PRINCIPAL-SUBSPACE NEWTON–GRADIENT METHODS WITH CURVATURE TRACKING AND REUSE

Anonymous authors
Paper under double-blind review

ABSTRACT

A few large Hessian eigenvalues can restrict the stepsize of gradient descent and slow progress in flatter directions, while full Newton steps can be too costly in high dimensions. We analyze principal-subspace Newton–gradient (PSNG) methods, which combine Newton steps in an approximate leading Hessian eigenspace with gradient steps in its orthogonal complement. On strongly convex quadratics, we show that a fixed, sufficiently accurate subspace preserves spectral acceleration: the iteration bound depends on the condition number of the remaining spectrum rather than that of the full Hessian. For objectives with a varying Hessian, we first consider updating the subspace and projected Hessian at every iteration. Under suitable local conditions, warm starts enable linear convergence with a fixed Hessian–vector product (HVP) budget per iteration after initialization. We then propose lazy PSNG, which reuses both quantities until accumulated step length reaches a prescribed threshold. Under corresponding local conditions, lazy PSNG retains linear convergence with a finite total HVP bound independent of the target accuracy, for fixed initialization and parameters. With additional global strong convexity and smoothness, an Armijo phase and a prescribed local restart extend this guarantee to arbitrary starting points. Experiments on quadratics and strongly convex logistic regression illustrate spectral acceleration and reduced HVP costs.

1 INTRODUCTION

First-order methods are inexpensive per iteration, but can converge slowly on ill-conditioned objectives. For a positive-definite Hessian with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d > 0$, large leading eigenvalues restrict the stepsize of gradient descent even when a larger step would be useful in flatter directions (Nesterov, 2004). Newton's method addresses this imbalance by scaling the gradient with curvature, at the cost of solving a potentially large linear system (Nocedal & Wright, 2006; Bottou et al., 2018). When only the leading $r < d$ eigendirections have large curvature, this suggests a less expensive alternative: apply a Newton step in their span and a gradient step in the orthogonal complement. For quadratic objectives, the remaining eigenvalues $\lambda_{r+1}, \ldots, \lambda_d$ determine the stepsize and contraction factor of the gradient update when the leading eigenspace is exact.

Hybrid methods already use this division of work. SONIA combines a second-order update in a sampled curvature subspace with a gradient step in the orthogonal complement (Jahani et al., 2021). FOSI uses approximate extreme Hessian eigenpairs to combine second-order updates with a first-order optimizer (Sivan et al., 2024). For such methods, however, solving a low-dimensional projected linear system does not by itself make the full iteration inexpensive. Computing an accurate eigenspace can require many Hessian–vector products (HVPs) (Saad, 2011; Halko et al., 2011). When the Hessian varies along the optimization trajectory, repeated curvature computation can be costly, while reusing an old subspace and projected Hessian introduces additional error.

We therefore study two questions: *how accurate must the principal subspace be to preserve the convergence benefit, and how long can its curvature information be reused?* We analyze principal-subspace Newton–gradient (PSNG) with two curvature update policies. *Fresh PSNG* updates the subspace and projected Hessian at every iteration. *Lazy PSNG* reuses both until the accumulated step length reaches a fixed threshold. After initialization, both policies start each subspace refresh from the previous basis and refine it by block power iteration.

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.

Contributions. We establish the following convergence guarantees and bounds on HVP counts.

- *Spectral acceleration survives subspace approximation.* On strongly convex quadratics, we give a sufficient accuracy condition under which a fixed, approximate subspace reduces the condition-number dependence of the iteration bound from $\kappa := \lambda_1/\lambda_d$ to $\kappa_\perp := \lambda_{r+1}/\lambda_d$. This can be a large reduction when a few leading eigenvalues cause poor conditioning.
- *Fresh PSNG: fixed HVP cost per iteration.* Under the stated local conditions, fresh PSNG converges linearly with a fixed HVP budget per iteration after initialization, even when the Hessian varies. The required subspace accuracy and refresh budget need not increase as the target error decreases.
- *Lazy PSNG: bounded total HVP cost.* We propose a refresh rule based on accumulated step length and prove local linear convergence with only finitely many HVPs. Under the stated local conditions, the total HVP bound, including initialization and all subsequent refreshes, is independent of the target accuracy ε when initialization and parameters are fixed.
- *Global convergence with bounded total HVP cost.* Under global strong convexity and smoothness and the stated local conditions, we extend the ε-independent total HVP guarantee to arbitrary starting points through a prescribed local restart. The bound covers both the global and local phases, with parameters and initialization bounds fixed.

Relation to prior work. SONIA and FOSI provide convergence guarantees for hybrid first- and second-order methods, but do not specify how accurately the subspace must be computed to retain convergence governed by the remaining spectrum (Jahani et al., 2021; Sivan et al., 2024). We address this gap by bounding the contraction loss due to subspace error on strongly convex quadratics, deriving an explicit accuracy condition for preserving spectral acceleration, and giving HVP budgets for attaining the required accuracy and maintaining it as the Hessian varies. Regularized lazy Newton methods refresh a full Hessian at fixed iteration intervals, with global complexity guarantees and, under strong convexity, local superlinear convergence (Doikov et al., 2023). Lazy PSNG instead uses accumulated step length to decide when to refresh its basis and projected Hessian. Under the stated local conditions, it converges linearly with finitely many curvature refreshes; subsequent iterations use the final stored pair and current gradients. Appendix A discusses further connections.

2 Principal-Subspace Newton–Gradient Methods

We consider $\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$ under the following basic assumptions.

Assumption 2.1. $f \in C^2$ has a minimizer $\mathbf{x}^\star$. Gradients and HVPs are evaluated exactly.

Write $\mathbf{H}(\mathbf{x}) := \nabla^2 f(\mathbf{x})$, $\mathbf{H}_t := \mathbf{H}(\mathbf{x}_t)$, and $\mathbf{g}_t := \nabla f(\mathbf{x}_t)$. For $\mathbf{H}_t \succ 0$, consider the quadratic model at $\mathbf{x}_t$: $\mathbf{g}_t^\top \mathbf{d} + \frac{1}{2}\mathbf{d}^\top \mathbf{H}_t \mathbf{d}$. Gradient descent on this model can be slow when $\mathbf{H}_t$ is ill-conditioned: large eigenvalues restrict the stepsize, slowing progress along directions of small curvature. Newton's method minimizes the model by solving $\mathbf{H}_t\mathbf{d} = -\mathbf{g}_t$, but this d-dimensional system can be prohibitively costly. This motivates hybrid methods that combine the two approaches by taking a Newton step in a leading eigenspace of dimension r and a gradient step in its orthogonal complement, with $1 \le r \ll d$. This reduces the Newton computation to an $r \times r$ system; with the exact leading eigenspace, the gradient step is governed only by the remaining spectrum.

Nevertheless, computing the exact leading eigenspace can be expensive. At an iterate $\mathbf{x}_s$, PSNG therefore constructs an orthonormal basis $\mathbf{P}_s \in \mathbb{R}^{d\times r}$ whose columns approximately span the leading eigenspace of $\mathbf{H}_s$, with $\mathbf{P}_s^\top \mathbf{P}_s = \mathbf{I}_r$. To derive the Newton correction within this approximate subspace, we seek a direction of the form $\mathbf{P}_s\mathbf{u}$, where $\mathbf{u} \in \mathbb{R}^r$. Substituting into the quadratic Taylor model gives, up to the constant $f(\mathbf{x}_s)$, the model

$$m_s(\mathbf{u}) := \mathbf{g}_s^\top \mathbf{P}_s \mathbf{u} + \tfrac{1}{2}\mathbf{u}^\top \widetilde{\mathbf{B}}_s \mathbf{u},$$

where $\widetilde{\mathbf{B}}_s := \mathbf{P}_s^\top \mathbf{H}_s \mathbf{P}_s$. Thus, $\widetilde{\mathbf{B}}_s$ is the Hessian of the restricted objective $\mathbf{u} \mapsto f(\mathbf{x}_s + \mathbf{P}_s\mathbf{u})$ at $\mathbf{u} = 0$. For $\mathbf{H}_s \succ 0$, the projected Hessian is also positive definite, so the minimizer of m_s satisfies $\widetilde{\mathbf{B}}_s\mathbf{u} = -\mathbf{P}_s^\top \mathbf{g}_s$, giving the subspace Newton direction $-\mathbf{P}_s\widetilde{\mathbf{B}}_s^{-1}\mathbf{P}_s^\top \mathbf{g}_s$. Once the basis and projected Hessian are available, this correction requires solving only an $r \times r$ system.

The subspace Newton direction leaves the orthogonal complement unchanged, so PSNG adds a gradient step there. To reduce repeated curvature computation, it also allows the pair $(\mathbf{P}_s, \widetilde{\mathbf{B}}_s)$ to be reused. At iteration t, let $s \le t$ denote the iteration at which the stored pair was last computed. With a gradient stepsize $\eta > 0$ in the orthogonal complement, the resulting preconditioner is

$$\mathbf{M}_t \coloneqq \mathbf{P}_s \widetilde{\mathbf{B}}_s^{-1} \mathbf{P}_s^\top + \eta(\mathbf{I} - \mathbf{P}_s \mathbf{P}_s^\top), \tag{2.1}$$

and the PSNG update uses the current gradient:

$$\mathbf{d}_t \coloneqq -\mathbf{M}_t \mathbf{g}_t, \qquad \mathbf{x}_{t+1} \coloneqq \mathbf{x}_t + \mathbf{d}_t. \tag{2.2}$$

When $s = t$, the subspace component is the Newton direction derived above. When $s < t$, $\widetilde{\mathbf{B}}_s$ replaces the current projected Hessian $\mathbf{P}_s^\top \mathbf{H}_t \mathbf{P}_s$, while the gradient remains current.

Subspace approximation and tracking. To compute the basis and track the leading eigenspace as the Hessian changes, we use block power iteration (Saad, 2011). At a point $\mathbf{x}_s$ where curvature information is to be computed, start from $\mathbf{Q}_{s,0} \in \mathbb{R}^{d\times r}$ with orthonormal columns and repeat

$$\mathbf{Q}_{s,k+1} \coloneqq \mathrm{orth}(\mathbf{H}_s \mathbf{Q}_{s,k}), \qquad k = 0, \dots, K-1,$$

where orth returns an orthonormal basis for the column space. Each block product requires r HVPs, one per column, without forming $\mathbf{H}_s$ explicitly (Pearlmutter, 1994). After K rounds, set $\mathbf{P}_s \coloneqq \mathbf{Q}_{s,K}$ and use another r HVPs to compute $\mathbf{Y} \coloneqq \mathbf{H}_s \mathbf{P}_s$, giving $\widetilde{\mathbf{B}}_s = \mathbf{P}_s^\top \mathbf{Y}$. The total cost is $K+1$ block HVPs, or $r(K+1)$ scalar HVPs; this count excludes gradient evaluations, orthonormalization, and the solution of projected linear systems.

Initialization uses $\mathbf{Q}_{0,0} \coloneqq \mathbf{Q}_0 \in \mathbb{R}^{d\times r}$ with orthonormal columns and K_{boot} rounds, under the initial-overlap and accuracy conditions specified in the analysis. For each later computation, let $j < s$ denote the index of the previous stored pair, set $\mathbf{Q}_{s,0} \coloneqq \mathbf{P}_j$, and run K_{ref} rounds. When the Hessian changes little relative to its eigengap, the previous basis remains close to the current leading eigenspace. Warm starts therefore let power iteration track changes in the principal directions by refining the stored basis.

Curvature refresh and reuse. With this tracking procedure in place, we must decide when to update the stored basis and projected Hessian; we call their joint update a *curvature refresh*. The *fresh* policy refreshes before every optimization step, so $s = t$. The *lazy* policy reuses the pair across multiple steps to reduce HVP cost when curvature changes little. Both policies use warm starts after initialization and the current gradient at every step.

For lazy PSNG, we propose triggering a refresh by *accumulated step length*. When the Hessian varies smoothly and has an eigengap, limiting this quantity controls changes in both the leading subspace and projected Hessian. We therefore maintain

$$\ell_{s,t} \coloneqq \sum_{k=s}^{t-1} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2$$

and refresh before step t whenever $\ell_{s,t} \ge \omega$, for a fixed threshold $\omega > 0$, resetting the accumulator. This rule uses only distances between successive iterates and requires no HVPs to monitor. As steps shrink, more updates fit below the same threshold, so the reuse interval adapts to optimization progress.

A larger ω permits longer reuse but allows more curvature change; a smaller value triggers more frequent curvature updates. Section 4.2 gives sufficient joint choices of ω, subspace accuracy, and K_{ref} that preserve convergence. Algorithm 1 summarizes both policies.

3 Convergence on Strongly Convex Quadratics

For the quadratic objective $f(\mathbf{x}) \coloneqq \frac{1}{2}(\mathbf{x}-\mathbf{x}^\star)^\top \mathbf{H}(\mathbf{x}-\mathbf{x}^\star)$ with $\mathbf{H} \succ 0$, we analyze lazy PSNG with curvature refreshes disabled after initialization. Since the Hessian is *constant*, the computed subspace and projected Hessian retain their accuracy throughout optimization. This setting isolates the effect of subspace error and shows that an *approximate* subspace can preserve spectral acceleration, with an iteration bound governed by the remaining spectrum.

Algorithm 1 PSNG with fresh or lazy update policy

Input: $\mathbf{x}_0$, orthonormal $\mathbf{Q}_0 \in \mathbb{R}^{d\times r}$, $\eta > 0$, $\texttt{mode} \in \{\texttt{fresh}, \texttt{lazy}\}$, $K_{\mathrm{boot}}, K_{\mathrm{ref}} \in \mathbb{Z}_{\geq 0}$, $\omega > 0$, iteration budget $T \in \mathbb{Z}_{\geq 0}$.

