad \varrho_{\mathcal{H}}(f) \ge \frac{2^{2/r}}{\pi}\, m^{1-2/r},$$
$$f(x_m) - \min_{X_R} f \ge \frac{LR^2}{32(r-1)\, m^{1+2/r}}.$$

It is a scaled, translated, signed-coordinate copy of (32) with exponent r, supported on m coordinates. For $r = 3\ln m$ the error bound is at least $LR^2/(96e^{2/3} m \ln m)$.

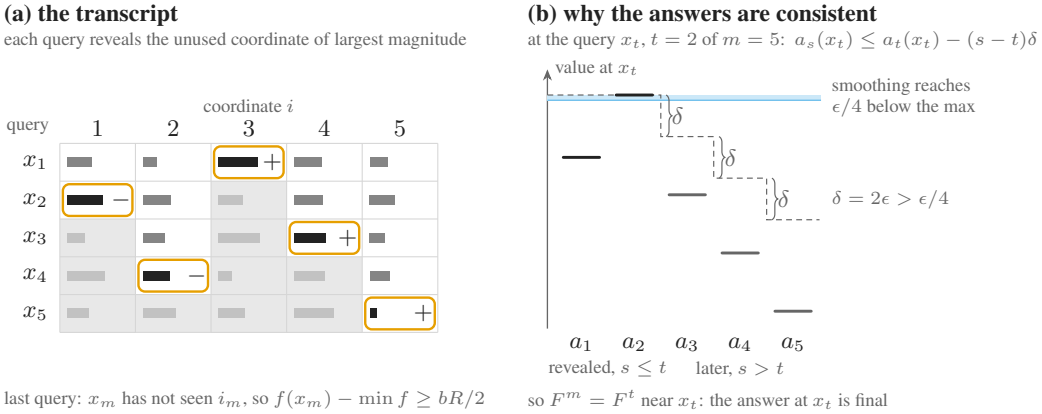

Figure 15: **The resisting oracle behind Proposition F.3** ($m = 5$, illustrative magnitudes, $R = 1$, so $\delta = 0.1$ and $\epsilon = 0.05$). **(a)** At each query the adversary commits the unused coordinate of largest magnitude (outlined), with the sign σ_t that makes the new affine piece $\sigma_t x_{i_t} - (t-1)\delta$ largest at x_t. Grey cells are coordinates already used; in rows 3–5 the largest entry is one of them. **(b)** At x_2 in the same transcript, every later piece lies below the staircase $a_t(x_t) - (s-t)\delta$, hence at least $\delta = 2\epsilon$ below an existing piece, eight times the smoothing width $\epsilon/4$ (band to scale). So the smoothed maximum agrees with its prefix near x_t, and every answer is consistent with the final function, even for an oracle that returns the whole germ.

Proof. Step 1: fix all numerical parameters before seeing the queries. With the given r, set

$$q = \frac{r}{r-1}, \quad \delta = \frac{R}{2m}, \quad \epsilon = \frac{R}{4m}, \quad M = 4(r-1)m^{2/r}, \quad b = \frac{L\epsilon}{M}.$$

The exponent q and scale b stay fixed throughout the construction, including for prefixes with fewer than m terms.

Step 2: construct a deterministic transcript. At query x_t, select, among the unused coordinates in $\{1,\dots,m\}$, an index i_t maximizing $|x_{t,i}|$. Let $\sigma_t \in \{-1,1\}$ satisfy $\sigma_t x_{t,i_t} = |x_{t,i_t}|$, taking $\sigma_t = 1$ at zero. Define

$$a_s(x) = \sigma_s x_{i_s} - (s-1)\delta, \qquad F^t(x) = \max_{\lambda\in\Delta_t} \left\{ \sum_{s=1}^{t} \lambda_s a_s(x) - \frac{\epsilon}{8}\|\lambda\|_q^2 \right\}.$$

Answer using the given local oracle applied to bF^t at x_t. These prefix functions are convex and C^2 by Lemma F.1; their smoothness is at most $bM/\epsilon = L$. Continue recursively through $t = m$ and freeze $f = bF^m$.

Step 3: verify consistency with the one final function. For every $s > t$, coordinate i_s was available when i_t was selected, and therefore

$$\begin{aligned} a_s(x_t) &\le |x_{t,i_s}| - (s-1)\delta \\ &\le |x_{t,i_t}| - (s-1)\delta = a_t(x_t) - (s-t)\delta. \end{aligned}$$

Since $\delta = 2\epsilon > \epsilon/4$, every later affine term is strictly more than $\epsilon/4$ below an existing term at x_t. This remains true on a neighbourhood of x_t. The locality part of Lemma F.1 implies $F^m = F^t$ on

that neighbourhood. Thus all oracle answers in the transcript agree with those of the single final f, even for an oracle returning the full germ. This is the point where locality, rather than agreement of gradients alone, is needed.

Step 4: the box optimum and the final error. Choose $x^* \in X_R$ with $x^*_{i_s} = -R\sigma_s$ for every s. This simultaneously minimizes all affine terms, so by their coordinatewise monotonicity in the dual maximum it minimizes F^m over X_R. At this point only the first term is active, since subsequent terms are separated from it by at least $\delta > \epsilon/4$. Consequently,

$$F^m(x^*) = -R - \epsilon/8.$$

At x_m, taking $\lambda = e_m$ in the dual and using $\sigma_m x_{m,i_m} \geq 0$ gives

$$F^m(x_m) \geq -(m-1)\delta - \epsilon/8.$$

Their difference is at least $R - (m-1)\delta \geq R/2$. Multiplying by b yields

$$f(x_m) - f(x^*) \geq \frac{L\epsilon R}{2M} = \frac{LR^2}{32(r-1)\, m^{1+2/r}},$$

which is at least $LR^2/(96e^{2/3} m \ln m)$ when $r = 3\ln m$, because then $m^{2/r} = e^{2/3}$ and $r - 1 \leq 3\ln m$.

