ACC-RSRN: ACCELERATED RANDOMIZED SUBSPACE REGULARIZED NEWTON

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose Accelerated Randomized Subspace Regularized Newton (Acc-RSRN) for smooth, possibly nonconvex optimization, with improved local linear convergence rates. Our improvement over the prior work has two components: a sharp analysis of the randomized subspace iterations replaces the dependence on the full Hessian condition number with a smaller tail-averaged quantity which leverages the preconditioning effect of randomized subspaces, and Nesterov acceleration further reduces the dependence on this quantity from linear to square root. For an n-dimensional problem and subspace dimension proportional to r, the local iteration complexity improves from $\widetilde{O}(n\kappa/r)$ in prior work to $\widetilde{O}(n\sqrt{\bar{\kappa}_r}/r)$, where $\bar{\kappa}_r$ is the average of the Hessian's tail eigenvalues at the minimizer divided by its smallest eigenvalue. Our results support two different Hessian access models: Hessian-vector products via Gaussian sketching, and entry-wise access via ridge leverage score sampling. Experiments on kernel logistic regression and a CNN objective illustrate the benefits of acceleration in RSRN.

1 INTRODUCTION

Newton's method for continuous minimization is known to attain fast local convergence by adapting its steps to the curvature of the minimized function (Nocedal & Wright, 2006). However, forming and solving a full n-dimensional Newton system is prohibitive for large problems. Randomized subspace Newton methods instead solve the local quadratic model only within a random r-dimensional subspace, where $r \ll n$ (Gower et al., 2019; Shang et al., 2024; Fuji et al., 2025). This reduces the cost of the linear solve, but the overall efficiency depends on how much progress each low-dimensional update can make.

In order to extend randomized subspace methods to smooth nonconvex problems, $\min_{x\in\mathbb{R}^n} f(x)$, Fuji et al. (2025) developed randomized subspace regularized Newton (RSRN), which uses a gradient-dependent regularizer on the subspace Hessian to handle indefinite curvature. They establish global stationarity guarantees and local linear convergence. Near a minimizer $\mathbf{x}_\star$ with positive definite Hessian $\mathbf{H}_\star = \nabla^2 f(\mathbf{x}_\star)$, their analysis yields a local iteration bound of $\widetilde{O}(n\kappa/r)$, where κ is the condition number of $\mathbf{H}_\star$, and $\widetilde{O}(\cdot)$ hides polylog terms. The factor n/r reflects the restriction to r-dimensional updates. The factor κ, however, depends only on the two extreme eigenvalues and can obscure substantial spectral structure: a few large eigenvalues can make κ arbitrarily large even when the remaining spectrum is well conditioned. The relevance of finer spectral structure is apparent in learning problems: outliers separated from a spectral bulk have been observed in neural network Hessians (Sagun et al., 2017; Ghorbani et al., 2019; Papyan, 2019), while decaying kernel spectra motivate low-rank approximations in kernel methods (Bach, 2013). This leads to our central question:

Can randomized subspace Newton exploit spectral structure to achieve faster local convergence?

Our results: Two complementary improvements. We answer this question by developing Accelerated Randomized Subspace Regularized Newton (Acc-RSRN), whose local convergence benefits from both a sharper spectral analysis and Nesterov acceleration (Nesterov, 1983). Let $\lambda_1 \geq \cdots \geq \lambda_n > 0$ be the eigenvalues of $\mathbf{H}_\star$. For a spectral cutoff $1 \leq r < n$, define the *tail-average condition number*

$$\bar{\kappa}_r := \frac{1}{n-r}\sum_{i=r+1}^{n}\frac{\lambda_i}{\lambda_n}. \tag{1.1}$$

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.

Unlike $\kappa = \lambda_1/\lambda_n$, this quantity depends on the average of the tail eigenvalues rather than the largest eigenvalue. For randomized subspaces based on Gaussian sketching with size $s = \Theta(r)$, under the parameter and initialization conditions of our local theory, we obtain the following progression of improvements in the local iteration complexity of RSRN (see Theorem 3.3):

$$\underbrace{\widetilde{O}\Big(\frac{n\kappa}{r}\Big)}_{\text{prior RSRN analysis}} \longrightarrow \underbrace{\widetilde{O}\Big(\frac{n\bar{\kappa}_r}{r}\Big)}_{\text{our RSRN bound}} \longrightarrow \underbrace{\widetilde{O}\Big(\frac{n\sqrt{\bar{\kappa}_r}}{r}\Big)}_{\text{Acc-RSRN}}.$$

Here $\widetilde{O}$ suppresses logarithmic factors in the problem dimension and target accuracy. The first improvement comes from a sharper analysis of the regularized subspace updates, replacing κ with the smaller quantity $\bar{\kappa}_r$. The second comes from Nesterov acceleration, which reduces this dependence from linear to square root without increasing the subspace dimension. To do so, we derive first- and second-moment bounds for the regularized projection at $\mathbf{H}_\star$, capturing average progress and the variance relevant to acceleration.

Why tail averaging matters. Since $1 \le \bar{\kappa}_r \le \kappa$, our refined analysis already improves the bound substantially before acceleration is applied. As an example, suppose only the leading r eigenvalues are large, while the remaining satisfy $\lambda_{r+1}/\lambda_n = O(1)$. Then, $\bar{\kappa}_r = O(1)$ and the local iteration bound becomes $\widetilde{O}(n/r)$, irrespective of the leading eigenvalues. This idealized example is motivated by the outlier-bulk structure observed in neural-network Hessians (Sagun et al., 2017; Ghorbani et al., 2019; Papyan, 2019). However, the tail need not be tightly clustered for the refinement to be useful. Polynomially decaying spectra arise in kernel methods and neural-network kernel models (Bach, 2013; Bietti & Bach, 2021). With polynomial spectral decay $\lambda_i \propto i^{-\alpha}$ for a fixed $\alpha > 0$, we have $\kappa = \lambda_1/\lambda_n \asymp n^\alpha$, but the tail-average condition number satisfies:

$$\bar{\kappa}_r \asymp \begin{cases} 1, & 0 < \alpha < 1, \\ 1 + \log(n/r), & \alpha = 1, \\ (n/r)^{\alpha-1}, & \alpha > 1. \end{cases}$$

Note that averaging the tail yields an improvement beyond merely excluding the leading eigenvalues: For $\alpha > 1$, $\bar{\kappa}_r$ is smaller than $\lambda_{r+1}/\lambda_n \asymp (n/r)^\alpha$ by a factor of order n/r. As a concrete example, $\alpha = 2$ gives an accelerated local iteration bound of $\widetilde{O}((n/r)^{3/2})$, compared with $\widetilde{O}((n/r)^2)$ from our nonaccelerated analysis and $\widetilde{O}(n^3/r)$ from the prior RSRN bound. Thus, the gains extend beyond isolated spectral outliers to full-rank Hessians with gradual spectral decay.

Two Hessian access models. Both our algorithmic framework and convergence analysis support a broad range of optimization settings via two different paradigms for constructing the randomized subspaces, neither of which requires forming the entire Hessian:

1. *Hessian-vector products*: Hessian-vector products can be evaluated efficiently without forming the Hessian (Pearlmutter, 1994), making this access model useful for neural-network optimization and curvature analysis (Martens, 2010; Ghorbani et al., 2019). In this model, we probe the Hessian with Gaussian vectors to construct the randomized subspaces and obtain the sharpest dependence on the subspace dimension.
2. *Entry-wise Hessian access*: When computing individual entries of the Hessian is more tractable than matrix-vector products, we construct the subspace using coordinate sampling based on ridge leverage scores (RLS) (Rudi et al., 2018; Rathore et al., 2026), attaining nearly matching convergence guarantees.

Safeguarded global convergence and experiments. Beyond the local acceleration guarantee, our further algorithmic contribution is a sufficient-decrease safeguard that globalizes the method for smooth nonconvex optimization. During the global phase, an accelerated trial is accepted only if it provides sufficient objective decrease; otherwise, the method takes a nonaccelerated RSRN step using the same sketch and resets momentum. Under the stated assumptions, this mechanism guarantees that the method reaches a stationary point or enters a certified local regime (Section 4). Once local entry is certified, the method switches to the accelerated local recursion and disables the global safeguard, so that objective monotonicity does not obstruct acceleration. This design connects global convergence

guarantees with the improved local rates. Experiments with practical accelerated variants on convex kernel logistic regression and a nonconvex CNN objective illustrate the benefits of acceleration in RSRN, and highlight the advantages of both Hessian access models enabled by our framework.

Additional related work. Prior works have combined randomized subspace Newton with cubic regularization and provide convergence guarantees for convex objectives (Doikov & Richtárik, 2018; Hanzely et al., 2020). Global guarantees for nonconvex objectives are also available for random-subspace trust-region and regularization frameworks (Cartis et al., 2022; 2025), and the homogenized trust-region method (Higuchi et al., 2025). Acceleration provides a complementary way to improve the progress made by low-dimensional updates. A closely related method is accelerated double-sketched subspace Newton (Shang et al., 2024), which combines double sketching with Nesterov acceleration, but it assumes access to the Hessian square root (which we do not) and does not provide any spectrum-dependent convergence rates (which we do). From a first-order perspective, a question related to ours is addressed for accelerated coordinate descent (Nesterov & Stich, 2017) and randomized-subspace Nesterov methods under matrix smoothness (Omiya et al., 2026). Our approach builds on the analysis of randomized linear system solvers (all of which apply to convex quadratic minimization only) including accelerated sketch-and-project (Tu et al., 2017; Gower et al., 2018) and randomized block Kaczmarz (Dereziński et al., 2025); our coordinate construction builds on approximate ridge leverage score sampling and projection bounds of Rathore et al. (2026). Further related work, including other sketched and subsampled Newton methods, can be found in Appendix G.

Notation. We write $[n] = \{1, \dots, n\}$ and let $\mathbf{I}$ denote the identity matrix of the appropriate dimension. The norm $\|\cdot\|$ is Euclidean for vectors and spectral for matrices. For real symmetric matrices, $\mathbf{A} \succeq 0$ means positive semidefinite (PSD), $\mathbf{A} \succ 0$ means positive definite (PD), and $\mathbf{A} \preceq \mathbf{B}$ means $\mathbf{B} - \mathbf{A} \succeq 0$. For $\mathbf{A} \succeq 0$, $\mathbf{A}^{1/2}$ denotes its unique PSD square root and $\|\mathbf{x}\|_{\mathbf{A}} = (\mathbf{x}^\top \mathbf{A}\mathbf{x})^{1/2}$ is the induced seminorm. For $B \subseteq [n]$, $\mathbf{I}_B \in \mathbb{R}^{|B| \times n}$ consists of the rows of $\mathbf{I}$ indexed by B, and $\mathbf{A}_{BB} = \mathbf{I}_B \mathbf{A} \mathbf{I}_B^\top$ is the corresponding principal submatrix. Finally, $B(\mathbf{x}, R) = \{\mathbf{y} : \|\mathbf{y} - \mathbf{x}\| \le R\}$ denotes the Euclidean ball, and $\mathbb{1}_{\mathcal{E}}$ denotes the indicator of an event $\mathcal{E}$.

2 Accelerated Randomized Subspace Regularized Newton

Acc-RSRN uses a regularized subspace Newton step to generate candidates, a momentum update to accelerate convergence, and an acceptance rule to ensure global progress. We first describe the candidate step, then explain when it is accepted and how the method switches to local acceleration.

Algorithm 1 RSRN-Trial$(\mathbf{z}, \mathbf{S}; \eta)$

Require: Base $\mathbf{z}$, sketch $\mathbf{S} \in \mathbb{R}^{s \times n}$, $\eta \ge 0$; $c_1 > 1$, $c_2 > 0$, $\gamma \ge 0$, $\alpha, \beta \in (0, 1)$.

1: $\mathbf{g} \leftarrow \nabla f(\mathbf{z})$
2: **if** $\|\mathbf{g}\| = 0$ **then return** $(\mathbf{z}, 0, 0, 0, 0)$
3: $\mathbf{B} \leftarrow \mathbf{S}\nabla^2 f(\mathbf{z})\mathbf{S}^\top$; $\Lambda \leftarrow \max\{0, -\lambda_{\min}(\mathbf{B})\}$
4: $\mathbf{M} \leftarrow \mathbf{B} + (\eta + c_1\Lambda + c_2\|\mathbf{g}\|^\gamma)\mathbf{I}$; $\mathbf{p} \leftarrow -\mathbf{S}^\top \mathbf{M}^{-1} \mathbf{S}\mathbf{g}$
5: Choose the first $t \in \{1, \beta, \beta^2, \dots\}$ satisfying
$f(\mathbf{z} + t\mathbf{p}) \le f(\mathbf{z}) + \alpha t \mathbf{g}^\top \mathbf{p}$
6: **return** $(\mathbf{z} + t\mathbf{p},\ t,\ \mathbf{p},\ -\alpha t \mathbf{g}^\top \mathbf{p},\ \|\mathbf{g}\|)$

Regularizing randomized subspace Newton. Given an iterate $\mathbf{z} \in \mathbb{R}^n$ and a sketching matrix $\mathbf{S} \in \mathbb{R}^{s \times n}$, randomized subspace Newton (RSN) minimizing a convex twice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ takes a step in $\mathrm{range}(\mathbf{S}^\top)$ and makes the following update (Gower et al., 2019):

$$\mathbf{z}^+ = \mathbf{z} - t\mathbf{S}^\top \left(\mathbf{S}\nabla^2 f(\mathbf{z})\mathbf{S}^\top\right)^{-1} \mathbf{S}\nabla f(\mathbf{z})$$

where $t > 0$ is the step size and the reduced Hessian $\mathbf{S}\nabla^2 f(\mathbf{z})\mathbf{S}^\top$ is assumed positive definite. For nonconvex objectives, however, this reduced Hessian may be singular or indefinite. To address this issue, Fuji et al. (2025) proposed randomized subspace regularized Newton (RSRN). We use their regularized step as the update routine in Algorithm 1, additionally allowing a ridge floor $\eta \ge 0$. Specifically, the routine replaces the reduced Hessian with

$$\mathbf{S}\nabla^2 f(\mathbf{z})\mathbf{S}^\top + \left(\eta + c_1\Lambda + c_2\|\nabla f(\mathbf{z})\|^\gamma\right)\mathbf{I}, \qquad \text{where } \Lambda = \max\{0, -\lambda_{\min}(\mathbf{S}\nabla^2 f(\mathbf{z})\mathbf{S}^\top)\}.$$

The terms inherited from RSRN have complementary roles: $c_1\Lambda$ offsets negative curvature, while $c_2\|\nabla f(\mathbf{z})\|^\gamma$ stabilizes the reduced solve and vanishes near stationarity for $\gamma > 0$. The additional ridge floor η is zero in the global phase and is chosen in the local phase to obtain our spectral-tail guarantees. We use Armijo search to select the step size t (see Algorithm 1).

Gaussian sketching and coordinate sampling. Our algorithms support different ways of choosing the subspace. A popular choice is Gaussian sketching, where $\mathbf{S} \in \mathbb{R}^{s\times n}$ has independent $\mathcal{N}(0, 1/s)$ entries. This approach is most effective in the Hessian-vector query model, where forming $\mathbf{S}\nabla^2 f(\mathbf{z})\mathbf{S}^\top$ uses s such queries (see Section 3.3.1). In the entry-wise access model, coordinate sampling is more practical. Here, $\mathbf{S} = \mathbf{I}_B$ selects a random set $B \subseteq \{1, ..., n\}$ of s coordinates, so that the reduced Hessian is the principal block $[\nabla^2 f(\mathbf{z})]_{BB}$. Our spectral guarantees use approximate ridge leverage score sampling to select the coordinates (see Section 3.3.2).

Algorithm 2 Safeguarded Acc-RSRN with Certified Local Entry

Require: $\mathbf{x}_0$; trial parameters; $\sigma \in (0, 1]$, $R_{\max} > 0$, $\varepsilon > 0$; GLOBAL law $\mathcal{D}_{\mathrm{g}}$, momentum $\{a_k, b_k, \theta_k\}$, and certified local-entry test LOCAL-CERT (Section 4; Definition C.3).

$\mathbf{v}_0 \leftarrow \mathbf{x}_0$, mode $\leftarrow$ GLOBAL, $\eta \leftarrow 0$
for $k = 0, 1, 2, \ldots$ **do**
  if mode $=$ GLOBAL **and** LOCAL-CERT$(\mathbf{x}_k) =$ **true then**
    mode $\leftarrow$ LOCAL, $\mathbf{v}_k \leftarrow \mathbf{x}_k$
    Freeze the returned ridge η, law $\mathcal{D}_{\mathrm{loc}}$, and bounds $(\underline{\mu}, \overline{\nu}, \delta_{\mathrm{loc}})$
    Set $(a_j, b_j, \theta_j) = (a, b, \theta)$ from (3.4) for all $j \geq k$
  end if
  $\mathbf{y}_k \leftarrow a_k\mathbf{v}_k + (1 - a_k)\mathbf{x}_k$
  if mode $=$ GLOBAL and $\|\mathbf{y}_k - \mathbf{x}_k\| > R_{\max}$ **then** $\mathbf{v}_k \leftarrow \mathbf{x}_k$, $\mathbf{y}_k \leftarrow \mathbf{x}_k$
  Draw $\mathbf{S}_k \sim \mathcal{D}_{\mathrm{g}}$ if GLOBAL, and $\mathbf{S}_k \sim \mathcal{D}_{\mathrm{loc}}$ if LOCAL
  $(\widetilde{\mathbf{x}}_{k+1}, t_k^A, \mathbf{p}_k^A, \mathsf{Dec}_k^A, h_k^A) \leftarrow$ RSRN-TRIAL$(\mathbf{y}_k, \mathbf{S}_k; \eta)$
  if $h_k^A \leq \varepsilon$ **then return** $\mathbf{y}_k$
  if mode $=$ LOCAL **or** $f(\widetilde{\mathbf{x}}_{k+1}) \leq f(\mathbf{x}_k) - \sigma\mathsf{Dec}_k^A$ **then**
    $\mathbf{x}_{k+1} \leftarrow \widetilde{\mathbf{x}}_{k+1}$
    $\mathbf{v}_{k+1} \leftarrow b_k\mathbf{v}_k + (1 - b_k)\mathbf{y}_k + \theta_k t_k^A \mathbf{p}_k^A$
  else
    $(\mathbf{x}_{k+1}, t_k^B, \mathbf{p}_k^B, \mathsf{Dec}_k^B, h_k^B) \leftarrow$ RSRN-TRIAL$(\mathbf{x}_k, \mathbf{S}_k; 0)$
    if $h_k^B \leq \varepsilon$ **then return** $\mathbf{x}_k$
    $\mathbf{v}_{k+1} \leftarrow \mathbf{x}_{k+1}$ ▷ same-sketch fallback and restart
  end if
end for

Momentum and safeguards. Acc-RSRN maintains an iterate $\mathbf{x}_k$ and a momentum state $\mathbf{v}_k$. When the momentum is not reset and the trial at the momentum point is accepted, the update has the form

$$\mathbf{y}_k = a_k\mathbf{v}_k + (1 - a_k)\mathbf{x}_k, \qquad \mathbf{x}_{k+1} = \mathbf{y}_k + t_k^A\mathbf{p}_k^A, \qquad \mathbf{v}_{k+1} = b_k\mathbf{v}_k + (1 - b_k)\mathbf{y}_k + \theta_k t_k^A\mathbf{p}_k^A,$$

where t_k^A and $\mathbf{p}_k^A$ are returned by Algorithm 1 at $\mathbf{y}_k$. The choice $a_k = b_k = 0, \theta_k = 1$ gives RSRN. The GLOBAL phase sets $\eta = 0$ and applies two safeguards. If $\|\mathbf{y}_k - \mathbf{x}_k\| > R_{\max}$, the momentum is reset before the trial. The accelerated candidate is accepted only if $f(\widetilde{\mathbf{x}}_{k+1}) \leq f(\mathbf{x}_k) - \sigma\mathsf{Dec}_k^A$, i.e., only if it decreases the objective sufficiently relative to the stored iterate $\mathbf{x}_k$. Otherwise, the method takes an RSRN step from $\mathbf{x}_k$ using the same sketch and resets momentum. This fallback is computed only after rejection. These rules preserve objective monotonicity during the global phase and supply the decrease mechanism for the stationarity-or-entry guarantee in Section 4.

