FINITE-TIME HESSIAN SKETCHING BELOW THE EFFECTIVE-DIMENSION THRESHOLD

Anonymous authors
Paper under double-blind review

ABSTRACT

Distributed Hessian sketching gives each worker only a small sketched matrix. Averaging the workers' Newton directions reduces random error, but not the bias that all small sketches share. Positive-shift debiasing requires the sketch size m to exceed the effective dimension $d_1(\lambda)$; we study memory-capped workers with smaller m. Below a stricter fold threshold $d_2(\lambda)$, no stable sketched inverse, with a positive or a negative shift, recovers the mean Newton direction once the gradient has components on two distinct Hessian eigenvalues, and adding workers cannot remove this error. We introduce *finite-time Hessian sketching* (FT): each worker runs a negatively shifted gradient flow on its sketched quadratic and stops it at a finite time, which keeps the worker operator positive semidefinite even when the shifted sketched Hessian is indefinite. For proportional Gaussian sketches of a Hessian with a few large eigenvalues over a tail of high effective rank, under weak regularization, let ζ be the ratio of the sketched-tail spread to the smallest large eigenvalue. When the large eigenvalues and the gradient's weights on them stay in fixed proportion, every stable shifted inverse keeps a mean-direction error of at least order ζ^2, whereas FT with explicit parameters reaches $O(\zeta^4 \log^2(1/\zeta))$ under conditions on the tail spread. Averaging q workers of a fixed policy reduces the remaining variance at rate $1/q$, which gives contraction bounds on quadratics without $m > d_1(\lambda)$. Experiments confirm the predicted response matching and worker crossover and show fewer outer iterations on synthetic and real-data problems with small local sketches.

1 INTRODUCTION

Newton's method scales each step by the curvature, but forming and factoring a large Hessian is often too expensive. At an iterate x of $F(x) = f(x) + \lambda\|x\|^2/2$, with f convex and twice differentiable and $\lambda > 0$, let $H = \nabla^2 f(x)$ and $g = \nabla F(x)$. The Newton direction is $d_\star = (H + \lambda I)^{-1} g$.

Massively parallel Hessian sketching splits this work across workers (Dereziński et al., 2020; Romanov et al., 2025). Worker j draws a Gaussian sketch $S_j \in \mathbb{R}^{m\times d}$, forms $B_j = S_j H S_j^\top$, and returns $S_j^\top (B_j + \alpha I)^{-1} S_j g$ for a local shift α; the server averages the q directions. More workers reduce random error but leave the mean response unchanged. Existing positive-shift corrections set the effective regularization to $\kappa = \lambda$ (Dereziński et al., 2020; Romanov et al., 2025), which is possible only when m exceeds the effective dimension $d_1(\lambda)$. A larger sketch costs m Hessian–vector products and $O(md)$ memory per worker, which a per-worker cap may not allow. In our memory-capped, massively parallel setting, synchronization rounds, not serial arithmetic, are the binding cost, so we count outer iterations at a fixed local sketch and claim no serial speedup.

Negative shifts help only up to a point. In the proportional Gaussian limit, LeJeune et al. (2024) characterize the stable negative branch of the sketched inverse, and a stricter fold threshold $d_2(\lambda)$ decides whether $\kappa = \lambda$ lies on it. For $d_2(\lambda) < m < d_1(\lambda)$ a stable negative shift still reaches the target; for $m < d_2(\lambda)$ no stable shift does, positive or negative (Proposition 3.1), and extra workers only estimate the biased mean more precisely (Figure 1a).

We get past this limit by replacing the shifted inverse with a finite-time filter, the early-stopped, negatively shifted gradient flow of Zhao (2026). Each worker runs this flow on its sketched quadratic for a time t and returns the result mapped back to $\mathbb{R}^d$. The negative shift counteracts the shrinkage of

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

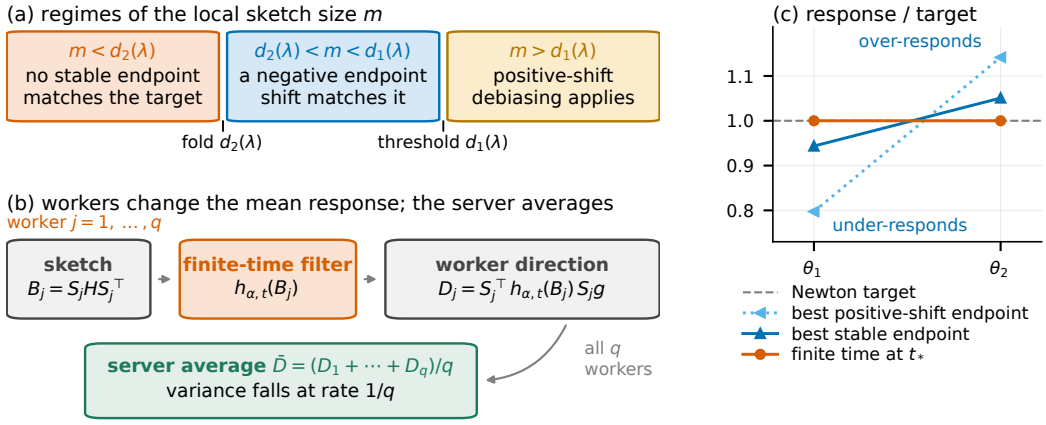


Figure 1: **Finite-time correction before worker averaging.** **(a)** Sketch-size regimes: a positive shift reaches the Newton target for $m > d_1(\lambda)$, a negative shift for $d_2(\lambda) < m < d_1(\lambda)$, and no stable shift for $m < d_2(\lambda)$. **(b)** FT corrects the mean response of each worker; averaging across workers reduces the variance. **(c)** After a common rescaling, FT matches the response on both spikes that carry the gradient, while the best stable endpoint leaves a mismatch.

the small sketch, and stopping at a finite t bounds the amplification. As $t \to \infty$ the flow converges to the shifted inverse above, which we therefore call a fixed-shift *endpoint*, but only when the shifted system is stable. At finite t the flow is defined for every shift, even past the endpoint's pole, and the operator it applies is positive semidefinite even when the shifted sketched Hessian is indefinite. With the stopping time as a second parameter, FT reaches mean directions that no stable endpoint can.

We compare directions by the scale-free Newton-energy ray loss of Section 3.1, which on a quadratic is the relative objective gap after the best step along the direction.

Contributions.

- **A limit of fixed-shift debiasing.** Below the fold, every stable fixed-shift endpoint, including the positive corrections of Dereziński et al. (2020) and Romanov et al. (2025) and the negative branch of LeJeune et al. (2024), keeps a mean-direction loss of at least $J_{\mathrm{end}}(\mu_0) > 0$ after the best rescaling when the gradient has components on two eigenvectors with distinct eigenvalues; more workers do not reduce it (Proposition 3.1).
- **Finite-time correction with sharper rates.** Under weak regularization and the high-effective-rank tail conditions of Theorem 3.2, an explicit FT pair whose shift lies beyond the stability boundary of every endpoint gives mean-direction loss $O(\zeta^4 \log^2(1/\zeta))$, compared with an endpoint floor of $\Omega(\zeta^2)$ for fixed relative head locations and weights. Here ζ is the ratio of the sketched-tail spread to the smallest head eigenvalue. Joint tuning can improve on this pair and matches both limiting responses exactly in a two-spike example (Section 5.1).
- **A finite-worker guarantee.** For any fixed positive-semidefinite worker rule, we bound the averaged direction's loss by its mean-direction error and a variance term that decreases as $1/q$, with no condition on the sketch size. This yields a sufficient worker count for FT to overtake an endpoint, a fixed-budget bound and contraction guarantees on quadratics (Theorem 3.3).
- **Simulation and real-data evidence.** Experiments on fixed spectra, sampled covariance models and real-data analysis show FT's advantage with small sketches and many workers. On sampled population-head problems, the gain persists with parameter reuse and without outer calibration; control experiments show how the gain depends on gradient geometry (Section 5).

2 Related Work

Randomized Newton and Hessian-sketch methods reduce the cost of second-order steps through low-dimensional projections (Pilanci & Wainwright, 2016; 2017; Lacotte & Pilanci, 2020). In

massively parallel variants, workers approximate a shared Newton direction independently and the server averages their outputs (Dereziński et al., 2020; Romanov et al., 2025). Existing corrections rescale the local regularization or the averaged direction (Dereziński et al., 2020; Zhang & Pilanci, 2023), and adaptive methods grow the sketch until an effective-dimension condition holds (Romanov et al., 2025). Resolvent equivalents give the limiting response of these shifted inverses, and ensemble analyses show that averaging reduces their fluctuations without changing that response (LeJeune et al., 2024; Du et al., 2023). We work below the fold, where no stable fixed-shift endpoint reaches the target response. The worker filter is the signed finite-time filter of negative-shifted gradient descent (Zhao, 2026), which established the negative-ridge barrier and early-stopped negative-shifted descent for regression. The novelty of this paper is the Hessian-sketch analogue: the fold barrier for distributed endpoints, the bias-cancelling shift that lies beyond their stability boundary, the Newton-direction rates under the scale-free ray loss, and the finite-worker averaging and optimization guarantees.

3 Theory

Standing assumptions. (A1) At the analyzed iterate, $H \succeq 0$, $\lambda > 0$ and $g \neq 0$; each statement concerns one outer step for the quadratic model below. (A2) Sketches have iid $N(0, 1/m)$ entries and are independent across workers and of any fixed policy assessed on them; selection on the same pool is not covered. (A3) Limits take $d, m \to \infty$ with $d/m \to \gamma > 1$ and a tail satisfying (G1)–(G3) of Appendix D. (A4) Separations take a gradient with positive Newton-energy weight on two distinct eigenvalues; the rate theorem adds an outlying head and $\lambda \leq \rho\mu_0$, $\rho < 1$. The finite-worker theorem needs only (A1)–(A2). Appendix B states the assumptions in full; proofs are in Appendices C and D.

3.1 The stable-endpoint barrier

At a fixed outer iterate, write $A = H + \lambda I$, $d_\star = A^{-1}g$, and $Q(d) = \frac{1}{2}d^\top Ad - g^\top d$. Worker j forms $B_j = S_j H S_j^\top$ and returns $D_j = M_j g$, where $M_j = S_j^\top h(B_j) S_j$ applies a scalar filter h to the eigenvalues of B_j: $h(b) = 1/(b+\alpha)$ for an endpoint, and the finite-time filter of Section 4 for FT. The server averages $\overline{D} = q^{-1}\sum_j D_j$.

The outer step size absorbs any positive scale, so we measure only the direction, through the Newton-energy ray loss

$$\mathcal{R}_{\text{ray}}(d) = \min_{c \geq 0} \frac{\|cd - d_\star\|_A^2}{g^\top A^{-1} g}. \tag{1}$$

When $g^\top d > 0$,

$$\mathcal{R}_{\text{ray}}(d) = 1 - \frac{(g^\top d)^2}{(d^\top Ad)(g^\top A^{-1} g)}. \tag{2}$$

On a quadratic objective this is exactly the relative objective gap after the best step along d.

Let $h_1, \ldots, h_d$ be the eigenvalues of H, and define

$$\begin{aligned} d_1(\mu) &= \sum_{i=1}^{d} \frac{h_i}{h_i + \mu}, \qquad & d_2(\mu) &= \sum_{i=1}^{d} \left(\frac{h_i}{h_i + \mu}\right)^2, \\ \ell(\mu) &= \mu\left(1 - \frac{d_1(\mu)}{m}\right), & \ell'(\mu) &= 1 - \frac{d_2(\mu)}{m}. \end{aligned} \tag{3}$$

If $\operatorname{rank}(H) > m$, the fold $\mu_0 > 0$ solves $d_2(\mu_0) = m$. The two thresholds are

$$\ell(\lambda) < 0 \iff m < d_1(\lambda), \qquad \lambda < \mu_0 \iff m < d_2(\lambda). \tag{4}$$

The first marks where the shift that targets λ turns negative; the second, where λ falls below every effective regularization a stable endpoint can reach (Proposition 3.1).

Write $H = U \operatorname{diag}(h_i) U^\top$ and express g in this basis; we call these coordinates the source. Its Newton-energy weights are

$$w_i = \frac{g_i^2/(h_i + \lambda)}{g^\top A^{-1} g}, \qquad \sum_i w_i = 1. \tag{5}$$

By Gaussian sign symmetry, $\mathbb{E}M = U\,\mathrm{diag}(\eta_i)U^\top$, so the mean direction multiplies the ith Newton coordinate by the relative response $r_i = (h_i + \lambda)\eta_i$. For a fixed positive-semidefinite spectral rule with nonzero mean action, the loss of the mean direction, $R_\infty = \mathcal{R}_{\mathrm{ray}}(\mathbb{E}[M]g)$, is the weighted relative variance of these responses:

$$R_\infty = 1 - \frac{\left(\sum_i w_i r_i\right)^2}{\sum_i w_i r_i^2} = \frac{\mathrm{Var}_w(r)}{\mathbb{E}_w[r^2]}. \tag{6}$$

Matching the Newton direction therefore means flattening the responses on the eigenvalues that carry the source (Proposition C.3 in Appendix C). For an endpoint with effective regularization κ, write

$$r_i(\kappa) = \frac{h_i + \lambda}{h_i + \kappa}, \qquad J_{\mathrm{end}}(\kappa) = \frac{\mathrm{Var}_w(r(\kappa))}{\mathbb{E}_w[r(\kappa)^2]}.$$

Proposition 3.1 (Stable-endpoint barrier). *Suppose* $\mathrm{rank}(H) > m$.

(i) Stable endpoint responses. *Under the proportional deterministic equivalent of LeJeune et al. (2024), stable endpoints have limiting mean action*

$$(H + \kappa I)^{-1}, \qquad \alpha = \ell(\kappa), \qquad \kappa > \mu_0.$$

The target $\kappa = \lambda$ requires a negative local shift when $d_2(\lambda) < m < d_1(\lambda)$; below the fold, $m < d_2(\lambda)$, every stable endpoint has $\kappa > \mu_0 > \lambda$.

(ii) Scalar-optimized floor. *If $0 < \lambda < \mu_0$, then*

$$\inf_{\kappa > \mu_0} J_{\mathrm{end}}(\kappa) = J_{\mathrm{end}}(\mu_0).$$

The floor is strictly positive whenever the source assigns positive Newton-energy weight to two distinct eigenvalues.

Below the fold, every stable endpoint has effective regularization $\kappa > \mu_0 > \lambda$, so its responses $(h_i + \lambda)/(h_i + \kappa)$ differ across distinct eigenvalues, and a common rescaling cannot make them equal. Proposition 3.1 therefore establishes a floor for every stable fixed-shift endpoint, negative shifts included. Appendix D.8 gives the limit of the guarded fixed-shift endpoints used in practice; positive mixtures also retain nonzero mean-direction loss (Remark C.4).

3.2 Finite time beats the endpoint floor

FT has two parameters, the shift and the stopping time, and can shape the relative responses in ways a single shift cannot. Let the spectrum of H_d, apart from a head of finitely many eigenvalues, converge to a limit τ_H as $d/m \to \gamma > 1$, with bounded norm, nonzero eigenvalues bounded away from zero, and limiting rank above m (conditions (G1)–(G3) of Appendix D).

Write $D_k(u) = \gamma \int \{h/(h+u)\}^k \, d\tau_H(h)$ for the limit of $d_k(u)/m$, so that the limiting fold solves $D_2(\mu_0) = 1$. Let $a = \gamma \int h \, d\tau_H(h)$, the limit of $\mathrm{tr}(H_d)/m$, be the tail floor, and let $a_- = \mu_0\{D_1(\mu_0) - 1\}$ be the lower edge of the limiting spectrum of $SH_dS^\top$; the stable shifts are those above $-a_-$. Let $\bar{h}$ be the largest tail eigenvalue. The fold μ_0 measures the spread of the sketched tail around a; Appendix D.1 gives the variance bounds and flat-tail formulas.

FT applies to each sketched Hessian the filter $h_{\alpha,t}(x) = \int_0^t e^{-(x+\alpha)s}\, ds$ with shift $\alpha = -\nu$ (Section 4). For a head eigenvalue θ with unit eigenvector e_θ, define

$$L_\theta(\nu, t) = \lim_{d\to\infty} \mathbb{E}\, e_\theta^\top S^\top h_{-\nu,t}(SH_dS^\top) S e_\theta, \qquad r_\theta(t) = (\theta + \lambda) L_\theta(\nu, t). \tag{7}$$

The source has Newton-energy weights w on the head, free of d, and $R_{\infty,\mathrm{FT}}(\nu, t)$ is the loss (6) of the responses $r_\theta(t)$. The next theorem says that when the sketched tail is narrow relative to the head, the best stable endpoint keeps a second-order directional error, while FT at a suitable signed level and stopping time reduces it to fourth order, up to logarithms.

Theorem 3.2 (Rate separation over a high-effective-rank tail). *Assume (G1)–(G3) and $0 < \lambda \le \rho\mu_0$ with $\rho < 1$. Let the source lie on head eigenvalues $\theta \in [\theta_{\min}, \theta_{\max}]$ that are outliers of the sketched tail, $\theta > \bar{h}$ and $m_\theta = \gamma \int h^2/(\theta - h)^2 \, d\tau_H < 1$, at least two distinct ones with positive weight, and put $\zeta = \mu_0/\theta_{\min}$.*

(i) Every stable fixed-shift endpoint has limiting mean-direction loss at least

$$J_{\mathrm{end}}(\mu_0) \geq (1-\rho)^2\,\zeta^2\,V, \qquad V = \mathrm{Var}_w\{\theta_{\min}/(\theta+\mu_0)\} > 0. \tag{8}$$

(ii) The level $\nu = a - \lambda$ exceeds a_- (Lemma D.4), so it is available to finite time but to no stable endpoint. At this level, for every $t > 0$, the head-response error is bounded below by a displacement defect and above by an exposure term: with $\varepsilon_\theta = \gamma \int h^2/(\theta - h)\, d\tau_H$,

$$-\big(m_\theta + \varepsilon_\theta/\theta + e^{-\theta t}\big) \leq r_\theta(t) - 1 \leq (\theta+\lambda)\, m_\theta\, t\, e^{(\nu - a_-)t}. \tag{9}$$

(iii) If the tail eigenvalues are at most μ_0 and $\zeta \leq 1/4$, then at $t = 2\log(1/\zeta)/\theta_{\min}$, $\sup_\theta |r_\theta(t) - 1| \leq \delta_\zeta = O(\zeta^2 \log(1/\zeta))$, with δ_ζ given explicitly in Appendix D, and, when $\delta_\zeta < 1$,

$$R_{\infty,\mathrm{FT}}(\nu, t) \leq \frac{\delta_\zeta^2}{(1-\delta_\zeta)^2} = O\big(\zeta^4 \log^2(1/\zeta)\big). \tag{10}$$

Along any joint sequence with $d, m, q \to \infty$ and $d/m \to \gamma$, the expected actual FT ray loss converges to $R_{\infty,\mathrm{FT}}(\nu, t)$.

Why finite time helps. Sketching moves an outlying head eigenvalue from θ to about $\theta + a$, almost the same displacement for every head eigenvalue. The level that cancels it, $\nu = a - \lambda$, lies at least $2\mu_0 - \lambda$ past the lower edge a_- of the sketched spectrum (Lemma D.4), where a shifted inverse has its pole inside the spectrum, so no stable endpoint can use it. Finite stopping removes the pole (Proposition C.1), so FT runs at exactly this level; the head then evolves at the Newton scale $\theta + \lambda$ (Lemma D.5) while finite time keeps the tail's contribution small and, under part (iii), the exposure at the edge bounded. The head responses are then within $O(\zeta^2 \log(1/\zeta))$ of one, whereas the best stable endpoint is off by a first-order amount that differs across distinct eigenvalues carrying the source. The loss is quadratic in the response error, which gives the fourth-order rate: the signed path removes the leading distortion (Zhao, 2026). For fixed relative head locations and weights, V stays positive and the endpoint floor is $\Omega(\zeta^2)$.

The explicit pair and the tuned method. The theorem's shift and stopping time certify what the finite-time family can reach. Joint tuning can improve response matching and achieves exact matching in the two-spike example of Section 5.1. The closed-form constants are conservative: exact limiting responses already separate at moderate spectral ratios (Appendix F.3). The separation also tolerates a fraction of tail Newton energy that is small relative to ζ^2 (Corollary D.6).

3.3 Workers reduce the remaining variance

Let $T_i = A^{1/2} M_i A^{1/2}$, $v = A^{-1/2} g / \sqrt{g^\top A^{-1} g}$ and $\overline{u}_q = q^{-1} \sum_{i=1}^q T_i v$.