Step 5: check membership and actual rotation. The function F^m is F_ϵ composed with a signed permutation of m coordinates and a translation. Lemma F.1 and the choice of b give $L_\infty(f) \leq L$, and Lemma F.2(c) together with the invariances above gives the bound on $\Lambda_{\mathcal{D}^+_d}(f)/L_\infty(f)$ and the lower bound on $\varrho_{\mathcal{H}}(f)$. All properties refer to the same final function that generated the transcript. □

What the construction does and does not say. The exponent r is fixed by the chosen horizon and rotation budget, and the signed coordinates and offsets produce one final objective after its interaction with the method has been specified. Its rotation is $\Theta(m^{1-2/r})$, and its error is $m^{1-2/r}/(32(r-1))$ times LR^2/m^2: the exponent trades non-accelerated hardness ($r \approx 3\ln m$) for accelerated hardness ($r \downarrow 2$) at exactly the exchange rate of the rotation. The result quantifies a worst-case family; it does not assert that one fixed function is hard for every method or that every high-rotation function exhibits a delayed improvement. For a model that counts N answered queries and then permits a separate output, the same construction uses $m = N + 1$ coordinates.

G THE SAME THREE LEVELS FOR GRADIENT NOISE

For stochastic gradients $\nabla f_t(x) = \nabla f(x) + \xi_t(x)$ with $\mathbb{E}\xi_t(x) = 0$ and covariance $\Sigma_t(x)$, Xie et al. (2026, Definition 4.1) compare a standard and an adaptive variance, and show that NSD with momentum has a dimension-free nonconvex rate in terms of the adaptive one but, by a lower bound, not in terms of the standard one. Both are costs of a noise-whitening preconditioner, and they differ in what the preconditioner may depend on. Using $\|\xi\|^2_{\mathcal{H},*} = \inf_H \|\xi\|^2_{H^{-1}}$ (Lemma B.2), where H ranges over $\mathcal{H}$ with $H \succ 0$ and $\operatorname{Tr} H \leq 1$,

$$\sigma^2_{\mathrm{std}} = \sup_{t,x} \mathbb{E}\Big[\inf_H \|\xi_t(x)\|^2_{H^{-1}}\Big], \quad \sigma^{\circ 2} = \sup_{t,x} \inf_H \mathbb{E}\|\xi_t(x)\|^2_{H^{-1}}, \quad \sigma^2_{\mathcal{H}} = \inf_H \sup_{t,x} \mathbb{E}\|\xi_t(x)\|^2_{H^{-1}}. \tag{39}$$

The preconditioner may depend on the noise realization, on the point, or on nothing. The middle quantity is new; we call $\gamma_{\mathcal{H}} := \sigma^{\circ 2}/\sigma^2_{\mathrm{std}}$ the pointwise gap and $\varrho^\sigma_{\mathcal{H}} := \sigma^2_{\mathcal{H}}/\sigma^{\circ 2}$ the noise rotation number.

Theorem G.1 (noise)**.** Let $\mathcal{H}$ be well-structured and assume $\sigma^2_{\mathrm{std}} < \infty$; set $\gamma_{\mathcal{H}} := \varrho^\sigma_{\mathcal{H}} := 1$ if $\sigma_{\mathrm{std}} = 0$.
(a) $\sigma^2_{\mathrm{std}} \leq \sigma^{\circ 2} \leq \sigma^2_{\mathcal{H}} \leq \operatorname{rk}\mathcal{K} \cdot \sigma^2_{\mathrm{std}}$, so $\gamma_{\mathcal{H}}, \varrho^\sigma_{\mathcal{H}} \in [1, \operatorname{rk}\mathcal{K}]$. Moreover $\sigma^{\circ 2} = \sup_{t,x} \big(\operatorname{Tr} \pi_{\mathcal{K}}(\Sigma_t(x))^{1/2}\big)^2$, and

$$\sigma^2_{\mathcal{H}} = \sup\Big\{\big(\operatorname{Tr}\, \pi_{\mathcal{K}}(\Sigma)^{1/2}\big)^2 : \Sigma \in \overline{\operatorname{conv}}\{\Sigma_t(x)\}_{t,x}\Big\}. \tag{40}$$

(b) For Gaussian noise, $\gamma_{\mathcal{H}} \le \frac{\pi}{2} b_{\max}$. For one simple component $\mathbb{M}_b(\mathbb{F})$, $\gamma_{\mathcal{H}} \le b$ for arbitrary noise. The constant $\pi/2$ is asymptotically sharp: in diagonal geometry with $\Sigma = I_d$, $\gamma_{\mathcal{H}} = d^2/(d + d(d-1)\frac{2}{\pi}) \to \frac{\pi}{2}$. The single-component bound is attained by Gaussian noise on one irreducible copy.
(c) In diagonal geometry, sparse noise $\xi = \sqrt{d}\,\epsilon\, e_I$ with I uniform on $[d]$ and ϵ a Rademacher sign has $\gamma_{\mathcal{H}} = d = \operatorname{rk}\mathcal{K}$.
(d) In diagonal geometry, the losses $f_t(x) = f(x) + \sqrt{d}\, g_t\, \operatorname{lse}_\tau(x)$ with $g_t \sim N(0,1)$ have Gaussian noise $\sqrt{d}\, g_t \operatorname{softmax}(x/\tau)$ of covariance $d\, pp^\top$ with $p = \operatorname{softmax}(x/\tau)$, and $\sigma^2_{\mathrm{std}} = \sigma^{\circ 2} = d$, $\sigma^2_{\mathcal{H}} = d^2$: $\gamma_{\mathcal{H}} = 1$ and $\varrho^\sigma_{\mathcal{H}} = d$.