$(\mathbf{P}, \widetilde{\mathbf{B}}) \leftarrow \textsc{Refresh}(\mathbf{x}_0, \mathbf{Q}_0, K_{\mathrm{boot}})$; $\ell \leftarrow 0$
for $t = 0, \ldots, T-1$ **do**
  if $t > 0$ **and** ($\texttt{mode} = \texttt{fresh}$ **or** $\ell \geq \omega$) **then**
    $(\mathbf{P}, \widetilde{\mathbf{B}}) \leftarrow \textsc{Refresh}(\mathbf{x}_t, \mathbf{P}, K_{\mathrm{ref}})$; $\ell \leftarrow 0$
  end if
  $\mathbf{g}_t \leftarrow \nabla f(\mathbf{x}_t)$; $\mathbf{b}_t \leftarrow \mathbf{P}^\top \mathbf{g}_t$; solve $\widetilde{\mathbf{B}}\mathbf{z}_t = \mathbf{b}_t$
  $\mathbf{d}_t \leftarrow -\mathbf{P}\mathbf{z}_t - \eta(\mathbf{g}_t - \mathbf{P}\mathbf{b}_t)$; $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t + \mathbf{d}_t$
  $\ell \leftarrow \ell + \|\mathbf{d}_t\|_2$
end for
return $\mathbf{x}_T$

function $\textsc{Refresh}(\mathbf{x}, \mathbf{Q}, K)$
for $k = 1, \ldots, K$ **do**
  $\mathbf{Q} \leftarrow \mathrm{orth}(\nabla^2 f(\mathbf{x})\mathbf{Q})$
end for
$\mathbf{Y} \leftarrow \nabla^2 f(\mathbf{x})\mathbf{Q}$
return $(\mathbf{Q}, \mathbf{Q}^\top \mathbf{Y})$
end function

Order the eigenvalues as $\lambda_1 \geq \cdots \geq \lambda_d > 0$. For an orthonormal $\mathbf{P} \in \mathbb{R}^{d\times r}$, form $\mathbf{M}$ as in (2.1) using $\mathbf{P}^\top \mathbf{H}\mathbf{P}$, and keep $\mathbf{M}$ fixed. Writing $\mathbf{e}_t := \mathbf{x}_t - \mathbf{x}^\star$, we have $\mathbf{g}_t = \mathbf{H}\mathbf{e}_t$, so the PSNG update gives $\mathbf{e}_{t+1} = (\mathbf{I} - \mathbf{M}\mathbf{H})\mathbf{e}_t$. To measure progress, we use a norm tied to the objective. For $\mathbf{A} \succ 0$, define $\|\mathbf{v}\|_{\mathbf{A}} := (\mathbf{v}^\top \mathbf{A}\mathbf{v})^{1/2}$ and let $\|\cdot\|_{\mathbf{A}}$ denote its induced operator norm. Here, $\|\mathbf{e}_t\|_{\mathbf{H}}^2 = 2[f(\mathbf{x}_t) - f(\mathbf{x}^\star)]$, so reducing the $\mathbf{H}$-norm error also reduces the objective gap. The quantity $q := \|\mathbf{I} - \mathbf{M}\mathbf{H}\|_{\mathbf{H}}$ is the largest possible ratio $\|\mathbf{e}_{t+1}\|_{\mathbf{H}} / \|\mathbf{e}_t\|_{\mathbf{H}}$ over nonzero errors. Proving $q < 1$ therefore gives a uniform reduction at each step and hence $\|\mathbf{e}_t\|_{\mathbf{H}} \leq q^t \|\mathbf{e}_0\|_{\mathbf{H}}$.

Spectral acceleration with approximate subspaces. Let $\mathbf{P}_\star$ have orthonormal columns spanning a leading rank-r eigenspace, and measure the subspace error by $\delta := \|\mathbf{P}\mathbf{P}^\top - \mathbf{P}_\star\mathbf{P}_\star^\top\|_2$, the sine of the largest principal angle. When $\delta = 0$, the Newton correction cancels the error components in the leading r eigendirections, while the gradient step multiplies each remaining component by $1 - \eta\lambda_i$. Hence, $q = \rho_\perp := \max_{i>r} |1 - \eta\lambda_i|$, which is less than one for $0 < \eta < 2/\lambda_{r+1}$. The following bound quantifies how subspace approximation affects this reference rate.

Theorem 3.1 (Contraction bound)**.** *If $\rho_\perp > 0$, then, for every $\delta \in [0, 1]$,*

$$q \leq \rho_\perp + C_\eta \delta^2, \tag{3.1}$$

where $C_\eta := \eta\left[(\lambda_1 - \lambda_{r+1}) + \frac{(\lambda_1 - \lambda_d)^2}{\rho_\perp \lambda_d}\right]$.

For a fixed positive exact-subspace factor, the bound limits the contraction loss to a term quadratic in the projector error. The bound needs no eigengap; its proof and the distinct $\rho_\perp = 0$ case are given in Appendix B.2. The next result shows when convergence is governed by $\kappa_\perp := \lambda_{r+1}/\lambda_d$, the condition number of the remaining spectrum.

Corollary 3.2 (Accuracy sufficient for spectral acceleration)**.** *Suppose $\kappa_\perp > 1$ and choose $\eta = 2/(\lambda_{r+1} + \lambda_d)$, which minimizes $\rho_\perp$ and gives $\rho_\perp = (\kappa_\perp - 1)/(\kappa_\perp + 1)$. If the fixed subspace satisfies*

$$C_\eta \delta^2 \leq \frac{1 - \rho_\perp}{2} = \frac{1}{\kappa_\perp + 1}, \tag{3.2}$$

then $q \leq \kappa_\perp/(\kappa_\perp + 1) < 1$. From any initial point, $O(\kappa_\perp \log(1/\varepsilon))$ iterations suffice to reach $\|\mathbf{e}_T\|_{\mathbf{H}} \leq \varepsilon$, for any target accuracy $0 < \varepsilon < \|\mathbf{e}_0\|_{\mathbf{H}}$.

For comparison, optimally stepped gradient descent (GD) has contraction factor $(\kappa - 1)/(\kappa + 1)$, where $\kappa := \lambda_1/\lambda_d$, and an iteration bound of $O(\kappa \log(1/\varepsilon))$ (Nesterov, 2004). Corollary 3.2 establishes *spectral acceleration with a fixed, approximate subspace*. Under (3.2), the condition-number

dependence of the iteration bound drops from κ to $\kappa_\perp$, a reduction by the potentially large factor $\kappa/\kappa_\perp = \lambda_1/\lambda_{r+1}$. The proof is in Appendix B.3. The required subspace accuracy is *independent* of ε: a single initial approximation suffices to retain this improved iteration bound for all target tolerances.

HVPs needed to compute the subspace. We next determine how many block power iterations suffice to meet (3.2). This construction requires the following additional conditions.

Assumption 3.3 (Quadratic subspace computation)**.** There is a positive eigengap at rank r, namely $\lambda_r > \lambda_{r+1}$, and the initial orthonormal basis $\mathbf{Q}_0$ has full-rank overlap with the leading eigenspace: $\mathbf{P}_\star^\top \mathbf{Q}_0$ is invertible.

Let $\mathbf{Q}_K$ be the basis after K rounds and define $\delta_K := \left\|\mathbf{Q}_K\mathbf{Q}_K^\top - \mathbf{P}_\star\mathbf{P}_\star^\top\right\|_2$. Set $\theta := \lambda_{r+1}/\lambda_r \in (0,1)$, and let Γ_0 be the finite tangent of the largest initial principal angle. Block power iteration gives $\delta_K \le \Gamma_0\theta^K$; substituting this estimate into (3.2) yields a sufficient K (Appendix B.4). Computing $\mathbf{P} := \mathbf{Q}_K$ and its projected Hessian then costs $r(K+1)$ scalar HVPs, with no further HVPs during optimization. This construction cost can still depend on the full spectrum and initial overlap.

4 Convergence with Curvature Tracking and Reuse

When the Hessian varies along the trajectory, its principal subspace and projected Hessian can change. We ask whether PSNG can retain the convergence benefit of an approximate principal subspace while controlling the cost of tracking and reusing this changing curvature.

Assumption 4.1 (Local regularity)**.** There is a convex neighborhood $\mathcal{N}$ of $\mathbf{x}^\star$ on which

$$\mu\mathbf{I} \preceq \mathbf{H}(\mathbf{x}) \preceq L\mathbf{I}, \qquad \|\mathbf{H}(\mathbf{x}) - \mathbf{H}(\mathbf{y})\|_2 \le L_H \|\mathbf{x}-\mathbf{y}\|_2, \quad \mathbf{x},\mathbf{y}\in\mathcal{N}, \tag{4.1}$$

for constants $0 < \mu \le L$ and $L_H \ge 0$.

Assumption 4.2 (Local spectral conditions)**.** The Hessian at the minimizer has a positive eigengap at rank r; write $\gamma := \frac{1}{2}[\lambda_r(\mathbf{H}(\mathbf{x}^\star)) - \lambda_{r+1}(\mathbf{H}(\mathbf{x}^\star))] > 0$ for half this gap. The fixed stepsize $\eta > 0$ satisfies $\rho_\perp(\mathbf{x}) := \max_{i>r}|1-\eta\lambda_i(\mathbf{H}(\mathbf{x}))| \le \bar\rho < 1$ throughout $\mathcal{N}$, for some $\bar\rho \in [0,1)$.

The eigengap controls subspace tracking, while $\bar\rho$ controls contraction in the orthogonal complement. Both fresh and lazy PSNG begin by computing an accurate subspace from an initial basis with sufficient overlap.

Assumption 4.3 (Initial subspace overlap)**.** The orthonormal initial basis $\mathbf{Q}_0$ has full-rank overlap with the leading eigenspace of $\mathbf{H}_0$, and the tangent of its largest principal angle is bounded by a known $\Gamma < \infty$.

Measure subspace accuracy at a refresh s by $\delta_s := \left\|\mathbf{P}_s\mathbf{P}_s^\top - \mathbf{\Pi}_{\star,s}\right\|_2$, where $\mathbf{\Pi}_{\star,s}$ is the leading rank-r projector of $\mathbf{H}_s$. After shrinking $\mathcal{N}$ if necessary, the eigengap at rank r is at least γ and the ratio λ_{r+1}/λ_r is at most $\bar\theta := 1-\gamma/L \in (0,1)$; this ratio bounds the contraction of block power iteration. To obtain $\delta_0 \le \delta_{\mathrm{tol}}$ for a prescribed tolerance $\delta_{\mathrm{tol}} \in (0,1)$, start from $\mathbf{Q}_0$ and run

$$K_{\mathrm{boot}} := \left\lceil \frac{\log\max\{1, \Gamma/\delta_{\mathrm{tol}}\}}{\log(1/\bar\theta)} \right\rceil \tag{4.2}$$

block-power rounds, followed by one block HVP to form the projected Hessian. This initialization is performed once; subsequent refreshes use K_{ref} rounds starting from the previous basis, with budgets specified below for fresh and lazy PSNG.

4.1 Local Convergence of Fresh PSNG

Fresh PSNG isolates the cost of tracking a changing principal subspace by refreshing curvature at every iteration. The key is that nearby iterates have nearby Hessians: the previous basis remains a useful approximation, so a fixed number of power iterations can restore its accuracy.

Proposition 4.4 (Local convergence of fresh PSNG)**.** *Under Assumptions 4.1 and 4.2, choose a target $q_{\mathrm{tar}} \in (\bar\rho, 1)$ and a tolerance $\delta_{\mathrm{tol}} \in (0,1)$ satisfying*

$$\bar\rho + \eta(L-\mu)\left(1 + \frac{L-\mu}{\mu q_{\mathrm{tar}}}\right)\delta_{\mathrm{tol}}^2 \le q_{\mathrm{tar}}. \tag{4.3}$$

For the error measure $V_t \coloneqq \|\mathbf{e}_t\|_{\mathbf{H}_t}$, define $c_s \coloneqq L\max\{\mu^{-1}, \eta\}/\sqrt{\mu}$ and $C_{\mathrm{loc}} \coloneqq L_H c_s(1/\mu + 1/(2\sqrt{\mu L}))$. Choose $R > 0$ satisfying

$$\overline{\mathbb{B}}_2\left(\mathbf{x}^\star, \frac{R}{\sqrt{\mu}} + c_s R\right) \subset \mathcal{N}, \qquad 2C_{\mathrm{loc}} R \le 1 - q_{\mathrm{tar}}. \tag{4.4}$$

Here, $\overline{\mathbb{B}}_2$ denotes a closed Euclidean ball. For the warm start from the preceding iterate, require $\bar{\delta} \coloneqq \delta_{\mathrm{tol}} + 2L_H c_s R/\gamma < 1$. Initialize with $\mathbf{x}_0 \in \mathcal{N}$, $V_0 \le R$, and $\delta_0 \le \delta_{\mathrm{tol}}$. Before each subsequent update at $t \ge 1$, refresh the basis by applying block power iteration to $\mathbf{H}_t$, starting from $\mathbf{P}_{t-1}$, for a fixed number of rounds

$$K_{\mathrm{ref}} \coloneqq \left\lceil \frac{\log\left(\bar{\delta}/[\delta_{\mathrm{tol}}\sqrt{1-\bar{\delta}^2}]\right)}{\log(1/\bar{\theta})} \right\rceil. \tag{4.5}$$

Form the projected Hessian in the refreshed basis before taking the PSNG step. Then, fresh PSNG maintains $\delta_t \le \delta_{\mathrm{tol}}$ and $V_t \le q_{\mathrm{loc}}^t V_0$, where $q_{\mathrm{loc}} \coloneqq (1 + q_{\mathrm{tar}})/2 < 1$.

For a fresh curvature pair with $\delta_t \le \delta_{\mathrm{tol}}$, condition (4.3) ensures $q_t \coloneqq \|\mathbf{I} - \mathbf{M}_t\mathbf{H}_t\|_{\mathbf{H}_t} \le q_{\mathrm{tar}}$, where q_t measures contraction with the Hessian held fixed at its current value. Warm starts maintain this accuracy, while the radius conditions control the nonlinear remainder and keep the iterates in $\mathcal{N}$ (Lemma C.1); the full proof is in Appendix C.3. Condition (4.3) allows nonzero subspace error within the margin between $\bar{\rho}$ and q_{tar}, so a fixed tolerance can retain a contraction target close to the ideal bound determined by the remaining spectrum.

After initialization, fresh PSNG uses $K_{\mathrm{ref}} + 1$ block HVPs at every iteration to update the basis and projected Hessian. Thus, with the initial state and all parameters fixed, $O(\log(1/\varepsilon))$ iterations and HVPs suffice to reach $V_t \le \varepsilon$.

4.2 Local Convergence of Lazy PSNG

The guarantee for fresh PSNG requires refreshing curvature at every step, making its HVP bound grow as the target tolerance ε decreases. It leaves open how long a computed curvature pair can be reused without losing convergence. The difficulty is that reuse introduces error in both the principal subspace and the stored projected Hessian.

