ANCHORED RANDOMIZED SUBSPACE NEWTON METHOD WITH LOCAL SUPERLINEAR CONVERGENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Randomized subspace projection techniques offer a scalable approach to Newton-type methods by restricting updates to low-dimensional randomized subspaces. However, unlike classical Newton-type methods, they generally fail to achieve superlinear convergence near a solution. In fact, when the subspace is constructed purely at random, such fast local rates are provably unattainable, even under local strong convexity. We show that this limitation can be overcome by a minimal modification of the subspace construction. Specifically, we augment the randomized subspace with a single *anchor direction* computed via a randomized Kaczmarz-type procedure. This simple addition improves the alignment of the subspace with local curvature. Under local-strongly-convex assumptions, we establish superlinear convergence for randomized subspace variants of quadratic and cubic regularized Newton methods. Our results demonstrate that fast local convergence is recoverable without abandoning the subspace framework. Numerical experiments further show that the proposed method often achieves improved performance despite the additional cost of computing the anchor.

1 INTRODUCTION

Large-scale optimization problems of the form

$$\min_{x\in\mathbb{R}^n} f(x) \tag{1}$$

are central to modern machine learning. Here, $f : \mathbb{R}^n \to \mathbb{R}$ is a C^2 function, which may be nonconvex globally. For such problems, first-order methods have been extensively studied due to their scalability and low per-iteration cost, leading to significant theoretical and practical advances. However, their convergence speed—especially in the vicinity of a solution—is inherently limited.

Second-order methods provide a compelling alternative, as they exploit curvature information and are known to achieve superlinear or even quadratic convergence under suitable conditions. In particular, near a nondegenerate local minimizer x_*, classical Newton-type methods enjoy rapid local convergence driven by the Hessian $\nabla^2 f(x_*)$. Despite this advantage, their applicability to large-scale problems remains limited due to the high computational cost associated with forming and solving systems involving the Hessian.

Prior work on randomized subspace Newton-type methods To bridge this gap, randomized subspace methods have recently emerged as a promising paradigm. These methods reduce computational complexity by performing the k-th updates in a low-dimensional subspace defined by a random matrix $S_k \in \mathbb{R}^{s\times n}$, where $s \ll n$. This line of research originates from the classical Newton method for convex optimization (e.g., Nocedal & Wright (2006)), for which Gower et al. (2019) proposed a randomized subspace variant. This work demonstrated that restricting the Newton step to a randomly selected low-dimensional subspace can significantly reduce computational cost while preserving desirable global convergence properties.

Building on this idea, randomized subspace variants have been extended to nonconvex optimization (see Appendix A). In particular, a randomized subspace version of the regularized Newton method (Ueda & Yamashita, 2014; 2010) was developed in Fuji et al. (2025). Similarly, randomized subspace variants of cubic regularization methods (Nesterov & Polyak, 2006; Cartis et al., 2011a;b)

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.

Table 1: Comparison of randomized-subspace Newton-type methods, with full-dimensional counterparts shown for reference. RS = randomized subspace; Reg. = regularized; TR = trust region; GN = Gauss–Newton; ARC = adaptive regularization using cubics; HTR = homogenized trust region. Local rate: 1 = linear, 1+ = superlinear, 2 = quadratic. The symbol "–" means that a local-rate result for a general objective function is not a main claim. The symbol † indicates that superlinear/near-quadratic convergence is shown only for a low-rank function with a rank-deficient Hessian.

Class	Method	Nonconvex	Subspace	Subproblem	Local
Full	Newton e.g.,(Nocedal & Wright, 2006)		Full space	Newton eq.	2
Full	RNM **(2RN)** (Ueda & Yamashita, 2014)	✓	Full space	Reg. Newton eq.	1+
Full	CR/ARC **(3RN)** (Nesterov & Polyak, 2006; Cartis et al., 2011a;b)	✓	Full space	Cubic model	2
RS	RSN (Gower et al., 2019)		RS	Newton model	1
RS	SSCN-cvx (Hanzely et al., 2020)		RS	Cubic model	1
RS	RSM/R-SGN (Cartis et al., 2020; 2022)	✓	RS	TR / quad. reg. / GN	–
RS	RSRNM **(RS-2RN)** (Fuji et al., 2025)	✓	RS	Reg. Newton	$1^{\dagger}$
RS	SSCN-ncvx **(RS-3RN)** (Zhao et al., 2025)	✓	Coordinate RS	Cubic model	–
RS	RS-ARC (Cartis et al., 2025)	✓	RS	ARC model	–
RS	RSHTR (Higuchi et al., 2025)	✓	RS	HTR model	$1^{\dagger}$
RS	SigmaSVD (Tsipinakis et al., 2026)	✓	RS	SVD approx.	$-^{\dagger}$
RS	Ours: **ARS-2RN / ARS-3RN**	✓	Anchored RS	Anchored Newton model	1+

have been proposed in Hanzely et al. (2020); Zhao et al. (2025); Cartis et al. (2025). A comparison of randomized subspace second-order methods, along with full-dimensional counterparts for reference, is provided in Table 1. These approaches offer favorable trade-offs between per-iteration cost and convergence guarantees. Beyond second-order methods, randomized subspace techniques have also been actively explored in first-order optimization (Kozak et al., 2021; Nozawa et al., 2023; Omiya et al., 2026) and zeroth-order optimization (Kozak et al., 2023; Hare et al., 2025; Cartis & Roberts, 2023; Nozawa et al., 2025; Nesterov & Spokoiny, 2017).

A limitation of randomized subspace Newton-type methods However, a fundamental limitation of randomized subspace second-order methods has been identified: they generally fail to inherit the fast local convergence of classical second-order methods. In particular, Fuji et al. (2025) shows that even when f is locally strongly convex around a strict local minimizer x_*, superlinear convergence cannot be achieved with high probability when the subspace is constructed purely at random.

This phenomenon is not specific to a particular algorithmic design. Similar observations have been reported for other randomized subspace Newton-type methods. For instance, Hanzely et al. (2020) explicitly argues that, although one might expect superlinear convergence, the obtained linear rate is essentially the best possible within the randomized subspace framework. As a result, existing works that establish faster local convergence rates do so only under restrictive structural assumptions on the objective function. In particular, superlinear or near-quadratic convergence has been shown only for functions with rank-deficient structure (Fuji et al., 2025; Higuchi et al., 2025; Tsipinakis et al., 2026).

These results naturally raise the following question:

> *Can randomized subspace Newton-type methods recover local superlinear convergence without restrictive low-rank assumptions?*

Contributions of this work In this paper, we answer this question in the affirmative. Our key idea is to enrich the randomized subspace with a single problem-dependent *anchor direction* that approximately tracks the local Newton direction. The anchor is computed via a randomized Kaczmarz-type procedure. Specifically, instead of using only the conventional random sketch $S_k \in \mathbb{R}^{s\times n}$, we construct the projection matrix $P_k \in \mathbb{R}^{(s+1)\times n}$ by augmenting S_k with the anchor as an additional row. Intuitively, while S_k explores randomized directions in the ambient space, the anchor injects a direction that captures the local Newton geometry, thereby overcoming the limitation of purely randomized subspaces.

Building on this construction, we establish local superlinear convergence under local strong convexity for randomized subspace variants of both quadratic and cubic regularized Newton methods,

Algorithm 1 Anchored Randomized Subspace Regularized Newton (ARS-RN)

Input: $x_{-1} \in \mathbb{R}^n,\ y_{-1} \in \mathbb{R}^n \setminus \{0\},\ s,\ r,\ \{T_k\}_{k\geq 0},\ \beta \in (0, 1/2),\ \tau \in (0, 1)$

1: $x_0 \leftarrow x_{-1}$, any $H_{-1} \in \mathbb{R}^{n\times n}$, any $g_{-1} \in \mathbb{R}^n \setminus \{0\}$ ▷ Initialization

2: **for** $k = 0, 1, 2, \ldots$ **do**

3: Set the initial point as ▷ Initialization for Randomized Kaczmarz method

$$y_{k,0} \leftarrow \left(1 - \frac{g_{k-1}^\top H_{k-1}(x_k - x_{k-1})}{\|g_{k-1}\|^2}\right) y_{k-1} + \frac{1}{\|g_{k-1}\|}(x_k - x_{k-1}) \tag{2}$$

4: Apply T_k steps of the Randomized Kaczmarz method (Algorithm 2) using sketches with r rows to (3) with initial point $y_{k,0}$, and set $y_k \leftarrow y_{k,T_k}$.

5: Sample $S_k \sim \mathcal{N}^{s\times n}(0, 1/s)$ and set $P_k \leftarrow \begin{bmatrix} y_k^\top/\|y_k\| \\ S_k \end{bmatrix} \in \mathbb{R}^{(s+1)\times n}$.

6: Compute u_k by solving the regularized subproblem (5) and $d_k \leftarrow P_k^\top u_k$.

7: $\alpha \leftarrow 1$ ▷ Armijo backtracking procedure

8: **while** $f(x_k + \alpha d_k) > f(x_k) + \beta\alpha\, g_k^\top d_k$ **do**

9: $\alpha \leftarrow \tau\,\alpha$

10: **end while**

11: Update $x_{k+1} \leftarrow x_k + \alpha d_k$

12: **end for**

including those studied in Fuji et al. (2025); Hanzely et al. (2020). Our analysis further shows that only a logarithmically increasing number of inner randomized Kaczmarz iterations is sufficient to achieve this rate, while preserving the global $O(\varepsilon^{-2})$ iteration complexity of existing randomized subspace regularized Newton methods. These results demonstrate that the lack of superlinear convergence is not an inherent limitation of subspace methods per se, but rather a consequence of how the subspace is constructed. Although computing the anchor incurs additional cost, we show both theoretically and empirically that this overhead is modest and leads to overall acceleration.

Notations Let I_s be the identity matrix of size s. For a positive semi-definite matrix H, we denote by $H^{1/2}$ its square root, i.e., $H^{1/2}H^{1/2} = H$. H-norm is defined as $\|x\|_H := \sqrt{x^\top Hx}$. For a symmetric matrix H, we denote by $\lambda_{\min}(H)$ and $\lambda_{\max}(H)$ its minimum and maximum eigenvalues, respectively. For $H_* \succ 0$, we write $\kappa(H_*) := \lambda_{\max}(H_*)/\lambda_{\min}(H_*)$. We denote by $\Pi(A) := A(A^\top A)^\dagger A^\top$ the orthogonal projection onto $\mathrm{range}(A)$ for a matrix A. We denote by $\mathcal{N}(0, 1)$ the standard normal distribution, and by $\mathcal{N}^{s\times n}(0, \sigma^2)$ the distribution of $s \times n$ matrices with i.i.d. $\mathcal{N}(0, \sigma^2)$ entries. We denote by $B(x, r)$ a neighborhood of x with radius r, and write $B(x)$ when only its existence matters. For related quantities α and β, we write $\alpha = O(\beta)$ if $\alpha \leq c\beta$ for some $c > 0$ as $\beta \to 0$, and $\alpha = \Theta(\beta)$ if $c_1\beta \leq \alpha \leq c_2\beta$ for some $c_1, c_2 > 0$ as $\beta \to 0$.

2 Proposed Algorithm

Throughout this paper, for each iterate x_k, we write

$$g_k := \nabla f(x_k), \qquad H_k := \nabla^2 f(x_k).$$

The proposed method, Anchored Randomized Subspace Regularized Newton (ARS-RN) stated in Algorithm 1, consists of the following four stages. The novelty of the method lies in (i) and (ii).

(i) lines 3 and 4: Warm-starting and computing a normalized Newton direction by a randomized Kaczmarz-type method We define the true normalized Newton direction $y_{k,*}$ by

$$H_k y_{k,*} = g_k/\|g_k\|. \tag{3}$$

ARS-RN approximately solves this system by a randomized Kaczmarz-type method (Algorithm 2) using only Hessian-vector products $v \mapsto H_k v$, and sets $y_k := y_{k,T_k}$. Since the target system $H_k y = g_k/\|g_k\|$ changes after every outer update, the initialization (2) acts as a warm start transported from the previous approximation y_{k-1} to the new system.This warm start is designed to facilitate tracking of the moving target $y_{k,*}$, as formalized in the local analysis. For $k = 0$, (2) gives $y_{0,0} = y_{-1}$.

We use this randomized Kaczmarz-type solver because its contraction has the same spectral form as randomized subspace Newton-type methods: both are governed by the smallest eigenvalue of an expected random projection matrix. The solver is based on randomized Kaczmarz methods (Gower & Richtárik, 2015; Rebrova & Needell, 2021); further details are given in Appendix B.

(ii) line 5: Construction of the subspace Fix the subspace dimension $s \in \mathbb{N}$ and sample $S_k \sim \mathcal{N}^{s\times n}(0, 1/s)$. We then define the anchored random matrix

$$P_k := \begin{bmatrix} y_k^\top / \|y_k\| \\ S_k \end{bmatrix} \in \mathbb{R}^{(s+1)\times n}. \tag{4}$$

In other words, the subspace $\mathrm{range}(P_k^\top)$ always contains $\mathrm{span}\{y_k\}$, while the remaining directions are supplied by the randomized subspace $\mathrm{range}(S_k^\top)$.

(iii) line 6: Computing the regularized subspace Newton direction The update direction d_k is obtained by solving a low-dimensional subproblem according to the chosen regularization design. Specifically, we consider the subproblem with regularization term $\mathcal{R}_k(\cdot)$,

$$u_k \in \mathrm{Argmin}_{u\in\mathbb{R}^{s+1}} \left\{ g_k^\top P_k^\top u + \tfrac{1}{2} u^\top P_k H_k P_k^\top u + \mathcal{R}_k(u) \right\} \tag{5}$$

and define d_k using its solution $u_k \in \mathbb{R}^{s+1}$ by $d_k := P_k^\top u_k$. In Section E, we present concrete choices of the regularizer $\mathcal{R}_k$: a quadratic regularizer following Fuji et al. (2025) and a reduced-coordinate cubic regularizer motivated by the cubic regularization used in Hanzely et al. (2020). The quadratic subproblem follows the existing regularized Newton form, while the cubic variant is adapted to the reduced-coordinate formulation of ARS-RN. Both subproblems fit the common regularized Newton form used below. For both regularization designs, the corresponding optimality condition can be written using a scalar value $\eta_k \geq 0$ as $(P_k H_k P_k^\top + \eta_k I_{s+1})u_k = -P_k g_k$, which yields the subproblem solution u_k of (5) and the search direction

$$d_k = -P_k^\top (P_k H_k P_k^\top + \eta_k I_{s+1})^{-1} P_k g_k. \tag{6}$$

For the analysis, we use the following condition on η_k.

Condition 2.1. There exist constants $\bar{\eta}, c_\eta, C_\eta > 0$ and $\delta \in (0, 1]$ such that

(a) $c_\eta \|d_k\| \leq \eta_k \leq \bar{\eta}$ holds for every iteration k.
(b) $\eta_k \leq C_\eta \|g_k\|^\delta$ holds for all sufficiently large k.

In this paper, we consider the quadratic and cubic regularizers described in Section E, which are shown there to satisfy the above condition.

(iv) line 7-line 9: Armijo line search and update We impose the Armijo condition

$$f(x_k + \alpha_k d_k) \leq f(x_k) + \beta\, \alpha_k\, g_k^\top d_k, \qquad \beta \in (0, 1/2)$$

and choose $\alpha_k = \tau^{m_k}$ for $\tau \in (0, 1)$, where m_k is the smallest nonnegative integer satisfying the condition. The iterate is updated by $x_{k+1} = x_k + \alpha_k d_k$.

HVP-based computational cost. We first compare the Hessian-related work of randomized subspace Newton-type methods and ARS-RN in terms of Hessian-vector products (HVPs), since both methods can construct their reduced Hessian matrices without explicitly forming H_k. Let ψ_{HVP} denote the cost of one HVP $v \mapsto H_k v$.

In randomized subspace Newton-type methods, $S_k H_k S_k^\top$ can be formed using s HVPs. In ARS-RN, the final reduced Hessian $P_k H_k P_k^\top$ can be formed using $s+1$ HVPs, while the T_k randomized Kaczmarz blocks use rT_k HVPs to form $R_{k,t} H_k R_{k,t}^\top$, $t = 0, \ldots, T_k - 1$. Thus, under this direct HVP accounting, the per-outer-iteration costs of randomized subspace Newton-type methods and ARS-RN are respectively $s\,\psi_{\mathrm{HVP}}$ and $(s + 1 + rT_k)\,\psi_{\mathrm{HVP}}$. Theorem 3.2 shows that a logarithmically increasing number of inner randomized Kaczmarz iterations is sufficient for ARS-RN to achieve local superlinear convergence, in contrast to the linear local convergence of existing purely randomized subspace methods.

This HVP count isolates second-order oracle work, but it does not capture all computational reuse. In ARS-RN, the inner Kaczmarz blocks and the final anchored subproblem are evaluated at the same base point x_k, so point-dependent derivative information can be reused across these computations; successive RSRNM steps, in contrast, update the base point after each subspace step. Appendix F gives a rate-aware total-work comparison, covering both explicitly assembled and matrix-free implementations, and identifies conditions under which this reuse yields a lower overall cost even when the HVP count itself is not smaller.

3 Theoretical Convergence Analysis

3.1 Global iteration complexity

We first state the global iteration complexity of the proposed method. For the global analysis, we impose only boundedness of the level set and Lipschitz continuity of the Hessian on a slightly enlarged level set. In particular, we do not assume convexity or strong convexity, and hence the result applies to nonconvex objectives.

Assumption 3.1. For the initial point x_0, let

$$\Omega := \{x \in \mathbb{R}^n \mid f(x) \le f(x_0)\}, \qquad \Omega_R := \Omega + B(0, \bar{\eta}/c_\eta)\,,$$

where $c_\eta, \bar{\eta}$ are the constants in Condition 2.1(a).

(a) Ω is bounded.
(b) There exists $L_H > 0$ such that $\|\nabla^2 f(x_1) - \nabla^2 f(x_2)\| \le L_H \|x_1 - x_2\|$ for all $x_1, x_2 \in \Omega_R$.

Let $U_H := \sup_{x\in\Omega_R} \|\nabla^2 f(x)\|$, $\alpha_{\min} := \tau \min\left\{1, \frac{s c_\eta}{\bar{\mathcal{C}} n L_H}\right\}$, where $\bar{\mathcal{C}}$ is the constant in Lemma C.4.

Although the proposed subspace P_k contains the anchor vector y_k, the estimates needed for the global analysis are obtained from the Gaussian sketch S_k. Indeed,

$$\|P_k g_k\|^2 = \left(y_k^\top g_k / \|y_k\|\right)^2 + \|S_k g_k\|^2 \ge \|S_k g_k\|^2, \quad \|P_k\|^2 \le 1 + \|S_k\|^2.$$

Therefore, the required lower bound on the subspace gradient and the operator-norm upper bound follow from standard concentration bounds for Gaussian sketches; these bounds for the anchored random matrix P_k are stated in Appendix C. Consequently, in the global analysis, we do not claim an essential improvement due to the anchor. Following the argument of Fuji et al. (2025), with the necessary adaptations to our setting, we obtain the following result. The complete proof is provided in Appendix D.1.

Theorem 3.1. *Suppose that Assumption 3.1 and Condition 2.1(a) hold. For any $\varepsilon > 0$ and fixed $\epsilon \in (0,1)$, if $K \ge \frac{\bar{\mathcal{C}}\frac{n}{s} U_H + \bar{\eta}}{\beta\, \alpha_{\min}(1-\epsilon)} \left(f(x_0) - \inf_{x\in\Omega} f(x)\right) \varepsilon^{-2}$, then*

$$\min_{0\le k\le K-1} \|\nabla f(x_k)\| \le \varepsilon$$

holds with probability at least $1 - 2K\exp(-\mathcal{C}_0 \epsilon^2 s) - 2K\exp(-s)$.

The anchor becomes useful in the local regime: y_k approximates the normalized Newton direction, and this anchored subspace structure yields the sharper local convergence behavior analyzed next.

3.2 Local convergence analysis

We next analyze the local convergence behavior of the proposed method. All statements below are understood to hold for sufficiently large k.

Assumption 3.2. (a) The generated sequence satisfies $x_k \to x_*$ a.s., where x_* is a stationary point with $g_* := \nabla f(x_*) = 0$ and $H_* := \nabla^2 f(x_*) \succ 0$.
(b) There exists $L_H > 0$ such that $\|\nabla^2 f(x) - \nabla^2 f(y)\| \le L_H \|x - y\|$ for all $x, y \in B(x_*)$.

Under Assumption 3.2, the gradient admits the local expansion

$$g_k = H_*(x_k - x_*) + O(\|x_k - x_*\|_{H_*}^2), \qquad \|g_k\| = \Theta(\|x_k - x_*\|_{H_*}).$$

Thus, the convergence rate of $\|g_k\|$ is equivalent to that of $\|x_k - x_*\|_{H_*}$ up to constant factors.

We use the following filtrations in the local analysis. For $t = 0, 1, \ldots, T_k$, define

$$\mathcal{F}_{k,t} \coloneqq \sigma\Big(S_{0:k-1},\ R_{j,\,0:T_j-1}\,(0 \le j < k),\ R_{k,\,0:t-1}\Big), \tag{7}$$

where $R_{k,0:t-1}$ is understood as the empty collection when $t = 0$. We then define

$$\mathcal{F}_k \coloneqq \mathcal{F}_{k,T_k}. \tag{8}$$

Thus, $\mathcal{F}_k$ represents the information available after the inner randomized Kaczmarz loop produces $y_k = y_{k,T_k}$ and before the new Gaussian block S_k is sampled. In particular, x_k and y_k are $\mathcal{F}_k$-measurable, while S_k is independent of $\mathcal{F}_k$.

We first recover the standard local linear convergence estimate of randomized subspace Newton-type methods. The contraction rate is described by the smallest eigenvalue of the expected random projection matrix $\mathbb{E}[\Pi(H_*^{1/2}P_k^\top) \mid \mathcal{F}_k]$.

Lemma 3.1. *Under Assumption 3.2, Condition 2.1(b), and $n > s + 2$, for all sufficiently large k, with $\lambda_k^P \coloneqq \lambda_{\min}\Big(\mathbb{E}[\Pi(H_*^{1/2}P_k^\top) \mid \mathcal{F}_k]\Big)$, we have*

$$\mathbb{E}\left[\|x_{k+1} - x_*\|_{H_*} \mid \mathcal{F}_k\right] \le \sqrt{1 - \lambda_k^P}\,\|x_k - x_*\|_{H_*} + O\big(\|x_k - x_*\|_{H_*}^{1+\delta}\big)\,. \tag{9}$$

Lemma 3.1 shows that the proposed method preserves the standard local linear convergence behavior of randomized subspace Newton-type methods. The additional effect of the proposed method appears only when we use the fact that the subspace P_k contains not only the random sketch S_k, but also the anchor y_k, which approximates the Newton direction. The following proposition makes this effect explicit: the local rate of $\|g_k\|$, equivalently that of $\|x_k - x_*\|_{H_*}$, depends on the approximation accuracy of y_k.

Proposition 3.1. *Suppose that Assumption 3.2, Condition 2.1(b), and $n > s + 2$ hold. Then, denoting $\lambda_k^P = \lambda_{\min}\Big(\mathbb{E}\Big[\Pi(H_*^{1/2}P_k^\top) \mid \mathcal{F}_k\Big]\Big)$, for all sufficiently large k,*

$$\mathbb{E}\left[\|g_{k+1}\| \mid \mathcal{F}_k\right] \le \left(\sqrt{\lambda_{\max}(H_*)(1 - \lambda_k^P)} + O(\|g_k\|)\right)\|g_k\|\|y_k - y_{k,*}\|_{H_*} + O(\|g_k\|^{1+\delta}) \tag{10}$$

Proof sketch. The key point is that the anchored random matrix contains y_k. Locally, $H_*^{-1/2}g_{k+1}$ is governed by the projected Newton residual, or equivalently,

$$g_{k+1} \simeq H_*^{1/2}\left(I - \Pi(H_*^{1/2}P_k^\top)\right)H_*^{-1/2}g_k.$$

By the definition of $y_{k,*}$, we have $H_*^{-1/2}g_k \simeq \|g_k\|H_*^{1/2}y_{k,*}$. Also, $y_k \in \mathrm{range}(P_k^\top)$ implies $H_*^{1/2}y_k \in \mathrm{range}(H_*^{1/2}P_k^\top)$, so the projector removes the $H_*^{1/2}y_k$-component. Hence $g_{k+1} \simeq \|g_k\|H_*^{1/2}\left(I - \Pi(H_*^{1/2}P_k^\top)\right)H_*^{1/2}y_{k,*} = \|g_k\|H_*^{1/2}\left(I - \Pi(H_*^{1/2}P_k^\top)\right)H_*^{1/2}(y_{k,*} - y_k)$. Therefore, the usual projection residual of order $\|g_k\|$ is multiplied by the anchor error $\|y_k - y_{k,*}\|_{H_*}$, which gives (10). The detailed decomposition is given in the appendix proof of Proposition 3.1. □

Proposition 3.1 is the key estimate in the local analysis. In a standard randomized subspace Newton method, the leading error remains of order $\|g_k\|$ through the projection residual. In contrast, because the proposed subspace contains the anchor y_k, the leading term is reduced to $\|g_k\|\|y_k - y_{k,*}\|_{H_*}$. Therefore, if y_k accurately approximates $y_{k,*}$, the next gradient norm becomes smaller than the usual linear-rate prediction.

Thus, the remaining task is to prove that the randomized Kaczmarz-generated anchor y_k tracks the moving target $y_{k,*} = H_k^{-1}g_k/\|g_k\|$. Unlike standard randomized Kaczmarz, the system $H_k y = g_k/\|g_k\|$ changes after every outer iteration. We use an inner contraction estimate for fixed k (Lemma 3.2; see the appendix proof of Lemma 3.2) and a warm-start estimate for the transfer from the k-th system to the $(k+1)$-st system (Lemma 3.3; see the appendix proof of Lemma 3.3).

First, the following lemma gives the one-step contraction of the inner randomized Kaczmarz iteration. Let $\lambda_R \coloneqq \lambda_{\min}\Big(\mathbb{E}\Big[\Pi(H_*^{1/2}R^\top)\Big]\Big)$.

Lemma 3.2. *Under Assumption 3.2, for all sufficiently large k,*

$$\mathbb{E}\left[\|y_{k,t+1} - y_{k,*}\|_{H_*} \mid \mathcal{F}_{k,t}\right] \le \left(\sqrt{1-\lambda_R} + O(\|x_k - x_*\|_{H_*})\right) \|y_{k,t} - y_{k,*}\|_{H_*}.$$

Next, we control the error introduced by the warm start at the next outer iteration. This term appears because the target vector changes from $y_{k,*}$ to $y_{k+1,*}$.

Lemma 3.3. *Under Assumption 3.2, Condition 2.1(b) and $n > s+2$, for all sufficiently large k,*

$$\mathbb{E}[\|y_{k+1,0} - y_{k+1,*}\|_{H_*} \mid \mathcal{F}_k] \le$$
$$\left(1 + \sqrt{\kappa(H_*)} + O(\|x_k - x_*\|_{H_*}^{\delta})\right) \|y_k - y_{k,*}\|_{H_*} + \frac{2 + \sqrt{\kappa(H_*)}}{\sqrt{\lambda_{\min}(H_*)}} + O(\|x_k - x_*\|_{H_*}).$$

Combining these two estimates over the inner iterations, together with a logarithmically increasing number of inner iterations T_k, yields asymptotic tracking of the moving target by the anchor; see the appendix proof of Proposition 3.2.