Certified local entry. Before drawing the next sketch, the test LOCAL-CERT checks whether the stored iterate admits a certified local configuration (Section 4). On success, the method enters the LOCAL mode, freezes the local ridge and sketch law, and fixes the momentum coefficients. Failed tests instead leave the global phase unchanged. The local configuration includes the moment bounds $\underline{\mu}, \overline{\nu}$ and a local energy scale δ_{loc}. The Hessians at subsequent base points continue to be updated.

In LOCAL, both the radius restart and the global sufficient-decrease safeguard are disabled, while Armijo backtracking is retained. Local acceleration contracts a two-state energy and need not decrease the objective at every iteration. The certified switch therefore recovers the pure accelerated recursion rather than relying on the safeguard to disappear on its own. Algorithm 2 gives the complete method.

3 LOCAL CONVERGENCE WITH ACCELERATION

We first analyze the LOCAL recursion of Algorithm 2, which is our main contribution. We state the assumptions and local theorem, give a proof sketch by explaining how nonlinear perturbations preserve quadratic acceleration, and derive explicit convergence rates for both Hessian access models.

3.1 ASSUMPTIONS AND MAIN CONVERGENCE THEOREM

Assumption 3.1 (Local objective). *The point $\mathbf{x}_\star$ is stationary, with $\mathbf{H}_\star = \nabla^2 f(\mathbf{x}_\star) \succeq h_- \mathbf{I} \succ 0$. On $B(\mathbf{x}_\star, R)$, the Hessian is L_H-Lipschitz and has norm at most U, where $R, h_-, U, L_H > 0$.*

For a fixed ridge $\eta \geq 0$ and sketch law $\mathcal{D}_{\mathrm{loc}}$, define

$$\mathbf{P}_\eta(\mathbf{S}) = \mathbf{H}_\star^{1/2}\mathbf{S}^\top(\mathbf{S}\mathbf{H}_\star\mathbf{S}^\top + \eta\mathbf{I})^{-1}\mathbf{S}\mathbf{H}_\star^{1/2}, \qquad \overline{\mathbf{P}}_\eta = \mathbb{E}_{\mathbf{S}\sim\mathcal{D}_{\mathrm{loc}}}\mathbf{P}_\eta(\mathbf{S}). \tag{3.1}$$

Note that $0 \preceq \mathbf{P}_\eta(\mathbf{S}) \preceq \mathbf{I}$ whenever the reduced inverse exists. These operators describe the key properties of the randomized subspace generated by $\mathbf{S} \sim \mathcal{D}_{\mathrm{loc}}$ relative to the Hessian $\mathbf{H}_\star$. We next delineate those properties in terms of the first and second order matrix moments of $\mathbf{P}_\eta(\mathbf{S})$.

Assumption 3.2 (Subspace moment bounds). *Each local sketch is a fresh draw from $\mathcal{D}_{\mathrm{loc}}$, independent of the preceding local history. Available scalar bounds $0 < \underline{\mu} < \overline{\nu} < \infty$ satisfy*

$$\overline{\mathbf{P}}_\eta \succeq \underline{\mu}\mathbf{I}, \qquad \mathbb{E}[\mathbf{P}_\eta(\mathbf{S})\overline{\mathbf{P}}_\eta^{-1}\mathbf{P}_\eta(\mathbf{S})] \preceq \overline{\nu}\overline{\mathbf{P}}_\eta. \tag{3.2}$$

For a sketch-independent constant $\ell_\eta > 0$, the reduced systems satisfy

$$\mathbf{S}\mathbf{H}_\star\mathbf{S}^\top + \eta\mathbf{I} \succeq \ell_\eta\mathbf{I} \quad \textit{almost surely}. \tag{3.3}$$

The first bound controls average progress, while the second controls the variability entering the momentum update. For $\eta > 0$, the reduced-system condition holds with $\ell_\eta = \eta$. Only the zero-floor case needs additional uniform conditioning. Set

$$q = \sqrt{\underline{\mu}/\overline{\nu}}, \qquad a = \frac{q}{1+q}, \qquad b = 1 - q, \qquad \theta = \frac{1}{\sqrt{\underline{\mu}\overline{\nu}}}. \tag{3.4}$$

For analysis, we use the fixed Hessian metric and denote the following:

$$\mathbf{z}_k = \mathbf{H}_\star^{1/2}(\mathbf{x}_k - \mathbf{x}_\star), \qquad \mathbf{w}_k = \mathbf{H}_\star^{1/2}(\mathbf{v}_k - \mathbf{x}_\star), \tag{3.5}$$

$$\Delta_k = \|\mathbf{w}_k\|^2_{\overline{\mathbf{P}}_\eta^{-1}} + \underline{\mu}^{-1}\|\mathbf{z}_k\|^2. \tag{3.6}$$

The fixed metric permits comparison with a single quadratic reference, even though the algorithm evaluates the Hessian at each current base point. Neither $\mathbf{H}_\star^{1/2}$ nor $\overline{\mathbf{P}}_\eta$ is computed.

Theorem 3.3 (Local acceleration). *Suppose Assumptions 3.1 and 3.2 hold. Run the* LOCAL *recursion of Algorithm 2 with $\alpha \in (0, 1/2)$, $c_2 > 0$, $0 < \gamma \leq 1/2$, and the parameters in (3.4). Set*

$$\rho = 1 - \tfrac{1}{2}q = 1 - \tfrac{1}{2}\sqrt{\underline{\mu}/\overline{\nu}}. \tag{3.7}$$

There exist $\delta, C_f > 0$, depending only on the fixed local problem and algorithmic parameters, such that for every fixed initial state with $\Delta_0 \leq \delta/100$, event $\mathcal{G} = \{\Delta_j \leq \delta \text{ for all } j \geq 0\}$ satisfies $\mathbb{P}(\mathcal{G}) \geq 0.99$, and for all $k \geq 0$,

$$\mathbb{E}[\Delta_k \mid \mathcal{G}] \leq 2\rho^k\Delta_0, \tag{3.8}$$

$$\mathbb{E}[f(\mathbf{x}_k) - f(\mathbf{x}_\star) \mid \mathcal{G}] \leq C_f\rho^k\Delta_0. \tag{3.9}$$

Conditional on $\mathcal{G}$, both states converge to $\mathbf{x}_\star$ almost surely.

The inverse contraction gap is $O(\sqrt{\overline{\nu}/\underline{\mu}})$, versus $O(1/\underline{\mu})$ for the nonaccelerated update of vanilla RSRN (Proposition E.2). Thus a sharper first moment improves the baseline bound, while a sufficiently small second moment enables a further acceleration gain. The radius δ is explicit in Appendix A; it can depend on the local problem and algorithmic parameters, so the rate alone does not specify the size of the admissible initialization neighborhood.

3.2 Outline of the convergence analysis

The proof of Theorem 3.3 proceeds from quadratic contraction to nonlinear stability, then to control of local exit. Let $\mathcal{F}_k$ denote the history before drawing $\mathbf{S}_k$; full proofs are in Appendix A.

Step 1: A contracting quadratic reference. We replace f by its quadratic Taylor expansion at $\mathbf{x}_\star$, retain η, omit the gradient-dependent ridge, and take unit steps. The resulting accelerated regularized sketch-and-project recursion contracts Δ_k by $1-q$ in conditional expectation, under the assumptions of Section 3.1 (see Proposition A.1; based on Gower et al., 2018; Dereziński et al., 2025).

Step 2: Nonlinear errors preserve the contraction. The comparison must account for Hessian variation, the gradient-dependent ridge, and line search. The following lemma shows that their combined effect preserves at least half of the quadratic contraction gap.

Lemma 3.4 (Uniform nonlinear and regularization perturbation)**.** *Under the hypotheses of Theorem 3.3, if $\Delta_k \le \delta$ with δ from (A.6), then $\Lambda_k = 0$ and the unit step passes Armijo for every admissible sketch. Let $\mathbf{\Phi}_{\mathbf{S}}$ be the quadratic reference map in (A.2), and set $\mathbf{G}_\Delta = \operatorname{diag}(\underline{\mu}^{-1}\mathbf{I}, \overline{\mathbf{P}}_\eta^{-1})$. Then*

$$\begin{bmatrix}\mathbf{z}_{k+1}\\ \mathbf{w}_{k+1}\end{bmatrix} = \mathbf{\Phi}_{\mathbf{S}_k}\begin{bmatrix}\mathbf{z}_k\\ \mathbf{w}_k\end{bmatrix} + \mathbf{e}_k, \qquad \|\mathbf{e}_k\|_{\mathbf{G}_\Delta} \le (q/8)\sqrt{\Delta_k}, \tag{3.10}$$

$$\mathbb{E}[\Delta_{k+1} \mid \mathcal{F}_k] \le \left(\sqrt{1-q} + q/8\right)^2 \Delta_k \le (1-q/2)\Delta_k. \tag{3.11}$$

Proof idea. The energy gives $\|\mathbf{y}_k - \mathbf{x}_\star\| \le \sqrt{\Delta_k/h_-}$. For small enough Δ_k, local positive curvature and a Taylor bound give $\Lambda_k = 0$ and unit Armijo acceptance. To compare the resulting step with the reference for the same sketch, we use $\|(\mathbf{S}\mathbf{H}_\star\mathbf{S}^\top + \eta\mathbf{I})^{-1/2}\mathbf{S}\mathbf{H}_\star^{1/2}\| \le 1$. This normalization and the uniform lower bound (3.3) control Hessian variation and the extra ridge uniformly over sketches. The gradient Taylor remainder then yields $\|\mathbf{e}_k\|_{\mathbf{G}_\Delta} = O(\Delta_k^{(1+\gamma)/2})$. Because $\gamma > 0$, the error is at most $(q/8)\sqrt{\Delta_k}$ for a sufficiently small threshold δ. The conditional L^2 triangle inequality combines this error with the reference contraction to give (3.11). The inverse comparison and Armijo estimates are detailed in Appendix A.2.

Step 3: Controlling exit before conditioning. The one-step bound is only local: a random update can still exit the neighborhood. Let $T = \inf\{k : \Delta_k > \delta\}$. The lemma applies before T, including the update that exits. A stopped-energy argument and a separate estimate on surviving trajectories give

$$\mathbb{P}(T < \infty) \le \frac{\Delta_0}{\delta}, \qquad \mathbb{E}[\Delta_k \mathbb{1}_{\{T>k\}}] \le \rho^k \Delta_0.$$

Hence $\Delta_0 \le \delta/100$ ensures $\mathbb{P}(\mathcal{G}) \ge 0.99$ for $\mathcal{G} = \{T = \infty\}$. Since $\mathcal{G} \subseteq \{T > k\}$, conditioning only after these estimates gives $\mathbb{E}[\Delta_k \mid \mathcal{G}] \le \rho^k \Delta_0/\mathbb{P}(\mathcal{G}) \le 2\rho^k\Delta_0$. Local curvature and summability give the remaining conclusions (Appendix A.3).

3.3 Spectral-tail rates under two Hessian access models

We now derive explicit local convergence rates for Gaussian sketching (motivated by Hessian-vector queries) and ridge leverage score sampling (motivated by entry-wise Hessian access). All eigenvalues in this section are those of $\mathbf{H}_\star$, and $\bar{\kappa}_r$ denotes the tail-average condition number defined in (1.1).

3.3.1 Gaussian sketches

Theorem 3.5 (Gaussian local convergence)**.** *Suppose Assumption 3.1 holds. Let $1 \le r \le s/2$ and let $\mathbf{S} \in \mathbb{R}^{s\times n}$ with i.i.d. entries $\mathbf{S}_{ij} \sim \mathcal{N}(0, 1/s)$. Set the ridge parameter $\eta_{\mathrm{G}} = \frac{1}{s}\sum_{i=r+1}^n \lambda_i$ and the momentum rule (3.4) with the bounds in (3.13) below. Under the remaining conditions of Theorem 3.3, we obtain local convergence result*

$$\mathbb{E}[\Delta_k \mid \mathcal{G}] \le 2\left(1 - c\frac{s}{n\sqrt{\bar{\kappa}_r}}\right)^k \Delta_0, \tag{3.12}$$

where $c > 0$ is an absolute constant. Consequently, given $0 < \varepsilon < 1$, $\mathbb{E}[\Delta_k \mid \mathcal{G}] \le \varepsilon\Delta_0$ is reached within $O(\frac{n\sqrt{\bar{\kappa}_r}}{s} \log \frac{2}{\varepsilon})$ iterations. Specifically, by choosing $s = 2r$ we obtain $\widetilde{O}(\frac{n\sqrt{\bar{\kappa}_r}}{r})$.

Proof sketch. With the choice of the ridge parameter η_{G}, Lemma B.1 gives $\overline{\mathbf{P}}_{\eta_{\mathrm{G}}} \succeq c_{\mathrm{G}}\mathbf{H}_\star(\mathbf{H}_\star + 2\eta_{\mathrm{G}}\mathbf{I})^{-1}$. The ridge choice also allows Lemma B.2 to bound the second moment independently of $\bar{\kappa}_r$. Together, these give valid bounds

$$\underline{\mu}_{\mathrm{G}} = \frac{c_{\mathrm{G}} s}{4n\bar{\kappa}_r}, \qquad \overline{\nu}_{\mathrm{G}} = C_{\mathrm{G}} \frac{n}{s}, \tag{3.13}$$

for constants $c_{\mathrm{G}} > 0$ and $C_{\mathrm{G}} \ge 1$. Substituting into Theorem 3.3 proves (3.12) with $c = \frac{1}{4}\sqrt{c_{\mathrm{G}}/C_{\mathrm{G}}}$. These Gaussian moment bounds are of independent interest as they improve on the convergence rates of other randomized subspace algorithms (Gower et al., 2018); see Appendices B.1.1–B.1.3.

3.3.2 Ridge leverage score sampling

For entry-wise access, we adapt the approximate ridge leverage score (ARLS) construction of Rathore et al. (2026), and sample from a shifted Hessian. Let $\mathbf{A} = \mathbf{H}_\star + \eta_{\mathrm{R}}\mathbf{I}$ with ridge parameter choice $\eta_{\mathrm{R}} = \frac{1}{n-r}\sum_{i=r+1}^{n} \lambda_i$. We draw coordinates independently by using upper approximations of the ridge leverage scores $\ell_i^{\zeta}(\mathbf{A}) = [\mathbf{A}(\mathbf{A} + \zeta\mathbf{I})^{-1}]_{ii}$ as sampling probabilities, see Lemma B.5.

Theorem 3.6 (ARLS local convergence)**.** *Suppose Assumption 3.1 holds, and let $C_0 \log n \le r < n/4$ for some constant C_0. At each local iteration, independently draw $s = \Omega(r \log^3 n)$ coordinates as in Lemma B.5, retain their union B, and define the subsampling matrix $\mathbf{S} = \mathbf{I}_B$. Use the momentum rule (3.4) with the bounds in (3.15) below. Under the remaining conditions of Theorem 3.3, we obtain*

$$\mathbb{E}[\Delta_k \mid \mathcal{G}] \le 2\left(1 - \frac{r}{8n\sqrt{\bar{\kappa}_r}}\right)^k \Delta_0. \tag{3.14}$$

Given $0 < \varepsilon < 1$, $\mathbb{E}[\Delta_k \mid \mathcal{G}] \le \varepsilon\Delta_0$ is reached within $O(\frac{n\sqrt{\bar{\kappa}_r}}{r} \log \frac{2}{\varepsilon})$ iterations.

Proof sketch. Since $\mathbf{A}_{BB} = (\mathbf{H}_\star)_{BB} + \eta_{\mathrm{R}}\mathbf{I}$, the ARLS projection bound of Rathore et al. (2026) transfers to our regularized Newton operator. Combining it with Lemma B.6, which adapts the regularized projection argument of Dereziński et al. (2025), we obtain

$$\underline{\mu}_{\mathrm{R}} = \frac{r}{4n\bar{\kappa}_r}, \qquad \overline{\nu}_{\mathrm{R}} = \frac{4n}{r}. \tag{3.15}$$

The stated rate now follows from Theorem 3.3; see Appendices B.2.1 and B.2.3 for details.

Comparison between the approaches. At a common spectral cutoff r, both methods attain a local iteration bound of $O((n\sqrt{\bar{\kappa}_r}/r)\log(2/\varepsilon))$ for reducing $\mathbb{E}[\Delta_k \mid \mathcal{G}]$ to $\varepsilon\Delta_0$, under the conditions of Theorem 3.3. Their Hessian access costs differ. Gaussian sketching uses $s = \Theta(r)$ Hessian-vector products per iteration, with $O(n\sqrt{\bar{\kappa}_r}\log(2/\varepsilon))$ products in total. For ARLS, constant-factor score approximations can be constructed with high probability using $\widetilde{O}(nr)$ Hessian-entry evaluations via BLESS (Rudi et al., 2018; Rathore et al., 2026); the entry count is derived in Appendix B.2.3. Each iteration draws $s = \Theta(r \log^3 n)$ coordinates and accesses $s^2 = O(r^2 \log^6 n)$ Hessian entries. In total, ARLS requires $\widetilde{O}(nr) + O\big(nr \log^6 n\sqrt{\bar{\kappa}_r} \log \frac{2}{\varepsilon}\big) = \widetilde{O}(nr\sqrt{\bar{\kappa}_r}\log(1/\varepsilon))$ Hessian entry evaluations, which for moderate $\bar{\kappa}_r$ is sublinear in the size of the Hessian.

4 Global Convergence and Switching to Local Acceleration

Section 3 assumes a sufficiently close starting point with fixed local parameters. We now overview the global convergence of our algorithm and describe the sufficient test for entering the local phase. See Appendix D for complete statements and proof details.

Convergence before the switch. In GLOBAL mode, the ridge floor is zero and an accelerated candidate is accepted only if

$$f(\widetilde{\mathbf{x}}_{k+1}) \le f(\mathbf{x}_k) - \sigma \mathrm{Dec}_k^A, \qquad \mathrm{Dec}_k^A = -\alpha t_k^A \nabla f(\mathbf{y}_k)^\top \mathbf{p}_k^A. \tag{4.1}$$

If the candidate is rejected, a nonaccelerated trial from $\mathbf{x}_k$ uses the same sketch and supplies the decrease instead: thus every completed global iteration satisfies $f(\mathbf{x}_{k+1}) \le f(\mathbf{x}_k)$. In Theorem D.7, under standard smoothness assumptions on the objective (Assumption D.1) and on the sketching law (Assumption D.2), we show that for any $\varepsilon > 0$ and $\delta_{\mathrm{g}} \in (0,1)$, with probability at least $1 - \delta_{\mathrm{g}}$ the method either queries a point $\mathbf{z}$ with $\|\nabla f(\mathbf{z})\| \le \varepsilon$, or starts the local phase within $O(\varepsilon^{-2} + \log(1/\delta_{\mathrm{g}}))$ iterations (see Corollary D.8 and Appendix D for the proof).

A sufficient test for local entry. A small gradient alone does not justify switching, since it may also occur near a saddle point. At a stored iterate $\mathbf{x}_c$, the test requires a valid positive lower bound on the full Hessian, $\mathbf{H}_c = \nabla^2 f(\mathbf{x}_c) \succeq m\mathbf{I}$, $m > 0$, and bounds U_{loc}, L_H on the Hessian norm and its Lipschitz constant on $B(\mathbf{x}_c, m/(2L_H))$. Proposition C.2 uses these bounds and spectral information at $\mathbf{H}_c$ to choose the ridge, sampling distribution, and moment bounds $\underline{\mu}, \overline{\nu}$, without knowing $\mathbf{x}_\star$ or $\mathbf{H}_\star$. The radius construction in Appendix A then supplies an energy threshold δ_{loc}, as specified in Definition C.3. For a chosen exit-probability bound $\delta_{\mathrm{exit}} \in (0,1)$, the test accepts when

$$\frac{8\|\nabla f(\mathbf{x}_c)\|^2}{\underline{\mu} m} \le \delta_{\mathrm{exit}} \delta_{\mathrm{loc}}.$$

Lemma C.4 shows that passing this test ensures a nearby nondegenerate minimizer and, after resetting $\mathbf{v} = \mathbf{x}_c$, bounds the initial two-state energy by the left-hand side, ensuring the local condition.

Convergence after the switch. The test is evaluated before drawing the next sketch. On success, Algorithm 2 resets momentum, fixes the ridge, sampling distribution, and momentum coefficients, and removes the global safeguard. It continues to draw fresh sketches and evaluate the Hessian at the current base point. Conditional on a valid switch, Theorem C.5 bounds the probability of leaving the local neighborhood by δ_{exit}; conditional on the switching history and no exit, the expected error decays geometrically with factor $1 - \frac{1}{2}\sqrt{\underline{\mu}/\overline{\nu}}$.