Formula (40) is the noise form of commitment: a fixed preconditioner must whiten the worst *mixture* of the noise distributions it will meet, and the concave functional $\Sigma \mapsto \operatorname{Tr} \pi_{\mathcal{K}}(\Sigma)^{1/2}$ can be much larger on mixtures than on each component. The Gaussian bound in (b) is the same rounding as in Theorem 3.1: $(\mathbb{E}\|\zeta\|_2)^2 \ge \frac{2}{\pi} \operatorname{Tr} \operatorname{Cov}(\zeta)$ for a Gaussian vector, and $\pi/2 = 1/(\mathbb{E}|g|)^2$. Unlike curvature, noise can exceed the Grothendieck bound at a single point, because the realization-dependent preconditioner sees *which* sparse coordinate fired; for curvature the matrix is deterministic, only second moments of the probe matter, and Theorem 3.1 applies.

Corollary G.2. If the noise is Gaussian with a covariance Σ independent of (t, x), then $\sigma^2_{\mathcal{H}} \le \frac{\pi}{2} b_{\max}\, \sigma^2_{\mathrm{std}}$, and $\sigma^2_{\mathcal{H}} \le \frac{\pi}{2} \sigma^2_{\mathrm{std}}$ in diagonal geometry. Hence, under the sampling and regularity assumptions of Theorem A.14 (Xie et al., 2026, Theorem 4.5), the dimension-free rate of NSD with momentum holds with σ_{std} in place of $\sigma_{\mathcal{H}}$, up to the factor $\sqrt{\pi b_{\max}/2}$, which is $\sqrt{\pi/2} < 1.26$ in diagonal geometry. In diagonal geometry, and more generally whenever $b_{\max}$ stays bounded, a dimension-growing separation with (t, x)-independent noise therefore requires non-Gaussian noise; the Bernoulli construction behind Theorem 4.7 of Xie et al. (2026) has $\gamma_{\mathcal{H}} \ge (d-1)/5$ (Appendix H), the mechanism of (c).

H NOISE

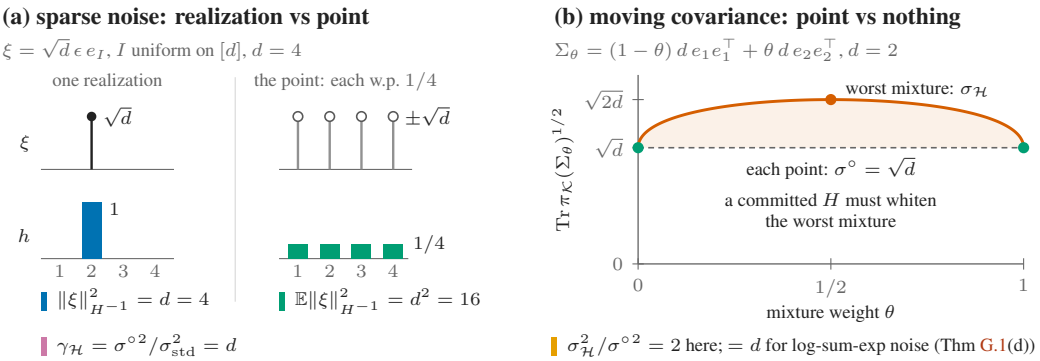

Figure 16: **The three noise levels** (Theorem G.1). **(a)** Sparse noise $\xi = \sqrt{d}\,\epsilon e_I$ with $d = 4$. A preconditioner that sees the realization puts its whole budget on the coordinate that fired (blue; a limit of admissible H) and pays $\|\xi\|^2_{H^{-1}} = d$, so $\sigma^2_{\mathrm{std}} = d$. One that sees only the point must spread its budget (green) and pays $\mathbb{E}\|\xi\|^2_{H^{-1}} = d^2$, so the pointwise gap is $\gamma_{\mathcal{H}} = d = \operatorname{rk}\mathcal{K}$. **(b)** A covariance that moves with the point, $\Sigma_\theta = (1-\theta)\, d\, e_1 e_1^\top + \theta\, d\, e_2 e_2^\top$ ($d = 2$). The functional $\Sigma \mapsto \operatorname{Tr} \pi_{\mathcal{K}}(\Sigma)^{1/2}$ is concave, so its value at the worst mixture ($\sqrt{2d}$ at $\theta = \frac{1}{2}$) exceeds its value at each point ($\sqrt{d}$). A committed preconditioner must whiten the worst mixture and pays $\sigma^2_{\mathcal{H}} = 2d$ against $\sigma^{\circ 2} = d$; for the log-sum-exp noise of Theorem G.1(d) the same mechanism gives a factor d.