Our analysis controls these errors through the *accumulated step length* since the last refresh. While the iterates remain in $\mathcal{N}$, the bound $\|\mathbf{H}_t - \mathbf{H}_s\|_2 \le L_H \ell_{s,t}$ controls the change in the projected Hessian and, through the eigengap, the principal subspace. Bounding the accumulated step length can therefore preserve contraction during reuse and keep the stored basis suitable for a warm start at the next refresh, using a trigger that requires no HVPs. To quantify the contraction during reuse, define

$$\Phi(\ell) \coloneqq \lambda_{\max}\begin{pmatrix} \dfrac{L_H\ell}{\mu} & \sqrt{\dfrac{\eta}{\mu}}((L-\mu)\delta_{\mathrm{tol}} + L_H\ell) \\ \sqrt{\dfrac{\eta}{\mu}}((L-\mu)\delta_{\mathrm{tol}} + L_H\ell) & \bar{\rho} + \eta(L-\mu)\delta_{\mathrm{tol}}^2 + \eta L_H \ell \end{pmatrix}. \tag{4.6}$$

If the basis has error at most δ_{tol} at a refresh, then $q_t \le \Phi(\ell_{s,t})$ throughout a reuse interval in $\mathcal{N}$ (see Lemma D.2). Since Φ is nondecreasing, choosing $\Phi(\omega) \le q_{\mathrm{tar}}$ controls contraction until the threshold is reached. The following theorem combines this rule with a warm-start budget that restores the required subspace accuracy at each refresh.

Theorem 4.5 (Local convergence of lazy PSNG)**.** *Under Assumptions 4.1 and 4.2, choose $q_{\mathrm{tar}} \in (\bar{\rho}, 1)$, $\delta_{\mathrm{tol}} \in (0, 1)$, $R > 0$ satisfying* (4.4)*, and $\omega > 0$ such that*

$$\Phi(\omega) \le q_{\mathrm{tar}}, \qquad \bar{\delta} \coloneqq \delta_{\mathrm{tol}} + \frac{2L_H(\omega + c_s R)}{\gamma} < 1. \tag{4.7}$$

Initialize with $\mathbf{x}_0 \in \mathcal{N}$, $V_0 \le R$, and a basis satisfying Assumption 4.3. Use K_{boot} from (4.2) *for initialization; at each later refresh, start from the previous basis and run the following number of block-power rounds:*

$$K_{\mathrm{ref}} \coloneqq \left\lceil \frac{\log\left(\bar{\delta}/[\delta_{\mathrm{tol}}\sqrt{1-\bar{\delta}^2}]\right)}{\log(1/\bar{\theta})} \right\rceil. \tag{4.8}$$

Run Algorithm 1 in lazy mode, refreshing before any step for which $\ell_{s,t} \geq \omega$. Then,

$$V_t \leq q_{\mathrm{loc}}^t V_0. \tag{4.9}$$

The total number of refreshes, including initialization, is at most $1 + c_s V_0/[(1 - q_{\mathrm{loc}})\omega]$, and the total number of block HVPs is at most

$$K_{\mathrm{boot}} + 1 + \frac{(K_{\mathrm{ref}} + 1)c_s V_0}{(1 - q_{\mathrm{loc}})\omega}. \tag{4.10}$$

The finite HVP bound follows from using accumulated step length to trigger refreshes. Linear convergence makes the sum of all step lengths finite. Each refresh after initialization requires an additional accumulated step length of at least ω, so only finitely many refreshes can occur, each with a fixed HVP budget. Thus, with the initial state and all parameters fixed, the same total HVP bound applies to every target accuracy ε, while gradient iterations retain an $O(\log(1/\varepsilon))$ bound and reuse the final curvature pair after the last refresh. The proof is in Appendix D.3.

The threshold balances reuse against the work needed at the next refresh. Choose the subspace tolerance small enough to leave a contraction margin for curvature changes, then select positive ω and R satisfying (4.7) and the radius conditions (4.4). A larger ω permits longer reuse but allows more curvature change, potentially increasing the warm-start budget. Appendix D.4 gives explicit sufficient choices.

4.3 Global Convergence under Global Strong Convexity

The local results require a sufficiently close starting point and an accurate subspace. To reach this regime from an arbitrary point, we use Armijo backtracking and then initialize the local method when the gradient is sufficiently small.

Assumption 4.6 (Global strong convexity and smoothness)**.** The Hessian bounds hold globally with the same constants $0 < \mu \leq L$: $\mu\mathbf{I} \preceq \mathbf{H}(\mathbf{x}) \preceq L\mathbf{I}$ for all $\mathbf{x} \in \mathbb{R}^d$.

These bounds make the PSNG preconditioner uniformly positive definite, so even an inaccurate or reused subspace gives a descent direction. For fixed $\eta > 0$, choose $c_A \in (0, 1/2)$ and $\beta \in (0, 1)$, and reduce α from 1 by factors of β until $f(\mathbf{x}_t + \alpha\mathbf{d}_t) \leq f(\mathbf{x}_t) + c_A\alpha\mathbf{g}_t^\top\mathbf{d}_t$ (Armijo, 1966; Nocedal & Wright, 2006). Denote the selected coefficient by α_t and set $\mathbf{x}_{t+1} = \mathbf{x}_t + \alpha_t\mathbf{d}_t$. The lazy refresh accumulator adds the step length $\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2 = \alpha_t \|\mathbf{d}_t\|_2$.

Theorem 4.7 (Global convergence with backtracking)**.** *Under Assumption 4.6, backtracking terminates for any sequence of curvature pairs formed as in Section 2. Set $m_{\min} \coloneqq \min\{L^{-1}, \eta\}$, $M_{\max} \coloneqq \max\{\mu^{-1}, \eta\}$, and $\Delta_t \coloneqq f(\mathbf{x}_t) - f(\mathbf{x}^\star)$. Then,*

$$\alpha_t \geq \alpha_{\min}, \qquad \Delta_{t+1} \leq q_{\mathrm{G}}\Delta_t, \tag{4.11}$$

where $\alpha_{\min} \coloneqq \beta \min\{1, 2(1 - c_A)/(LM_{\max})\}$ and $q_{\mathrm{G}} \coloneqq 1 - 2c_A\alpha_{\min}\mu m_{\min} \in (0, 1)$. Thus, $\mathbf{x}_t \to \mathbf{x}^\star$ from any starting point, without an eigengap or subspace-accuracy requirement.

To recover the local guarantees, also impose Assumptions 4.1 and 4.2, and choose the parameters of Proposition 4.4 or Theorem 4.5 with $q_{\mathrm{tar}} < 1 - 2c_A$ and $L_H M_{\max} c_s R \leq 3(1 - 2c_A - q_{\mathrm{tar}})$. These conditions ensure that the local updates satisfy the Armijo decrease condition with $\alpha_t = 1$. The gradient test $\|\mathbf{g}_t\|_2 \leq \mu R/\sqrt{L}$ certifies $\mathbf{x}_t \in \mathcal{N}$ and $V_t \leq R$. Its first passing index T_{ent} satisfies

$$T_{\mathrm{ent}} \leq \left\lceil \frac{\log \max\{1,\ 2L^2\Delta_0/(\mu^2 R^2)\}}{\log(1/q_{\mathrm{G}})} \right\rceil. \tag{4.12}$$

At this point, restart with a basis satisfying Assumption 4.3, the prescribed local budgets, and a reset path accumulator. The Armijo line search then accepts $\alpha_t = 1$ at every subsequent iteration, so the local convergence and HVP bounds apply (see Proposition E.2; proofs are in Appendix E).

In particular, the global phase may keep a single curvature pair, formed with one block HVP, until the entry test passes. Restarting lazy PSNG then gives a total block-HVP bound of one plus (4.10), with V_0 replaced by R. Under these conditions, this two-stage method converges from any starting point with a total curvature cost independent of ε, for fixed problem parameters, algorithm parameters, and initialization bounds.

5 EXPERIMENTS

We evaluate spectral acceleration and the benefit of curvature reuse. Appendices F.1 and F.2 specify the data and implementation.

Baselines are gradient descent (GD), Nesterov's accelerated gradient (NAG) (Nesterov, 2004), limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) (Liu & Nocedal, 1989), Newton's method using conjugate gradients (Newton–CG) or dense Cholesky factorization (Newton) (Nocedal & Wright, 2006), and FOSI with GD as its base optimizer (FOSI-GD) (Sivan et al., 2024).

Spectral acceleration with principal subspaces. We first isolate the effect of subspace selection on a quadratic with $d = 2000$, $r = 50$, $\kappa = 1000$, and $\kappa_\perp = 10$. Exact PSNG uses the supplied principal eigenspace. Our random-subspace PSNG variant instead chooses an orthonormal basis of the same rank at random and keeps it fixed, using the same update (2.2) with a separately tuned gradient stepsize. Exact PSNG converges much faster than GD, NAG, and this random variant (Figure 1, left), illustrating the benefit of reducing the relevant condition number from κ to $\kappa_\perp$.

To isolate approximation error, we also vary the rank-one subspace on the two-dimensional quadratic with $\mathbf{H} = \mathrm{diag}(4, 1)$. At $\eta = 1/2$, the increase in the contraction factor scales with the square of the subspace error, consistent with Section 3. Plots and further accuracy diagnostics are in Appendix F.3.

We next test computed subspaces on ℓ_2-regularized logistic regression ($n = 4000$, $d = 120$, $\lambda = 10^{-3}$), starting from zero. We multiply the six largest eigenvalues of the Gaussian feature covariance by a scale factor $a \in \{0.1, 0.3, 1\}$, keeping the remaining eigenvalues, latent samples, and labels fixed within each seed. PSNG uses $r = 6$, exact gradients and HVPs, and block power iteration. Unless stated otherwise, logistic iteration counts refer to $\|\mathbf{g}_t\|_2 / \|\mathbf{g}_0\|_2 \le 10^{-6}$. As the full Hessian condition number at the reference minimizer rises roughly tenfold, the remaining condition number stays near five. GD slows substantially, whereas the iteration counts for fresh and lazy PSNG remain nearly unchanged (Figure 1, right): approximate principal subspaces retain the acceleration despite larger leading eigenvalues.

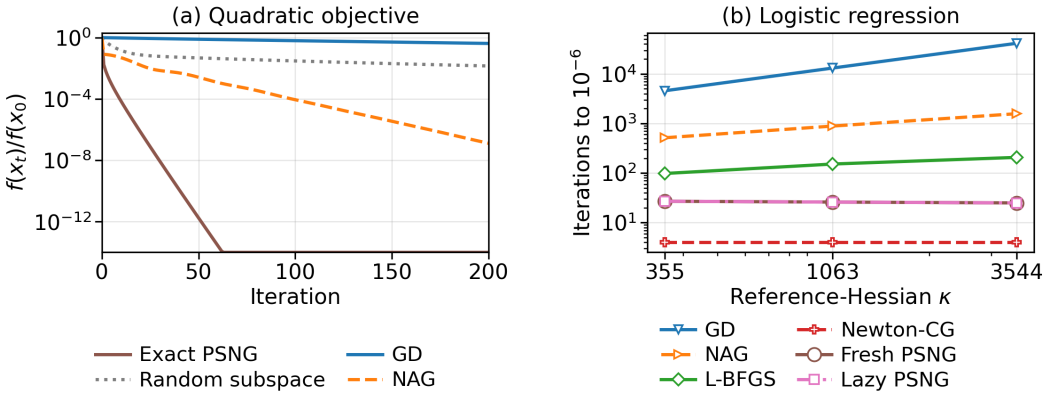

Figure 1: Spectral acceleration. Left: relative quadratic objective gaps; the random-subspace band shows mean ± one standard deviation over three bases. Right: logistic-regression iterations to relative gradient 10^{-6}, with medians and interquartile ranges over three data seeds.

Maintaining convergence with fewer curvature refreshes. Fresh and lazy PSNG use the same initial basis and Armijo line-search settings, with $K_{\mathrm{boot}} = 5$ and $K_{\mathrm{ref}} = 2$; lazy PSNG uses $\omega = 1$. FOSI-GD refreshes every ten updates. All logistic HVP and timing costs include spectral parameter selection; a rank-r block product counts as r HVPs. At $a = 1$, lazy matches fresh's iteration count with substantially fewer HVPs. Its HVP count stays constant as the relative-gradient tolerance tightens from 10^{-4} to 10^{-8}. The full comparison, including timing, is in Appendix F.3; Newton–CG is fastest on these $d = 120$ problems, and FOSI-GD has similar runtime to lazy PSNG.

On the original data, keeping the initial curvature pair fixed also gives nearly the same convergence. To examine when refreshing matters, we increase the label-generating signal strength fivefold and

allocate 80% of its latent energy to the six leading covariance directions. Across ten new seeds, lazy needs only four or five refreshes after initialization and nearly matches fresh's iteration count, while the fixed pair takes about 3.7 times as many iterations. Lazy saves about 79% of fresh's HVPs, with both comparisons computed per seed before taking medians.

Its HVP cost stops increasing by tolerance 10^{-6} and remains unchanged through 10^{-8} (Figure 2). Appendix F.4 gives the full comparison and diagnostics, which attribute the fixed pair's slower convergence mainly to its outdated projected Hessian. A rank sweep further shows that capturing the six dominant directions gives most of the iteration benefit, while larger subspaces add HVP cost (Appendix F.5).

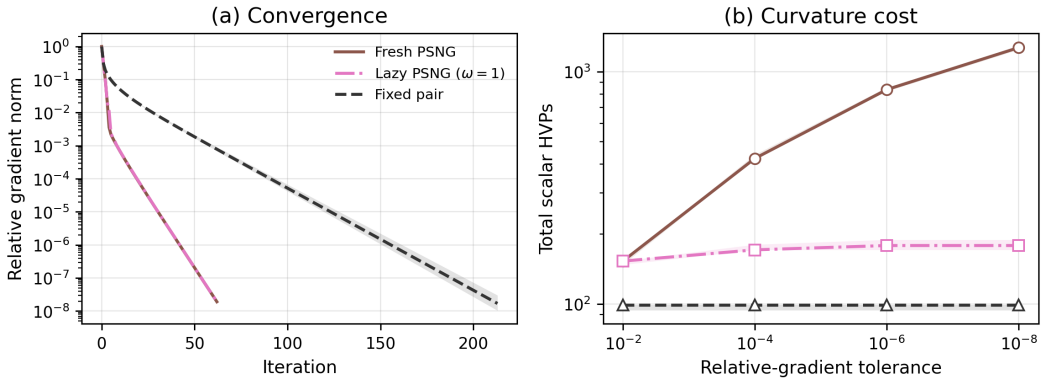

Figure 2: Few refreshes preserve fast convergence under stronger curvature changes. Relative gradient (left) and total HVPs to each tolerance (right), with medians and interquartile ranges over ten validation seeds. Lazy uses $\omega = 1$; Fixed pair never refreshes after initialization. Convergence curves use iterations recorded for every seed.