Proposition 3.2. *Suppose that Assumption 3.2, Condition 2.1(b), and $n > s+2$ hold. Suppose that $\{T_k\}$ is a sequence of positive integers satisfying, for all sufficiently large k, $T_k \ge \left\lceil \frac{4}{-\log(1-\lambda_R)} \log(k+1) \right\rceil$. Then $\|y_k - y_{k,*}\|_{H_*} \longrightarrow 0$ $(k \to \infty)$ a.s.*

Proposition 3.2 shows that a logarithmically increasing number of inner randomized Kaczmarz iterations is sufficient for the anchor y_k to track the moving target $y_{k,*}$ almost surely. The inner iterations contract the anchor error at a rate governed locally by λ_R, while the growth of T_k makes this contraction sufficiently strong to suppress the propagated anchor error and the warm-start error. Consequently, $\|y_k - y_{k,*}\|_{H_*} \to 0$ almost surely.

We now combine the local linear bound for x_k (Lemma 3.1), the anchor-dependent gradient decomposition (Proposition 3.1), and the anchor-tracking result (Proposition 3.2).

Theorem 3.2. *Suppose that Assumption 3.2 and Condition 2.1(b) hold, and let $n > s+2$. Furthermore, assume that $\{T_k\}$ satisfies the condition in Proposition 3.2. Then*

$$\frac{\|x_{k+1} - x_*\|_{H_*}}{\|x_k - x_*\|_{H_*}} \longrightarrow 0 \quad \text{a.s.}. \tag{11}$$

Consequently, the sequence $\{x_k\}$ generated by Algorithm 1 converges locally superlinearly to the stationary point x_*.

Proof sketch. By Proposition 3.1, after dividing by $\|g_k\|$, for all sufficiently large k,

$$\mathbb{E}\left[\|g_{k+1}\|/\|g_k\| \mid \mathcal{F}_k\right] \le C_1 \|y_k - y_{k,*}\|_{H_*} + C_2 \|x_k - x_*\|_{H_*}^{\delta}$$

for some constants $C_1, C_2 > 0$. Thus, it suffices to show that the two terms on the right-hand side are summable almost surely.

By Lemma 3.1, once x_k is in a sufficiently small neighborhood of x_*, its conditional expected error contracts by a uniform factor $\rho_x < 1$. Since $x_k \to x_*$ almost surely by Assumption 3.2, a stopping-time localization argument therefore yields $\sum_{k=0}^{\infty} \|x_k - x_*\|_{H_*}^{\delta} < \infty$ a.s. Similarly, under the condition in Proposition 3.2, the contraction accumulated over the T_k inner randomized Kaczmarz iterations is summable in k. Combining this contraction with the warm-start estimate and applying the same localization argument gives $\sum_{k=0}^{\infty} \|y_k - y_{k,*}\|_{H_*} < \infty$ a.s. Hence, $\sum_{k=0}^{\infty} \mathbb{E}\left[\|g_{k+1}\|/\|g_k\| \mid \mathcal{F}_k\right] < \infty$ a.s. Conditional Markov's inequality and the conditional Borel–Cantelli lemma then imply $\|g_{k+1}\|/\|g_k\| \longrightarrow 0$ a.s. Finally, the local equivalence $\|g_k\| = \Theta(\|x_k - x_*\|_{H_*})$ transfers this result to (11). The stopping-time localization, summability arguments, and conditional Borel–Cantelli step are given in the appendix proof of Theorem 3.2. □

Sharp bounds for the projection constants. The local rate involves $\lambda_{\min}(\mathbb{E}[\Pi(H_*^{1/2} P_k^{\top}) \mid \mathcal{F}_k])$ and $\lambda_{\min}(\mathbb{E}[\Pi(H_*^{1/2} R^{\top})])$. Since P_k contains the Gaussian block S_k,

$$\lambda_{\min}\left(\mathbb{E}[\Pi(H_*^{1/2} P_k^{\top}) \mid \mathcal{F}_k]\right) \ge \lambda_{\min}\left(\mathbb{E}[\Pi(H_*^{1/2} S_k^{\top})]\right).$$

Smallest-eigenvalue estimates for expected sketched projection matrices are standard in sketch-and-project analyses (Gower & Richtárik, 2015; Dereziński & Rebrova, 2024). Sharp Gaussian projection estimates give, up to the explicit multiplicative factors stated in Appendix D.3,

$$\lambda_{\min}\Big(\mathbb{E}[\Pi(H_*^{1/2}S_k^\top)]\Big) \gtrsim \frac{s\lambda_{\min}(H_*)}{\mathrm{Err}(H_*^{1/2}, s-1)}, \qquad \lambda_{\min}\Big(\mathbb{E}[\Pi(H_*^{1/2}R^\top)]\Big) \gtrsim \frac{r\lambda_{\min}(H_*)}{\mathrm{Err}(H_*^{1/2}, r-1)}.$$

Here r is the row dimension of R, while $\mathrm{Err}(H_*^{1/2}, \cdot)$ denotes the randomized SVD residual error defined in (83) of Appendix D.3; the two inequalities are stated precisely in (84) and (85).

The important point is that these lower bounds are expressed through randomized SVD residual errors, rather than only through a worst-case condition-number term. Existing results show that such residual errors become small when the underlying matrix is well captured by a low-dimensional spectral subspace, for example under spectral decay or flat-tailed spectra (Derezinski et al., 2020; Dereziński & Rebrova, 2024; Halko et al., 2011). Consequently, in such regimes, the two eigenvalue terms in the local rate can yield meaningful contraction estimates. Appendix D.3 gives the precise inequalities and compares them with the more pessimistic trace- and condition-number-based estimates.

4 Numerical Experiments

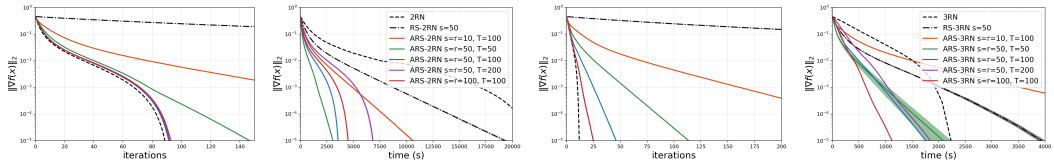
We evaluate whether the anchored subspace construction improves practical convergence behavior. We consider two settings. First, on a synthetic strongly convex logistic regression problem, we test whether the method behaves consistently with the local theory. Second, on real-world convex empirical risk minimization problems, we test whether the anchor remains useful beyond the idealized strongly convex setting. Experimental details, including hardware, preprocessing, timing protocol, objective definitions, and parameter choices, are given in Appendix G.

In synthetic experiments, we compare full-space methods, randomized subspace variants, and their anchored counterparts. In the real-world experiments, we focus on randomized subspace methods and evaluate ARS-3RN against purely randomized baselines. For quadratic regularization, we denote the full-space, randomized subspace, and anchored variants by 2RN, RS-2RN, and ARS-2RN, respectively, and refer to them collectively as the 2RN family. Analogously, for cubic regularization, we use 3RN, RS-3RN, and ARS-3RN, and refer to them as the 3RN family. See Table 1 for their characteristics. All methods are initialized at $x_0 = 0$ and are terminated when $\|\nabla f(x_k)\|_2 \le 10^{-5}$ or when the iteration limit is reached. For randomized methods, curves show the mean over independent random seeds, with shaded regions indicating the minimum and maximum values.

Synthetic strongly convex logistic regression. Figure 1 reports results on a dense strongly convex logistic regression problem. This experiment examines the finite-iteration behavior associated with the mechanism predicted by the local theory: improving the anchor accuracy should make the anchored subspace more informative. The iteration-wise plots show that larger T, corresponding to more inner anchor updates, leads to a faster local decrease. In particular, sufficiently accurate anchors substantially accelerate the local convergence of ARS-2RN. Since T is fixed within each run, these experiments do not directly test the asymptotic superlinear-convergence result, which requires an increasing inner iteration budget T_k. The time plots show that the improvement in local convergence can offset the additional anchor cost for suitable choices of s, r, and T.

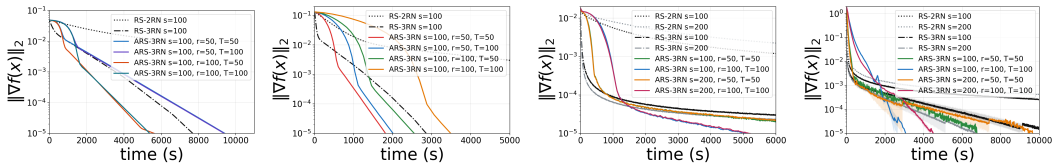
Regression problems with real-world datasets downloadable from Chang & Lin (2011). We next evaluate the anchored construction on real-world convex empirical risk minimization problems: multi-label logistic regression, binary logistic regression, and softmax regression. The multi-label problems are regularized and are closer to the locally strongly convex regime considered in the theory, whereas the epsilon and USPS experiments are unregularized and therefore are not strongly convex in general.

Figure 2 compares ARS-3RN with purely randomized subspace baselines. On the regularized multi-label problems, sufficiently informative anchors lead to a faster decrease over the reported optimization horizon. On the unregularized epsilon and USPS problems, which lie outside the strong-convexity setting of the local convergence theory, ARS-3RN also exhibits improved finite-iteration convergence behavior for suitable choices of s, r, and T. Since T is fixed within each run, these



(a) 2RN-family (Iteration) (b) 2RN-family (Time) (c) 3RN-family (Iteration) (d) 3RN-family (Time)

Figure 1: Synthetic strongly convex logistic regression ($\dim = 5000, \lambda = 10^{-3}$). Panels (a)–(b) compare the 2RN family, and panels (c)–(d) compare the 3RN family. The vertical axis shows $\|\nabla f(x_k)\|_2$ on a logarithmic scale. For randomized methods, solid curves report the mean over five random seeds, and shaded bands indicate the minimum–maximum range across these seeds. Iteration-wise plots show the effect of anchor accuracy, while time plots show the tradeoff between anchor cost and faster local decrease. The choices of s, r, and T are shown in the legends of (b) and (d), shared by panels (a)–(b) and (c)–(d), respectively; additional details are given in Appendix G.



(a) flickr_deepwalk (dim: 25,155) (b) youtube_node2vec (dim: 6,063) (c) epsilon (dim: 2,001) (d) USPS (dim: 2,570)

Figure 2: Real-world convex empirical risk minimization problems. See Appendix G.3 for dataset details. We compare randomized subspace Newton-type methods based on purely randomized subspaces with the proposed ARS-3RN. Panels (a) and (b) report independent multi-label logistic regression on flickr_deepwalk ($\lambda = 10^{-3}$) and youtube_node2vec ($\lambda = 10^{-3}$), respectively. Panel (c) reports binary logistic regression on epsilon ($\lambda = 0$), and panel (d) reports softmax regression on USPS ($\lambda = 0$). The vertical axis shows $\|\nabla f(x_k)\|_2$ on a logarithmic scale, and the horizontal axis shows cumulative computation time. For randomized methods, curves show the mean over three independent random seeds, and shaded regions indicate the minimum and maximum values across these seeds. The choices of s, r, and T are shown in the legends; T is fixed throughout each run.

experiments do not test the asymptotic superlinear convergence result; rather, they evaluate whether the anchored construction can provide practical benefits under a fixed computational budget. Overall, the results indicate that anchoring can improve practical convergence on these convex empirical risk minimization problems, although the benefit depends on the balance between anchor accuracy and its additional computational cost.

5 Conclusion

We proposed an anchored subspace design for randomized subspace Newton-type methods and introduced ARS-RN, achieving superlinear convergence under local strong convexity. While purely randomized subspaces are known to yield at most linear convergence, we showed that incorporating a low-cost approximation of the normalized Newton direction restores fast local convergence. Our results suggest a design principle for randomized subspace methods: augmenting randomized subspaces with problem-dependent directions to overcome the limitations of purely random constructions.

A limitation of the proposed method is the additional cost of computing the anchor direction. Extending the approach to sub-sampled Newton (Roosta-Khorasani & Mahoney, 2019) and mini-batch settings for computational efficiency is an important direction for future work. In such settings, it remains to understand how the interaction between sampling noise and subspace randomness affects the global complexity and local convergence properties.

REFERENCES

Robert Anderson, Julian Andrej, Andrew Barker, Jamie Bramwell, Jean-Sylvain Camier, Jakub Cerveny, Veselin Dobrev, Yohann Dudouit, Aaron Fisher, Tzanio Kolev, et al. Mfem: A modular finite element methods library. *Computers & Mathematics with Applications*, 81:42–74, 2021.

Julian Andrej, Nabil Atallah, Jan-Phillip Bäcker, Jean-Sylvain Camier, Dylan Copeland, Veselin Dobrev, Yohann Dudouit, Tobias Duswald, Brendan Keith, Dohyun Kim, et al. High-performance finite elements with mfem. *The International Journal of High Performance Computing Applications*, 38(5):447–467, 2024.

Coralia Cartis and Lindon Roberts. Scalable subspace methods for derivative-free nonlinear least-squares optimization. *Mathematical Programming*, 199(1–2):461–524, 2023.

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, 2011a. doi: 10.1007/s10107-009-0286-5.

Coralia Cartis, Nicholas I. M. Gould, and Philippe L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. part ii: Worst-case function- and derivative-evaluation complexity. *Mathematical Programming*, 130(2):295–319, 2011b. doi: 10.1007/s10107-009-0337-y.

Coralia Cartis, Jaroslav Fowkes, and Zhen Shao. A randomised subspace Gauss–Newton method for nonlinear least-squares. *in Workshop on "Beyond First Order Methods in ML Systems" at the 37 th International Conference on Machine Learning*, 2020. URL https://arxiv.org/abs/2211.05727.

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

Coralia Cartis, Zhen Shao, and Edward Tansley. Random subspace cubic-regularization methods, with applications to low-rank functions. *arXiv: 2501.09734*, 2025. URL https://arxiv.org/abs/2501.09734.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, 2011. doi: 10.1145/1961189.1961199.

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.

Michal Derezinski, Feynman T Liang, Zhenyu Liao, and Michael W Mahoney. Precise expressions for random projections: Low-rank approximation and randomized newton. *Advances in Neural Information Processing Systems*, 33:18272–18283, 2020.

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*, 6:1–35, 2025.

Robert Gower, Dmitry Kovalev, Felix Lieder, and Peter Richtárik. RSN: randomized subspace Newton. *Advances in Neural Information Processing Systems*, 32, 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.

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 855–864, 2016. doi: 10.1145/2939672.2939754.

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

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

Warren Hare, Lindon Roberts, and Clément W. Royer. Expected decrease for derivative-free algorithms using random subspaces. *Mathematics of Computation*, 94(351):277–304, 2025.

Rei Higuchi, Pierre-Louis Poirion, and Akiko Takeda. Improving convergence guarantees of random subspace second-order algorithm for nonconvex optimization. *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/pdf?id=tuu4de7HL1.

Grant Hillier and Raymond Kan. Properties of the inverse of a noncentral wishart matrix. *Econometric Theory*, 38(6):1092–1116, 2022.

Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994. doi: 10.1109/34.291440.

David Kozak, Stephen Becker, Alireza Doostan, and Luis Tenorio. A stochastic subspace approach to gradient-free optimization in high dimensions. *Computational Optimization and Applications*, 79(2):339–368, 2021.

David Kozak, Cesare Molinari, Lorenzo Rosasco, Luis Tenorio, and Silvia Villa. Zeroth-order optimization with orthogonal random directions. *Mathematical Programming*, 199(1–2):1179–1219, 2023.

Boris S. Mityagin. The zero set of a real analytic function. *Mathematical Notes*, 107:529–530, 2020. doi: 10.1134/S0001434620030189.

Deanna Needell and Joel A. Tropp. Paved with good intentions: Analysis of a randomized block kaczmarz method. *Linear Algebra and its Applications*, 441:199–221, 2014. doi: 10.1016/j.laa.2012.12.022.

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 Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, 2 edition, 2006. doi: 10.1007/978-0-387-40065-5.

Ryota Nozawa, Pierre-Louis Poirion, and Akiko Takeda. Randomized subspace gradient method for constrained optimization, 2023. arXiv:2307.03335.

Ryota Nozawa, Pierre-Louis Poirion, and Akiko Takeda. Zeroth-order random subspace algorithm for non-smooth convex optimization. *Journal of Optimization Theory and Applications*, 204(53), 2025.

Gaku Omiya, Pierre-Louis Poirion, and Akiko Takeda. Convergence analysis of randomized subspace normalized SGD under heavy-tailed noise, 2026. arXiv:2601.20399.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. DeepWalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 701–710, 2014. doi: 10.1145/2623330.2623732.

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

Elizaveta Rebrova and Deanna Needell. On block gaussian sketching for the kaczmarz method. *Numerical Algorithms*, 86(1):443–473, 2021.

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

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

Nick Tsipinakis, Panagiotis Tigas, and Panos Parpas. A multilevel low-rank newton method with super-linear convergence rate and its application to non-convex problems. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=PKakPzVVja.

Kenji Ueda and Nobuo Yamashita. Convergence properties of the regularized newton method for the unconstrained nonconvex optimization. *Applied Mathematics and Optimization*, 62(1):27–46, 2010.

Kenji Ueda and Nobuo Yamashita. A regularized Newton method without line search for unconstrained optimization. *Computational Optimization and Applications*, 59(1):321–351, 2014.

Roman Vershynin. High-dimensional probability. *University of California, Irvine*, 10(11):31, 2020.

Umberto Villa, Noemi Petra, and Omar Ghattas. hippylib: An extensible software framework for large-scale inverse problems governed by pdes: Part i: Deterministic inversion and linearized bayesian inference. *ACM Transactions on Mathematical Software (TOMS)*, 47(2):1–34, 2021.

David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1–2):1–157, 2014. doi: 10.1561/0400000060.

Jim Zhao, Aurelien Lucchi, and Nikita Doikov. Cubic regularized subspace Newton for non-convex optimization. *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025.

A Prior work on Subspace Newton methods

Prior work on Subspace Newton methods and their significance The classical Newton method solves the Newton equation at an iterate x_k,

$$\nabla^2 f(x_k) d_k = -\nabla f(x_k)$$

and updates $x_{k+1} = x_k + d_k$. However, when n is large, forming and storing the Hessian and solving the resulting linear system become bottlenecks. In contrast, Newton-type methods based on randomized subspaces restrict the search to the subspace $\mathrm{range}(S_k^\top)$ spanned by a random matrix $S_k \in \mathbb{R}^{s\times n}$ with $s \ll n$, and use the reduced Hessian

$$S_k \nabla^2 f(x_k) S_k^\top \in \mathbb{R}^{s\times s}$$

to solve a low-dimensional subproblem. By reducing the dimension of the linear system solved at each iteration from n to s, it can substantially reduce both computational and memory costs, especially when $s \ll n$. This dimension-reduction viewpoint is closely related to the broader randomized numerical linear algebra literature, where sketching is used to compress large linear algebra problems so that expensive computations can be performed on smaller projected matrices; see, e.g., Woodruff (2014).

Related but distinct are sketched Newton methods such as Newton Sketch (Pilanci & Wainwright, 2017). Newton Sketch reduces the cost of a Newton-type step by replacing the Hessian, or a Hessian factor, with a randomly projected or subsampled version, while still computing an approximate full-space Newton direction. In contrast, the randomized subspace Newton methods discussed here restrict the search direction itself to $\mathrm{range}(S_k^\top)$ and solve a reduced s-dimensional subproblem.

Moreover, RSN has been shown to achieve global linear convergence under assumptions such as strong convexity (Gower et al., 2019).

Regularization in Subspace Newton methods When (1) is globally nonconvex, $\nabla^2 f(x_k)$ may be indefinite, and hence the reduced Hessian $S_k \nabla^2 f(x_k) S_k^\top$ need not be positive definite. Thus, the RSN direction is not guaranteed to be a descent direction in general. Regularization is therefore essential in the nonconvex setting to ensure descent and global convergence. Fuji et al. proposed the Randomized Subspace Regularized Newton Method (RSRNM), which applies quadratic regularization so that the Hessian on the subspace becomes positive definite,

$$d_k^{\mathrm{RSRNM}} \;=\; -S_k^\top \Big(S_k \nabla^2 f(x_k) S_k^\top + \eta_k I_s \Big)^{-1} S_k \nabla f(x_k), \qquad x_{k+1} = x_k + \alpha_k d_k^{\mathrm{RSRNM}}$$

and updates the iterate accordingly. Here, $\eta_k > 0$ is chosen so that descent holds, and α_k is determined by an Armijo line search. They establish global convergence, with an $O(\varepsilon^{-2})$ iteration complexity, as well as local linear convergence (Fuji et al., 2025). As another line of regularized Newton methods, subspace cubic Newton-type methods, which perform cubic regularization on a subspace, have also been proposed. For high-dimensional convex minimization, Hanzely et al. proposed Stochastic Subspace Cubic Newton (SSCN), which minimizes a cubic regularization model on a randomly chosen subspace (Hanzely et al., 2020). Furthermore, Zhao et al. extended the SSCN framework to nonconvex functions, providing global guarantees for arbitrary subspace sizes in terms of stationarity complexity, as well as guarantees that recover rates consistent with classical cubic regularization, namely $O(\varepsilon^{-3/2})$, as the subspace dimension increases (Zhao et al., 2025).

B Randomized Kaczmarz

The Kaczmarz method is a classical projection method for solving a linear system

$$Az = b,$$

where $A \in \mathbb{R}^{m \times n}$, $z \in \mathbb{R}^n$, and $b \in \mathbb{R}^m$. Given the current iterate z_t, the classical Kaczmarz update projects z_t onto the solution hyperplane of one row equation $a_i^\top z = b_i$:

$$z_{t+1} = z_t - \frac{a_i^\top z_t - b_i}{\|a_i\|^2} a_i.$$

Throughout this section, we use randomized Kaczmarz in a broad sense to refer to randomized Kaczmarz-type projection methods, including single-row, block, and Gaussian-sketch variants. The classical single-row randomized Kaczmarz method was shown by Strohmer & Vershynin (2009) to converge exponentially in expectation for consistent overdetermined linear systems. A modern and useful viewpoint is the sketch-and-project framework of Gower & Richtárik (2015), in which one samples a sketching matrix S_t and computes

$$z_{t+1} = \operatorname*{Argmin}_{z \in \mathbb{R}^n} \|z - z_t\|_B^2 \quad \text{subject to} \quad S_t A z = S_t b.$$

This framework contains randomized Kaczmarz-type methods as special cases, as well as randomized coordinate descent, randomized Newton-type methods, and related randomized iterative methods. For example, row-sampling choices of S_t recover single-row and block randomized Kaczmarz updates, whereas Gaussian choices of S_t lead to Gaussian sketch-and-project variants; block variants of randomized Kaczmarz were studied by Needell & Tropp (2014), and block Gaussian variants were further analyzed by Rebrova & Needell (2021).

We use this block Gaussian randomized Kaczmarz-type iteration as an inner anchor solver. This choice is motivated by its spectral compatibility with randomized subspace Newton-type methods. Both the progress of randomized subspace Newton updates and the worst-case contraction of randomized Kaczmarz-type methods are governed by the smallest eigenvalue of an expected randomized subspace projection matrix. Thus, a subspace distribution that is favorable for randomized subspace Newton-type methods in this spectral sense also yields fast contraction for the randomized Kaczmarz-based approximation of the normalized Newton direction. The anchor exploits this compatibility by transporting the previous approximate normalized Newton direction to the next time-varying linear system, while the randomized Kaczmarz iterations reduce the resulting tracking error. This viewpoint is consistent with recent sharp spectral estimates for expected sketched projection matrices in sketch-and-project methods (Dereziński & Rebrova, 2024).

Algorithm 2 Randomized Kaczmarz method

Input: $y_{k,0} \in \mathbb{R}^n,\ T_k \in \mathbb{N},\ r \in \mathbb{N}$
Output: y_{k,T_k}

1: **for** $t = 0, 1, \ldots, T_k - 1$ **do**
2: Sample $R_{k,t} \in \mathbb{R}^{r\times n}$ by $R_{k,t} \sim \mathcal{N}^{r\times n}(0, 1/r)$
3: Solve the linear system

$$R_{k,t} H_k R_{k,t}^\top u_{k,t} = R_{k,t} \left(H_k y_{k,t} - g_k/\|g_k\| \right). \tag{13}$$

4: Update

$$y_{k,t+1} \leftarrow y_{k,t} - R_{k,t}^\top u_{k,t}. \tag{14}$$

5: **end for**

At the outer iteration k, the target anchor direction is defined through the normalized Newton system

$$H_k y_{k,*} = \frac{g_k}{\|g_k\|}. \tag{12}$$

The normalization in (12) is used only when the algorithm has not terminated. Indeed, if $\|g_k\| = 0$, then x_k is already a stationary point and the outer algorithm terminates before calling the inner anchor solver. Therefore, we consider only iterations for which

$$\|g_k\| > 0.$$

In particular, $g_k/\|g_k\|$ is well-defined.

The inner randomized Kaczmarz loop computes an approximation y_{k,T_k} of $y_{k,*}$. For a block Gaussian sketch $R_{k,t} \in \mathbb{R}^{r\times n}$, we write the update in the following H_k-metric sketch-and-project form:

$$y_{k,t+1} = \text{Argmin}_{y\in\mathbb{R}^n} \|y - y_{k,t}\|_{H_k}^2 \quad \text{subject to} \quad R_{k,t} H_k y = R_{k,t} \frac{g_k}{\|g_k\|}.$$

This variational form should be understood through its closed-form implementation in Algorithm 2. In particular, the actual update requires solving the linear system (13), or equivalently applying the inverse of $R_{k,t} H_k R_{k,t}^\top$. Thus, for the update to be well-defined, it is sufficient that

$$R_{k,t} H_k R_{k,t}^\top$$

is invertible almost surely. We therefore impose the following rank condition.

Assumption B.1. At each iteration k, for the number of rows r used in Algorithm 2,

$$r \le \text{rank}(H_k)$$

holds.

Under Assumption B.1, the Gaussian compression $R_{k,t} H_k R_{k,t}^\top$ is invertible almost surely. This follows from the rotational invariance of Gaussian matrices (Vershynin, 2020) and the standard measure-zero argument for nontrivial analytic determinants (Mityagin, 2020). Consequently,

$$(R_{k,t} H_k R_{k,t}^\top)^{-1}$$

exists almost surely, and hence the linear system in (13) and the closed-form update (14) are well-defined almost surely.

C Properties of Gaussian random matrices

In this section, we recall two standard high-probability estimates for scaled Gaussian random matrices. These estimates are used in the global analysis. We then show that analogous estimates hold for the anchored random matrix used in our method,

$$P_k = \begin{bmatrix} y_k^\top/\|y_k\| \\ S_k \end{bmatrix},$$

where $S_k \in \mathbb{R}^{s\times n}$ is a scaled Gaussian matrix. The key observation is that the Gaussian part S_k alone already provides the required estimates, and the additional anchor row does not destroy these bounds.

We first recall a standard Johnson–Lindenstrauss type estimate for scaled Gaussian matrices.