We put everything together in the combined global and local convergence guarantee of Theorem D.9. Under the aforementioned assumptions we show that, with high probability, either an ε-stationary point is queried within the global iteration bound, or a valid switch occurs within that bound and the continued local iteration converges to a nearby minimizer without leaving the local neighborhood.

5 Experiments

We compare Acc-RSRN with RSRN, accelerated gradient descent (AGD), and Adam (Kingma & Ba, 2014) on convex kernel logistic regression (KLR) and a nonconvex CNN objective. We evaluate convergence in terms of both iterations and wall-clock time, under matrix-vector product and entry-wise Hessian access models. In these experiments, Acc-RSRN uses empirical momentum choices without the certified local-entry test. Randomized experiments are averaged over 10 runs with different random seeds. Full experimental details are in Appendix F.

5.1 Kernel logistic regression

We study KLR as an example where entry-wise Hessian access is the most practical computing model. The objective uses 5,000 `real-sim` data points, an RBF kernel, and regularizer $\lambda_{\mathrm{reg}} = 10^{-4}$. Both Acc-RSRN and RSRN use mixed uniform/Hessian-diagonal coordinate sampling (which is cheaper to compute than the ARLS sampling used in our theory) with $s \in \{100, 400\}$, gradient-dependent damping, and Armijo backtracking. Acc-RSRN adapts momentum from independent gradient probes and uses monotone restarts with resampled fallback blocks.

Figure 1 shows that acceleration improves convergence for both block sizes under the same sampling policy. For $s = 400$, all ten runs of both methods reach an objective gap of 10^{-8}, with Acc-RSRN reducing the median solver time from 91.2 to 58.6 seconds. In Appendix F.1, we report paired 95% confidence intervals over the ten runs to quantify the benefit of acceleration, supporting a statistically significant improvement for both block sizes.

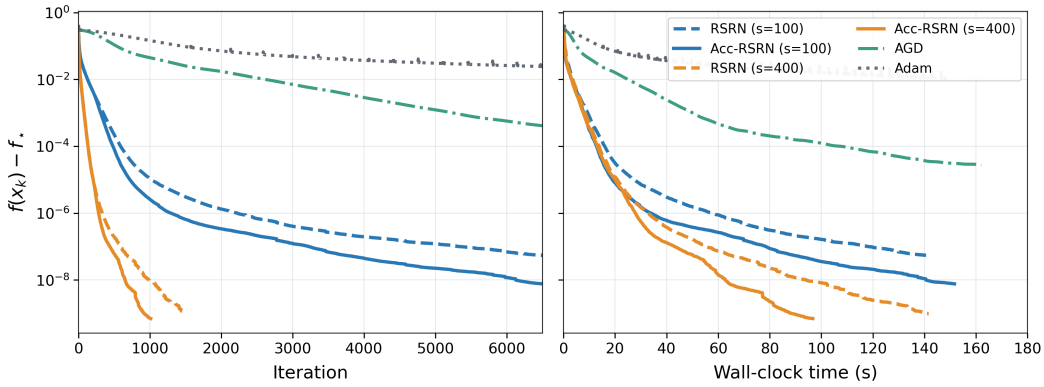

Figure 1: KLR objective difference versus iterations and solver time for Acc-RSRN and baselines.

5.2 Fixed-batch CNN optimization

We next consider a nonconvex CNN objective, where Hessian is most effectively accessed via matrix-vector products. Concretely, we minimize mean cross-entropy on 256 fixed MNIST images using a 1,710-parameter ReLU/max-pool CNN. RSRN and Acc-RSRN both use Gaussian sketching to generate random subspaces, with $s \in \{50, 100\}$ and $c_2 = 1/s$. Acc-RSRN uses fixed momentum $(q, \theta) = (0.03, 3)$ with same-sketch fallback and radius restart. Appendix F.2 gives the parameter selection details.

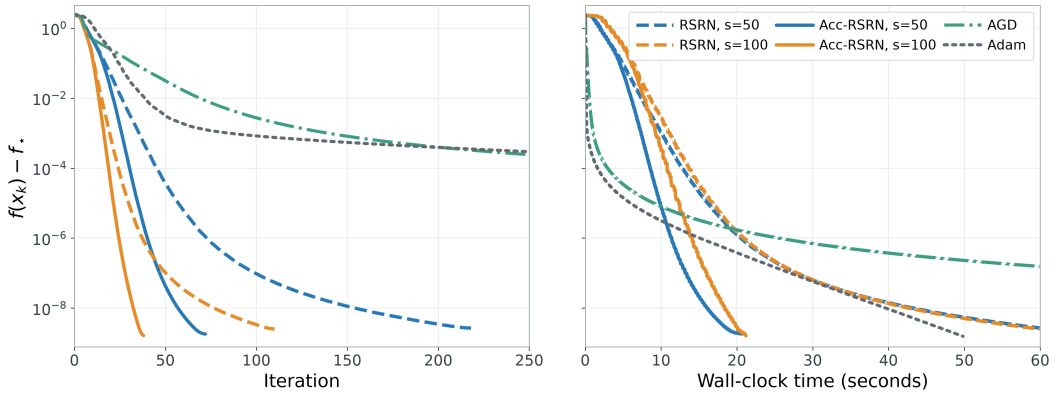

Figure 2: CNN objective difference versus iterations and solver time for Acc-RSRN and baselines.

Figure 2 shows that acceleration improves the convergence of Acc-RSRN for both sketch sizes even on this nonconvex objective. At the objective gap 10^{-8}, Acc-RSRN reduces mean solver time over RSRN from 43.04 to 15.99 seconds for $s = 50$, and from 42.68 to 18.19 seconds for $s = 100$. In Appendix F.2 we present 95% bootstrap confidence intervals over the 10 independent runs, showing that there is a statistically significant improvement from acceleration of at least factor 2 for both sketch sizes (see Table 2 for details). In comparison, Adam takes 39.63 seconds, while AGD does not attain the target objective value within 60 seconds.

6 Conclusions

We proposed Acc-RSRN, a randomized subspace Newton method with Nesterov acceleration and regularized projections, which attains a fast local linear convergence rate depending on a tail-average condition number of the Hessian at the optimum and a global convergence to a stationary point for nonconvex objectives. Our theory encompasses two different Hessian access models, and our experiments verify the effectiveness of local acceleration for both convex and nonconvex objectives.

AI USE STATEMENT

Generative AI tools assisted manuscript editing, refining mathematical derivations and proof arguments, and implementing and checking code for supplementary diagnostic experiments. AI-generated suggestions were independently evaluated and verified by the authors rather than treated as evidence of correctness. The authors take full responsibility for the paper's final text, mathematical claims, experimental results, and conclusions.

REPRODUCIBILITY STATEMENT

Appendices A–E contain complete proofs of all theoretical results, including the certification details for the local switch. Appendix F specifies the experimental protocols: problem instances, all hyperparameters, sketch and sampling distributions, stopping rules, seeds, and the statistical procedures behind the reported confidence intervals.

REFERENCES

Francis Bach. Sharp analysis of low-rank kernel matrix approximations. In *Conference on Learning Theory*, pp. 185–209. PMLR, 2013.

Alberto Bietti and Francis Bach. Deep equals shallow for ReLU networks in kernel regimes. In *International Conference on Learning Representations*, 2021.

Coralia Cartis, Nicholas I. M. Gould, and Philippe L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results. *Mathematical Programming*, 127(2):245–295, 2011.

Coralia Cartis, Jaroslav Fowkes, and Zhen Shao. Randomised subspace methods for non-convex optimization, with applications to nonlinear least-squares. *arXiv preprint arXiv:2211.09873*, 2022.

Coralia Cartis, Zhen Shao, and Edward Tansley. Random subspace cubic-regularization methods, with applications to low-rank functions. *arXiv preprint arXiv:2501.09734*, 2025.

Kenneth R. Davidson and Stanislaw J. Szarek. Local operator theory, random matrices and Banach spaces. In *Handbook of the Geometry of Banach Spaces*, volume 1, chapter 8, pp. 317–366. North-Holland, 2001. doi: 10.1016/S1874-5849(01)80010-3.

Michał Dereziński and Elizaveta Rebrova. Sharp analysis of sketch-and-project methods via a connection to randomized singular value decomposition. *SIAM Journal on Mathematics of Data Science*, 6(1):127–153, 2024.

Michał Dereziński, Jonathan Lacotte, Mert Pilanci, and Michael W. Mahoney. Newton-LESS: Sparsification without trade-offs for the sketched Newton update. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Michał Dereziński, Deanna Needell, Elizaveta Rebrova, and Jiaming Yang. Randomized Kaczmarz methods with beyond-Krylov convergence. *SIAM Journal on Matrix Analysis and Applications*, 46(4):2558–2588, 2025.

Nikita Doikov and Peter Richtárik. Randomized block cubic Newton method. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

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

Terunari Fuji, Pierre-Louis Poirion, and Akiko Takeda. Theoretical analysis of the randomized subspace regularized Newton method for non-convex optimization. *Open Journal of Mathematical Optimization*, 6:8, 2025. doi: 10.5802/ojmo.46. URL https://ojmo.centre-mersenne.org/articles/10.5802/ojmo.46/.

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via Hessian eigenvalue density. In *International Conference on Machine Learning*, pp. 2232–2241. PMLR, 2019.

Robert M. Gower and Peter Richtárik. Randomized iterative methods for linear systems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1660–1690, 2015.

Robert M. Gower, Filip Hanzely, Peter Richtárik, and Sebastian U. Stich. Accelerated stochastic matrix inversion: general theory and speeding up BFGS rules for faster second-order optimization. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

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

Filip Hanzely, Nikita Doikov, Yurii Nesterov, and Peter Richtárik. Stochastic subspace cubic Newton method. In *International Conference on Machine Learning*, pp. 4027–4038. PMLR, 2020.

Rei Higuchi, Pierre-Louis Poirion, and Akiko Takeda. Improving convergence guarantees of random subspace second-order algorithm for nonconvex optimization. In *International Conference on Learning Representations*, 2025.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Yurii Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

Yurii Nesterov. *Lectures on Convex Optimization*. Springer, 2nd edition, 2018.

Yurii Nesterov and Boris T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Yurii Nesterov and Sebastian U. Stich. Efficiency of the accelerated coordinate descent method on structured optimization problems. *SIAM Journal on Optimization*, 27(1):110–123, 2017.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, 2nd edition, 2006.

Brendan O'Donoghue and Emmanuel Candès. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15(3):715–732, 2015.

Gaku Omiya, Pierre-Louis Poirion, and Akiko Takeda. Randomized subspace Nesterov accelerated gradient. *arXiv preprint arXiv:2605.00740*, 2026.

Vardan Papyan. Measurements of three-level hierarchical structure in the outliers in the spectrum of deepnet Hessians. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 5012–5021, 2019.

Barak A. Pearlmutter. Fast exact multiplication by the Hessian. *Neural Computation*, 6(1):147–160, 1994.

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.

Pratik Rathore, Zachary Frangella, Jiaming Yang, Michał Dereziński, and Madeleine Udell. Have ASkotch: A neat solution for large-scale kernel ridge regression. *Journal of Machine Learning Research*, 27(162):1–48, 2026.

Farbod Roosta-Khorasani and Michael W. Mahoney. Sub-sampled Newton methods. *Mathematical Programming*, 174:293–326, 2019.

Alessandro Rudi, Daniele Calandriello, Luigi Carratino, and Lorenzo Rosasco. On fast leverage score sampling and optimal learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Levent Sagun, Utku Evci, V. Ugur Guney, Yann Dauphin, and Léon Bottou. Empirical analysis of the Hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

Jun Shang, Haishan Ye, and Xiangyu Chang. Accelerated double-sketching subspace Newton. *European Journal of Operational Research*, 319(2):484–493, 2024. doi: 10.1016/j.ejor.2024.04.002.

Thomas Strohmer and Roman Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2009.

Stephen Tu, Shivaram Venkataraman, Ashia C. Wilson, Alex Gittens, Michael I. Jordan, and Benjamin Recht. Breaking locality accelerates block Gauss–Seidel. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

A Proofs for the Local Analysis

This section proves Theorem 3.3. Appendix A.1 establishes the quadratic acceleration, Appendix A.2 chooses an explicit neighborhood and transfers the result to nonlinear iteration, and Appendix A.3 converts the stopped estimates into conditional rates. The certified entry into this recursion is treated separately in Appendix C.

A.1 Quadratic acceleration from moment bounds

We first isolate the fixed-Hessian recursion. Its first and second moments determine the momentum parameters and the one-step contraction that will serve as the reference for the nonlinear proof.

Given $\eta \geq 0$, the exact first and second order moments are

$$\begin{aligned} \mu_\eta &= \lambda_{\min}(\overline{\mathbf{P}}_\eta), \\ \nu_\eta &= \lambda_{\max}\left(\overline{\mathbf{P}}_\eta^{-1/2}\mathbb{E}_{\mathbf{S}\sim\mathcal{D}_{\mathrm{loc}}}[\mathbf{P}_\eta(\mathbf{S})\overline{\mathbf{P}}_\eta^{-1}\mathbf{P}_\eta(\mathbf{S})]\overline{\mathbf{P}}_\eta^{-1/2}\right). \end{aligned} \tag{A.1}$$

When $\overline{\mathbf{P}}_\eta \succ 0$, they satisfy $0 < \mu_\eta \leq 1 \leq \nu_\eta \leq 1/\mu_\eta$. The inequalities in (3.2) are equivalent to $0 < \underline{\mu} \leq \mu_\eta$ and $\overline{\nu} \geq \nu_\eta$. The following proposition records the exact-update quadratic argument underlying Dereziński et al. (2025, Lemma 2.3), which already treats regularized projections and bounds in place of exact moments and builds on Gower et al. (2018). We give the conditional one-step proof in our notation for positive contractions, which uses the fact that $\mathbf{P}_\eta^2 \preceq \mathbf{P}_\eta$.

Proposition A.1 (Quadratic acceleration using certified moment bounds)**.** *For a quadratic objective with Hessian $\mathbf{H}_\star$, fixed ridge η, use (3.4) and omit only the vanishing extra term $c_2\|\nabla f\|^\gamma$ in the recursion. Writing $\mathbf{u}_k = a\mathbf{w}_k + (1-a)\mathbf{z}_k$, its unit-step updates are*

$$\begin{aligned} \mathbf{z}_{k+1} &= (\mathbf{I} - \mathbf{P}_\eta(\mathbf{S}_k))\mathbf{u}_k, \\ \mathbf{w}_{k+1} &= b\mathbf{w}_k + (1-b)\mathbf{u}_k - \theta\mathbf{P}_\eta(\mathbf{S}_k)\mathbf{u}_k. \end{aligned} \tag{A.2}$$

Every unit step satisfies Armijo for $\alpha \leq 1/2$, and

$$\mathbb{E}[\Delta_{k+1} \mid \mathcal{F}_k] \leq (1-q)\Delta_k, \qquad \mathbb{E}[f(\mathbf{x}_k) - f(\mathbf{x}_\star)] \leq \frac{\mu}{2}(1-q)^k\Delta_0. \tag{A.3}$$

For $\mathbf{v}_0 = \mathbf{x}_0$, the last bound is at most $2(1-q)^k(f(\mathbf{x}_0) - f(\mathbf{x}_\star))$.

Proof of Proposition A.1. Condition on $\mathcal{F}_k$ and suppress subscripts. Write $\mathbf{P} = \mathbf{P}_\eta(\mathbf{S}_k)$, $\mathbf{M} = \overline{\mathbf{P}}_\eta$, $\mu_b = \underline{\mu}$, $V = \overline{\nu}$, and $\mathbf{u} = a\mathbf{w} + (1-a)\mathbf{z}$. Independence gives $\mathbb{E}[\mathbf{P} \mid \mathcal{F}_k] = \mathbf{M}$ and $\mathbb{E}[\mathbf{P}\mathbf{M}^{-1}\mathbf{P} \mid \mathcal{F}_k] \preceq V\mathbf{M}$. All expectations in this one-step calculation are conditional on $\mathcal{F}_k$. Since $\mathbf{P}^2 \preceq \mathbf{P}$,

$$\mathbb{E}\|\mathbf{z}^+\|^2 \leq \|\mathbf{u}\|^2 - \|\mathbf{u}\|_{\mathbf{M}}^2. \tag{A.4}$$

By using $b+q=1$, $\mathbb{E}[\mathbf{P}\mathbf{M}^{-1}\mathbf{P}] \preceq V\mathbf{M}$ and the convex-combination identity, we have

$$\mathbb{E}\|\mathbf{w}^+\|^2_{\mathbf{M}^{-1}} \le b\|\mathbf{w}\|^2_{\mathbf{M}^{-1}} + q\|\mathbf{u}\|^2_{\mathbf{M}^{-1}} - bq\|\mathbf{w}-\mathbf{u}\|^2_{\mathbf{M}^{-1}} + \theta^2 V\|\mathbf{u}\|^2_{\mathbf{M}} - 2\theta\langle b\mathbf{w}+q\mathbf{u},\mathbf{u}\rangle. \tag{A.5}$$

Use $\mathbf{M}^{-1} \preceq \mu_b^{-1}\mathbf{I}$ and $\theta^2 V = 1/\mu_b$ to add μ_b^{-1} times (A.4), we obtain

$$\mathbb{E}\Delta^+ \le b\|\mathbf{w}\|^2_{\mathbf{M}^{-1}} + \frac{1+q}{\mu_b}\|\mathbf{u}\|^2 - 2\theta\langle b\mathbf{w}+q\mathbf{u},\mathbf{u}\rangle - bq\|\mathbf{w}-\mathbf{u}\|^2_{\mathbf{M}^{-1}}.$$

The choices $a = q/(1+q)$, $b = 1-q$, and $\theta = q/\mu_b$ imply $b\mathbf{w}+q\mathbf{u} = (\mathbf{u}-b\mathbf{z})/q$. Thus the preceding right-hand side equals

$$b\Delta - \frac{b}{\mu_b}\|\mathbf{z}-\mathbf{u}\|^2 - bq\|\mathbf{w}-\mathbf{u}\|^2_{\mathbf{M}^{-1}} \le b\Delta.$$

This proves the one-step claim without using equality of either bound with its exact moment. For a quadratic, the reduced direction satisfies $\mathbf{p}^\top\mathbf{H}_\star\mathbf{p} \le -\mathbf{g}^\top\mathbf{p}$, so $f(\mathbf{y}+\mathbf{p}) \le f(\mathbf{y}) + \mathbf{g}^\top\mathbf{p}/2$ and Armijo accepts the unit step. Since $f(\mathbf{x})-f(\mathbf{x}_\star) = \|\mathbf{z}\|^2/2 \le \mu_b\Delta/2$, iteration yields (A.3). A reset gives $\Delta_0 \le 2\|\mathbf{z}_0\|^2/\mu_b$. Finally, the general inequalities on the moments follow from $0 \preceq \mathbf{P} \preceq \mathbf{I}$, Jensen's inequality for $\mathbf{M}^{-1/2}\mathbf{P}\mathbf{M}^{-1/2}$, and that $\mathbb{E}[\mathbf{P}\mathbf{M}^{-1}\mathbf{P}] \preceq \mu_\eta^{-1}\mathbb{E}[\mathbf{P}^2] \preceq \mu_\eta^{-1}\mathbf{M}$. □

A.2 Nonlinear transfer in the local neighborhood

We first explicitly choose the neighborhood before bounding the nonlinear error. In (A.6), the first three constraints on $r_\star$ keep the trial segment inside the regularity ball, ensure positive curvature, and ensure unit-step Armijo acceptance. The final two constraints control Hessian variation and the gradient-dependent ridge.

For *every positive ridge* we may take $\ell_\eta = \eta$. For the original zero-floor variant, (3.3) instead requires $\lambda_{\min}(\mathbf{S}\mathbf{H}_\star\mathbf{S}^\top) \ge \ell_0 > 0$ almost surely. We take $L_H > 0$, and set

$$J = \sqrt{\frac{1+\theta^2}{\underline{\mu}}}, \qquad \varepsilon_\star = \min\{1/2, q/(32J)\}, \qquad \delta = h_- r_\star^2,$$
$$r_\star = \min\left\{\frac{R}{1+2U/h_-}, \frac{h_-}{2L_H}, \frac{3h_-^2(1/2-\alpha)}{2L_H U}, \frac{\varepsilon_\star h_-}{2L_H}, \frac{1}{U}\left(\frac{\varepsilon_\star\ell_\eta}{2c_2}\right)^{1/\gamma}\right\}. \tag{A.6}$$

We now use the chosen neighborhood to prove two facts: Armijo accepts a unit step for every sketch, and the nonlinear update is a small perturbation of the quadratic reference. Together they give the robust one-step contraction in Lemma 3.4. The nonlinear algorithm retains the vanishing extra ridge, and uses the exact LOCAL-mode updates of Algorithm 2: its cross-state safeguard and radius restart are disabled after the certified reset. Next we give the proof of Lemma 3.4.