6 Conclusion

In this work, we propose PSNG and show that approximate principal subspaces can provide acceleration without becoming increasingly accurate as optimization proceeds. Under the stated local conditions, curvature computation can even stop while linear convergence continues toward arbitrarily high accuracy. These guarantees rely on exact gradients and HVPs, strong convexity, and an eigengap for subspace tracking. In stochastic settings, an open question is whether variance reduction (Johnson & Zhang, 2013; Shamir, 2016) can stabilize tracking and prevent noise from causing persistent refreshes. For nonconvex objectives, a promising direction is to incorporate cubic regularization (Nesterov & Polyak, 2006; Doikov et al., 2023), with the additional challenge of detecting and exploiting negative curvature outside the tracked subspace.

AI use statement

In this work, we used generative AI tools to assist with manuscript editing and language polishing, literature retrieval and reference verification, refining mathematical derivations, and supporting the implementation of experiments. We did not use generative AI for research ideation, research execution, or drafting substantive sections of the paper. We have reviewed all AI-assisted work. Specifically, we checked the mathematical arguments, verified the cited sources, and checked the reported experimental results against the experimental outputs. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Reproducibility statement

The proofs and explicit parameter conditions are given in Appendices B–E. Appendix F specifies the experimental objectives, data generation, algorithm settings, evaluation procedures, and computational-cost accounting. The anonymous supplementary material provides the experiment

and plotting code, fixed random seeds, dependency versions, and reference numerical results, together with instructions for rerunning all experiments and regenerating the figures. All experimental data are generated by the supplied scripts.

REFERENCES

Larry Armijo. Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16(1):1–3, 1966.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

R. H. Byrd, S. L. Hansen, Jorge Nocedal, and Yoram Singer. A stochastic quasi-Newton method for large-scale optimization. *SIAM Journal on Optimization*, 26(2):1008–1031, 2016.

Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

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

Nikita Doikov, El Mahdi Chayti, and Martin Jaggi. Second-order optimization with lazy Hessians. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8138–8161. PMLR, 2023.

Nikita Doikov, Sebastian U. Stich, and Martin Jaggi. Spectral preconditioning for gradient methods on graded non-convex functions. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11227–11252. PMLR, 2024.

Murat A. Erdogdu and Andrea Montanari. Convergence rates of sub-sampled Newton methods. In *Advances in Neural Information Processing Systems*, volume 28, pp. 3052–3060, 2015.

Zachary Frangella, Joel A. Tropp, and Madeleine Udell. Randomized Nyström preconditioning. *SIAM Journal on Matrix Analysis and Applications*, 44(2):718–752, 2023.

Robert M. Gower, Dmitry Kovalev, Felix Lieder, and Peter Richtárik. RSN: Randomized subspace Newton. In *Advances in Neural Information Processing Systems*, volume 32, pp. 616–625, 2019.

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

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.

Filip Hanzely, Nikita Doikov, Peter Richtárik, and Yurii Nesterov. Stochastic subspace cubic Newton method. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4027–4038. PMLR, 2020.

Slavomir Hanzely. Sketch-and-project meets Newton method: Global $O(1/k^2)$ convergence with low-rank updates. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 3205–3213. PMLR, 2025.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, second edition, 2013.

Majid Jahani, MohammadReza Nazari, Rachael Tappenden, Albert Berahas, and Martin Takac. SONIA: A symmetric blockwise truncated optimization algorithm. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 487–495. PMLR, 2021.

Ruichen Jiang, Parameswaran Raman, Shoham Sabach, Aryan Mokhtari, Mingyi Hong, and Volkan Cevher. Krylov cubic regularized Newton: A subspace second-order method with dimension-free convergence rate. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4411–4419. PMLR, 2024.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, volume 26, pp. 315–323. Curran Associates, Inc., 2013.

Jonathan Lacotte, Yifei Wang, and Mert Pilanci. Adaptive Newton sketch: Linear-time optimization with quadratic convergence and effective Hessian dimensionality. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5926–5936. PMLR, 2021.

Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45:503–528, 1989.

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

James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2408–2417. PMLR, 2015.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2004.

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

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, second edition, 2006.

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

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

Yousef Saad. *Numerical Methods for Large Eigenvalue Problems*. Society for Industrial and Applied Mathematics, revised edition, 2011.

Yousef Saad. Analysis of subspace iteration for eigenvalue problems with evolving matrices. *SIAM Journal on Matrix Analysis and Applications*, 37(1):103–122, 2016.

Ohad Shamir. Fast stochastic algorithms for SVD and PCA: Convergence properties and convexity. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 248–256. PMLR, 2016.

Hadar Sivan, Moshe Gabel, and Assaf Schuster. FOSI: Hybrid first and second order optimization. In *International Conference on Learning Representations*, 2024.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020.

A Extended Related Work

PSNG connects hybrid first- and second-order updates with spectral approximation and curvature reuse. We discuss these connections below, focusing on how subspace accuracy affects convergence and curvature cost.

A.1 Hybrid first- and second-order methods

SONIA combines a second-order update in a Hessian-sampled subspace with a scaled gradient step in its orthogonal complement (Jahani et al., 2021). Its main algorithm rebuilds sampled curvature at every iteration and has convergence guarantees for strongly convex and nonconvex objectives. FOSI uses approximate extreme Hessian eigenpairs to combine second-order subspace updates with a first-order optimizer, refreshing the spectral estimates at prescribed iteration intervals (Sivan et al., 2024). It provides convergence guarantees and discusses practical safeguards for eigenpair approximation errors.

For PSNG on strongly convex quadratics, we bound the contraction loss in terms of the projector error and show that this loss is quadratic in the error when the exact-subspace contraction factor is fixed and positive. This gives an explicit accuracy condition under which a fixed, approximate subspace retains an iteration bound governed by the remaining spectrum. We then relate this condition to the HVP work needed to construct and update the subspace.

A.2 Spectral preconditioning and subspace approximation

NewSamp combines Hessian subsampling with spectral truncation; its convergence bound reflects the spectrum below the truncation threshold as well as sampling error (Erdogdu & Montanari, 2015). Spectral preconditioning with inexact leading eigenvectors also improves convergence guarantees for graded nonconvex objectives (Doikov et al., 2024). For linear systems, randomized Nyström preconditioning uses low-rank spectral approximation to accelerate iterative methods (Frangella et al., 2023).

PSNG uses exact HVPs to isolate the effects of principal-subspace approximation and curvature reuse. Standard block power iteration and low-rank approximation techniques provide the spectral computation tools (Saad, 2011; Halko et al., 2011); our analysis connects their accuracy to the contraction factor of the combined Newton and gradient update. The resulting iteration bound can depend on the remaining spectrum, while the cost of constructing the subspace can still depend on the full spectrum and initial overlap.

A.3 Approximate and subspace Newton methods

Inexact Newton methods allow approximate solutions of Newton systems (Dembo et al., 1982), while Hessian-free optimization obtains curvature information through matrix–vector products without forming a dense Hessian (Martens, 2010; Pearlmutter, 1994). K-FAC and Shampoo reduce computational cost through structured preconditioners (Martens & Grosse, 2015; Gupta et al., 2018).

Newton sketches (Pilanci & Wainwright, 2017; Lacotte et al., 2021) and random or Krylov subspace Newton methods (Gower et al., 2019; Hanzely et al., 2020; Jiang et al., 2024) use sketching or restricted search directions to make Newton-type computations less expensive. Sketch-and-project Newton methods also admit global rates under semi-strong self-concordance and suitable sketching assumptions (Hanzely, 2025). PSNG couples a subspace Newton step with a gradient step in the orthogonal complement, making both subspace accuracy and the remaining spectrum relevant to convergence.

A.4 Curvature tracking and reuse

Stochastic quasi-Newton methods collect curvature information at spaced intervals (Byrd et al., 2016), and subspace iteration for evolving matrices refines an existing eigenspace estimate as the matrix changes (Saad, 2016). Regularized lazy Newton methods reuse a full Hessian for a fixed number of iterations, with global complexity guarantees and local superlinear convergence under strong convexity (Doikov et al., 2023).

For PSNG, warm starts connect changes in the Hessian to the HVP budget needed to restore subspace accuracy. The accumulated-step-length trigger controls how long both the basis and projected Hessian can be reused, without requiring HVPs to monitor the change. Under the stated local conditions, linear convergence makes the sum of step lengths finite, so only finitely many refreshes occur. With initialization and parameters fixed, the total HVP bound is independent of the target accuracy, while gradient iterations continue using the final curvature pair. This guarantee concerns curvature computation; gradient evaluations remain necessary as accuracy increases.

B Proofs for Strongly Convex Quadratics

This appendix proves the results in Section 3, using the same fixed Hessian and curvature pair. Appendix B.1 establishes a block contraction bound, which yields Theorem 3.1 in Appendix B.2. We then prove Corollary 3.2, derive the HVP budget for computing the subspace, and give examples that show the order of the error bounds cannot generally be improved.

B.1 Block error representation and contraction bounds

The block representation below applies to any selected subspace, whether or not it approximates a leading eigenspace. Fix any $\mathbf{P} \in \mathbb{R}^{d\times r}$ with orthonormal columns and let $\mathbf{U} \in \mathbb{R}^{d\times(d-r)}$ have orthonormal columns spanning $\mathrm{range}(\mathbf{P})^{\perp}$. In the basis $[\mathbf{P}, \mathbf{U}]$, write

$$\begin{pmatrix} \mathbf{B} & \mathbf{E} \\ \mathbf{E}^\top & \mathbf{C} \end{pmatrix} \coloneqq [\mathbf{P}, \mathbf{U}]^\top \mathbf{H} [\mathbf{P}, \mathbf{U}], \qquad \mathbf{B} \coloneqq \mathbf{P}^\top \mathbf{H} \mathbf{P}. \tag{B.1}$$

For the PSNG update with the exact projected Hessian $\mathbf{B}$, decompose the error as $\mathbf{e} \coloneqq \mathbf{x} - \mathbf{x}^\star = \mathbf{P}\mathbf{u} + \mathbf{U}\mathbf{v}$. A direct calculation gives

$$\begin{pmatrix} \mathbf{u}^+ \\ \mathbf{v}^+ \end{pmatrix} = \begin{pmatrix} 0 & -\mathbf{B}^{-1}\mathbf{E} \\ -\eta \mathbf{E}^\top & \mathbf{I} - \eta \mathbf{C} \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix}. \tag{B.2}$$

The Newton correction removes the diagonal error term within the selected subspace. The remaining error depends on the block $\mathbf{C}$ on the orthogonal complement and the coupling block $\mathbf{E}$, both of which depend on the chosen subspace. The coupling can also be written as the invariant-subspace residual

$$\mathcal{R}(\mathbf{P}; \mathbf{H}) \coloneqq (\mathbf{I} - \mathbf{P}\mathbf{P}^\top)\mathbf{H}\mathbf{P} = \mathbf{U}\mathbf{E}^\top, \qquad \|\mathcal{R}(\mathbf{P}; \mathbf{H})\|_2 = \|\mathbf{E}\|_2 . \tag{B.3}$$

To obtain a scalar contraction bound, define the contraction factor in the orthogonal complement and the normalized residual by

$$\rho_C \coloneqq \|\mathbf{I} - \eta \mathbf{C}\|_2 , \qquad \zeta \coloneqq \sqrt{\eta}\, \left\| \mathbf{B}^{-1/2} \mathbf{E} \right\|_2 . \tag{B.4}$$

For $a, c, z \geq 0$, define $\Psi(a, c, z) \coloneqq (a + c + \sqrt{(a-c)^2 + 4z^2})/2$, the largest eigenvalue of $\left(\begin{smallmatrix} a & z \\ z & c \end{smallmatrix}\right)$.

Theorem B.1 (Contraction bound for a quadratic objective)**.** *For every orthonormal* $\mathbf{P}$ *and* $\eta > 0$*,*

$$q \leq \widehat{q}(\mathbf{P}, \eta) \coloneqq \Psi(0, \rho_C, \zeta). \tag{B.5}$$

Consequently, $\|\mathbf{e}^+\|_{\mathbf{H}} \leq \widehat{q}\, \|\mathbf{e}\|_{\mathbf{H}}$ *and* $f(\mathbf{x}^+) - f(\mathbf{x}^\star) \leq \widehat{q}^{\,2}[f(\mathbf{x}) - f(\mathbf{x}^\star)]$*. For a target* $q_{\mathrm{tar}} \in (\rho_C, 1)$*,*

$$\zeta^2 \leq q_{\mathrm{tar}}(q_{\mathrm{tar}} - \rho_C) \quad \Longrightarrow \quad q \leq \widehat{q} \leq q_{\mathrm{tar}}. \tag{B.6}$$

Proof of Theorem B.1. Let $\mathbf{Q} \coloneqq [\mathbf{P}, \mathbf{U}] \in \mathbb{R}^{d\times d}$ and $\mathbf{T} \coloneqq \mathbf{I} - \mathbf{M}\mathbf{H}$. In this basis, $\mathbf{M}$ has diagonal blocks $\mathbf{B}^{-1}$ and $\eta\mathbf{I}$. Consequently,

$$\mathbf{S} \coloneqq \mathbf{Q}^\top \mathbf{M}^{-1/2} \mathbf{T} \mathbf{M}^{1/2} \mathbf{Q} = \mathbf{Q}^\top (\mathbf{I} - \mathbf{M}^{1/2} \mathbf{H} \mathbf{M}^{1/2}) \mathbf{Q},$$

where

$$\mathbf{S} = \begin{pmatrix} 0 & -\sqrt{\eta}\, \mathbf{B}^{-1/2}\mathbf{E} \\ -\sqrt{\eta}\, \mathbf{E}^\top \mathbf{B}^{-1/2} & \mathbf{I} - \eta \mathbf{C} \end{pmatrix}, \qquad q = \|\mathbf{S}\|_2 . \tag{B.7}$$

Also, $\mathbf{H}\mathbf{T} = \mathbf{H} - \mathbf{H}\mathbf{M}\mathbf{H}$ is symmetric, so $\mathbf{T}$ is self-adjoint in the $\mathbf{H}$ inner product. Both weighted representations have the eigenvalues of $\mathbf{T}$; hence,

$$\|\mathbf{T}\|_{\mathbf{H}} = \|\mathbf{T}\|_{\mathbf{M}^{-1}} = \rho(\mathbf{T}) = \|\mathbf{S}\|_2 = \max_i \left| 1 - \lambda_i(\mathbf{M}^{1/2} \mathbf{H} \mathbf{M}^{1/2}) \right|. \tag{B.8}$$