Lemma C.1 (Gaussian Johnson–Lindenstrauss estimate; (Vershynin, 2020, Lemma 5.3.2 and Exercise 5.3.3))**.** *Let $S \in \mathbb{R}^{s\times n}$ be a Gaussian matrix whose entries are independent and distributed as $\mathcal{N}(0, 1/s)$. Then there exists a universal constant $\mathcal{C}_0 > 0$ such that, for any $z \in \mathbb{R}^n$ and any $\epsilon \in (0, 1)$,*

$$\mathbb{P}(1-\epsilon)\|z\|^2 \le \|Sz\|^2 \le (1+\epsilon)\|z\|^2 \ge 1 - 2\exp(-\mathcal{C}_0\epsilon^2 s).$$

We also use the following standard operator-norm bound for scaled Gaussian matrices.

Lemma C.2 (Gaussian operator-norm bound; (Vershynin, 2020, Theorem 4.6.1), (Fuji et al., 2025, Lemma 2.2))**.** *Let $S \in \mathbb{R}^{s\times n}$ be a Gaussian matrix whose entries are independent and distributed as $\mathcal{N}(0, 1/s)$. Then there exists a constant $\bar{C} > 0$ such that*

$$\|S\|^2 = \|SS^\top\| = \|S^\top S\| \le \bar{C}\,\frac{n}{s}$$

holds with probability at least $1 - 2\exp(-s)$.

The next two lemmas transfer these estimates to the anchored random matrix P_k.

Lemma C.3. *For every iteration k and every $\epsilon \in (0, 1)$,*

$$\|P_k g_k\|^2 \ge (1-\epsilon)\|g_k\|^2$$

holds with probability at least

$$1 - 2\exp(-\mathcal{C}_0\epsilon^2 s).$$

Proof. By the definition (4) of the anchored random matrix,

$$\begin{aligned}\|P_k g_k\|^2 &= g_k^\top P_k^\top P_k g_k \\ &= g_k^\top \left(\frac{y_k y_k^\top}{\|y_k\|^2} + S_k^\top S_k\right) g_k \\ &= \left(\frac{y_k^\top g_k}{\|y_k\|}\right)^2 + \|S_k g_k\|^2 \\ &\ge \|S_k g_k\|^2. \end{aligned} \tag{15}$$

Applying Lemma C.1 with $S = S_k$ and $z = g_k$, we have

$$\|S_k g_k\|^2 \ge (1-\epsilon)\|g_k\|^2$$

with probability at least

$$1 - 2\exp(-\mathcal{C}_0\epsilon^2 s).$$

Combining this estimate with (15) gives the claim. □

Lemma C.4. *For every iteration k,*

$$\|P_k^\top P_k\| = \|P_k P_k^\top\| = \|P_k\|^2 \le \bar{\mathcal{C}}\frac{n}{s}$$

holds with probability at least

$$1 - 2\exp(-s),$$

where $\bar{\mathcal{C}} := \bar{C} + 1$.

Proof. By the definition (4) of the anchored random matrix,

$$P_k^\top P_k = \frac{y_k y_k^\top}{\|y_k\|^2} + S_k^\top S_k, \qquad \left\| \frac{y_k y_k^\top}{\|y_k\|^2} \right\| = 1.$$

Therefore,

$$\begin{aligned} \|P_k\|^2 = \|P_k^\top P_k\| &\le \left\| \frac{y_k y_k^\top}{\|y_k\|^2} \right\| + \|S_k^\top S_k\| \\ &= 1 + \|S_k\|^2. \end{aligned}$$

By Lemma C.2,

$$\|S_k\|^2 \le \bar{C}\frac{n}{s}$$

holds with probability at least $1 - 2\exp(-s)$. Hence, on this event,

$$\|P_k\|^2 \le 1 + \bar{C}\frac{n}{s} \le (\bar{C}+1)\frac{n}{s},$$

where the last inequality uses $s \le n$. Letting $\bar{\mathcal{C}} \coloneqq \bar{C} + 1$, we obtain

$$\|P_k\|^2 \le \bar{\mathcal{C}}\frac{n}{s}$$

with probability at least $1 - 2\exp(-s)$. Since

$$\|P_k^\top P_k\| = \|P_k P_k^\top\| = \|P_k\|^2,$$

the proof is complete. □

D Proofs for Theoretical Convergence Analysis

D.1 Proof of the Global Iteration Complexity Result

In this subsection, we prove the global iteration complexity result stated in the main text. Throughout the proof, we work under Assumption 3.1 and Condition 2.1(a) on the regularization coefficient η_k.

Recall that

$$\Omega \coloneqq \{x \in \mathbb{R}^n \mid f(x) \le f(x_0)\}.$$

By Assumption 3.1(a), Ω is bounded, and so is Ω_R. Thus, by the continuity of $f \in C^2$, we can define

$$U_g \coloneqq \sup_{x\in\Omega} \|\nabla f(x)\| < \infty, \qquad U_H \coloneqq \sup_{x\in\Omega_R} \|\nabla^2 f(x)\| < \infty.$$

In particular, for every iteration k,

$$\|g_k\| = \|\nabla f(x_k)\| \le U_g, \qquad \|H_k\| = \|\nabla^2 f(x_k)\| \le U_H$$

hold.

Hereafter, for each iteration k and each $\epsilon \in (0,1)$, define

$$\mathcal{E}_k^{P,\mathrm{JL}}(\epsilon) \coloneqq \left\{ \|P_k g_k\|^2 \ge (1-\epsilon)\|g_k\|^2 \right\}, \qquad \mathcal{E}_k^{P,\mathrm{OP}} \coloneqq \left\{ \|P_k\|^2 \le \bar{\mathcal{C}}\frac{n}{s} \right\}.$$

By Lemmas C.3 and C.4,

$$\mathbb{P}\mathcal{E}_k^{P,\mathrm{JL}}(\epsilon) \ge 1 - 2\exp(-\mathcal{C}_0\epsilon^2 s), \qquad \mathbb{P}\mathcal{E}_k^{P,\mathrm{OP}} \ge 1 - 2\exp(-s)$$

hold. The following estimates are deterministic on these events. We first use the operator-norm event $\mathcal{E}_k^{P,\mathrm{OP}}$ to show that the Armijo line search returns a sufficiently large step.

Lemma D.1. *Suppose that Assumption 3.1 and Condition 2.1(a) hold. Let $\beta \in (0, 1/2)$ and $\tau \in (0, 1)$. For every iteration k,*

$$\alpha_k \geq \alpha_{\min}, \qquad \alpha_{\min} \coloneqq \tau \min\left\{1,\ \frac{sc_\eta}{\bar{\mathcal{C}}\, n\, L_H}\right\}$$

holds with probability at least

$$1 - 2\exp(-s).$$

Proof. We prove the claim on the event $\mathcal{E}_k^{P,\mathrm{OP}}$. Let

$$\mathcal{H}_k \coloneqq P_k H_k P_k^\top + \eta_k I_{s+1},$$

and define

$$d_k \coloneqq -P_k^\top \mathcal{H}_k^{-1} P_k g_k.$$

By Taylor's theorem, there exists $\theta \in (0, 1)$ such that

$$f(x_k + \alpha d_k) = f(x_k) + \alpha g_k^\top d_k + \frac{1}{2}\alpha^2\, d_k^\top \nabla^2 f(x_k + \theta\alpha d_k)\, d_k$$

holds. Therefore,

$$\begin{aligned}
&f(x_k) - f(x_k + \alpha d_k) + \beta\alpha g_k^\top d_k \\
&= (1-\beta)\alpha(-g_k^\top d_k) - \frac{1}{2}\alpha^2\, d_k^\top \nabla^2 f(x_k + \theta\alpha d_k)\, d_k \\
&\overset{(A)}{=} (1-\beta)\alpha\,(P_k g_k)^\top \mathcal{H}_k^{-1}(P_k g_k) \\
&\qquad - \frac{1}{2}\alpha^2\,(P_k g_k)^\top \mathcal{H}_k^{-1} P_k \nabla^2 f(x_k + \theta\alpha d_k) P_k^\top \mathcal{H}_k^{-1}(P_k g_k) \\
&\overset{(B)}{\geq} \frac{1}{2}\alpha^2\,(P_k g_k)^\top \mathcal{H}_k^{-1}(P_k g_k) \\
&\qquad - \frac{1}{2}\alpha^2\,(P_k g_k)^\top \mathcal{H}_k^{-1} P_k \nabla^2 f(x_k + \theta\alpha d_k) P_k^\top \mathcal{H}_k^{-1}(P_k g_k) \\
&= \frac{1}{2}\alpha^2 \underbrace{(P_k g_k)^\top \left(\mathcal{H}_k^{-1} - \mathcal{H}_k^{-1} P_k H_k P_k^\top \mathcal{H}_k^{-1}\right)(P_k g_k)}_{\text{(I)}} \\
&\qquad - \frac{1}{2}\alpha^2 \underbrace{(P_k g_k)^\top \mathcal{H}_k^{-1} P_k \left(\nabla^2 f(x_k + \theta\alpha d_k) - H_k\right) P_k^\top \mathcal{H}_k^{-1}(P_k g_k)}_{\text{(II)}}.
\end{aligned} \tag{16}$$

(A) follows by substituting $d_k = -P_k^\top \mathcal{H}_k^{-1} P_k g_k$, and (B) follows from $\beta < 1/2$ and $\alpha \leq 1$, which imply

$$(1-\beta)\alpha \geq \frac{1}{2}\alpha \geq \frac{1}{2}\alpha^2.$$

We now bound (I) and (II), respectively.

First, for (I), since

$$P_k H_k P_k^\top = \mathcal{H}_k - \eta_k I_{s+1},$$

we have

$$\begin{aligned}
&\mathcal{H}_k^{-1} - \mathcal{H}_k^{-1} P_k H_k P_k^\top \mathcal{H}_k^{-1} \\
&= \mathcal{H}_k^{-1} - \mathcal{H}_k^{-1}(\mathcal{H}_k - \eta_k I_{s+1})\mathcal{H}_k^{-1} \\
&= \eta_k \mathcal{H}_k^{-2}.
\end{aligned}$$

Hence,

$$\text{(I)} = \eta_k \|\mathcal{H}_k^{-1} P_k g_k\|^2.$$

Moreover, Condition 2.1(a) gives $\eta_k \geq c_\eta \|d_k\|$, and therefore

$$\text{(I)} \geq c_\eta \|d_k\| \, \|\mathcal{H}_k^{-1} P_k g_k\|^2.$$

Next, for (II), Condition 2.1(a) gives

$$\|d_k\| \leq \frac{\eta_k}{c_\eta} \leq \frac{\bar{\eta}}{c_\eta}.$$

Thus, for $\alpha \in (0, 1]$ and $\theta \in (0, 1)$,

$$\|x_k + \theta\alpha d_k - x_k\| \leq \theta\alpha \|d_k\| \leq \frac{\bar{\eta}}{c_\eta}.$$

Since $x_k \in \Omega$, we also have $x_k + \theta\alpha d_k \in \Omega_R$. Hence, by Assumption 3.1(b),

$$\|\nabla^2 f(x_k + \theta\alpha d_k) - H_k\| \leq L_H \|\theta\alpha d_k\| \leq L_H \, \alpha \|d_k\|.$$

Moreover, on $\mathcal{E}_k^{P,\mathrm{OP}}$,

$$\|P_k\|^2 \leq \bar{\mathcal{C}} \frac{n}{s}$$

holds, and therefore

$$\begin{aligned} \text{(II)} &\leq \|P_k\|^2 \, \|\nabla^2 f(x_k + \theta\alpha d_k) - H_k\| \, \|\mathcal{H}_k^{-1} P_k g_k\|^2 \\ &\leq \bar{\mathcal{C}} \frac{n}{s} \, L_H \, \alpha \|d_k\| \, \|\mathcal{H}_k^{-1} P_k g_k\|^2. \end{aligned}$$

Substituting these estimates into (16), we obtain, on $\mathcal{E}_k^{P,\mathrm{OP}}$,

$$f(x_k) - f(x_k + \alpha d_k) + \beta\alpha g_k^\top d_k \geq \frac{1}{2}\alpha^2 \Big(c_\eta - \bar{\mathcal{C}} \frac{n}{s} L_H \, \alpha \Big) \, \|d_k\| \, \|\mathcal{H}_k^{-1} P_k g_k\|^2.$$

Therefore, if

$$0 < \alpha \leq \min \left\{ 1, \; \frac{s c_\eta}{\bar{\mathcal{C}} \, n \, L_H} \right\},$$

then

$$f(x_k + \alpha d_k) \leq f(x_k) + \beta\alpha g_k^\top d_k$$

holds on $\mathcal{E}_k^{P,\mathrm{OP}}$.

The Armijo backtracking procedure starts from $\alpha = 1$, decreases α by a factor of τ, and accepts the first α satisfying the Armijo condition. Hence, on $\mathcal{E}_k^{P,\mathrm{OP}}$,

$$\alpha_k \geq \tau \min \left\{ 1, \; \frac{s c_\eta}{\bar{\mathcal{C}} \, n \, L_H} \right\} = \alpha_{\min}$$

follows.

Finally, since

$$\mathbb{P}\mathcal{E}_k^{P,\mathrm{OP}} \geq 1 - 2\exp(-s),$$

the claimed probability estimate follows. □

We have now obtained the step-size estimate needed for one iteration. The next lemma combines the lower bound on $\|P_k g_k\|$ on $\mathcal{E}_k^{P,\mathrm{JL}}(\epsilon)$ and the lower bound on the step size on $\mathcal{E}_k^{P,\mathrm{OP}}$ to quantify the objective decrease per iteration.

Lemma D.2. *Suppose that Assumption 3.1 and Condition 2.1(a) hold. Fix $\epsilon \in (0,1)$. Then*

$$f(x_k) - f(x_{k+1}) \geq \frac{\beta\, \alpha_{\min}(1-\epsilon)}{\bar{\mathcal{C}}\frac{n}{s}U_H + \bar{\eta}} \|g_k\|^2$$

holds with probability at least

$$1 - 2\exp(-\mathcal{C}_0\epsilon^2 s) - 2\exp(-s).$$

Proof. We prove the claim on the event

$$\mathcal{E}_k^{P,\mathrm{JL}}(\epsilon) \cap \mathcal{E}_k^{P,\mathrm{OP}}.$$

The decrease after one iteration satisfies

$$\begin{aligned}
f(x_k) - f(x_{k+1}) &\overset{(A)}{\geq} \beta\alpha_k(-g_k^\top d_k) \\
&= \beta\alpha_k(P_k g_k)^\top (P_k H_k P_k^\top + \eta_k I_{s+1})^{-1}(P_k g_k) \\
&\overset{(B)}{\geq} \beta\alpha_{\min}(P_k g_k)^\top (P_k H_k P_k^\top + \eta_k I_{s+1})^{-1}(P_k g_k) \\
&\overset{(C)}{\geq} \beta\alpha_{\min}\frac{1}{\lambda_{\max}(P_k H_k P_k^\top + \eta_k I_{s+1})}\|P_k g_k\|^2 \\
&\overset{(D)}{\geq} \frac{\beta\,\alpha_{\min}}{\bar{\mathcal{C}}\frac{n}{s}U_H + \bar{\eta}}\|P_k g_k\|^2 \\
&\overset{(E)}{\geq} \frac{\beta\,\alpha_{\min}(1-\epsilon)}{\bar{\mathcal{C}}\frac{n}{s}U_H + \bar{\eta}}\|g_k\|^2. \qquad (17)
\end{aligned}$$

(A) follows from the Armijo condition

$$f(x_{k+1}) \leq f(x_k) + \beta\alpha_k g_k^\top d_k.$$

(B) follows from Lemma D.1, since on $\mathcal{E}_k^{P,\mathrm{OP}}$,

$$\alpha_k \geq \alpha_{\min}$$

holds. (C) follows from

$$(P_k g_k)^\top (P_k H_k P_k^\top + \eta_k I_{s+1})^{-1}(P_k g_k) \geq \lambda_{\max}(P_k H_k P_k^\top + \eta_k I_{s+1})^{-1}\|P_k g_k\|^2.$$

(D) follows because, on $\mathcal{E}_k^{P,\mathrm{OP}}$,

$$\lambda_{\max}(P_k H_k P_k^\top + \eta_k I_{s+1}) \leq \|P_k H_k P_k^\top\| + \eta_k \leq \|P_k\|^2\|H_k\| + \eta_k \leq \bar{\mathcal{C}}\frac{n}{s}U_H + \bar{\eta}$$

holds. (E) follows from the definition of $\mathcal{E}_k^{P,\mathrm{JL}}(\epsilon)$.

Thus, (17) holds on

$$\mathcal{E}_k^{P,\mathrm{JL}}(\epsilon) \cap \mathcal{E}_k^{P,\mathrm{OP}}.$$

Hence,

$$\mathbb{P}f(x_k) - f(x_{k+1}) \geq \frac{\beta\,\alpha_{\min}(1-\epsilon)}{\bar{\mathcal{C}}\frac{n}{s}U_H + \bar{\eta}}\|g_k\|^2 \geq \mathbb{P}\mathcal{E}_k^{P,\mathrm{JL}}(\epsilon) \cap \mathcal{E}_k^{P,\mathrm{OP}}.$$

Furthermore, by the union bound,

$$\begin{aligned}
\mathbb{P}\mathcal{E}_k^{P,\mathrm{JL}}(\epsilon) \cap \mathcal{E}_k^{P,\mathrm{OP}} &= 1 - \mathbb{P}\Big(\mathcal{E}_k^{P,\mathrm{JL}}(\epsilon) \cap \mathcal{E}_k^{P,\mathrm{OP}}\Big)^c \\
&\geq 1 - \mathbb{P}\Big(\mathcal{E}_k^{P,\mathrm{JL}}(\epsilon)\Big)^c - \mathbb{P}\Big(\mathcal{E}_k^{P,\mathrm{OP}}\Big)^c \\
&\geq 1 - 2\exp(-\mathcal{C}_0\epsilon^2 s) - 2\exp(-s).
\end{aligned}$$

□

We have thus obtained the per-iteration decrease estimate needed for the global iteration complexity result (Theorem 3.1). It remains to sum this estimate over $k = 0, \ldots, K-1$ and to combine the probability bounds.

Proof of Theorem 3.1

Proof. Define

$$f_{\inf} \coloneqq \inf_{x\in\Omega} f(x)$$

and

$$\mathcal{E}_K(\epsilon) \coloneqq \bigcap_{k=0}^{K-1} \left(\mathcal{E}_k^{P,\mathrm{JL}}(\epsilon) \cap \mathcal{E}_k^{P,\mathrm{OP}}\right).$$

On $\mathcal{E}_K(\epsilon)$, applying Lemma D.2 to each $k = 0, \dots, K-1$ gives

$$f(x_k) - f(x_{k+1}) \geq \frac{\beta\, \alpha_{\min}(1-\epsilon)}{\bar{\mathcal{C}}\frac{n}{s} U_H + \bar{\eta}} \|g_k\|^2. \tag{18}$$

Summing (18) over $k = 0, \dots, K-1$, and using $f(x_K) \geq f_{\inf}$ since $x_k \in \Omega$, we obtain

$$\begin{aligned} f(x_0) - f_{\inf} &\geq f(x_0) - f(x_K) \\ &= \sum_{k=0}^{K-1} \big(f(x_k) - f(x_{k+1})\big) \\ &\geq \frac{\beta\, \alpha_{\min}(1-\epsilon)}{\bar{\mathcal{C}}\frac{n}{s} U_H + \bar{\eta}} \sum_{k=0}^{K-1} \|g_k\|^2. \end{aligned}$$

Therefore,

$$\sum_{k=0}^{K-1} \|g_k\|^2 \leq \frac{\bar{\mathcal{C}}\frac{n}{s} U_H + \bar{\eta}}{\beta\, \alpha_{\min}(1-\epsilon)} \big(f(x_0) - f_{\inf}\big),$$

and hence

$$\min_{0\leq k\leq K-1} \|g_k\|^2 \leq \frac{1}{K} \sum_{k=0}^{K-1} \|g_k\|^2 \leq \frac{\bar{\mathcal{C}}\frac{n}{s} U_H + \bar{\eta}}{\beta\, \alpha_{\min}(1-\epsilon)} \big(f(x_0) - f_{\inf}\big) \frac{1}{K}.$$

Choosing K so that the right-hand side is at most ε^2,

$$\min_{0\leq k\leq K-1} \|\nabla f(x_k)\| = \min_{0\leq k\leq K-1} \|g_k\| \leq \varepsilon$$

follows.

It remains to establish the probability estimate. By the union bound,

$$\begin{aligned} \mathbb{P}\mathcal{E}_K(\epsilon)^c &= \mathbb{P} \bigcup_{k=0}^{K-1} \left(\mathcal{E}_k^{P,\mathrm{JL}}(\epsilon) \cap \mathcal{E}_k^{P,\mathrm{OP}}\right)^c \\ &\leq \sum_{k=0}^{K-1} \mathbb{P}\Big(\mathcal{E}_k^{P,\mathrm{JL}}(\epsilon)\Big)^c + \sum_{k=0}^{K-1} \mathbb{P}\Big(\mathcal{E}_k^{P,\mathrm{OP}}\Big)^c. \end{aligned}$$

By Lemmas C.3 and C.4,

$$\mathbb{P}\Big(\mathcal{E}_k^{P,\mathrm{JL}}(\epsilon)\Big)^c \leq 2\exp(-\mathcal{C}_0\epsilon^2 s), \qquad \mathbb{P}\Big(\mathcal{E}_k^{P,\mathrm{OP}}\Big)^c \leq 2\exp(-s).$$

Therefore,

$$\mathbb{P}\mathcal{E}_K(\epsilon)^c \leq 2K\exp(-\mathcal{C}_0\epsilon^2 s) + 2K\exp(-s).$$

Hence,

$$\mathbb{P}\mathcal{E}_K(\epsilon) \geq 1 - 2K\exp(-\mathcal{C}_0\epsilon^2 s) - 2K\exp(-s).$$

□

D.2 Proof of the Local Superlinear Convergence Result

In this subsection, we prove the local superlinear convergence result stated in the main text.

By the continuity of $\nabla^2 f$, if $B(x_*)$ is chosen sufficiently small, then there exist constants $\mu, L > 0$ such that

$$\mu I_n \preceq \nabla^2 f(x) \preceq L I_n, \qquad \forall x \in B(x_*) \tag{19}$$

holds. Moreover,

$$\sqrt{\lambda_{\min}(H_*)}\,\|v\| \le \|v\|_{H_*} \le \sqrt{\lambda_{\max}(H_*)}\,\|v\| \tag{20}$$

holds.

We first record two basic local estimates. They show that g_k is approximated to first order by $H_*(x_k - x_*)$, and that $\|g_k\|$ and $\|x_k - x_*\|_{H_*}$ are of the same order. These estimates will be used repeatedly to rewrite the remainder terms below in terms of $\|x_k - x_*\|_{H_*}$ or $\|g_k\|$.

Lemma D.3. *Suppose that Assumption 3.2 holds and that k is sufficiently large so that $x_k \in B(x_*)$. Let*

$$r_k \coloneqq g_k - H_*(x_k - x_*).$$

Then

$$\|r_k\| = O(\|x_k - x_*\|_{H_*}^2) \tag{21}$$

$$\|g_k\| = \Theta(\|x_k - x_*\|_{H_*}) \tag{22}$$

hold.

Proof. Since $x_k \in B(x_*)$ and $B(x_*)$ is convex, for every $\theta \in [0,1]$,

$$x_* + \theta(x_k - x_*) \in B(x_*)$$

holds. Therefore, Assumption 3.2(b) and (19) can be applied on the line segment $\{x_* + \theta(x_k - x_*) : \theta \in [0,1]\}$.

We first prove (21). Using the mean-value representation of $g_k = \nabla f(x_k)$,

$$\begin{aligned} r_k &= \nabla f(x_k) - H_*(x_k - x_*) \\ &= \nabla f(x_k) - \nabla f(x_*) - H_*(x_k - x_*) \\ &= \int_0^1 \nabla^2 f(x_* + \theta(x_k - x_*))(x_k - x_*)\, d\theta - H_*(x_k - x_*) \\ &= \int_0^1 \big(\nabla^2 f(x_* + \theta(x_k - x_*)) - H_*\big)(x_k - x_*)\, d\theta. \end{aligned}$$

Hence, by the triangle inequality,

$$\begin{aligned} \|r_k\| &\le \int_0^1 \big\|\nabla^2 f(x_* + \theta(x_k - x_*)) - H_*\big\|\, \|x_k - x_*\|\, d\theta \\ &\overset{\text{(A)}}{\le} \int_0^1 L_H\, \theta \|x_k - x_*\|\, \|x_k - x_*\|\, d\theta \\ &= \frac{L_H}{2}\|x_k - x_*\|^2 \\ &\overset{\text{(B)}}{\le} \frac{L_H}{2\,\lambda_{\min}(H_*)}\|x_k - x_*\|_{H_*}^2. \end{aligned}$$

Here, (A) follows from Assumption 3.2(b). Moreover, (B) follows from (20). Thus (21) holds.

Next, we prove (22). Noticing

$$\|g_k\| = \left\| \int_0^1 \nabla^2 f(x_* + \theta(x_k - x_*))(x_k - x_*)\, d\theta \right\|$$

and

$$\mu I_n \preceq \nabla^2 f(x_* + \theta(x_k - x_*)) \preceq L I_n \qquad (\theta \in [0,1])$$

from (19), we have

$$\mu \|x_k - x_*\| \le \|g_k\| \le L \|x_k - x_*\|.$$

Moreover, by (20),

$$\frac{\mu}{\sqrt{\lambda_{\max}(H_*)}} \|x_k - x_*\|_{H_*} \le \|g_k\| \le \frac{L}{\sqrt{\lambda_{\min}(H_*)}} \|x_k - x_*\|_{H_*}.$$

Therefore (22) holds. □

Next, we show that locally the line search eventually always accepts $\alpha_k = 1$. This is needed in order to handle the subsequent updates in a concise Newton-type form.

Lemma D.4. *Suppose that Assumption 3.2 holds and that the step size α_k is chosen by the Armijo line search*

$$f(x_k + \alpha_k d_k) \le f(x_k) + \beta \alpha_k g_k^\top d_k, \qquad 0 < \beta < \tfrac{1}{2}, \quad 0 < \tau < 1. \tag{23}$$

Then, for all sufficiently large k,

$$\alpha_k = 1 \tag{24}$$

holds with probability 1.

Proof. By the definition of d_k,

$$d_k = -P_k^\top (P_k H_k P_k^\top + \eta_k I_{s+1})^{-1} P_k g_k$$

we have

$$\begin{aligned} -g_k^\top d_k &= g_k^\top P_k^\top \left(P_k H_k P_k^\top + \eta_k I_{s+1}\right)^{-1} P_k g_k \\ &= g_k^\top P_k^\top \left(P_k H_k P_k^\top + \eta_k I_{s+1}\right)^{-1} \left(P_k H_k P_k^\top + \eta_k I_{s+1}\right) \left(P_k H_k P_k^\top + \eta_k I_{s+1}\right)^{-1} P_k g_k \\ &= d_k^\top H_k d_k + \eta_k \left\| \left(P_k H_k P_k^\top + \eta_k I_{s+1}\right)^{-1} P_k g_k \right\|^2 \\ &\ge d_k^\top H_k d_k. \end{aligned} \tag{25}$$

Using Taylor's expansion, there exists $\theta \in (0,1)$ such that

$$\begin{aligned} f(x_k + d_k) &= f(x_k) + g_k^\top d_k + \frac{1}{2} d_k^\top \nabla^2 f(x_k + \theta d_k) d_k \\ &= f(x_k) + g_k^\top d_k + \frac{1}{2} d_k^\top H_k d_k + \frac{1}{2} d_k^\top (\nabla^2 f(x_k + \theta d_k) - H_k) d_k \\ &\le f(x_k) + g_k^\top d_k + \frac{1}{2} d_k^\top H_k d_k + \frac{L_H}{2} \theta \|d_k\|^3 \qquad \text{(by Assumption 3.2(b))} \\ &\le f(x_k) + \frac{1}{2} g_k^\top d_k + \frac{L_H}{2} \|d_k\|^3 \quad \text{(by (25) and } 0 \le \theta \le 1) \\ &= f(x_k) + \beta g_k^\top d_k + \underbrace{\left(\tfrac{1}{2} - \beta\right) g_k^\top d_k + \frac{L_H}{2} \|d_k\|^3}_{\text{(I)}}. \end{aligned}$$

Thus, to prove the Armijo condition (23) with $\alpha = 1$, it suffices to show (I)≤ 0.