Lemma H.1 (Gaussian norm lower bound)**.** If $\zeta \sim N(0, \Sigma)$ in $\mathbb{R}^m$ then $\mathbb{E}\|\zeta\|_2 \ge \sqrt{2/\pi}\, \sqrt{\operatorname{Tr}\Sigma}$.

Proof. Step 1: reduction to eigenvalues. Diagonalizing Σ, $\|\zeta\|_2 \overset{d}{=} \left(\sum_{j=1}^m \lambda_j g_j^2\right)^{1/2}$ with $\lambda_j \ge 0$ the eigenvalues of Σ and g_j i.i.d. standard normal.

Step 2: concavity and symmetry. The function $\phi(\lambda) := \mathbb{E}\big(\sum_j \lambda_j g_j^2\big)^{1/2}$ on $\mathbb{R}_+^m$ is concave (the square root of a linear function of λ is concave, and expectations preserve concavity) and invariant under permutations of the coordinates of λ.

Step 3: minimization over the simplex. On the simplex $\{\lambda \geq 0 : \sum_j \lambda_j = s\}$, a concave function attains its minimum at a vertex; at the vertex $\lambda = se_1$ one has $\phi = \sqrt{s}\,\mathbb{E}|g_1| = \sqrt{2s/\pi}$ (Lemma A.15). With $s = \operatorname{Tr}\Sigma$ this is the claim. □

Proof of Theorem G.1. Throughout, $\xi = \xi_t(x)$ has mean zero and covariance $\Sigma = \Sigma_t(x)$, suprema are over (t, x), and H ranges over $\mathcal{P} := \{H \in \mathcal{H} : H \succ 0,\ \operatorname{Tr} H \leq 1\}$. (Definition 4.1 of Xie et al. (2026) takes the minimum over all $H \in \mathcal{H}$ with $\operatorname{Tr} H \leq 1$, including singular H with the convention $\|\xi\|_{H^{-1}} = +\infty$ off the range of H; since $H_\delta = (1-\delta)H + \delta I/d$ is positive definite with $\|\xi\|^2_{H_\delta^{-1}} \leq \|\xi\|^2_{H^{-1}}/(1-\delta)$, the two conventions give the same infimum.) The expressions in (39) agree with the definitions of Xie et al. (2026) because $\|\xi\|^2_{\mathcal{H},*} = \inf_{H\in\mathcal{P}} \|\xi\|^2_{H^{-1}}$ (Lemma B.2(c)).

(a) The chain of inequalities.

Step 1: $\sigma^2_{\mathrm{std}} \leq \sigma^{\circ 2}$. An infimum of linear functionals is concave, so

$$\mathbb{E}\|\xi\|^2_{\mathcal{H},*} = \mathbb{E}\inf_H \xi^\top H^{-1}\xi \;\leq\; \inf_H \mathbb{E}\,\xi^\top H^{-1}\xi = \inf_H \langle \Sigma, H^{-1}\rangle;$$

take suprema over (t, x).

Step 2: $\sigma^{\circ 2} \leq \sigma^2_{\mathcal{H}}$ is $\sup\inf \leq \inf\sup$.

Step 3: $\sigma^2_{\mathcal{H}} \leq \operatorname{rk}\mathcal{K} \cdot \sigma^2_{\mathrm{std}}$. With the notation of Lemma C.4, put $H_0 := \sum_i P_i/(\operatorname{rk}\mathcal{K}\cdot r_i)$. It lies in $\mathcal{H}$, is positive definite, and $\operatorname{Tr} H_0 = \sum_i b_i r_i/(\operatorname{rk}\mathcal{K}\cdot r_i) = 1$. Moreover $\|\xi\|^2_{H_0^{-1}} = \operatorname{rk}\mathcal{K}\sum_i r_i\|P_i\xi\|_2^2 = \operatorname{rk}\mathcal{K}\,\|S\xi\|_2^2 \leq \operatorname{rk}\mathcal{K}\,\|\xi\|^2_{\mathcal{H},*}$ by Lemma C.4(a). Taking expectations and suprema, $\sigma^2_{\mathcal{H}} \leq \sup \mathbb{E}\|\xi\|^2_{H_0^{-1}} \leq \operatorname{rk}\mathcal{K}\cdot\sigma^2_{\mathrm{std}}$.

Step 4: the formula for $\sigma^{\circ 2}$ is (19). Under the assumption $\sigma^2_{\mathrm{std}} < \infty$, the chain makes all three quantities finite. If $\sigma_{\mathrm{std}} > 0$, division gives $\gamma_{\mathcal{H}}, \varrho^\sigma_{\mathcal{H}} \in [1, \operatorname{rk}\mathcal{K}]$; if $\sigma_{\mathrm{std}} = 0$, all three vanish and the stated convention applies.

Step 5: the mixture formula (40). Let $\mathcal{C}$ be the closed convex hull of $\{\Sigma_t(x)\}$. It is bounded, because $\operatorname{Tr}\Sigma = \mathbb{E}\|\xi\|_2^2 \leq \mathbb{E}\|\xi\|^2_{\mathcal{H},*} \leq \sigma^2_{\mathrm{std}}$ by Lemma B.2(b), hence compact. For fixed H, $\Sigma \mapsto \langle\Sigma, H^{-1}\rangle$ is linear and continuous, so $\sup_{t,x}\langle\Sigma_t(x), H^{-1}\rangle = \sup_{\Sigma\in\mathcal{C}}\langle\Sigma, H^{-1}\rangle$. For fixed Σ, $H \mapsto \langle\Sigma, H^{-1}\rangle$ is convex and continuous on the convex set $\mathcal{P}$. Sion's minimax theorem (Theorem A.4) therefore gives

$$\sigma^2_{\mathcal{H}} = \inf_{H\in\mathcal{P}}\sup_{\Sigma\in\mathcal{C}}\langle\Sigma, H^{-1}\rangle = \sup_{\Sigma\in\mathcal{C}}\inf_{H\in\mathcal{P}}\langle\Sigma, H^{-1}\rangle = \sup_{\Sigma\in\mathcal{C}}\big(\operatorname{Tr}\pi_{\mathcal{K}}(\Sigma)^{1/2}\big)^2,$$

the last step by (19).