Theorem 3.3 (Calibrated loss with independent workers)**.** *Let $T_1, \ldots, T_q \succeq 0$ be iid with $\mathbb{E}\|T_i v\|^2 < \infty$. Define $\mu = \mathbb{E}[T_i v] \neq 0$, $\sigma^2 = \mathbb{E}\|T_i v - \mu\|^2$, $\omega = \sigma^2/\|\mu\|^2$,*

$$R_q = \min_{c \geq 0} \|v - c\overline{u}_q\|^2, \qquad R_\infty = 1 - \frac{(v^\top \mu)^2}{\|\mu\|^2}.$$

Then

$$\mathbb{E}R_q \leq \frac{R_\infty + \omega/q}{1 + \omega/q} \leq R_\infty + \frac{\omega}{q}, \qquad |\mathbb{E}R_q - R_\infty| \leq \frac{6\omega}{q}. \tag{11}$$

The filter sets R_∞ and averaging sets ω/q, so a filter with the better mean direction can still lose at small q if its workers are noisier. For fixed FT and endpoint policies with $\Delta = R_{\infty,\mathrm{end}} - R_{\infty,\mathrm{FT}} > 0$, FT has the smaller expected loss once $q > (\omega_{\mathrm{FT}} + 6\omega_{\mathrm{end}})/\Delta$. At a total budget of $N = qm$ sketch columns, the FT bound reads $\mathbb{E}R_{N/m,\mathrm{FT}}(m) \leq R_{\infty,\mathrm{FT}}(m) + \omega_{\mathrm{FT}}(m)\, m/N$, with no requirement $m > d_1(\lambda)$ (Corollary C.5 and Proposition C.6).

On a quadratic objective the ray loss is the next relative objective gap, so Corollary C.7 gives the contraction $\mathbb{E}[\Gamma_{k+1} \mid x_k] \leq \{R_\infty(w^{(k)}) + \omega(w^{(k)})/q\}\, \Gamma_k / \{1 + \omega(w^{(k)})/q\}$. The guarantee covers fixed rules and policies chosen independently of the sketches they are applied to.

Table 1: **Methods compared.** The sketch baselines backtrack from step one and L-BFGS uses its native line search; our calibration c_Q may exceed one. The loss column refers to the head-source setting of Theorem 3.2, where every stable endpoint has limiting mean ray loss at least $(1-\rho)^2\zeta^2 V$. The FT rate is that of the theorem's explicit pair; the implemented method tunes its signed level and stopping time.

Method	Direction	Mean-direction loss below the fold	Step
RZP, limited m	resolvent $S^\top(B+\hat{\lambda}I)^{-1}Sg$ with the shift $\hat{\lambda} \in [5\lambda/12, \lambda]$ of Romanov et al. (2025)	floor for the fixed fallback shift	from 1, shrink only
Undebiased endpoint	resolvent, $\alpha = \lambda$	positive	from 1, shrink only
L-BFGS	quasi-Newton, one worker, no HVPs	not applicable	own line search
Tuned stable endpoint (ours)	resolvent, best of 65 positive and 65 admissible negative shifts	positive, at least $J_{\mathrm{end}}(\mu_0)$ (Prop. 3.1)	$c_Q > 1$, then backtracking
FT, no c_Q (ours)	finite-time filter, model value at unit scale	$O(\zeta^4 \log^2(1/\zeta))$	from 1, shrink only
FT (ours)	finite-time filter, calibrated model value	$O(\zeta^4 \log^2(1/\zeta))$	$c_Q > 1$, then backtracking

4 Finite-Time Hessian Sketching

One outer iteration works as follows, in the notation of Section 3.1. The server coordinates q workers; each can call the same Hessian–vector product (HVP) oracle $v \mapsto Hv$ but stores only an m-row sketch.

Finite-time worker filter. Worker j draws $S_j \in \mathbb{R}^{m\times d}$ with iid $N(0, 1/m)$ entries and forms the sketched Hessian $B_j = S_j H S_j^\top$ with m HVPs. For a shift $\alpha \in \mathbb{R}$ and a stopping time $t > 0$, define

$$h_{\alpha,t}(b) = \int_0^t e^{-(b+\alpha)s}\, ds = \begin{cases} \dfrac{1-e^{-(b+\alpha)t}}{b+\alpha}, & b+\alpha \neq 0, \\ t, & b+\alpha = 0. \end{cases} \tag{12}$$

The worker runs the shifted gradient flow of its sketched quadratic up to time t and returns the state mapped back to $\mathbb{R}^d$:

$$D_j(\alpha,t) = S_j^\top h_{\alpha,t}(B_j) S_j g = S_j^\top z_j(t), \qquad \dot{z}_j(s) = S_j g - (B_j+\alpha I) z_j(s), \quad z_j(0) = 0. \tag{13}$$

Proposition C.2 identifies this trajectory with negative-shifted gradient descent on an auxiliary regression problem.

A negative shift counteracts the shrinkage of the small sketch, and finite stopping bounds the amplification. Since $h_{\alpha,t}(b) > 0$ for every real $b+\alpha$, the worker operator $S_j^\top h_{\alpha,t}(B_j) S_j$ is positive semidefinite even when $B_j + \alpha I$ is indefinite. The exposure $E_j(\alpha,t) = t[-\lambda_{\min}(B_j) - \alpha]_+$ controls unstable amplification (Proposition C.1):

$$0 \preceq S_j^\top h_{\alpha,t}(B_j) S_j \preceq t e^{E_j(\alpha,t)} S_j^\top S_j. \tag{14}$$

The experiments bound the search by a cap on E_j (12 in the sampled-covariance and corn studies) that fixes the grid's range and guards reuse but is not a requirement of the filter: the calibrated step is scale-invariant and the model score selects (Appendix E.3).

Worker averaging and calibration. For each candidate (α, t), the server averages $\overline{D}(\alpha,t) = q^{-1}\sum_{j=1}^q D_j(\alpha,t)$. The filter sets the shape of the direction, and the local quadratic model Q of Section 3.1 fixes its scale: for a nonzero direction D,

$$c_Q(D) = \frac{[g^\top D]_+}{D^\top A D} = \frac{[g^\top D]_+}{D^\top (HD + \lambda D)} \tag{15}$$

minimizes $Q(cD)$ over $c \geq 0$ and attains the ray loss (1). The server keeps the candidate with the smallest calibrated model value and sets $p = c_Q(\overline{D})\overline{D}$. On a quadratic objective the step $x^+ = x - p$ is optimal along the chosen ray, and the next relative gap equals the realized ray loss; on other objectives we apply Armijo backtracking to $F(x - sp)$ from $s = 1$. The scalar c_Q can exceed one, so we also report FT without it, with the baselines' step rule of backtracking only from $c_Q = 1$.

Small-matrix implementation. Because $m \ll d$, each worker diagonalizes $B_j = U_j \operatorname{diag}(b_{j,1}, \ldots, b_{j,m}) U_j^\top$ once and caches $L_j = S_j^\top U_j$ and $\widetilde{K}_j = H S_j^\top U_j$. With $a_j = (h_{\alpha,t}(b_{j,i}))_{i \leq m}$,

$$D_j(\alpha, t) = L_j \operatorname{diag}(a_j) L_j^\top g, \qquad HD_j(\alpha, t) = \widetilde{K}_j \operatorname{diag}(a_j) L_j^\top g, \tag{16}$$

so extra candidates cost no further HVPs, and returning both D_j and HD_j lets the server evaluate (15) without another HVP.

Parameter selection and cost. The shift and time are selected on the calibrated quadratic model, without knowledge of μ_0, $\theta_{\min}$ or the tail law. The explicit pair of Theorem 3.2 certifies what the family can reach; joint tuning over an expanded grid of shifts and times is the experimental method; quality-aware reuse, the low-overhead variant, keeps the previous pair and searches again only when a nearby probe improves the model score materially or the stored pair leaves the grid's exposure range (Appendix E.3). Every round draws fresh sketches. Theorem 3.3 covers a stored pair conditional on past sketches; selection on the current pool is evaluated experimentally. With C candidates, each worker spends m HVPs, one $O(m^3)$ factorization, $O(Cmd)$ filter work and $O(md)$ storage, and sends $m + 2Cd$ numbers. Table 1 defines the comparators; the full algorithms and resource accounting are in Appendix E.

5 Experiments

All sketched methods use Gaussian sketches, the same HVP oracle, and fresh sketches at every outer iteration. The target is a certified relative objective gap of 10^{-4}; a run that does not reach it within the study's budget counts as no hit (protocols in Appendices F–I).

Regime. The local sketch size is $m = 16$ unless stated otherwise, and the problems are weakly regularized with a tail of high effective rank, so that this sketch lies below the fold. Fixed-Hessian studies isolate the mechanism (Section 5.1); sampled covariance models and real ridge regressions on corn spectroscopy, CT slice location and superconductivity test the optimization benefit (Sections 5.2 and 5.3). Nonlinear and broad-source controls are in Appendix H. The $d = 128$ closed loops are controlled studies with paired seeds and certified gaps; the $d = 10^4$ power-law head sweep is the larger-dimensional spectral test and keeps FT ahead of the endpoint family (Appendix H.7).

5.1 Response matching and worker averaging

Joint tuning matches the two limiting head responses exactly in a model with two outlying spikes over a flat tail. This model satisfies Theorem 3.2(i)–(ii); the tuned shift and stopping time give zero FT mean-direction loss while every stable endpoint retains a positive floor. The exposure stays below six, and stopping 20% early or late still keeps the loss below the endpoint floor. The model and tuned pair are given in Appendix F, with timing robustness in Appendix F.2.

The prediction holds at finite size. At fixed $d/m = 16$, FT's estimated mean-direction loss falls to its noise level as d grows from 256 to 8,192, while the stable endpoints stay above their floor (Figure 6). Across flat, geometric and power-law tails with aspect ratios from 16 to 1024 and $\lambda \leq \mu_0/4$, the endpoint floor is 9.9 to 405 times the limiting FT loss at the theorem's explicit pair (Figure 2a and Appendix F.3).

Figure 2c is the theorem-aligned test of the finite-worker bound: on a polynomial spectrum with $d = 500$, $\lambda = 10^{-4}$ and $m = 16$, policies chosen on an independent pilot are frozen and assessed on disjoint worker pools, the fixed-rule setting of Theorem 3.3, unlike the same-pool closed loops of Section 5.2. FT has the smaller estimated mean-direction loss but the noisier workers (Appendix G). Averaging turns its smaller mean error into a gain: FT overtakes the endpoint between $q = 16$ and

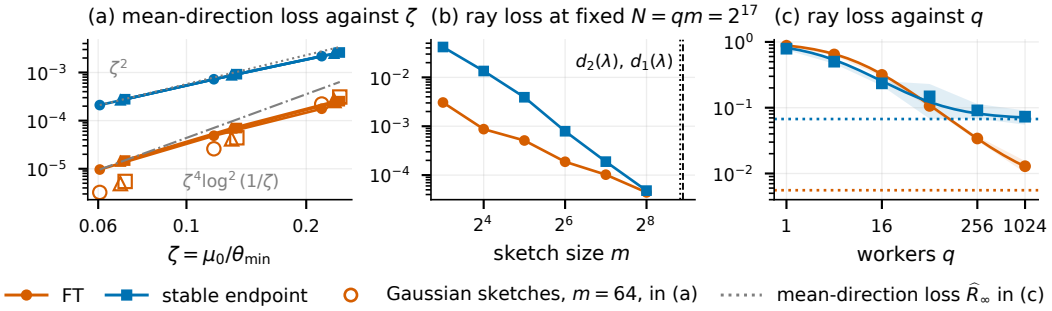


Figure 2: **Finite time corrects the mean response before averaging.** (a) Limiting floor of the stable endpoints and limiting FT loss at the explicit pair of Theorem 3.2, for flat, geometric and power-law tails (circles, squares, triangles) at $\lambda = \mu_0/4$. Open markers are the same losses measured from the mean responses of Gaussian sketches with $m = 64$; grey lines are the guides ζ^2 and $\zeta^4 \log^2(1/\zeta)$. (b) Ray loss of FT and the stable endpoint (ours) at a fixed total budget $N = qm$, for a source on the head; vertical lines mark $d_2(\lambda)$ and $d_1(\lambda)$. (c) Independently selected, frozen policies against the number of workers: markers are measured losses, curves the bound of Theorem 3.3 at independently estimated moments, and dotted lines the mean-direction losses.

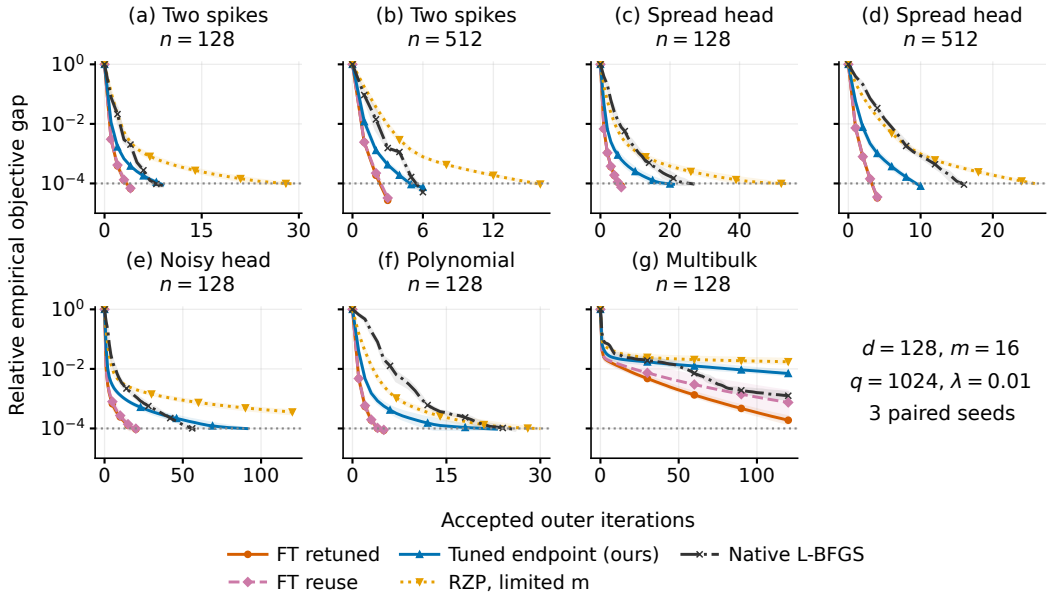


Figure 3: **FT reduces outer iterations on sampled covariance objectives.** All seven models, $d = 128$, $m = 16$, $q = 1024$, $\lambda = 0.01$. Lines are mean relative objective gaps over three paired seeds, shading their range; the dotted line marks 10^{-4}, and converged runs keep their final gap. Quality-aware reuse closely tracks full retuning. The horizontal axis counts accepted outer iterations, not time.

64, consistent with the independently estimated moment-curve crossing near $q = 49$. At a fixed total budget $N = qm$, FT gains most for sketches below the fold, and both methods improve as m grows (Figure 2b); in the well-regularized control the endpoint matches FT as the sketch grows (Appendix I).

5.2 Closed-loop optimization

Each model fixes a covariance Σ and a signal β mainly on the population head, then samples $X = Z\Sigma^{1/2}$ and $y = X\beta + \varepsilon$. Every method minimizes the same ridge objective $F(x) =$

Table 2: **FT reduces outer iterations on corn and CT slice.** Iterations to relative gap 10^{-4} at $m = 16$, $q = 256$: mean [range] over three paired runs, which are three sketch seeds on one split for corn and three grouped data splits for CT slice and superconductivity. Corn L-BFGS is deterministic and ran with an extended budget. A fraction marks partial success, > 120 no hit within 120 iterations, and a dash a method not run. Bold marks the fewest iterations in each column. Counts are outer iterations, not time.

	Corn spectroscopy				UCI regression	
Method	Moisture	Oil	Protein	Starch	CT slice	Supercond.
FT, retuned	**4.7 [4–5]**	43.7 [40–47]	38.3 [36–40]	37.3 [35–40]	**36.7 [34–39]**	1/3 hit
FT, quality reuse	7.3 [6–9]	70.7 [68–73]	60.7 [57–63]	56.7 [54–61]	37.7 [35–40]	1/3 hit
FT, no c_Q	5.3 [5–6]	**36.3 [35–38]**	**32.3 [31–34]**	**30.0 [29–31]**	37.0 [35–38]	1/3 hit
Tuned endpoint (ours)	12.3 [10–15]	1/3 hit	93.7 [82–100]	97.0 [86–117]	1/3 hit	> 120
Endpoint reuse (ours)	15.0 [12–20]	> 120	> 120	> 120	—	—
RZP, limited m	27.7 [27–28]	> 120	> 120	> 120	> 120	> 120
Undebiased endpoint	—	—	—	—	> 120	> 120
Native L-BFGS	1,125	5,657	5,055	5,018	60.7 [56–65]	> 120

$\|Xx - y\|^2/(2n) + \lambda\|x\|^2/2$ from $x_0 = 0$. Sampling couples head and tail, so the empirical gradient is not confined to the population head. We use $d = 128$, $n \in \{128, 512\}$, two-spike, spread-head and polynomial covariances, a noisy-response case and a multibulk control, all below the fold at $m = 16$ (Appendix H.2).

At $q = 1024$, FT achieves a 1.8- to 4.3-fold reduction in iterations relative to the tuned stable endpoint on the six non-multibulk models, and takes 50% to 82% fewer iterations than native L-BFGS (Figure 3, Table 3). The gain holds under quality-aware reuse, and FT without c_Q beats L-BFGS in all 18 paired runs, so it comes from the filter rather than the calibration. At $q = 256$, retuning and reuse again win all 18 comparisons with L-BFGS (Table 4). No method reaches the target on the multibulk control within 120 iterations.

Appendix H extends the above experiments to nonlinear objectives and to broad-source quadratics at $d = 10^4$: on logistic and softmax regression FT still reaches the target first, in 14 and 28 rounds against 56 and 45 for native L-BFGS, while neither sketched endpoint baseline reaches it (Appendix H.3); the $d = 10^4$ power-law head sweep is the larger-dimensional spectral test and keeps FT ahead of the endpoint family (Appendix H.7).

5.3 Real-data regression

On corn spectroscopy, FT takes fewer outer iterations on all four outcomes (Table 2). The common $\lambda = 10^{-8}$ was fixed by a validation screen before any optimizer was run. At $m = 16$ and $q = 256$, plain FT and quality-aware reuse both beat the tuned endpoint and native L-BFGS. Across $q \in \{64, 256, 1024\}$, all 108 FT runs reach 10^{-4}. On CT slice regression, retuned FT and reuse take 36.7 and 37.7 iterations, against 60.7 for L-BFGS. Superconductivity is harder: each FT variant reaches the target on one of three splits, and no endpoint and no L-BFGS run reaches it within 120 iterations. Data, selection and tuning protocols are in Appendices H.4 and H.5.

6 Conclusion

Below the fold, the shift that cancels sketch bias lies past the stability edge of every shifted inverse, so no stable endpoint gives the mean Newton direction, however many workers; a negatively shifted flow stopped at finite time crosses that edge. In closed loops on sampled covariance models and real data sets, FT needs fewer outer iterations than every stable shift, also with parameter reuse and without the outer calibration. It helps most with small local sketches, a gradient spread over several large eigenvalues, and many workers.

Reproducibility Statement

The code, the frozen experiment protocols, the saved campaign summaries and the table and figure generators are available in an anonymous repository at https://anonymous.4open.science/r/finite-time-hessian-sketching-2C7E; the code and protocols are also in the supplementary material. Its README explains how to regenerate every table and figure of the paper and how to rerun each experiment from its registered protocol. Appendix J lists data sources, preprocessing, objective and stopping conventions, and the replay audits. Complete proofs are in Appendices C and D, with the standing assumptions collected in Appendix B and restated in each theorem.

AI Use Statement

The authors supplied the core mathematical mechanism and directed the research questions, theoretical development and experimental design. Generative AI was not used to originate the mechanism, the theoretical framework or the research questions. After the authors completed the proofs, generative AI assistants helped with several rounds of mathematical proofreading, checking logical steps, assumptions, notation and consistency between statements and proofs. The authors evaluated the feedback and verified the revised arguments. During the preceding, author-directed development, AI also helped develop and refine intermediate mathematical claims, proof arguments and experimental protocols. Further assistance included implementing and testing code, executing the frozen protocols, generating figures and tables from saved results, and editing the manuscript. AI-generated code was tested against exact reference implementations and independent replay audits. The authors take full responsibility for all content.

References

Naman Agarwal, Brian Bullins, and Elad Hazan. Second-order stochastic optimization for machine learning in linear time. *Journal of Machine Learning Research*, 18(116):1–40, 2017. URL https://jmlr.org/papers/v18/16-491.html.

Alnur Ali, J. Zico Kolter, and Ryan J. Tibshirani. A continuous-time view of early stopping for least squares regression. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1370–1378. PMLR, 2019. URL https://proceedings.mlr.press/v89/ali19a.html.

Zhidong Bai and Jack W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer Series in Statistics. Springer, 2nd edition, 2010. doi: 10.1007/978-1-4419-0661-8.

Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005. doi: 10.1214/009117905000000233.

Burak Bartan and Mert Pilanci. Distributed sketching for randomized optimization: Exact characterization, concentration, and lower bounds. *IEEE Transactions on Information Theory*, 69(6): 3850–3879, 2023. doi: 10.1109/TIT.2023.3247559.

Albert S. Berahas, Raghu Bollapragada, and Jorge Nocedal. An investigation of Newton-sketch and subsampled Newton methods. *Optimization Methods and Software*, 35(4):661–680, 2020. doi: 10.1080/10556788.2020.1725751.

Raghu Bollapragada, Richard H. Byrd, and Jorge Nocedal. Exact and inexact subsampled Newton methods for optimization. *IMA Journal of Numerical Analysis*, 39(2):545–578, 2019. doi: 10.1093/imanum/dry009.