By (25) and (19),

$$-g_k^\top d_k \ge d_k^\top H_k d_k \ge \mu \|d_k\|^2$$

and hence

$$(I) \le \left(-\left(\tfrac{1}{2} - \beta\right) \mu + \frac{L_H}{2} \|d_k\| \right) \|d_k\|^2.$$

Therefore, if $\|d_k\| \le \frac{(1-2\beta)\mu}{L_H}$, then (I)$\le 0$ follows.

Finally, we show that $\|d_k\| \to 0$. Since $H_k \succeq \mu I_n$ by (19),

$$\|d_k\| = \left\|P_k^\top (P_k H_k P_k^\top + \eta_k I_{s+1})^{-1} P_k g_k\right\| \le \left\|P_k^\top (\mu P_k P_k^\top)^{-1} P_k\right\| \, \|g_k\| \le \frac{1}{\mu}\|g_k\|. \quad (26)$$

By (22), $\|g_k\| \to 0$, and hence $\|d_k\| \to 0$. Thus, for all sufficiently large k, $\|d_k\| \le \frac{(1-2\beta)\mu}{L_H}$, so $\alpha = 1$ satisfies the Armijo condition. Since the line search starts from $\alpha = 1$, (24) follows. □

We have therefore shown that, for sufficiently large k, the update can be treated with $\alpha_k = 1$.

Next, we quantify the effect of the regularization term η_k and the error incurred when replacing H_k by the limiting Hessian H_*.

Recall the filtrations defined in (7) and (8). In particular, x_k and y_k are $\mathcal{F}_k$-measurable, while the new Gaussian block S_k is independent of $\mathcal{F}_k$.

Lemma D.5. *Suppose that Assumption 3.2 and Condition 2.1(b) hold. For sufficiently large k, assume that $x_k \in B(x_*)$, and let $n > s + 2$. Then,*

$$\mathbb{E}\left[\left\|d_k + P_k^\top (P_k H_k P_k^\top)^{-1} P_k g_k\right\| \,\middle|\, \mathcal{F}_k\right] = O(\|x_k - x_*\|_{H_*}^{1+\delta}) \quad (27)$$

$$\mathbb{E}\left[\left\|d_k + P_k^\top (P_k H_* P_k^\top)^{-1} P_k g_k\right\|_{H_*} \,\middle|\, \mathcal{F}_k\right] = O(\|x_k - x_*\|_{H_*}^{1+\delta}). \quad (28)$$

Proof. In this proof, for nonsingular matrices A and B, we use the identity

$$A^{-1} - B^{-1} = A^{-1}(I - AB^{-1}) = A^{-1}(B - A)B^{-1}. \quad (29)$$

We argue below conditional on $\mathcal{F}_k$, so that y_k, H_k, g_k can be regarded as fixed, while S_k, P_k, and possibly η_k remain random. We use Condition 2.1(b) pathwise.

We first prove (27). From the definition of d_k,

$$\begin{aligned}
&d_k + P_k^\top (P_k H_k P_k^\top)^{-1} P_k g_k \\
&= P_k^\top \Big((P_k H_k P_k^\top)^{-1} - (P_k H_k P_k^\top + \eta_k I_{s+1})^{-1} \Big) P_k g_k \\
&\overset{(A)}{=} P_k^\top (P_k H_k P_k^\top)^{-1} \Big(P_k H_k P_k^\top + \eta_k I_{s+1} - P_k H_k P_k^\top \Big) (P_k H_k P_k^\top + \eta_k I_{s+1})^{-1} P_k g_k \\
&= \eta_k \, P_k^\top (P_k H_k P_k^\top)^{-1} (P_k H_k P_k^\top + \eta_k I_{s+1})^{-1} P_k g_k.
\end{aligned}$$

(A) uses (29). Therefore,

$$\begin{aligned}
&\left\|d_k + P_k^\top (P_k H_k P_k^\top)^{-1} P_k g_k\right\| \\
&\le \eta_k \, \|H_k^{-1/2}\| \cdot \left\| H_k^{1/2} P_k^\top (P_k H_k P_k^\top)^{-1} (P_k H_k P_k^\top + \eta_k I_{s+1})^{-1} P_k H_k^{1/2} \right\| \cdot \|H_k^{-1/2}\| \cdot \|g_k\| \\
&\le \eta_k \, \|H_k^{-1/2}\|^2 \cdot \left\| H_k^{1/2} P_k^\top (P_k H_k P_k^\top)^{-1} (P_k H_k P_k^\top + \eta_k I_{s+1})^{-1} P_k H_k^{1/2} \right\| \cdot \|g_k\|. \quad (30)
\end{aligned}$$

Let the thin singular value decomposition of $P_k H_k^{1/2} \in \mathbb{R}^{(s+1)\times n}$ be

$$P_k H_k^{1/2} = U \Sigma V^\top,$$

where $\Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_{s+1})$ with $\sigma_i > 0$. Then

$$P_k H_k P_k^\top = U \Sigma^2 U^\top,$$

and hence

$$(P_k H_k P_k^\top)^{-1} = U \Sigma^{-2} U^\top, \qquad (P_k H_k P_k^\top + \eta_k I_{s+1})^{-1} = U (\Sigma^2 + \eta_k I_{s+1})^{-1} U^\top.$$

Thus,

$$\begin{aligned}
&H_k^{1/2} P_k^\top (P_k H_k P_k^\top)^{-1} (P_k H_k P_k^\top + \eta_k I_{s+1})^{-1} P_k H_k^{1/2} \\
&= V \Sigma U^\top U \Sigma^{-2} U^\top U (\Sigma^2 + \eta_k I_{s+1})^{-1} U^\top U \Sigma V^\top \\
&= V (\Sigma^2 + \eta_k I_{s+1})^{-1} V^\top.
\end{aligned}$$

Therefore,

$$\left\| H_k^{1/2} P_k^\top (P_k H_k P_k^\top)^{-1} (P_k H_k P_k^\top + \eta_k I_{s+1})^{-1} P_k H_k^{1/2} \right\| = \max_{1 \le i \le s+1} \frac{1}{\sigma_i^2 + \eta_k} = \frac{1}{\lambda_{\min}(P_k H_k P_k^\top) + \eta_k} \le \frac{1}{\lambda_{\min}(P_k H_k P_k^\top)}.$$

Moreover, since $H_k \succeq \mu I_n$ by (19),

$$P_k H_k P_k^\top \succeq \mu P_k P_k^\top, \qquad \lambda_{\min}(P_k H_k P_k^\top) \ge \mu \, \lambda_{\min}(P_k P_k^\top),$$

and

$$\|H_k^{-1/2}\|^2 \le \frac{1}{\mu}.$$

Substituting these bounds into (30) gives

$$\left\| d_k + P_k^\top (P_k H_k P_k^\top)^{-1} P_k g_k \right\| \le \frac{\eta_k}{\mu^2} \cdot \frac{1}{\lambda_{\min}(P_k P_k^\top)} \, \|g_k\|. \tag{31}$$

Here, let $Q \in \mathbb{R}^{n \times n}$ be an orthogonal matrix satisfying

$$Q \frac{y_k}{\|y_k\|} = e_1$$

and define

$$\widetilde{S}_k \coloneqq S_k Q^\top, \qquad \widetilde{P}_k \coloneqq \begin{bmatrix} e_1^\top \\ \widetilde{S}_k \end{bmatrix}.$$

By the rotational invariance of Gaussian matrices, conditionally on $\mathcal{F}_k$,

$$\widetilde{S}_k \sim \mathcal{N}^{s \times n}(0, 1/s).$$

Moreover,

$$P_k Q^\top = \begin{bmatrix} (y_k/\|y_k\|)^\top Q^\top \\ S_k Q^\top \end{bmatrix} = \begin{bmatrix} e_1^\top \\ \widetilde{S}_k \end{bmatrix} = \widetilde{P}_k.$$

Hence,

$$P_k P_k^\top = (P_k Q^\top)(P_k Q^\top)^\top = \widetilde{P}_k \widetilde{P}_k^\top.$$

Furthermore, decomposing $\widetilde{S}_k = [\, a_k, \; B_k \,]$ by columns with $a_k \in \mathbb{R}^s$ and $B_k \in \mathbb{R}^{s \times (n-1)}$,

$$\widetilde{P}_k \widetilde{P}_k^\top = \begin{bmatrix} 1 & a_k^\top \\ a_k & a_k a_k^\top + B_k B_k^\top \end{bmatrix}.$$

By the Schur complement formula,

$$(\widetilde{P}_k \widetilde{P}_k^\top)^{-1} = \begin{bmatrix} 1 + a_k^\top (B_k B_k^\top)^{-1} a_k & -a_k^\top (B_k B_k^\top)^{-1} \\ -(B_k B_k^\top)^{-1} a_k & (B_k B_k^\top)^{-1} \end{bmatrix}.$$

Therefore,

$$\frac{1}{\lambda_{\min}(P_k P_k^\top)} = \frac{1}{\lambda_{\min}(\widetilde{P}_k \widetilde{P}_k^\top)} \le \mathrm{tr}\big((\widetilde{P}_k \widetilde{P}_k^\top)^{-1}\big) = 1 + a_k^\top (B_k B_k^\top)^{-1} a_k + \mathrm{tr}\big((B_k B_k^\top)^{-1}\big). \tag{32}$$

Here, a_k is independent of B_k, and conditionally $a_k \sim N(0, I_s/s)$. Hence,

$$\mathbb{E}\left[a_k^\top (B_k B_k^\top)^{-1} a_k \;\middle|\; \mathcal{F}_k, B_k \right] = \mathrm{tr}\left((B_k B_k^\top)^{-1} \mathbb{E}[a_k a_k^\top \mid \mathcal{F}_k, B_k] \right) = \frac{1}{s} \mathrm{tr}\big((B_k B_k^\top)^{-1}\big).$$

Moreover, the columns of B_k are conditionally independent given $\mathcal{F}_k$ and distributed as $N(0, I_s/s)$. Therefore, by the standard construction of the Wishart distribution,

$$B_k B_k^\top \sim W_s\left(n-1, \frac{1}{s} I_s\right).$$

For a central Wishart matrix, Hillier & Kan (2022) gives

$$W \sim W_m(\nu, \Sigma) \quad \Longrightarrow \quad \mathbb{E}[W^{-1}] = \frac{1}{\nu - m - 1}\Sigma^{-1}, \qquad \nu > m+1.$$

Substituting $m = s,\ \nu = n-1,\ \Sigma = I_s/s$, we obtain

$$\mathbb{E}\left[(B_k B_k^\top)^{-1} \mid \mathcal{F}_k\right] = \frac{1}{(n-1)-s-1}\left(\frac{1}{s} I_s\right)^{-1} = \frac{s}{n-s-2} I_s \qquad (n-s>2).$$

Therefore,

$$\mathbb{E}\left[\mathrm{tr}\big((B_k B_k^\top)^{-1}\big) \mid \mathcal{F}_k\right] = \frac{s^2}{n-s-2}.$$

Hence, from (32),

$$\begin{aligned}\mathbb{E}\left[\frac{1}{\lambda_{\min}(P_k P_k^\top)} \,\middle|\, \mathcal{F}_k\right] &\le 1 + \left(1 + \frac{1}{s}\right)\frac{s^2}{n-s-2} \\ &= \frac{n+s^2-2}{n-s-2}. \end{aligned} \tag{33}$$

Taking conditional expectations on both sides of (31), and using (33) together with Condition 2.1(b), we obtain

$$\mathbb{E}\left[\left\|d_k + P_k^\top (P_k H_k P_k^\top)^{-1} P_k g_k\right\| \mid \mathcal{F}_k\right] \le \frac{C_\eta}{\mu^2}\frac{n+s^2-2}{n-s-2}\|g_k\|^{1+\delta}.$$

Moreover, by (22), $\|g_k\| = O(\|x_k - x_*\|_{H_*})$, so

$$\mathbb{E}\left[\left\|d_k + P_k^\top (P_k H_k P_k^\top)^{-1} P_k g_k\right\| \mid \mathcal{F}_k\right] = O(\|x_k - x_*\|_{H_*}^{1+\delta}),$$

that is, (27) follows.

Next, we prove (28). By the triangle inequality,

$$\begin{aligned}&\left\|d_k + P_k^\top (P_k H_* P_k^\top)^{-1} P_k g_k\right\|_{H_*} \\ &\qquad \le \underbrace{\left\|H_*^{1/2}\left(d_k + P_k^\top (P_k H_k P_k^\top)^{-1} P_k g_k\right)\right\|}_{\text{(II)}} \\ &\qquad\quad + \underbrace{\left\|H_*^{1/2} P_k^\top \left((P_k H_* P_k^\top)^{-1} - (P_k H_k P_k^\top)^{-1}\right) P_k g_k\right\|}_{\text{(III)}}. \end{aligned} \tag{34}$$

The conditional expectation of (II) is $O(\|x_k - x_*\|_{H_*}^{1+\delta})$ by (27).

On the other hand, (III) can be written using (29) as

$$(P_k H_* P_k^\top)^{-1} - (P_k H_k P_k^\top)^{-1} = (P_k H_* P_k^\top)^{-1} P_k (H_k - H_*) P_k^\top (P_k H_k P_k^\top)^{-1}$$

and therefore

$$\begin{aligned}\text{(III)} &\le \left\|H_*^{1/2} P_k^\top (P_k H_* P_k^\top)^{-1} P_k\right\| \|H_k - H_*\| \left\|P_k^\top (P_k H_k P_k^\top)^{-1} P_k\right\| \|g_k\| \\ &\le \frac{1}{\sqrt{\lambda_{\min}(H_*)}} \cdot \|H_k - H_*\| \cdot \frac{1}{\mu} \cdot \|g_k\|. \end{aligned}$$

Moreover,

$$\|H_k - H_*\| = O(\|x_k - x_*\|_{H_*}), \qquad \|g_k\| = O(\|x_k - x_*\|_{H_*})$$

imply

$$\text{(III)} = O(\|x_k - x_*\|_{H_*}^2).$$

Returning this to (34), we obtain

$$\begin{aligned}\mathbb{E}\left[\left\|d_k + P_k^\top (P_k H_* P_k^\top)^{-1} P_k g_k\right\|_{H_*} \,\middle|\, \mathcal{F}_k\right] &\le O(\|x_k - x_*\|_{H_*}^{1+\delta}) + O(\|x_k - x_*\|_{H_*}^2) \\ &= O(\|x_k - x_*\|_{H_*}^{1+\delta}),\end{aligned}$$

that is, (28). □

Lemma D.5 shows that both the error caused by introducing regularization and the error from replacing H_k by H_* can be treated as higher-order terms. The next lemma goes one step further and directly bounds the difference between the projection-type operators

$$P_k^\top (P_k H P_k^\top)^{-1} P_k H$$

themselves. This estimate will be used repeatedly later when organizing the update equations for x_k, g_k, and $y_{k,t}$.

Lemma D.6. *Suppose that Assumption 3.2 holds. Then*

$$\left\| P_k^\top (P_k H_k P_k^\top)^{-1} P_k H_k - P_k^\top (P_k H_* P_k^\top)^{-1} P_k H_* \right\| = O(\|x_k - x_*\|_{H_*})$$

holds.

Proof.

$$\begin{aligned}&P_k^\top (P_k H_k P_k^\top)^{-1} P_k H_k - P_k^\top (P_k H_* P_k^\top)^{-1} P_k H_* \\ &= P_k^\top (P_k H_k P_k^\top)^{-1} P_k H_k - P_k^\top (P_k H_* P_k^\top)^{-1} P_k H_k + P_k^\top (P_k H_* P_k^\top)^{-1} P_k (H_k - H_*) \\ &= -P_k^\top (P_k H_k P_k^\top)^{-1} P_k (H_k - H_*) P_k^\top (P_k H_* P_k^\top)^{-1} P_k H_k + P_k^\top (P_k H_* P_k^\top)^{-1} P_k (H_k - H_*).\end{aligned}$$

In the last equality, we used (29). Therefore,

$$\begin{aligned}&\left\| P_k^\top (P_k H_k P_k^\top)^{-1} P_k H_k - P_k^\top (P_k H_* P_k^\top)^{-1} P_k H_* \right\| \\ &\quad \le \left\| P_k^\top (P_k H_k P_k^\top)^{-1} P_k \right\| \|H_k - H_*\| \left\| P_k^\top (P_k H_* P_k^\top)^{-1} P_k H_k \right\| \\ &\qquad + \left\| P_k^\top (P_k H_* P_k^\top)^{-1} P_k \right\| \|H_k - H_*\|.\end{aligned}$$

Here, by the projection property,

$$\begin{aligned}\left\| P_k^\top (P_k H P_k^\top)^{-1} P_k \right\| &\le \|H^{-1/2}\| \left\| H^{1/2} P_k^\top (P_k H P_k^\top)^{-1} P_k H^{1/2} \right\| \|H^{-1/2}\| \\ &\le \|H^{-1/2}\|^2 = \frac{1}{\lambda_{\min}(H)}\end{aligned}$$

holds. Applying this with $H = H_k$ and $H = H_*$, and further using $\|H_k\| \le L$ and $\lambda_{\min}(H_k) \ge \mu$, we obtain

$$\left\| P_k^\top (P_k H_k P_k^\top)^{-1} P_k H_k - P_k^\top (P_k H_* P_k^\top)^{-1} P_k H_* \right\| \le \left(\frac{1}{\lambda_{\min}(H_*)} + \frac{L}{\mu\, \lambda_{\min}(H_*)} \right) \|H_k - H_*\|.$$

Finally, Assumption 3.2(b) gives $\|H_k - H_*\| \le \frac{L_H}{\sqrt{\lambda_{\min}(H_*)}} \|x_k - x_*\|_{H_*}$, which proves the claim. □

We have now collected the perturbation estimates needed for the local analysis. We next record the projection inequality used to control the leading terms.

The outer update uses the anchored matrix

$$P_k = \begin{bmatrix} y_k^\top / \|y_k\| \\ S_k \end{bmatrix},$$

where y_k is $\mathcal{F}_k$-measurable and S_k is independent of $\mathcal{F}_k$. Therefore, for any vector v,

$$v^\top H_*^{1/2}\mathbb{E}\Big[I - \Pi(H_*^{1/2}P_k^\top)\,\Big|\,\mathcal{F}_k\Big]\,H_*^{1/2}v \le \Big(1 - \lambda_{\min}\Big(\mathbb{E}\Big[\Pi(H_*^{1/2}P_k^\top)\,\Big|\,\mathcal{F}_k\Big]\Big)\Big)\,\|v\|_{H_*}^2. \tag{35}$$

Similarly, since $R_{k,t} \sim \mathcal{N}^{r\times n}(0, 1/r)$ is independent of $\mathcal{F}_{k,t}$, for any vector v,

$$\begin{aligned} v^\top H_*^{1/2}\mathbb{E}\Big[I - \Pi(H_*^{1/2}R_{k,t}^\top)\,\Big|\,\mathcal{F}_{k,t}\Big]\,H_*^{1/2}v &= v^\top H_*^{1/2}\Big(I - \mathbb{E}\Big[\Pi(H_*^{1/2}R_{k,t}^\top)\Big]\Big)\,H_*^{1/2}v \\ &\le \Big(1 - \lambda_{\min}\Big(\mathbb{E}\Big[\Pi(H_*^{1/2}R_{k,t}^\top)\Big]\Big)\Big)\,\|v\|_{H_*}^2. \end{aligned} \tag{36}$$

We now prove Lemma 3.1. The proof estimates the update equation for $x_{k+1} - x_*$ and derives a local linear bound in conditional expectation. This estimate will later be used to control the error terms in y_k.

Proof of Lemma 3.1

Proof. For sufficiently large k, we have $\alpha_k = 1$, and hence $x_{k+1} = x_k + d_k$. Therefore,

$$\begin{aligned} x_{k+1} - x_* &= x_k - x_* - P_k^\top(P_kH_*P_k^\top)^{-1}P_kg_k + \big(d_k + P_k^\top(P_kH_*P_k^\top)^{-1}P_kg_k\big) \\ &\overset{(A)}{=} x_k - x_* - P_k^\top(P_kH_*P_k^\top)^{-1}P_k\big(H_*(x_k - x_*) + r_k\big) + \big(d_k + P_k^\top(P_kH_*P_k^\top)^{-1}P_kg_k\big) \\ &= \big(I_n - P_k^\top(P_kH_*P_k^\top)^{-1}P_kH_*\big)(x_k - x_*) - P_k^\top(P_kH_*P_k^\top)^{-1}P_kr_k \\ &\quad + \big(d_k + P_k^\top(P_kH_*P_k^\top)^{-1}P_kg_k\big). \end{aligned}$$

Here, in (A), we used $g_k = H_*(x_k - x_*) + r_k$, which follows from the definition of r_k.

Multiplying both sides by $H_*^{1/2}$ yields

$$\begin{aligned} H_*^{1/2}(x_{k+1} - x_*) = &\underbrace{\Big(I_n - \Pi(H_*^{1/2}P_k^\top)\Big)\,H_*^{1/2}(x_k - x_*)}_{\text{(I)}} \\ &- \underbrace{H_*^{1/2}P_k^\top(P_kH_*P_k^\top)^{-1}P_kr_k}_{\text{(II)}} \\ &+ \underbrace{H_*^{1/2}\big(d_k + P_k^\top(P_kH_*P_k^\top)^{-1}P_kg_k\big)}_{\text{(III)}}. \end{aligned} \tag{37}$$

Taking norms on both sides and then taking conditional expectations gives

$$\mathbb{E}\left[\|x_{k+1} - x_*\|_{H_*} \mid \mathcal{F}_k\right] \le \mathbb{E}[\|\text{(I)}\| \mid \mathcal{F}_k] + \mathbb{E}[\|\text{(II)}\| \mid \mathcal{F}_k] + \mathbb{E}[\|\text{(III)}\| \mid \mathcal{F}_k].$$

We bound each term below. First, for (I), Jensen's inequality gives

$$\begin{aligned} \mathbb{E}[\|\text{(I)}\| \mid \mathcal{F}_k] &\le \left(\mathbb{E}\left[\left\|\Big(I_n - \Pi(H_*^{1/2}P_k^\top)\Big)\,H_*^{1/2}(x_k - x_*)\right\|^2\middle|\mathcal{F}_k\right]\right)^{1/2} \\ &= \Big((x_k - x_*)^\top H_*^{1/2}\mathbb{E}\left[I_n - \Pi(H_*^{1/2}P_k^\top) \mid \mathcal{F}_k\right]H_*^{1/2}(x_k - x_*)\Big)^{1/2} \\ &\le \Big(1 - \lambda_{\min}\Big(\mathbb{E}\Big[\Pi(H_*^{1/2}P_k^\top)\,\Big|\,\mathcal{F}_k\Big]\Big)\Big)^{1/2}\,\|x_k - x_*\|_{H_*}. \end{aligned} \tag{38}$$

The final inequality follows from (35).

Next, for (II),

$$\begin{aligned}\|\text{(II)}\| &\le \left\| H_*^{1/2} P_k^\top (P_k H_* P_k^\top)^{-1} P_k \right\| \|r_k\| \\ &\le \left\| H_*^{1/2} P_k^\top (P_k H_* P_k^\top)^{-1} P_k H_*^{1/2} \right\| \|H_*^{-1/2}\| \, \|r_k\| \\ &\le \frac{1}{\sqrt{\lambda_{\min}(H_*)}} \|r_k\| \\ &= O(\|x_k - x_*\|_{H_*}^2). \end{aligned} \tag{39}$$

In the last estimate, we used (21). Therefore,

$$\mathbb{E}[\|\text{(II)}\| \mid \mathcal{F}_k] \le O(\|x_k - x_*\|_{H_*}^2).$$

Finally, for (III), by (28),

$$\mathbb{E}\left[\left\| H_*^{1/2} \big(d_k + P_k^\top (P_k H_* P_k^\top)^{-1} P_k g_k \big) \right\| \middle| \mathcal{F}_k \right] = O(\|x_k - x_*\|_{H_*}^{1+\delta}). \tag{40}$$

Substituting (38), (39), and (40) into (37), we obtain

$$\begin{aligned}&\mathbb{E}\left[\|x_{k+1} - x_*\|_{H_*} \mid \mathcal{F}_k \right] \\ &\le \left(1 - \lambda_{\min}\Big(\mathbb{E}\Big[\Pi(H_*^{1/2} P_k^\top) \,\Big|\, \mathcal{F}_k \Big] \Big) \right)^{1/2} \|x_k - x_*\|_{H_*} \\ &\qquad\qquad + O(\|x_k - x_*\|_{H_*}^2) + O(\|x_k - x_*\|_{H_*}^{1+\delta}) \\ &\le \left(1 - \lambda_{\min}\Big(\mathbb{E}\Big[\Pi(H_*^{1/2} P_k^\top) \,\Big|\, \mathcal{F}_k \Big] \Big) \right)^{1/2} \|x_k - x_*\|_{H_*} + O(\|x_k - x_*\|_{H_*}^{1+\delta}), \end{aligned}$$

that is, (9). □

We now turn to the gradient estimate used in the local superlinear convergence proof. The key point is that the leading term of g_{k+1} is controlled by the anchor error $y_k - y_{k,*}$. This estimate allows us to combine the outer update with the error contraction of the randomized Kaczmarz-type inner loop.

Proof of Proposition 3.1 Taking conditional expectations in (41) and (43), and using Lemma D.5, we obtain that the regularization error is of order $O(\|g_k\|^{1+\delta})$, so it remains to bound the unregularized projected Newton residual. Exploiting the fact that the anchored subspace contains y_k, and replacing H_k by H_* up to higher-order terms, we obtain the desired bound in (10).