(b) Gaussian noise and one simple component. Use the real block representation of Lemma B.5, with $q_i = \dim_{\mathbb{R}}\mathbb{F}_i$, $r_i = q_i n_i$, and $d_i = b_i r_i$. Write $\xi_i = P_i\xi$ and $\Sigma_{ii} = P_i\Sigma P_i$, considered as a covariance on V_i.

Step 1: one simple component, arbitrary mean-zero noise. Write $d = br$. The inclusion $\sqrt{r}B_2 \subseteq B_{\mathcal{H}}$ gives $\|\xi\|_{\mathcal{H},*} \geq \sqrt{r}\|\xi\|_2$, while Cauchy–Schwarz and trace preservation give

$$\big(\operatorname{Tr}\pi_{\mathcal{K}}(\Sigma)^{1/2}\big)^2 \leq d\operatorname{Tr}\pi_{\mathcal{K}}(\Sigma) = d\operatorname{Tr}\Sigma, \qquad \mathbb{E}\|\xi\|^2_{\mathcal{H},*} \geq r\operatorname{Tr}\Sigma. \tag{41}$$

Thus the pointwise adaptive variance is at most b times the standard variance at every (t, x), and taking suprema gives $\gamma_{\mathcal{H}} \leq b$.

This bound is attained by Gaussian noise supported on one irreducible copy $W \cong \mathbb{F}^b$, with covariance $\Sigma = E_W$. Indeed, Lemma B.5(c) gives $\pi_{\mathcal{K}}(E_W) = I/n$ and $\|\xi\|_{\mathcal{H},*} = \sqrt{r}\|\xi\|_2$ on W, so

$$\big(\operatorname{Tr}\pi_{\mathcal{K}}(E_W)^{1/2}\big)^2 = d^2/n, \qquad \mathbb{E}\|\xi\|^2_{\mathcal{H},*} = r\dim_{\mathbb{R}} W = rqb,$$

whose ratio is b because $r = qn$ and $d = bqn$. For a real split block, $E_W = I_b \otimes e_1 e_1^\top$.

Step 2: arbitrary components, Gaussian noise. The product-ball inclusion in Lemma B.5(b) implies

$$\|\xi\|_{\mathcal{H},*} \geq \sum_i \sqrt{r_i}\|\xi_i\|_2.$$

Every marginal ξ_i is Gaussian, so Lemma H.1 and Jensen yield

$$\mathbb{E}\|\xi\|_{\mathcal{H},*}^2 \geq \left(\sum_i \sqrt{r_i}\,\mathbb{E}\|\xi_i\|_2\right)^2 \geq \frac{2}{\pi}\left(\sum_i \sqrt{r_i}\sqrt{\operatorname{Tr}\Sigma_{ii}}\right)^2.$$

No independence among components is required. On the other hand, the bimodule and trace-preservation properties of $\pi_{\mathcal{K}}$ imply that $\pi_{\mathcal{K}}(\Sigma)$ is block diagonal and $\operatorname{Tr}(\pi_{\mathcal{K}}(\Sigma)|_{V_i}) = \operatorname{Tr}\Sigma_{ii}$. Therefore

$$\operatorname{Tr}\pi_{\mathcal{K}}(\Sigma)^{1/2} \leq \sum_i \sqrt{d_i}\sqrt{\operatorname{Tr}\Sigma_{ii}} \leq \sqrt{b_{\max}}\sum_i \sqrt{r_i}\sqrt{\operatorname{Tr}\Sigma_{ii}}.$$

Combining these inequalities gives the pointwise comparison with factor $(\pi/2)b_{\max}$; taking suprema proves $\gamma_{\mathcal{H}} \leq (\pi/2)b_{\max}$. The coefficient $\pi/2$ is asymptotically sharp for the diagonal algebra: with $\Sigma = I_d$, $\sigma^{\circ 2} = d^2$ and $\mathbb{E}\|\xi\|_1^2 = d + d(d-1)2/\pi$ (Lemma A.15).

(c) Sparse noise. $\Sigma = \mathbb{E}\xi\xi^\top = I_d$, so $\sigma^{\circ 2} = (\sum_i \sqrt{\Sigma_{ii}})^2 = d^2$, while $\|\xi\|_1 = \sqrt{d}$ almost surely, so $\sigma_{\mathrm{std}}^2 = d$.

(d) Rotating noise. Here $\nabla f_t(x) = \nabla f(x) + \sqrt{d}\,g_t\,p(x)$ with $p(x) = \operatorname{softmax}(x/\tau) \in \Delta^\circ$, so $\xi_t(x) = \sqrt{d}\,g_t p(x)$ is a centred Gaussian vector with covariance $\Sigma(x) = d\,p(x)p(x)^\top$.

Step 1: $\sigma_{\mathrm{std}}^2 = d$. $\|\xi_t(x)\|_1 = \sqrt{d}\,|g_t|\,\|p(x)\|_1 = \sqrt{d}\,|g_t|$.