Rixon Crane and Fred Roosta. DINGO: Distributed Newton-type method for gradient-norm optimization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Ron S. Dembo, Stanley C. Eisenstat, and Trond Steihaug. Inexact Newton methods. *SIAM Journal on Numerical Analysis*, 19(2):400–408, 1982. doi: 10.1137/0719025.

Michał Dereziński and Michael W. Mahoney. Distributed estimation of the inverse hessian by determinantal averaging. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Michał Dereziński, Burak Bartan, Mert Pilanci, and Michael W. Mahoney. Debiasing distributed second order optimization with surrogate sketching and scaled regularization. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Michał Dereziński, Zhenyu Liao, Edgar Dobriban, and Michael W. Mahoney. Sparse sketches with small inversion bias. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*. PMLR, 2021.

Edgar Dobriban and Yue Sheng. WONDER: Weighted one-shot distributed ridge regression in high dimensions. *Journal of Machine Learning Research*, 21(66):1–52, 2020. URL https://jmlr.org/papers/v21/19-277.html.

Edgar Dobriban and Yue Sheng. Distributed linear regression by averaging. *The Annals of Statistics*, 49(2):918–943, 2021. doi: 10.1214/20-AOS1984.

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018. doi: 10.1214/17-AOS1549.

Jin-Hong Du, Pratik Patil, and Arun K. Kuchibhotla. Subsample ridge ensembles: Equivalences and generalized cross-validation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8585–8631. PMLR, 2023. URL https://proceedings.mlr.press/v202/du23d.html.

Eigenvector Research. NIR of corn samples for standardization benchmarking. Public spectroscopy dataset, n.d. URL https://www.eigenvector.com/data/Corn/. Accessed September 21, 2026.

Tobias Emrich, Franz Graf, Hans-Peter Kriegel, Matthias Schubert, Markus Thoma, and Alexander Cavallaro. CT slice localization via instance-based regression. In *Medical Imaging 2010: Image Processing*, volume 7623 of *Proceedings of SPIE*, pp. 762320, 2010. doi: 10.1117/12.841877.

Murat A. Erdogdu and Andrea Montanari. Convergence rates of sub-sampled Newton methods. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Sachin Garg, Kevin Tan, and Michał Dereziński. Distributed least squares in small space via sketching and bias reduction. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/86c283920335ed1fec3edee227e05fbf-Abstract-Conference.html.

Robert M. Gower, Dmitry Kovalev, Felix Lieder, and Peter Richtárik. RSN: Randomized subspace Newton. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Franz Graf, Hans-Peter Kriegel, Matthias Schubert, Sebastian Poelsterl, and Alexander Cavallaro. Relative location of CT slices on axial axis. UCI Machine Learning Repository, 2011. URL https://doi.org/10.24432/C5CP6G. Dataset.

Sunny Gupta, Mohit Jindal, Pankhi Kashyap, Pranav Jeevan, and Amit Sethi. FLeNS: Federated learning with enhanced Nesterov-Newton sketch. *arXiv preprint arXiv:2409.15216*, 2024.

Kam Hamidieh. A data-driven statistical model for predicting the critical temperature of a superconductor. *Computational Materials Science*, 154:346–354, 2018a. doi: 10.1016/j.commatsci.2018.07.052.

Kam Hamidieh. Superconductivty data. UCI Machine Learning Repository, 2018b. URL https://doi.org/10.24432/C53P47. Dataset; repository title preserves the original spelling.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022. doi: 10.1214/21-AOS2133.

Rustem Islamov, Xun Qian, and Peter Richtárik. Distributed second order methods with fast rates and compressed communication. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*. PMLR, 2021.

Iain M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327, 2001. doi: 10.1214/aos/1009210544.

Antti Knowles and Jun Yin. Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, 169(1–2):257–352, 2017. doi: 10.1007/s00440-016-0730-4.

Dmitry Kobak, Jonathan Lomond, and Benoit Sanchez. The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *Journal of Machine Learning Research*, 21(169):1–16, 2020. URL https://jmlr.org/papers/v21/19-844.html.

Jonathan Lacotte and Mert Pilanci. Effective dimension adaptive sketching methods for faster regularized least-squares optimization. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/e105b88b3e1ac23ec811a708cd7edebf-Abstract.html.

Jonathan Lacotte and Mert Pilanci. Adaptive and oblivious randomized subspace methods for high-dimensional optimization: Sharp analysis and lower bounds. *IEEE Transactions on Information Theory*, 68(5):3281–3303, 2022.

Olivier Ledoit and Sandrine Péché. Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151(1–2):233–264, 2011. doi: 10.1007/s00440-010-0298-3.

Daniel LeJeune, Hamid Javadi, and Richard G. Baraniuk. The implicit regularization of ordinary least squares ensembles. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*. PMLR, 2020. URL https://proceedings.mlr.press/v108/lejeune20b.html.

Daniel LeJeune, Pratik Patil, Hamid Javadi, Richard G. Baraniuk, and Ryan J. Tibshirani. Asymptotics of the sketched pseudoinverse. *SIAM Journal on Mathematics of Data Science*, 6(1):199–225, 2024. doi: 10.1137/22M1530264.

Jian Li, Yong Liu, and Weiping Wang. FedNS: A fast sketching Newton-type algorithm for federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 13509–13517, 2024. doi: 10.1609/aaai.v38i12.29254.

Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1–3):503–528, 1989. doi: 10.1007/BF01589116.

Vladimir A. Marchenko and Leonid A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967. doi: 10.1070/SM1967v001n04ABEH001994.

James Martens. Deep learning via Hessian-free optimization. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 735–742, 2010.

Sen Na, Michał Dereziński, and Michael W. Mahoney. Hessian averaging in stochastic Newton methods achieves superlinear convergence. *Mathematical Programming*, 201(1–2):473–520, 2023. doi: 10.1007/s10107-022-01913-5.

Chengmei Niu, Zhenyu Liao, Zenan Ling, and Michael W. Mahoney. Fundamental bias in inverting random sampling matrices with application to sub-sampled Newton. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 46649–46692. PMLR, 2025. URL https://proceedings.mlr.press/v267/niu25c.html.

Pratik Patil and Jin-Hong Du. Generalized equivalences between subsampling and ridge regularization. In *Advances in Neural Information Processing Systems*, volume 36, pp. 78926–78963, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f976982cd1c1b9e076c096787ef6652e-Abstract-Conference.html.

Pratik Patil and Daniel LeJeune. Asymptotically free sketched ridge ensembles: Risks, cross-validation, and tuning. In *International Conference on Learning Representations*, 2024. URL https://iclr.cc/virtual/2024/poster/18074.

Mert Pilanci and Martin J. Wainwright. Iterative hessian sketch: Fast and accurate solution approximation for constrained least-squares. *Journal of Machine Learning Research*, 17(53):1–38, 2016. URL https://jmlr.org/papers/v17/14-460.html.

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

Sashank J. Reddi, Jakub Konečný, Peter Richtárik, Barnabás Póczos, and Alexander J. Smola. AIDE: Fast and communication efficient distributed optimization. *arXiv preprint arXiv:1608.06879*, 2016. doi: 10.48550/arXiv.1608.06879. URL https://arxiv.org/abs/1608.06879.

Elad Romanov, Fangzhao Zhang, and Mert Pilanci. Newton meets Marchenko–Pastur: Massively parallel second-order optimization with hessian sketching and debiasing. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ed9dcde1eb9c597f68c1d375bbecf3fc-Abstract-Conference.html.

Farbod Roosta-Khorasani and Michael W. Mahoney. Sub-sampled Newton methods. *Mathematical Programming*, 174(1–2):293–326, 2019. doi: 10.1007/s10107-018-1346-5.

Ohad Shamir, Nati Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate Newton-type method. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1000–1008. PMLR, 2014. URL https://proceedings.mlr.press/v32/shamir14.html.

Jack W. Silverstein. Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *Journal of Multivariate Analysis*, 55(2):331–339, 1995. doi: 10.1006/jmva.1995.1083.

Jack W. Silverstein and Sang-Il Choi. Analysis of the limiting spectral distribution of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):295–309, 1995. doi: 10.1006/jmva.1995.1058.

Alexander Tsigler and Peter L. Bartlett. Benign overfitting in ridge regression. *Journal of Machine Learning Research*, 24(123):1–76, 2023. URL https://jmlr.org/papers/v24/22-1398.html.

Shusen Wang, Fred Roosta, Peng Xu, and Michael W. Mahoney. GIANT: Globally improved approximate Newton method for distributed optimization. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Fangzhao Zhang and Mert Pilanci. Optimal shrinkage for distributed second-order optimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 41523–41549. PMLR, 2023. URL https://proceedings.mlr.press/v202/zhang23ah.html.

Yuchen Zhang and Xiao Lin. DiSCO: Distributed optimization for self-concordant empirical loss. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 362–370. PMLR, 2015. URL https://proceedings.mlr.press/v37/zhangb15.html.

Yuchen Zhang, John C. Duchi, and Martin J. Wainwright. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14:3321–3363, 2013. URL https://jmlr.org/papers/v14/zhang13b.html.

Peng Zhao. Beyond negative-ridge endpoints: Mixed-sign spectral regularization via negative-shifted gradient descent. *arXiv preprint arXiv:2607.22474*, 2026. doi: 10.48550/arXiv.2607.22474. URL https://arxiv.org/abs/2607.22474.

A Extended Related Work

Randomized and distributed second-order methods. Newton sketch and iterative Hessian sketch replace a Hessian or its square root with a randomized low-dimensional approximation (Pilanci & Wainwright, 2017; 2016); subsampled Newton, LiSSA and randomized subspace Newton are related routes (Erdogdu & Montanari, 2015; Roosta-Khorasani & Mahoney, 2019; Bollapragada et al., 2019; Berahas et al., 2020; Agarwal et al., 2017; Gower et al., 2019; Lacotte & Pilanci, 2022). Hessian-free inexact Newton methods spend Krylov depth (Dembo et al., 1982; Martens, 2010), whereas our setting spends parallel width: each of q workers computes an m-row Hessian sketch before one aggregation. DANE, AIDE, DiSCO, GIANT and DINGO aggregate local second-order information of sharded data (Shamir et al., 2014; Reddi et al., 2016; Zhang & Lin, 2015; Wang et al., 2018; Crane & Roosta, 2019), compressed Newton methods reduce their communication (Islamov et al., 2021), one-shot averaging of local estimators is analyzed by Zhang et al. (2013); Dobriban & Sheng (2021; 2020), and federated Newton-sketch methods pool sketched square-root Hessians at a server, with a sketch size matched to the effective dimension (Li et al., 2024; Gupta et al., 2024). We instead average directions from sketches below the effective dimension of a shared objective, the model of Dereziński et al. (2020); Zhang & Pilanci (2023); Romanov et al. (2025); Garg et al. (2024); exact bias characterizations of averaged sketched solutions are given by Bartan & Pilanci (2023). Methods that aggregate Hessian information across iterations (Na et al., 2023) enlarge the accumulated sketch, whereas FT modifies the spectral response produced within each round. Averaging directions also differs from pooling all qm sketch columns and solving over their span, which has different server-memory requirements.

Inversion bias, debiasing and deterministic equivalents. The inverse of a sketched matrix is biased even when the sketch is unbiased. Determinantal averaging removes this bias exactly (Dereziński & Mahoney, 2019), sparse sketches make it small (Dereziński et al., 2021), surrogate sketches with scaled regularization debias distributed Newton directions (Dereziński et al., 2020), and related work studies bias in sampled matrices (Niu et al., 2025) and deterministic shrinkage of the averaged direction (Zhang & Pilanci, 2023). These corrections select an endpoint of a shifted sketched inverse, and adaptive variants increase m until the effective-dimension condition holds. A deterministic scalar adjusts the overall magnitude but cannot remove the variation of the relative response across source-bearing eigenspaces; FT changes the response family itself. The endpoint equivalent used here, including its stable negative branch, is due to LeJeune et al. (2024), and ensembles of subsampled or sketched ridge estimators are equivalent to a ridge estimator with modified regularization (LeJeune et al., 2020; Du et al., 2023; Patil & Du, 2023; Patil & LeJeune, 2024). These results characterize the limiting mean response of endpoint ensembles: averaging reduces fluctuations around that mean but does not enlarge the family of mean responses. The sketched spectrum is described by the equation of Marchenko & Pastur (1967) (Silverstein & Choi, 1995; Bai & Silverstein, 2010), with eigenvector and anisotropic refinements (Ledoit & Péché, 2011; Knowles & Yin, 2017) and applications to ridge regression (Dobriban & Wager, 2018; Hastie et al., 2022), and spiked covariance models with the BBP transition (Johnstone, 2001; Baik et al., 2005) supply the structure of Appendix D: over a tail of high effective rank, a head eigenvalue appears in the sketched spectrum as an outlier displaced by the tail floor, with a mass defect of second order in the spread of the sketched tail.

Negative regularization and early stopping. The statistically optimal ridge parameter can be negative (Kobak et al., 2020; Wu & Xu, 2020; Tsigler & Bartlett, 2023). Early-stopped gradient flow acts as a spectral filter and tracks ridge risk along a regularization path (Ali et al., 2019). Negative-shifted gradient descent (NSGD) combines a negative shift with early stopping and thereby avoids the negative-ridge pole at finite time (Zhao, 2026). FT inherits this signed finite-time filtering. The

floor-critical level $\nu = a - \lambda$ of Theorem 3.2 certifies the path family, while the method tunes signed level and stopping time for the sketched Hessian. The new elements are its use on sketched Hessians, the scale-free Newton-energy criterion, the below-fold comparison with stable endpoint means, and the finite-worker analysis that separates the error of the mean direction from the $1/q$ variance.

B Assumptions and Selection Modes

B.1 Standing assumptions

The results use the following assumptions only where explicitly cited.

(A1) At the analyzed iterate, $H \succeq 0$, $\lambda > 0$, and $g \neq 0$, so $A = H + \lambda I \succ 0$. Nonlinear guarded-descent statements additionally assume that F is twice continuously differentiable, strongly convex, and smooth.

(A2) Worker sketches have iid entries $S_{ab} \sim \mathcal{N}(0, 1/m)$, are independent across workers, and are independent of any fixed policy being assessed. A policy selected from the same assessment pool is not covered by this independence condition.

(A3) For random-matrix limits, $d, m \to \infty$ with $d/m \to \gamma \in (1, \infty)$, and the spectrum of H_d satisfies conditions (G1)–(G3) of Appendix D: bounded norm, nonzero eigenvalues bounded away from zero, a limiting spectral distribution, and limiting rank above m.

(A4) For a strict endpoint separation, the source has positive Newton-energy weight on at least two distinct eigenvalues. The rate theorem takes a source on finitely many head eigenvalues that are outliers of the sketched tail, and $\lambda \leq \rho\mu_0$ with $\rho < 1$; its explicit rates also use tail eigenvalues at most μ_0 and $\mu_0/\theta_{\min} \leq 1/4$.

Discussion. (A1) makes each result a statement about one outer step on the local quadratic model; Corollary C.7 chains these steps on a quadratic objective, and on other objectives FT takes the same step with a line search, which Appendix H tests on logistic and softmax regression. Gaussian sketches in (A2) make the mean worker operator diagonal in the eigenbasis of H (proof of Proposition C.3). The finite-worker theorem, Theorem 3.3, needs neither (A3) nor (A4) and uses (A2) only through the independence of the workers: any iid positive-semidefinite worker operators with finite second moment and nonzero mean action on g suffice, for any sketch distribution and any policy fixed independently of its sketches. Policies selected on the same pool are treated in Appendix B.2. (A3) is the proportional regime of the imported equivalent, Theorem D.1, and finite sketches already follow it (Figures 4 and 6). The two distinct eigenvalues in (A4) are necessary: a source on one eigenvalue is matched by rescaling alone, so no floor exists. The condition $\lambda \leq \rho\mu_0$ places the problem below the fold, the regime this paper targets; the additional bounds $\bar{h} \leq \mu_0$ and $\mu_0/\theta_{\min} \leq 1/4$ give the explicit constants of part (iii) of Theorem 3.2. Fixed m with $d \to \infty$, and tails whose nonzero eigenvalues are not bounded away from zero, such as an infinite power law, lie outside that theorem (Appendix D.9).

B.2 Limit orders and selection modes

We distinguish three limit orders:

1. fixed d, m, followed by $q \to \infty$;
2. proportional $d, m \to \infty$ at fixed $d/m \to \gamma$;
3. a joint $d, m, q \to \infty$ limit in which the proportional mean response converges and the worker variance vanishes.

We also distinguish three policy-selection modes:

Independent pilot.
A disjoint worker pool selects one filter, which is frozen before assessment. The fixed-policy worker theorem applies conditionally on the pilot.

Same pool.
Candidate filters are selected using the workers that produce the reported direction. The fixed-policy iid theorem does not directly apply to the selected estimator.

Closed-loop adaptive.
Candidates are evaluated at each outer iteration. Exact model-based selection and line-search guarantees remain valid, but the fixed-policy $1/q$ theorem applies only when the current policy is independent of the current assessment sketches.

The fixed-policy worker theorem cannot be transferred to a filter selected on the same worker pool by conditioning on the selected parameters, because selection changes the conditional distribution of the worker directions. Two guarantees remain: exact evaluation of the local quadratic model selects the best candidate among those computed, and Armijo backtracking gives descent for an accepted step under its usual assumptions. The independent-pilot experiment of Appendix G isolates the fixed-policy setting, and the closed-loop experiments report same-pool adaptive selection as an empirical procedure outside that theorem. The budget $N = qm$ counts sketch rows or HVP columns; it is not equivalent to storing and solving over a pooled N-column basis at the server.

C Structural and Finite-Worker Results

C.1 Filter positivity and exposure

Proposition C.1 (Filter positivity, exposure, and rank)**.** *For $t > 0$, define*

$$h_{\alpha,t}(b) = \int_0^t e^{-(b+\alpha)s}\, ds,$$

with continuous value t at $b + \alpha = 0$. For $B \succeq 0$, let

$$E = t[-\lambda_{\min}(B) - \alpha]_+, \qquad M = S^\top h_{\alpha,t}(B) S.$$

Then

$$h_{\alpha,t}(b) > 0, \qquad 0 \preceq M \preceq t e^E S^\top S, \qquad \operatorname{rank}(M) \le m.$$

Proof. Put $c = b + \alpha$. The integral representation is positive for every $c \in \mathbb{R}$ and has value t at $c = 0$. For $0 \le s \le t$,

$$e^{-cs} \le e^{[-ct]_+},$$

so

$$0 < h_{\alpha,t}(b) \le t e^{[-ct]_+}.$$

Applying this scalar inequality in a spectral decomposition of B, then taking a congruence with S, gives the operator bounds. The rank bound follows because M factors through an m-dimensional space. □

The exposure E controls amplification caused by eigenvalues below the negative shift. It is a numerical safety parameter, not a claim that the shifted matrix $B + \alpha I$ is positive semidefinite.

C.2 Ray calibration

For

$$Q(d) = \frac{1}{2} d^\top A d - g^\top d, \qquad d_\star = A^{-1} g,$$

completing the square gives

$$Q(d) - Q(d_\star) = \frac{1}{2} \|d - d_\star\|_A^2. \tag{17}$$

For a nonzero direction d, the optimal nonnegative scalar is

$$c_Q(d) = \frac{[g^\top d]_+}{d^\top A d}. \tag{18}$$

When $g^\top d > 0$, substitution yields

$$\mathcal{R}_{\text{ray}}(d) = 1 - \frac{(g^\top d)^2}{(d^\top A d)(g^\top A^{-1} g)}. \tag{19}$$

For a quadratic objective,

$$F(x) - F(x^\star) = \frac{1}{2} \|x - x^\star\|_A^2.$$

Therefore, after optimizing the step along d, the next relative objective gap equals $\mathcal{R}_{\text{ray}}(d)$.

C.3 AUXILIARY NSGD REPRESENTATION

Proposition C.2 (Exact sample-covariance representation)**.** *Let*

$$C = SH^{1/2}, \qquad W = C^\top C, \qquad f_{\alpha,t}(b) = b\, h_{\alpha,t}(b).$$

For every $H \succeq 0$,

$$H^{1/2} M H^{1/2} = f_{\alpha,t}(W). \tag{20}$$

If $H \succ 0$, consider the auxiliary noiseless regression problem

$$X = \sqrt{m}\, C, \qquad \beta_0 = H^{-1/2} g, \qquad y = X\beta_0.$$

The negative-shifted gradient-flow estimate satisfies

$$\widehat{\beta} = f_{\alpha,t}(W)\beta_0 = H^{1/2} M g.$$

With $\beta_\lambda = H^{1/2} A^{-1} g$,

$$\|Mg - A^{-1}g\|_A^2 = \|\widehat{\beta} - \beta_\lambda\|_{I+\lambda H^{-1}}^2. \tag{21}$$

Proof. An SVD of C gives

$$C^\top h_{\alpha,t}(CC^\top) C = (C^\top C) h_{\alpha,t}(C^\top C).$$

Zero singular values contribute zero because $f_{\alpha,t}(0) = 0$, proving (20). Multiplication by $H^{-1/2} g$ gives the auxiliary estimator. If $z = Mg - A^{-1}g$, then

$$\widehat{\beta} - \beta_\lambda = H^{1/2} z$$

and

$$H^{1/2}(I + \lambda H^{-1}) H^{1/2} = H + \lambda I = A.$$

This proves (21). □

The identity transfers the NSGD filter exactly, but not every NSGD risk theorem. In particular, Newton-energy loss corresponds to a weighted coefficient error rather than the auxiliary prediction error. The auxiliary sample size is m per worker; the product qm enters only through worker averaging.