Proof. Expanding $g_{k+1} = \nabla f(x_{k+1})$ around x_k by Taylor's theorem gives

$$\begin{aligned} g_{k+1} &= g_k + H_k(x_{k+1} - x_k) + O(\|x_{k+1} - x_k\|^2) \\ &= g_k + H_k d_k + O(\|d_k\|^2). \end{aligned} \tag{41}$$

Since

$$\|d_k\| = \|P_k^\top (P_k H_k P_k^\top + \eta_k I_{s+1})^{-1} P_k g_k\| \le \frac{1}{\mu} \|g_k\|,$$

we have

$$O(\|d_k\|^2) = O(\|g_k\|^2). \tag{42}$$

Next,

$$g_k + H_k d_k = g_k - H_k P_k^\top (P_k H_k P_k^\top)^{-1} P_k g_k + H_k \big(d_k + P_k^\top (P_k H_k P_k^\top)^{-1} P_k g_k \big). \tag{43}$$

Therefore, by (41), (42), and the triangle inequality,

$$\begin{aligned}\mathbb{E}[\|g_{k+1}\| \mid \mathcal{F}_k] \le{}& \mathbb{E}\left[\left\| g_k - H_k P_k^\top (P_k H_k P_k^\top)^{-1} P_k g_k \right\| \middle| \mathcal{F}_k \right] \\ &+ \|H_k\| \, \mathbb{E}\left[\left\| d_k + P_k^\top (P_k H_k P_k^\top)^{-1} P_k g_k \right\| \middle| \mathcal{F}_k \right] + O(\|g_k\|^2). \end{aligned}$$

By Lemma D.5, the local boundedness of $\|H_k\|$, and (22), the second term on the right-hand side satisfies

$$\|H_k\|\,\mathbb{E}\left[\left\|d_k + P_k^\top (P_k H_k P_k^\top)^{-1} P_k g_k\right\| \middle| \mathcal{F}_k\right] = O(\|g_k\|^{1+\delta}).$$

Hence, since $\delta \in (0,1]$,

$$\mathbb{E}[\|g_{k+1}\| \mid \mathcal{F}_k] \le \mathbb{E}\left[\left\|g_k - H_k P_k^\top (P_k H_k P_k^\top)^{-1} P_k g_k\right\| \middle| \mathcal{F}_k\right] + O(\|g_k\|^{1+\delta}). \tag{44}$$

Here,

$$y_{k,*} = \frac{H_k^{-1} g_k}{\|g_k\|}$$

and hence

$$g_k = \|g_k\| H_k y_{k,*}.$$

Also, since

$$P_k^\top e_1 = \frac{y_k}{\|y_k\|},$$

we have

$$H_k P_k^\top (P_k H_k P_k^\top)^{-1} P_k H_k y_k = H_k y_k.$$

Using this identity,

$$\begin{aligned} &g_k - H_k P_k^\top (P_k H_k P_k^\top)^{-1} P_k g_k \\ &\quad = \|g_k\| H_k (y_{k,*} - y_k) - \|g_k\| H_k P_k^\top (P_k H_k P_k^\top)^{-1} P_k H_k (y_{k,*} - y_k). \end{aligned}$$

Below,

$$\begin{aligned} &- \|g_k\| H_k (y_k - y_{k,*}) + \|g_k\| H_k P_k^\top (P_k H_k P_k^\top)^{-1} P_k H_k (y_k - y_{k,*}) \\ &= \underbrace{-\|g_k\| H_*^{1/2} \left(I_n - \Pi(H_*^{1/2} P_k^\top)\right) H_*^{1/2} (y_k - y_{k,*})}_{\text{(I)}} \\ &\quad + \underbrace{\|g_k\| (H_* - H_k)(y_k - y_{k,*})}_{\text{(II)}} \\ &\quad - \underbrace{\|g_k\| \left(H_* P_k^\top (P_k H_* P_k^\top)^{-1} P_k H_* - H_k P_k^\top (P_k H_k P_k^\top)^{-1} P_k H_k\right)(y_k - y_{k,*})}_{\text{(III)}}. \end{aligned} \tag{45}$$

Here, since

$$H_* P_k^\top (P_k H_* P_k^\top)^{-1} P_k H_* = H_*^{1/2} \Pi(H_*^{1/2} P_k^\top) H_*^{1/2},$$

we obtain the form of (I).

We first bound (I). By Jensen's inequality,

$$\mathbb{E}[\|\text{(I)}\| \mid \mathcal{F}_k] \le \|g_k\| \left(\mathbb{E}\left[\left\|H_*^{1/2} \left(I_n - \Pi(H_*^{1/2} P_k^\top)\right) H_*^{1/2} (y_k - y_{k,*})\right\|^2 \middle| \mathcal{F}_k\right]\right)^{1/2}.$$

Moreover,

$$\begin{aligned} &\left\|H_*^{1/2} \left(I_n - \Pi(H_*^{1/2} P_k^\top)\right) H_*^{1/2} (y_k - y_{k,*})\right\|^2 \\ &= (y_k - y_{k,*})^\top H_*^{1/2} \left(I_n - \Pi(H_*^{1/2} P_k^\top)\right) H_* \left(I_n - \Pi(H_*^{1/2} P_k^\top)\right) H_*^{1/2} (y_k - y_{k,*}) \\ &\le \lambda_{\max}(H_*)\, (y_k - y_{k,*})^\top H_*^{1/2} \left(I_n - \Pi(H_*^{1/2} P_k^\top)\right) H_*^{1/2} (y_k - y_{k,*}), \end{aligned}$$

where, in the last inequality, we used

$$H_* \preceq \lambda_{\max}(H_*) I$$

and

$$\left(I_n - \Pi(H_*^{1/2}P_k^\top)\right)^2 = I_n - \Pi(H_*^{1/2}P_k^\top).$$

Therefore,

$$\begin{aligned}
&\mathbb{E}[\|\text{(I)}\| \mid \mathcal{F}_k] \\
&\le \sqrt{\lambda_{\max}(H_*)}\,\|g_k\|\left((y_k - y_{k,*})^\top H_*^{1/2}\mathbb{E}\left[I_n - \Pi(H_*^{1/2}P_k^\top) \mid \mathcal{F}_k\right] H_*^{1/2}(y_k - y_{k,*})\right)^{1/2} \\
&\le \left(\lambda_{\max}(H_*)\left(1 - \lambda_{\min}\left(\mathbb{E}\left[\Pi(H_*^{1/2}P_k^\top)\,\middle|\,\mathcal{F}_k\right]\right)\right)\right)^{1/2}\|g_k\|\,\|y_k - y_{k,*}\|_{H_*}. \qquad (46)
\end{aligned}$$

The final inequality follows from (35).

Next, we bound (II). By (22) and Assumption 3.2(b),

$$\|H_k - H_*\| = O(\|x_k - x_*\|) = O(\|g_k\|)$$

and hence

$$\begin{aligned}
\mathbb{E}[\|\text{(II)}\| \mid \mathcal{F}_k] &\le \|g_k\|\,\|H_k - H_*\|\,\|y_k - y_{k,*}\| \\
&= O(\|g_k\|^2)\|y_k - y_{k,*}\|_{H_*}. \qquad (47)
\end{aligned}$$

Next, we bound (III). We first write

$$\begin{aligned}
&H_*P_k^\top(P_kH_*P_k^\top)^{-1}P_kH_* - H_kP_k^\top(P_kH_kP_k^\top)^{-1}P_kH_k \\
&= H_*\left[P_k^\top(P_kH_*P_k^\top)^{-1}P_kH_* - P_k^\top(P_kH_kP_k^\top)^{-1}P_kH_k\right] \\
&\quad + (H_* - H_k)P_k^\top(P_kH_kP_k^\top)^{-1}P_kH_k.
\end{aligned}$$

Hence, by the triangle inequality,

$$\begin{aligned}
&\left\|H_*P_k^\top(P_kH_*P_k^\top)^{-1}P_kH_* - H_kP_k^\top(P_kH_kP_k^\top)^{-1}P_kH_k\right\| \\
&\le \|H_*\|\left\|P_k^\top(P_kH_*P_k^\top)^{-1}P_kH_* - P_k^\top(P_kH_kP_k^\top)^{-1}P_kH_k\right\| \\
&\quad + \|H_* - H_k\|\left\|P_k^\top(P_kH_kP_k^\top)^{-1}P_kH_k\right\|.
\end{aligned}$$

The first term is $O(\|x_k - x_*\|_{H_*})$ by Lemma D.6 and the boundedness of $\|H_*\|$. For the second term, the local Lipschitz continuity of the Hessian gives

$$\|H_k - H_*\| = O(\|x_k - x_*\|) = O(\|x_k - x_*\|_{H_*}),$$

while

$$\left\|P_k^\top(P_kH_kP_k^\top)^{-1}P_kH_k\right\| = O(1)$$

for all sufficiently large k. Therefore,

$$\begin{aligned}
&\left\|H_*P_k^\top(P_kH_*P_k^\top)^{-1}P_kH_* - H_kP_k^\top(P_kH_kP_k^\top)^{-1}P_kH_k\right\| \\
&= O(\|x_k - x_*\|_{H_*}).
\end{aligned}$$

Thus, by (22),

$$\begin{aligned}
\mathbb{E}[\|\text{(III)}\| \mid \mathcal{F}_k] &\le \|g_k\|\,O(\|x_k - x_*\|_{H_*})\,\|y_k - y_{k,*}\| \\
&= O(\|g_k\|^2)\|y_k - y_{k,*}\|_{H_*}. \qquad (48)
\end{aligned}$$

Taking the conditional expectation of the norm in (45), and using (46), (47), and (48), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left\|g_k - H_kP_k^\top(P_kH_kP_k^\top)^{-1}P_kg_k\right\|\middle|\mathcal{F}_k\right] \\
&\le \left(\left(\lambda_{\max}(H_*)\left(1 - \lambda_{\min}\left(\mathbb{E}\left[\Pi(H_*^{1/2}P_k^\top)\,\middle|\,\mathcal{F}_k\right]\right)\right)\right)^{1/2} + O(\|g_k\|)\right)\|g_k\|\,\|y_k - y_{k,*}\|_{H_*} \qquad (49)
\end{aligned}$$

Substituting (49) into (44), we obtain

$$\mathbb{E}[\|g_{k+1}\| \mid \mathcal{F}_k]$$
$$\leq \left(\left(\lambda_{\max}(H_*)\left(1-\lambda_{\min}\left(\mathbb{E}\left[\Pi(H_*^{1/2}P_k^\top)\,\middle|\,\mathcal{F}_k\right]\right)\right)\right)^{1/2}+O(\|g_k\|)\right)\|g_k\|\,\|y_k-y_{k,*}\|_{H_*}+O(\|g_k\|^{1+\delta}).$$

This is (10). □

We have shown that the gradient error is governed by $y_k - y_{k,*}$. It remains to show that the vector y_k, produced by the randomized Kaczmarz-type method, tracks the moving target $y_{k,*}$. We first establish the contraction of the anchor error within each outer iteration and then bound the error introduced by the warm start. Combining these estimates with a logarithmically increasing number of inner iterations T_k will yield almost-sure anchor tracking.

For a fixed outer iteration k, we first prove Lemma 3.2, which estimates the change in the error $y_{k,t} - y_{k,*}$ after one inner iteration. The proof shows that the leading term contracts with the factor determined by

$$\lambda_{\min}\left(\mathbb{E}\left[\Pi(H_*^{1/2}R_{k,t}^\top)\right]\right),$$

while the error caused by replacing H_k with H_* is higher order.

Proof of Lemma 3.2

Proof. For all sufficiently large k, the randomized Kaczmarz update (14) gives

$$\begin{aligned}
y_{k,t+1}-y_{k,*} &= y_{k,t}-R_{k,t}^\top\left(R_{k,t}H_kR_{k,t}^\top\right)^{-1}R_{k,t}\left(H_ky_{k,t}-\frac{g_k}{\|g_k\|}\right)-y_{k,*}\\
&\overset{(A)}{=} y_{k,t}-y_{k,*}-R_{k,t}^\top\left(R_{k,t}H_kR_{k,t}^\top\right)^{-1}R_{k,t}H_k(y_{k,t}-y_{k,*})\\
&= \left(I_n-R_{k,t}^\top\left(R_{k,t}H_*R_{k,t}^\top\right)^{-1}R_{k,t}H_*\right)(y_{k,t}-y_{k,*})\\
&\quad+\left(R_{k,t}^\top\left(R_{k,t}H_*R_{k,t}^\top\right)^{-1}R_{k,t}H_*-R_{k,t}^\top\left(R_{k,t}H_kR_{k,t}^\top\right)^{-1}R_{k,t}H_k\right)(y_{k,t}-y_{k,*}).
\end{aligned}$$

Here, (A) follows from the definition of $y_{k,*}$, namely

$$H_ky_{k,*}=\frac{g_k}{\|g_k\|}.$$

Multiplying both sides by $H_*^{1/2}$ gives

$$\begin{aligned}
&H_*^{1/2}(y_{k,t+1}-y_{k,*})\\
&=\underbrace{\left(I_n-\Pi(H_*^{1/2}R_{k,t}^\top)\right)H_*^{1/2}(y_{k,t}-y_{k,*})}_{\text{(I)}}\\
&\quad+\underbrace{H_*^{1/2}\left(R_{k,t}^\top\left(R_{k,t}H_*R_{k,t}^\top\right)^{-1}R_{k,t}H_*-R_{k,t}^\top\left(R_{k,t}H_kR_{k,t}^\top\right)^{-1}R_{k,t}H_k\right)(y_{k,t}-y_{k,*})}_{\text{(II)}}.
\end{aligned}$$

Therefore, taking norms and conditional expectations gives

$$\mathbb{E}\left[\|y_{k,t+1}-y_{k,*}\|_{H_*}\mid\mathcal{F}_{k,t}\right]\leq\mathbb{E}[\|\text{(I)}\|\mid\mathcal{F}_{k,t}]+\mathbb{E}[\|\text{(II)}\|\mid\mathcal{F}_{k,t}]. \tag{50}$$

First, we bound (I). By Jensen's inequality,

$$\begin{aligned}
\mathbb{E}[\|\text{(I)}\|\mid\mathcal{F}_{k,t}] &\leq \left(\mathbb{E}\left[\|\text{(I)}\|^2\middle|\mathcal{F}_{k,t}\right]\right)^{1/2}\\
&= \left((y_{k,t}-y_{k,*})^\top H_*^{1/2}\mathbb{E}\left[I_n-\Pi(H_*^{1/2}R_{k,t}^\top)\mid\mathcal{F}_{k,t}\right]H_*^{1/2}(y_{k,t}-y_{k,*})\right)^{1/2}\\
&\leq \left(1-\lambda_{\min}\left(\mathbb{E}\left[\Pi(H_*^{1/2}R_{k,t}^\top)\right]\right)\right)^{1/2}\|y_{k,t}-y_{k,*}\|_{H_*}.
\end{aligned} \tag{51}$$

The final inequality follows from (36).

Next, we bound (II). Applying Lemma D.6 with $R_{k,t}$ in place of P_k, we obtain

$$\left\| R_{k,t}^\top \left(R_{k,t} H_* R_{k,t}^\top\right)^{-1} R_{k,t} H_* - R_{k,t}^\top \left(R_{k,t} H_k R_{k,t}^\top\right)^{-1} R_{k,t} H_k \right\| = O(\|x_k - x_*\|_{H_*}).$$

Therefore,

$$\begin{aligned}\|\text{(II)}\| &\le \|H_*^{1/2}\| \left\| R_{k,t}^\top \left(R_{k,t} H_* R_{k,t}^\top\right)^{-1} R_{k,t} H_* - R_{k,t}^\top \left(R_{k,t} H_k R_{k,t}^\top\right)^{-1} R_{k,t} H_k \right\| \|y_{k,t} - y_{k,*}\| \\ &= O(\|x_k - x_*\|_{H_*}) \, \|y_{k,t} - y_{k,*}\|_{H_*}.\end{aligned}$$

Since $y_{k,t} - y_{k,*}$ is $\mathcal{F}_{k,t}$-measurable, it follows that

$$\mathbb{E}[\|\text{(II)}\| \mid \mathcal{F}_{k,t}] = O(\|x_k - x_*\|_{H_*}) \, \|y_{k,t} - y_{k,*}\|_{H_*}. \tag{52}$$

Substituting (51) and (52) into (50), we obtain

$$\begin{aligned}&\mathbb{E}\left[\|y_{k,t+1} - y_{k,*}\|_{H_*} \mid \mathcal{F}_{k,t}\right] \\ &\le \left(\left(1 - \lambda_{\min}\left(\mathbb{E}\left[\Pi(H_*^{1/2} R_{k,t}^\top)\right]\right)\right)^{1/2} + O(\|x_k - x_*\|_{H_*}) \right) \|y_{k,t} - y_{k,*}\|_{H_*} \\ &= \left(\sqrt{1 - \lambda_R} + O(\|x_k - x_*\|_{H_*}) \right) \|y_{k,t} - y_{k,*}\|_{H_*},\end{aligned}$$

which proves the claim. □

Lemma 3.2 controls the contraction within a fixed outer iteration. When moving from iteration k to $k+1$, however, the next initial value $y_{k+1,0}$ is supplied by the warm start, and its error must be controlled separately. We next prove Lemma 3.3, which gives this warm-start error estimate.

Proof of Lemma 3.3

Proof. First, by the definition (2) of $y_{k+1,0}$,

$$\begin{aligned}&y_{k+1,0} - y_{k+1,*} \\ &= \left(1 - \frac{g_k^\top H_k (x_{k+1} - x_k)}{\|g_k\|^2}\right) y_k + \frac{1}{\|g_k\|}(x_{k+1} - x_k) - y_{k+1,*} \\ &= \underbrace{\left(1 - \frac{g_k^\top H_k (x_{k+1} - x_k)}{\|g_k\|^2}\right)(y_k - y_{k,*})}_{\text{(I)}} \\ &\qquad - \underbrace{\left(y_{k+1,*} - \left(1 - \frac{g_k^\top H_k (x_{k+1} - x_k)}{\|g_k\|^2}\right) y_{k,*} - \frac{1}{\|g_k\|}(x_{k+1} - x_k) \right)}_{\text{(II)}}.\end{aligned}$$

Therefore, by the triangle inequality,

$$\|y_{k+1,0} - y_{k+1,*}\|_{H_*} \le \|\text{(I)}\|_{H_*} + \|\text{(II)}\|_{H_*}. \tag{53}$$

We first bound the coefficient of (I). For sufficiently large k, $\alpha_k = 1$ and $x_{k+1} - x_k = d_k$, and hence

$$\begin{aligned}1 - \frac{g_k^\top H_k (x_{k+1} - x_k)}{\|g_k\|^2} &= 1 - \frac{g_k^\top H_k d_k}{\|g_k\|^2} \\ &= 1 - \frac{g_k^\top H_* d_k}{\|g_k\|^2} - \frac{g_k^\top (H_k - H_*) d_k}{\|g_k\|^2} \\ &= 1 + \frac{g_k^\top H_* P_k^\top (P_k H_* P_k^\top)^{-1} P_k g_k}{\|g_k\|^2} \\ &\qquad - \frac{g_k^\top H_* \left(d_k + P_k^\top (P_k H_* P_k^\top)^{-1} P_k g_k\right)}{\|g_k\|^2} \\ &\qquad - \frac{g_k^\top (H_k - H_*) d_k}{\|g_k\|^2}.\end{aligned} \tag{54}$$

Here, for the second term,

$$H_* P_k^\top (P_k H_* P_k^\top)^{-1} P_k = H_*^{1/2} \Pi(H_*^{1/2} P_k^\top) H_*^{-1/2},$$

and hence

$$\frac{g_k^\top H_* P_k^\top (P_k H_* P_k^\top)^{-1} P_k g_k}{\|g_k\|^2} = \frac{g_k^\top}{\|g_k\|^2} H_*^{1/2} \Pi(H_*^{1/2} P_k^\top) H_*^{-1/2} g_k.$$

Since $\Pi(H_*^{1/2} P_k^\top)$ is an orthogonal projection,

$$\|\Pi(H_*^{1/2} P_k^\top)\| \le 1.$$

Therefore,

$$\begin{aligned}\left| \frac{g_k^\top}{\|g_k\|^2} H_*^{1/2} \Pi(H_*^{1/2} P_k^\top) H_*^{-1/2} g_k \right| &\le \frac{\|H_*^{1/2} g_k\|}{\|g_k\|} \frac{\|H_*^{-1/2} g_k\|}{\|g_k\|} \\ &\le \sqrt{\lambda_{\max}(H_*)} \frac{1}{\sqrt{\lambda_{\min}(H_*)}} = \sqrt{\kappa(H_*)}. \end{aligned} \tag{55}$$

Moreover, by (28),

$$\mathbb{E}\left[\left\| d_k + P_k^\top (P_k H_* P_k^\top)^{-1} P_k g_k \right\|_{H_*} \,\middle|\, \mathcal{F}_k \right] = O(\|x_k - x_*\|_{H_*}^{1+\delta}).$$

Since $\|g_k\| = \Theta(\|x_k - x_*\|_{H_*})$ by (22), we have

$$\begin{aligned}&\mathbb{E}\left[\left| \frac{g_k^\top H_* \left(d_k + P_k^\top (P_k H_* P_k^\top)^{-1} P_k g_k \right)}{\|g_k\|^2} \right| \,\middle|\, \mathcal{F}_k \right] \\ &\le \frac{\|g_k\|_{H_*}}{\|g_k\|^2} \mathbb{E}\left[\left\| d_k + P_k^\top (P_k H_* P_k^\top)^{-1} P_k g_k \right\|_{H_*} \,\middle|\, \mathcal{F}_k \right] \\ &= O(\|x_k - x_*\|_{H_*}^{\delta}). \end{aligned} \tag{56}$$

Furthermore, by Assumption 3.2(b),

$$\|H_k - H_*\| = O(\|x_k - x_*\|_{H_*}).$$

Also, locally $\|d_k\| = O(\|x_k - x_*\|_{H_*})$. Hence, again using $\|g_k\| = \Theta(\|x_k - x_*\|_{H_*})$,

$$\left| \frac{g_k^\top (H_k - H_*) d_k}{\|g_k\|^2} \right| \le \frac{\|g_k\| \, \|H_k - H_*\| \, \|d_k\|}{\|g_k\|^2} = O(\|x_k - x_*\|_{H_*}). \tag{57}$$

Substituting (55), (56), and (57) into (54), we obtain

$$\mathbb{E}\left[\left| 1 - \frac{g_k^\top H_k (x_{k+1} - x_k)}{\|g_k\|^2} \right| \,\middle|\, \mathcal{F}_k \right] \le 1 + \sqrt{\kappa(H_*)} + O(\|x_k - x_*\|_{H_*}^{\delta}). \tag{58}$$

Here, $y_k - y_{k,*}$ is $\mathcal{F}_k$-measurable. Therefore, by (58),

$$\begin{aligned}\mathbb{E}[\|(\mathrm{I})\|_{H_*} \mid \mathcal{F}_k] &= \mathbb{E}\left[\left| 1 - \frac{g_k^\top H_k (x_{k+1} - x_k)}{\|g_k\|^2} \right| \,\middle|\, \mathcal{F}_k \right] \|y_k - y_{k,*}\|_{H_*} \\ &\le \left(1 + \sqrt{\kappa(H_*)} + O(\|x_k - x_*\|_{H_*}^{\delta}) \right) \|y_k - y_{k,*}\|_{H_*}. \end{aligned} \tag{59}$$

Next, we bound (II).

$$\|(\mathrm{II})\|_{H_*} = \left\| y_{k+1,*} - \left(1 - \frac{g_k^\top H_k (x_{k+1} - x_k)}{\|g_k\|^2} \right) y_{k,*} - \frac{1}{\|g_k\|} (x_{k+1} - x_k) \right\|_{H_*}.$$

Using

$$y_{k,*} = H_k^{-1} \frac{g_k}{\|g_k\|}, \qquad y_{k+1,*} = H_{k+1}^{-1} \frac{g_{k+1}}{\|g_{k+1}\|},$$

we obtain the exact decomposition

$$\begin{aligned}
&\|\text{(II)}\|_{H_*} \\
&\le \underbrace{\left\| H_*^{1/2} \left(H_{k+1}^{-1} - H_k^{-1}\right) \frac{g_{k+1}}{\|g_{k+1}\|} \right\|}_{\text{(II-a)}} \\
&\quad + \underbrace{\left\| H_*^{1/2} H_k^{-1} \left[\frac{g_{k+1}}{\|g_{k+1}\|} - \frac{g_k}{\|g_k\|} - \frac{1}{\|g_k\|} \left(I_n - \frac{g_k g_k^\top}{\|g_k\|^2} \right) H_k (x_{k+1} - x_k) \right] \right\|}_{\text{(II-b)}}.
\end{aligned}$$

First, for (II-a),

$$H_{k+1}^{-1} - H_k^{-1} = H_{k+1}^{-1}(H_k - H_{k+1})H_k^{-1}.$$

Therefore,

$$\begin{aligned}
\text{(II-a)} &\le \|H_*^{1/2} H_{k+1}^{-1}\| \, \|H_{k+1} - H_k\| \, \|H_k^{-1}\| \\
&= O(\|x_{k+1} - x_k\|) \\
&= O(\|x_k - x_*\|_{H_*}),
\end{aligned} \tag{60}$$

where we used (19), Assumption 3.2(b), and $\|x_{k+1} - x_k\| = O(\|g_k\|) = O(\|x_k - x_*\|_{H_*})$.

Next, for (II-b), by the triangle inequality,

$$\text{(II-b)} \le \|H_*^{1/2} H_k^{-1}\| \left(\left\| \frac{g_{k+1}}{\|g_{k+1}\|} - \frac{g_k}{\|g_k\|} \right\| + \frac{1}{\|g_k\|} \left\| \left(I_n - \frac{g_k g_k^\top}{\|g_k\|^2} \right) H_k (x_{k+1} - x_k) \right\| \right). \tag{61}$$

The first term satisfies

$$\left\| \frac{g_{k+1}}{\|g_{k+1}\|} - \frac{g_k}{\|g_k\|} \right\| \le 2. \tag{62}$$

For the second term, since

$$\left\| I_n - \frac{g_k g_k^\top}{\|g_k\|^2} \right\| \le 1,$$

we have

$$\frac{1}{\|g_k\|} \left\| \left(I_n - \frac{g_k g_k^\top}{\|g_k\|^2} \right) H_k (x_{k+1} - x_k) \right\| \le \frac{\|H_k(x_{k+1} - x_k)\|}{\|g_k\|}.$$

For sufficiently large k, $\alpha_k = 1$, and hence $x_{k+1} - x_k = d_k$. By (25),

$$-g_k^\top (x_{k+1} - x_k) \ge (x_{k+1} - x_k)^\top H_k (x_{k+1} - x_k) = \|x_{k+1} - x_k\|_{H_k}^2.$$

Moreover, by the Cauchy–Schwarz inequality in the H_k-metric,

$$-g_k^\top (x_{k+1} - x_k) \le \|g_k\|_{H_k^{-1}} \|x_{k+1} - x_k\|_{H_k}.$$

Therefore,

$$\|x_{k+1} - x_k\|_{H_k} \le \|g_k\|_{H_k^{-1}}.$$

It follows that

$$\begin{aligned}
\|H_k(x_{k+1} - x_k)\| &\le \sqrt{\lambda_{\max}(H_k)} \|x_{k+1} - x_k\|_{H_k} \\
&\le \sqrt{\lambda_{\max}(H_k)} \|g_k\|_{H_k^{-1}} \\
&\le \sqrt{\frac{\lambda_{\max}(H_k)}{\lambda_{\min}(H_k)}} \, \|g_k\| \\
&= \sqrt{\kappa(H_k)} \, \|g_k\|.
\end{aligned}$$

Hence,

$$\frac{1}{\|g_k\|}\left\|\left(I_n-\frac{g_kg_k^\top}{\|g_k\|^2}\right)H_k(x_{k+1}-x_k)\right\|\le\sqrt{\kappa(H_k)}. \tag{63}$$

Since $H_k\to H_*$,

$$\sqrt{\kappa(H_k)}=\sqrt{\kappa(H_*)}+O(\|x_k-x_*\|_{H_*}).$$

Moreover,

$$\begin{aligned}\|H_*^{1/2}H_k^{-1}\|&\le\|H_*^{1/2}H_*^{-1}\|+\|H_*^{1/2}(H_k^{-1}-H_*^{-1})\|\\&=\frac{1}{\sqrt{\lambda_{\min}(H_*)}}+\|H_*^{1/2}(H_k^{-1}-H_*^{-1})\|.\end{aligned}$$

Using

$$H_k^{-1}-H_*^{-1}=H_k^{-1}(H_*-H_k)H_*^{-1},$$

together with (19) and Assumption 3.2(b), we obtain

$$\|H_*^{1/2}(H_k^{-1}-H_*^{-1})\|=O(\|x_k-x_*\|_{H_*}).$$

Therefore,

$$\|H_*^{1/2}H_k^{-1}\|=\frac{1}{\sqrt{\lambda_{\min}(H_*)}}+O(\|x_k-x_*\|_{H_*}).$$

Thus, from (61), (62), and (63),

$$\begin{aligned}\text{(II-b)}&\le\left(\frac{1}{\sqrt{\lambda_{\min}(H_*)}}+O(\|x_k-x_*\|_{H_*})\right)\left(2+\sqrt{\kappa(H_*)}+O(\|x_k-x_*\|_{H_*})\right)\\&=\frac{2+\sqrt{\kappa(H_*)}}{\sqrt{\lambda_{\min}(H_*)}}+O(\|x_k-x_*\|_{H_*}).\end{aligned} \tag{64}$$

Combining (60) and (64), we obtain

$$\|\text{(II)}\|_{H_*}\le\frac{2+\sqrt{\kappa(H_*)}}{\sqrt{\lambda_{\min}(H_*)}}+O(\|x_k-x_*\|_{H_*}). \tag{65}$$

Consequently,

$$\mathbb{E}[\|\text{(II)}\|_{H_*}\mid\mathcal{F}_k]\le\frac{2+\sqrt{\kappa(H_*)}}{\sqrt{\lambda_{\min}(H_*)}}+O(\|x_k-x_*\|_{H_*}).$$

Finally, taking conditional expectations in (53), and using (59) and (65), we obtain

$$\begin{aligned}\mathbb{E}[\|y_{k+1,0}-y_{k+1,*}\|_{H_*}\mid\mathcal{F}_k]\le&\left(1+\sqrt{\kappa(H_*)}+O(\|x_k-x_*\|_{H_*}^{\delta})\right)\|y_k-y_{k,*}\|_{H_*}\\&+\frac{2+\sqrt{\kappa(H_*)}}{\sqrt{\lambda_{\min}(H_*)}}+O(\|x_k-x_*\|_{H_*}).\end{aligned}$$

This proves the claim. □

We have now obtained both the one-step contraction of the inner iteration and the warm-start error estimate. We next combine these estimates with the growth condition on T_k to prove Proposition 3.2. The key idea is that a logarithmically increasing number of inner iterations makes the accumulated contraction sufficiently strong to overcome the warm-start error, yielding $\|y_k-y_{k,*}\|_{H_*}\to 0$ almost surely.