Proof of Lemma 3.4. Let $r \le r_\star$, $\Delta_k \le h_- r^2$, and $\mathbf{y} = \mathbf{y}_k$. Since $\overline{\mathbf{P}}_\eta \preceq \mathbf{I}$ and $\underline{\mu} \le 1$, we have $\Delta_k \ge \|\mathbf{z}_k\|^2 + \|\mathbf{w}_k\|^2$. Hence $\mathbf{x}_k, \mathbf{v}_k, \mathbf{y}$ lie in $B(\mathbf{x}_\star, r)$. Write $\mathbf{H}_y = \nabla^2 f(\mathbf{y})$ and $\mathbf{g} = \nabla f(\mathbf{y})$. The radius constraints imply $\mathbf{H}_y \succeq (h_-/2)\mathbf{I}$, $\|\mathbf{g}\| \le Ur$, and $\Lambda_k = 0$.

Let $\mathbf{d} = -\mathbf{p}_k$ and $D = \mathbf{g}^\top\mathbf{d} \ge 0$. The reduced solve gives $\mathbf{d}^\top\mathbf{H}_y\mathbf{d} \le D$, so

$$\|\mathbf{d}\|^2 \le 2D/h_-, \qquad \|\mathbf{d}\| \le 2\|\mathbf{g}\|/h_- \le 2Ur/h_-.$$

If $D = 0$ then $\mathbf{d} = 0$ and Armijo holds. Otherwise the segment $[\mathbf{y}, \mathbf{y}-\mathbf{d}]$ stays in $B(\mathbf{x}_\star, R)$ by the first constraint on $r_\star$. Hessian Lipschitz continuity and the third radius constraint give

$$\begin{aligned} f(\mathbf{y}-\mathbf{d}) &\le f(\mathbf{y}) - D + \tfrac{1}{2}\mathbf{d}^\top\mathbf{H}_y\mathbf{d} + \tfrac{L_H}{6}\|\mathbf{d}\|^3 \\ &\le f(\mathbf{y}) - \left(\tfrac{1}{2} - \frac{2L_H U r}{3h_-^2}\right) D \le f(\mathbf{y}) - \alpha D. \end{aligned}$$

Thus the first Armijo trial is accepted uniformly over the sketch law.

For the perturbation, abbreviate $\mathbf{H} = \mathbf{H}_\star$ and define

$$\mathbf{B} = \mathbf{S}\mathbf{H}^{1/2}, \qquad \mathbf{A}_0 = \mathbf{B}\mathbf{B}^\top + \eta\mathbf{I}, \qquad \mathbf{T} = \mathbf{A}_0^{-1/2}\mathbf{B},$$
$$\mathbf{E}_y = \mathbf{H}^{-1/2}(\mathbf{H}_y - \mathbf{H})\mathbf{H}^{-1/2}, \qquad \xi_y = c_2\|\mathbf{g}\|^\gamma.$$

Then $\|\mathbf{T}\| \le 1$, $\mathbf{A}_0 \succeq \ell_\eta \mathbf{I}$, and

$$\mathbf{F} = \mathbf{A}_0^{-1/2}(\mathbf{B}\mathbf{E}_y\mathbf{B}^\top + \xi_y\mathbf{I})\mathbf{A}_0^{-1/2}, \qquad \|\mathbf{F}\| \le \varepsilon_r := \frac{L_H r}{h_-} + \frac{c_2(Ur)^\gamma}{\ell_\eta}.$$

Both summands are at most $\varepsilon_\star/2$ for $r \le r_\star$. Set $\mathbf{P} = \mathbf{T}^\top\mathbf{T}$ and $\mathbf{Q} = \mathbf{T}^\top(\mathbf{I}+\mathbf{F})^{-1}\mathbf{T}$; these are the reference and current lifted inverses in the fixed metric. For $\varepsilon_r < 1$,

$$\|\mathbf{Q} - \mathbf{P}\| \le \frac{\varepsilon_r}{1-\varepsilon_r}, \qquad \|\mathbf{Q}\| \le \frac{1}{1-\varepsilon_r}.$$

The integral Taylor identity along $[\mathbf{x}_\star, \mathbf{y}]$ gives, with $\mathbf{u} = \mathbf{H}^{1/2}(\mathbf{y} - \mathbf{x}_\star)$,

$$\mathbf{H}^{-1/2}\mathbf{g} = \mathbf{u} + \mathbf{e}_g, \qquad \|\mathbf{e}_g\| \le \frac{L_H r}{h_-}\|\mathbf{u}\|.$$

Consequently

$$\|\mathbf{H}^{1/2}\mathbf{d} - \mathbf{P}\mathbf{u}\| = \|\mathbf{Q}(\mathbf{u} + \mathbf{e}_g) - \mathbf{P}\mathbf{u}\| \le \frac{2\varepsilon_r}{1-\varepsilon_r}\|\mathbf{u}\|. \tag{A.7}$$

The error in the two updated states is $(-\mathbf{e}_d, -\theta\mathbf{e}_d)$, where $\mathbf{e}_d = \mathbf{H}^{1/2}\mathbf{d} - \mathbf{P}\mathbf{u}$. Since $\overline{\mathbf{P}}_\eta^{-1} \preceq \underline{\mu}^{-1}\mathbf{I}$ and $\|\mathbf{u}\|^2 \le \Delta_k$,

$$\|\mathbf{e}_k\|_{\mathbf{G}_\Delta} \le J\|\mathbf{e}_d\| \le \frac{2J\varepsilon_r}{1-\varepsilon_r}\sqrt{\Delta_k} \le 4J\varepsilon_\star\sqrt{\Delta_k} \le (q/8)\sqrt{\Delta_k}.$$

Minkowski's inequality in conditional L^2 and Proposition A.1 now prove (3.11), where we also use the inequality $(\sqrt{1-q} + q/8)^2 \le 1 - q + q/4 + q^2/64 \le 1 - q/2$. □

A.3 Proof of Theorem 3.3

The one-step bound holds only inside the chosen neighborhood. We bound the probability of exit and the energy before exit, then obtain the theorem's conditional expectations by conditioning on no exit.

Proof of Theorem 3.3. Let δ be as in (A.6), and set $\rho = 1 - \frac{1}{2}\sqrt{\underline{\mu}/\overline{\nu}}$. First fix an arbitrary $\delta_{\text{exit}} \in (0,1)$ and a fixed initial state satisfying $\Delta_0 \le \delta_{\text{exit}}\delta$. For the proof only, define

$$T = \inf\{k \ge 0 : \Delta_k > \delta\}, \qquad \mathcal{E}_\delta = \{T = \infty\},$$

with $\inf \varnothing = \infty$, and freeze the process after its first exit. This leaves all trajectories on $\mathcal{E}_\delta$ unchanged. The event $\{T > k\}$ is $\mathcal{F}_k$-measurable. On this event, Lemma 3.4 applies even if the next state exits. Nonnegativity and conditioning give

$$\mathbb{E}[\Delta_{k+1}\mathbb{1}_{\{T>k+1\}}] \le \mathbb{E}[\mathbb{1}_{\{T>k\}}\mathbb{E}[\Delta_{k+1} \mid \mathcal{F}_k]] \le \rho\mathbb{E}[\Delta_k\mathbb{1}_{\{T>k\}}].$$

Since $\Delta_0 < \delta$, induction gives

$$\mathbb{E}[\Delta_k\mathbb{1}_{\{T>k\}}] \le \rho^k\Delta_0. \tag{A.8}$$

The stopped process $Z_k = \Delta_{k\wedge T}$ is a nonnegative supermartingale: it is constant after exit and satisfies the one-step inequality before exit. Integrability follows inductively from that inequality, so $\mathbb{E}Z_k \le \Delta_0$. Since $Z_k > \delta$ on $\{T \le k\}$, $\mathbb{P}(T \le k) \le \Delta_0/\delta$. Letting $k \to \infty$ yields

$$\mathbb{P}(T < \infty) \le \Delta_0/\delta \le \delta_{\text{exit}}. \tag{A.9}$$

Conditional expectations. Since $\mathcal{E}_\delta \subseteq \{T > k\}$ and $\mathbb{P}(\mathcal{E}_\delta) \geq 1 - \delta_{\text{exit}} > 0$, the survival-weighted estimate above implies

$$\mathbb{E}[\Delta_k \mid \mathcal{E}_\delta] = \frac{\mathbb{E}[\Delta_k \mathbb{1}_{\mathcal{E}_\delta}]}{\mathbb{P}(\mathcal{E}_\delta)} \leq \frac{\mathbb{E}[\Delta_k \mathbb{1}_{\{T>k\}}]}{\mathbb{P}(\mathcal{E}_\delta)} \leq \frac{\rho^k \Delta_0}{1 - \delta_{\text{exit}}}.$$

Choose $C_f := \frac{U\mu}{h_-}$. On $\{T > k\}$, the local curvature bounds and the definition of Δ_k give

$$0 \leq f(\mathbf{x}_k) - f(\mathbf{x}_\star) \leq \frac{U}{2}\|\mathbf{x}_k - \mathbf{x}_\star\|^2 \leq \frac{U\mu}{2h_-}\Delta_k = \frac{C_f}{2}\Delta_k.$$

Consequently,

$$\mathbb{E}[(f(\mathbf{x}_k) - f(\mathbf{x}_\star))\mathbb{1}_{\{T>k\}}] \leq \frac{C_f}{2}\rho^k \Delta_0,$$
$$\mathbb{E}[f(\mathbf{x}_k) - f(\mathbf{x}_\star) \mid \mathcal{E}_\delta] \leq \frac{C_f}{2(1 - \delta_{\text{exit}})}\rho^k \Delta_0.$$

Constant success probability. For the initialization in the theorem, take $\delta_{\text{exit}} = 1/100$ and $\mathcal{G} = \mathcal{E}_\delta$. The exit estimate gives $\mathbb{P}(\mathcal{G}) \geq 0.99$, which proves the probability statement in Theorem 3.3. Since

$$\frac{1}{1 - \delta_{\text{exit}}} = \frac{100}{99} \leq 2, \qquad \frac{1}{2(1 - \delta_{\text{exit}})} = \frac{50}{99} \leq 1,$$

the preceding conditional estimates prove (3.8) and (3.9).

Almost-sure convergence. For every $a > 0$, the survival-weighted estimate gives

$$\sum_{k\geq 0} \mathbb{P}(\Delta_k > a, \mathcal{E}_\delta) \leq \frac{1}{a}\sum_{k\geq 0} \mathbb{E}[\Delta_k \mathbb{1}_{\{T>k\}}] \leq \frac{\Delta_0}{a(1-\rho)} < \infty.$$

Borel-Cantelli, applied to a countable sequence $a \downarrow 0$, yields $\Delta_k \to 0$ almost surely on $\mathcal{E}_\delta$. Since Δ_k controls both squared state errors, both states converge to $\mathbf{x}_\star$ almost surely conditional on $\mathcal{G}$. The general stopping and conditional-expectation estimates, together with the same almost-sure convergence argument, apply for every $\delta_{\text{exit}} \in (0, 1)$ with $\Delta_0 \leq \delta_{\text{exit}}\delta$. □

B SKETCH-MOMENT BOUNDS AND SPECTRAL COROLLARIES

We prove the local analysis: Gaussian moments and their spectral-tail rate in Appendix B.1, and the ARLS construction, its two moment bounds, and its shifted-Hessian rate in Appendix B.2.

Common notation. For a fixed $\mathbf{H} \succ 0$, order its eigenvalues as $\lambda_1(\mathbf{H}) \geq \cdots \geq \lambda_n(\mathbf{H}) > 0$ and set $h = \lambda_n(\mathbf{H})$. For an integer $0 \leq r < n$, define sum of the tail eigenvalues as $T_r(\mathbf{H}) := \sum_{i=r+1}^n \lambda_i(\mathbf{H})$, and have $T_0(\mathbf{H}) = \operatorname{tr}\mathbf{H}$. When the matrix is clear we write T_r. Write μ, ν for the exact moments in (A.1), with $\mathbf{H}$ replacing $\mathbf{H}_\star$. The common first-moment is:

$$\overline{\mathbf{P}}_\eta \succeq c_\mu \mathbf{H}(\mathbf{H} + \tau\mathbf{I})^{-1}, \qquad \mu \geq \frac{c_\mu h}{h + \tau}. \tag{B.1}$$

Here $c_\mu > 0$ and $\tau > 0$ are comparison parameters. The two sketch families below use different laws and different second-moment arguments.

B.1 GAUSSIAN SKETCHES: FIRST AND SECOND MOMENTS

Throughout this track, $\mathbf{S}_{ij} \sim \mathcal{N}(0, 1/s)$ independently and $1 \leq s \leq n$. We first establish the mean operator bound, then control the second moment without excluding rare draws, and finally combine them to prove Theorem 3.5.

B.1.1 First-moment lower bound

Lemma B.1 (Gaussian first moment). *For $\mathbf{H} \succ 0$, $\eta > 0$, and $0 \le r \le s/2$, define $\tau = \eta + T_r(\mathbf{H})/s$. Then*

$$\mathbb{E}\mathbf{P}_\eta \succeq c_{\mathrm{G}}\mathbf{H}(\mathbf{H} + \tau\mathbf{I})^{-1}. \tag{B.2}$$

Proof of Lemma B.1. By rotational invariance we may assume $\mathbf{H} = \mathrm{diag}(h_1, \dots, h_n)$ with $h_1 \ge \cdots \ge h_n > 0$. Let $\mathbf{s}_i$ be column i of $\mathbf{S}$, and set

$$\mathbf{A}_{-i} = \eta\mathbf{I} + \sum_{j \ne i} h_j \mathbf{s}_j \mathbf{s}_j^\top, \qquad t_i = \mathbf{s}_i^\top \mathbf{A}_{-i}^{-1} \mathbf{s}_i.$$

Sherman-Morrison gives

$$(\mathbf{P}_\eta)_{ii} = \frac{h_i t_i}{1 + h_i t_i}. \tag{B.3}$$

Let $\mathbf{Q}_i$ be the Euclidean orthogonal projector onto the orthogonal complement of $\mathrm{span}\{\mathbf{s}_j : j \le r, j \ne i\}$. Conditionally on these head columns, $\mathbf{Q}_i$ is independent of $\mathbf{s}_i$ and all remaining tail columns. Its rank q is at least $s - r \ge s/2$. Set $\mathbf{u}_i = \mathbf{Q}_i\mathbf{s}_i$ and $X_i = \|\mathbf{u}_i\|^2$. Conditionally on the head columns, sX_i follows the χ^2_q law, so

$$\mathbb{E}X_i = \frac{q}{s}, \qquad \mathbb{E}X_i^2 = \frac{q(q+2)}{s^2}, \qquad \mathbb{P}\left(X_i \ge \frac{q}{2s}\right) \ge \frac{q}{4(q+2)} \ge \frac{1}{12}.$$

Given the head columns and $\mathbf{s}_i$, independence of the remaining columns implies

$$\mathbb{E}[\mathbf{u}_i^\top \mathbf{A}_{-i}\mathbf{u}_i \mid \mathbf{Q}_i, \mathbf{s}_i] = \left(\eta + \frac{1}{s}\sum_{j > r, j \ne i} h_j\right) X_i \le \tau X_i.$$

Consequently $\mathbf{u}_i^\top \mathbf{A}_{-i}\mathbf{u}_i \le 2\tau X_i$ with conditional probability at least $1/2$. On the intersection with $X_i \ge q/(2s)$, Cauchy-Schwarz in the $\mathbf{A}_{-i}$ metric yields

$$t_i \ge \frac{(\mathbf{s}_i^\top \mathbf{u}_i)^2}{\mathbf{u}_i^\top \mathbf{A}_{-i}\mathbf{u}_i} \ge \frac{X_i}{2\tau} \ge \frac{1}{8\tau}.$$

This intersection has probability at least $1/24$. Hence

$$\mathbb{E}(\mathbf{P}_\eta)_{ii} \ge \frac{1}{24}\frac{h_i}{h_i + 8\tau} \ge c_{\mathrm{G}}\frac{h_i}{h_i + \tau}$$

for a sufficiently small absolute constant $c_{\mathrm{G}} > 0$. Column sign symmetry makes $\mathbb{E}\mathbf{P}_\eta$ diagonal in this eigenbasis: flipping the sign of column i reverses every off-diagonal entry in row i and leaves the distribution unchanged. The diagonal inequalities therefore establish the Loewner bound. □

B.1.2 Second moment upper bound

Lemma B.2 (Gaussian second moment). *Suppose (B.1) holds for a Gaussian sketch with $1 \le s \le n$, $\eta > 0$, and $\tau > 0$. Put $R_\eta = \tau/\eta$. Then*

$$\nu \le \frac{C}{c_\mu^2}(1 + R_\eta)^2 \frac{n}{s}, \tag{B.4}$$

where $C > 0$ is an absolute constant. In particular, if $c_\mu = c_{\mathrm{G}}$ and $\tau \le 3\eta$, then $\nu \le C_{\mathrm{G}} n/s$ for the fixed absolute constant C_{G} specified below. This includes both the choice $\tau = 2\eta$ below and the frozen-anchor choice $\tau = 3\eta$.

Lemma B.3 (Two Gaussian norm moments). *For the normalized Gaussian sketch with $1 \le s \le n$, write*

$$a_S = \mathbb{E}\|\mathbf{S}\|^2, \qquad j_S = \left\|\mathbb{E}[\|\mathbf{S}\|^2 \mathbf{S}^\top \mathbf{S}]\right\|.$$

Then

$$a_S \le 12n/s, \qquad \mathbb{E}[\|\mathbf{S}\|^2 \mathbf{S}^\top \mathbf{S}] = (1 + 2/(sn))a_S\mathbf{I}, \qquad j_S \le 36n/s. \tag{B.5}$$

Proof. The Gaussian spectral-norm inequality (Davidson & Szarek, 2001) gives

$$\mathbb{P}\left(\|\mathbf{S}\| > 1 + \sqrt{n/s} + t/\sqrt{s}\right) \le e^{-t^2/2}.$$

Put $a = 1 + \sqrt{n/s}$ and $X = (\|\mathbf{S}\| - a)_+$. Integration gives $\mathbb{E}X^2 \le 2/s$, whence $\mathbb{E}\|\mathbf{S}\|^2 \le 2a^2 + 2\mathbb{E}X^2 \le 12n/s$. For the matrix identity, write $\mathbf{S} = R\mathbf{U}/\sqrt{s}$, where R^2 has the χ^2_{sn} law and $\mathbf{U}$ is independent and uniform on the unit Frobenius sphere. Right orthogonal invariance makes $\mathbb{E}[\|\mathbf{U}\|^2\mathbf{U}^\top\mathbf{U}]$ a scalar multiple of $\mathbf{I}$; its trace equals $\mathbb{E}\|\mathbf{U}\|^2$. Thus

$$\mathbb{E}[\|\mathbf{S}\|^2\mathbf{S}^\top\mathbf{S}] = \frac{\mathbb{E}R^4}{s^2 n}\mathbb{E}\|\mathbf{U}\|^2\mathbf{I} = \left(1 + \frac{2}{sn}\right)\mathbb{E}\|\mathbf{S}\|^2\mathbf{I}.$$

□

Proof of Lemma B.2. We first prove a bound for any isotropic sketch with $\mathbb{E}[\mathbf{S}^\top\mathbf{S}] = \mathbf{I}$ and finite a_S, j_S. Write $c = c_\mu$, $R_\eta = \tau/\eta$, and define the deterministic matrices

$$\mathbf{C} = (\mathbf{H} + \tau\mathbf{I})^{-1/2}\mathbf{H}^{1/2}\overline{\mathbf{P}}_\eta^{-1/2}, \qquad \mathbf{D}_1 = \mathbf{H}^{1/2}(\mathbf{H} + \tau\mathbf{I})^{-1/2}, \quad \mathbf{D}_2 = \sqrt{\tau}(\mathbf{H} + \tau\mathbf{I})^{-1/2}.$$

The first-moment bound implies $\mathbf{C}^\top\mathbf{C} \preceq c^{-1}\mathbf{I}$. Also $\|\mathbf{D}_1\|, \|\mathbf{D}_2\| \le 1$ and

$$(\mathbf{H} + \tau\mathbf{I})^{1/2} = \mathbf{H}^{1/2}\mathbf{D}_1 + \sqrt{\tau}\mathbf{D}_2.$$