C.4 SCALE-FREE MEAN RESPONSE

Proposition C.3 (Loss of the mean direction)**.** *For a positive-semidefinite spectral worker rule selected independently of the sketches on which it is assessed, with nonzero mean action, the relative mean responses satisfy (6). Consequently, $R_\infty = 0$ exactly when the relative responses are constant on the source-bearing eigenvalues.*

Proof of Proposition C.3. Work in an eigenbasis of H. For every diagonal sign matrix D, the Gaussian sketches S and SD have the same distribution, while

$$SDHDS^\top = SHS^\top.$$

A fixed spectral worker rule therefore satisfies

$$M(SD) = DM(S)D.$$

Taking expectations removes off-diagonal entries of $\mathbb{E}M$. Orthogonal invariance makes the mean scalar within repeated-eigenvalue blocks.

Let

$$T = A^{1/2} M A^{1/2}, \qquad v = \frac{A^{-1/2} g}{\sqrt{g^\top A^{-1} g}}, \qquad w_i = v_i^2.$$

Then $\sum_i w_i = 1$. Write

$$\mathbb{E}T = \operatorname{diag}(r_1, \ldots, r_d).$$

The ray loss of the mean worker direction is

$$R_\infty = 1 - \frac{(v^\top \mathbb{E}Tv)^2}{\|\mathbb{E}Tv\|^2} = 1 - \frac{(\sum_i w_i r_i)^2}{\sum_i w_i r_i^2}.$$

Consequently,

$$R_\infty = \frac{\mathrm{Var}_w(r)}{\mathbb{E}_w[r^2]}. \tag{22}$$

Thus $R_\infty = 0$ exactly when the relative response is constant on all source-bearing eigenspaces. □

C.5 Stable-endpoint barrier

Proof of Proposition 3.1(i). The sign of the conventional correction follows directly from

$$\ell(\lambda) = \lambda \left(1 - \frac{d_1(\lambda)}{m}\right).$$

For the stable endpoint branch, LeJeune et al. (2024) identify the limiting mean action as

$$(H + \kappa I)^{-1}, \qquad \alpha = \ell(\kappa).$$

Differentiating the fixed-point map gives

$$\ell'(\mu) = 1 - \frac{d_2(\mu)}{m}.$$

Since d_2 is decreasing, the increasing branch is $\kappa > \mu_0$, where $d_2(\mu_0) = m$. The target $\kappa = \lambda$ is on this branch exactly when $m > d_2(\lambda)$. If $m < d_2(\lambda)$, every stable endpoint has $\kappa > \mu_0 > \lambda$. □

Proof of Proposition 3.1(ii). For a stable endpoint, define

$$r_i(\kappa) = \frac{h_i + \lambda}{h_i + \kappa}, \qquad J_{\mathrm{end}}(\kappa) = 1 - \frac{(\mathbb{E}_w r)^2}{\mathbb{E}_w r^2}.$$

Let

$$a = \mathbb{E}_w r, \qquad b_2 = \mathbb{E}_w r^2, \qquad c_3 = \mathbb{E}_w r^3, \qquad \Delta = \kappa - \lambda.$$

Direct differentiation gives

$$r_i' = \frac{r_i^2 - r_i}{\Delta}, \qquad a' = \frac{b_2 - a}{\Delta}, \qquad b_2' = \frac{2(c_3 - b_2)}{\Delta}.$$

Therefore,

$$J_{\mathrm{end}}'(\kappa) = \frac{2a(ac_3 - b_2^2)}{\Delta b_2^2} \geq 0,$$

where the inequality follows from

$$(\mathbb{E}_w r^2)^2 \leq (\mathbb{E}_w r)(\mathbb{E}_w r^3).$$

Hence

$$\inf_{\kappa > \mu_0} J_{\mathrm{end}}(\kappa) = J_{\mathrm{end}}(\mu_0).$$

If the source assigns positive weight to two distinct eigenvalues, their responses at μ_0 are distinct and the weighted variance is positive. For a source supported on one eigenvalue, the calibrated loss is zero. □

C.6 Endpoint mixtures and scalar corrections

Remark C.4 (Scalar corrections and endpoint mixtures). Multiplying an averaged direction by a positive deterministic scalar leaves its ray loss unchanged. Scalar shrinkage or bias correction therefore cannot eliminate a nonconstant relative response.

The same conclusion holds for positive mixtures of stable endpoint means. Suppose

$$\overline{M} = \sum_k a_k (H + \kappa_k I)^{-1}, \qquad a_k > 0, \qquad \kappa_k > \mu_0 > \lambda.$$

Its relative response at eigenvalue h is

$$r(h) = \sum_k a_k \frac{h+\lambda}{h+\kappa_k}.$$

Since

$$r'(h) = \sum_k a_k \frac{\kappa_k - \lambda}{(h+\kappa_k)^2} > 0,$$

the response is nonconstant across distinct source-bearing eigenvalues. Positive averaging over shifts, regularizations, or sketch sizes therefore retains positive mean-direction loss below the fold.

C.7 Finite-worker theorem

Proof of Theorem 3.3. Let

$$v = \frac{A^{-1/2}g}{\sqrt{g^\top A^{-1} g}}, \qquad u_j = T_j v, \qquad T_j = A^{1/2} M_j A^{1/2},$$

and define

$$\mu = \mathbb{E} u_j, \qquad \sigma^2 = \mathbb{E}\|u_j - \mu\|^2, \qquad \overline{u}_q = \frac{1}{q}\sum_{j=1}^{q} u_j.$$

Positive semidefiniteness and $\mu \neq 0$ imply $a := v^\top \mu > 0$. Independence gives

$$\mathbb{E}\overline{u}_q = \mu, \qquad \mathbb{E}\|\overline{u}_q\|^2 = \|\mu\|^2 + \frac{\sigma^2}{q}.$$

The realization-wise optimal scalar performs no worse in expectation than the deterministic scalar

$$c_0 = \frac{a}{\|\mu\|^2 + \sigma^2/q}.$$

Expanding at c_0 yields

$$\mathbb{E}R_q \leq 1 - \frac{a^2}{\|\mu\|^2 + \sigma^2/q}.$$

With

$$R_\infty = 1 - \frac{a^2}{\|\mu\|^2}, \qquad \omega = \frac{\sigma^2}{\|\mu\|^2},$$

we obtain

$$\mathbb{E}R_q \leq \frac{R_\infty + \omega/q}{1+\omega/q} \leq R_\infty + \frac{\omega}{q}. \tag{23}$$

For the two-sided estimate, define

$$\Phi(u) = \begin{cases} (v^\top u)^2/\|u\|^2, & u \neq 0, \\ 0, & u = 0. \end{cases}$$

At a nonzero point, $\|\nabla\Phi(u)\| \leq 1/\|u\|$ and $\|\nabla^2\Phi(u)\| \leq 2/\|u\|^2$. To verify the latter, write $u = \rho n$ with $\|n\| = 1$ and $v = an + bw$ with $w \perp n$ and $a^2 + b^2 = 1$. On the span of n and w, the Hessian of Φ is

$$\frac{2}{\rho^2}\begin{pmatrix} 0 & -ab \\ -ab & b^2 - a^2 \end{pmatrix},$$

with eigenvalues $2b^2/\rho^2$ and $-2a^2/\rho^2$; directions orthogonal to this span have eigenvalue $-2a^2/\rho^2$, and degenerate spans obey the same bound. If $\|u-\mu\| < \|\mu\|/2$, the segment from μ to u stays at norm at least $\|\mu\|/2$, so the Taylor remainder is at most $4\|u-\mu\|^2/\|\mu\|^2$. Otherwise, with $s = \|u-\mu\|/\|\mu\| \geq 1/2$, the bounds $0 \leq \Phi \leq 1$ and $\|\nabla\Phi(\mu)\| \leq 1/\|\mu\|$ bound the remainder by $1 + s \leq 6s^2$. Hence, for every u, including $u = 0$,

$$\left|\Phi(u) - \Phi(\mu) - \nabla\Phi(\mu)^\top(u-\mu)\right| \leq 6\frac{\|u-\mu\|^2}{\|\mu\|^2}.$$

Since $v^\top \overline{u}_q \geq 0$, we have $R_q = 1 - \Phi(\overline{u}_q)$ and $R_\infty = 1 - \Phi(\mu)$. Applying the estimate to $u = \overline{u}_q$, taking expectations, in which the linear term vanishes, and using

$$\mathbb{E}\|\overline{u}_q - \mu\|^2 = \frac{\sigma^2}{q}$$

gives

$$|\mathbb{E}R_q - R_\infty| \leq \frac{6\omega}{q}. \tag{24}$$

□

The theorem requires finite second moments and a fixed worker policy independent of the assessment sketches. It does not require a proportional random-matrix limit. The quantity R_∞ measures the error of the mean direction, whereas ω/q measures finite-worker fluctuation.

C.8 Worker crossover and fixed-budget consequences

Corollary C.5 (Sufficient worker crossover)**.** *For fixed FT and endpoint policies, suppose*

$$\Delta = R_{\infty,\mathrm{end}} - R_{\infty,\mathrm{FT}} > 0.$$

Then FT has lower expected ray loss whenever

$$q > \frac{\omega_{\mathrm{FT}} + 6\omega_{\mathrm{end}}}{\Delta}. \tag{25}$$

Proof. The upper half of Theorem 3.3 gives

$$\mathbb{E}R_{q,\mathrm{FT}} \leq R_{\infty,\mathrm{FT}} + \frac{\omega_{\mathrm{FT}}}{q},$$

while the two-sided estimate gives

$$\mathbb{E}R_{q,\mathrm{end}} \geq R_{\infty,\mathrm{end}} - \frac{6\omega_{\mathrm{end}}}{q}.$$

The claimed condition follows by subtraction. □

The condition is sufficient rather than necessary, and the constant six makes it conservative. Positive deterministic scaling changes neither R_∞ nor ω.

Proposition C.6 (Threshold-free fixed-budget bound)**.** *Fix N and use $q = N/m$ workers. Suppose*

$$R_{\infty,\mathrm{FT}}(m) \leq \varepsilon, \qquad \omega_{\mathrm{FT}}(m) \leq \overline{\omega}.$$

Then

$$\mathbb{E}R_{N/m,\mathrm{FT}}(m) \leq \varepsilon + \overline{\omega}\frac{m}{N}. \tag{26}$$

If $m < d_2(\lambda)$ and an endpoint policy satisfies

$$R_{\infty,\mathrm{end}}(m) \geq J_{\mathrm{end}}(\mu_0(m)) - \delta_m,$$

then

$$\mathbb{E}R_{N/m,\mathrm{end}}(m) \geq J_{\mathrm{end}}(\mu_0(m)) - \delta_m - 6\omega_{\mathrm{end}}(m)\frac{m}{N}. \tag{27}$$

The allowance $\delta_m \geq 0$ bounds the finite-dimensional endpoint's possible shortfall below the limiting mean-response floor, as required by the displayed premise.

Proof. Substitute $q = N/m$ into (23) and (24). The endpoint floor follows from Proposition 3.1. No condition of the form $m > d_1(\lambda)$ is needed for the FT upper bound. □

The bound does not imply that smaller m is always preferable. Both $R_\infty(m)$ and $\omega(m)$ depend on the local sketch size, and the experiments show that larger sketches can reduce both sufficiently to offset the smaller worker count N/m.

C.9 Quadratic contraction

Corollary C.7 (Worker-improving contraction on quadratics)**.** *Let F be quadratic with Hessian A, and let*

$$\Gamma_k = F(x_k) - F(x^\star).$$

At round k, select a positive semidefinite worker policy independently of the current sketches, average q worker directions, and apply the ray-optimal scalar. Then

$$\mathbb{E}[\Gamma_{k+1} \mid x_k] \le \frac{R_\infty(w^{(k)}) + \omega(w^{(k)})/q}{1 + \omega(w^{(k)})/q}\Gamma_k. \tag{28}$$

If $R_\infty \le \overline{R} < 1$ and $\omega \le \overline{\omega} < \infty$ uniformly along the iterates, then

$$\mathbb{E}\Gamma_K \le \left(\frac{\overline{R} + \overline{\omega}/q}{1 + \overline{\omega}/q}\right)^K \Gamma_0.$$

Proof. For a quadratic objective,

$$\Gamma_k = \frac{1}{2} g_k^\top A^{-1} g_k.$$

Equation (17) shows that the relative next-step gap after ray calibration is exactly the realized ray loss:

$$\Gamma_{k+1} = R_q\, \Gamma_k.$$

Conditionally on x_k, the workers are iid under the selected fixed policy. Applying Theorem 3.3 gives (28); the uniform bound then follows by the tower property. □

Below the fold, a stable endpoint may have a positive source-dependent limit R_∞, so increasing q cannot improve its contraction beyond that floor. For FT at the pair of Theorem 3.2, R_∞ is of the smaller order $\zeta^4 \log^2(1/\zeta)$ in the proportional limit, besides finite-worker and finite-dimensional error.

D Rate Separation over a High-Effective-Rank Tail

D.1 Setting and limiting quantities

Throughout this appendix $H_d \succeq 0$ is a sequence of $d \times d$ Hessians and S has iid $\mathcal{N}(0, 1/m)$ entries; the sketches for different d are defined on one probability space. We assume:

- **(G1)** $\|H_d\| \le C_H$, and the nonzero eigenvalues of H_d are at least $c_H > 0$, uniformly in d.
- **(G2)** $d/m \to \gamma \in (1, \infty)$.
- **(G3)** The empirical spectral distribution of H_d converges weakly to a probability measure τ_H with $\gamma\, \tau_H((0, \infty)) > 1$.

Conditions (G1) and (G2) are the hypotheses of the imported equivalent (Theorem D.1), and (G3) fixes the limit; its inequality says that the limiting rank exceeds the sketch size. The *tail* is τ_H. The *head* consists of finitely many eigenvalues of H_d outside the support of τ_H, with limits in $[\theta_{\min}, \theta_{\max}]$; they do not change any limit below. For $u > 0$ and $k = 1, 2$ let

$$D_k(u) = \gamma \int \left(\frac{h}{h+u}\right)^k d\tau_H(h) = \lim_{d\to\infty} \frac{d_k(u)}{m}, \qquad \ell_\infty(u) = u\{1 - D_1(u)\}. \tag{29}$$

The limit holds by (G2) and (G3) because the integrand is bounded and continuous. The limiting fold is the unique $\mu_0 > 0$ with $D_2(\mu_0) = 1$: it exists because D_2 decreases strictly from $\gamma\, \tau_H((0, \infty)) > 1$ to zero. The finite-d folds of Section 3.1 converge to it, because $d_2(u)/m \to D_2(u)$ pointwise, both sides decrease in u, and D_2 is continuous and strictly decreasing. Put

$$a_- = -\ell_\infty(\mu_0) = \mu_0\{D_1(\mu_0) - 1\}, \qquad Z(v) = -\frac{1}{v} + \gamma \int \frac{h}{1 + hv}\, d\tau_H(h). \tag{30}$$

Since $0 < h/(h+\mu_0) < 1$ on the positive part of the support, $D_1(\mu_0) > D_2(\mu_0) = 1$, so $a_- > 0$. The function Z is analytic away from 0 and the points $-1/h$ with $h \in \operatorname{supp}\tau_H$, $h > 0$; $Z(1/u) = -\ell_\infty(u)$; and, for $v > 0$,

$$Z'(v) = v^{-2}\{1 - D_2(1/v)\}. \tag{31}$$

The quantities of the rate theorem are the tail floor, the largest tail eigenvalue, and, for $\theta > \bar{h}$, the mass and the displacement defects of a head eigenvalue:

$$\begin{aligned} a &= \gamma \int h \, d\tau_H(h), \qquad \bar{h} = \sup \operatorname{supp} \tau_H, \\ m_\theta &= \gamma \int \frac{h^2}{(\theta - h)^2} \, d\tau_H(h), \qquad \varepsilon_\theta = \gamma \int \frac{h^2}{\theta - h} \, d\tau_H(h). \end{aligned} \tag{32}$$

The tail floor a is the limit of $\operatorname{tr}(H_d)/m$. Since $D_2(\mu_0) = 1$ and $\mu_0 \le h + \mu_0 \le \bar{h} + \mu_0$,

$$\mu_0^2 \le \gamma \int h^2 \, d\tau_H(h) \le (\mu_0 + \bar{h})^2. \tag{33}$$

The middle term is the variance of the limiting spectrum μ_B of the sketched tail, whose mean is a and whose second moment is $\gamma \int h^2 d\tau_H + a^2$. Thus μ_0 measures the spread of the sketched tail around the floor.

Flat tail. For $H_d = \operatorname{diag}(\theta_1, \dots, \theta_n, b, \dots, b)$ with fixed n and $b > 0$, $\tau_H = \delta_b$, $a = \gamma b$ and $D_k(u) = \gamma\{b/(b+u)\}^k$, so that

$$\mu_0 = b(\sqrt{\gamma} - 1), \qquad a_- = b(\sqrt{\gamma} - 1)^2, \tag{34}$$

and $\bar{h} \le \mu_0$ holds for $\gamma \ge 4$. This is the model of the two-spike experiment of Section 5.1.

Mean responses. For a signed level ν and a stopping time t let $h_{-\nu,t}(x) = \int_0^t e^{-s(x-\nu)} \, ds$. For an eigenvalue $\theta_d \to \theta \ge 0$ of H_d with unit eigenvector e_θ, define

$$L_\theta(\nu, t) = \lim_{d \to \infty} \mathbb{E}\left[e_\theta^\top S^\top h_{-\nu,t}(S H_d S^\top) S e_\theta\right], \qquad r_\theta(t) = (\theta + \lambda) L_\theta(\nu, t). \tag{35}$$

The limit exists by Lemma D.3, and the dependence of $r_\theta(t)$ on ν is suppressed. By orthogonal invariance the mean worker operator is a multiple of the identity on every eigenspace of H_d (proof of Proposition C.3), so the choice of e_θ within an eigenspace is immaterial, and $r_\theta(t)$ is the limiting relative Newton response on the eigenvalue θ. Sources are taken with finitely many source-bearing eigenvalues whose limits and Newton-energy weights do not depend on d, and $R_{\infty,\mathrm{FT}}(\nu, t)$ denotes the mean-direction loss (22) of the responses $r_\theta(t)$ under these weights.

D.2 Imported stable-endpoint equivalent

The stable endpoint branch used in the comparison is due to LeJeune et al. (2024). We restate the two results we use in our notation: their $(p, q, \mathbf{A}, \mathbf{S}, \lambda, \mu, \lambda_0, \mu_0)$ are our $(d, m, H, S^\top, \alpha, \kappa, \alpha_0, \mu_0)$, and their worker-sized symbol q is not our number of workers.

Theorem D.1 (Imported: isotropic sketching equivalence and monotonicity; LeJeune et al. (2024, Theorem 4.1 and Proposition 5.3))**.** *Let $H \succeq 0$ have $\|H\|_{\mathrm{op}}$ uniformly bounded in d and $\liminf \lambda_{\min}^+(H) > 0$, and let $\sqrt{m}\, S^\top \in \mathbb{R}^{d \times m}$ have iid entries with mean zero, variance one and a finite $8 + \delta$ moment. Let $\mu_0 > -\lambda_{\min}^+(H)$ and α_0 be the unique solutions of $d_2(\mu_0) = m$ and $\alpha_0 = \ell(\mu_0)$. As $d, m \to \infty$ with $0 < \liminf m/d \le \limsup m/d < \infty$:*

(i) for every $\alpha > \limsup \alpha_0$, $H^{1/2} S^\top (SHS^\top + \alpha I_m)^{-1} S \simeq H^{1/2}(H + \kappa I_d)^{-1}$;

(ii) if in addition $\alpha \neq 0$ or $\limsup m/d < \liminf \operatorname{rank}(H)/d$, then $S^\top (SHS^\top + \alpha I_m)^{-1} S \simeq (H + \kappa I_d)^{-1}$, where κ is the unique solution in (μ_0, ∞) of $\alpha = \ell(\kappa)$; moreover $\kappa \simeq 1/\tilde{v}(\alpha)$ with $\tilde{v}(\alpha) = m^{-1} \operatorname{tr}[(SHS^\top + \alpha I_m)^{-1}]$, and $\alpha_0 \simeq -\lambda_{\min}^+(SHS^\top)$;

(iii) for a fixed aspect ratio, $\alpha \mapsto \kappa(\alpha)$ is increasing on (α_0, ∞), with $\kappa(\alpha) \to \mu_0$ as $\alpha \downarrow \alpha_0$ and $\kappa(\alpha) \to \infty$ as $\alpha \to \infty$.