Proof of Proposition 3.2

Proof. We prove a slightly more general statement. Let $\gamma > 0$ and suppose that, for all sufficiently large k,

$$T_k \geq \left\lceil \frac{2(1+\gamma)}{-\log(1-\lambda_R)} \log(k+1) \right\rceil. \tag{66}$$

The condition stated in the proposition corresponds to the particular choice $\gamma = 1$.

We first make the local contraction estimate uniform on a deterministic neighborhood. By Lemma 3.2, the $O(\|x_k - x_*\|_{H_*})$ term is a local estimate with a deterministic constant independent of the inner index t. Since

$$\frac{1+\gamma/2}{2(1+\gamma)} < \frac{1}{2}$$

and $0 < 1 - \lambda_R < 1$, we have

$$\sqrt{1-\lambda_R} < (1-\lambda_R)^{\frac{1+\gamma/2}{2(1+\gamma)}}.$$

Hence, by choosing the local neighborhood sufficiently small, whenever $x_k \in B(x_*)$,

$$\sqrt{1-\lambda_R} + O(\|x_k - x_*\|_{H_*}) \leq (1-\lambda_R)^{\frac{1+\gamma/2}{2(1+\gamma)}}. \tag{67}$$

Therefore, whenever $x_k \in B(x_*)$, applying Lemma 3.2 repeatedly over $t = 0, \ldots, T_k - 1$ and using the tower property gives

$$\mathbb{E}\left[\|y_k - y_{k,*}\|_{H_*} \mid \mathcal{F}_{k,0}\right] \leq (1-\lambda_R)^{\frac{1+\gamma/2}{2(1+\gamma)} T_k} \|y_{k,0} - y_{k,*}\|_{H_*}.$$

By (66), for all sufficiently large k,

$$(1-\lambda_R)^{\frac{1+\gamma/2}{2(1+\gamma)} T_k} \leq (k+1)^{-(1+\gamma/2)}.$$

Thus,

$$\mathbb{E}\left[\|y_k - y_{k,*}\|_{H_*} \mid \mathcal{F}_{k,0}\right] \leq (k+1)^{-(1+\gamma/2)} \|y_{k,0} - y_{k,*}\|_{H_*} \tag{68}$$

whenever $x_k \in B(x_*)$ and k is sufficiently large.

We next justify applying this estimate at iteration $k+1$ while conditioning on $\mathcal{F}_k$. By (26), (22), and $\alpha_k \leq 1$,

$$\begin{aligned} \|x_{k+1} - x_*\|_{H_*} &\leq \|x_k - x_*\|_{H_*} + \sqrt{\lambda_{\max}(H_*)} \|d_k\| \\ &\leq \left(1 + \frac{L}{\mu} \sqrt{\frac{\lambda_{\max}(H_*)}{\lambda_{\min}(H_*)}}\right) \|x_k - x_*\|_{H_*}. \end{aligned}$$

Consequently, after shrinking $B(x_*)$ if necessary, we may ensure that whenever $x_k \in B(x_*)$, the next point x_{k+1} remains in the neighborhood on which (67), Lemma 3.2, and Lemma 3.3 are valid. This choice is deterministic and does not depend on the realization of the random sketches.

Hence, whenever $x_k \in B(x_*)$ and k is sufficiently large, (68) can be applied at iteration $k+1$. Since

$$\mathcal{F}_k \subseteq \mathcal{F}_{k+1,0},$$

the tower property gives

$$\begin{aligned} \mathbb{E}\left[\|y_{k+1} - y_{k+1,*}\|_{H_*} \mid \mathcal{F}_k\right] &= \mathbb{E}\left[\mathbb{E}\left[\|y_{k+1} - y_{k+1,*}\|_{H_*} \mid \mathcal{F}_{k+1,0}\right] | \mathcal{F}_k\right] \\ &\leq (k+2)^{-(1+\gamma/2)} \mathbb{E}\left[\|y_{k+1,0} - y_{k+1,*}\|_{H_*} \mid \mathcal{F}_k\right]. \end{aligned}$$

By Lemma 3.3,

$$\begin{aligned} &\mathbb{E}\left[\|y_{k+1,0} - y_{k+1,*}\|_{H_*} \mid \mathcal{F}_k\right] \\ &\leq \left(1 + \sqrt{\kappa(H_*)} + O(\|x_k - x_*\|_{H_*}^{\delta})\right) \|y_k - y_{k,*}\|_{H_*} + \frac{2 + \sqrt{\kappa(H_*)}}{\sqrt{\lambda_{\min}(H_*)}} + O(\|x_k - x_*\|_{H_*}). \end{aligned}$$

Since all the $O(\cdot)$ terms above are uniform local estimates, after shrinking $B(x_*)$ if necessary, there exists a deterministic constant $C > 0$ such that, whenever $x_k \in B(x_*)$ and k is sufficiently large,

$$\mathbb{E}\left[\|y_{k+1} - y_{k+1,*}\|_{H_*} \mid \mathcal{F}_k\right] \le C(k+2)^{-(1+\gamma/2)}\|y_k - y_{k,*}\|_{H_*} + C(k+2)^{-(1+\gamma/2)}. \quad (69)$$

Since

$$\sum_{k=0}^{\infty}(k+2)^{-(1+\gamma/2)} < \infty$$

and

$$C(k+2)^{-(1+\gamma/2)} \longrightarrow 0,$$

there exists a deterministic integer such that, from that index onward,

$$C(k+2)^{-(1+\gamma/2)} \le \frac{1}{2}. \quad (70)$$

It remains to take care of the fact that the iteration from which x_k remains in $B(x_*)$ is sample-path dependent. Fix any deterministic integer m sufficiently large that (66) and (70) hold for every $k \ge m$, and define

$$\tau_m \coloneqq \inf\{j \ge m : x_j \notin B(x_*)\},$$

with $\inf \emptyset = \infty$.

For $k \ge m$, (69) and (70) imply

$$\mathbb{E}\left[\frac{\|y_{(k+1)\wedge\tau_m} - y_{(k+1)\wedge\tau_m,*}\|_{H_*}}{1+\|y_m - y_{m,*}\|_{H_*}}\middle|\mathcal{F}_k\right] \le \frac{\|y_{k\wedge\tau_m} - y_{k\wedge\tau_m,*}\|_{H_*}}{1+\|y_m - y_{m,*}\|_{H_*}} - \frac{1}{2}\mathbf{1}_{\{k<\tau_m\}}\frac{\|y_k - y_{k,*}\|_{H_*}}{1+\|y_m - y_{m,*}\|_{H_*}} + C(k+2)^{-(1+\gamma/2)}. \quad (71)$$

Indeed, on $\{\tau_m \le k\}$ the two stopped terms are identical, whereas on $\{k < \tau_m\}$, (69) applies.

The normalized stopped process is nonnegative and its initial value is at most one. Moreover,

$$\sum_{k=m}^{\infty} C(k+2)^{-(1+\gamma/2)} < \infty.$$

Therefore, the Robbins–Siegmund theorem applied to (71) gives

$$\sum_{k=m}^{\infty}\mathbf{1}_{\{k<\tau_m\}}\frac{\|y_k - y_{k,*}\|_{H_*}}{1+\|y_m - y_{m,*}\|_{H_*}} < \infty \qquad \text{a.s.}$$

Now Assumption 3.2 gives

$$\|x_k - x_*\|_{H_*} \longrightarrow 0 \qquad \text{a.s.}$$

Therefore,

$$\mathbb{P}\left(\bigcup_{m=0}^{\infty}\{\tau_m = \infty\}\right) = 1.$$

Since there are only countably many deterministic integers m, the preceding summability result may be taken to hold simultaneously for all such m on an event of probability one.

On this probability-one event, choose an m for which $\tau_m = \infty$. Then

$$\sum_{k=m}^{\infty}\|y_k - y_{k,*}\|_{H_*} < \infty. \quad (72)$$

Consequently,

$$\|y_k - y_{k,*}\|_{H_*} \longrightarrow 0 \qquad (k \to \infty) \quad \text{a.s.}$$

Finally, choosing $\gamma = 1$ in (66) gives

$$T_k \geq \left\lceil \frac{4}{-\log(1-\lambda_R)} \log(k+1) \right\rceil,$$

which is exactly the condition stated in the proposition. □

We are now ready to prove the local superlinear convergence result stated in the main text. The proof combines the local linear contraction of the outer iterates, the anchor-dependent gradient estimate in Proposition 3.1, and the anchor-tracking mechanism established above.

Proof. We prove the theorem in three steps. First, using the local linear contraction together with a stopping-time argument, we show that

$$\sum_{k=0}^{\infty} \|x_k - x_*\|_{H_*}^{\delta} < \infty \qquad \text{a.s.}$$

Second, we use the tail-summability result established in the proof of Proposition 3.2. Finally, these two summability properties imply that the conditional probabilities of nonvanishing gradient ratios are summable. Conditional Markov's inequality and the conditional Borel–Cantelli lemma then yield

$$\frac{\|g_{k+1}\|}{\|g_k\|} \to 0 \qquad \text{a.s.},$$

which, by the local equivalence between $\|g_k\|$ and $\|x_k - x_*\|_{H_*}$, gives the desired superlinear convergence.

The condition on T_k in Theorem 3.2 corresponds to the choice $\gamma = 1$ in the more general condition below. We prove the result for an arbitrary fixed $\gamma > 0$, and recover the theorem by setting $\gamma = 1$ at the end.

Suppose that, for all sufficiently large k,

$$T_k \geq \left\lceil \frac{2(1+\gamma)}{-\log(1-\lambda_R)} \log(k+1) \right\rceil. \tag{73}$$

Then

$$(1-\lambda_R)^{T_k/2} \leq (k+1)^{-(1+\gamma)},$$

and therefore

$$\sum_{k=0}^{\infty} (1-\lambda_R)^{T_k/2} < \infty.$$

Step 1: Summability of the outer iterate errors.

By (81),

$$\lambda_k^P \geq \lambda_{\min}\Big(\mathbb{E}\Big[\Pi(H_*^{1/2} S_k^\top)\Big]\Big).$$

Since $H_* \succ 0$ and S_k is Gaussian,

$$\lambda_{\min}\Big(\mathbb{E}\Big[\Pi(H_*^{1/2} S_k^\top)\Big]\Big) > 0.$$

Hence, choose a deterministic constant $\rho_x \in (0,1)$ satisfying

$$\sqrt{1 - \lambda_{\min}\Big(\mathbb{E}\Big[\Pi(H_*^{1/2} S_k^\top)\Big]\Big)} < \rho_x < 1.$$

The $O(\|x_k - x_*\|_{H_*}^{1+\delta})$ term in (9) is a local estimate with a deterministic constant. Therefore, there exist deterministic constants $C_x > 0$ and $r_x > 0$ such that, whenever

$$\|x_k - x_*\|_{H_*} \le r_x,$$

we have

$$\mathbb{E}\left[\|x_{k+1} - x_*\|_{H_*} \,|\, \mathcal{F}_k\right] \le \sqrt{1 - \lambda_{\min}\Big(\mathbb{E}\Big[\Pi(H_*^{1/2} S_k^\top)\Big]\Big)}\|x_k - x_*\|_{H_*} + C_x\|x_k - x_*\|_{H_*}^{1+\delta}.$$

Shrinking r_x if necessary, we may assume

$$\sqrt{1 - \lambda_{\min}\Big(\mathbb{E}\Big[\Pi(H_*^{1/2} S_k^\top)\Big]\Big)} + C_x r_x^\delta \le \rho_x.$$

Thus, whenever $\|x_k - x_*\|_{H_*} \le r_x$,

$$\mathbb{E}\left[\|x_{k+1} - x_*\|_{H_*} \,|\, \mathcal{F}_k\right] \le \rho_x\|x_k - x_*\|_{H_*}.$$

Since $\delta \in (0, 1]$, conditional Jensen's inequality gives

$$\begin{aligned}\mathbb{E}\left[\|x_{k+1} - x_*\|_{H_*}^\delta \,\middle|\, \mathcal{F}_k\right] &\le \left(\mathbb{E}\left[\|x_{k+1} - x_*\|_{H_*} \,|\, \mathcal{F}_k\right]\right)^\delta \\ &\le \rho_x^\delta \|x_k - x_*\|_{H_*}^\delta \end{aligned} \tag{74}$$

whenever $\|x_k - x_*\|_{H_*} \le r_x$.

The difficulty is that the iteration at which the sequence eventually enters this neighborhood is sample-path dependent. We therefore localize the argument using stopping times. For each deterministic integer $m \ge 0$, define

$$\tau_m := \inf\left\{j \ge m : \|x_j - x_*\|_{H_*} > r_x\right\},$$

with $\inf \emptyset = \infty$.

For $k \ge m$, (74) implies

$$\begin{aligned}&\mathbb{E}\left[\|x_{(k+1)\wedge\tau_m} - x_*\|_{H_*}^\delta \,\middle|\, \mathcal{F}_k\right] \\ &\le \|x_{k\wedge\tau_m} - x_*\|_{H_*}^\delta - (1 - \rho_x^\delta)\mathbf{1}_{\{k<\tau_m\}}\|x_k - x_*\|_{H_*}^\delta.\end{aligned} \tag{75}$$

Indeed, on $\{\tau_m \le k\}$ both stopped variables are identical, whereas on $\{k < \tau_m\}$, (74) applies.

Applying the Robbins–Siegmund theorem to (75), for every deterministic m,

$$\sum_{k=m}^{\infty} \mathbf{1}_{\{k<\tau_m\}}\|x_k - x_*\|_{H_*}^\delta < \infty \qquad \text{a.s.} \tag{76}$$

Now Assumption 3.2 gives

$$\|x_k - x_*\|_{H_*} \longrightarrow 0 \qquad \text{a.s.}$$

Therefore,

$$\mathbb{P}\left(\bigcup_{m=0}^{\infty}\{\tau_m = \infty\}\right) = 1.$$

On $\{\tau_m = \infty\}$, (76) becomes

$$\sum_{k=m}^{\infty} \|x_k - x_*\|_{H_*}^\delta < \infty.$$

Since only finitely many terms precede m, we conclude that

$$\sum_{k=0}^{\infty} \|x_k - x_*\|_{H_*}^\delta < \infty \qquad \text{a.s.} \tag{77}$$

Step 2: Summability of the anchor errors.

Under (73), the proof of Proposition 3.2 establishes the stronger intermediate result

$$\sum_{k=m}^{\infty} \|y_k - y_{k,*}\|_{H_*} < \infty$$

almost surely for some finite index m; see (72). Since only finitely many terms precede m, it follows that

$$\sum_{k=0}^{\infty} \|y_k - y_{k,*}\|_{H_*} < \infty \qquad \text{a.s.} \tag{78}$$

Step 3: From summability to superlinear convergence.

We now apply Proposition 3.1. If $g_k = 0$ at some iteration, the algorithm has already reached a stationary point. Hence it suffices to consider the case $\|g_k\| > 0$. Dividing (10) by $\|g_k\|$ gives

$$\begin{aligned}&\mathbb{E}\left[\frac{\|g_{k+1}\|}{\|g_k\|}\,\middle|\,\mathcal{F}_k\right]\\ &\le \left(\sqrt{\lambda_{\max}(H_*)(1-\lambda_k^P)} + O(\|g_k\|)\right)\|y_k - y_{k,*}\|_{H_*} + O(\|g_k\|^\delta).\end{aligned} \tag{79}$$

Since $0 \le \lambda_k^P \le 1$ and $\|g_k\| \to 0$, the coefficient multiplying $\|y_k - y_{k,*}\|_{H_*}$ is almost surely bounded for all sufficiently large k. Furthermore, by (22),

$$\|g_k\|^\delta = O(\|x_k - x_*\|_{H_*}^\delta).$$

Therefore, by (77) and (78), the right-hand side of (79) is summable along the tail almost surely.

For any $\varepsilon > 0$, conditional Markov's inequality therefore gives, for all sufficiently large k,

$$\mathbb{P}\left(\frac{\|g_{k+1}\|}{\|g_k\|} > \varepsilon \,\middle|\, \mathcal{F}_k\right) \le \frac{1}{\varepsilon}\mathbb{E}\left[\frac{\|g_{k+1}\|}{\|g_k\|}\,\middle|\,\mathcal{F}_k\right].$$

Consequently,

$$\sum_{k=0}^{\infty} \mathbb{P}\left(\frac{\|g_{k+1}\|}{\|g_k\|} > \varepsilon \,\middle|\, \mathcal{F}_k\right) < \infty \qquad \text{a.s.},$$

because the tail is summable by (77) and (78), while the finitely many preceding conditional probabilities are bounded by one.

Hence, by the conditional Borel–Cantelli lemma, for every $\varepsilon > 0$,

$$\frac{\|g_{k+1}\|}{\|g_k\|} > \varepsilon$$

occurs only finitely often almost surely. Taking a countable sequence of positive ε decreasing to zero, we obtain

$$\frac{\|g_{k+1}\|}{\|g_k\|} \longrightarrow 0 \qquad (k \to \infty) \quad \text{a.s.} \tag{80}$$

Finally, by (22), there exist constants $c_1, c_2 > 0$ such that, for all sufficiently large k,

$$c_1\|x_k - x_*\|_{H_*} \le \|g_k\| \le c_2\|x_k - x_*\|_{H_*}.$$

Therefore,

$$\frac{\|x_{k+1} - x_*\|_{H_*}}{\|x_k - x_*\|_{H_*}} \le \frac{c_2}{c_1}\frac{\|g_{k+1}\|}{\|g_k\|}.$$

Combining this with (80),

$$\frac{\|x_{k+1} - x_*\|_{H_*}}{\|x_k - x_*\|_{H_*}} \longrightarrow 0 \qquad (k \to \infty) \quad \text{a.s.}$$

This proves (11).

Finally, taking $\gamma = 1$ in (73) gives

$$T_k \geq \left\lceil \frac{4}{-\log(1 - \lambda_R)} \log(k+1) \right\rceil,$$

This completes the proof. □

D.3 Evaluating the local-rate projection constants

We collect existing estimates for the smallest eigenvalues of the expected projection matrices that appear in the local convergence rate. Throughout this subsection, let

$$\lambda_{\max}(H_*) = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n = \lambda_{\min}(H_*) > 0$$

be the eigenvalues of H_*. We denote the row dimension of S_k by s, and the row dimension of the inner Gaussian sketch R by r.

We first relate the projection constant for the anchored random matrix P_k to that for its Gaussian block S_k. Since

$$P_k = \begin{bmatrix} y_k^\top / \|y_k\| \\ S_k \end{bmatrix},$$

we have

$$\operatorname{range}(H_*^{1/2} S_k^\top) \subseteq \operatorname{range}(H_*^{1/2} P_k^\top).$$

Therefore, the orthogonal projection onto $\operatorname{range}(H_*^{1/2} P_k^\top)$ dominates the orthogonal projection onto $\operatorname{range}(H_*^{1/2} S_k^\top)$ in the Loewner order:

$$\Pi(H_*^{1/2} P_k^\top) \succeq \Pi(H_*^{1/2} S_k^\top).$$

Taking conditional expectation with respect to $\mathcal{F}_k$, and using the independence of S_k from $\mathcal{F}_k$, gives

$$\begin{aligned} \mathbb{E}\Big[\Pi(H_*^{1/2} P_k^\top) \,\Big|\, \mathcal{F}_k\Big] &\succeq \mathbb{E}\Big[\Pi(H_*^{1/2} S_k^\top) \,\Big|\, \mathcal{F}_k\Big] \\ &= \mathbb{E}\Big[\Pi(H_*^{1/2} S_k^\top)\Big]. \end{aligned}$$

Hence,

$$\lambda_{\min}\Big(\mathbb{E}\Big[\Pi(H_*^{1/2} P_k^\top) \,\Big|\, \mathcal{F}_k\Big]\Big) \geq \lambda_{\min}\Big(\mathbb{E}\Big[\Pi(H_*^{1/2} S_k^\top)\Big]\Big). \tag{81}$$

Thus, the anchored projection constant is at least as large as the Gaussian projection constant associated with S_k.

We next recall the classical convergence-rate estimate for randomized projection methods. For the randomized Kaczmarz method applied to a consistent linear system $Az = b$, the standard mean-square contraction is controlled by the scaled condition number

$$\frac{\sigma_{\min}^2(A)}{\|A\|_F^2},$$

namely,

$$\mathbb{E}\|z_t - z_*\|^2 \leq \left(1 - \frac{\sigma_{\min}^2(A)}{\|A\|_F^2}\right)^t \|z_0 - z_*\|^2$$

under the usual row-norm sampling rule Strohmer & Vershynin (2009). This rate is also recovered as a special case of the more general sketch-and-project framework Gower & Richtárik (2015); Dereziński & Rebrova (2024).

If we apply this classical scaled-condition-number viewpoint with $A = H_*^{1/2}$, then

$$\sigma_{\min}^2(H_*^{1/2}) = \lambda_{\min}(H_*), \qquad \|H_*^{1/2}\|_F^2 = \operatorname{tr}(H_*).$$

Therefore, the classical rate constant becomes

$$\frac{\sigma_{\min}^2(H_*^{1/2})}{\|H_*^{1/2}\|_F^2} = \frac{\lambda_{\min}(H_*)}{\operatorname{tr}(H_*)}.$$

For a block sketch of row dimension q, this trace-based scaling gives the baseline quantity

$$\frac{q\lambda_{\min}(H_*)}{\operatorname{tr}(H_*)}, \qquad q \in \{s, r\}.$$

This estimate should not be confused with its worst-case condition-number lower bound. Indeed, since

$$\operatorname{tr}(H_*) = \sum_{i=1}^{n} \lambda_i \le n\lambda_{\max}(H_*),$$

we have

$$\frac{q\lambda_{\min}(H_*)}{\operatorname{tr}(H_*)} \ge \frac{q}{n\kappa(H_*)}.$$

However, the right-hand side is only a pessimistic floor. The ratio between the trace-based estimate and this condition-number bound is

$$\frac{q\lambda_{\min}(H_*)/\operatorname{tr}(H_*)}{q/(n\kappa(H_*))} = \frac{n\lambda_{\max}(H_*)}{\operatorname{tr}(H_*)}.$$

This ratio can be much larger than one whenever the trace is substantially smaller than $n\lambda_{\max}(H_*)$. Thus, even the classical trace-based estimate can be significantly stronger than the crude condition-number estimate.

The trace-based estimate is already more informative than the crude condition-number floor, but it still compresses the spectrum of H_* into only $\lambda_{\min}(H_*)$ and $\operatorname{tr}(H_*)$. It therefore cannot fully exploit finer spectral structure, such as rapid spectral decay or a small randomized SVD residual tail. We therefore turn to sharper Gaussian estimates for expected sketched projection matrices.

Let $G_q \in \mathbb{R}^{q\times n}$ be a Gaussian sketch, where $q \in \{s, r\}$. The precise variance normalization of G_q is irrelevant here, because multiplying G_q by a nonzero scalar does not change the subspace $\operatorname{range}(H_*^{1/2}G_q^\top)$. Applying the sharp Gaussian projection estimate of (Dereziński & Rebrova, 2024, Theorem 3.1) with $A = H_*^{1/2}$ gives

$$\lambda_{\min}\Big(\mathbb{E}\Big[\Pi(H_*^{1/2}G_q^\top)\Big]\Big) \ge (1-\varepsilon_{n,q})\frac{q\lambda_{\min}(H_*)}{\operatorname{Err}(H_*^{1/2}, q-1)}, \tag{82}$$

where

$$\varepsilon_{n,q} \le \frac{4q}{n} + \frac{8\log(3n)}{n}.$$

Here

$$\operatorname{Err}(H_*^{1/2}, \ell) := \mathbb{E}_{\Omega_\ell}\left[\left\|H_*^{1/2}\big(I_n - \Pi(H_*^{1/2}\Omega_\ell^\top)\big)\right\|_F^2\right], \tag{83}$$

where $\Omega_\ell \in \mathbb{R}^{\ell\times n}$ is a Gaussian sketch. This quantity is the randomized SVD residual error: it measures the part of $H_*^{1/2}$ that remains after projecting onto the Gaussian sketched subspace Dereziński & Rebrova (2024); Halko et al. (2011).

Specializing (82) to $G_q = S_k$ and $G_q = R$, we obtain

$$\lambda_{\min}\Big(\mathbb{E}\Big[\Pi(H_*^{1/2}S_k^\top)\Big]\Big) \ge (1-\varepsilon_{n,s})\frac{s\lambda_{\min}(H_*)}{\operatorname{Err}(H_*^{1/2}, s-1)}, \tag{84}$$

$$\lambda_{\min}\Big(\mathbb{E}\Big[\Pi(H_*^{1/2}R^\top)\Big]\Big) \ge (1-\varepsilon_{n,r})\frac{r\lambda_{\min}(H_*)}{\operatorname{Err}(H_*^{1/2}, r-1)}. \tag{85}$$

Unlike the classical trace-based estimate, these bounds depend on the randomized SVD residual errors $\mathrm{Err}(H_*^{1/2}, s-1)$ and $\mathrm{Err}(H_*^{1/2}, r-1)$, rather than on the full trace $\mathrm{tr}(H_*)$. This is the key connection between sketch-and-project convergence rates and randomized SVD errors established in Dereziński & Rebrova (2024). It is also consistent with the full-spectrum analysis of expected random projection matrices in Derezinski et al. (2020) and the randomized low-rank approximation framework of Halko et al. (2011).