For $\mathbf{A}_S = \mathbf{S}\mathbf{H}\mathbf{S}^\top + \eta\mathbf{I}$, let

$$\mathbf{B} = \mathbf{A}_S^{-1/2}\mathbf{S}\mathbf{H}^{1/2}\overline{\mathbf{P}}_\eta^{-1/2} = \mathbf{U}_S + \mathbf{V}_S,$$

where

$$\mathbf{U}_S = \mathbf{A}_S^{-1/2}\mathbf{S}\mathbf{H}^{1/2}\mathbf{D}_1\mathbf{C}, \qquad \mathbf{V}_S = \sqrt{\tau}\mathbf{A}_S^{-1/2}\mathbf{S}\mathbf{D}_2\mathbf{C}.$$

Then $\mathbf{W} := \overline{\mathbf{P}}_\eta^{-1/2}\mathbf{P}_\eta\overline{\mathbf{P}}_\eta^{-1/2} = \mathbf{B}^\top\mathbf{B}$. Since $\|\mathbf{A}_S^{-1/2}\mathbf{S}\mathbf{H}^{1/2}\| \le 1$ and $\mathbf{A}_S \succeq \eta\mathbf{I}$,

$$\|\mathbf{U}_S\|^2 \le 1/c, \qquad \|\mathbf{V}_S\|^2 \le (R_\eta/c)\|\mathbf{S}\|^2,$$
$$\|\mathbf{V}_S\mathbf{x}\|^2 \le R_\eta\|\mathbf{S}\mathbf{D}_2\mathbf{C}\mathbf{x}\|^2, \qquad \mathbb{E}\|\mathbf{V}_S\mathbf{x}\|^2 \le (R_\eta/c)\|\mathbf{x}\|^2.$$

Moreover,

$$\mathbb{E}[\|\mathbf{V}_S\|^2\|\mathbf{V}_S\mathbf{x}\|^2] \le \frac{R_\eta^2}{c}\mathbf{x}^\top\mathbf{C}^\top\mathbf{D}_2\mathbb{E}[\|\mathbf{S}\|^2\mathbf{S}^\top\mathbf{S}]\mathbf{D}_2\mathbf{C}\mathbf{x} \le \frac{R_\eta^2 j_S}{c^2}\|\mathbf{x}\|^2.$$

For every fixed vector $\mathbf{x}$,

$$\begin{aligned}\|\mathbf{W}\mathbf{x}\|^2 &\le \|\mathbf{U}_S + \mathbf{V}_S\|^2\|(\mathbf{U}_S + \mathbf{V}_S)\mathbf{x}\|^2 \\ &\le 4(c^{-1} + \|\mathbf{V}_S\|^2)(c^{-1}\|\mathbf{x}\|^2 + \|\mathbf{V}_S\mathbf{x}\|^2).\end{aligned}$$

Taking expectations, we obtain

$$\nu \le \frac{4}{c^2}\left[1 + R_\eta(1 + a_S) + R_\eta^2 j_S\right]. \tag{B.6}$$

Substituting (B.5) and using $n/s \ge 1$ gives

$$\nu \le \frac{C}{c^2}(1 + R_\eta)^2\frac{n}{s}$$

for an absolute constant $C > 0$, proving (B.4) without excluding any Gaussian sketches. Since c_{G} is a fixed absolute constant, choosing $C_{\mathrm{G}} \ge 16C/c_{\mathrm{G}}^2$ ensures $\nu \le C_{\mathrm{G}}n/s$ whenever $c = c_{\mathrm{G}}$ and $R_\eta \le 3$. This is one common choice of C_{G} for all Gaussian statements. □

B.1.3 Gaussian convergence rate

Proof of Theorem 3.5. Apply the two moment bounds at $\mathbf{H} = \mathbf{H}_\star$, with $\eta_{\mathrm{G}} = T_r(\mathbf{H})/s$ and $\tau = 2\eta_{\mathrm{G}}$. Here $1 \le r \le s/2$ and $s \le n$, as required by the Gaussian second-moment bound. Since $r \le s/2 \le n/2$, $T_r \ge (n - r)h \ge nh/2$, so $h + 2T_r/s \le 4T_r/s$. Consequently

$$\mu \ge \frac{c_{\mathrm{G}}h}{h + 2T_r/s} \ge \frac{c_{\mathrm{G}}s}{4(n - r)\bar{\kappa}_r} \ge \underline{\mu}_{\mathrm{G}}.$$

The second-moment claim follows from Lemma B.2. □

B.2 Approximate ridge leverage score sampling

There are three distinct ingredients in this track. The rounded ARLS sampling law and its mean projection bound are imported from Rathore et al. (2026, Definition 5 and Lemma 10). The regularized coordinate second-moment estimate specializes the variance argument of Dereziński et al. (2025, Theorem 3.8). We then apply these ingredients to a shifted Hessian, transfer the mean bound to the Newton operator, and choose the shift using the spectral tail. Neither the imported sampling law nor the underlying variance argument is claimed as a new ingredient here.

B.2.1 Imported sampling law and first-moment lower bound

Definition B.4 (Scores, normalization, and coordinate sketches)**.** *For $\mathbf{A} \succ 0$ and $\zeta > 0$, define*

$$\ell_i^\zeta(\mathbf{A}) = [\mathbf{A}(\mathbf{A}+\zeta\mathbf{I})^{-1}]_{ii}, \qquad d^\zeta(\mathbf{A}) = \sum_i \ell_i^\zeta(\mathbf{A}).$$

A c_{A}-approximation consists of upper scores $w_i \geq \ell_i^\zeta(\mathbf{A})$ with $L = \sum_i w_i \leq c_{\mathrm{A}} d^\zeta(\mathbf{A})$. For the rounded ARLS law, set

$$\widehat{w}_i = \frac{L}{n}\left\lceil \frac{n w_i}{L} \right\rceil, \qquad q_i = \frac{\widehat{w}_i}{\sum_j \widehat{w}_j}. \tag{B.7}$$

Write $\mathbf{q} = (q_1, \ldots, q_n)$ for the probability vector. Draw s independent indices from $\mathbf{q}$ and let B be their union. The sketch is the unscaled, distinct-row *selector $\mathbf{S} = \mathbf{I}_B$; thus $\mathbf{I}_B\mathbf{I}_B^\top = \mathbf{I}_{|B|}$ and $|B| \leq s$. The rounding preserves upper scores and increases their total by at most L, which only changes the approximation factor by two.*

Lemma B.5 (Imported ARLS projection bound)**.** *Let $\mathbf{A} \succ 0$, $r \geq C_0 \log n$, $4r < n$, and choose positive ridges τ_{A}, ζ by*

$$d^{\tau_{\mathrm{A}}}(\mathbf{A}) = 2r, \qquad d^\zeta(\mathbf{A}) = 4r. \tag{B.8}$$

Suppose $1 + \tau_{\mathrm{A}}/\lambda_{\min}(\mathbf{A}) \leq n^{C_*}$ for a fixed constant C_*. For the ARLS construction above with $s \geq C c_{\mathrm{A}} r \log^3 n$,

$$\mathbb{E}\big[\mathbf{A}^{1/2}\mathbf{I}_B^\top(\mathbf{A}_{BB})^{-1}\mathbf{I}_B\mathbf{A}^{1/2}\big] \succeq \tfrac{1}{2}\mathbf{A}(\mathbf{A}+\tau_{\mathrm{A}}\mathbf{I})^{-1}. \tag{B.9}$$

This is the fixed-effective-dimension specialization of Lemma 10 in Rathore et al. (2026), including its ARLS rounding construction. The polynomial-ratio hypothesis makes explicit the inverse-polynomial coupling accuracy used in that proof. We use $d^\zeta = 4r$, not an arbitrarily large effective dimension at the sampling ridge. Both restrictions hold for the shifted construction below. This is the imported first-moment result used below.

B.2.2 Second moment upper bound for distinct coordinate updates

The following bound specializes the regularized-projection variance argument of Dereziński et al. (2025, Theorem 3.8) to distinct coordinate updates. We include the derivation and its constant.

Lemma B.6 (Coordinate second moment)**.** *Let $\mathbf{S} = \mathbf{I}_B$ select distinct coordinates, $\eta > 0$, and suppose* (B.1) *holds. Then*

$$\nu \leq \frac{1}{c_\mu}\left(1 + \frac{\tau}{\eta}\right). \tag{B.10}$$

Proof of Lemma B.6. Inverting the first-moment bound gives $\overline{\mathbf{P}}_\eta^{-1} \preceq c_\mu^{-1}(\mathbf{I} + \tau\mathbf{H}^{-1})$. Write $\mathbf{A}_B = \mathbf{H}_{BB} + \eta\mathbf{I}$. Since $\mathbf{I}_B\mathbf{I}_B^\top = \mathbf{I}$,

$$\mathbf{P}_\eta\mathbf{H}^{-1}\mathbf{P}_\eta = \mathbf{H}^{1/2}\mathbf{I}_B^\top\mathbf{A}_B^{-2}\mathbf{I}_B\mathbf{H}^{1/2} \preceq \eta^{-1}\mathbf{P}_\eta.$$

Together with $\mathbf{P}_\eta^2 \preceq \mathbf{P}_\eta$, this implies $\mathbb{E}[\mathbf{P}_\eta\overline{\mathbf{P}}_\eta^{-1}\mathbf{P}_\eta] \preceq c_\mu^{-1}(1+\tau/\eta)\overline{\mathbf{P}}_\eta$. □

B.2.3 SHIFTED-HESSIAN RATE

The modification for our Newton update is to construct ARLS on $\mathbf{A} = \mathbf{H} + \eta_{\mathrm{R}}\mathbf{I}$, where $\eta_{\mathrm{R}} = T_r(\mathbf{H})/(n-r)$, and to use the same η_{R} once in the reduced solve for $\mathbf{H}$. We transfer the imported unregularized projection bound for $\mathbf{A}$ to the regularized operator for $\mathbf{H}$, then apply the coordinate second-moment bound.

Proof of Theorem 3.6. For $\mathbf{A} = \mathbf{H} + \eta_{\mathrm{R}}\mathbf{I}$ and $\eta = \eta_{\mathrm{R}}$, write $\mathbf{\Pi}_B(\mathbf{A}) = \mathbf{A}^{1/2}\mathbf{I}_B^\top \mathbf{A}_{BB}^{-1}\mathbf{I}_B\mathbf{A}^{1/2}$. Since $\mathbf{A}_{BB} = \mathbf{H}_{BB} + \eta_{\mathrm{R}}\mathbf{I}$, we have

$$\mathbf{P}_\eta(\mathbf{H}) = \mathbf{H}^{1/2}\mathbf{A}^{-1/2}\mathbf{\Pi}_B(\mathbf{A})\mathbf{A}^{-1/2}\mathbf{H}^{1/2}. \tag{B.11}$$

With $\tau = \eta_{\mathrm{R}} + \tau_{\mathrm{A}}$, this identity transfers (B.9) to the first-moment inequality in (B.13) below.

For any $t > 0$,

$$d^t(\mathbf{A}) \le r + \frac{\sum_{i>r}\lambda_i(\mathbf{A})}{t}.$$

Since $d^{\tau_{\mathrm{A}}}(\mathbf{A}) = 2r$ and $\sum_{i>r}\lambda_i(\mathbf{A}) = T_r(\mathbf{H}) + (n-r)\eta_{\mathrm{R}} = 2T_r(\mathbf{H})$,

$$\tau_{\mathrm{A}} \le \frac{2T_r(\mathbf{H})}{r}, \qquad \frac{\tau}{\eta_{\mathrm{R}}} \le \frac{2n}{r} - 1. \tag{B.12}$$

In particular, $1 + \tau_{\mathrm{A}}/\lambda_{\min}(\mathbf{A}) \le 1 + 2(n-r)/r$, so the polynomial-ratio requirement in Lemma B.5 is automatic. The first moment and $\eta_{\mathrm{R}} \ge h$ give

$$\mu \ge \frac{h}{2(h+\tau)} \ge \frac{rh}{4n\eta_{\mathrm{R}}} = \frac{r}{4n\bar{\kappa}_r}.$$

The second-moment lemma gives $\nu \le 2(1 + \tau/\eta_{\mathrm{R}}) \le 4n/r$. Together, these estimates yield

$$\overline{\mathbf{P}}_\eta \succeq \tfrac{1}{2}\mathbf{H}(\mathbf{H}+\tau\mathbf{I})^{-1}, \qquad \mu \ge \frac{r}{4n\bar{\kappa}_r}, \qquad \nu \le \frac{4n}{r}. \tag{B.13}$$

Substitute these bounds in Proposition A.1 and Theorem 3.3.

□

What is sampled, and what is optimized? The shift is only a device for constructing the distribution and damping the reduced solve. The gradient remains $\nabla f(\mathbf{y})$: no quadratic penalty is added to f, and the limiting stationary point is not shifted. The needed scores can also be expressed using the unshifted Hessian:

$$\ell_i^\zeta(\mathbf{H} + \eta\mathbf{I}) = \frac{\eta}{\eta+\zeta} + \frac{\zeta}{\eta+\zeta}\ell_i^{\eta+\zeta}(\mathbf{H}). \tag{B.14}$$

Thus exact shifted RLS is a particular *uniform–RLS mixture*, not a requirement to compute a second unrelated set of scores. Replacing the last score by a certified upper approximation preserves an upper approximation with no larger factor. Rounding and normalization then use (B.7).

Cost and interpretation. At fixed cutoff r, Gaussian sketches use $s = \Theta(r)$ rows, whereas this ARLS guarantee uses $s = \widetilde{O}(c_{\mathrm{A}} r)$ draws. At a fixed draw budget s, the RLS cutoff is only $r = \Theta(s/(c_{\mathrm{A}}\log^3 n))$ in this certificate; it must not be replaced by s inside the spectral tail. Both methods solve a reduced system, but a coordinate sketch accesses Hessian entries/principal blocks rather than dense Hessian-matrix products. Which is cheaper depends on the objective and its derivative access model. In the entry-wise access model, constant-factor RLS approximations at effective dimension $4r$ can be obtained with high probability in $\widetilde{O}(nr^2)$ arithmetic work using the BLESS algorithm of Rudi et al. (2018); evaluating application-specific Hessian entries must also be accounted for. The shift only changes diagonal entries of this preprocessing matrix. This score-computation cost is not free for a general nonlinear Hessian. For clarity, entry queries and arithmetic work are different costs. At $d^\zeta(\mathbf{A}) = 4r$, the BLESS dictionary guarantee and score formula (Rudi et al., 2018, Theorem 1 and Eq. (3)) use a dictionary J of size $\widetilde{O}(r)$. Evaluating the approximate scores for all coordinates queries the diagonal and the blocks $\mathbf{A}_{J,[n]}$ and $\mathbf{A}_{JJ}$, hence $O(n|J| + |J|^2 + n) = \widetilde{O}(nr)$ entries; the intermediate dictionaries obey the same bound up to

logarithmic factors. Their score evaluations and reduced factorizations cost $\widetilde{O}(nr^2)$ arithmetic. A fixed multiplicative rescaling turns the two-sided score estimates into certified upper scores. These are preprocessing bounds under the cited entry-access model, with confidence and ridge-scale logarithms suppressed; they exclude the cost of choosing or certifying the spectral parameters. Randomized score failures must be included in δ_{cert}. The regime $s \ll n$ is where the sampling construction is computationally useful.

C Proofs for the Certified Switch

C.1 From global to local: certified entry

The spectral rates above are stated at a fixed reference Hessian. The following construction closes the implementation interface: the law, ridge, moment bounds, momentum, and entry radius can all be fixed using data at the candidate switch point, without evaluating $\mathbf{H}_\star$.

Lemma C.1 (Certificate for a locally unique minimizer)**.** *Suppose $L_H > 0$, $\nabla^2 f$ is L_H-Lipschitz on $B(\mathbf{x}_c, m/(2L_H))$, and*

$$\mathbf{H}_c = \nabla^2 f(\mathbf{x}_c) \succeq m\mathbf{I} \succ 0, \qquad \|\nabla f(\mathbf{x}_c)\| < \frac{m^2}{8L_H}. \tag{C.1}$$

Then f is $m/2$-strongly convex on that ball and has a unique stationary point $\mathbf{x}_\star$ in its interior. Moreover

$$\|\mathbf{x}_c - \mathbf{x}_\star\| \le \frac{2\|\nabla f(\mathbf{x}_c)\|}{m} < \frac{m}{4L_H}, \qquad \tfrac{1}{2}\mathbf{H}_\star \preceq \mathbf{H}_c \preceq \tfrac{3}{2}\mathbf{H}_\star. \tag{C.2}$$

If $\|\nabla^2 f\| \le U_{\mathrm{loc}}$ on the candidate ball, the hypotheses on the objective in Assumption 3.1 hold on $B(\mathbf{x}_\star, R)$ with

$$h_- = m/2, \qquad U = U_{\mathrm{loc}}, \qquad R = m/(4L_H). \tag{C.3}$$

A positive *algorithmic* ridge cannot replace the full-space curvature condition in this lemma. It makes a reduced solve invertible but does not make the objective locally strongly convex or distinguish a saddle from a minimizer. Thus the m^2/L_H scale is not removed by adding η.

Proposition C.2 (Frozen-anchor Gaussian and ARLS parameters)**.** *Suppose that (C.1) holds and set $h_- = m/2$. Then the following frozen-anchor constructions provide a local sketch law $\mathcal{D}_{\mathrm{loc}}$ and certified bounds $(\underline{\mu}, \overline{\nu}, \ell_\eta)$ satisfying Assumption 3.2.*

Gaussian. *Choose $1 \le r \le s/2$, $s \le n$, take $\eta = T_r(\mathbf{H}_c)/s$, and freeze the normalized Gaussian law. Then*

$$\underline{\mu} = \frac{c_{\mathrm{G}} h_-}{h_- + 3\eta}, \qquad \overline{\nu} = C_{\mathrm{G}} \frac{n}{s}, \qquad \ell_\eta = \eta \tag{C.4}$$

are valid bounds in Assumption 3.2.

Shifted-Hessian ARLS. *Choose $C_0 \log n \le r < n/4$, set $\eta = T_r(\mathbf{H}_c)/(n-r)$ and $\mathbf{A}_c = \mathbf{H}_c + \eta\mathbf{I}$. Compute upper scores and ridges $d^{\tau_{\mathrm{A},c}}(\mathbf{A}_c) = 2r$, $d^{\zeta}(\mathbf{A}_c) = 4r$ as in Lemma B.5. Freeze the rounded, normalized law with $s \ge C c_{\mathrm{A}} r \log^3 n$ draws, deduplicated to $\mathbf{S} = \mathbf{I}_B$. Set $\tau_c = \eta + \tau_{\mathrm{A},c}$. Then*

$$\underline{\mu} = \frac{h_-}{6(h_- + \tau_c)}, \overline{\nu} = 6(1 + \tau_c/\eta), \qquad \ell_\eta = \eta \tag{C.5}$$

are valid bounds in Assumption 3.2.

Definition C.3 (Certified local-entry test)**.** *At a stored point $\mathbf{x}_c$, the test proceeds as follows.*

1. **Certify local curvature.** *Let $\mathbf{H}_c = \nabla^2 f(\mathbf{x}_c)$. Obtain $m > 0$ such that $\mathbf{H}_c \succeq m\mathbf{I}$, together with constants $U_{\mathrm{loc}}, L_H > 0$ bounding the Hessian norm and its Lipschitz constant on $B(\mathbf{x}_c, m/(2L_H))$. If no such certified bounds are available, return* ***false****.*

2. **Construct a candidate local configuration.** *Using* $\mathbf{H}_c, m, U_{\mathrm{loc}}, L_H$*, form a candidate configuration* $(\eta, \mathcal{D}_{\mathrm{loc}}, \underline{\mu}, \overline{\nu}, \ell_\eta)$ *according to Proposition C.2. Set* $h_- = m/2$*,* $U = U_{\mathrm{loc}}$*,* $R = m/(4L_H)$*, compute* (a, b, θ) *from (3.4), and* $r_\star$ *from (A.6). Define* $\delta_{\mathrm{loc}} = (m/2)r_\star^2$*.*

3. **Test local entry.** *For a budget* $\delta_{\mathrm{exit}} \in (0, 1)$*, return* ***true*** *if*

$$\|\nabla f(\mathbf{x}_c)\| \le \frac{m r_\star}{4}\sqrt{\delta_{\mathrm{exit}}\underline{\mu}}, \tag{ENTRY}$$

and return ***false*** *otherwise. On success, return the certified local configuration* $(\eta, \mathcal{D}_{\mathrm{loc}}, \underline{\mu}, \overline{\nu}, \delta_{\mathrm{loc}})$ *and the momentum parameters* (a, b, θ)*. All certification and preprocessing randomness is generated before the next trial sketch.*

Since $\delta_{\mathrm{loc}} = (m/2)r_\star^2$, condition (ENTRY) is equivalent to

$$\frac{8\|\nabla f(x_c)\|^2}{\underline{\mu} m} \le \delta_{\mathrm{exit}}\,\delta_{\mathrm{loc}}.$$

This is the energy form used in Lemma C.4. Moreover, no separate localization check is needed. Indeed, for either construction in Proposition C.2, the candidate lower moment bound satisfies $\underline{\mu} \le 1$ directly from (C.4) – (C.5). Hence, since $r_\star \le m/(4L_H)$ and $\delta_{\mathrm{exit}} < 1$, (ENTRY) implies

$$\|\nabla f(\mathbf{x}_c)\| \le \frac{m^2}{16L_H} < \frac{m^2}{8L_H}. \tag{LOC}$$

Hence a successful LOCAL-CERT test automatically verifies the small-gradient condition in Lemma C.1; this lemma then guarantees the localization properties required to validate the candidate moment bounds constructed in Proposition C.2.