Here $\simeq$ is their asymptotic equivalence: for deterministic test matrices Θ_d of bounded trace norm,

$$A_d \simeq B_d \quad \Longleftrightarrow \quad \operatorname{tr}[\Theta_d(A_d - B_d)] \to 0 \text{ almost surely for every such } \Theta_d.$$

In the theorem, μ_0, α_0 and κ are the finite-d quantities of Section 3.1; under (G2) and (G3) they converge to the limits denoted by the same symbols, the limiting fold μ_0 above and $\kappa(\alpha)$ of Lemma D.2, with $\alpha_0 \to -a_-$.

Correspondence of the defining equations. Their defining equations for the fold and the pole,

$$1 = \frac{1}{q}\operatorname{tr}\big[\mathbf{A}^2(\mathbf{A}+\mu_0\mathbf{I})^{-2}\big], \qquad \lambda_0 = \mu_0\Big(1 - \frac{1}{q}\operatorname{tr}\big[\mathbf{A}(\mathbf{A}+\mu_0\mathbf{I})^{-1}\big]\Big),$$

read $d_2(\mu_0) = m$ and $\alpha_0 = \ell(\mu_0)$ in our notation, and their fixed-point equation reads $\alpha = \ell(\kappa)$. Part (i) is their (4.2), part (ii) their (4.3)–(4.4), and part (iii) the first clause of their Proposition 5.3. Gaussian entries have moments of every order, and (G1) is their spectral hypothesis.

The imported result identifies the endpoint response family. The scale-free loss, its minimization over this family, and the finite-time crossing are derived in this paper. The equivalent is a statement about bilinear forms; it has no operator-norm or unguarded finite-sample counterpart, and it does not interchange a finite-dimensional infimum over data-dependent shifts with the high-dimensional limit.

Gaussian bounds used below. Let $G = \sqrt{m}\,S$ and $d/m \le C_\gamma$. Quarter-nets of the two unit spheres have at most 9^m and 9^d points, and approximating both vectors in the variational expression for $\|G\|$ gives

$$\|G\| \le 2 \max_{u,v} |u^\top G v|$$

over these nets. Each tested scalar is standard normal, so the Gaussian tail and a union bound yield

$$\Pr\Big\{\|S\| > 2\sqrt{2(1+C_\gamma)\log 9} + 2s\Big\} \le 2e^{-ms^2/2}, \qquad s \ge 0.$$

Three consequences are used below:

$$\sup_d \mathbb{E}\|S\|^4 < \infty, \qquad \|S\| = O(1) \text{ almost surely}, \qquad \|Se_j\|^2 \to 1 \text{ almost surely}.$$

The first follows by tail integration, with a constant depending only on C_γ; the second from Borel–Cantelli, because for fixed positive s the probabilities are summable along the sequence $m \asymp d$; the third in the same way from chi-square Chernoff tails.

D.3 The sketched spectrum below its lower edge

Lemma D.2 (Sketched spectrum and its lower edge)**.** *Assume (G1)–(G3). The empirical spectral distribution of $SH_dS^\top$ converges almost surely to a probability measure μ_B with compact support, whose Stieltjes transform $v(z) = \int d\mu_B(x)/(x-z)$ is, for every z in the upper half-plane, the unique solution of $Z(v) = z$ in the upper half-plane. Moreover $a_- = \inf \operatorname{supp} \mu_B$; on $(-\infty, a_-)$, v is real and increasing with values in $(0, 1/\mu_0)$, $Z(v(x)) = x$, and $v(x) \uparrow 1/\mu_0$ as $x \uparrow a_-$; and for $\alpha > 0$, $v(-\alpha) = 1/\kappa(\alpha)$, where $\kappa(\alpha) > \mu_0$ solves $\alpha = \ell_\infty(\kappa)$.*

Proof. Limit law. The matrix $SH_dS^\top = m^{-1}GH_dG^\top$, with G an $m \times d$ standard Gaussian matrix, is a sample covariance matrix in companion form. Under (G2) and (G3) its empirical spectral distribution converges almost surely to a probability measure μ_B whose Stieltjes transform is the unique solution in the upper half-plane of the equation $Z(v) = z$ of Marchenko & Pastur (1967) (Silverstein, 1995). This uses only the weak limit of the spectrum of H_d, so the head does not matter. The support is compact because $\|SH_dS^\top\| \le C_H\|S\|^2$.

The negative axis. For $x < 0$, $v(x) > 0$ because μ_B is carried by $[0, \infty)$, and $Z(v(x)) = x$ by analytic continuation. By (31),

$$Z' > 0 \text{ on } (0, 1/\mu_0), \qquad Z'(1/\mu_0) = 0, \qquad Z(v) \to -\infty \text{ as } v \downarrow 0,$$

where $D_2(1/v) < 1$ on $(0, 1/\mu_0)$, and $Z(1/\mu_0) = -\ell_\infty(\mu_0) = a_-$; hence Z is an increasing bijection from $(0, 1/\mu_0)$ onto $(-\infty, a_-)$. The map $x \mapsto v(x)$ is continuous on $(-\infty, 0)$ and tends to

zero at $-\infty$. If it reached $1/\mu_0$ at some $x < 0$, then $x = Z(1/\mu_0) = a_- > 0$, a contradiction; hence $v(x) \in (0, 1/\mu_0)$ and

$$v = Z^{-1} \quad \text{on } (-\infty, 0).$$

Continuation to the edge. This inverse is real-analytic on $(-\infty, a_-)$, so it extends to an analytic function $\hat{v}$ with $Z(\hat{v}(z)) = z$ on a complex neighbourhood of $(-\infty, a_-)$, which we take of the form $U = \{x + iy : x < a_-,\ |y| < r(x)\}$ with $r > 0$ continuous. The functions v and $\hat{v}$ agree on the negative axis, hence on the connected set $U \cap \{\operatorname{Re} z < 0\}$, hence on the connected set $U \cap \{\operatorname{Im} z > 0\}$. Therefore v continues analytically to U with real values on $(-\infty, a_-)$, and the Stieltjes inversion formula gives

$$\mu_B\big((-\infty, a_-)\big) = 0.$$

If a_- were not in the support, v would be analytic and real near a_- with $v(a_-) = 1/\mu_0$ and $Z(v(x)) = x$, and differentiation would give $Z'(1/\mu_0)v'(a_-) = 1$, which contradicts $Z'(1/\mu_0) = 0$. Hence $a_- = \inf \operatorname{supp} \mu_B$.

The stable branch. The identity

$$v(-\alpha) = \frac{1}{\kappa(\alpha)}, \qquad \alpha > 0,$$

is $Z(1/u) = -\ell_\infty(u)$ read on this branch, on which ℓ_∞ is increasing because $\ell'_\infty = 1 - D_2 > 0$ on (μ_0, ∞). □

D.4 Lifted spectral measure of an eigen-direction

Let e_θ be a unit eigenvector of H_d for the eigenvalue $\theta_d \to \theta \geq 0$, put $s_\theta = Se_\theta$, let E_B be the projection-valued spectral measure of $B = SH_dS^\top$, and define the positive measure $\omega_\theta^{(d)}(dx) = s_\theta^\top E_B(dx) s_\theta$. Its total mass is $\|s_\theta\|^2$, a random quantity with mean one.

Lemma D.3 (Lifted measure)**.** *Assume (G1)–(G3). For every real $\theta \geq 0$ there is a probability measure ω_θ with compact support in $[a_-, \infty)$ such that*

$$\int \frac{d\omega_\theta(x)}{x - z} = F_\theta(z) := \frac{v(z)}{1 + \theta v(z)}, \quad z \in \mathbb{C} \setminus [a_-, \infty), \qquad L_\theta(\nu, t) = \int h_{-\nu,t}(x)\, d\omega_\theta(x). \tag{36}$$

If θ is the limit of eigenvalues θ_d of H_d, the limit in (35) exists for every $\nu \in \mathbb{R}$ and $t > 0$ and is given by the second identity; for every other $\theta \geq 0$ the second identity defines L_θ. For an eigenvalue limit, ω_θ is the limit of the spectral measure of B at the vector Se_θ.

Proof. Eigenvalue limits. Let first θ be the limit of eigenvalues θ_d of H_d. The coupling of the sketches across d is immaterial for L_θ, which is a limit of expectations. For each fixed $\alpha > 0$, Theorem D.1(ii), tested against the deterministic matrix $e_\theta e_\theta^\top$ of trace norm one, gives

$$\int \frac{d\omega_\theta^{(d)}(x)}{x + \alpha} - \frac{1}{\theta_d + \kappa_d(\alpha)} \longrightarrow 0 \quad \text{almost surely,}$$

where $\kappa_d(\alpha)$ solves $\alpha = \kappa\{1 - d_1(\kappa)/m\}$ on the stable branch. For $\alpha > 0$ this is the only positive root, because the map $\kappa \mapsto \kappa\{1 - d_1(\kappa)/m\}$ is negative up to the finite-d fold and increasing beyond it.

Identification of the limit transform. By (29) these maps converge to ℓ_∞, locally uniformly because their derivatives $1 - d_2(\kappa)/m$ lie in $[1 - d/m, 1]$, and ℓ_∞ is strictly increasing on (μ_0, ∞), so $\kappa_d(\alpha) \to \kappa(\alpha)$. With Lemma D.2,

$$\int \frac{d\omega_\theta^{(d)}(x)}{x + \alpha} \longrightarrow \frac{1}{\theta + \kappa(\alpha)} = F_\theta(-\alpha) \quad \text{almost surely, for every positive rational } \alpha,$$

on the intersection of these probability-one events.

Tightness and uniqueness. The total mass $\|s_\theta\|^2 \to 1$ almost surely, and the Gaussian bounds above give $\|S\| = O(1)$ almost surely, so $\|B\| \leq C_H \|S\|^2$ is eventually bounded and the measures are tight. Every subsequential weak limit has the Stieltjes transform F_θ on the negative rationals. The

function F_θ is analytic on $\mathbb{C} \setminus [a_-, \infty)$: off the real axis $\operatorname{Im} v \neq 0$, and on $(-\infty, a_-)$ we have $v > 0$, so $1 + \theta v$ does not vanish. The transform of a limit point is analytic off $[0, \infty)$, so by the identity theorem it equals F_θ there; since F_θ is real and analytic on $(0, a_-)$, Stieltjes inversion shows that the limit point is carried by $[a_-, \infty)$, and Stieltjes uniqueness identifies a single limit ω_θ, a probability measure.

Convergence of the response. For fixed ν and t, the filter $h_{-\nu,t}$ is continuous on $[0, \infty)$ and

$$0 < h_{-\nu,t}(x) \le t e^{\nu_+ t}, \qquad \nu_+ = \max\{\nu, 0\},$$

so its integral against $\omega_\theta^{(d)}$ converges almost surely. The bound $t e^{\nu_+ t} \|s_\theta\|^2$ has a uniformly bounded second moment, because $m\|s_\theta\|^2$ is chi-square with m degrees of freedom, which gives uniform integrability and the convergence of expectations.

General θ. For a general real $\theta \ge 0$, replace one eigenvalue of H_d by θ. This preserves (G1)–(G3) and the limit of the spectrum, hence μ_B, v and F_θ, and the argument above applied to the modified sequence produces ω_θ. □

D.5 The head above the sketched tail

Lemma D.4 (Tail below the fold)**.** *Assume (G1)–(G3). Then $a - a_- \ge 2\mu_0$. If $\bar{h} \le \mu_0$, then $a - a_- \le 3\mu_0$ and, for $\theta \ge 4\mu_0$,*

$$m_\theta \le \frac{64}{9}\Big(\frac{\mu_0}{\theta}\Big)^2 \le \frac{4}{9}, \qquad \varepsilon_\theta \le \frac{16}{3}\,\frac{\mu_0^2}{\theta}.$$

Proof. Since $\mu_0 D_1(\mu_0) = \gamma \int h\mu_0/(h+\mu_0)\, d\tau_H = a - \gamma \int h^2/(h+\mu_0)\, d\tau_H$,

$$a - a_- = \mu_0 + \gamma \int \frac{h^2}{h+\mu_0}\, d\tau_H = \mu_0 + \gamma \int \Big(\frac{h}{h+\mu_0}\Big)^2 (h+\mu_0)\, d\tau_H.$$

Since $h + \mu_0 \ge \mu_0$ and $D_2(\mu_0) = 1$, the last integral is at least μ_0; if $\bar{h} \le \mu_0$, then $h + \mu_0 \le 2\mu_0$ and it is at most $2\mu_0$. In the same way, if $\bar{h} \le \mu_0$,

$$m_\theta = \gamma \int \Big(\frac{h}{h+\mu_0}\Big)^2 \Big(\frac{h+\mu_0}{\theta - h}\Big)^2 d\tau_H \le \Big(\frac{2\mu_0}{\theta - \mu_0}\Big)^2, \qquad \varepsilon_\theta \le \frac{4\mu_0^2}{\theta - \mu_0},$$

and $\theta - \mu_0 \ge \frac{3}{4}\theta$. □

Lemma D.5 (Atom of the lifted measure)**.** *Assume (G1)–(G3), and let $\theta > \bar{h}$ with $m_\theta < 1$. Put $x_\theta = \theta + a + \varepsilon_\theta$. Then $x_\theta \notin \operatorname{supp} \mu_B$, $v(x_\theta) = -1/\theta$, and*

$$\omega_\theta = (1 - m_\theta)\, \delta_{x_\theta} + \omega_\theta^{\mathrm{rest}},$$

where $\omega_\theta^{\mathrm{rest}}$ is a positive measure of mass m_θ carried by $[a_-, \infty)$.

Proof. Put $v_\theta = -1/\theta \in (-1/\bar{h}, 0)$. The function Z is analytic near v_θ, and a direct computation gives

$$Z(v_\theta) = \theta + \gamma \int \frac{h\theta}{\theta - h}\, d\tau_H = \theta + a + \varepsilon_\theta = x_\theta,$$

$$Z'(v_\theta) = \theta^2 - \gamma \int \frac{h^2\theta^2}{(\theta - h)^2}\, d\tau_H = \theta^2(1 - m_\theta) > 0.$$

By the inverse function theorem there is an analytic $\hat{v}$ near x_θ with $Z(\hat{v}(z)) = z$, $\hat{v}(x_\theta) = v_\theta$, real on the real axis, and $\hat{v}'(x_\theta) = 1/Z'(v_\theta) > 0$. Hence $\operatorname{Im} \hat{v}(x + iy) = y\,\hat{v}'(x) + O(y^2) > 0$ for small $y > 0$ and x near x_θ. By the uniqueness in Lemma D.2, $v = \hat{v}$ there. Thus v continues analytically across the real axis near x_θ with real values; by Stieltjes inversion μ_B has no mass near x_θ, and $v(x_\theta) = v_\theta$. Near x_θ, $1 + \theta v(z) = \theta \hat{v}'(x_\theta)(z - x_\theta) + O((z - x_\theta)^2)$, so

$$F_\theta(z) = \frac{v_\theta}{\theta \hat{v}'(x_\theta)}\,\frac{1}{z - x_\theta} + O(1) = \frac{1 - m_\theta}{x_\theta - z} + O(1),$$

which is an atom of mass $1 - m_\theta$ at x_θ. Since ω_θ is a probability measure carried by $[a_-, \infty)$ (Lemma D.3), the rest has mass m_θ and the same carrier. □

D.6 Proof of the rate-separation theorem

Proof of Theorem 3.2. Step 1: endpoints. For $\kappa > \mu_0$ the endpoint response is

$$r(\theta;\kappa) = \frac{\theta+\lambda}{\theta+\kappa},$$

and $\inf_{\kappa>\mu_0} J_{\mathrm{end}}(\kappa) = J_{\mathrm{end}}(\mu_0)$ by Proposition 3.1(ii), applied with the limiting fold to the limiting mean action $(H+\kappa I)^{-1}$ of Proposition 3.1(i); for a negative shift this is the limit of the guarded mean of Section D.8, because the unguarded mean does not exist at finite d. At $\kappa = \mu_0$,

$$r_0(\theta) = 1 - \frac{\mu_0-\lambda}{\theta+\mu_0} \le 1, \qquad \text{so} \qquad \mathbb{E}_w r_0^2 \le 1,$$

and therefore

$$J_{\mathrm{end}}(\mu_0) \ge \mathrm{Var}_w(r_0) = (\mu_0-\lambda)^2 \,\mathrm{Var}_w\Big(\frac{1}{\theta+\mu_0}\Big) \ge (1-\rho)^2\zeta^2 V,$$

which is (8).

Step 2: the finite-time level. By Lemma D.4,

$$\nu - a_- = (a-a_-) - \lambda \ge 2\mu_0 - \lambda > 0,$$

so the level $\nu = a - \lambda$ lies above the lower edge a_- of the sketched spectrum and is available to no stable endpoint.

Step 3: atom and remainder. Every head eigenvalue satisfies $\theta > \bar{h}$ and $m_\theta < 1$ by hypothesis, so Lemma D.5 applies. Integrating the filter against $\omega_\theta = (1-m_\theta)\delta_{x_\theta} + \omega_\theta^{\mathrm{rest}}$, with $x_\theta - \nu = \theta+\lambda+\varepsilon_\theta$, gives

$$r_\theta(t) = (1-m_\theta)\,\frac{\theta+\lambda}{\theta+\lambda+\varepsilon_\theta}\,\big(1-e^{-(\theta+\lambda+\varepsilon_\theta)t}\big) + (\theta+\lambda)\int h_{-\nu,t}\,d\omega_\theta^{\mathrm{rest}}. \tag{37}$$

Step 4: the sandwich. The first term of (37) is a product of three factors in $[0,1]$, so

$$1 - m_\theta - \frac{\varepsilon_\theta}{\theta} - e^{-\theta t} \le (1-m_\theta)\,\frac{\theta+\lambda}{\theta+\lambda+\varepsilon_\theta}\,\big(1-e^{-(\theta+\lambda+\varepsilon_\theta)t}\big) \le 1.$$

The second term is nonnegative, and on the carrier $[a_-,\infty)$ of $\omega_\theta^{\mathrm{rest}}$ the filter is bounded by

$$0 \le h_{-\nu,t}(x) = \int_0^t e^{-(x-\nu)s}\,ds \le \int_0^t e^{(\nu-a_-)s}\,ds \le t\,e^{(\nu-a_-)t}, \qquad x \ge a_-,$$

so, since $\omega_\theta^{\mathrm{rest}}$ has mass m_θ, the second term is at most $(\theta+\lambda)m_\theta t e^{(\nu-a_-)t}$. Hence, for every $t>0$,

$$-\big(m_\theta + \varepsilon_\theta/\theta + e^{-\theta t}\big) \le r_\theta(t) - 1 \le (\theta+\lambda)\,m_\theta\,t\,e^{(\nu-a_-)t}, \tag{38}$$

which is (9).

Step 5: rates. Let now $\bar{h} \le \mu_0$ and $\zeta \le 1/4$, so that every head eigenvalue has $\theta \ge 4\mu_0$ and, by Lemma D.4, $\nu - a_- \le 3\mu_0$. The same lemma bounds the two defects in (38): since $\lambda/\theta \le \frac{1}{4}$,

$$m_\theta + \frac{\varepsilon_\theta}{\theta} \le \frac{112}{9}\Big(\frac{\mu_0}{\theta}\Big)^2 \le 13\zeta^2, \qquad (\theta+\lambda)\,m_\theta\,t \le \frac{5}{4}\cdot\frac{64}{9}\,\frac{\mu_0^2}{\theta}\,t \le 9\zeta^2\,\theta_{\min}t.$$

At the stopping time $\theta_{\min}t = 2\log(1/\zeta)$,

$$e^{-\theta t} \le e^{-\theta_{\min}t} = \zeta^2, \qquad (\nu-a_-)t \le 3\mu_0 t = 6\zeta\log(1/\zeta),$$

and (38) gives

$$\sup_\theta |r_\theta(t) - 1| \le 14\zeta^2 + 18\zeta^2\log(1/\zeta)\,e^{6\zeta\log(1/\zeta)} =: \delta_\zeta,$$

which is the explicit constant of part (iii), of order $\zeta^2\log(1/\zeta)$. Finally, when $\delta_\zeta < 1$,

$$\mathrm{Var}_w(r) \le \mathbb{E}_w(r-1)^2 \le \delta_\zeta^2, \qquad \mathbb{E}_w r^2 \ge (1-\delta_\zeta)^2,$$

and Proposition C.3 gives (10).

Step 6: expected ray loss. Column-sign symmetry makes $\mathbb{E}M$ exactly diagonal in the eigenbasis of H_d for every finite d, so a source on the head has no mean leakage into the rest of the spectrum. Put $A = H_d + \lambda I$, normalize $\xi_d = A^{-1/2}g$ to unit norm and put $u_d = A^{1/2}MA^{1/2}\xi_d$. The Gaussian bounds give $\sup_d \mathbb{E}\|S\|^4 < \infty$, and Proposition C.1 gives

$$\|u_d\| \le \|A\|\, te^{\nu t}\|S\|^2, \qquad \text{hence} \qquad \sup_d \mathbb{E}\|u_d\|^2 < \infty.$$

Moreover

$$\|\mathbb{E}u_d\|^2 \longrightarrow \sum_j w_j r_{\theta_j}(t)^2 > 0,$$

because every $L_\theta(\nu, t)$ is positive, so the variance ratio $\omega_d = \mathbb{E}\|u_d - \mathbb{E}u_d\|^2/\|\mathbb{E}u_d\|^2$ of Theorem 3.3 (not the lifted measure) is eventually bounded. The two-sided bound of that theorem and the convergence of the mean-direction loss,

$$|\mathbb{E}R_{d,q} - R_{d,\infty}| \le \frac{6\omega_d}{q}, \qquad R_{d,\infty} \longrightarrow R_{\infty,\mathrm{FT}}(\nu, t),$$

give the convergence for every $q(d) \to \infty$. This variance bound is numerically loose and gives no useful value of q. □

Each side of (38) drops one of two terms of opposite signs, the mass defect and the remainder, which partly cancel at moderate ζ. The passage to (10) further bounds a variance by a squared sup norm and uses the crude constants of Lemma D.4. The theorem is therefore a rate statement: with these constants (10) falls below (8) only for ζ of the order of 10^{-3} on the sources of Appendix F.3, which reports the exact limiting values.