Step 2: $\sigma^{\circ 2} = d$. Pointwise, $\big(\sum_i \sqrt{\Sigma_{ii}(x)}\big)^2 = d\big(\sum_i p_i(x)\big)^2 = d$. Hence $\gamma_{\mathcal{H}} = 1$.

Step 3: $\sigma_{\mathcal{H}}^2 = d^2$. For $H = \operatorname{diag}(h)$ with $h > 0$, $\sum_i h_i \leq 1$, $\langle\Sigma(x), H^{-1}\rangle = d\sum_i p_i(x)^2/h_i$ is a convex function of p, so its supremum over the simplex is attained at a vertex:

$$\sup_x \langle\Sigma(x), H^{-1}\rangle = d\max_k \frac{1}{h_k} \geq d^2,$$

because some $h_k \leq 1/d$; equality holds for $h = \mathbf{1}/d$. Hence $\sigma_{\mathcal{H}}^2 = d^2$ and $\varrho_{\mathcal{H}}^\sigma = d$. In the language of (40), the uniform mixture of the vertex covariances $d\,e_k e_k^\top$ is I_d, whose value d^2 exceeds the value d of every component. □

Proof of Corollary G.2. Step 1: no rotation. If $\Sigma_t(x) \equiv \Sigma$ then $\sigma_{\mathcal{H}}^2 = \inf_H\langle\Sigma, H^{-1}\rangle = \sigma^{\circ 2}$, so $\varrho_{\mathcal{H}}^\sigma = 1$ and $\sigma_{\mathcal{H}}^2 = \gamma_{\mathcal{H}}\sigma_{\mathrm{std}}^2 \leq \frac{\pi}{2}b_{\max}\sigma_{\mathrm{std}}^2$ by Theorem G.1(b); in diagonal geometry $b_{\max} = 1$.

Step 2: transfer to Theorem 4.5 of Xie et al. (2026). That bound (Theorem A.14) is increasing in $\sigma_{\mathcal{H}}$, so it remains valid with $\sigma_{\mathcal{H}}$ replaced by $\sqrt{\pi b_{\max}/2}\,\sigma_{\mathrm{std}}$; the tuned regimes stated there pick up at most this factor.

Step 3: the Bernoulli instance. In the construction of their Theorem 4.7, let $R_{t,j}$ be i.i.d. Bernoulli(δ) variables with $d\delta \leq 1$, and write the centred noise as $\xi_t = -C(R_t - \delta\mathbf{1})$. The raw Bernoulli vector has on average at most one nonzero coordinate. Its centred covariance is $\Sigma = C^2\delta(1-\delta)I$, so $\sigma^{\circ 2} = d^2C^2\delta(1-\delta)$, while their eq. (60) gives $\mathbb{E}\|\xi_t\|_1^2 \leq 5C^2d\delta$; hence

$$\gamma_{\mathcal{H}} \geq \frac{d^2C^2\delta(1-\delta)}{5C^2d\delta} = \frac{d(1-\delta)}{5} \geq \frac{d-1}{5}:$$

a sparse non-Gaussian noise with a pointwise gap linear in d, the same mechanism as in (c). □

I Reproducible numerical illustrations

The supplementary directory `experiments/` contains every script, configuration, raw array and hash used for Figure 4, together with the archived diagnostic run behind Figures 17 and 18. No hyperparameter search, run selection or curve fitting is performed. Semidefinite programs are solved with CVXPY and CLARABEL; solver feasibility is checked in floating point, not certified by interval arithmetic.

I.1 HESSIANS ALONG TRAINING (FIGURE 4)

Models. All computations use float64, full-batch losses, exact Hessians (PyTorch automatic differentiation), a ridge term $\frac{\lambda}{2}\|w\|_2^2$ with $\lambda = 10^{-3}$, and Adam with learning rate 0.05 and default moment parameters for 600 iterations. Features are standardized, projected by exact PCA, and scaled to unit mean-square row norm. *Binary logistic regression:* the breast-cancer data of scikit-learn, 10 principal components, $w_0 = 0$, $d = 10$. *Softmax regression:* the digit images of scikit-learn restricted to the classes $\{0, \dots, C-1\}$ for $C \in \{3, 5, 8\}$, 6 principal components, $W_0 = 0$, $d = 6C$. *MLP:* a one-hidden-layer $\tanh$ network $6 \to 6 \to 3$ with biases on the digit classes $\{0, 1, 2\}$, $d = 63$, weights initialized with the fixed seed 20260926 at scale $1/\sqrt{\text{fan-in}}$ and zero biases. Hessians are recorded at the 22 distinct iterations $\{0, 1, 2, 3, 4, 6, 8, 10, 14, \dots, 449, 600\}$ of a geometric grid.

Constants. In diagonal geometry, for every recorded Hessian A we compute $\nu(A) = \min\{\sum_i h_i : -\operatorname{diag}(h) \preceq A \preceq \operatorname{diag}(h)\}$, using only the upper constraint when all Hessians of the model are positive semidefinite (the four convex models), and then shift the solver's h by the amount needed to make the imposed constraints hold in floating point, so every reported ν and Λ is a feasible upper value. For $d \le 18$ we compute $\|A\|_{\infty\to 1}$ exactly by enumerating sign vectors. For larger d we report the best of 400 Goemans–Williamson roundings of the semidefinite relaxation followed by coordinate-flip local search (convex models), or of 400 random sign starts followed by alternating maximization (MLP); these are values of explicit sign vectors, hence lower bounds on $\|A\|_{\infty\to 1}$, and the plotted pointwise gaps are upper bounds. For the prefix S_t of recorded Hessians we compute $\Lambda(S_t)$, the least trace of a diagonal matrix dominating all of them, and $\varrho(S_t) = \Lambda(S_t)/\max_{A\in S_t} \nu(A)$.