Lemma C.4 (Soundness of LOCAL-CERT)**.** *For a valid successful certificate,* $\mathbf{x}_\star$ *exists as above, its Hessian satisfies* $(m/2)\mathbf{I} \preceq \mathbf{H}_\star \preceq U_{\mathrm{loc}}\mathbf{I}$*, and resetting* $\mathbf{v} = \mathbf{x}_c$ *gives*

$$\Delta_{\mathrm{entry}} \le \frac{8}{\underline{\mu} m}\|\nabla f(\mathbf{x}_c)\|^2 \le \delta_{\mathrm{exit}}\delta_{\mathrm{loc}} < \delta_{\mathrm{loc}}. \tag{C.6}$$

The one-way switch and its stopping time. The certificate LOCAL-CERT is evaluated only in GLOBAL mode, only at stored iterates, and before the next sketch is drawn; hence the switch time

$$K_{\mathrm{loc}} := \inf\{k \ge 0 : \textsc{Local-Cert}(\mathbf{x}_k) = \textbf{true}\} \tag{C.7}$$

with $\inf \varnothing = \infty$, is a stopping time for the pre-sketch filtration. At K_{loc} the momentum state is reset, $\mathbf{v}_{K_{\mathrm{loc}}} = \mathbf{x}_{K_{\mathrm{loc}}}$, which, for a valid successful certificate, gives a bound on the two-state entry energy (Lemma C.4). The certificate is never re-evaluated, and the method never returns to GLOBAL mode. A GLOBAL iteration uses at most two reduced solves; a LOCAL iteration uses at most one.

Let $\mathcal{F}_k$ include the decision and configuration produced by any certificate computation at iteration k, then K_{loc} is a stopping time. Conditional on $\mathcal{F}_{K_{\mathrm{loc}}}$, the local configuration is fixed and subsequent sketches are independent draws from its frozen law.

For the asymptotic local conclusions, continue the local recursion after any gradient-tolerance stopping time, as in Appendix A. Before termination its iterates coincide with Algorithm 2; continuation is an analysis convention.

Theorem C.5 (Local phase after the certified switch)**.** *On every history with* $K_{\mathrm{loc}} < \infty$ *and a valid successful certificate, let* $\Delta_{K_{\mathrm{loc}}}$ *refer to the reset state, and set*

$$T_\delta^{\mathrm{loc}} = \inf\{j \ge 0 : \Delta_{K_{\mathrm{loc}}+j} > \delta_{\mathrm{loc}}\}, \qquad \rho = 1 - \tfrac{1}{2}\sqrt{\underline{\mu}/\overline{\nu}}. \tag{C.8}$$

Conditional on that history,

$$\mathbb{E}[\Delta_{K_{\mathrm{loc}}+j}\mathbb{1}_{\{T_\delta^{\mathrm{loc}} > j\}} \mid \mathcal{F}_{K_{\mathrm{loc}}}] \le \rho^j \Delta_{K_{\mathrm{loc}}}, \tag{C.9}$$

$$\mathbb{P}(T_\delta^{\mathrm{loc}} < \infty \mid \mathcal{F}_{K_{\mathrm{loc}}}) \le \Delta_{K_{\mathrm{loc}}}/\delta_{\mathrm{loc}} \le \delta_{\mathrm{exit}}. \tag{C.10}$$

The analogous survival-weighted function bound has prefactor $C_f\Delta_{K_{\mathrm{loc}}}$*. On no exit the continued local recursion converges almost surely. Its function-gap expectation conditioned on no exit is at most* $C_f\rho^j\Delta_{K_{\mathrm{loc}}}/(1 - \delta_{\mathrm{exit}})$*.*

Thus the inverse local contraction gap is $O(\sqrt{\overline{\nu}/\underline{\mu}})$, measured from the *first* certified switch, not a future-dependent last-entry index. No certificate is re-evaluated and no safeguard is reintroduced after switching. Local exit remains an explicit failure event, not a pathwise-impossibility claim.

Proposition C.6 (Eventual activation near a nondegenerate minimizer)**.** *Suppose a global-mode trajectory converges to a nondegenerate minimizer, certification is attempted at all sufficiently late stored points, and the certificates are valid with $m, \underline{\mu}, \delta_{\mathrm{loc}}$ bounded below by positive constants and U_{loc}, L_H bounded above. Then* LOCAL-CERT *eventually succeeds.*

C.2 PROOFS

Proof of Lemma C.1. Let $r_c = m/(2L_H)$. On the candidate ball, $\nabla^2 f \succeq (m - L_H r_c)\mathbf{I} = (m/2)\mathbf{I}$. For $\|\mathbf{z} - \mathbf{x}_c\| = r_c$, Taylor's lower bound gives

$$f(\mathbf{z}) - f(\mathbf{x}_c) \geq -\|\nabla f(\mathbf{x}_c)\| r_c + (m/4) r_c^2 > 0.$$

The minimum on the closed ball is therefore attained in its interior; strong convexity gives uniqueness and stationarity. Strong monotonicity of the gradient then gives $\|\mathbf{x}_c - \mathbf{x}_\star\| \leq 2\|\nabla f(\mathbf{x}_c)\|/m < m/(4L_H)$. In particular, $\|\mathbf{H}_c - \mathbf{H}_\star\| < m/4$ and $\mathbf{H}_\star \succeq (m/2)\mathbf{I}$, so $-\frac{1}{2}\mathbf{H}_\star \preceq \mathbf{H}_c - \mathbf{H}_\star \preceq \frac{1}{2}\mathbf{H}_\star$. Finally $B(\mathbf{x}_\star, m/(4L_H))$ is contained in the candidate ball, which proves (C.3). □

Proof of Proposition C.2. In the Gaussian case, (C.2) implies $T_r(\mathbf{H}_\star) \leq 2T_r(\mathbf{H}_c)$. Lemma B.1 therefore implies the mean interface with $c_\mu = c_{\mathrm{G}}$ and $\tau = 3\eta$. The $\tau/\eta \leq 3$ case of Lemma B.2, with the same absolute constant C_{G}, proves (C.4).

For ARLS set $\mathbf{A}_\star = \mathbf{H}_\star + \eta\mathbf{I}$. The same localization gives $\frac{1}{2}\mathbf{A}_\star \preceq \mathbf{A}_c \preceq \frac{3}{2}\mathbf{A}_\star$. Compression, inversion, and Lemma B.5 yield

$$\begin{aligned}\mathbb{E}[\mathbf{I}_B^\top(\mathbf{A}_{\star,BB})^{-1}\mathbf{I}_B] &\succeq \tfrac{1}{2}\mathbb{E}[\mathbf{I}_B^\top(\mathbf{A}_{c,BB})^{-1}\mathbf{I}_B] \\ &\succeq \tfrac{1}{4}(\mathbf{A}_c + \tau_{\mathrm{A},c}\mathbf{I})^{-1} \succeq \tfrac{1}{6}(\mathbf{A}_\star + \tau_{\mathrm{A},c}\mathbf{I})^{-1}.\end{aligned}$$

Congruence by $\mathbf{H}_\star^{1/2}$ and Lemma B.6 prove (C.5). Since $\eta > 0$ in both constructions, the reduced-system bound (3.3) holds with $\ell_\eta = \eta$. □

Proof of Lemma C.4. As shown after Definition C.3, the residual test implies (LOC). Lemma C.1 therefore supplies the minimizer, localization, and regularity ball, and Proposition C.2 validates the computed moment bounds. Write $\mathbf{e} = \mathbf{x}_c - \mathbf{x}_\star$, $\mathbf{g}_c = \nabla f(\mathbf{x}_c)$, and

$$\widehat{\mathbf{H}} = \int_0^1 \nabla^2 f(\mathbf{x}_\star + t\mathbf{e})\, dt, \qquad \mathbf{g}_c = \widehat{\mathbf{H}}\mathbf{e}.$$

The entire segment lies in the candidate ball, so $\widehat{\mathbf{H}} \succeq (m/2)\mathbf{I}$ and $\|\widehat{\mathbf{H}} - \mathbf{H}_\star\| \leq L_H\|\mathbf{e}\| < m/4$. In particular $\mathbf{H}_\star \preceq 2\widehat{\mathbf{H}}$. For the reset $\mathbf{v} = \mathbf{x}_c$, the moment lower bound and $\|\mathbf{e}\| \leq 2\|\mathbf{g}_c\|/m$ yield

$$\begin{aligned}\Delta_{\mathrm{entry}} &\leq \frac{2}{\underline{\mu}}\mathbf{e}^\top\mathbf{H}_\star\mathbf{e} \leq \frac{4}{\underline{\mu}}\mathbf{e}^\top\widehat{\mathbf{H}}\mathbf{e} = \frac{4}{\underline{\mu}}\mathbf{e}^\top\mathbf{g}_c \\ &\leq \frac{8}{\underline{\mu} m}\|\mathbf{g}_c\|^2 \leq \delta_{\mathrm{exit}}\delta_{\mathrm{loc}}.\end{aligned}$$

The proof bounds the Hessian energy through $\mathbf{e}^\top\mathbf{g}_c$, rather than through $U_{\mathrm{loc}}\|\mathbf{e}\|^2$, avoiding a U_{loc}/m loss. No full Newton solve is needed by the test: $\widehat{\mathbf{H}}$ occurs only in this proof. □

Proof of Theorem C.5. Condition on a history at $K_{\mathrm{loc}} < \infty$ with a valid successful certificate. Its minimizer, law, ridge, moment bounds, and radius are now fixed, and future sketches use fresh independent randomness from this law. Lemma C.4 gives $\Delta_{K_{\mathrm{loc}}} \leq \delta_{\mathrm{exit}}\delta_{\mathrm{loc}} < \delta_{\mathrm{loc}}$. Apply the general-budget stopping and conditional-expectation estimates established in the proof of Theorem 3.3 to the reindexed local recursion, with initial energy $\Delta_{K_{\mathrm{loc}}}$ and threshold δ_{loc}. Conditional on $\mathcal{F}_{K_{\mathrm{loc}}}$, these estimates give (C.9) and (C.10). For the installed local configuration, use $C_f = U\underline{\mu}/h_-$ as in that proof. The deterministic conversion $f(\mathbf{x}) - f(\mathbf{x}_\star) \leq (C_f/2)\Delta$ gives the stated survival-weighted function bound, with the harmless relaxation $C_f/2 \leq C_f$. Dividing by $\mathbb{P}(T_\delta^{\mathrm{loc}} = \infty \mid \mathcal{F}_{K_{\mathrm{loc}}}) \geq 1 - \delta_{\mathrm{exit}}$ gives the stated function-gap expectation conditioned on the switch history and no exit. The same Borel–Cantelli argument proves almost-sure convergence on no exit. □

Proof of Proposition C.6. Continuity gives $\nabla f(\mathbf{x}_k) \to 0$. The uniform positive lower bounds and finite upper bounds make the energy form of (ENTRY) hold at all sufficiently late certification attempts. Condition (LOC) then follows automatically. □

D GLOBAL GUARANTEES AND PROOFS

This appendix follows the safeguarded phase from its assumptions to the combined guarantee. Appendix D.1 fixes the trial region and conditional sketch event; Appendix D.2 proves descent and monotonicity; Appendix D.3 establishes uniform quantitative decrease; Appendix D.4 derives the global horizon; Appendix D.5 combines it with certified local retention. Each proof follows its statement.

D.1 GLOBAL ASSUMPTIONS

We now analyze the safeguarded phase that precedes the certified local entry of Section C.1. Throughout this section every trial uses ridge floor $\eta = 0$, as in the original global method. Positive local ridge floors introduced at the switch do not change the constants or proofs in this section.

The safeguard can be analyzed without assuming convexity. Throughout this section we consider the GLOBAL phase of Algorithm 2, that is, all iterations $k < K_{\mathrm{loc}}$; once the method switches, monotonicity of the stored objective is intentionally no longer enforced. We first isolate the regularity of the objective and the probabilistic property of the sketch. The assumptions are stated on the points and line-search segments that the algorithm actually queries.

Assumption D.1 (Global trial region)**.** *The function is bounded below by $f_{\inf}$ on a compact set $\mathcal{K}$ containing the initial sublevel set $\{\mathbf{x} : f(\mathbf{x}) \le f(\mathbf{x}_0)\}$, every momentum point allowed by $R_{\max}$, and every trial segment generated by a sketch satisfying $\|\mathbf{S}\| \le L_S$, where L_S is the constant in Assumption D.2. On $\mathcal{K}$,*

$$\|\nabla f(\mathbf{z})\| \le G, \qquad \|\nabla^2 f(\mathbf{z})\| \le U_H, \qquad \|\nabla^2 f(\mathbf{z}) - \nabla^2 f(\mathbf{y})\| \le L_H \|\mathbf{z} - \mathbf{y}\|. \tag{D.1}$$

Assumption D.2 (Sketch capture)**.** *Let $\mathcal{F}_k$ contain the history before $\mathbf{S}_k$ is sampled, so $\mathbf{x}_k$ and $\mathbf{y}_k$ are $\mathcal{F}_k$-measurable. There are constants $L_S < \infty$, $\chi > 0$, and $\delta_{\mathrm{sk}} \in [0, 1)$ such that, conditionally on $\mathcal{F}_k$, with probability at least $1 - \delta_{\mathrm{sk}}$,*

$$\|\mathbf{S}_k\| \le L_S, \qquad \|\mathbf{S}_k \nabla f(\mathbf{z})\|^2 \ge \chi \|\nabla f(\mathbf{z})\|^2 \tag{D.2}$$

simultaneously for every nonstationary $\mathbf{z} \in \{\mathbf{x}_k, \mathbf{y}_k\}$.

Assumption D.2 is a finite-iterate, conditional event. It does not replace a dependent sequence by independent copies: the base points may depend on the full past, and only the newly sampled sketch is conditioned upon. For a particular sketch family, L_S, χ, and δ_{sk} can be obtained from its own tail bounds.

Why only bounded-norm trial segments need regularity bounds. The compact-region assumption is needed for quantitative decrease on good sketches, not for monotonicity on all sketches. In particular, $\mathcal{K}$ need not contain rejected segments with $\|\mathbf{S}\| > L_S$. This distinction matters for untruncated Gaussian sketches: even $f(x) = x^4 + x$ at $x = 0$ gives $p = -S^2/c_2$ in the scalar global trial, so the first trial segments are not uniformly bounded over all draws.

To see a sufficient setting for Assumption D.1, suppose the initial sublevel set is compact and the Hessian is locally Lipschitz. Let $\mathcal{B}$ be its closed $R_{\max}$-neighborhood and $G_0 = \sup_{\mathcal{B}} \|\nabla f\| < \infty$. Every trial base lies in $\mathcal{B}$. For $\|\mathbf{S}\| \le L_S$,

$$\|\mathbf{p}\| \le \frac{L_S^2}{c_2} \|\mathbf{g}\|^{1-\gamma} \le \frac{L_S^2}{c_2} G_0^{1-\gamma}.$$

Thus all relevant trial segments, whose tested steps have $t \le 1$, lie in a fixed compact enlargement of $\mathcal{B}$; finite regularity constants can be chosen there. This argument first bounds gradients at bases and then bounds segments, so it is not circular. On other sketches, a nonzero direction is a strict descent direction and Armijo terminates by differentiability; a zero direction is accepted with equality. The safeguard or fallback still keeps the stored iterate in the initial sublevel set. No bound on those rejected segments is used.

For example, normalized Gaussian sketches satisfy the capture assumption with fixed positive probability without truncation: take $L_S = 2\sqrt{n}$, so Markov gives $\mathbb{P}(\|\mathbf{S}\| > L_S) \le 1/4$ using $\mathbb{E}\|\mathbf{S}\|_F^2 = n$. Choose $\chi > 0$ so that $\mathbb{P}(\chi_s^2 < s\chi) \le 1/8$. For each pre-sketch nonzero gradient, $\|\mathbf{Sg}\|^2/\|\mathbf{g}\|^2 \sim \chi_s^2/s$. A union bound for the norm and the two gradients gives joint capture probability at least $1/2$, conditionally on the past; their mutual independence is unnecessary.

D.2 Safeguarded monotonicity

We first relate the trial direction to its decrease certificate, then show that the safeguard preserves stored-objective monotonicity even when a sketch fails the capture event. For one call to Algorithm 1, abbreviate

$$\mathbf{B} = \mathbf{S}\nabla^2 f(\mathbf{z})\mathbf{S}^\top, \qquad \mathbf{M} = \mathbf{B} + c_1\Lambda\mathbf{I}_s + c_2\|\nabla f(\mathbf{z})\|^\gamma\mathbf{I}_s,$$
$$q(\mathbf{z},\mathbf{S}) = (\mathbf{S}\nabla f(\mathbf{z}))^\top\mathbf{M}^{-1}\mathbf{S}\nabla f(\mathbf{z}). \tag{D.3}$$

Lemma D.3 (Descent and decrease certificate)**.** *If $\nabla f(\mathbf{z}) \neq 0$, then $\mathbf{M} \succ 0$ and*

$$\nabla f(\mathbf{z})^\top\mathbf{p} = -q(\mathbf{z},\mathbf{S}) \le 0, \qquad \mathsf{Dec}(\mathbf{z},\mathbf{S}) = \alpha t q(\mathbf{z},\mathbf{S}). \tag{D.4}$$

The inequality is strict whenever $\mathbf{S}\nabla f(\mathbf{z}) \neq 0$.

Proof of Lemma D.3. Let $\mathbf{g} = \nabla f(\mathbf{z})$. If $\lambda_{\min}(\mathbf{B}) \ge 0$, then $\Lambda = 0$ and $\mathbf{M} \succeq c_2\|\mathbf{g}\|^\gamma\mathbf{I}_s \succ 0$. If $\lambda_{\min}(\mathbf{B}) < 0$, then

$$\lambda_{\min}(\mathbf{M}) \ge (c_1 - 1)\Lambda + c_2\|\mathbf{g}\|^\gamma > 0. \tag{D.5}$$

Thus $\mathbf{M}$ is positive definite in either case. Since $\mathbf{p} = -\mathbf{S}^\top\mathbf{M}^{-1}\mathbf{Sg}$,

$$\mathbf{g}^\top\mathbf{p} = -(\mathbf{Sg})^\top\mathbf{M}^{-1}\mathbf{Sg} = -q(\mathbf{z},\mathbf{S}). \tag{D.6}$$

The quadratic form is nonnegative and is positive exactly when $\mathbf{Sg} \neq 0$. Substitution in the definition of the Armijo certificate gives $\mathsf{Dec}(\mathbf{z},\mathbf{S}) = \alpha t q(\mathbf{z},\mathbf{S})$. □

Lemma D.4 (Monotonicity of the stored iterates in GLOBAL mode)**.** *Every* GLOBAL*-mode iteration of Algorithm 2 satisfies $f(\mathbf{x}_{k+1}) \le f(\mathbf{x}_k)$. If the accelerated trial is accepted, then*

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \ge \sigma\mathsf{Dec}_k^A; \tag{D.7}$$

otherwise, $f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \ge \mathsf{Dec}_k^B$.

Proof of Lemma D.4. If the accelerated trial is accepted, its acceptance test is exactly (D.7). If it is rejected, the fallback Armijo condition and Lemma D.3 give

$$f(\mathbf{x}_{k+1}) \le f(\mathbf{x}_k) + \alpha t_k^B\nabla f(\mathbf{x}_k)^\top\mathbf{p}_k^B = f(\mathbf{x}_k) - \mathsf{Dec}_k^B. \tag{D.8}$$

Both certificates are nonnegative, proving monotonicity. □

On bad sketch events the stored objective still does not increase. Corollary D.8 below exploits this fact using indicators.