More concretely, the randomized SVD residual error admits a spectral-tail bound. By (Dereziński & Rebrova, 2024, Lemma 3.3), which follows from the standard randomized low-rank approximation analysis of (Halko et al., 2011, Theorem 10.5), if $q \geq 5$ and $2 \leq p \leq q-3$, then

$$\mathrm{Err}(H_*^{1/2}, q-1) \leq \frac{q-2}{p-1} \sum_{i \geq q-1-p} \lambda_i.$$

Indeed, this is obtained by applying the bound to $A = H_*^{1/2}$, whose squared singular values are the eigenvalues λ_i of H_*. Thus, the denominator in the sharp estimate is controlled by a residual spectral tail of H_*, rather than by the full trace $\sum_{i=1}^n \lambda_i$.

Consequently, the sharp estimates become favorable when the residual spectral tail is small. For $A = H_*^{1/2}$, the squared singular values of A are precisely the eigenvalues of H_*. Therefore, the polynomial and exponential spectral-decay results of (Dereziński & Rebrova, 2024, Corollary 3.4) translate into corresponding decay assumptions on $\{\lambda_i\}_{i=1}^n$. Moreover, the flat-tailed case of (Dereziński & Rebrova, 2024, Corollary 3.5) applies when the eigenvalues after the leading directions are comparable to $\lambda_{\min}(H_*)$, with a sketch dimension sufficiently large relative to the number of leading directions. In these regimes, the bounds in (84)–(85) can be much stronger than what is suggested by the classical condition-number-based viewpoint.

Combining (81) with (84), we finally obtain the lower bound relevant to the anchored subspace:

$$\lambda_{\min}\Big(\mathbb{E}\Big[\Pi(H_*^{1/2}P_k^\top)\,\Big|\,\mathcal{F}_k\Big]\Big) \geq (1-\varepsilon_{n,s})\frac{s\lambda_{\min}(H_*)}{\mathrm{Err}(H_*^{1/2}, s-1)}.$$

Thus, the local-rate projection constant for the anchored random matrix is at least as large as the Gaussian projection constant of its random block. Moreover, this constant can be evaluated through randomized SVD residual errors, rather than only through the trace-based estimate or its much more pessimistic condition-number lower bound.

E Examples of subproblems that satisfy Condition 2.1

In this section, we present concrete implementations of the regularized update direction in ARS-RN. We also explain how the conditions imposed on the regularization coefficient η_k in the global and local convergence analyses are satisfied, or what additional assumptions are required.

Some of the arguments below use an operator-norm bound for

$$P_k = \begin{bmatrix} y_k^\top/\|y_k\| \\ S_k \end{bmatrix},$$

where S_k is Gaussian. Such a bound is not deterministic in general. Therefore, whenever a proof uses the estimate

$$\|P_k\|^2 \leq \bar{C}\frac{n}{s},$$

the corresponding conclusion is to be understood on this high-probability event. By Lemma C.4, this event holds, for each fixed iteration k, with probability at least $1 - 2\exp(-s)$. Claims whose proofs do not use this bound remain deterministic under their stated assumptions.

The convergence theorems are stated under Condition 2.1. For the concrete ARS-2RN and ARS-3RN choices below, this condition is verified on the same high-probability operator-norm event used in the global analysis.

E.1 ARS-2RN with Quadratic Regularizer

Following the RSRNM of Fuji et al. (2025), we use a quadratic regularization parameter determined by the negative curvature of the projected Hessian and the gradient norm. For the anchored random matrix P_k, define

$$\Lambda_k \coloneqq \max\{0, -\lambda_{\min}(P_k H_k P_k^\top)\}.$$

Then, for parameters $c_1 > 1$, $c_2 > 0$, and $\delta \in (0, 1/2]$, we set

$$\eta_k \coloneqq c_1 \Lambda_k + c_2 \|g_k\|^\delta. \tag{86}$$

In our proposed Algorithm 1, we use this η_k in the solution formula (6) for the subproblem, and call the resulting algorithm ARS-2RN. We show that the η_k in (86) satisfies Condition 2.1(a) under Assumption 3.1 for the global convergence analysis, and Condition 2.1(b) under Assumption 3.2 for the local convergence analysis.

E.1.1 Checking Condition 2.1(a) under Assumption 3.1

First, for the global convergence analysis, we show that η_k can be uniformly controlled by d_k and the boundedness of $\nabla^2 f$.

Lemma E.1. *Under Assumption 3.1, for any iteration k with $g_k \neq 0$, $P_k H_k P_k^\top + c_1 \Lambda_k I_{s+1} \succeq 0$ holds, and in particular*

$$P_k H_k P_k^\top + \eta_k I_{s+1} \succeq c_2 \|g_k\|^\delta I_{s+1} \succ 0. \tag{87}$$

Hence, the subproblem has a unique solution, and

$$u_k = -(P_k H_k P_k^\top + \eta_k I_{s+1})^{-1} P_k g_k, \qquad d_k = -P_k^\top (P_k H_k P_k^\top + \eta_k I_{s+1})^{-1} P_k g_k \tag{88}$$

hold.

Proof. By $\Lambda_k = \max\{0, -\lambda_{\min}(P_k H_k P_k^\top)\}$, we have $\lambda_{\min}(P_k H_k P_k^\top) + \Lambda_k \geq 0$. Thus, since $c_1 > 1$,

$$\lambda_{\min}(P_k H_k P_k^\top + c_1 \Lambda_k I_{s+1}) = \lambda_{\min}(P_k H_k P_k^\top) + c_1 \Lambda_k \geq \lambda_{\min}(P_k H_k P_k^\top) + \Lambda_k \geq 0,$$

that is, $P_k H_k P_k^\top + c_1 \Lambda_k I_{s+1} \succeq 0$. Therefore,

$$P_k H_k P_k^\top + \eta_k I_{s+1} = (P_k H_k P_k^\top + c_1 \Lambda_k I_{s+1}) + c_2 \|g_k\|^\delta I_{s+1} \succeq c_2 \|g_k\|^\delta I_{s+1}.$$

If $g_k \neq 0$, then $c_2 \|g_k\|^\delta > 0$, and so (87) follows. Finally, the first-order optimality condition

$$P_k g_k + (P_k H_k P_k^\top + \eta_k I_{s+1}) u_k = 0$$

gives (88). □

Next, we establish Condition 2.1(a), namely upper and lower bounds on η_k.

Proposition E.1. *Suppose that Assumption 3.1 holds and that $\delta \in (0, 1/2]$. Then, for each fixed iteration k with $g_k \neq 0$, with probability at least $1 - 2\exp(-s)$, there exist constants $c_\eta, \bar{\eta} > 0$, independent of k, such that*

$$c_\eta \|d_k\| \leq \eta_k \leq \bar{\eta}.$$

Hence, Condition 2.1(a) holds with probability at least $1 - 2\exp(-s)$ for each fixed iteration k.

Proof. By Assumption 3.1, at every iteration k,

$$\|g_k\| \leq U_g, \qquad \|H_k\| \leq U_H$$

holds. We argue below on the event $\mathcal{E}_k^{P,\mathrm{OP}}$, that is, $\|P_k\|^2 \leq \bar{\mathcal{C}} \frac{n}{s}$.

(Upper bound) First,

$$\Lambda_k \coloneqq \max\{0, -\lambda_{\min}(P_k H_k P_k^\top)\} \leq \|P_k H_k P_k^\top\|$$

and

$$\|P_k H_k P_k^\top\| \le \|P_k\|^2 \|H_k\| \le \bar{\mathcal{C}} \frac{n}{s} U_H \qquad (\text{on } \mathcal{E}_k^{P,\mathrm{OP}})$$

imply

$$\Lambda_k \le \bar{\mathcal{C}} \frac{n}{s} U_H.$$

Therefore,

$$\eta_k = c_1 \Lambda_k + c_2 \|g_k\|^\delta \le c_1 \bar{\mathcal{C}} \frac{n}{s} U_H + c_2 U_g^\delta =: \bar{\eta}.$$

(Lower bound) By Lemma E.1 and (88),

$$\|u_k\| \le \|(P_k H_k P_k^\top + \eta_k I)^{-1}\| \, \|P_k\| \, \|g_k\|.$$

Here, by the definition of Λ_k, we have $\lambda_{\min}(P_k H_k P_k^\top) + \Lambda_k \ge 0$, and therefore

$$\lambda_{\min}(P_k H_k P_k^\top + \eta_k I) = \lambda_{\min}(P_k H_k P_k^\top) + c_1 \Lambda_k + c_2 \|g_k\|^\delta \ge c_2 \|g_k\|^\delta,$$

Hence,

$$\|(P_k H_k P_k^\top + \eta_k I)^{-1}\| = \frac{1}{\lambda_{\min}(P_k H_k P_k^\top + \eta_k I)} \le \frac{1}{c_2 \|g_k\|^\delta}.$$

Therefore,

$$\|u_k\| \le \frac{\|P_k\|}{c_2} \|g_k\|^{1-\delta}, \qquad \|d_k\| = \|P_k^\top u_k\| \le \|P_k\| \, \|u_k\| \le \frac{\|P_k\|^2}{c_2} \|g_k\|^{1-\delta}.$$

Moreover, since $\delta \in (0, 1/2]$, we have $1 - 2\delta \ge 0$, and

$$\|g_k\|^{1-\delta} = \|g_k\|^{1-2\delta} \|g_k\|^\delta \le U_g^{1-2\delta} \|g_k\|^\delta.$$

Hence,

$$\|d_k\| \le \frac{\|P_k\|^2}{c_2} U_g^{1-2\delta} \|g_k\|^\delta \le \frac{\bar{\mathcal{C}}}{c_2} \frac{n}{s} U_g^{1-2\delta} \|g_k\|^\delta \qquad (\text{on } \mathcal{E}_k^{P,\mathrm{OP}}).$$

On the other hand, since $\eta_k \ge c_2 \|g_k\|^\delta$, the above inequality gives

$$\eta_k \ge c_2 \|g_k\|^\delta \ge \frac{c_2^2}{\bar{\mathcal{C}}} \frac{s}{n} U_g^{-(1-2\delta)} \, \|d_k\|.$$

Therefore,

$$c_\eta := \frac{c_2^2}{\bar{\mathcal{C}}} \frac{s}{n} U_g^{-(1-2\delta)}$$

then $c_\eta \|d_k\| \le \eta_k$ holds on $\mathcal{E}_k^{P,\mathrm{OP}}$.

Combining these estimates, on $\mathcal{E}_k^{P,\mathrm{OP}}$,

$$c_\eta \|d_k\| \le \eta_k \le \bar{\eta}$$

holds. Since

$$\Pr(\mathcal{E}_k^{P,\mathrm{OP}}) \ge 1 - 2\exp(-s),$$

the claim follows. □

E.1.2 Checking Condition 2.1(b) under Assumption 3.2

Next, we show the local convergence condition $\eta_k \le C_\eta \|g_k\|^\delta$.

Proposition E.2. *Suppose that Assumption 3.2 holds. Then, for any iteration k such that $x_k \in B(x_*)$, there exists a constant $C_\eta > 0$ such that*

$$\eta_k \le C_\eta \|g_k\|^\delta.$$

Hence, Condition 2.1(b) holds.

Proof. If $x_k \in B(x_*)$, then by (19), $H_k \succeq \mu I_n$, and for any $v \in \mathbb{R}^{s+1} \setminus \{0\}$,

$$v^\top (P_k H_k P_k^\top) v = (P_k^\top v)^\top H_k (P_k^\top v) \ge \mu \|P_k^\top v\|^2 > 0$$

so $P_k H_k P_k^\top \succ 0$, and hence $\lambda_{\min}(P_k H_k P_k^\top) > 0$. By the definition (86) of η_k,

$$\eta_k = c_2 \|g_k\|^\delta.$$

Thus, taking $C_\eta := c_2$ yields $\eta_k \le C_\eta \|g_k\|^\delta$. □

E.2 ARS-3RN WITH A REDUCED-COORDINATE CUBIC REGULARIZER

In this section, motivated by cubic regularization of Newton's method (Nesterov & Polyak, 2006) and its randomized subspace variants (Hanzely et al., 2020; Zhao et al., 2025), we introduce a reduced-coordinate cubic regularizer for ARS-RN and show that the induced η_k satisfies Condition 2.1. Whereas Zhao et al. (2025) consider randomized coordinate subspaces, we use the anchored Gaussian subspace

$$P_k = \begin{bmatrix} y_k^\top / \|y_k\| \\ S_k \end{bmatrix}.$$

For a general non-orthonormal matrix P_k, the reduced-coordinate norm $\|u\|$ does not in general coincide with the ambient-step norm $\|P_k^\top u\|$. In particular, subspace cubic models based on a Taylor upper bound, such as that of Hanzely et al. (2020), regularize the ambient displacement. Here, instead, we deliberately regularize the reduced coordinate u. This choice preserves the scalar diagonal-shift form of the generic ARS-RN update in (6).

Specifically, let u_k be a solution of the cubic subproblem

$$u_k \in \operatorname{Argmin}_{u \in \mathbb{R}^{s+1}} \left\{ g_k^\top P_k^\top u + \frac{1}{2} u^\top (P_k H_k P_k^\top) u + \frac{\widetilde{M}_k}{6} \|u\|^3 \right\}, \tag{89}$$

where $\widetilde{M}_k > 0$ is the reduced-coordinate cubic regularization coefficient. We define the corresponding model by

$$m_k(u) \coloneqq f(x_k) + g_k^\top P_k^\top u + \frac{1}{2} u^\top (P_k H_k P_k^\top) u + \frac{\widetilde{M}_k}{6} \|u\|^3.$$

The optimality condition for (89) is

$$P_k g_k + (P_k H_k P_k^\top) u_k + \frac{\widetilde{M}_k}{2} \|u_k\| \, u_k = 0. \tag{90}$$

Therefore, defining

$$\eta_k \coloneqq \frac{\widetilde{M}_k}{2} \|u_k\|, \tag{91}$$

we obtain

$$(P_k H_k P_k^\top + \eta_k I) u_k = -P_k g_k.$$

Thus, (89) is a concrete cubic regularization design within the generic ARS-RN subproblem (5). We call Algorithm 1 with this choice of η_k ARS-3RN.

The reduced-coordinate form in (89) is important for this connection. Indeed, an ambient-step cubic term $\frac{M_k}{6} \|P_k^\top u\|^3$ would have gradient

$$\frac{M_k}{2} \|P_k^\top u\| P_k P_k^\top u,$$

and would therefore induce a metric regularization proportional to $P_k P_k^\top$, rather than the scalar shift $\eta_k I$ used in (6). Our reduced-coordinate model is therefore not identified with the Taylor upper models used in standard subspace cubic Newton methods. Its role is instead to generate a step-dependent scalar regularization parameter compatible with the common ARS-RN framework. The convergence analysis below does not require m_k to upper-bound $f(x_k + P_k^\top u)$; globalization is handled separately by the Armijo line search in Algorithm 1.

For the analysis below, we directly assume that the reduced-coordinate cubic coefficient is uniformly bounded away from zero and infinity. Namely, there exist constants $0 < \underline{M} \le \overline{M} < \infty$ such that

$$\underline{M} \le \widetilde{M}_k \le \overline{M} \qquad \text{for all relevant iterations } k.$$

This is an algorithmic condition imposed on the present reduced-coordinate variant and is not derived from the iteration-dependent parameter choices in existing SSCN methods. For example, it is satisfied by choosing a fixed positive coefficient, or by safeguarding an adaptive coefficient within a prescribed positive interval.

E.2.1 Checking Condition 2.1(a) under Assumption 3.1

The global upper and lower bounds are obtained using the boundedness of $\|g_k\|$ and $\|H_k\|$, together with a high-probability upper bound on $\|P_k\|$.

Proposition E.3. *Suppose that Assumption 3.1 holds and, moreover, that*

$$0 < \underline{M} \le \widetilde{M}_k \le \overline{M}.$$

Then, for each fixed iteration k, with probability at least $1-2\exp(-s)$, there exist constants $c_\eta, \bar{\eta} > 0$, independent of k, such that

$$c_\eta \|d_k\| \le \eta_k = \frac{\widetilde{M}_k}{2}\|u_k\| \le \bar{\eta}.$$

Hence, the inequalities required in Condition 2.1(a) hold with probability at least $1-2\exp(-s)$ for each fixed iteration k.

Proof. By Assumption 3.1, $\|g_k\| \le U_g$ and $\|H_k\| \le U_H$. On $\mathcal{E}_k^{P,\mathrm{OP}}$, we also have

$$\|P_k\|^2 \le \bar{\mathcal{C}}\frac{n}{s}.$$

By (91) and $\widetilde{M}_k \ge \underline{M}$,

$$\eta_k \ge \frac{\underline{M}}{2}\|u_k\|.$$

On the other hand, since $d_k = P_k^\top u_k$,

$$\begin{aligned}\|d_k\| &\le \|P_k\|\,\|u_k\| \\ &\le \sqrt{\bar{\mathcal{C}}\frac{n}{s}}\,\|u_k\|.\end{aligned}$$

Therefore,

$$\begin{aligned}\eta_k &\ge \frac{\underline{M}}{2}\|u_k\| \\ &\ge \frac{\underline{M}}{2}\sqrt{\frac{s}{\bar{\mathcal{C}}n}}\,\|d_k\| =: c_\eta\|d_k\|.\end{aligned}$$

For convenience, define the cubic objective without the constant term by

$$\widehat{m}_k(u) \coloneqq g_k^\top P_k^\top u + \frac{1}{2}u^\top(P_k H_k P_k^\top)u + \frac{\widetilde{M}_k}{6}\|u\|^3.$$

Since u_k is a minimizer, $\widehat{m}_k(u_k) \le \widehat{m}_k(0) = 0$. Therefore,

$$\begin{aligned}\frac{\widetilde{M}_k}{6}\|u_k\|^3 &\le -g_k^\top P_k^\top u_k - \frac{1}{2}u_k^\top(P_k H_k P_k^\top)u_k \\ &\le \|P_k g_k\|\,\|u_k\| + \frac{1}{2}\|P_k H_k P_k^\top\|\,\|u_k\|^2 \\ &\le \|P_k\|U_g\,\|u_k\| + \frac{1}{2}\|P_k\|^2 U_H\,\|u_k\|^2.\end{aligned}$$

If $u_k \neq 0$, dividing both sides by $\|u_k\|$ and using $\widetilde{M}_k \ge \underline{M}$ gives

$$\frac{\underline{M}}{6}\|u_k\|^2 - \frac{1}{2}\|P_k\|^2 U_H\,\|u_k\| - \|P_k\|U_g \le 0.$$

Substituting $\|P_k\| \le \sqrt{\bar{\mathcal{C}}n/s}$ on $\mathcal{E}_k^{P,\mathrm{OP}}$, we can bound $\|u_k\|$ from above by the positive root of the corresponding quadratic equation. Thus, $\|u_k\| \le C_u$ for a constant C_u independent of k. Hence, by (91) and $\widetilde{M}_k \le \overline{M}$,

$$\eta_k = \frac{\widetilde{M}_k}{2}\|u_k\| \le \frac{\overline{M}}{2}C_u =: \bar{\eta}$$

on $\mathcal{E}_k^{P,\mathrm{OP}}$. If $u_k = 0$, then $d_k = 0$ and $\eta_k = 0$, so $c_\eta\|d_k\| \le \eta_k \le \bar{\eta}$ also holds.

Since

$$\Pr(\mathcal{E}_k^{P,\mathrm{OP}}) \ge 1 - 2\exp(-s),$$

the claim follows. □

E.2.2 Checking Condition 2.1(b) under Assumption 3.2

We show that $\eta_k = O(\|g_k\|^{2/3})$ using $H_k \succeq \mu I$.

Proposition E.4. *Under Assumption 3.2, let $x_k \in B(x_*)$, and assume that $\widetilde{M}_k \le \overline{M}$. Then*

$$\eta_k \le \frac{\overline{M}^{2/3}}{2^{4/3}\mu^{1/3}}\|g_k\|^{2/3}.$$

Hence, Condition 2.1(b) holds with $\delta = \frac{2}{3}$.

Proof. Multiplying both sides of (90) by $u_k^\top$ and using $d_k = P_k^\top u_k$, we obtain

$$d_k^\top g_k + d_k^\top H_k d_k + \frac{\widetilde{M}_k}{2}\|u_k\|^3 = 0.$$

Since $x_k \in B(x_*)$, we have $H_k \succeq \mu I_n$, and hence

$$\begin{aligned}\frac{\widetilde{M}_k}{2}\|u_k\|^3 &= -d_k^\top g_k - d_k^\top H_k d_k \\ &\le \|d_k\|\|g_k\| - \mu\|d_k\|^2 \\ &= -\left(\mu^{1/2}\|d_k\| - \frac{\|g_k\|}{2\mu^{1/2}}\right)^2 + \frac{\|g_k\|^2}{4\mu} \\ &\le \frac{\|g_k\|^2}{4\mu}.\end{aligned}$$

Therefore,

$$\|u_k\| \le \left(\frac{\|g_k\|^2}{2\mu\widetilde{M}_k}\right)^{1/3}.$$

Using (91),

$$\begin{aligned}\eta_k &= \frac{\widetilde{M}_k}{2}\|u_k\| \\ &\le \frac{\widetilde{M}_k^{2/3}}{2^{4/3}\mu^{1/3}}\|g_k\|^{2/3} \\ &\le \frac{\overline{M}^{2/3}}{2^{4/3}\mu^{1/3}}\|g_k\|^{2/3}.\end{aligned}$$

Thus, Condition 2.1(b) holds with $\delta = \frac{2}{3}$. □

F Total-work complexity beyond HVP counts

Hessian-vector product (HVP) counts are a useful implementation-independent proxy for second-order work, but they do not distinguish between derivative information that must be refreshed when the optimization point changes and derivative queries that can reuse the same point-dependent state. This distinction is particularly relevant for ARS-RN. With a fixed inner budget T and $r = s$, the T randomized Kaczmarz blocks and the final anchored outer step are all performed at the same outer point x_k, whereas $T + 1$ consecutive RSRNM steps are performed at $T + 1$ different base points.

The resulting computational advantage is not tied to whether the Hessian is formed explicitly. Explicit-Hessian and matrix-free implementations are two realizations of the same reuse principle.

A unified base-point cost model. Assume that moving the derivative oracle to a new point x requires an irreducible point-dependent setup of cost C_0. After this setup has been performed, let C_s denote the incremental cost of one ordinary s-dimensional random-subspace Newton block at this fixed point. This includes whatever directional or subspace first- and second-order information is actually used by the implementation, together with the associated $s \times s$ reduced linear algebra. No full-gradient or full-Hessian computation is assumed.

Because $r = s$ and the inner and outer Gaussian sketches have the same width, each of the T Kaczmarz refinement blocks has the same leading derivative-query structure as one ordinary s-dimensional RSRNM block. Let Δ_T denote the total cost specific to anchoring beyond these $T+1$ ordinary blocks. It includes, for example, the extra work required to maintain the anchor-Hessian product, the enlargement of the final reduced system from dimension s to $s+1$, and any anchor-specific normalization or scalar information required by the implementation. Under the HVP accounting used in the main text, the dominant second-order contribution to Δ_T is one additional HVP, plus the small s versus $s+1$ reduced-system overhead.

With this notation, $T+1$ RSRNM steps cost

$$(T+1)(C_0 + C_s),$$

whereas one fixed-T ARS-RN outer iteration costs

$$C_0 + (T+1)C_s + \Delta_T.$$

Therefore,

$$\boxed{(T+1)(C_0 + C_s) - \big(C_0 + (T+1)C_s + \Delta_T\big) = TC_0 - \Delta_T.}$$

Hence, whenever one ARS-RN outer step provides optimization progress comparable to $T+1$ RSRNM steps, ARS-RN has strictly lower total work as soon as

$$\boxed{TC_0 > \Delta_T.} \tag{92}$$

This condition is weaker than requiring ARS-RN to use fewer HVPs. In particular, under the HVP accounting of the main text, the two methods perform essentially the same $s(T+1)$ ordinary HVPs over this comparison, while ARS-RN replaces T base-point refreshes by a single refresh and pays only the anchor-specific overhead Δ_T.

Rate-aware comparison. The previous comparison isolates the computational effect of derivative reuse. To combine it with convergence, let $\rho_{\mathrm{RS}} \in (0,1)$ denote a local contraction bound for one RSRNM step and let $\rho_{\mathrm{A}}(T) \in (0,1)$ denote a local contraction bound for one fixed-T ARS-RN outer step. Up to integer-rounding and local-transient terms, the corresponding total-work upper bounds scale as

$$\frac{C_0 + C_s}{-\log \rho_{\mathrm{RS}}} \log \frac{1}{\varepsilon}, \qquad \frac{C_0 + (T+1)C_s + \Delta_T}{-\log \rho_{\mathrm{A}}(T)} \log \frac{1}{\varepsilon}.$$

Thus the anchored method has the smaller local total-work upper bound whenever

$$\frac{C_0 + (T+1)C_s + \Delta_T}{-\log \rho_{\mathrm{A}}(T)} < \frac{C_0 + C_s}{-\log \rho_{\mathrm{RS}}}. \tag{93}$$

In particular, if the fixed-T convergence analysis yields

$$\rho_{\mathrm{A}}(T) \le \rho_{\mathrm{RS}}^{T+1},$$

then (92) is sufficient for a strict total-work advantage. More generally, (93) shows that base-point reuse can compensate for some loss in HVP efficiency: ARS-RN need not have a smaller HVP count in order to have a smaller overall computational complexity.

Explicit-Hessian specialization. When the dimension and sparsity structure permit the Hessian, or an explicitly assembled sparse tangent operator, to be stored, the point-dependent setup may include gradient evaluation and Hessian assembly. In that case, C_0 contains the irreducible cost of refreshing this information at a new iterate, whereas C_s contains the cost of projecting the already available derivative information onto the random subspace and solving the reduced system. Under the current implementation, Δ_T is dominated by one additional Hessian-vector (or Hessian-matrix)

product together with the incremental s versus $s+1$ reduced-system cost. Hence (92) states that ARS-RN is preferable whenever the cost of avoiding T derivative refreshes exceeds this anchor-specific overhead. This regime can yield a substantial advantage when assembly is expensive relative to repeated products with the assembled operator.

The explicit-Hessian case is not intended as the primary large-scale regime. For very high dimensional problems, storing an $n \times n$ Hessian is typically infeasible, which leads to the matrix-free specialization below.

Matrix-free specialization. In a matrix-free implementation, C_0 represents reusable point-dependent state rather than a stored Hessian. Examples include a nonlinear forward state and adjoint state, point-dependent linearized operators, cached quadrature or constitutive data, a simulation state, or a reusable automatic-differentiation linearization. Conditional on this state, C_s is the cost of obtaining only the s-dimensional first-order information and matrix-free second-order actions required by one random-subspace block; neither the full gradient nor the full Hessian needs to be formed.