For any block vector $(\mathbf{u}, \mathbf{v})$,

$$\left| \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix}^\top \mathbf{S} \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} \right| \le 2\zeta \|\mathbf{u}\|_2 \|\mathbf{v}\|_2 + \rho_C \|\mathbf{v}\|_2^2 .$$

The right side is the Rayleigh form of $\left(\begin{smallmatrix} 0 & \zeta \\ \zeta & \rho_C \end{smallmatrix}\right)$ at $(\|\mathbf{u}\|_2, \|\mathbf{v}\|_2)$. Its largest eigenvalue is $\Psi(0, \rho_C, \zeta)$, proving (B.5). The error bound follows from the induced norm; the objective bound follows from $f(\mathbf{x}) - f(\mathbf{x}^\star) = \frac{1}{2}\|\mathbf{e}\|_{\mathbf{H}}^2$. For $q_{\mathrm{tar}} > \rho_C$,

$$\Psi(0, \rho_C, \zeta) \le q_{\mathrm{tar}} \quad \Longleftrightarrow \quad \zeta^2 \le q_{\mathrm{tar}}(q_{\mathrm{tar}} - \rho_C),$$

by squaring $\sqrt{\rho_C^2 + 4\zeta^2} \le 2q_{\mathrm{tar}} - \rho_C$. This proves (B.6). For $\rho_C > 0$,

$$\Psi(0, \rho_C, \zeta) - \rho_C = \frac{2\zeta^2}{\sqrt{\rho_C^2 + 4\zeta^2} + \rho_C} \le \frac{\zeta^2}{\rho_C}.$$

□

Exact stability criterion. The preceding scalar bound is convenient for controlling a target rate. The same block representation also gives an exact criterion for contraction. Since $\mathbf{A} \coloneqq \mathbf{M}^{1/2}\mathbf{H}\mathbf{M}^{1/2} \succ 0$, (B.8) gives $q < 1$ if and only if $2\mathbf{I} - \mathbf{A} \succ 0$. In the basis $\mathbf{Q}$, this matrix is

$$\begin{pmatrix} \mathbf{I} & -\sqrt{\eta}\,\mathbf{B}^{-1/2}\mathbf{E} \\ -\sqrt{\eta}\,\mathbf{E}^\top\mathbf{B}^{-1/2} & 2\mathbf{I} - \eta\mathbf{C} \end{pmatrix}.$$

Taking the Schur complement of its identity block proves

$$q < 1 \quad \Longleftrightarrow \quad \eta\lambda_{\max}(\mathbf{C} + \mathbf{E}^\top\mathbf{B}^{-1}\mathbf{E}) < 2. \tag{B.9}$$

The simpler sufficient condition $\rho_C < 1$ and $\zeta^2 < 1 - \rho_C$ follows from the scalar upper bound. When $\mathbf{P}$ is a leading eigenspace, $\mathbf{E} = 0$ and $q = \rho_\perp$. Thus, $0 < \eta < 2/\lambda_{r+1}$ is necessary and sufficient for contraction in that case, and $\eta = 2/(\lambda_{r+1} + \lambda_d)$ gives $\rho_\perp = (\kappa_\perp - 1)/(\kappa_\perp + 1)$ with $\kappa_\perp \coloneqq \lambda_{r+1}/\lambda_d$. A small residual alone, however, does not ensure that the selected subspace contains the leading eigenvectors. For example, $\mathbf{H} = \mathrm{diag}(100, 1)$ and $\mathbf{P} = \mathbf{e}_2$ have zero residual, but leave the largest eigenvalue in the orthogonal complement. The bound on ρ_C in (B.6) is therefore essential. The next subsection controls both terms by measuring distance to the leading subspace.

B.2 Proof of Theorem 3.1

We now specialize the block bound to a subspace whose projector is within δ of the leading projector. The key distinction is that the diagonal block bounds change by $O(\delta^2)$, whereas the coupling is $O(\delta)$. When the exact-subspace contraction factor is positive, the scalar comparison bound makes the latter contribution quadratic as well.

Proof of Theorem 3.1. Put $\mathbf{\Pi}_\star \coloneqq \mathbf{P}_\star\mathbf{P}_\star^\top$. For a unit vector $\mathbf{p} \in \mathrm{range}(\mathbf{P})$, $\|(\mathbf{I} - \mathbf{\Pi}_\star)\mathbf{p}\|_2 \le \delta$; decomposing along the leading eigenspace and its orthogonal complement gives $\mathbf{p}^\top\mathbf{H}\mathbf{p} \ge b_\delta$, where $b_\delta \coloneqq \lambda_r - (\lambda_r - \lambda_d)\delta^2$. Similarly, a unit $\mathbf{u} \in \mathrm{range}(\mathbf{U})$ satisfies $\|\mathbf{\Pi}_\star\mathbf{u}\|_2 \le \delta$, so $\mathbf{u}^\top\mathbf{H}\mathbf{u} \le c_\delta$, where $c_\delta \coloneqq \lambda_{r+1} + (\lambda_1 - \lambda_{r+1})\delta^2$. Therefore,

$$\lambda_{\min}(\mathbf{B}) \ge b_\delta \ge \lambda_d, \qquad \lambda_d\mathbf{I} \preceq \mathbf{C} \preceq c_\delta\mathbf{I}. \tag{B.10}$$

Write $\mathbf{\Pi} \coloneqq \mathbf{P}\mathbf{P}^\top$ and $[\mathbf{A}, \mathbf{D}] \coloneqq \mathbf{A}\mathbf{D} - \mathbf{D}\mathbf{A}$. The commutator $[\mathbf{H}, \mathbf{\Pi}]$ has blocks $\left(\begin{smallmatrix} 0 & -\mathbf{E} \\ \mathbf{E}^\top & 0 \end{smallmatrix}\right)$ in the $\mathbf{Q}$ basis, and $[\mathbf{H}, \mathbf{\Pi}_\star] = 0$. It follows that

$$\begin{aligned} \|\mathbf{E}\|_2 &= \|[\mathbf{H}, \mathbf{\Pi} - \mathbf{\Pi}_\star]\|_2 \\ &\le 2\left\|\mathbf{H} - \tfrac{1}{2}(\lambda_1 + \lambda_d)\mathbf{I}\right\|_2 \|\mathbf{\Pi} - \mathbf{\Pi}_\star\|_2 = (\lambda_1 - \lambda_d)\delta. \end{aligned}$$

The maximum of $|1 - \eta\lambda|$ over $[\lambda_d, \lambda_{r+1}]$ is $\rho_\perp$. Since this function is η-Lipschitz and the spectrum of $\mathbf{C}$ is contained in $[\lambda_d, c_\delta]$,

$$\rho_C \le \rho_\perp + \eta(\lambda_1 - \lambda_{r+1})\delta^2, \qquad \zeta^2 \le \frac{\eta(\lambda_1 - \lambda_d)^2\delta^2}{b_\delta}.$$

Since $\Psi(0, a, b)$ is nondecreasing in a and b,

$$\widehat{q}(\mathbf{P}, \eta) \le \Psi\left(0, \rho_\perp + \eta(\lambda_1 - \lambda_{r+1})\delta^2, \frac{\sqrt{\eta}\,(\lambda_1 - \lambda_d)\delta}{\sqrt{b_\delta}}\right). \tag{B.11}$$

For $\rho_\perp > 0$, apply $\Psi(0, a, b) \le a + b^2/a$ to this upper bound, using $a \ge \rho_\perp$ and $b_\delta \ge \lambda_d$, to obtain (3.1). This argument does not require $\rho_C \ge \rho_\perp$. □

Subspace coupling when the exact factor is zero. The exact block dynamics also give

$$\left\|\mathbf{u}^+\right\|_2 \le \frac{(\lambda_1 - \lambda_d)\delta}{b_\delta} \left\|\mathbf{v}\right\|_2.$$

For $\rho_\perp = 0$, the last two arguments in (B.11) are $O(\delta^2)$ and $O(\delta)$, respectively. Since $\Psi(0, a, b) \le a + b$, the resulting rate bound is $O(\delta)$. Thus, the second-order bound requires a positive exact-subspace factor.

B.3 Proof of Corollary 3.2

Proof of Corollary 3.2. Condition (3.2) gives $q \le (1 + \rho_\perp)/2 = \kappa_\perp/(\kappa_\perp + 1)$. Consequently, for $0 < \varepsilon < \|\mathbf{e}_0\|_{\mathbf{H}}$, the target $\|\mathbf{e}_T\|_{\mathbf{H}} \le \varepsilon$ in Corollary 3.2 is reached from any initial point within

$$T = \left\lceil \frac{\log(\|\mathbf{e}_0\|_{\mathbf{H}}/\varepsilon)}{\log(1 + 1/\kappa_\perp)} \right\rceil = O(\kappa_\perp \log(\|\mathbf{e}_0\|_{\mathbf{H}}/\varepsilon)) \tag{B.12}$$

iterations. If $\|\mathbf{e}_0\|_{\mathbf{H}} \le \varepsilon$, no outer iteration is needed. □

B.4 Subspace computation and HVP budgets

We next relate the required subspace accuracy to the work used to compute the subspace. For the fixed Hessian, write $\mathbf{Q}_k$ for the basis after k power iterations, starting from $\mathbf{Q}_0$. Suppose $\lambda_r > \lambda_{r+1}$, set $\theta := \lambda_{r+1}/\lambda_r \in (0, 1)$, and let $\Gamma_0 < \infty$ be the tangent of the largest principal angle between the initial and leading subspaces. This requires $\mathbf{P}_\star^\top \mathbf{Q}_0$ to be invertible. The block-power estimate $\delta_K \le \Gamma_0 \theta^K$ (Saad, 2011), proved below, and Theorem 3.1 give, for $\rho_\perp > 0$,

$$q_K \le \rho_\perp + C_\eta \Gamma_0^2 \theta^{2K}, \tag{B.13}$$

where q_K is the contraction factor for the computed subspace and its projected Hessian. Under the stepsize and $\kappa_\perp > 1$ conditions of Corollary 3.2, the accuracy condition (3.2) is ensured by

$$K \ge \left\lceil \frac{\log \max\{1,\, 2C_\eta \Gamma_0^2/(1 - \rho_\perp)\}}{2\log(1/\theta)} \right\rceil. \tag{B.14}$$

Forming the pair costs $r(K + 1)$ scalar HVPs: K block-power rounds and one final block HVP for the projected Hessian.

Derivation of the block-power bound. Choose an orthogonal eigenbasis $[\mathbf{P}_\star, \mathbf{U}_\star]$, with eigenvalue blocks $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$. Write $\mathbf{Z}_1 := \mathbf{P}_\star^\top \mathbf{Q}_0$ and $\mathbf{Z}_2 := \mathbf{U}_\star^\top \mathbf{Q}_0$. Full-rank overlap means $\mathbf{Z}_1$ is invertible. The initial subspace is the graph of $\mathbf{F}_0 := \mathbf{Z}_2 \mathbf{Z}_1^{-1}$, with $\|\mathbf{F}_0\|_2 = \Gamma_0$. Orthonormalization preserves the range of $\mathbf{H}^K \mathbf{Q}_0$, whose graph is

$$\mathbf{F}_K := \mathbf{\Lambda}_2^K \mathbf{F}_0 \mathbf{\Lambda}_1^{-K}, \qquad \|\mathbf{F}_K\|_2 \le \Gamma_0 \theta^K.$$

The singular values of $\mathbf{F}_K$ are the tangents of the principal angles. Consequently,

$$\delta_K = \frac{\|\mathbf{F}_K\|_2}{\sqrt{1 + \|\mathbf{F}_K\|_2^2}} \le \frac{\Gamma_0 \theta^K}{\sqrt{1 + \Gamma_0^2 \theta^{2K}}} \le \Gamma_0 \theta^K. \tag{B.15}$$

Substitution into (3.1) proves (B.13). The budget (B.14) ensures $C_\eta \Gamma_0^2 \theta^{2K} \le (1 - \rho_\perp)/2$ and hence (3.2), including when $\Gamma_0 = 0$. □

B.5 SHARPNESS EXAMPLES

The following family shows that the quadratic dependence cannot in general be improved, and that the zero-factor case has a different order. Take $\mathbf{H} = \mathrm{diag}(a, b)$ with $a > b > 0$, $r = 1$, and $\mathbf{P}_\delta := (\sqrt{1-\delta^2}, \delta)^\top$. Let $\Delta := a - b$ and $\mathbf{U}_\delta := (-\delta, \sqrt{1-\delta^2})^\top$. The blocks associated with $\mathbf{P}_\delta$ are

$$B_\delta = a - \Delta\delta^2, \qquad C_\delta = b + \Delta\delta^2, \qquad E_\delta = -\Delta\delta\sqrt{1-\delta^2}. \tag{B.16}$$

The two eigenvalues of the symmetric error map are $(R_\delta \pm \sqrt{R_\delta^2 + 4g_\delta^2})/2$, where

$$R_\delta := 1 - \eta C_\delta, \qquad g_\delta^2 := \eta E_\delta^2 / B_\delta.$$

Thus, the exact factor, for either sign of R_δ, is

$$q(\delta) = \frac{|R_\delta| + \sqrt{R_\delta^2 + 4g_\delta^2}}{2}. \tag{B.17}$$

If $1/a < \eta < 1/b$, then $\rho := 1 - \eta b > 0$ and, as $\delta \downarrow 0$,

$$R_\delta = \rho - \eta\Delta\delta^2, \qquad g_\delta^2 = \frac{\eta\Delta^2}{a}\delta^2 + O(\delta^4).$$

Expansion of (B.17) gives

$$q(\delta) = \rho + c\delta^2 + O(\delta^4), \qquad c := \frac{\eta b(a-b)(a\eta - 1)}{a(1-\eta b)} > 0. \tag{B.18}$$

Here, $c = -\eta\Delta + \eta\Delta^2/(a\rho)$. Thus, a fixed positive exact-subspace factor can incur a strictly positive loss of order δ^2. When $\eta = 1/b$ and hence $\rho_\perp = 0$, we instead have $R_\delta = -\Delta\delta^2/b$ and $g_\delta = \Delta\delta/\sqrt{ab} + O(\delta^3)$. Substitution into (B.17) proves $q(\delta) = \Delta\delta/\sqrt{ab} + O(\delta^2)$. This linear dependence explains why the assumption $\rho_\perp > 0$ must be retained in the second-order result.