D.3 Uniform Armijo acceptance and quantitative decrease

The next two lemmas turn a good sketch into a uniform gradient-based decrease. The Armijo proof retains the cubic Taylor remainder; this is where the global restriction $\gamma \le 1/2$ is used.

Lemma D.5 (Uniform Armijo step)**.** *Suppose Assumption D.1 holds and $0 \le \gamma \le \frac{1}{2}$. On the event $\|\mathbf{S}\| \le L_S$, Armijo backtracking accepts a step satisfying $t \ge \underline{t}$, where*

$$t_{\mathrm{safe}} = \min\left\{1 - \alpha, \sqrt{\frac{3(1-\alpha)c_2^2}{L_H L_S^4 G^{1-2\gamma}}}\right\}, \qquad \underline{t} = \beta\min\{1, t_{\mathrm{safe}}\}. \tag{D.9}$$

When $L_H = 0$, the second term in the minimum is interpreted as $+\infty$.

Proof of Lemma D.5. Fix a nonstationary trial base $\mathbf{z}$, set $h = \|\nabla f(\mathbf{z})\|$, and write $\mathbf{u} = -\mathbf{M}^{-1}\mathbf{S}\nabla f(\mathbf{z})$, so $\mathbf{p} = \mathbf{S}^\top \mathbf{u}$ and $q = \mathbf{u}^\top \mathbf{M}\mathbf{u}$. Since $\mathbf{M} \succeq c_2 h^\gamma \mathbf{I}_s$,

$$\|\mathbf{p}\|^2 \le L_S^2 \|\mathbf{u}\|^2 \le \frac{L_S^2}{c_2 h^\gamma} q, \tag{D.10}$$

$$q \le \frac{\|\mathbf{S}\nabla f(\mathbf{z})\|^2}{c_2 h^\gamma} \le \frac{L_S^2}{c_2} h^{2-\gamma}. \tag{D.11}$$

Combining these inequalities yields

$$\frac{\|\mathbf{p}\|^3}{q} \le \frac{L_S^4}{c_2^2} h^{1-2\gamma} \le \frac{L_S^4}{c_2^2} G^{1-2\gamma}. \tag{D.12}$$

The last inequality is where $0 \le \gamma \le \frac{1}{2}$ is used.

Hessian Lipschitz continuity on the trial segment gives

$$f(\mathbf{z}+t\mathbf{p}) \le f(\mathbf{z}) + t\nabla f(\mathbf{z})^\top \mathbf{p} + \frac{t^2}{2}\mathbf{p}^\top \nabla^2 f(\mathbf{z})\mathbf{p} + \frac{L_H t^3}{6}\|\mathbf{p}\|^3. \tag{D.13}$$

Moreover,

$$\mathbf{p}^\top \nabla^2 f(\mathbf{z})\mathbf{p} = \mathbf{u}^\top \mathbf{B}\mathbf{u} \le \mathbf{u}^\top \mathbf{M}\mathbf{u} = q, \tag{D.14}$$

because $\mathbf{M} - \mathbf{B} \succeq 0$. Using $\nabla f(\mathbf{z})^\top \mathbf{p} = -q$, the Armijo condition follows whenever

$$(1-\alpha) - \frac{t}{2} - \frac{L_H t^2}{6}\frac{\|\mathbf{p}\|^3}{q} \ge 0. \tag{D.15}$$

For $t \le 1-\alpha$, the first two terms are at least $(1-\alpha)/2$. Equation (D.12) then shows that every $t \le t_{\text{safe}}$ is accepted. Backtracking from one by factors of β therefore returns $t \ge \beta \min\{1, t_{\text{safe}}\} = \underline{t}$. The case $q = 0$ has $\mathbf{p} = 0$ and satisfies Armijo with equality. □

Lemma D.6 (Quantitative trial decrease)**.** *Under Assumptions D.1–D.2, every good trial based at a point* $\mathbf{z}$ *satisfies*

$$\mathsf{Dec}(\mathbf{z}, \mathbf{S}) \ge c_{\text{dec}} \|\nabla f(\mathbf{z})\|^2, \qquad c_{\text{dec}} := \frac{\alpha \underline{t}\, \chi}{(1+c_1)L_S^2 U_H + c_2 G^\gamma}. \tag{D.16}$$

Proof of Lemma D.6. On a good sketch event,

$$\|\mathbf{B}\| \le L_S^2 U_H, \qquad \Lambda \le \|\mathbf{B}\|, \tag{D.17}$$

and hence

$$\lambda_{\max}(\mathbf{M}) \le (1+c_1)L_S^2 U_H + c_2 G^\gamma. \tag{D.18}$$

It follows from (D.3), the sketch-capture inequality, and Lemma D.5 that

$$\mathsf{Dec}(\mathbf{z}, \mathbf{S}) = \alpha t q(\mathbf{z}, \mathbf{S}) \tag{D.19}$$

$$\ge \frac{\alpha \underline{t} \|\mathbf{S}\nabla f(\mathbf{z})\|^2}{(1+c_1)L_S^2 U_H + c_2 G^\gamma} \tag{D.20}$$

$$\ge c_{\text{dec}} \|\nabla f(\mathbf{z})\|^2. \tag{D.21}$$

□

D.4 A GLOBAL HORIZON FOR STATIONARITY OR CERTIFIED ENTRY

We count good iterations using conditional probability bounds. This yields a horizon for a stationary query or a switch, without requiring independence of the state-dependent capture indicators.

The certified switch can only shorten the safeguarded phase: on any sample path, every iteration before K_{loc} is a GLOBAL-mode iteration and is governed by the safeguard. Define the first ε-stationary query time

$$T_\varepsilon := \inf\{k \ge 0 : \text{iteration } k \text{ queries a base point } \mathbf{z} \text{ with } \|\nabla f(\mathbf{z})\| \le \varepsilon\}, \tag{D.22}$$

and recall the switch time K_{loc} from (C.7). Because the certificate is checked before $\mathbf{S}_k$ is sampled, K_{loc} is a stopping time for the pre-sketch filtration $(\mathcal{F}_k)$.

Theorem D.7 (Global phase: stationarity or certified local entry)**.** *Let Assumptions D.1–D.2 hold for the* GLOBAL*-mode iterations, let* $0 \le \gamma \le \frac{1}{2}$*, and let* $\varepsilon > 0$*. Let* $c_{\mathrm{dec}} > 0$ *be the uniform trial-decrease constant defined in* (D.16) *of Lemma D.6; explicitly,*

$$c_{\mathrm{dec}} = \frac{\alpha \underline{t}\, \chi}{(1+c_1)L_S^2 U_H + c_2 G^{\gamma}},$$

where $\underline{t}$ *is the uniform Armijo lower bound in* (D.9)*. Set* $p_{\mathrm{sk}} := 1 - \delta_{\mathrm{sk}} > 0$*. For any* $\vartheta \in (0,1)$*, define*

$$N_\varepsilon(\vartheta) = \left\lfloor \frac{f(\mathbf{x}_0) - f_{\inf}}{(1-\vartheta)p_{\mathrm{sk}}\sigma c_{\mathrm{dec}}\varepsilon^2} \right\rfloor + 1. \tag{D.23}$$

Then, for every integer $N \ge N_\varepsilon(\vartheta)$*, Algorithm 2, with the gradient norm checked at every queried base point, satisfies*

$$\mathbb{P}\big(\min\{T_\varepsilon, K_{\mathrm{loc}}\} < N\big) \ge 1 - \exp\left(-\frac{\vartheta^2 p_{\mathrm{sk}} N}{2}\right). \tag{D.24}$$

Thus, with this probability, one of the first N *iterations* $k = 0, \dots, N-1$ *queries an* ε*-stationary base point or starts the certified local phase. Up to this first event, at most* $2N$ *reduced linear systems are solved, excluding certification and preprocessing.*

Proof of Theorem D.7. Write $F_0 = f(\mathbf{x}_0) - f_{\inf}$, $d_\varepsilon = \sigma c_{\mathrm{dec}}\varepsilon^2$, and $\tau_\varepsilon = \min\{T_\varepsilon, K_{\mathrm{loc}}\}$. Let $\mathcal{E}_k$ be the joint good-sketch event of Assumption D.2 for a global iteration that is started. It covers all nonstationary trial bases queried in that iteration, not an event conditioned on which branch is eventually accepted.

Predictable continuation of the indicators. Include the certification decision and the pre-trial radius restart in $\mathcal{F}_k$, but not the new trial sketch. The event

$$\mathcal{A}_k := \{T_\varepsilon \ge k\} \cap \{K_{\mathrm{loc}} > k\}$$

is $\mathcal{F}_k$-measurable: it says that no previous iteration has returned an ε-stationary point and that the algorithm has not switched before drawing the current sketch. Define

$$I_k = \begin{cases} \mathbb{1}_{\mathcal{E}_k}, & \text{on } \mathcal{A}_k, \\ 1, & \text{on } \mathcal{A}_k^c. \end{cases} \tag{D.25}$$

After a return, extend the history trivially. The artificial indicators are used only for the proof and do not modify the algorithm. Assumption D.2 implies

$$0 \le I_k \le 1, \qquad \mathbb{E}[I_k \mid \mathcal{F}_k] \ge p_{\mathrm{sk}} \quad \text{almost surely for every } k. \tag{D.26}$$

Notice that $\{T_\varepsilon \ge k\}$ is predictable even though a stationary fallback query in the current iteration can depend on $\mathbf{S}_k$. We do not assert that T_ε itself is a stopping time for the pre-sketch filtration.

Descent counts only good iterations. For every completed, nonterminal global iteration $k < \tau_\varepsilon$, Lemma D.4 guarantees nonnegative objective decrease on all sketches. If $\mathcal{E}_k$ holds and the accelerated trial is accepted, Lemma D.6 and the safeguard give

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \ge \sigma c_{\mathrm{dec}} \|\nabla f(\mathbf{y}_k)\|^2 \ge d_\varepsilon.$$

If the trial is rejected, the fallback base is actually queried and its gradient norm also exceeds ε. Armijo and $\sigma \le 1$ then give

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \ge c_{\mathrm{dec}} \|\nabla f(\mathbf{x}_k)\|^2 \ge d_\varepsilon.$$

Hence, pathwise on $\{k < \tau_\varepsilon\}$,

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \ge d_\varepsilon I_k.$$

On $\{\tau_\varepsilon \ge N\}$ the first N iterations are all completed nonterminal global iterations. Summing only these global updates yields

$$d_\varepsilon \sum_{k=0}^{N-1} I_k \le f(\mathbf{x}_0) - f(\mathbf{x}_N) \le F_0. \tag{D.27}$$

Thus $N \ge N_\varepsilon(\vartheta)$ implies the event inclusion

$$\{\tau_\varepsilon \ge N\} \subseteq \left\{ \sum_{k=0}^{N-1} I_k < (1-\vartheta)p_{\mathrm{sk}} N \right\}. \tag{D.28}$$

In particular, no descent is summed after the local switch.

Conditional Chernoff bound. For any $\lambda > 0$, since $I_k \in \{0, 1\}$,

$$\begin{aligned}\mathbb{E}[e^{-\lambda I_k} \mid \mathcal{F}_k] &= 1 - (1 - e^{-\lambda})\mathbb{E}[I_k \mid \mathcal{F}_k] \\ &\le 1 - p_{\mathrm{sk}} + p_{\mathrm{sk}} e^{-\lambda} \le \exp\bigl(p_{\mathrm{sk}}(e^{-\lambda} - 1)\bigr).\end{aligned} \tag{D.29}$$

Because $\sum_{j<k} I_j$ is $\mathcal{F}_k$-measurable, repeated conditioning gives

$$\mathbb{E}\exp\left(-\lambda \sum_{k=0}^{N-1} I_k\right) \le \exp\bigl(N p_{\mathrm{sk}}(e^{-\lambda} - 1)\bigr).$$

Markov's inequality with $\lambda = -\log(1 - \vartheta)$ therefore yields

$$\begin{aligned}\mathbb{P}\left(\sum_{k=0}^{N-1} I_k < (1 - \vartheta) p_{\mathrm{sk}} N\right) &\le \exp[-N p_{\mathrm{sk}}\{\vartheta + (1 - \vartheta)\log(1 - \vartheta)\}] \\ &\le \exp\bigl(-\tfrac{1}{2}\vartheta^2 p_{\mathrm{sk}} N\bigr)\,.\end{aligned} \tag{D.30}$$

The last inequality follows by integrating $-\log(1 - \vartheta) \ge \vartheta$ from 0 to ϑ. Combining this estimate with (D.28) proves (D.24). The concentration estimate is taken on the whole probability space, not conditionally on the future event $\{\tau_\varepsilon \ge N\}$, and requires no independence between the indicators. The solve count follows from at most two reduced solves per global iteration, with the same sketch reused in the fallback. □

The conclusion is deliberately stated for queried base points. This includes the stored iterate when the fallback is used and the momentum point when the accelerated trial is queried; no inequality between $\|\nabla f(\mathbf{x}_k)\|$ and $\|\nabla f(\mathbf{y}_k)\|$ is assumed. The proof follows the good-iteration counting idea of Fuji et al. (2025, Theorem 8), using a conditional exponential-moment argument rather than asserting that the state-dependent indicators are independent. Bad sketches cannot undo previously obtained objective decrease. The additional factor $((1 - \vartheta)p_{\mathrm{sk}})^{-1}$ in (D.23) allows enough iterations to accumulate the required number of good sketches; the old all-good horizon cannot simply be retained while replacing its probability bound. If no certificate ever fires, $K_{\mathrm{loc}} = \infty$ and (D.24) is a stationarity guarantee. The theorem does not assert that the global phase must reach LOCAL mode.

Corollary D.8 (An explicit confidence-dependent horizon)**.** *Under the hypotheses of Theorem D.7, let $\delta_{\mathrm{g}} \in (0, 1)$ and set*

$$M_\varepsilon(\delta_{\mathrm{g}}) = \max\left\{N_\varepsilon(1/2),\ \left\lceil \frac{8}{p_{\mathrm{sk}}} \log \frac{1}{\delta_{\mathrm{g}}} \right\rceil\right\}. \tag{D.31}$$

Then $\mathbb{P}(\min\{T_\varepsilon, K_{\mathrm{loc}}\} < M_\varepsilon(\delta_{\mathrm{g}})) \ge 1 - \delta_{\mathrm{g}}$. On this event the method uses at most $2M_\varepsilon(\delta_{\mathrm{g}})$ reduced solves before the stationary query or switch, excluding certification and preprocessing, where

$$M_\varepsilon(\delta_{\mathrm{g}}) = O\left(1 + \frac{f(\mathbf{x}_0) - f_{\inf}}{p_{\mathrm{sk}} \sigma c_{\mathrm{dec}} \varepsilon^2} + \frac{1}{p_{\mathrm{sk}}} \log \frac{1}{\delta_{\mathrm{g}}}\right).$$

Proof of Corollary D.8. Use Theorem D.7 with $\vartheta = 1/2$ and $N = M_\varepsilon(\delta_{\mathrm{g}})$. The definition of this horizon gives $N \ge N_\varepsilon(1/2)$ and $p_{\mathrm{sk}} N/8 \ge \log(1/\delta_{\mathrm{g}})$. Thus the failure probability is at most $e^{-p_{\mathrm{sk}} N/8} \le \delta_{\mathrm{g}}$. The complexity and solve bounds follow immediately. □

Interpretation of the global complexity bound. Corollary D.8 converts Theorem D.7 into an iteration bound for a prescribed failure probability δ_{g}. The two terms defining $M_\varepsilon(\delta_{\mathrm{g}})$ account for sufficient objective decrease and the desired confidence, respectively. Crucially, bad sketches cannot increase the stored objective, so a uniformly positive conditional capture probability suffices; it need not approach one as $\varepsilon \to 0$. For fixed problem, sketch, and algorithmic constants, the method therefore queries an ε-stationary point or triggers the certified local switch within $O(\varepsilon^{-2} + \log(1/\delta_{\mathrm{g}}))$ iterations with probability at least $1 - \delta_{\mathrm{g}}$.

D.5 Combined guarantee with a conditional local rate

The preceding global horizon and the certified-entry results now describe the same two-mode algorithm. The local rate below is stated as a conditional expectation on retained trajectories; its proof still uses the stopped estimates of Theorem C.5.

Theorem D.9 (Combined global-to-local guarantee)**.** *Suppose Assumptions D.1–D.2 hold for the* GLOBAL *phase of Algorithm 2, with $\alpha \in (0, 1/2)$, $0 < \gamma \le 1/2$, and $\varepsilon > 0$. Fix $\delta_{\mathrm{g}}, \delta_{\mathrm{exit}} \in (0,1)$ and $\delta_{\mathrm{cert}} \ge 0$ such that $\delta_{\mathrm{g}} + \delta_{\mathrm{cert}} + \delta_{\mathrm{exit}} < 1$. Assume the total probability of an invalid certificate across all attempted switches is at most δ_{cert}, and every valid successful configuration satisfies Definition C.3 with exit budget δ_{exit}. Let $M = M_\varepsilon(\delta_{\mathrm{g}})$ from Corollary D.8.*

On each valid switching history with $K_{\mathrm{loc}} < \infty$, consider the continued local recursion without gradient-tolerance stopping, with its certified configuration and associated minimizer $\mathbf{x}_\star$ fixed conditionally on $\mathcal{F}_{K_{\mathrm{loc}}}$. Let $\Delta_{K_{\mathrm{loc}}} \le \delta_{\mathrm{exit}}\delta_{\mathrm{loc}}$ be the reset energy and set $\mathcal{G}_{\mathrm{loc}} = \{T_\delta^{\mathrm{loc}} = \infty\}$, with T_δ^{loc} as in (C.8).

Then, with probability at least $1 - \delta_{\mathrm{g}} - \delta_{\mathrm{cert}} - \delta_{\mathrm{exit}}$, either an ε-stationary point is queried before iteration M, or a valid switch occurs at $K_{\mathrm{loc}} < M$ and $\mathcal{G}_{\mathrm{loc}}$ holds. On every valid switching history, $\mathbb{P}(\mathcal{G}_{\mathrm{loc}} \mid \mathcal{F}_{K_{\mathrm{loc}}}) \ge 1 - \delta_{\mathrm{exit}}$. Moreover, with $\rho = 1 - \frac{1}{2}\sqrt{\underline{\mu}/\overline{\nu}}$ and $C_f = U\underline{\mu}/h_-$ for the installed configuration, for every $j \ge 0$,

$$\mathbb{E}[f(\mathbf{x}_{K_{\mathrm{loc}}+j}) - f(\mathbf{x}_\star) \mid \mathcal{F}_{K_{\mathrm{loc}}}, \mathcal{G}_{\mathrm{loc}}] \le \frac{C_f}{2(1-\delta_{\mathrm{exit}})}\rho^j \Delta_{K_{\mathrm{loc}}}. \tag{D.32}$$

Proof of Theorem D.9. Global horizon and local retention. Corollary D.8 gives $\mathbb{P}(\min\{T_\varepsilon, K_{\mathrm{loc}}\} \ge M) \le \delta_{\mathrm{g}}$. The probability of any invalid certification is at most δ_{cert}. On a valid switching history, Theorem C.5 bounds the conditional local exit probability by δ_{exit}, proving the stated retention bound. Sum these conditional bounds over the disjoint valid switching events $\{K_{\mathrm{loc}} = k\}$, $k < M$. The tower property bounds the probability of a valid switch before M followed by exit by δ_{exit}. A union bound over these three failure types proves the claimed global-to-local probability bound.