D.7 Sources with tail energy

Corollary D.6 (Source with tail energy)**.** *In the setting of Theorem 3.2, let the source put Newton-energy weight $1 - \rho_T$ on the head, with conditional weights w, and $\rho_T < 1$ on finitely many eigenvalues $h \le \mu_0$. If $4\zeta \log(1/\zeta) e^{6\zeta \log(1/\zeta)} \le 1$, which holds for $\zeta \le 0.036$ and implies $\delta_\zeta < 1$, then*

$$R_{\infty,\mathrm{FT}}(\nu, t) \le \frac{\delta_\zeta^2 + \rho_T}{(1 - \rho_T)(1 - \delta_\zeta)^2}, \qquad \inf_{\kappa > \mu_0} J_{\mathrm{end}}(\kappa) \ge (1 - \rho_T)(1 - \rho)^2 \zeta^2 V.$$

The separation persists when ρ_T is small relative to ζ^2.

Proof. Tail responses stay in $[0, 1]$. For $h \le \mu_0$ the measure ω_h is a probability measure carried by $[a_-, \infty)$ (Lemma D.3), so the filter bound of Step 4 in the proof of Theorem 3.2 gives

$$0 \le r_h(t) \le (h + \lambda)\, t\, e^{(\nu - a_-)t} \le 2\mu_0 t\, e^{3\mu_0 t} = 4\zeta \log(1/\zeta)\, e^{6\zeta \log(1/\zeta)} \le 1,$$

hence $|r_h(t) - 1| \le 1$ on the tail, while $|r_\theta(t) - 1| \le \delta_\zeta$ on the head by the theorem.

Finite time. Splitting the source into its head part, of weight $1 - \rho_T$, and its tail part, of weight ρ_T,

$$\mathrm{Var}(r) \le \mathbb{E}(r - 1)^2 \le (1 - \rho_T)\delta_\zeta^2 + \rho_T, \qquad \mathbb{E}r^2 \ge (1 - \rho_T)(1 - \delta_\zeta)^2,$$

and the ratio of the two bounds is the first claim.

Endpoints. Proposition 3.1 applies to the full source, all endpoint responses at $\kappa = \mu_0$ are at most one, and the variance of a mixture is at least the weighted variance within the head, so

$$\inf_{\kappa > \mu_0} J_{\mathrm{end}}(\kappa) \ge \mathrm{Var}(r_0) \ge (1 - \rho_T)\, \mathrm{Var}_w(r_0) \ge (1 - \rho_T)(1 - \rho)^2 \zeta^2 V,$$

the last step being Step 1 of the proof of Theorem 3.2. □

D.8 Negative-endpoint safety and limit boundaries

A fixed negative endpoint may approach a random finite-dimensional eigenvalue. To define a safe fixed policy, choose a denominator margin $\eta > 0$ and use the negative endpoint only when

$$\lambda_{\min}(B) + \alpha \geq \eta.$$

Otherwise, use a fixed positive-shift fallback. The resulting worker operator is positive semidefinite and satisfies

$$\|M\| \leq \max\{\eta^{-1}, \lambda^{-1}\}\|S\|^2.$$

For a fixed shift $\alpha > -a_-$ and a margin $\eta < a_- + \alpha$, the fallback probability vanishes, because $\lambda_{\min}(B) \to a_-$ almost surely by Theorem D.1(ii). On the safe event the bilinear forms of the guarded operator converge almost surely to those of $(H + \kappa(\alpha)I)^{-1}$ by the same theorem, and the norm bound with $\sup_d \mathbb{E}\|S\|^4 < \infty$ makes them uniformly integrable, so the guarded mean converges to the stable endpoint equivalent.

This argument applies to each fixed safe endpoint. It is not uniform over shifts approaching the random edge and does not cover a shift selected on the assessment pool. In particular, for every finite Gaussian sketch and fixed $\nu > 0$,

$$\Pr\{\lambda_{\min}(B) \leq \nu\} > 0.$$

Consequently, an all-workers-safe event can vanish as $q \to \infty$ unless one controls the tail probability or specifies a worker-local fallback.

D.9 Scope of the rate-separation theorem

Parts (i)–(ii) of Theorem 3.2 apply to outlying heads over tails satisfying (G1)–(G3). Its explicit rate in part (iii) also uses $\bar{h} \leq \mu_0$ and $\zeta \leq 1/4$. Extensions beyond these results include:

- fixed m with $d \to \infty$;
- tails whose nonzero eigenvalues are not bounded away from zero, such as an infinite power law, for which the imported equivalent is not available;
- explicit rates for a head close to the spread of the sketched tail ($\zeta > 1/4$), where part (ii) still holds; a head whose size grows with d; a regularization at or above the fold, where a stable endpoint reaches the target;
- sharp constants: the explicit bound (10) is below one for $\zeta < 0.057$ and falls below the endpoint bound (8) only for ζ of the order of 10^{-3};
- filters selected adaptively on the assessment workers;
- unrestricted exposure-capped policies without fallback analysis;
- a uniform finite-dimensional comparison over near-edge shifts; or
- distributed wall-clock complexity.

The theorem concerns limiting mean responses. The experiments of Appendix F check the limit against finite sketches and report exact limiting values, starting with the two-spike instance of Section 5.1.

E Algorithm and Resource Model

E.1 Access model

Workers access the same objective and Hessian–vector product oracle $v \mapsto Hv$. This is a shared-objective model rather than a data-sharded local-Hessian model; a data-sharded implementation would need additional communication to evaluate global HVPs. With a worker memory budget M_w, local storage has the form $W(m, d) = \gamma_{\text{mem}} md + O(m^2) + W_{\text{scratch}} \leq M_w$. The strict local-cap experiments fix $m = 16$ while increasing q, and do not retain an unrestricted qm-column basis at the server.

Algorithm 1 Worker j: finite-time sketched directions

Require: gradient g, HVP oracle $v \mapsto Hv$, sketch size m

Draw $S_j \in \mathbb{R}^{m\times d}$ with iid $\mathcal{N}(0, 1/m)$ entries
Compute $K_j = HS_j^\top$ with m HVPs and form $B_j = S_jK_j$
Diagonalize $B_j = U_j \operatorname{diag}(b_{j,i})U_j^\top$; cache $L_j = S_j^\top U_j$ and $\widetilde{K}_j = K_jU_j$
Send the eigenvalues $\{b_{j,i}\}_{i\le m}$ to the server ▷ summary stage
Receive a candidate set $\mathcal{C}$ of pairs (α, t)
for each $(\alpha, t) \in \mathcal{C}$ **do**
    $a_j \leftarrow (h_{\alpha,t}(b_{j,1}), \dots, h_{\alpha,t}(b_{j,m}))^\top$
    $D_j(\alpha, t) \leftarrow L_j \operatorname{diag}(a_j)L_j^\top g$ and $HD_j(\alpha, t) \leftarrow \widetilde{K}_j \operatorname{diag}(a_j)L_j^\top g$
end for
Send $\{D_j(\alpha, t), HD_j(\alpha, t)\}_{(\alpha,t)\in\mathcal{C}}$ to the server

Algorithm 2 Server: calibrated finite-time sketched Newton under one policy (retune, safety reuse or quality-aware reuse)

Require: x_0, q workers, grid rule $\mathcal{G}(b_{\min}, r_{\mathrm{med}})$ of Section E.3, exposure cap $E_{\max}$, policy

for $k = 0, 1, 2, \dots$ **do**
    $g_k \leftarrow \nabla F(x_k)$; broadcast g_k; workers run lines 1–4 of Algorithm 1 on fresh sketches
    $b_{\min} \leftarrow \min_{j,i} b_{j,i}$ and $r_{\mathrm{med}} \leftarrow \max\{\mathrm{median}_{j,i}\, b_{j,i}, \lambda, 10^{-12}\}$
    if $k = 0$, or the policy is retune, or the stored pair (α, t) has exposure $t[-b_{\min} - \alpha]_+ > E_{\max}$ **then**
        $\mathcal{C} \leftarrow \mathcal{G}(b_{\min}, r_{\mathrm{med}})$ ▷ full grid; the last case is a safety refresh
    else if the policy is safety reuse **then**
        $\mathcal{C} \leftarrow \{(\alpha, t)\}$
    else ▷ quality-aware reuse
        $\mathcal{C} \leftarrow \mathcal{P}$, the stored pair and its probes $(\alpha, t/2)$, $(\alpha, 2t)$, $(\alpha \pm r_{\mathrm{med}}/4, t)$ and $(\lambda, 1/r_{\mathrm{med}})$
    end if
    Broadcast $\mathcal{C}$; average the returned directions, $\overline{D} = q^{-1}\sum_j D_j$ and $H\overline{D} = q^{-1}\sum_j HD_j$, for each candidate
    Score each candidate by $Q(c_Q\overline{D})$ with $c_Q = [g_k^\top \overline{D}]_+/\{\overline{D}^\top(H\overline{D} + \lambda\overline{D})\}$, or $c_Q = 1$ for FT without c_Q
    if this round evaluated $\mathcal{P}$, and either the stored pair is invalid or the best feasible probe improves its score by more than 10% of its absolute value **then**
        Repeat the last two steps with $\mathcal{C} \leftarrow \mathcal{G}(b_{\min}, r_{\mathrm{med}}) \cup \mathcal{P}$, so that the scored probes stay eligible ▷ quality refresh
    end if
    Select the candidate (α_k, t_k) with the smallest score among those with exposure at most $E_{\max}$, and store it
    $x_{k+1} \leftarrow x_k - s\, c_Q\overline{D}(\alpha_k, t_k)$, with Armijo backtracking on F from $s = 1$
end for

E.2 Worker and server algorithms

Algorithms 1 and 2 give one outer round as implemented. The candidate grid depends on two pooled summaries of the current sketches, so each round has a summary stage, in which every worker sends the m eigenvalues of its sketched Hessian, before the candidate stage. The safety check, the probes and any refresh act in the current round, before the direction is selected.

Near $b + \alpha = 0$, the filter is evaluated as $-\operatorname{expm1}(-(b + \alpha)t)/(b + \alpha)$, with limiting value t, to retain numerical accuracy.

E.3 Candidate grid and reuse

Let $b_{\min}$ be the smallest eigenvalue among the current worker small matrices, and let $r_{\mathrm{med}} = \max\{\mathrm{median}_{j,i}(b_{j,i}), \lambda, 10^{-12}\}$, the median taken over the eigenvalues of all current workers. The sampled-covariance, corn and power-law-tail head studies use 25 geometrically spaced stopping

times in $[0.03/r_{\mathrm{med}}, 128/r_{\mathrm{med}}]$, the nonnegative shifts $\{0, \lambda, 0.1r_{\mathrm{med}}, r_{\mathrm{med}}, 10r_{\mathrm{med}}\}$, and negative shifts parameterized through the exposure $E(\alpha, t) = t[-b_{\min} - \alpha]_+ \in \{0, 1.5, 3, 4.5, 6, 9, 12\}$, that is $\alpha = -b_{\min} - E/t$. This gives at most 300 candidates with $E_{\max} = 12$, for both retuning and reuse. The earlier nonlinear, broad-source fixed-Hessian and UCI studies use the 90-candidate grid: nine times in $[0.03/r_{\mathrm{med}}, 16/r_{\mathrm{med}}]$, the same five nonnegative shifts, and exposures through six. Duplicate and above-cap candidates are removed.

Full retuning scores this grid at every round. Safety reuse scores it at the first round and keeps the selected pair while its exposure on the current pool stays within the cap; otherwise the grid is refreshed in that round. Between full searches, quality-aware reuse scores at most six candidates per round (Algorithm 2); probe times are clipped to the study's search interval, duplicates and probes above the cap are removed, and every probe, duplicate evaluation and candidate transmission is charged. Quality-aware reuse keeps the previous shift and time, probes five nearby candidates, and searches the full grid again when a probe improves the model score by more than 10% or the stored pair exceeds the exposure cap. Full retuning and reuse are same-pool adaptive procedures (Appendix B).

E.4 Calibration and line search

For each averaged candidate $\overline{D}$, the server computes $c_Q(\overline{D}) = [g^\top \overline{D}]_+ / \{\overline{D}^\top (H + \lambda I)\overline{D}\}$, scores $Q(c_Q(\overline{D})\overline{D})$, and selects $p = c_Q(\overline{D})\overline{D}$. On a quadratic objective the unit update $x^+ = x - p$ is optimal along the selected ray. On a nonquadratic objective the implementation applies the Armijo backtracking of Appendix J to $F(x - sp)$ from $s = 1$; the same line search serves every sketched method. FT without c_Q sets the scalar to one: candidates are scored by $Q(\overline{D})$ and the step is $s\overline{D}$, again by backtracking from one.

E.5 Methods compared

The reference procedures retain their prescribed direction and step rules; the two matched controls isolate finite stopping and outer calibration (Section 4).

RZP, limited m. The debiased sketch of Romanov et al. (2025) at the same local sketch size as FT. Each worker solves $\hat{s}(-\hat{\lambda}) = 1/\lambda$ for a shift $\hat{\lambda} \in [5\lambda/12, \lambda]$, where $\hat{s}(-a) = m^{-1} \sum_i (b_{j,i} + a)^{-1}$ (their Algorithm 2), and returns the fallback $\hat{\lambda} = 5\lambda/12$ when the equation has no root in that interval. At $m = 16$ every worker returns the fallback at every round of every closed loop reported here. The step is their backtracking line search (their Algorithm 3).

RZP, adaptive. The code also implements their Algorithm 1, which doubles m until $\hat{s}(-5\lambda/12) > 1/\lambda$ and then applies Algorithm 2 on an independent sketch; all attempts are charged. Its resource-relaxed comparisons are available with the code.

Undebiased endpoint. The fixed endpoint with local shift $\alpha = \lambda$, with backtracking from step one and halving as in the experiments of Dereziński et al. (2020); the sufficient-decrease constant is the 10^{-4} used for every sketched method here.

Tuned stable endpoint (ours). Our extension of the endpoint family to the stable branch of LeJeune et al. (2024): the best of 65 positive shifts and 65 admissible negative shifts $\alpha = -b_{\min} + \delta$, with $\delta > 0$ on a geometric grid so that every current worker denominator stays positive. It scores 130 candidates per full search and is calibrated by c_Q like FT, which isolates finite stopping from shift tuning and common rescaling.

L-BFGS. SciPy L-BFGS-B (Liu & Nocedal, 1989) without bounds, with ten correction pairs and its own line search, one worker, no HVPs and no outer scalar. Its history holds $20d$ numbers, within the $2md = 32d$ numbers that a sketched worker needs at $m = 16$, so it satisfies the same memory cap. The code records gradient calls, HVP columns and candidate traffic separately from accepted iterations.

E.6 Complexity and communication

For C candidates, each worker performs m HVP columns to form $HS^\top$, one $O(m^3)$ eigendecomposition and $O(Cmd)$ candidate-application work, with $O(md)$ principal storage. If each worker returns both D_j and HD_j for every candidate, a double-precision communication model is $8qd(2C+1)$ bytes per round, including one gradient broadcast; the summary stage adds $8qm$ bytes, which is negligible. Candidate reuse reduces application and communication costs but not the m-HVP sketch construction.

Workers could instead send their factors L_j and $\widetilde{K}_j$, $2md$ numbers, and let the server apply every candidate, which is less traffic than $2Cd$ numbers whenever $C > m$, as under full retuning ($C = 300$, $m = 16$). This exchanges communication for $2qmd$ server storage and $O(Cqmd)$ serial application work. Our implementation keeps the factors and candidate application distributed across workers. Under parameter reuse without a full refresh, $C \le 6 < m$, so sending candidate directions has the smaller payload.

Each Armijo trial costs one objective evaluation at the server and each round one gradient evaluation. Both are recorded for every sketched method; the resource logs also identify the serial timing environment.

F Controlled Spectral Experiments

These experiments isolate the two parts of the proposed mechanism. The two-spike and three-tail studies test whether finite time changes the mean response left by stable endpoints; fixed-aspect-ratio scaling checks the proportional prediction. Appendix G then asks how many independent workers are needed to offset variance.

F.1 Two-spike protocol

The two-spike experiment isolates response matching on two head eigenvalues. It uses

$$b = 1, \qquad \gamma = \frac{d}{m} = 16, \qquad (\theta_1, \theta_2) = (7.5, 15), \qquad \lambda = 0.75, \qquad \nu = 18.$$

Equation (34) gives $\mu_0 = 3$, $a_- = 9$ and the tail floor $a = 16$. A flat tail has full effective rank, the regularization is weak, $\lambda = \mu_0/4$, and both spikes are outliers of the sketched tail ($m_\theta = 0.38$ and 0.08). This configuration instantiates parts (i)–(ii) of Theorem 3.2 at $\zeta = 0.4$; the smaller-ζ experiment below tests its rate clause. The limiting responses are computed from the lifted measure of Lemma D.3, and at the level $\nu = 18$ they cross at

$$t_* = 0.2654,$$

where FT has zero limiting mean-direction loss on the two-spike source.

Pilot and assessment workers are disjoint. Filter parameters selected on the pilot are frozen before the worker-count sweep. Every direction is scored by its ray loss, the loss after the best common scalar; these experiments take no optimization step.

The source controls include:

- equal Newton-energy weight on the two spikes;
- a source supported on one spike;
- a mixture of spike and background coordinates;
- a background-only source; and
- isotropic or smoothly weighted sources.

F.2 Response matching and source controls

The limiting FT responses cross at t_*, producing zero two-spike mean-direction loss. The stable endpoint envelope retains limiting loss 0.00288, while the positive-endpoint envelope has loss 0.0306 in this configuration.

The matching improvement tolerates timing error. Evaluating the saved limiting response path at $t/t_* \in \{0.8, 0.9, 1.1, 1.2\}$ gives losses 0.00180, 0.00041, 0.00035 and 0.00131, all below 0.00288. Positive semidefiniteness gives the uniform worker exposure bound $E_j \le \nu t$, ranging from 3.82 to 5.73 at these times. More generally, smooth responses with a positive common value at t_* give $R_{\infty,\mathrm{FT}}(\nu, t) = O(|t - t_*|^2)$ locally by (6): response differences are $O(|t - t_*|)$, their weighted variance is quadratic, and the denominator stays positive.

A source supported on one eigenspace has zero calibrated endpoint floor: any positive response on one coordinate agrees with the target after a common scalar. Background-only sources similarly need not favor a filter selected to match the upper spikes. Mixed sources can penalize a head-specific filter matched in the limit, although source-specific pilot selection may recover an advantage.

These controls show that a broad spectrum alone is insufficient. The source must expose variation in the endpoint's relative response.

F.3 Rate separation over three tails

This study tests the rate separation of Theorem 3.2 across three tail shapes, under a protocol frozen before the runs. The tail floor is fixed at $a = 1$, with one of three shapes: flat, 16 levels spaced geometrically in $[0.25, 2]$, and 32 levels proportional to $k^{-1/2}$, each level with the same multiplicity. The aspect ratio γ is 16, 64, 256 or 1024, the head sources are $(0.5, 1, 2, 4)$ and $(1, 3)$ with equal Newton-energy weights, and λ/μ_0 is 0.001, 0.25 or 0.9. Of the 24 combinations of tail, aspect ratio and source, 20 satisfy the hypotheses of part (iii) of the theorem; at $\gamma = 16$ the four-eigenvalue source has $\zeta > 1/4$, and the top level of the power-law tail lies above the fold. The limiting quantities are computed from the spectrum alone by (30), (32) and (36), at the pair of the theorem, $\nu = a - \lambda$ and $\theta_{\min} t = 2\log(1/\zeta)$, and the mean responses are compared with 2,000 Gaussian sketches at $m = 64$ (1,000 at $\gamma = 1024$) and 1,000 at $m = 128$ for $\gamma \le 256$.