Sharpness of the dependence on the power budget. The same two-dimensional family also shows that the rate of improvement with K is attained. With fixed $0 < \Gamma_0 < \infty$ and $\theta = b/a$, the graph is scalar and its norm is exactly $\Gamma_0\theta^K$. Hence,

$$\delta_K^2 = \frac{\Gamma_0^2\theta^{2K}}{1 + \Gamma_0^2\theta^{2K}} = \Gamma_0^2\theta^{2K} + O(\theta^{4K}).$$

Combining this identity with (B.18) yields

$$q_K - \rho = c\Gamma_0^2\theta^{2K} + O(\theta^{4K}), \qquad c > 0.$$

This family attains the θ^{2K} order for block power. For a fixed Hessian, the resulting pair can be used throughout optimization without additional HVPs. Appendix C addresses the extra work required to track the principal subspace when the Hessian varies.

C PROOFS FOR LOCAL CONVERGENCE OF FRESH PSNG

This appendix proves Proposition 4.4 and the HVP bound in Section 4.1. We first establish local contraction and warm-start tracking, then combine them to prove the proposition and verify parameter feasibility. The two supporting lemmas also apply to lazy PSNG and will be used in Appendix D. We retain the notation and constants of Section 4.

C.1 LOCAL REGULARITY AND CONTRACTION

We first justify the uniform spectral bounds used in Section 4. Under Assumptions 4.1 and 4.2, let $\mathbf{H}_\star := \mathbf{H}(\mathbf{x}^\star)$, whose eigengap at rank r is 2γ. Restricting to $\|\mathbf{x} - \mathbf{x}^\star\|_2 \le \gamma/(2L_H)$ when $L_H > 0$ ensures $\|\mathbf{H}(\mathbf{x}) - \mathbf{H}_\star\|_2 \le \gamma/2$. Weyl's inequality (Horn & Johnson, 2013) then gives

$$\lambda_r(\mathbf{H}(\mathbf{x})) - \lambda_{r+1}(\mathbf{H}(\mathbf{x})) \ge \gamma, \qquad \frac{\lambda_{r+1}(\mathbf{H}(\mathbf{x}))}{\lambda_r(\mathbf{H}(\mathbf{x}))} \le 1 - \frac{\gamma}{L} = \bar{\theta} \in (0, 1).$$

When $L_H = 0$, the Hessian is constant and no such radius restriction is needed. The leading rank-r projector is unique on this neighborhood, although eigenvalues may coincide within either cluster.

The bound on the contraction factor in the orthogonal complement follows locally: choose a fixed $0 < \eta < 2/\lambda_{r+1}(\mathbf{H}_\star)$, so that $\rho_\perp(\mathbf{x}^\star) < 1$. For any $\bar{\rho} \in (\rho_\perp(\mathbf{x}^\star), 1)$, continuity of the ordered eigenvalues gives $\rho_\perp(\mathbf{x}) \le \bar{\rho}$ on a sufficiently small neighborhood. Alternatively, known valid bounds $m_\perp \le \lambda_d(\mathbf{H}(\mathbf{x})) \le \lambda_{r+1}(\mathbf{H}(\mathbf{x})) \le L_\perp$ give

$$\bar{\rho} \coloneqq \max\{|1 - \eta m_\perp|, |1 - \eta L_\perp|\} < 1$$

when $0 < m_\perp \le L_\perp$ and $0 < \eta < 2/L_\perp$. Intersecting the neighborhoods with a small ball preserves convexity; we use $\mathcal{N}$ for the resulting neighborhood in Section 4. These are assumptions on local regularity and spectral information; observed step lengths alone do not establish them.

The next lemma transfers a uniform bound on the linearized contraction factor to convergence of the nonlinear iterates. It uses only local regularity; an eigengap is needed later to maintain subspace accuracy.

Lemma C.1 (Local contraction)**.** *Under Assumption 4.1, consider the PSNG update* (2.2) *from $\mathbf{x}_t \in \mathcal{N}$, using a fresh or stored curvature pair formed in $\mathcal{N}$. Then,*

$$\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2 \le c_s V_t, \tag{C.1}$$

where c_s is defined in Proposition 4.4. For $q_{\mathrm{tar}} \in (0,1)$, let $R > 0$ satisfy (4.4)*, with C_{loc} as defined there. If $V_t \le R$, the full trial segment lies in $\mathcal{N}$. If also $q_t \le q_{\mathrm{tar}}$, then*

$$V_{t+1} \le q_{\mathrm{tar}} V_t + C_{\mathrm{loc}} V_t^2 \le q_{\mathrm{loc}} V_t \le R, \qquad q_{\mathrm{loc}} = \frac{1 + q_{\mathrm{tar}}}{2} < 1.$$

Consequently, if $\mathbf{x}_0 \in \mathcal{N}$, $V_0 \le R$, and $q_t \le q_{\mathrm{tar}}$ at every iteration, then $\mathbf{x}_t \in \mathcal{N}$ and $V_t \le q_{\mathrm{loc}}^t V_0$ for all $t \ge 0$.

The radius conditions keep every trial segment in $\mathcal{N}$ and control both the Taylor remainder and the change in the Hessian norm. The proof bounds the proposed step before using these conditions to establish that its endpoint remains in $\mathcal{N}$. Their product form also covers vanishing constants without division by zero.

Proof of Lemma C.1. For a current fresh or stored pair, the projected Hessian used in the preconditioner is at least $\mu\mathbf{I}$, so $\|\mathbf{M}_t\|_2 \le M_{\max}$, where $M_{\max} = \max\{\mu^{-1}, \eta\}$. For $\mathbf{x}_t \in \mathcal{N}$, convexity of $\mathcal{N}$ and stationarity give

$$\|\mathbf{x}_t - \mathbf{x}^\star\|_2 \le V_t/\sqrt{\mu}, \qquad \|\mathbf{g}_t\|_2 \le L \|\mathbf{x}_t - \mathbf{x}^\star\|_2 .$$

Thus, since $c_s = M_{\max} L/\sqrt{\mu}$, the proposed update $\mathbf{d}_t$ has length at most $c_s V_t$ even before its endpoint is known to lie in $\mathcal{N}$. Taylor's integral formula bounds the gradient Taylor remainder:

$$\mathbf{r}_t \coloneqq \mathbf{g}_t - \mathbf{H}_t \mathbf{e}_t, \qquad \|\mathbf{r}_t\|_2 \le \frac{L_H}{2} \|\mathbf{e}_t\|_2^2 \le \frac{L_H}{2\mu} V_t^2.$$

Therefore,

$$\|\mathbf{e}_{t+1}\|_{\mathbf{H}_t} \le q_t V_t + \sqrt{L}\, M_{\max} \|\mathbf{r}_t\|_2 \le q_t V_t + c_0 V_t^2,$$

where $c_0 \coloneqq L_H \sqrt{L}\, M_{\max}/(2\mu)$. If the line segment from $\mathbf{x}_t$ to $\mathbf{x}_{t+1}$ lies in $\mathcal{N}$, Hessian Lipschitz continuity implies $\mathbf{H}_{t+1} \preceq (1 + c_d V_t)\mathbf{H}_t$, where $c_d \coloneqq (L_H/\mu) c_s$. Taking square roots and using $\sqrt{1+z} \le 1 + z/2$ gives the full recursion

$$V_{t+1} \le (1 + c_d V_t/2)(q_t V_t + c_0 V_t^2). \tag{C.2}$$

The term $c_0 V_t^2$ bounds the Taylor remainder, and the prefactor accounts for the change from the $\mathbf{H}_t$-norm to the $\mathbf{H}_{t+1}$-norm.

If $V_t \le R$, the step-length estimate and the first condition in (4.4) put this entire segment in $\mathcal{N}$. If also $q_t \le q_{\mathrm{tar}} < 1$, expansion of the recursion gives

$$\begin{aligned} V_{t+1} &\le q_{\mathrm{tar}} V_t + (c_0 + c_d q_{\mathrm{tar}}/2) V_t^2 + (c_d c_0/2) V_t^3 \\ &\le q_{\mathrm{tar}} V_t + C_{\mathrm{loc}} V_t^2 \le \big[q_{\mathrm{tar}} + C_{\mathrm{loc}} R\big] V_t \le \frac{1 + q_{\mathrm{tar}}}{2} V_t. \end{aligned}$$

The second inequality uses $C_{\mathrm{loc}} = c_0 + c_d$ and $c_0 V_t \le c_0 R \le (1 - q_{\mathrm{tar}})/2 < 1$, which follows from (4.4) and $c_0 \le C_{\mathrm{loc}}$. Starting from $\mathbf{x}_0 \in \mathcal{N}$ and $V_0 \le R$, induction gives $\mathbf{x}_t \in \mathcal{N}$ and $V_t \le q_{\mathrm{loc}}^t V_0$ whenever $q_t \le q_{\mathrm{tar}}$ at every iteration. If $V_t = 0$, the iterate remains at $\mathbf{x}^\star$. □

The initialization bound $\|\mathbf{x}_0 - \mathbf{x}^\star\|_2 \le R/\sqrt{L}$ ensures $\mathbf{x}_0 \in \mathcal{N}$ and $V_0 \le R$. Since $\sqrt{\mu}\,\|\mathbf{e}_t\|_2 \le V_t \le \sqrt{L}\,\|\mathbf{e}_t\|_2$, the lemma also gives $\|\mathbf{x}_t - \mathbf{x}^\star\|_2 \le q_{\mathrm{loc}}^t V_0/\sqrt{\mu}$ and $f(\mathbf{x}_t) - f(\mathbf{x}^\star) \le L q_{\mathrm{loc}}^{2t} V_0^2/(2\mu)$.

Gradient stopping test. Although V_t is convenient for analysis, stopping can be based on the observable gradient. For either local mode, strong convexity in $\mathcal{N}$ gives $\|\mathbf{x}_t - \mathbf{x}^\star\|_2 \le \|\mathbf{g}_t\|_2/\mu$. The gradient test $\|\mathbf{g}_t\|_2 \le \mu\varepsilon/\sqrt{L}$ ensures $V_t \le \varepsilon$, since $V_t \le \sqrt{L}\,\|\mathbf{g}_t\|_2/\mu$ in $\mathcal{N}$.

It remains to ensure $q_t \le q_{\mathrm{tar}}$ throughout the trajectory. For fresh updates this follows from a sufficiently accurate current subspace; for lazy updates we must also control the change in the stored projected Hessian. The following tracking bound supports both cases; Appendix D adds the estimates needed during reuse.

C.2 Initialization and warm-start tracking

Under Assumption 4.3, the block-power estimate (B.15) gives $\delta_0 \le \Gamma\bar{\theta}^K$ after K rounds at $\mathbf{x}_0$. The budget (4.2) therefore ensures $\delta_0 \le \delta_{\mathrm{tol}}$; it gives zero rounds when $\Gamma \le \delta_{\mathrm{tol}}$, including $\Gamma = 0$. One additional block HVP forms the projected Hessian.

At a later computation, the previous basis remains a useful starting point when the intervening Hessian change is small. The following lemma quantifies this fact for both fresh and lazy PSNG.

Lemma C.2 (Warm-start subspace tracking)**.** *Under Assumptions 4.1 and 4.2, let $\mathbf{x}_s, \mathbf{x}_t \in \mathcal{N}$ and $h \coloneqq \|\mathbf{H}_t - \mathbf{H}_s\|_2 < \gamma/2$. Suppose that the stored orthonormal basis $\mathbf{P}_s$ has projector error δ_s at s and that $\bar{\delta}_{s,t} \coloneqq \delta_s + 2h/\gamma < 1$. Starting block power iteration for $\mathbf{H}_t$ from $\mathbf{P}_s$, after any integer $K \ge 0$ the output has projector error*

$$\delta_t^{\mathrm{new}} \le \frac{\bar{\delta}_{s,t}}{\sqrt{1 - \bar{\delta}_{s,t}^2}}\bar{\theta}^K. \tag{C.3}$$

Proof. The local spectral bounds established in Appendix C.1 imply that the leading eigenvalues of $\mathbf{H}_t$ and the remaining eigenvalues of $\mathbf{H}_s$ are separated by at least $\gamma - h > \gamma/2$. The Davis–Kahan sine theorem (Davis & Kahan, 1970) therefore gives

$$\|\mathbf{\Pi}_{\star,t} - \mathbf{\Pi}_{\star,s}\|_2 \le \frac{2h}{\gamma}. \tag{C.4}$$

The triangle inequality bounds the stored basis's error relative to the current leading projector by $\bar{\delta}_{s,t}$. Since this is less than one, the overlap is full rank and its tangent factor is at most $\bar{\delta}_{s,t}/\sqrt{1 - \bar{\delta}_{s,t}^2}$. Applying (B.15) with the eigenvalue ratio bounded by $\bar{\theta}$ proves (C.3). □

For fresh updates with $\mathbf{x}_t, \mathbf{x}_{t+1} \in \mathcal{N}$ and $\|\mathbf{H}_{t+1} - \mathbf{H}_t\|_2 < \gamma/2$, combining this estimate with Hessian Lipschitz continuity gives

$$\left\|\mathbf{P}_t\mathbf{P}_t^\top - \mathbf{\Pi}_{\star,t+1}\right\|_2 \le \delta_t + \frac{2L_H}{\gamma}\,\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2\,. \tag{C.5}$$

This bounds the starting error for each fresh update. For lazy updates, the same lemma uses $h \le L_H \ell_{s,t}$; Appendix D.3 verifies its hypotheses even when a step crosses the refresh threshold.

C.3 Proof of Proposition 4.4

To prove Proposition 4.4, we also need the contraction bound for a fresh pair. The quadratic block estimates in Appendices B.1–B.2 give

$$q_t \le \Psi\Big(0, \bar{\rho} + \eta(L-\mu)\delta_t^2, \sqrt{\eta/\mu}\,(L-\mu)\delta_t\Big)\,. \tag{C.6}$$

For $a, b \ge 0$ and $q_{\mathrm{tar}} > 0$, the condition $\Psi(0,a,b) \le q_{\mathrm{tar}}$ is equivalent to $a + b^2/q_{\mathrm{tar}} \le q_{\mathrm{tar}}$. Applying this with $a = \bar{\rho} + \eta(L-\mu)\delta_{\mathrm{tol}}^2$ and $b = \sqrt{\eta/\mu}\,(L-\mu)\delta_{\mathrm{tol}}$ gives exactly (4.3).

The initialization budget is justified in Appendix C.2. We now verify that the refresh budget (4.5) in Proposition 4.4 restores the required subspace accuracy at every iteration.