The important assumption is that the reusable state fits in memory and must be materially refreshed when the nonlinear base point changes. Accordingly, C_0 should be interpreted as the *irreducible incremental cost of moving the derivative oracle from one base point to the next*, after allowing any warm starts, incremental updates, symbolic reuse, or other implementation-level reuse available to RSRNM. Under this interpretation, the same condition (92) applies unchanged. Thus the reuse argument survives even when both methods operate exclusively through subspace gradients and matrix-free Hessian actions.

Conversely, if the derivative oracle is effectively stateless so that $C_0 \approx 0$, this particular advantage disappears. In that regime, any superiority of ARS-RN must come from its convergence behavior rather than derivative reuse.

Representative large-scale problem classes. *PDE-constrained inverse problems and optimal control.* At a new parameter value, one typically solves the forward problem and the corresponding adjoint problem and evaluates the point-dependent linearized operators. Hessian actions in different parameter directions can then be computed matrix-free through incremental forward and adjoint solves using this fixed base-point information. This computational structure is explicit in scalable PDE-constrained inverse-problem solvers such as hIPPYlib (Villa et al., 2021). In this setting, C_0 represents the unavoidable cost of moving the nonlinear state, adjoint, and reusable linearization to a new parameter point, whereas C_s contains the directional incremental solves at that fixed point. ARS-RN amortizes the former over all T inner blocks and the final outer step.

Simulation-based nonlinear least squares and data assimilation. For objectives of the form

$$f(x) = \frac{1}{2}\|F(x) - b\|^2 + \mathcal{R}(x),$$

evaluating $F(x)$ may require a large simulation. At a fixed x, tangent and adjoint models can often be reused to obtain Jacobian-vector, transposed-Jacobian, Gauss–Newton, or Hessian actions in several directions without forming the associated matrices. Global strong convexity is not automatic for nonlinear least squares; the relevant regime here is local, where the reduced Hessian is positive definite as assumed by our local convergence analysis, or where a positive-definite Gauss–Newton model is used.

Nonlinear finite elements with partial assembly. In high-order finite-element computations, explicitly assembling and storing the global tangent matrix can be prohibitively expensive. MFEM's partial-assembly model separates an operator setup phase, which computes and stores quadrature-level data, from repeated operator applications (Anderson et al., 2021; Andrej et al., 2024). For a tangent operator evaluated at a fixed nonlinear state, this setup/apply decomposition allows the same point-dependent data to be reused across several tangent directions. This again matches the decomposition into C_0 and C_s. By contrast, in a fully stateless matrix-free implementation that recomputes all point-dependent information on every application, C_0 may be negligible and this reuse mechanism need not provide an advantage.

Reusable automatic-differentiation or implicit-model linearizations. The same abstraction applies whenever a computational graph, fixed-point model, or implicit layer admits a point-dependent lin-

earization that can be retained and applied repeatedly to different tangent directions. The size of the resulting C_0/C_s ratio is implementation- and memory-dependent, so this should be viewed as a compatible regime rather than as a universal claim about neural-network training.

Scope and interpretation. The explicit and matrix-free specializations are complementary rather than competing assumptions. Both are instances of the same principle: ARS-RN performs several low-dimensional derivative queries at one nonlinear base point, whereas successive RSRNM steps move the base point after every update. When derivative information is explicitly assembled, ARS-RN amortizes assembly. When the computation is matrix-free, it amortizes reusable base-point state.

The claim is therefore strongest for large-scale *nonlinear* optimization problems with expensive reusable base-point state. It does not rely on forming a full gradient or a full Hessian, and it remains meaningful when storing an $n \times n$ Hessian is impossible. It should not be interpreted as a universal runtime advantage: the benefit vanishes when base-point refresh is negligible, when the reusable state cannot be retained, or when RSRNM can update that state essentially for free. A pure quadratic is also not a favorable example for this reuse argument because its Hessian is constant globally and can be shared across all iterates.

For experiments, the most informative quantities to report are therefore not only HVP count and wall-clock time, but also the measured base-point refresh cost, the marginal cost of one s-dimensional derivative block at fixed state, the anchor-specific overhead Δ_T, and the reduced-system cost. These measurements directly test whether the sufficient condition (92) holds in the target application.

G Additional Experimental Details

This appendix provides the experimental details omitted from the main text: stopping criteria, randomness, timing measurements, problem formulations, dataset dimensions, method hyperparameters, and implementation details for the experiments in Section 4.

G.1 Common experimental protocol

Unless otherwise stated, all methods are initialized at $x_0 = 0$ and use

$$\|\nabla f(x_k)\|_2 \le 10^{-5}$$

as the gradient-based termination criterion. Additional stopping rules, the number of random seeds, and the visualization of variability are specified separately for each experimental setting below.

For randomized methods, we run multiple independent random seeds and report the mean curve over seeds. In the original synthetic and real-world experiments, shaded regions indicate the minimum–maximum range across seeds. In the additional sparse real-world experiments, we instead report the mean plus or minus one sample standard deviation over five random seeds, as described in Section G.5.

The reported wall-clock time is measured by the cumulative time inside the optimization loop. This includes sketch construction, anchor/randomized Kaczmarz updates, Hessian-vector product calls, subproblem solves, objective and gradient evaluations, and line-search or accept/reject logic. It excludes data generation, data loading, and the initial logged row.

For the 2RN, RS-2RN, and ARS-2RN variants, we use an Armijo line search consistent with Section 2. The initial stepsize is $\alpha_0 = 1$, the Armijo parameter is $\beta = 10^{-4}$, the shrinkage factor is $\tau = 0.5$, and the maximum number of backtracking iterations is 25. Thus, starting from $\alpha = 1$, we repeatedly replace

$$\alpha \leftarrow \tau\alpha$$

until

$$f(x_k + \alpha d_k) \le f(x_k) + \beta\alpha g_k^\top d_k$$

is satisfied, or until the maximum number of backtracking iterations is reached.

For the 2RN-type methods, we use the same negative-curvature shift notation as in the main text. Let B_k denote the matrix defining the search space:

$$B_k = \begin{cases} I_n, & \text{for 2RN,} \\ S_k, & \text{for RS-2RN,} \\ P_k, & \text{for ARS-2RN.} \end{cases}$$

We define

$$\Lambda_k := \max\{0, -\lambda_{\min}(B_k H_k B_k^\top)\}.$$

The damping parameter is set as

$$\eta_k = 2\Lambda_k + \|g_k\|^{1/2}$$

for 2RN and RS-2RN, and as

$$\eta_k = \max\{10^{-8},\, 2\Lambda_k + \|g_k\|^{1/2}\}$$

for ARS-2RN. This corresponds to the quadratic regularization form used in (86), with $c_1 = 2$, $c_2 = 1$, and $\delta = 1/2$. These values follow the numerical experiments of Fuji et al. (2025), which use the same regularization parameters for RNM and RSRNM. The additional lower bound 10^{-8} for ARS-2RN is only a numerical safeguard for the anchored variant.

For the 3RN, RS-3RN, and ARS-3RN variants, we use an ARC-type accept/reject rule. Given a trial step d_k, let $m_k(d)$ denote the corresponding cubic model and define

$$\rho_k = \frac{f(x_k) - f(x_k + d_k)}{m_k(0) - m_k(d_k)}.$$

The trial step is accepted if $\rho_k \geq \eta_1$ and rejected otherwise:

$$x_{k+1} = \begin{cases} x_k + d_k, & \rho_k \geq \eta_1, \\ x_k, & \rho_k < \eta_1. \end{cases}$$

The cubic regularization parameter is initialized and safeguarded by

$$\sigma_0 = 1, \qquad \sigma_{\min} = 10^{-8}, \qquad \sigma_{\max} = 10^{8},$$

and the ARC update parameters are

$$\eta_1 = 0.05, \qquad \eta_2 = 0.9, \qquad \gamma_1 = 1.5, \qquad \gamma_2 = 2.0.$$

The update of σ_k is

$$\sigma_{k+1} = \begin{cases} \max\{\sigma_{\min},\, \sigma_k/\gamma_2\}, & \rho_k \geq \eta_2, \\ \sigma_k, & \eta_1 \leq \rho_k < \eta_2, \\ \min\{\sigma_{\max},\, \gamma_1\sigma_k\}, & \rho_k < \eta_1. \end{cases}$$

Thus, very successful steps decrease the cubic regularization parameter, unsuccessful steps increase it, and intermediate accepted steps leave it unchanged.

These ARC parameters were chosen to match the public implementation of Zhao et al. (2025).[1] In particular, their implementation uses the same values $\sigma_0 = 1$, $\eta_1 = 0.05$, $\eta_2 = 0.9$, $\gamma_1 = 1.5$, and $\gamma_2 = 2.0$. We keep these parameters fixed for all 3RN-type methods so that the comparison reflects the effect of the subspace construction rather than dataset-specific tuning of the ARC globalization rule. The bounds $\sigma_{\min} = 10^{-8}$ and $\sigma_{\max} = 10^{8}$ are only numerical safeguards. Unlike the SSCN implementation of Zhao et al., which samples coordinate subspaces, our RS-3RN uses Gaussian sketch matrices.

G.2 Problem formulations

Binary logistic regression. For the synthetic logistic regression experiment and the epsilon experiment, we use the binary logistic regression objective

$$\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{m} \sum_{i=1}^{m} \log\left(1 + \exp(-b_i a_i^\top x)\right) + \frac{\lambda}{2}\|x\|^2, \tag{94}$$

where $a_i \in \mathbb{R}^n$, $b_i \in \{-1, 1\}$, and $\lambda \geq 0$.

[1] https://github.com/jydzhao/cubic_subspace_newton

Table 2: Datasets and problem dimensions used in the numerical experiments.

Setting	Dataset	Problem	d	C	m	n	λ
Synthetic	synthetic-logistic	binary logistic	5000	2	25000	5000	10^{-3}
Real-world	flickr_deepwalk	multi-label logistic	129	195	20000	25155	10^{-3}
Real-world	youtube_node2vec	multi-label logistic	129	47	31703	6063	10^{-3}
Real-world	epsilon	binary logistic	2001	2	10000	2001	0
Real-world	usps	softmax	257	10	7291	2570	0
Real-world	real-sim	binary logistic	20958	2	72309	20958	10^{-3}
Real-world	news20	binary logistic	20001	2	19996	20001	10^{-3}
Real-world	rcv1	binary logistic	20001	2	20242	20001	10^{-3}

Multi-label logistic regression. For the flickr_deepwalk and youtube_node2vec datasets, we solve the independent multi-label logistic regression problem

$$\min_W f(W) = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{q}\log\left(1+\exp(-y_{ij}a_i^\top w_j)\right) + \frac{\lambda}{2}\|W\|_F^2, \tag{95}$$

where $a_i \in \mathbb{R}^d$, $y_i \in \{-1,1\}^q$, and w_j is the j-th column of W. The flickr_deepwalk features are associated with DeepWalk graph representations (Perozzi et al., 2014), while the youtube_node2vec features are associated with node2vec graph representations (Grover & Leskovec, 2016).

Softmax regression. For the USPS dataset, we solve the multiclass softmax regression problem

$$\min_W f(W) = \frac{1}{m}\sum_{i=1}^{m}\left[\log\left(\sum_{\ell=1}^{K}\exp(a_i^\top w_\ell)\right) - a_i^\top w_{y_i}\right] + \frac{\lambda}{2}\|W\|_F^2, \tag{96}$$

where $y_i \in \{1,\dots,K\}$ and w_ℓ is the parameter vector for class ℓ. The USPS experiment uses the USPS handwritten digit dataset (Hull, 1994).

G.3 Datasets and problem dimensions

Table 2 summarizes the datasets and problem dimensions used in the experiments. Binary logistic regression uses (94), multi-label logistic regression uses (95), and softmax regression uses (96). The column d denotes the number of input features after any bias feature is appended, C denotes the number of classes or labels, m denotes the number of samples used in the experiment, and n denotes the optimization dimension used by the optimizer.

When a bias feature is appended in preprocessing, the internal implementation dimension can differ from the raw dimension dC. The table reports the dimensions used to describe the plotted problem instances; implementation-specific preprocessing details are fixed across methods within each experiment.

Dataset sources and terms of use. The real-world datasets used in our experiments were downloaded from the LIBSVM dataset repository (Chang & Lin, 2011). The raw files are used only for research evaluation in our numerical experiments and are not redistributed. For datasets whose features are generated from representation learning methods, we also cite the corresponding method papers.

The LIBSVM dataset pages used in our experiments are:

- Binary classification datasets: `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary/`.
- Multi-label datasets: `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multilabel/`.
- Multi-class classification datasets: `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass/`.

- **epsilon.** We used `epsilon_normalized.bz2` from the LIBSVM binary classification datasets. The corresponding LIBSVM page lists the source as PASCAL Challenge 2008.
- **flickr_deepwalk.** We used `flickr_deepwalk.svm.bz2` from the LIBSVM multi-label datasets. Since this file uses DeepWalk features, we cite DeepWalk (Perozzi et al., 2014).
- **youtube_node2vec.** We used `youtube_node2vec.svm.bz2` from the LIBSVM multi-label datasets. Since this file uses node2vec features, we cite node2vec (Grover & Leskovec, 2016).
- **usps.** We used `usps.bz2` from the LIBSVM multi-class classification datasets. We cite the USPS dataset reference (Hull, 1994).

G.4 Synthetic logistic regression

The synthetic experiment uses the binary logistic regression objective (94). We generate a dense Gaussian dataset with $m = 25000$ samples and $n = 5000$ variables. The regularization parameter is set to $\lambda = 10^{-3}$, so the objective is strongly convex. The initial point is $x_0 = 0$.

The purpose of this experiment is to evaluate whether the proposed anchored randomized subspace methods behave consistently with the local theory. We therefore vary the subspace dimension s, the anchor sketch dimension r, and the number T of inner randomized Kaczmarz updates.

G.5 Additional sparse real-world binary logistic regression experiments

We additionally evaluate the methods on three high-dimensional sparse binary-classification datasets: real-sim, news20, and rcv1. All three experiments use the binary logistic regression objective (94) with $\lambda = 10^{-3}$. These experiments evaluate the practical effect of the anchored construction on sparse real-world problems with approximately 2×10^4 optimization variables.

For real-sim, we use all $20{,}958$ input features without adding an intercept, resulting in $m = 72{,}309$ samples and $n = 20{,}958$ optimization variables. For news20 and rcv1, we retain the $20{,}000$ features with the largest document frequencies, with ties broken by the original feature index. This feature-selection procedure does not use the class labels. We then append a constant intercept feature, resulting in $n = 20{,}001$ optimization variables. The news20 and rcv1 experiments use $m = 19{,}996$ and $m = 20{,}242$ samples, respectively. No additional feature standardization or rescaling is applied, and the intercept coefficient is regularized by the same ℓ_2 penalty as the other coefficients.

For the outer randomized subspace, we use coordinate sampling motivated by the block-coordinate sketches employed in the numerical experiments of Gower et al. (2019). At each outer iteration, s coordinate directions are sampled uniformly without replacement from the n coordinates. Thus, RS-2RN and RS-3RN optimize over an s-dimensional random coordinate subspace.

For ARS-3RN, the same s randomly sampled coordinate directions are augmented by a single anchor direction, resulting in an $(s+1)$-dimensional reduced subproblem. The anchor is constructed using independent Gaussian sketches in the inner randomized Kaczmarz procedure: the entries of each inner sketch matrix are sampled independently from $N(0, 1/r)$, where r denotes the inner sketch dimension. Thus, coordinate sampling is used for the outer randomized subspace, whereas Gaussian sketching is used for the inner anchor construction. We use the coordinate-sketching setup of Gower et al. only as motivation for the outer subspace construction; RS-2RN and RS-3RN remain the regularized Newton-type variants considered in this paper rather than implementations of RSN itself.

We compare RS-2RN, RS-3RN, and ARS-3RN. For RS-2RN and RS-3RN, we use

$$s \in \{128, 512, 1024\}.$$

For ARS-3RN, we use

$$s = 128, \qquad r = 64, \qquad T \in \{8, 16, 32\}.$$

Here, s denotes the dimension of the outer random coordinate subspace, whereas r denotes the dimension of the Gaussian sketches used in the inner anchor computation. The value of T is fixed throughout each run. Hence, these experiments evaluate finite-iteration behavior rather than directly testing the asymptotic superlinear-convergence result that requires an increasing inner-iteration schedule T_k.

Hessian–vector product implementation. For these experiments, we exploit the generalized-linear-model structure of binary logistic regression and compute Hessian–vector products analytically without forming the full Hessian. For a block of directions V, we evaluate

$$\nabla^2 f(x)V = \frac{1}{m} A^\top D(x)(AV) + \lambda V,$$

where $D(x)$ is the diagonal matrix containing the second-derivative weights of the logistic loss.

All Hessian–vector products are computed using the full dataset. Thus, the randomness in these experiments arises from the outer coordinate sampling and the Gaussian sketches used for anchor construction, rather than from subsampling the data or approximating the Hessian from a subset of samples. All methods use the same analytic block-HVP implementation.

Stopping criteria and aggregation. All runs are initialized at $x_0 = 0$ and terminate when

$$\|\nabla f(x_k)\|_2 \le 10^{-5}$$

or when the dataset-specific wall-clock limit is reached. The limits are 1200 seconds for real-sim, 600 seconds for news20, and 400 seconds for rcv1. No additional outer-iteration limit is imposed.

For every method configuration, we use random seeds $0, 1, 2, 3, 4$. Let

$$G_k^{(j)} = \|\nabla f(x_k^{(j)})\|_2$$

denote the gradient norm at outer iteration k for seed j. The curves report the mean

$$\bar{G}_k = \frac{1}{5} \sum_{j=0}^{4} G_k^{(j)},$$

and the sample standard deviation is

$$\widehat{\mathrm{sd}}_k = \left(\frac{1}{4} \sum_{j=0}^{4} \left(G_k^{(j)} - \bar{G}_k \right)^2 \right)^{1/2}.$$

The shaded regions therefore represent

$$\bar{G}_k \pm \widehat{\mathrm{sd}}_k.$$

For the time-axis plots, let $t_k^{(j)}$ denote the cumulative optimization time at outer iteration k for seed j. The point corresponding to iteration k is plotted at

$$\bar{t}_k = \frac{1}{5} \sum_{j=0}^{4} t_k^{(j)}.$$

Thus, the time-axis curves are aggregated at a common outer-iteration index rather than by interpolating the individual runs onto a common wall-clock time grid. Only iterations available for all five seeds are included in the aggregated curves.

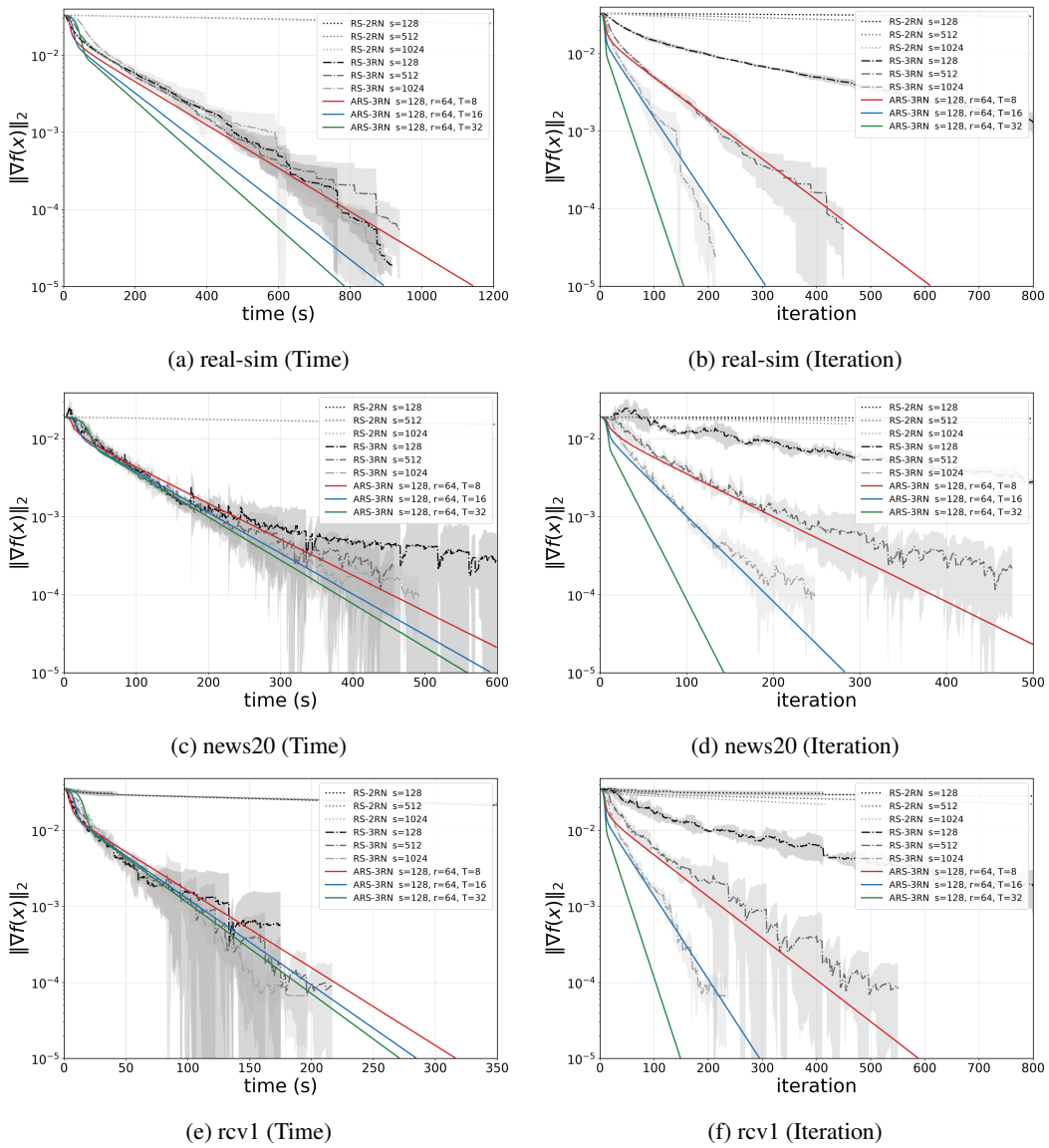
G.6 Method hyperparameters

Table 3 summarizes the method hyperparameters used in the synthetic logistic regression experiment.

G.7 Implementation and compute environment

The synthetic and original real-world experiments were implemented in Python using NumPy, SciPy, Matplotlib, and PyYAML. The recorded software environment used NumPy 2.0.2, SciPy 1.13.1, Matplotlib 3.9.4, and PyYAML 6.0.3.

The synthetic experiments were conducted on a MacBook Pro equipped with an Apple M3 Pro CPU with 12 cores and 18 GB of RAM. The original real-world experiments on flickr_deepwalk and



(a) real-sim (Time)

(b) real-sim (Iteration)

(c) news20 (Time)

(d) news20 (Iteration)

(e) rcv1 (Time)

(f) rcv1 (Iteration)

Figure 3: Additional sparse real-world binary logistic regression experiments with $\lambda = 10^{-3}$. Panels (a)–(b), (c)–(d), and (e)–(f) report results on real-sim, news20, and rcv1, respectively. The left column shows $\|\nabla f(x_k)\|_2$ against cumulative computation time, and the right column shows $\|\nabla f(x_k)\|_2$ against outer iteration. RS-2RN and RS-3RN use random coordinate subspaces with $s \in \{128, 512, 1024\}$, while ARS-3RN uses $s = 128$, $r = 64$, and $T \in \{8, 16, 32\}$. For ARS-3RN, the outer random component is a coordinate subspace and the anchor is constructed using Gaussian sketches. Curves report the mean over five independent random seeds, and shaded regions indicate one sample standard deviation.

epsilon were conducted on the same MacBook Pro, while those on youtube_node2vec and USPS were conducted on an x86-64 workstation equipped with an AMD EPYC 7413 24-Core Processor with 24 CPU cores. Unless otherwise stated, all methods compared within the same experiment were run on the same machine.

The additional sparse real-world experiments on real-sim, news20, and rcv1, described in Section G.5, were implemented separately in Python 3.12.13 using PyTorch 2.13.0. All computations were performed in double precision on CPU, and the runs were executed sequentially using a single thread. The recorded operating system was macOS 26.6.2 on arm64. The sparse experiments use

Table 3: Method hyperparameters for the synthetic logistic regression experiment. The synthetic logistic problem has dimension 5000 and regularization parameter $\lambda = 10^{-3}$.

Dataset	Method	Hyperparameters	Seeds	Solver / globalization
synthetic-logistic	2RN	base	1	Armijo
synthetic-logistic	RS-2RN	$s = 50$	5	Armijo
synthetic-logistic	ARS-2RN	$s = r = 10, \quad T = 100,$ $s = r = 50, \quad T \in \{50, 100, 200\},$ $s = r = 100, \quad T = 100$	5	Armijo
synthetic-logistic	3RN	base	1	ARC, exact solve
synthetic-logistic	RS-3RN	$s = 50$	5	ARC, Lanczos
synthetic-logistic	ARS-3RN	$s = r = 10, \quad T = 100,$ $s = r = 50, \quad T \in \{50, 100, 200\},$ $s = r = 100, \quad T = 100$	5	ARC, Lanczos

Table 4: Method hyperparameters for the real-world experiments.

Dataset	Method	Hyperparameters	Seeds	Solver / globalization
flickr_deepwalk	RS-2RN	$s = 100$	3	Armijo
flickr_deepwalk	RS-3RN	$s = 100$	3	ARC, Lanczos
flickr_deepwalk	ARS-3RN	$s = 100,$ $r \in \{50, 100\},$ $T \in \{50, 100\}$	3	ARC, Lanczos
youtube_node2vec	RS-2RN	$s = 100$	3	Armijo
youtube_node2vec	RS-3RN	$s = 100$	3	ARC, Lanczos
youtube_node2vec	ARS-3RN	$s = 100,$ $r \in \{50, 100\},$ $T \in \{50, 100\}$	3	ARC, Lanczos
epsilon	RS-2RN	$s \in \{100, 200\}$	3	Armijo
epsilon	RS-3RN	$s \in \{100, 200\}$	3	ARC, Lanczos
epsilon	ARS-3RN	$(s, r, T) \in \{(100, 50, 50),$ $(100, 100, 100),$ $(200, 50, 50),$ $(200, 100, 100)\}$	3	ARC, Lanczos
usps	RS-2RN	$s \in \{100, 200\}$	3	Armijo
usps	RS-3RN	$s \in \{100, 200\}$	3	ARC, Lanczos
usps	ARS-3RN	$(s, r, T) \in \{(100, 50, 50),$ $(100, 100, 100),$ $(200, 50, 50),$ $(200, 100, 100)\}$	3	ARC, Lanczos

the analytic block Hessian–vector product implementation described in Section G.5, rather than the NumPy/SciPy implementation used in the preceding experiments.

We also report the computation required for the experiments included in the paper. On our machines, the synthetic logistic experiment required approximately 93 CPU-hours in total for all reported configurations and random seeds. The original real-world experiments required approximately 40 CPU-hours for flickr_deepwalk, 15 CPU-hours for youtube_node2vec, 38 CPU-hours for epsilon, and 50 CPU-hours for USPS.

The additional sparse real-world experiments required approximately 13 hours for real-sim, 7 hours for news20, and 4 hours for rcv1, including all reported method configurations and random seeds. Because these runs were executed sequentially using a single thread, these measured execution times

correspond to approximately 24 additional CPU-hours in total. Combining the original and additional experiments, the total reported computational cost is therefore approximately 260 CPU-hours.

Preliminary experiments, failed runs, HVP-fusion preprocessing benchmarks, data loading, and figure generation are not included in these estimates. No GPU was used.