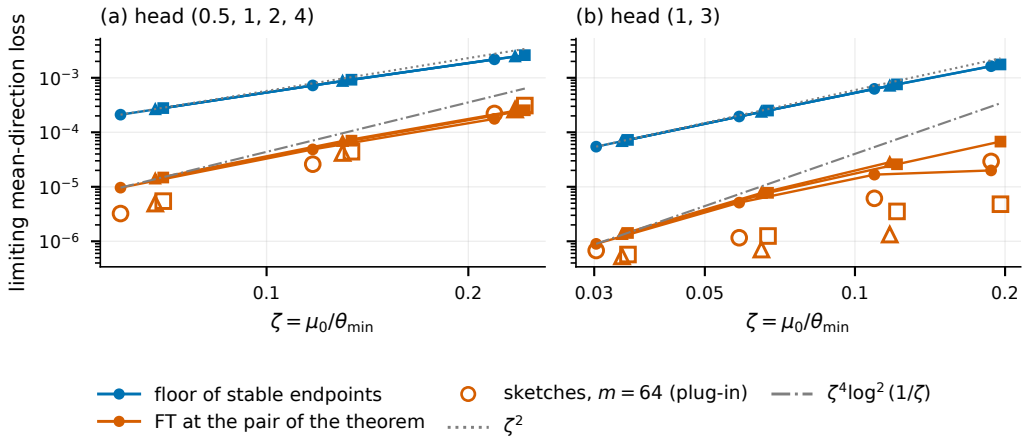


Figure 4: **Endpoint floor and FT loss against** ζ at $\lambda = \mu_0/4$, for the flat (circles), geometric (squares) and power-law (triangles) tails. Filled markers and lines are limits; open markers are the plug-in loss of the mean responses of Gaussian sketches at $m = 64$. The guides pass through the smallest ζ of each panel.

Figure 4 compares the limiting losses across three tail shapes. For $\lambda/\mu_0 \le 0.25$, the endpoint floor is 9.9 to 405 times the loss of the theorem's explicit FT pair over these 20 configurations. The plotted rate curves are guides, not fitted exponents. A fine population-response search under the same exposure cap finds still smaller FT losses, consistent with the value of jointly tuning shift and time; this grid is a diagnostic, whereas the implemented selector scores candidates by the local quadratic model (Section 4).

Figure 5 separates the size of the endpoint obstruction from FT's relative performance over nine aspect ratios and ten values of λ/μ_0 for each tail. The endpoint floor is largest on this grid under weak regularization and at larger ζ. In all 90 cells with $\lambda/\mu_0 \le 0.3$ and $\zeta \le 0.15$, it exceeds the

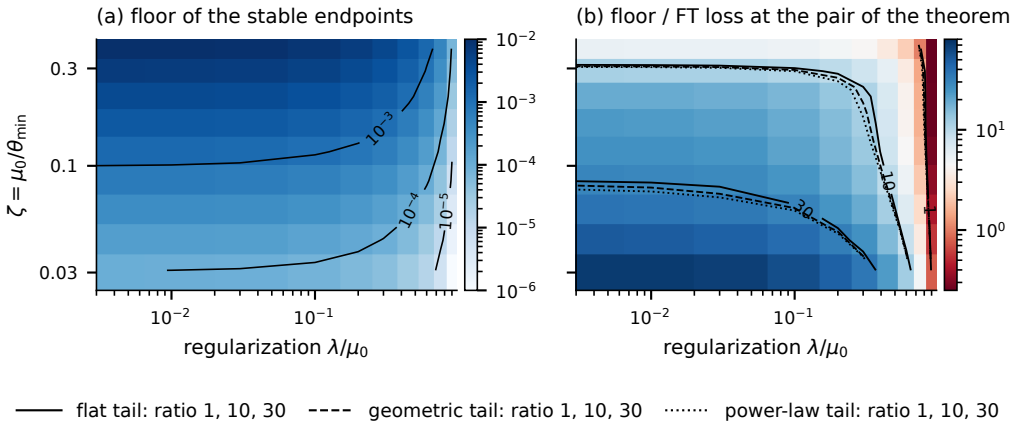


Figure 5: **Endpoint obstruction and FT's relative loss, in the limit.** Head $(0.5, 1, 2, 4)$ over a tail with floor one, aspect ratios 16 to 4096. (a) Stable-endpoint mean-direction floor for the flat tail. (b) The floor divided by FT loss at the theorem's explicit pair; ratios above one favor FT. Contours at 1, 10 and 30 nearly coincide for the three tails.

explicit FT loss by at least a factor of 11. The ratio falls below 0.8 at $\lambda/\mu_0 = 0.9$, where that pair offers no gain. Thus a large relative advantage need not mean a large absolute reduction: as ζ shrinks, the endpoint floor also becomes small.

The open markers show the plug-in losses of the Gaussian-sketch mean responses. The complete grid and uncertainty diagnostics accompany the code.

F.4 Fixed-aspect-ratio scaling

To test the proportional prediction, we hold $d/m = 16$ and use

$$(m, d) \in \{(16, 256), (32, 512), (64, 1024), (128, 2048), (256, 4096), (512, 8192)\}.$$

The filter matched in the limit ($\nu = 18$, $t = t_*$) and a grid of stable fixed-shift endpoints are evaluated on $q = 2^{17}/m$ disjoint workers at every size, with no pilot selection. The quantity of the theorem is the mean-direction loss, so we estimate it from the measured mean responses $\hat{r}_1, \hat{r}_2$ of the two spikes: for the source with equal Newton-energy weights, $\widehat{R}_\infty = \{(\hat{r}_1 - \hat{r}_2)^2 - \mathrm{se}^2\}/\{2(\hat{r}_1^2 + \hat{r}_2^2)\}$, with se the standard error of the paired difference, whose square is the noise carried by the squared gap. This estimate has no ω/q term, unlike the loss of the averaged direction, which at a fixed total budget is dominated by that term once q falls to 256. Each endpoint uses the worker-local guard of the experiments, and its fallback rate is recorded.

The floor is a property of the stable endpoint family, not of one selected shift: over the whole grid of admissible negative shifts, the smallest mean-direction loss is 1.4 to 1.7 times the floor 0.0029 at every dimension, and it is approached only as the shift tends to the limiting edge $-a_-$, where the worker fallback rate rises (Figure 6b). The filter matched in the limit, by contrast, has zero limiting loss: its estimate falls from 2.5×10^{-4} at $d = 256$ to the noise level 10^{-6} from $d = 2{,}048$ on (Figure 6a), so the measured separation from the best endpoint grows from a factor 20 to more than 10^3. Its normalized worker variance is stable under proportional scaling (3.75 at $d = 256$, 3.45 at $d = 8{,}192$), comparable to the endpoint's 4.4 to 3.0, so the advantage in the loss of the averaged direction appears once $\widehat{\omega}/q$ falls below the endpoint loss, about five hundred workers at every size (Theorem 3.3). The per-shift and per-source tables are in the supplementary material.

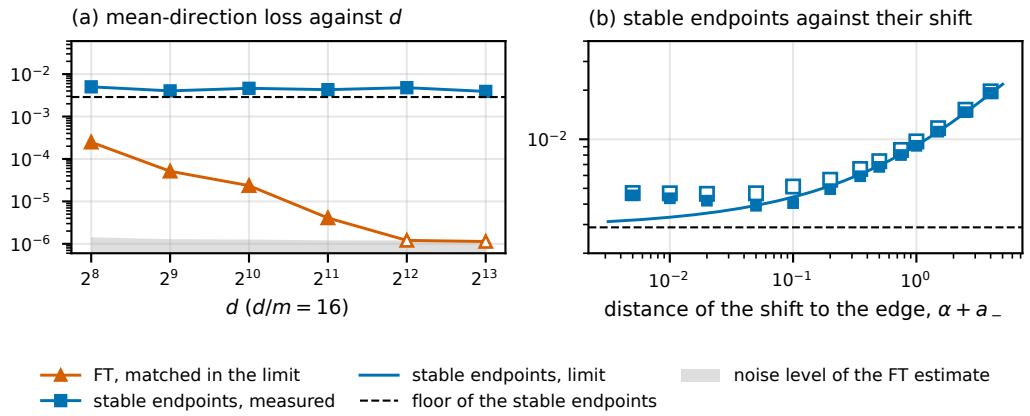


Figure 6: **Fixed aspect ratio** $d/m = 16$**.** (a) Mean-direction loss of the measured mean responses against d: the filter matched in the limit falls to the noise level of the estimate (grey; an open marker is a value below it), while every stable endpoint stays above the floor of the family (dashed). (b) The mean-direction loss of the whole stable endpoint grid against its shift $\alpha + a_-$ at $d = 1{,}024$ (open) and $d = 8{,}192$ (filled), with the limit (line): no shift reaches the floor, which the family approaches only as $\alpha \to -a_-$.

G Independent-Pilot Worker Crossover

This experiment tests how many workers are needed to turn FT's lower mean-direction error into a gain for independently selected filters. It fixes $H = \mathrm{diag}(a_i - \lambda)$, where $a_i = i^{-\log(1000)/\log(500)}$, $d = 500$, $m = 16$ and $\lambda = 10^{-4}$. For this spectrum, $d_1(\lambda) = 476.3$, $d_2(\lambda) = 454.1$ and $\mu_0 = 0.0391$. In each of five replications, an independent 256-worker pilot selects one FT filter and one guarded endpoint for each source and target $q = 1024$. The selected physical parameters are frozen. A disjoint worker pool estimates R_∞ and ω, and separate assessment pools measure the actual ray loss at increasing q. This design isolates worker averaging from filter retuning; pilot and assessment costs are recorded separately.

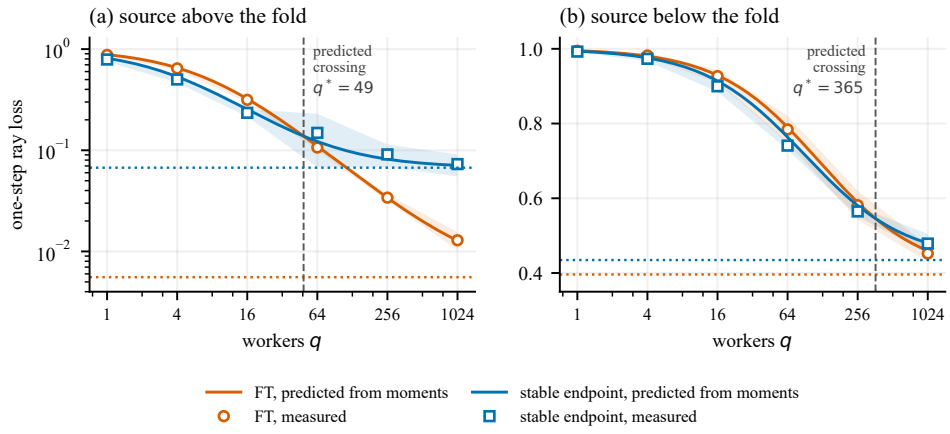


Figure 7: **Worker crossover: measurement and theorem.** Markers are measured ray losses of the frozen filters, with the variation across replications shaded. Curves are $(\widehat{R}_\infty + \widehat{\omega}/q)/(1 + \widehat{\omega}/q)$, the bound of Theorem 3.3 at the moments of a disjoint worker pool; dotted lines are $\widehat{R}_\infty$, and the dashed line is the crossing of the two curves.

On the above-fold source, FT has the smaller estimated mean-direction loss, $\widehat{R}_{\infty,\mathrm{FT}} = 0.0056$ against $\widehat{R}_{\infty,\mathrm{end}} = 0.0673$, but the larger normalized worker variance, $\widehat{\omega}_{\mathrm{FT}} = 7.5$ against $\widehat{\omega}_{\mathrm{end}} = 4.0$. The curve of Theorem 3.3 at these moments tracks the measured FT loss within 3% at every q from 1 to 1024, where it gives 0.0128 against a measured 0.0129, and it places the crossing at $q^* = 49$; FT is measured worse at $q = 16$ and better at $q = 64$, and reaches 0.0129 against 0.0731 at $q = 1024$ (Figure 7). The endpoint is measured up to 22% above its curve, at $q = 64$; for both families the 95% interval across replications covers the curve at every q. The sufficient count of Corollary C.5, whose constant six holds for every direction distribution, is 587 on this source.

Below the fold the worker variances are large, 120 against 89, and the mean-direction losses close, 0.40 against 0.43, so the predicted crossing moves to $q^* = 365$ and is observed between 256 and 1024; the isotropic source behaves in the same way ($q^* = 259$). The source decides how many workers finite time needs. The FT fallback rate is zero in the moment-assessment sample and that of the guarded endpoint is about 0.12%, so the trend is not explained by changes in endpoint admissibility.

H Closed-Loop Optimization and Data Experiments

H.1 Common protocol

All closed-loop experiments use fresh sketches at every outer iteration. Unless otherwise stated, the local sketch size is $m = 16$, the primary target is a certified relative objective gap of 10^{-4}, and the prespecified secondary target is 10^{-2}. Runs terminate after 120 accepted rounds or the specified time budget. At each tolerance, a run that does not reach the target before termination is counted as no hit.

Synthetic results use three independent seeds. Corn uses three sketch seeds on one fixed data split. Comparisons use paired sketch streams whenever the methods permit it. Parameter reuse retains only physical shift and stopping-time values; no sketch or factorization is reused across outer iterations.

RZP at limited m and the undebiased endpoint use backtracking from step one; native L-BFGS uses its own line search. None receives an outer scalar. The tuned stable endpoint (ours) searches over 65 positive and 65 admissible negative shifts and receives the same calibration c_Q as FT, which isolates finite stopping from shift tuning and common rescaling. FT is also reported without c_Q. Method definitions and candidate-selection rules appear in Appendix E.

H.2 Sampled covariance models

This study defines a population signal before drawing the empirical objective. Write $\Sigma = \mathrm{diag}(s_1, \ldots, s_d)$, let Z_{ij} be independent standard Gaussians, and generate

$$X = Z\Sigma^{1/2}, \qquad y = X\beta + \varepsilon, \qquad F(x) = \frac{\|Xx - y\|^2}{2n} + \frac{\lambda}{2}\|x\|^2.$$

We use $d = 128$, $\lambda = 0.01$, $m = 16$, and $q \in \{256, 1024\}$. All methods start at zero and receive the same raw-data gradient and HVP oracle $Hv = X^\top(Xv)/n$. The empirical head–tail block $X_T^\top X_H/n$ is retained, so a population-head signal produces tail components in the empirical gradient. Separate random streams generate the signal, data, response noise and paired worker sketches.

Seven prespecified cases. The two-spike covariance has head eigenvalues $7.5, 15$; the spread-head covariance has 12 logarithmically spaced eigenvalues from 4 to 64. Both have tail eigenvalues proportional to $z^{-3/4}$ for equally spaced $z \in [1, 4]$, normalized to mean one, and are tested at $n = 128, 512$. The fifth case adds independent $N(0, 0.1^2)$ response noise to the spread-head model at $n = 128$. The polynomial case uses $s_j \propto j^{-\log(1000)/\log(128)}$, normalized to unit mean over coordinates $13, \ldots, 128$, with signal on the first 12 coordinates. In these six cases, $\beta_j = \xi_j/\sqrt{ks_j}$ on the k head coordinates, with independent random signs ξ_j, and zero elsewhere; thus $\beta^\top\Sigma\beta = 1$. The remaining case uses three population eigenvalue levels $1, 0.01, 0.0001$ on $16, 42, 70$ coordinates, again normalized to unit tail mean, and allocates 90% of the population signal energy to the head and 10% to a random tail direction. The polynomial and multibulk cases use $n = 128$. Responses are noiseless except in the designated noisy case.

Selection and measurement. Three paired seeds are run to relative empirical objective gap 10^{-4}, with a 120-iteration budget and no time truncation. FT uses the 300-candidate shift–time grid of Appendix E.3. Quality-aware reuse receives the same initial and refresh grid. The tuned endpoint searches its 130-shift grid; endpoint reuse retains the physical shift until a stability refresh. Native L-BFGS is run once per input and appears at both worker counts. Dense reference solves and spectral diagnostics are evaluator-only; population prediction risk is logged separately from the empirical optimization target and never used for selection or stopping.

Table 3: **Sampled covariance models at $q = 1024$.** Accepted outer iterations to relative gap 10^{-4}, mean $\pm$ sample standard deviation over three paired seeds. "No hit" means none reaches the target within 120 iterations; bold marks the fewest iterations in each row. FT and the tuned endpoint use c_Q except in the indicated column; FT reuse is quality-aware and endpoint reuse is stability-triggered.

	Finite time (ours)			Endpoint (ours)			
Model	retuned	reuse	no c_Q	retuned	reuse	RZP	L-BFGS
Two spikes, $n = 128$	**4.0±0.0**	**4.0±0.0**	**4.0±0.0**	9.0±0.0	9.7±0.6	27.0±2.6	8.0±1.0
Two spikes, $n = 512$	**3.0±0.0**	**3.0±0.0**	**3.0±0.0**	5.3±0.6	5.3±0.6	16.0±0.0	6.0±0.0
Spread head, $n = 128$	**5.7±0.6**	6.0±0.0	6.7±1.2	18.0±2.6	19.3±3.1	47.7±6.0	24.0±2.6
Spread head, $n = 512$	**4.0±0.0**	**4.0±0.0**	**4.0±0.0**	10.0±0.0	10.7±0.6	26.0±0.0	16.0±0.0
Noisy head, $n = 128$	**18.3±2.5**	19.0±3.0	21.3±2.5	78.7±12.5	99.3±16.8	no hit	56.3±2.1
Polynomial, $n = 128$	**4.3±0.6**	**4.3±0.6**	5.0±0.0	16.7±5.5	18.3±6.1	25.3±4.2	24.0±2.0
Multibulk, $n = 128$	no hit	no hit	no hit	no hit	no hit	no hit	no hit

Table 4: **The same sampled inputs at $q = 256$.** Reporting conventions match Table 3. FT retuning and quality-aware reuse beat native L-BFGS in all 18 paired runs across the six non-multibulk cases.

	Finite time (ours)			Endpoint (ours)			
Model	retuned	reuse	no c_Q	retuned	reuse	RZP	L-BFGS
Two spikes, $n = 128$	**5.0±1.0**	5.3±0.6	6.0±1.0	11.3±1.5	12.0±2.0	31.0±5.6	8.0±1.0
Two spikes, $n = 512$	**3.7±0.6**	4.0±0.0	4.0±0.0	6.3±0.6	6.7±0.6	18.3±0.6	6.0±0.0
Spread head, $n = 128$	**9.3±0.6**	9.7±1.2	11.0±1.7	24.0±4.6	33.3±16.4	51.3±8.1	24.0±2.6
Spread head, $n = 512$	**4.7±0.6**	**4.7±0.6**	5.7±0.6	11.0±0.0	11.3±1.2	28.7±0.6	16.0±0.0
Noisy head, $n = 128$	**21.7±3.2**	22.7±2.5	23.7±2.5	77.7±13.5	100.0±14.9	no hit	56.3±2.1
Polynomial, $n = 128$	**8.0±0.0**	8.3±0.6	8.7±0.6	16.3±1.5	18.7±2.5	30.7±4.7	24.0±2.0
Multibulk, $n = 128$	no hit	no hit	no hit	no hit	no hit	no hit	no hit

Every sampled input has $m < d_2(\lambda)$. The code provides the per-input geometry, trajectories and resource counts alongside these iteration tables.

H.3 Nonlinear objectives

Two sampled-data classification models test the method with a varying Hessian. Both start at $x_0 = 0$, and all generated inputs are saved.

Logistic. $n = 1500$ observations and $p = 150$ features. Rows are Gaussian with correlation 0.85 between neighbouring features, and feature j is scaled by a geometric sequence from 1 to 0.03. Labels follow the logistic model whose coefficient vector has norm 3 and a uniformly random direction; $\lambda = 10^{-3}$.

Softmax. $n = 1000$, $p = 60$ and four classes, that is 180 parameters. Features are Gaussian plus a rank-six factor component, with the same geometric scaling; labels follow the softmax model with coefficient columns of norm 3; $\lambda = 10^{-3}$.

At $q = 1024$, retuned FT reaches the target in 14.0 iterations on logistic and 28.3 on softmax; quality-aware reuse takes 15.0 and 31.7. FT without c_Q takes 12.0 and 27.0, showing that the filter retains its advantage without outer calibration. The tuned endpoint takes 46.3 and 91.7 iterations (Table 5).

Table 5: **Nonlinear objectives.** Mean first-hit iterations to relative gap 10^{-4}, with three seeds. Parentheses count successful runs; means in these cells use successes only. "No hit" denotes zero successes under the common iteration/time budget. Bold marks the fewest iterations in each row. Here FT uses the 90-candidate grid.