Proof of Proposition 4.4. The initialization gives $\delta_0 \le \delta_{\mathrm{tol}}$. For any current fresh pair of this accuracy, the angle-block estimates give $\rho_C \le \bar{\rho} + \eta(L-\mu)\delta_{\mathrm{tol}}^2$ and $\zeta \le \sqrt{\eta/\mu}\,(L-\mu)\delta_{\mathrm{tol}}$. Condition (4.3) and (C.6) therefore imply $q_t \le q_{\mathrm{tar}}$. Lemma C.1 keeps the trial segment in $\mathcal{N}$, gives $V_{t+1} \le q_{\mathrm{loc}} V_t \le R$, and bounds its length by $c_s R$. Because $\bar{\delta} < 1$, $L_H c_s R < \gamma/2$; Lemma C.2 bounds the next warm-start error by $\bar{\delta}$ and gives

$$\delta_{t+1} \le \frac{\bar{\delta}}{\sqrt{1-\bar{\delta}^2}}\bar{\theta}^{K_{\mathrm{ref}}}.$$

The budget (4.5) makes this upper bound at most δ_{tol}. Induction proves the result. □

For $T \ge 1$ outer updates, fresh PSNG uses

$$K_{\mathrm{boot}} + 1 + (T-1)(K_{\mathrm{ref}} + 1)$$

block HVPs. Every refresh includes the final projected-Hessian product. For $0 < \varepsilon < V_0$, taking $T = \lceil \log(V_0/\varepsilon)/\log(1/q_{\mathrm{loc}}) \rceil$ is sufficient. Thus, the fresh-mode upper bound grows as $O(\log(1/\varepsilon))$. The next subsection verifies that the accuracy and radius conditions can be met with parameters independent of the target tolerance.

C.4 Feasible parameters and HVP budgets

Fix $q_{\mathrm{tar}} \in (\bar{\rho}, 1)$. The left side of (4.3) tends to $\bar{\rho}$ as $\delta_{\mathrm{tol}} \downarrow 0$, so a tolerance $\delta_{\mathrm{tol}} \in (0,1)$ satisfying that condition exists. Let $r_{\mathcal{N}} > 0$ satisfy $\overline{\mathbb{B}}_2(\mathbf{x}^\star, r_{\mathcal{N}}) \subseteq \mathcal{N}$. For $L_H > 0$, a sufficient radius choice is

$$0 < R < \min\left\{ \frac{r_{\mathcal{N}}}{\mu^{-1/2} + c_s}, \quad \frac{1-q_{\mathrm{tar}}}{2C_{\mathrm{loc}}}, \quad \frac{\gamma(1-\delta_{\mathrm{tol}})}{2L_H c_s} \right\}.$$

The first two bounds imply (4.4); the third ensures $\bar{\delta} = \delta_{\mathrm{tol}} + 2L_H c_s R/\gamma < 1$ in Proposition 4.4. If $L_H = 0$, then $C_{\mathrm{loc}} = 0$ and $\bar{\delta} = \delta_{\mathrm{tol}}$, so only the neighborhood bound on R is needed. In either case, $\|\mathbf{x}_0 - \mathbf{x}^\star\|_2 \le R/\sqrt{L}$ ensures a valid local starting point.

Since $\log(1/\bar{\theta}) = -\log(1-\gamma/L) \ge \gamma/L$, the ceiling formulas (4.2) and (4.5) give

$$K_{\mathrm{boot}} \le 1 + \frac{L}{\gamma}\log\max\{1, \Gamma/\delta_{\mathrm{tol}}\},$$

$$K_{\mathrm{ref}} \le 1 + \frac{L}{\gamma}\log\left(\frac{\bar{\delta}}{\delta_{\mathrm{tol}}\sqrt{1-\bar{\delta}^2}}\right).$$

Here, the second logarithm is positive since $0 < \delta_{\mathrm{tol}} \le \bar{\delta} < 1$. A small relative eigengap bound γ/L, poor initial overlap (large Γ), or $\bar{\delta}$ close to one can make these budgets large. The subspace tolerance δ_{tol} can be fixed independently of ε.

These choices depend on the regularity, spectral, neighborhood, and initialization bounds, but not on the target accuracy. They establish the fixed per-iteration HVP budget claimed in Section 4.1. Appendix D extends the argument to reuse between refreshes.

D Proofs for Local Convergence of Lazy PSNG

This appendix proves Theorem 4.5 from Section 4.2. We use the local contraction and warm-start tracking lemmas from Appendix C. The additional argument bounds the error from reusing both the subspace and projected Hessian by accumulated step length. We first derive this bound, then prove convergence and the finite total HVP bound, and finally verify parameter feasibility.

D.1 Block contraction with a stored projected Hessian

Reusing a pair introduces an additional error because its projected Hessian no longer equals the current compression onto the stored subspace. We isolate this error before relating it to the accumulated step length since the last refresh. The following bound allows any current $\mathbf{H} \succ 0$, orthonormal splitting $\mathbf{Q} \coloneqq [\mathbf{P}, \mathbf{U}]$, stored $\widetilde{\mathbf{B}} \succ 0$, and $\eta > 0$. Set $\mathbf{B} \coloneqq \mathbf{P}^\top \mathbf{H}\mathbf{P}$, $\mathbf{C} \coloneqq \mathbf{U}^\top \mathbf{H}\mathbf{U}$, and $\mathbf{E} \coloneqq \mathbf{P}^\top \mathbf{H}\mathbf{U}$.

Define ρ_B for the projected-Hessian reuse error, ρ_C for the contraction factor in the orthogonal complement, and ζ for their coupling:

$$\rho_B \coloneqq \left\| \mathbf{I} - \widetilde{\mathbf{B}}^{-1/2} \mathbf{B} \widetilde{\mathbf{B}}^{-1/2} \right\|_2, \quad \rho_C \coloneqq \|\mathbf{I} - \eta \mathbf{C}\|_2, \\ \zeta \coloneqq \sqrt{\eta} \left\| \widetilde{\mathbf{B}}^{-1/2} \mathbf{E} \right\|_2. \tag{D.1}$$

We use Ψ as defined in Appendix B.1.

Lemma D.1 (Block contraction bound)**.** *For* $\widetilde{\mathbf{M}} \coloneqq \mathbf{P}\widetilde{\mathbf{B}}^{-1}\mathbf{P}^\top + \eta(\mathbf{I} - \mathbf{P}\mathbf{P}^\top)$*,*

$$\left\| \mathbf{I} - \widetilde{\mathbf{M}}\mathbf{H} \right\|_{\mathbf{H}} \leq \Psi(\rho_B, \rho_C, \zeta). \tag{D.2}$$

Proof. Let $\widetilde{\mathbf{T}} \coloneqq \mathbf{I} - \widetilde{\mathbf{M}}\mathbf{H}$. Then,

$$\mathbf{Q}^\top \widetilde{\mathbf{M}}^{-1/2} \widetilde{\mathbf{T}} \widetilde{\mathbf{M}}^{1/2} \mathbf{Q} = \begin{pmatrix} \mathbf{I} - \widetilde{\mathbf{B}}^{-1/2} \mathbf{B} \widetilde{\mathbf{B}}^{-1/2} & -\sqrt{\eta}\, \widetilde{\mathbf{B}}^{-1/2} \mathbf{E} \\ -\sqrt{\eta}\, \mathbf{E}^\top \widetilde{\mathbf{B}}^{-1/2} & \mathbf{I} - \eta \mathbf{C} \end{pmatrix}. \tag{D.3}$$

The norm of this symmetric matrix equals $\left\|\widetilde{\mathbf{T}}\right\|_{\mathbf{H}} = \left\|\widetilde{\mathbf{T}}\right\|_{\widetilde{\mathbf{M}}^{-1}}$ by the weighted self-adjointness argument in Appendix B.1. For a block vector $(\mathbf{u}, \mathbf{v})$, the absolute Rayleigh form is at most

$$\rho_B \|\mathbf{u}\|_2^2 + 2\zeta \|\mathbf{u}\|_2 \|\mathbf{v}\|_2 + \rho_C \|\mathbf{v}\|_2^2.$$

The comparison matrix $\left(\begin{smallmatrix} \rho_B & \zeta \\ \zeta & \rho_C \end{smallmatrix}\right)$ has largest eigenvalue $\Psi(\rho_B, \rho_C, \zeta)$, which proves (D.2). □

Conditions for a prescribed contraction factor. The comparison bound can be checked without computing the eigenvalues of the full preconditioned Hessian. For $a, c, z \geq 0$ and $q_{\mathrm{tar}} \geq 0$, the scalar condition is

$$\Psi(a, c, z) \leq q_{\mathrm{tar}} \quad \Longleftrightarrow \quad a \leq q_{\mathrm{tar}}, \quad c \leq q_{\mathrm{tar}}, \quad z^2 \leq (q_{\mathrm{tar}} - a)(q_{\mathrm{tar}} - c). \tag{D.4}$$

Indeed, these are exactly the conditions for $\left(\begin{smallmatrix} q_{\mathrm{tar}} - a & -z \\ -z & q_{\mathrm{tar}} - c \end{smallmatrix}\right)$ to be positive semidefinite. This also shows that Ψ is nondecreasing in its three nonnegative arguments, alternatively by the Rayleigh characterization of the nonnegative comparison matrix.

For completeness, the exact contraction criterion is

$$2\widetilde{\mathbf{B}} - \mathbf{B} \succ 0, \qquad \eta \lambda_{\max}\Big(\mathbf{C} + \mathbf{E}^\top (2\widetilde{\mathbf{B}} - \mathbf{B})^{-1} \mathbf{E}\Big) < 2. \tag{D.5}$$

To see this, the positive-definite matrix $\widetilde{\mathbf{M}}^{1/2} \mathbf{H} \widetilde{\mathbf{M}}^{1/2}$ must have all eigenvalues below two. In the $\mathbf{Q}$ basis, the Schur complement of the leading block of $2\mathbf{I} - \widetilde{\mathbf{M}}^{1/2} \mathbf{H} \widetilde{\mathbf{M}}^{1/2}$ gives precisely (D.5). For the refresh policy, we use the sufficient comparison bound instead: each of its three arguments admits a direct upper bound in terms of the accumulated step length.

D.2 Contraction controlled by accumulated step length

We now apply the block bound to a stored pair along the optimization trajectory. The accumulated step length controls each error component, yielding the function Φ used in Theorem 4.5.

Lemma D.2 (Contraction bound with reused curvature)**.** *Under Assumptions 4.1 and 4.2, suppose that the pair formed at s satisfies $\delta_s \leq \delta_{\mathrm{tol}}$ and is used at $t \geq s$. If $\mathbf{x}_k \in \mathcal{N}$ for $s \leq k \leq t$, then*

$$q_t \leq \Phi(\ell_{s,t}). \tag{D.6}$$

Proof of Lemma D.2. For a curvature pair formed at s, let $\widetilde{\mathbf{M}}_s$ denote its preconditioner and define $q_{t|s} \coloneqq \left\| \mathbf{I} - \widetilde{\mathbf{M}}_s \mathbf{H}_t \right\|_{\mathbf{H}_t}$. This equals q_t whenever that pair is used at iteration t. Write $\Omega_{s,t} \coloneqq L_H \ell_{s,t}$. Fix a refresh pair $(\mathbf{P}_s, \widetilde{\mathbf{B}}_s)$ with $\widetilde{\mathbf{B}}_s \coloneqq \mathbf{P}_s^\top \mathbf{H}_s \mathbf{P}_s$. Choose $\mathbf{U}_s$ with orthonormal columns spanning $\mathrm{range}(\mathbf{P}_s)^\perp$. For $t \geq s$, write $\mathbf{B}_{t|s} \coloneqq \mathbf{P}_s^\top \mathbf{H}_t \mathbf{P}_s$, $\mathbf{C}_{t|s} \coloneqq \mathbf{U}_s^\top \mathbf{H}_t \mathbf{U}_s$, and $\mathbf{E}_{t|s} \coloneqq$

$\mathbf{P}_s^\top \mathbf{H}_t \mathbf{U}_s$ for the current blocks. Subscripts $t \mid s$ on the quantities in (D.1) indicate this current Hessian and stored pair. At the last refresh, define

$$b_s \coloneqq \lambda_{\min}(\widetilde{\mathbf{B}}_s), \qquad r_s \coloneqq \|\mathcal{R}(\mathbf{P}_s; \mathbf{H}_s)\|_2, \qquad \rho_{C,s|s} \coloneqq \left\|\mathbf{I} - \eta \mathbf{C}_{s|s}\right\|_2.$$

If $\mathbf{x}_k \in \mathcal{N}$ along the refresh interval, then $\|\mathbf{H}_t - \mathbf{H}_s\|_2 \le \Omega_{s,t}$. Compression onto the fixed subspaces cannot increase the operator norm, and the invariance residual satisfies

$$\|\mathcal{R}(\mathbf{P}_s; \mathbf{H}_t)\|_2 \le r_s + \|\mathbf{H}_t - \mathbf{H}_s\|_2.$$

It follows that

$$\rho_{B,t|s} \le \frac{\Omega_{s,t}}{b_s}, \quad \rho_{C,t|s} \le \rho_{C,s|s} + \eta\Omega_{s,t}, \quad \zeta_{t|s} \le \sqrt{\frac{\eta}{b_s}}(r_s + \Omega_{s,t}). \tag{D.7}$$

Here, the first inequality uses $\mathbf{B}_{t|s} - \widetilde{\mathbf{B}}_s = \mathbf{P}_s^\top(\mathbf{H}_t - \mathbf{H}_s)\mathbf{P}_s$. Consequently, a potentially tighter bound for the current refresh interval is

$$q_{t|s} \le \Psi\left(\frac{\Omega_{s,t}}{b_s}, \rho_{C,s|s} + \eta\Omega_{s,t}, \sqrt{\frac{\eta}{b_s}}(r_s + \Omega_{s,t})\right).$$

If $\delta_s \le \delta_{\mathrm{tol}}$, the spectral-block estimates in Appendix B.2 give the uniform refresh data

$$b_s \ge \mu, \qquad r_s \le (L-\mu)\delta_{\mathrm{tol}}, \qquad \rho_{C,s|s} \le \bar{\rho} + \eta(L-\mu)\delta_{\mathrm{tol}}^2. \tag{D.8}$$

Applying Lemma D.1 and monotonicity of Ψ gives

$$q_{t|s} \le \Psi\left(\frac{\Omega_{s,t}}{\mu}, \bar{\rho} + \eta(L-\mu)\delta_{\mathrm{tol}}^2 + \eta\Omega_{s,t}, \sqrt{\frac{\eta}{\mu}}((L-\mu)\delta_{\mathrm{tol}} + \Omega_{s,t})\right) = \Phi(\ell_{s,t}),$$

using $\Omega_{s,t} = L_H \ell_{s,t}$ and the definition of Ψ as the largest eigenvalue of the comparison matrix. This proves (D.6). Since $L_H \ge 0$, the entries of the nonnegative comparison matrix are nondecreasing in ℓ, so $\Phi(\ell)$ is also nondecreasing. □

In the matrix defining Φ in (4.6), the upper-left entry accounts for the mismatch between the stored and current projected Hessians; the lower-right entry bounds the contraction factor of the gradient update in the orthogonal complement. The off-diagonal entry bounds how errors in one subspace affect the update in the other. The terms proportional to $L_H\ell$ account for Hessian changes since the last refresh.