From the stopped estimate to a conditional rate. Fix a valid switching history. Since $\mathcal{G}_{\mathrm{loc}} \subseteq \{T_\delta^{\mathrm{loc}} > j\}$ and the energy is nonnegative, (C.9) implies

$$\mathbb{E}[\Delta_{K_{\mathrm{loc}}+j}\mathbb{1}_{\mathcal{G}_{\mathrm{loc}}} \mid \mathcal{F}_{K_{\mathrm{loc}}}] \le \mathbb{E}[\Delta_{K_{\mathrm{loc}}+j}\mathbb{1}_{\{T_\delta^{\mathrm{loc}} > j\}} \mid \mathcal{F}_{K_{\mathrm{loc}}}] \le \rho^j \Delta_{K_{\mathrm{loc}}}.$$

Conditioning on the event means normalizing by its conditional probability, which is at least $1 - \delta_{\mathrm{exit}} > 0$ by the stated retention bound:

$$\begin{aligned}\mathbb{E}[\Delta_{K_{\mathrm{loc}}+j} \mid \mathcal{F}_{K_{\mathrm{loc}}}, \mathcal{G}_{\mathrm{loc}}] &= \frac{\mathbb{E}[\Delta_{K_{\mathrm{loc}}+j}\mathbb{1}_{\mathcal{G}_{\mathrm{loc}}} \mid \mathcal{F}_{K_{\mathrm{loc}}}]}{\mathbb{P}(\mathcal{G}_{\mathrm{loc}} \mid \mathcal{F}_{K_{\mathrm{loc}}})} \\ &\le \frac{\rho^j \Delta_{K_{\mathrm{loc}}}}{1-\delta_{\mathrm{exit}}}.\end{aligned}$$

This gives the conditional energy estimate. On retained trajectories, $f(\mathbf{x}) - f(\mathbf{x}_\star) \le (C_f/2)\Delta$, as proved in Appendix A.3, giving (D.32). Almost-sure convergence follows from Theorem C.5.

The event $\mathcal{G}_{\mathrm{loc}}$ depends on future sketches. We first establish the stopped estimates under the original conditional sketch law and only then condition on this event. □

E The original zero-floor RLS branch

Lemma E.1 (Local quadratic equivalence and preconditioned PL)**.** *Suppose $(1-\varepsilon)\mathbf{H}_\star \preceq \nabla^2 f(\mathbf{z}) \preceq (1+\varepsilon)\mathbf{H}_\star$ on a ball centered at $\mathbf{x}_\star$, $\nabla f(\mathbf{x}_\star) = 0$, and $0 < \varepsilon < 1$. For every $\mathbf{x}$ in that ball and $\mathbf{M} \succ 0$,*

$$\frac{1-\varepsilon}{2}\|\mathbf{x}-\mathbf{x}_\star\|_{\mathbf{H}_\star}^2 \le f(\mathbf{x}) - f(\mathbf{x}_\star) \le \frac{1+\varepsilon}{2}\|\mathbf{x}-\mathbf{x}_\star\|_{\mathbf{H}_\star}^2, \tag{E.1}$$

$$\frac{1}{2}\|\nabla f(\mathbf{x})\|_{\mathbf{M}^{-1}}^2 \ge (1-\varepsilon)\mu_M(f(\mathbf{x}) - f(\mathbf{x}_\star)), \quad \mu_M = \lambda_{\min}(\mathbf{H}_\star^{1/2}\mathbf{M}^{-1}\mathbf{H}_\star^{1/2}). \tag{E.2}$$

Proof. Integrating the Hessian on the segment from $\mathbf{x}_\star$ to $\mathbf{x}$ gives (E.1). Also $\mathbf{H}_\star \succeq \mu_M \mathbf{M}$. The local strong-convexity inequality applied from $\mathbf{x}$ to $\mathbf{x}_\star$ gives

$$f(\mathbf{x}) - f(\mathbf{x}_\star) \le \nabla f(\mathbf{x})^\top (\mathbf{x} - \mathbf{x}_\star) - \frac{(1-\varepsilon)\mu_M}{2} \|\mathbf{x} - \mathbf{x}_\star\|_{\mathbf{M}}^2 \le \frac{\|\nabla f(\mathbf{x})\|_{\mathbf{M}^{-1}}^2}{2(1-\varepsilon)\mu_M}.$$

□

Proposition E.2 (First moments suffice without acceleration)**.** *Write* $\mathbf{H} = \mathbf{H}_\star$ *and* $\mathbf{Z}_\eta(\mathbf{S}) = \mathbf{S}^\top(\mathbf{S}\mathbf{H}\mathbf{S}^\top + \eta\mathbf{I})^{-1}\mathbf{S}$*, and let* $\mu = \lambda_{\min}(\mathbb{E}[\mathbf{H}^{1/2}\mathbf{Z}_\eta(\mathbf{S})\mathbf{H}^{1/2}]) > 0$*. For the quadratic objective,* $\mathbf{x}^+ = \mathbf{x} - \mathbf{Z}_\eta(\mathbf{S})\nabla f(\mathbf{x})$ *obeys*

$$\mathbb{E}[f(\mathbf{x}^+) - f(\mathbf{x}_\star) \mid \mathbf{x}] \le (1-\mu)(f(\mathbf{x}) - f(\mathbf{x}_\star)).$$

For the nonlinear nonaccelerated method, suppose the local spectral error is at most $\varepsilon < 1$*, the current reduced inverse satisfies* $\mathbf{Z}_k(\mathbf{S}) \succeq (1+\varepsilon)^{-1}\mathbf{Z}_\eta(\mathbf{S})$*, and the unit Armijo step is accepted. If* (B.1) *holds, then*

$$\mathbb{E}[f(\mathbf{x}^+) - f(\mathbf{x}_\star) \mid \mathbf{x}] \le \left(1 - \frac{2\alpha(1-\varepsilon)}{1+\varepsilon}\frac{c_\mu h}{h+\tau}\right)(f(\mathbf{x}) - f(\mathbf{x}_\star)). \tag{E.3}$$

Proof. For the quadratic, $(\mathbf{I} - \mathbf{P}_\eta)^2 \preceq \mathbf{I} - \mathbf{P}_\eta$. For the nonlinear case, use Armijo, the inverse comparison, the mean projection bound, and (E.2) with $\mathbf{M} = \mathbf{H}_\star + \tau\mathbf{I}$. No sketch-dependent line-search step is pulled outside an expectation. □

Corollary E.3 (The old RLS result, with its sampling budget retained)**.** *Apply Lemma B.5 directly to* $\mathbf{A} = \mathbf{H}$ *and use the unregularized coordinate solve,* $\eta = 0$*. Under that lemma's hypotheses,*

$$\overline{\mathbf{P}}_0 \succeq \tfrac{1}{2}\mathbf{H}(\mathbf{H} + \tau_{\mathrm{A}}\mathbf{I})^{-1}, \qquad \tau_{\mathrm{A}} \le T_r(\mathbf{H})/r, \qquad \mu_0 \ge \frac{r}{2n\bar{\kappa}_r}.$$

Thus the nonaccelerated quadratic iteration has contraction at most $1 - r/(2n\bar{\kappa}_r)$*, with* $s \ge Cc_{\mathrm{A}} r \log^3 n$ *draws. The same first-moment dependence holds locally by Proposition E.2; a vanishing additional ridge is allowed once* $c_2\|\mathbf{g}\|^\gamma/h$ *is sufficiently small. For superpolynomial* $1 + \tau_{\mathrm{A}}/h$*, the stated* $\log^3 n$ *budget is not justified by the quoted ARLS lemma without tracking its accuracy dependence. The positive shifted construction avoids that issue.*

Remark E.4 (Why the shift and ridge cannot simply be deleted)**.** *For diagonal* $\mathbf{H}$*, an unregularized coordinate projector is* $\mathbf{P}_0 = \mathrm{diag}(\mathbb{1}_{\{i \in B\}})$*. If* $\pi_i = \mathbb{P}(i \in B)$*, then*

$$\mu_0 = \min_i \pi_i, \qquad \nu_0 = \max_i \frac{1}{\pi_i} = \frac{1}{\mu_0}, \qquad \sqrt{\mu_0/\nu_0} = \mu_0.$$

This includes blocks obtained from nonuniform RLS draws and shows that the modular acceleration guarantee can give no improvement *for the old unregularized sampling law. The positive-ridge, shifted-score result in Theorem 3.6 is therefore a substantive addition, not a renaming of the old theorem or a claim of square-root improvement for every RLS variant.*

F Experimental protocols and supplementary results

F.1 Kernel logistic regression

Problem definition. We use a fixed stratified subset of $n = 5000$ examples from `real-sim`, selected with subset seed 0, and labels $y_i \in \{-1, +1\}$. We construct RBF kernel $K_{ij} = \exp(-\gamma\|x_i - x_j\|^2)$ with $\gamma = 0.5$. All methods start from $w_0 = 0$ and minimize the same objective:

$$f(w) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp[-y_i(\mathbf{K}\mathbf{w})_i]) + \frac{\lambda_{\mathrm{reg}}}{2}\mathbf{w}^\top\mathbf{K}\mathbf{w}, \qquad \lambda_{\mathrm{reg}} = 10^{-4}.$$

Sampling and Newton trials. For $s \in \{100, 400\}$, both RSRN variants sample coordinates without replacement using

$$p_{k,j} = \frac{1-\omega_k}{n} + \omega_k \frac{h_{k,j}}{\sum_\ell h_{k,\ell}}, \qquad \omega_k = 0.1 + 0.4\,[1 - \cos(\pi \min\{k/2000, 1\})]\,,$$

where $h_{k,j}$ is the latest Hessian-diagonal estimate. It is computed at the stored iterate, refreshed every 25 iterations through iteration 2000, and subsequently frozen. Thus the Hessian-diagonal mixture weight increases smoothly from 0.1 to 0.9.

Variation across repeated runs. To check whether the improvement depends on a particular random seed, we repeat each randomized method ten times using the same set of seeds. Each run is allowed up to 6500 iterations for $s = 100$ or 1900 iterations for $s = 400$, and may stop earlier when the full-gradient norm falls below 10^{-8}.

Table 1 summarizes both the remaining objective gap and the time needed to reach a gap of 10^{-8}. Smaller gaps indicate better solutions. The gap column uses a geometric mean, which averages the errors on a logarithmic scale, together with a 95% confidence interval describing uncertainty across random seeds. The "Reached target" column counts successful runs, and its brackets give a confidence interval for the probability of success. The time column reports the median time among successful runs, with brackets showing the middle half of their observed times.

Table 1: Results over ten runs. Gaps are measured at iteration k; success means reaching an objective gap of 10^{-8} within the run budget. Times include successful runs only, and "–" means that no run reached the target.

Method	s	k	Gap [95% CI]	Reached target [95% CI]	Time (s) [middle 50%]
RSRN	100	5000	9.89×10^{-8} [6.24×10^{-8}, 1.57×10^{-7}]	0/10 [0.00, 0.28]	–
Acc-RSRN	100	5000	1.71×10^{-8} [1.02×10^{-8}, 2.87×10^{-8}]	7/10 [0.40, 0.89]	124.4 [111.8, 132.2]
RSRN	400	800	2.26×10^{-8} [1.29×10^{-8}, 3.96×10^{-8}]	10/10 [0.72, 1.00]	91.2 [74.2, 103.0]
Acc-RSRN	400	800	1.81×10^{-9} [7.54×10^{-10}, 4.32×10^{-9}]	10/10 [0.72, 1.00]	58.6 [55.5, 69.1]

The benefit of acceleration is visible in both solution quality and target attainment. For $s = 100$, Acc-RSRN reaches the target in seven runs, whereas RSRN does not reach it in any run. For $s = 400$, both methods reach the target in all ten runs, and acceleration reduces the median time from 91.2 to 58.6 seconds. We also compare the two methods seed by seed, rather than relying only on their separate averages. The resulting 95% confidence intervals for the RSRN-to-Acc-RSRN gap ratio are $[3.38, 9.84]$ for $s = 100$ and $[8.45, 18.45]$ for $s = 400$. A ratio above one favors acceleration; both intervals lie entirely above one, supporting an improvement at the tested checkpoints.

F.2 CNN: OBJECTIVE, PAIRED DESIGN, AND REPORTING

Task and model. We fix 256 MNIST training images and one initialization, using data seed 42 and initialization seed 42 with separate random streams. Images are 28×28 grayscale tensors scaled to $[0, 1]$, without additional standardization. The architecture is a 3×3 convolution with one input and one output channel and a bias, followed by ReLU, 2×2 max pooling, and a biased affine map from 169 features to ten logits. There are $d = 1710$ trainable parameters. We minimize

$$f(x) = \frac{1}{256} \sum_{i=1}^{256} -\log p_x(y_i \mid u_i),$$

without weight decay or other regularization of the objective. Every derivative evaluation uses all 256 images. Hessian-vector products are computed by automatic differentiation.

Implementation details. For each $s \in \{50, 100\}$, RSRN and Acc-RSRN use the same sequence of Gaussian sketches $\mathbf{S} \in \mathbb{R}^{s \times d}$ with independent $\mathbf{S}_{ij} \sim \mathcal{N}(0, 1/s)$. The algorithms follow Algorithm 2 with choice $c_1 = 2$, $\gamma = 1/2$, and the additional ridge floor is zero. We set $c_2 = 1/s$, giving 0.02 and 0.01, respectively. Backtracking starts at $t = 1$ and halves the step until $f(\mathbf{z} + t\mathbf{p}) \le f(\mathbf{z}) + 0.3t\nabla f(\mathbf{z})^\top p$; the limits are 80 backtracks and minimum step 10^{-16}. We further use $q = 0.03$, $a = q/(1+q)$, $b = 1 - q$, $\theta = 3$, and $\sigma = 0.1$.

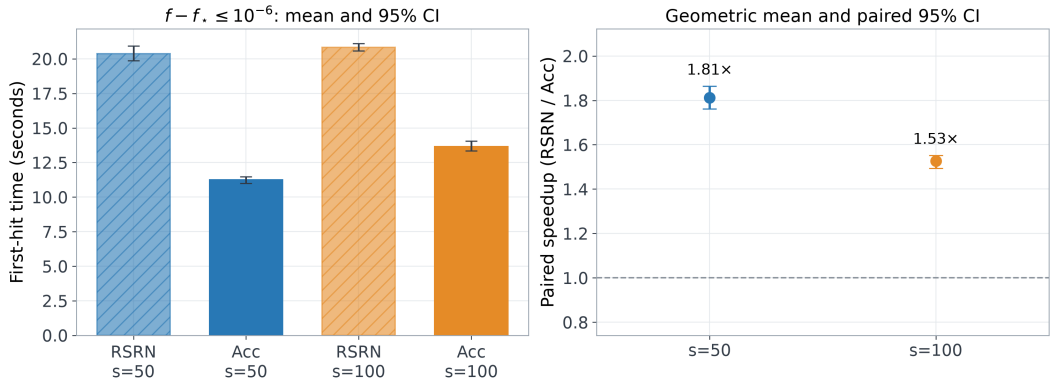

Figure 3: Uncertainty at the primary target $f - f_\star \leq 10^{-6}$. Left: mean first-hit solver time with 95% bootstrap confidence intervals. Right: geometric mean of paired RSRN/Acc-RSRN time ratios with 95% paired-bootstrap confidence intervals; values above one favor acceleration. Each comparison uses ten paired sketch seeds on the same fixed data and initialization.

First-hit statistics. For target ε, let $T_{M,s,j}(\varepsilon)$ be the first recorded solver time at which method M reaches $f - f_\star \leq \varepsilon$ for seed j, within 60 seconds. We report the sample mean, sample standard deviation, and a 95% percentile-bootstrap confidence interval for the mean. The paired acceleration statistic is

$$G_s(\varepsilon) = \exp\left[\frac{1}{10}\sum_{j=1}^{10} \log \frac{T_{\mathrm{RSRN},s,j}(\varepsilon)}{T_{\text{Acc-RSRN},s,j}(\varepsilon)}\right].$$

The statistical configuration was fixed before the ten-seed runs: $\varepsilon = 10^{-6}$ is the primary target and 10^{-8} is an additional target. We use 20,000 bootstrap resamples, resampling seed pairs jointly for G_s.

Table 2: First-hit solver time to $f - f_\star \leq \varepsilon$ on the fixed 256-image CNN instance. Every row attains its target in 10/10 runs within 60 seconds. SD is the sample standard deviation across sketch seeds. The intervals are 95% percentile-bootstrap confidence intervals for the mean, using 20,000 resamples.

Target ε	Method	s	Mean $\pm$ SD (s)	95% CI for mean (s)
10^{-6}	RSRN	50	20.37 ± 0.88	$[19.88, 20.92]$
	Acc-RSRN	50	11.24 ± 0.41	$[10.98, 11.47]$
	RSRN	100	20.83 ± 0.45	$[20.59, 21.11]$
	Acc-RSRN	100	13.66 ± 0.61	$[13.33, 14.05]$
10^{-8}	RSRN	50	43.04 ± 1.53	$[42.15, 43.97]$
	Acc-RSRN	50	15.99 ± 0.62	$[15.62, 16.35]$
	RSRN	100	42.68 ± 0.99	$[42.06, 43.22]$
	Acc-RSRN	100	18.19 ± 0.71	$[17.79, 18.63]$

Table 2 reports both target levels. At 10^{-6}, the paired speedups are 1.811 (95% CI $[1.761, 1.863]$) for $s = 50$ and 1.526 ($[1.494, 1.551]$) for $s = 100$; Figure 3 visualizes these estimates. At 10^{-8}, they increase to 2.692 ($[2.635, 2.755]$) and 2.348 ($[2.294, 2.401]$), respectively. All four randomized groups attain both targets in 10/10 runs. Adam takes 15.12 and 39.63 seconds at the two targets, while AGD takes 25.53 seconds at 10^{-6} and does not attain 10^{-8} within budget. No seed-based confidence intervals are assigned to these single baseline runs. The primary conclusion is improved empirical performance of the accelerated method relative to its matched RSRN control on this instance, with a larger time advantage at the tighter target.

G Additional Related Work

Sketched and subsampled Newton methods. A related line of work approximates the Newton system while retaining a full-dimensional parameter update. Subsampled Newton methods construct

Hessian approximations by sampling components of a finite-sum objective (Erdogdu & Montanari, 2015; Roosta-Khorasani & Mahoney, 2019). Newton Sketch instead sketches an accessible factor of the Hessian (Pilanci & Wainwright, 2017), and Newton-LESS sparsifies this sketch while retaining nearly the same local convergence guarantees as Gaussian sketches (Dereziński et al., 2021). Thus, subsampled Newton methods exploit finite-sum structure, whereas Newton Sketch and Newton-LESS operate with an accessible Hessian factor. Randomized subspace Newton takes a different approach by restricting the update itself to a random subspace of the parameters (Gower et al., 2019; Fuji et al., 2025). In our setting, the corresponding reduced Hessians can be formed using either Hessian-vector products or entry-wise Hessian access.

Subspace cubic-regularization and trust-region methods. Cubic regularization stabilizes the quadratic model with a cubic penalty (Nesterov & Polyak, 2006; Cartis et al., 2011). For convex objectives, Hanzely et al. (2020) show that the local rate of stochastic subspace cubic Newton matches that of a subspace method applied to the quadratic model at $\mathbf{x}_\star$, and hence depends only on $\mathbf{H}_\star$. Proposition E.2 gives the analogous fixed-Hessian first-moment contraction for the nonaccelerated regularized RSRN step. Our spectral results further bound this rate explicitly through $\bar{\kappa}_r$ for Gaussian and ARLS subspaces and accelerate this dependence. For nonconvex objectives, the random-subspace homogenized trust-region method of Higuchi et al. (2025) reaches an ϵ-approximate first-order stationary point in $O(\epsilon^{-3/2})$ iterations and converges locally at a linear rate, while random-subspace cubic regularization retains the global complexity orders of full-dimensional cubic regularization under suitable assumptions (Cartis et al., 2025); see also the general random-subspace framework of Cartis et al. (2022). Our safeguarded GLOBAL analysis instead gives an $O(\epsilon^{-2})$ stationarity-or-entry bound, matching the first-order complexity order of RSRN; its role is to reach a certified LOCAL regime rather than improve the global complexity order.

Spectral-tail rates, acceleration, and restarts. Sketch-and-project methods for linear systems encompass randomized Kaczmarz and randomized coordinate descent (Gower & Richtárik, 2015; Strohmer & Vershynin, 2009). Dereziński & Rebrova (2024) show that their convergence can improve under spectral decay through sharp bounds on the expected sketched projection. Acceleration of randomized projection and block methods has been studied using Nesterov-type momentum (Nesterov, 1983; 2018; Tu et al., 2017; Gower et al., 2018), while regularized block projections with adaptive momentum can exploit outlying singular values (Dereziński et al., 2025). Accelerated coordinate descent (Nesterov & Stich, 2017) and randomized-subspace Nesterov methods (Omiya et al., 2026) provide related first-order mechanisms for subspace acceleration. Our fixed-Hessian reference recursion is an accelerated regularized projection method in the $\mathbf{H}_\star$ metric; the nonlinear analysis transfers this mechanism to regularized Newton updates while accounting for Hessian variation, Armijo line search, and the probability of leaving the local neighborhood. Adaptive momentum restarts are common in accelerated first-order methods (O'Donoghue & Candès, 2015). In the analyzed two-mode algorithm, restarts are confined to the GLOBAL phase; after a successful one-time certificate, the method switches to the unrestarted LOCAL recursion.