		Baselines			Ours			
Problem	q	RZP	endpoint	L-BFGS	FT	FT quality	FT, no c_Q	tuned endpoint
Logistic	1024	no hit	no hit	56.0	14.0	15.0	**12.0**	46.3
Softmax	1024	no hit	no hit	45.0	28.3	31.7	**27.0**	91.7

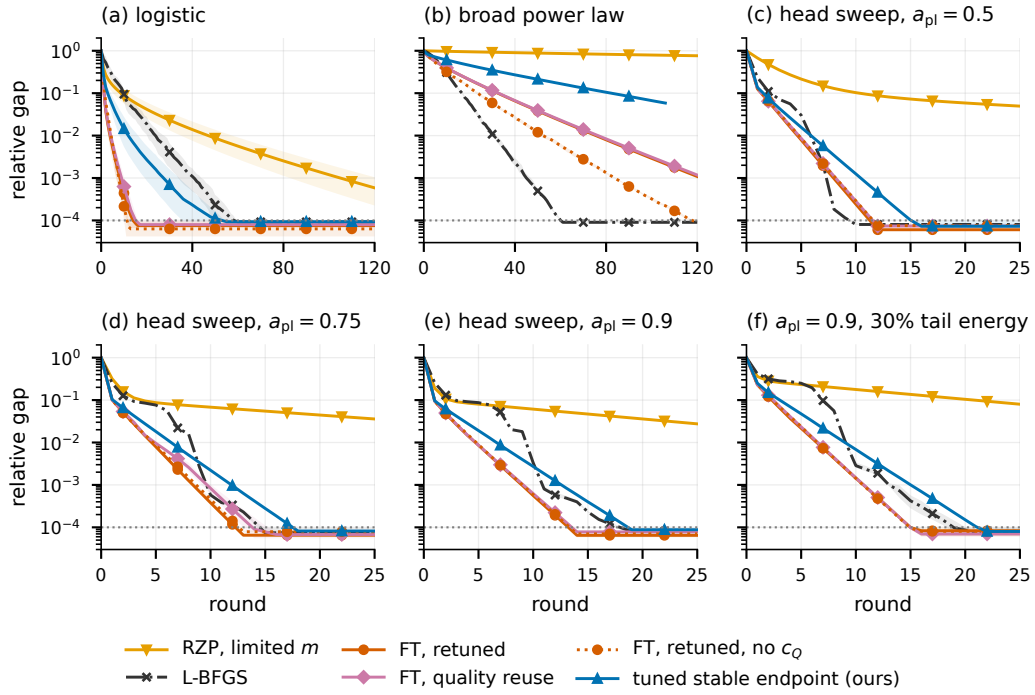


Figure 8: **Closed-loop trajectories at $m = 16$.** Certified relative gap against round, geometric mean over three seeds, seed range shaded. A run that reaches 10^{-4} (dotted) is held at its final gap, and a curve ends where a seed exhausted its budget. (a) Logistic at $q = 1024$; (b) the fixed-Hessian, broad-source power law with $d = 10^4$, $\kappa(A) = 10^3$ and $q = 1024$; (c)–(f) the power-law-tail head sweep of Appendix H.7 at $d = 10^4$ and $q = 1024$, first 25 of 100 rounds: tail exponents $a_{\mathrm{pl}} = 0.5$, 0.75 and 0.9 with 10% tail energy, and $a_{\mathrm{pl}} = 0.9$ with 30%. RZP at limited m reaches the target in none of them.

H.4 Corn: data, selection and protocol

Corn tests FT on real spectra with regularization selected by predictive validation before the optimizer comparison.

Data and preprocessing. The Eigenvector corn dataset contains near-infrared spectra of 80 corn samples and four measured properties: moisture, oil, protein and starch (Eigenvector Research, n.d.). We use the `m5` instrument's 700 channels, covering 1100–2498 nm in 2 nm increments, and retain all four responses. A fixed split assigns 64 samples to development and 16 to held-out evaluation. Each predictor and response is centered and scaled using development-set statistics; cross-validation fits these transformations separately within each training fold. Each response defines a ridge objective $\|Xx - y\|^2/(2n) + \lambda\|x\|^2/2$, with no separate intercept after centering.

Weak regularization selected before optimization. An independent dataset screen evaluates predictive error, without running FT or its competitors. Three repeats of five-fold cross-validation use the grid $\{0\} \cup \{10^{-10+j/2} : j = 0, \ldots, 32\}$. For every corn response the pooled minimum occurs at zero. We fix the common positive value $\lambda = 10^{-8}$: each response's CV error is within 0.03% of its minimum. This choice and the retention of all four outcomes precede the optimizer comparison. The held-out samples are scored after optimization and do not enter parameter selection or stopping.

Matched comparison. All methods start at zero. Each of the six sketched methods evaluated on corn in Table 2 uses $m = 16$, $q \in \{64, 256, 1024\}$, three paired sketch seeds and a 120-iteration budget. Streams are nested across worker counts and refreshed at each iteration. FT uses the expanded 300-candidate grid with exposure at most 12 (Section 4). Quality-aware reuse stores the selected physical shift and time, checks up to five nearby candidates and refreshes when the quadratic-model score improves by more than 10%. The tuned endpoint and its reuse arm use the same 130-shift family and outer calibration as FT. RZP keeps its original root bracket and backtracking; this limited-size arm is below its sketch-selection region. Native L-BFGS uses SciPy's L-BFGS-B with ten pairs, its native line search and no HVP or added scalar. Its four completion runs change only the iteration budget to 10,000; their accepted-state prefixes agree exactly with the original 120-iteration runs.

Iteration gains and tuning work. All three FT policies reach the primary target in every tested outcome/worker/seed combination, earlier than the paired tuned endpoint. At $q = 256$, reuse reduces candidate evaluations by 76%–95% relative to full retuning while taking more outer iterations. Every sketched iteration costs qm HVP columns; searching its cached small-matrix factors adds no HVPs. Recorded solver times for retuned FT are 1.09, 10.10, 8.92, and 8.71 seconds on moisture, oil, protein and starch, versus 0.19, 1.02, 1.14, and 1.11 for native L-BFGS to the same target. These host-contended measurements accompany the iteration comparison; full work, timing and worker-sweep tables are included in the reproducibility records.

H.5 CT SLICE AND SUPERCONDUCTIVITY: DATA AND PROTOCOL

These studies test whether FT's iteration gains extend to CT slice prediction and superconductivity regression.

Data and grouped splits. The UCI CT slice dataset contains $53{,}500$ rows with 384 histogram features and a slice-location response (Emrich et al., 2010; Graf et al., 2011). Superconductivity contains $21{,}263$ rows with 81 numerical features and critical temperature as the response (Hamidieh, 2018a;b). Patient identifiers and chemical formulas define the respective groups. Three seeds give paired 60/20/20 grouped training/validation/test splits, subsampled to 2000/500/1000 rows. Features are standardized and responses centered using training statistics only.

Selection and optimization. An optimizer-free geometry screen selects these two datasets. Each split selects a common ridge penalty from $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$ by validation loss and freezes it across methods: CT uses 10^{-2} on all splits, and superconductivity uses $10^{-4}, 10^{-4}, 10^{-3}$. The saved studies use $m = 16$, $q \in \{64, 256\}$, 120 outer iterations and a 180-second solver cap. Table 2 reports $q = 256$; none of these runs is time-censored. These studies use the original 90-candidate FT grid: nine times in $[0.03/r_{\text{med}}, 16/r_{\text{med}}]$, ten shift choices and exposure at most 6, with r_{med} the current median worker eigenvalue. Corn uses the expanded grid of Appendix H.4. Reuse and baseline rules follow Section 4; endpoint reuse was not run here. Test data enter neither selection nor stopping.

Primary and secondary targets. CT reaches 10^{-4} on every FT run. Superconductivity reaches this target on one split for each FT policy, while the endpoints and L-BFGS exhaust the iteration budget. At the prespecified secondary target 10^{-2}, retuned FT, quality reuse and plain FT succeed on all three superconductivity splits in 16.7, 24.3 and 13.7 iterations on average, against 38.0 for the tuned endpoint and 53.7 for L-BFGS.

H.6 HIGH-DIMENSIONAL FIXED-HESSIAN BOUNDARY

This control tests performance when most gradient energy lies below the fold. It uses a fixed Hessian with $d = 10^4$, $A = \mathrm{diag}(a_i)$, $a_i = i^{-\log \kappa(A)/\log d}$, $\lambda = 0.1/\kappa(A)$, and $H = A - \lambda I$, with

$$\kappa(A) \in \{250, 1000, 4000\}.$$

The solution is proportional to $(\xi_i/\sqrt{a_i})_i$, where the ξ_i are independent standard Gaussians, normalized to unit A-norm. Thus all d distinct curvature modes are active at $x_0 = 0$, with equal expected Newton energy. The local sketch remains $m = 16$, and the source initially places at most 0.1% of its Newton energy above the fold. These experiments therefore lie outside the setting of Theorem 3.2, whose source sits on a head above the tail.

Runs use $q \in \{256, 1024\}$, three seeds, and an extended time cap. At $q = 1024$, a 100-round budget permits $1{,}638{,}400$ sketch columns. Table 6 compares gaps under this common sketch-work budget.

Table 6: **Fixed-Hessian broad-source control at $d = 10^4$.** Certified relative gap after 100 rounds, averaged over three seeds, under a budget of $1{,}600q$ HVP columns. Runs that converged earlier retain their final gap. FT uses quality-aware reuse with c_Q; the tuned stable endpoint (ours) uses the same calibration. The ratio compares their gaps; larger than one favors FT. The released full table includes every FT variant.

$\kappa(A)$	q	RZP	tuned endpoint	FT quality	endpoint / FT
250	256	0.760	0.00548	2.42×10^{-4}	22.6
250	1024	0.759	0.00414	1.55×10^{-4}	26.7
1,000	256	0.794	0.071	0.00594	11.9
1,000	1024	0.794	0.067	0.00311	21.5
4,000	256	0.833	0.244	0.067	3.6
4,000	1024	0.833	0.245	0.034	7.2

At $q = 1024$, FT with quality-aware reuse has gaps 7 to 27 times smaller than the tuned stable endpoint (Table 6). The full table also shows that FT without c_Q improves over RZP. These are objective gaps over a measured horizon; the endpoint floor of Proposition 3.1 concerns the mean direction of one round and does not by itself imply a limiting objective gap.

At condition number 250, calibrated FT variants and uncalibrated retuning reach the primary target at $q = 1024$. At 1000, only FT retuning without c_Q reaches it; at 4000, no sketched method does. The calibration optimizes the current step, not necessarily the full trajectory.

FT improves on the endpoint even for this broad source. L-BFGS reaches the primary target in 32.0, 60.7, and 112.3 rounds on the three problems, respectively, and remains a strong practical baseline.

H.7 POWER-LAW-TAIL HEAD SWEEP AT $d = 10^4$

This sweep tests whether FT's head-response advantage persists when the gradient also carries tail energy. It adds a six-eigenvalue head to the power-law family used in the broad-source control above.

Setup. Under a protocol frozen before the runs, we use tail eigenvalues $h_i = i^{-a_{\mathrm{pl}}}$ for $a_{\mathrm{pl}} \in \{0.5, 0.6, 0.75, 0.9\}$ on $d - 6$ coordinates, with six head eigenvalues geometrically spaced from 1.5 to 6 times the tail's top eigenvalue, $\lambda = \mu_0/4$ and $m = 16$. The head lies above the sketched tail: the largest outlier statistic is 0.60, at $a_{\mathrm{pl}} = 0.5$. Every source places 10% or 30% of its Newton energy on a Gaussian tail vector. The initial gradient therefore has a nonzero component on every eigenvector, rather than being confined to a small invariant subspace. This tests response matching across six head eigenvalues and a tail, beyond the two-spike example of Section 5.1.

Heavier tails move the fold toward the tail's top eigenvalue: $\zeta = \mu_0/1.5$ is 0.49, 0.29, 0.15 and 0.086, and $\kappa(A)$ is 32, 54, 105 and 186, for the four exponents. We use three paired seeds, $q = 1024$, a 100-round cap and the 300-candidate grid.

Iteration gains. Table 7 and Figure 8(c)–(f) give the results. Retuned FT reaches the target in fewer rounds than the endpoint with per-round retuning in all 24 paired runs, as does FT without c_Q; the ratio of their rounds is 1.3 to 1.4 for every exponent, and RZP at limited m reaches the target in none.

After one round FT leaves 0.2% to 4% of the remaining Newton energy on the head, the endpoint 1.8% to 14%. The ratio does not grow for heavier tails: with 10% tail energy the tail's response mismatch dominates the endpoint's mean-direction loss (floor proxy 0.050 to 0.054 at every exponent).

L-BFGS tracks $\kappa(A)$, from 10 rounds at $a_{\mathrm{pl}} = 0.5$ to 18.7 at $a_{\mathrm{pl}} = 0.9$; retuned FT is ahead of it for $a_{\mathrm{pl}} = 0.75$ (13 against 15) and $a_{\mathrm{pl}} = 0.9$ (14 against 18.7) and behind it for $a_{\mathrm{pl}} = 0.5$ (12 against 10). With 30% tail energy every sketched arm needs one to three more rounds and the ordering is unchanged.

Table 7: **Power-law-tail head sweep at $d = 10^4$, 10% tail energy, $q = 1024$.** Accepted outer iterations to relative gap 10^{-4}, mean $\pm$ sample standard deviation over three paired seeds; "no hit" means none within 100 iterations; bold marks the fewest iterations in each column.

	$q = 1024$			
Method	$a_{\mathrm{pl}} = 0.5$	$a_{\mathrm{pl}} = 0.6$	$a_{\mathrm{pl}} = 0.75$	$a_{\mathrm{pl}} = 0.9$
FT, retuned	12.0 ± 0.0	$\mathbf{12.0 \pm 0.0}$	$\mathbf{13.0 \pm 0.0}$	$\mathbf{14.0 \pm 0.0}$
FT, quality reuse	12.0 ± 0.0	$\mathbf{12.0 \pm 0.0}$	14.7 ± 1.2	$\mathbf{14.0 \pm 0.0}$
FT, no c_Q	12.0 ± 0.0	12.3 ± 0.6	$\mathbf{13.0 \pm 0.0}$	$\mathbf{14.0 \pm 0.0}$
Tuned endpoint (ours)	15.7 ± 0.6	17.0 ± 0.0	18.0 ± 0.0	19.0 ± 0.0
Endpoint reuse (ours)	16.0 ± 0.0	17.0 ± 0.0	18.3 ± 0.6	20.0 ± 0.0
RZP, limited m	no hit	no hit	no hit	no hit
Native L-BFGS	$\mathbf{10.0 \pm 0.0}$	$\mathbf{12.0 \pm 0.0}$	15.0 ± 0.0	18.7 ± 0.6

I Thresholds and Fixed-Budget Allocation

I.1 Threshold map

Table 8 records the effective-dimension thresholds of the spiked mechanism models and the corn objective. Each lies below the fold at $m = 16$. These scales locate the local sketch cap relative to endpoint matching and RZP's sketch-selection guarantee.

Table 8: **Thresholds of the spiked models and corn data.** At the initial iterate and stated regularization: the fold scale $d_2(\lambda)$, the debiasing threshold $d_1(\lambda)$, the lower scale $1.5\, d_1(\lambda)$ in RZP's sketch-selection guarantee, and the fraction of Newton energy above the fold of a sketch of size 16. Corn shares one Hessian across four outcomes; its head-energy range covers all four. A dash marks a quantity not evaluated for that experiment.

Problem	d	λ	$d_2(\lambda)$	$d_1(\lambda)$	$1.5\, d_1(\lambda)$	Head energy
Spiked model, $d = 256$	256	0.75	85	147	221	–
Spiked model, $d = 1024$	1,024	0.75	335	586	879	–
Corn (m5)	700	10^{-8}	62.9	63.0	94.4	95.0–99.8%

I.2 Budget-matched sketch-size sweep

For each local sketch size m, an independent pilot selects one FT filter and one endpoint filter. The policies are then frozen and assessed on disjoint workers at total budgets $N = qm \in \{2^{13}, 2^{15}, 2^{17}\}$. The assessment budget excludes pilot selection, whose cost is reported separately in the supplementary material. Figure 2b shows the head source of the polynomial spectrum at the largest budget; the supplementary material tabulates the other budgets, problems and sources.

On the weakly regularized problems, FT has lower loss throughout the tested below-fold range, with the largest relative advantage at small m. The gap narrows as m approaches the fold or effective-dimension scale.

The strongly regularized control locates the transition to endpoint matching. At $\lambda = 1$, FT has higher loss at the smallest sketch and becomes indistinguishable from the endpoint as m grows. The recorded RZP fallback fraction is 100% through $m = 128$ and 0% at $m = 256, 512$. The scale $1.5 d_1(\lambda)$ is the lower end of RZP's selection guarantee; acceptance uses the empirical Stieltjes test.

Their test accepts with high probability once $m \geq 2d_1(\lambda)$, under its stated assumptions (Romanov et al., 2025, Section 3).

J Reproducibility

J.1 Software, provenance and objectives

The implementation uses Python 3.11 or later with NumPy, SciPy, pandas, Matplotlib, scikit-learn, threadpoolctl, and psutil; exact package versions, platform information and dependency hashes are in the supplementary environment files. The supplementary material contains the source code, the frozen experiment configurations, dataset and output checksums, generated tables and figures, raw per-seed and per-split trajectories, pilot, moment and assessment outputs, the table-generation scripts, read-only audit reports and automated tests of the filter, calibration, HVP accounting and table-generation pipelines. Every experiment has a protocol frozen before execution, a runner and an independent audit script. Manuscript tables are generated from saved experiment outputs by scripts that verify input hashes and basic invariants (worker counts, HVP totals, first-hit definitions, zero solver HVPs for L-BFGS). The audits reconstruct saved worker-prefix averages, filter evaluations, ray scores and selected parameters within floating-point tolerance; they support implementation and provenance and do not replace the mathematical arguments.

All objectives have the form $F(x) = f(x) + \frac{\lambda}{2}\|x\|^2$. Quadratics use $f(x) = \frac{1}{2}x^\top Hx - b^\top x$ with $H \succeq 0$ and the exact solution $x^\star = (H + \lambda I)^{-1}b$. Binary logistic regression uses $f(x) = n^{-1}\sum_i \log(1 + e^{-y_i a_i^\top x})$ with $y_i \in \{-1, 1\}$. Softmax regression uses the mean cross-entropy; the final class coefficient is fixed to zero for identifiability, all retained coefficients are regularized, and no unpenalized intercept is used. Ridge regression uses $f(x) = (2n)^{-1}\|Xx - y\|^2$.

J.2 Data and preprocessing

Corn is distributed by Eigenvector Research at `https://www.eigenvector.com/data/Corn/`. The source archive is `corn.mat_.zip`; its `m5spec.data` and `propvals.data` arrays provide the inputs and four responses. The download checksum, original row indices, cross-validation folds and fitted transformations are saved with the independent screen and corn comparison. Appendix H.4 gives the data and selection protocol. The comparison records include all 220 primary runs and the four native completion runs. A separate audit replays the full ridge objective from the saved accepted states and checks the SVD reference against an augmented least-squares solve using pivoted QR.

CT slice location (UCI 206, DOI 10.24432/C5CP6G) uses `slice_localization_data.csv`; superconductivity (UCI 464, DOI 10.24432/C53P47) uses `train.csv` and `unique_m.csv`. The supplementary material preserves archive hashes, group assignments, split indices, preprocessing statistics and the original experiment protocols. Appendix H.5 gives their common-objective comparison and the grid used by these saved runs.

J.3 Reference solutions and certified gaps

Reference solutions are analytic for diagonal quadratics, obtained by a separately charged direct solve for ridge objectives, and obtained by a high-accuracy reference optimizer for nonlinear objectives. For the computed reference point x_{ref}, strong convexity gives $F(x_{\mathrm{ref}}) - F(x^\star) \leq \Delta_{\mathrm{ref}} := \|\nabla F(x_{\mathrm{ref}})\|^2/(2\lambda)$. With $G_0 = F(x_0) - F(x_{\mathrm{ref}})$, the recorded certified relative gap is

$$\widehat{G}_k = \max\left\{0,\ \frac{F(x_k) - F(x_{\mathrm{ref}}) + \Delta_{\mathrm{ref}}}{G_0}\right\}, \tag{39}$$

which upper-bounds the true objective gap relative to the same normalization. First-hit times are computed from $\widehat{G}_k$, not from an interpolated trajectory.

J.4 Line search and failure handling

Every sketched method uses the same Armijo backtracking on the objective. For the selected direction p, which includes c_Q when the method uses it, the trial steps are $s = 1, \frac{1}{2}, \frac{1}{4}, \ldots$, and a step is accepted

when

$$F(x - sp) \le F(x) - 10^{-4}\, s\, g^\top p.$$

The step is halved at most 35 times, so the smallest trial step is $2^{-35} \approx 2.9 \times 10^{-11}$; there is no other lower bound on the step. If $g^\top p \le 0$ or is not finite, or if no trial step is accepted, the run stops with a recorded failure status, the failed round is not counted, and every target not yet reached is recorded as a miss. L-BFGS uses the line search of SciPy's L-BFGS-B.

J.5 Randomness, pairing and resource accounting

Every run records the random seeds of problem or split construction, worker sketches, pilot selection, moment estimation and ray-loss assessment. Competing sketch methods use paired sketches whenever their admissibility rules allow a common worker stream, and outer rounds use fresh sketches. Independent-pilot experiments use disjoint pilot, moment and assessment streams; filters selected on a pilot are frozen before assessment and are not retuned as q changes. Same-pool and independent-pilot results are stored separately and are not pooled in summaries.

One HVP equivalent is one column passed through the oracle; an m-column block counts as m. Filtering uses cached factors and adds no HVPs. Records separately count HVP work, block-HVP depth, candidate evaluations, lifting operations, communication, pilot selection and reference fitting. Data preparation, held-out evaluation and reference fitting sit outside solver HVP totals. L-BFGS uses no HVPs; its gradient calls and rounds are recorded. Process resident-set size includes Python, data and libraries. Serial runs use a Windows 11 desktop (ten-core Intel Core Ultra 5 225F, 32 GB RAM), Python 3.12–3.13, NumPy 2.1–2.5, SciPy 1.15–1.18 and one to four BLAS threads. The sampled-covariance study runs up to three single-threaded replicates concurrently. Corn runs up to four independent two-threaded replicates concurrently; these timings are not CPU-isolated. Manifests record exact versions, and Appendix E models distributed costs.