What is and is not measured. $\varrho(S_t)$ is the rotation of the finite set of curvatures recorded at the checkpoints; recording more iterates can only enlarge $\Lambda(S_t)$ and may change the ratio slightly. It is neither a lower nor an upper bound on the global rotation $\varrho_{\mathcal{H}}(f)$: a subset can rotate less than the landscape, and it can also rotate more, because $\max_{A\in S_t} \nu(A)$ may miss the landscape's largest pointwise constant. It is the commitment cost that an accumulated preconditioner faced along this particular trajectory.

Results. At the last iteration the pointwise gap $\max_{A\in S_t} \nu(A)/\max_{A\in S_t} \|A\|_{\infty\to 1}$, the rotation $\varrho(S_t)$ and the total factor $\Lambda(S_t)/\max_{A\in S_t} \|A\|_{\infty\to 1}$ are, respectively: logistic regression 1.000, 1.000, 1.000 (PCA makes its dominating initial Hessian diagonal, so even its pointwise gap is one); softmax with $C = 3$: 1.098, 1.009, 1.108; $C = 5$: 1.039, 1.025, 1.065; $C = 8$: 1.033, 1.010, 1.043; MLP: 1.058, 1.082, 1.145. The largest pointwise gap of a single recorded Hessian is 1.123 (softmax, $C = 3$, at initialization). For $d > 18$ the standard norms are lower bounds, so the corresponding gaps and totals are upper bounds; for $d \le 18$ all values are exact up to solver tolerance.

I.2 RANDOM MATRICES (FIGURE 4A)

For $d = 8$ the archived diagnostic run uses four families of 12 matrices each: a symmetrized Gaussian matrix; $BB^\top$ for a Gaussian $B \in \mathbb{R}^{8\times 8}$; Gram matrices of eight random unit vectors in $\mathbb{R}^2$; and symmetric block matrices $\left(\begin{smallmatrix} 0 & C \\ C^\top & 0 \end{smallmatrix}\right)$ with $C_{ij} = \langle p_i, q_j\rangle$ for independent unit vectors $p_i, q_j \in \mathbb{R}^3$, $i, j \le 4$. All are normalized in Frobenius norm and generated in this order by one NumPy generator with seed 20260912. The standard norm is computed by enumerating the 2^7 sign vectors with $u_1 = 1$; the semidefinite program is solved with gap and feasibility tolerances 10^{-9} and repaired into feasible primal and dual points. The largest relative primal–dual interval was below $1.2 \cdot 10^{-8}$. The largest ratios were 1.233 for the symmetric families and 1.061 for the positive semidefinite families (rounded upward).

I.3 SUPPLEMENTARY DIAGNOSTICS FROM THE ARCHIVED RUN

Resisting-oracle protocol. We fix $d = 64$, $R = L = 1$ and $k \in \{4, 8, 16, 32, 64\}$, set $r_k = 3\ln k$, $\delta_k = R/(2k)$, $\epsilon_k = R/(4k)$, $M_k = 4(r_k - 1)k^{2/r_k}$ and $\beta_k = L\epsilon_k/M_k$, and run the construction of Proposition F.3 against projected sign descent $x_{t+1} = \operatorname{clip}_{[-R,R]^d}(x_t - R\operatorname{sign}(\nabla f^t(x_t)))$ from $x_1 = 0$, breaking ties by the smallest index. The dual maximization defining f^t is evaluated by

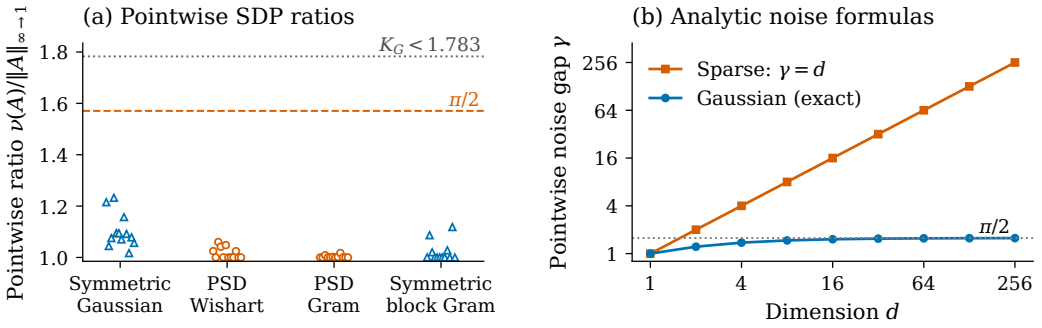

Figure 17: **Archived diagnostic. (a)** The random-matrix ratios of Figure 4a by family; markers are numerical feasible upper bounds, and the lines are the dimension-free bounds of Theorem 3.1. **(b)** Exact evaluations of the pointwise noise gap $\gamma_{\mathcal{H}}$ for isotropic Gaussian noise, $(\pi/2)/(1+(\pi/2-1)/d)$, and for sparse coordinate noise, d (Theorem G.1); these are analytic curves, not Monte Carlo estimates.

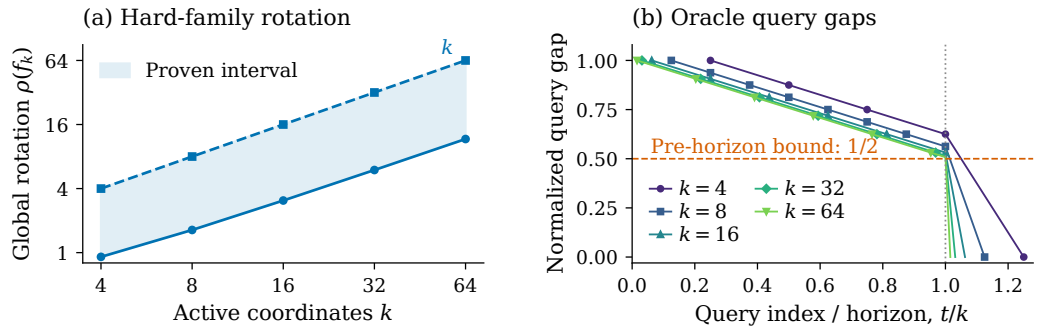

Figure 18: **The resisting oracle in action.** Ambient dimension $d = 64$, $k \in \{4, 8, 16, 32, 64\}$ active coordinates. **(a)** The proven lower bound of (14) with $r = 3 \ln k$ and the trivial upper bound k for the rotation of the smoothed-max family; it is not a measured rotation. **(b)** Query gaps of projected sign descent with fixed step R against the resisting oracle, replayed on the final function, normalized by $\beta_k R$ with $\beta_k = \epsilon_k / M_k$. Every query $t \le k$ stays above $1/2$; the kth update reaches a box minimizer, first shown as a query at $t = k + 1$.

a 100-step bisection for the multiplier of the simplex constraint; at 40 deterministic near-tie inputs with at least three active weights, the primal and dual values agreed to within $2 \cdot 10^{-16} \epsilon_k$. Along each trajectory the gradient has one active coordinate, so the query gaps equal $\beta_k[R-(t-1)\delta_k] \ge \beta_k R/2$ for $t \le k$; the normalized gaps at the horizon are 0.625, 0.5625, 0.53125, 0.515625 and 0.5078125. Replaying all historical queries on the final function reproduces every stored value and gradient exactly, so the evolving construction is one fixed function along the whole transcript. The sign method only exposes the information pattern; it is not an accelerated method.

J Grothendieck's rising sea and this paper

The quotation under the abstract comes from a passage of *Récoltes et Semailles* in which Grothendieck contrasts two ways of opening a nut, his image for a theorem to be proved (Grothendieck, 1985–1987, pp. 552–555). The first is the hammer and chisel: strike the shell at many points until it cracks. In the second, the *rising sea* ("la mer qui monte"), the problem is "submerged and dissolved by some more or less vast theory, going well beyond the results originally to be established" (p. 555; translations by McLarty, 2007). This paper tries to take the second road, and it meets Grothendieck twice on the way.

The resistant substance. The question we started from invites the hammer and chisel. The smoothness constant that governs Adam can exceed the one that governs sign descent by a factor of the dimension: is that factor real, and what is it made of? Attacked directly, the question compares two unrelated definitions, an operator norm $\|A\|_{\infty\to 1}$ that is NP-hard to compute and the least trace of a diagonal majorant, a semidefinite program over the whole landscape.

The sea. Instead of comparing the two constants, we look for the object of which both are readings. A random probe, a point together with a direction, reads the curvature $\mathbb{E}[u^\top \nabla^2 f(X)u]$, and the standard, pointwise and adaptive constants are its largest readings under a sure, a pointwise and a global budget (Theorem 2.1). The generality costs nothing: the statement holds in every well-structured geometry and, with two directions, for indefinite Hessians. It changes the question. The factor between the two constants is no longer a comparison of two definitions; it is the price of relaxing one budget, twice.

The problem dissolves. At this level each relaxation has a classical answer, and neither upper bound needs a new inequality. Relaxing the sure budget to a pointwise one is a semidefinite relaxation, priced by Grothendieck's inequality for the amplitudes of the probe and by the block size for its directions (Theorem 3.1); for convex objectives in diagonal geometry it is Grothendieck's theorem on 2-summing operators, with the budget as a Pietsch measure (Remark 2.2). Relaxing the pointwise budget to a global one asks one preconditioner to serve every point, and its price, the rotation number, is bounded by the rank of the algebra through Cauchy–Schwarz in a rescaled Euclidean metric (Theorem 4.1). The same budgets reappear in a minimax problem, where rotation becomes the acceleration horizon (Theorem 5.2), and in the variance of stochastic gradients (Theorem G.1). The accelerated method of Theorem 5.1 is analysed in the same spirit: its proof never chooses coordinates, and the unknown majorant enters only through the trace of the global budget.

Grothendieck twice. The inequality that prices the first relaxation is the result Grothendieck himself described as "the fundamental theorem in the metric theory of tensor products" (Grothendieck, 1953; Pisier, 2012); the method of the paragraphs above is his rising sea. The sharp constructions, the saturated functions of Theorem 4.1 and the smoothed maxima of Proposition 5.3, are hammer and chisel. So are the questions the sea has not yet reached: the exact Grothendieck constant of an algebra with several interacting blocks, the logarithm in the lower bound of Theorem 5.2, and a rate for Adam's own update rule (Section 7).