At $\ell = 0$, the upper-left entry vanishes, so (D.4) shows that $\Phi(0) \le q_{\mathrm{tar}}$ is equivalent to (4.3). Thus, Φ accounts for both the initial subspace approximation and the subsequent curvature change. This contraction bound does not require the Hessian change to be small or the current subspace angle to be known; small Hessian changes are needed to control the warm-start budget at the next refresh.

D.3 Proof of Theorem 4.5

The proof has three parts: contraction during reuse, restoration of subspace accuracy when the threshold is crossed, and a bound on the number of refreshes obtained from the sum of step lengths. The first two are proved jointly by induction, so the argument does not assume that future iterates remain in $\mathcal{N}$.

Proof of Theorem 4.5. We prove jointly that the iterates remain in $\mathcal{N}$, their error V_t is at most R, and each refresh interval begins with $\delta_s \le \delta_{\mathrm{tol}}$. The prescribed initialization establishes these properties at $t = s = 0$. For an iterate with $\ell_{s,t} < \omega$, Lemma D.2 and monotonicity give

$$q_{t|s} \le \Phi(\ell_{s,t}) \le \Phi(\omega) \le q_{\mathrm{tar}}.$$

The preconditioner was formed at a point in $\mathcal{N}$. Lemma C.1 therefore places the trial segment inside $\mathcal{N}$ and gives $V_{t+1} \le q_{\mathrm{loc}} V_t \le R$.

Suppose next that the step-length threshold is first reached at $t > s$. The preceding update started with $\ell_{s,t-1} < \omega$ and moved the iterate by $\|\mathbf{x}_t - \mathbf{x}_{t-1}\|_2 \le c_s R$. Hence,

$$\ell_{s,t} < \omega + c_s R, \qquad \|\mathbf{H}_t - \mathbf{H}_s\|_2 \le L_H \ell_{s,t} \le L_H(\omega + c_s R). \tag{D.9}$$

Since $\bar{\delta} < 1$, this Hessian drift is strictly less than $\gamma/2$. Lemma C.2 bounds the current projector error of the stored basis by $\bar{\delta}$ and its tangent factor by $\bar{\delta}/\sqrt{1-\bar{\delta}^2}$. The budget (4.8) therefore returns a basis with $\delta_t \le \delta_{\mathrm{tol}}$. Form its exact projected Hessian and reset the refresh interval before taking a step. At this point the bound is evaluated at zero, and

$$q_{t|t} \le \Phi(0) \le \Phi(\omega) \le q_{\mathrm{tar}}.$$

Lemma C.1 gives the same contraction for the refreshed step. This closes the induction: the iterate reached by the threshold-crossing step remains in the local region, and the refreshed pair gives a valid contraction bound for the following step. This proves (4.9).

The convergence estimate now bounds the number of refreshes. For every finite horizon T, the sum of the first T step lengths satisfies

$$\sum_{t=0}^{T-1} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2 \le c_s \sum_{t=0}^{T-1} V_t \le \frac{c_s V_0}{1 - q_{\mathrm{loc}}}.$$

Each post-initialization refresh closes a disjoint refresh interval with accumulated step length at least ω. Thus, the number of such refreshes through any T is at most $c_s V_0/[(1-q_{\mathrm{loc}})\omega]$. Let N_{ref} count all refreshes, including initialization. The same finite bound holds for the infinite trajectory, so $N_{\mathrm{ref}} \le 1 + c_s V_0/[(1-q_{\mathrm{loc}})\omega]$, proving the stated refresh-count bound. If $c_s V_0/(1-q_{\mathrm{loc}}) < \omega$, the entire path is shorter than the trigger threshold, so no refresh occurs after initialization.

Finally, we translate the refresh count into a curvature cost. The initialization costs $K_{\mathrm{boot}} + 1$ block HVPs, and each subsequent refresh costs $K_{\mathrm{ref}} + 1$; the final $+1$ always forms the projected Hessian for the returned basis. There are no HVPs between refreshes under the stated policy, so the total block-HVP count is at most

$$K_{\mathrm{boot}} + 1 + (K_{\mathrm{ref}} + 1)(N_{\mathrm{ref}} - 1).$$

Substituting the refresh-count bound above proves (4.10). Multiplying by r gives the number of scalar HVPs. For $0 < \varepsilon < V_0$, taking

$$T = \left\lceil \frac{\log(V_0/\varepsilon)}{\log(1/q_{\mathrm{loc}})} \right\rceil$$

outer iterations ensures $V_T \le \varepsilon$. If $V_0 \le \varepsilon$, no outer step is needed. The curvature bound may still include the initialization and remains a valid upper bound. □

D.4 Feasible parameters and HVP budgets

We verify that the parameters of Theorem 4.5 can be chosen independently of the final accuracy. Fix $q_{\mathrm{tar}} \in (\bar{\rho}, 1)$ and choose $\delta_{\mathrm{tol}} \in (0,1)$ so that $\Phi(0) < q_{\mathrm{tar}}$. Such a choice exists because $\Phi(0) \to \bar{\rho}$ as $\delta_{\mathrm{tol}} \downarrow 0$. Let $r_{\mathcal{N}} > 0$ satisfy $\overline{\mathbb{B}}_2(\mathbf{x}^\star, r_{\mathcal{N}}) \subseteq \mathcal{N}$.

For $L_H > 0$, continuity of Φ permits a fixed $\omega > 0$ satisfying $\Phi(\omega) \le q_{\mathrm{tar}}$ and $2L_H\omega < \gamma(1-\delta_{\mathrm{tol}})$. A sufficient radius choice is

$$0 < R < \min\left\{ \frac{r_{\mathcal{N}}}{\mu^{-1/2} + c_s}, \quad \frac{1 - q_{\mathrm{tar}}}{2C_{\mathrm{loc}}}, \quad \frac{\gamma(1-\delta_{\mathrm{tol}}) - 2L_H\omega}{2L_H c_s} \right\}.$$

The first two bounds imply (4.4); the third ensures $\bar{\delta} = \delta_{\mathrm{tol}} + 2L_H(\omega + c_s R)/\gamma < 1$ as required by (4.7). When $L_H = 0$, we have $C_{\mathrm{loc}} = 0$, $\bar{\delta} = \delta_{\mathrm{tol}}$, and $\Phi(\omega) = \Phi(0)$ for every $\omega > 0$. Only the neighborhood bound on R is then needed. With the prescribed subspace initialization, $\|\mathbf{x}_0 - \mathbf{x}^\star\|_2 \le R/\sqrt{L}$ is sufficient for the theorem to apply.

The initialization budget is justified in Appendix C.2. At a later refresh, (D.9) and Lemma C.2 bound the initial tangent factor by $\bar{\delta}/\sqrt{1-\bar{\delta}^2}$. Applying (B.15) gives exactly (4.8). The budget estimates in Appendix C.4 hold with this lazy-policy value of $\bar{\delta}$ as well. A larger ω permits longer reuse but increases the bound on the starting subspace error and may increase K_{ref}.

All these choices use fixed regularity, spectral, neighborhood, and initialization bounds. Together with the finite refresh count, they give a total HVP bound independent of the target accuracy. Appendix E explains how a global Armijo phase reaches a valid local starting point and how to initialize the method there.

E PROOFS FOR GLOBAL CONVERGENCE

This appendix proves Theorem 4.7 and the transition to local convergence described in Section 4.3. Uniform positive definiteness of the preconditioner suffices for global Armijo convergence. The local spectral and curvature-cost guarantees additionally require accurate subspace initialization, control of curvature changes, and a margin that ensures Armijo acceptance of $\alpha_t = 1$. We prove these claims in that order and then account for the work before and after the local restart.

E.1 PROOF OF THEOREM 4.7

We first verify that every stored pair yields a uniformly bounded positive-definite preconditioner under the global Hessian bounds. The standard descent argument then applies regardless of the age or accuracy of its subspace.

Proof of Theorem 4.7. Fresh and stored projected Hessians have spectra in $[\mu, L]$. If the pair used at t was formed at s, the blocks of $\mathbf{M}_t$ are $\widetilde{\mathbf{B}}_s^{-1}$ and $\eta\mathbf{I}$. Hence, $m_{\min}\mathbf{I} \preceq \mathbf{M}_t \preceq M_{\max}\mathbf{I}$, independently of subspace accuracy or age. Set $h_t := \mathbf{g}_t^\top \mathbf{M}_t \mathbf{g}_t$. Then, $\mathbf{g}_t^\top \mathbf{d}_t = -h_t$ and $\|\mathbf{d}_t\|_2^2 \le M_{\max} h_t$. Global smoothness gives

$$f(\mathbf{x}_t + \alpha \mathbf{d}_t) \le f(\mathbf{x}_t) - \alpha \left(1 - \frac{L M_{\max} \alpha}{2}\right) h_t.$$

Thus, every $\alpha \le 2(1 - c_A)/(L M_{\max})$ satisfies Armijo. Backtracking from $\alpha = 1$ terminates and returns a coefficient $\alpha_t \ge \alpha_{\min}$, as stated in (4.11). For $\mathbf{g}_t = 0$, $\alpha_t = 1$ is accepted immediately.

Strong convexity implies $\|\mathbf{g}_t\|_2^2 \ge 2\mu\Delta_t$. Together with Armijo and $h_t \ge m_{\min}\|\mathbf{g}_t\|_2^2$, this proves $\Delta_{t+1} \le q_{\mathrm{G}}\Delta_t$. Since $\mu m_{\min} \le 1$, $\alpha_{\min} \le \beta < 1$, and $2c_A < 1$, we have $0 < q_{\mathrm{G}} < 1$. Finally, $\|\mathbf{x}_t - \mathbf{x}^\star\|_2^2 \le 2\Delta_t/\mu$ implies convergence of the iterates to the unique minimizer. □

E.2 ARMIJO ACCEPTANCE WITHOUT DAMPING

Global convergence brings the iterates close to the minimizer; applying the local results additionally requires $\alpha_t = 1$, so that $\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{d}_t$. The next proposition supplies the additional spectral margin and bound on $\|\mathbf{d}_t\|_2$ needed to identify the Armijo iteration with the local iteration.

Proposition E.1 (Sufficient conditions for accepting a unit stepsize)**.** *Under the assumptions of Theorem 4.7, suppose that the Hessian is L_H-Lipschitz on the local neighborhood $\mathcal{N}$ of Section 4. Fix $q_{\mathrm{tar}} \ge 0$. If the line segment from $\mathbf{x}_t$ to $\mathbf{x}_t + \mathbf{d}_t$ lies in $\mathcal{N}$ and*

$$q_t \le q_{\mathrm{tar}} < 1 - 2c_A, \qquad L_H M_{\max}\|\mathbf{d}_t\|_2 \le 3(1 - 2c_A - q_{\mathrm{tar}}), \tag{E.1}$$

then the Armijo line search accepts $\alpha_t = 1$. If $q_t \le q_{\mathrm{tar}} < 1 - 2c_A$ holds for all sufficiently large t, $\alpha_t = 1$ at every sufficiently large iteration.

Proof of Proposition E.1. Let $\mathbf{p}_t := \mathbf{M}_t^{1/2}\mathbf{g}_t$ and $\mathbf{A}_t := \mathbf{M}_t^{1/2}\mathbf{H}_t\mathbf{M}_t^{1/2}$. The weighted self-adjointness argument in Appendix B.1 gives $q_t = \|\mathbf{I} - \mathbf{A}_t\|_2$, so $\mathbf{A}_t \preceq (1 + q_{\mathrm{tar}})\mathbf{I}$. Taylor's theorem on the full trial segment yields

$$\begin{aligned} f(\mathbf{x}_t + \mathbf{d}_t) &\le f(\mathbf{x}_t) - \|\mathbf{p}_t\|_2^2 + \tfrac{1}{2}\mathbf{p}_t^\top \mathbf{A}_t \mathbf{p}_t + \tfrac{L_H}{6}\|\mathbf{d}_t\|_2^3 \\ &\le f(\mathbf{x}_t) - \tfrac{1 - q_{\mathrm{tar}}}{2}\|\mathbf{p}_t\|_2^2 + \tfrac{L_H M_{\max}}{6}\|\mathbf{d}_t\|_2 \|\mathbf{p}_t\|_2^2 . \end{aligned}$$

The second line uses $\|\mathbf{d}_t\|_2^2 \le M_{\max}\|\mathbf{p}_t\|_2^2$. The bound on $\|\mathbf{d}_t\|_2$ in (E.1) makes the right side at most $f(\mathbf{x}_t) - c_A\|\mathbf{p}_t\|_2^2$, showing that $\alpha_t = 1$ satisfies the Armijo condition. The product form covers $L_H = 0$ without division by zero.

For the eventual statement, Theorem 4.7 gives $\mathbf{x}_t \to \mathbf{x}^\star$ and $\mathbf{g}_t \to 0$. The uniform preconditioner bound implies $\|\mathbf{d}_t\|_2 \le M_{\max}\|\mathbf{g}_t\|_2 \to 0$. Because $\mathcal{N}$ is a neighborhood of $\mathbf{x}^\star$, full trial segments eventually lie in $\mathcal{N}$ and the second condition in (E.1) holds. The persistent spectral margin then ensures $\alpha_t = 1$ at every subsequent iteration. □

Applying the condition throughout the local phase. For fresh PSNG, (4.3) and (C.6) ensure $q_t \le q_{\mathrm{tar}}$; for lazy PSNG, Lemma D.2 and $\Phi(\omega) \le q_{\mathrm{tar}}$ give the same bound. Lemma C.1 keeps each trial segment in $\mathcal{N}$ and bounds its length by $c_s R$. If $\bar{\rho} < q_{\mathrm{tar}} < 1 - 2c_A$ and $L_H M_{\max} c_s R \le 3(1 - 2c_A - q_{\mathrm{tar}})$, Proposition E.1 guarantees $\alpha_t = 1$. Thus, after a valid local initialization, the Armijo iterates coincide with the corresponding local PSNG iterates. The next proposition gives an observable entry test and specifies the restart.

E.3 Local entry, restart, and total curvature accounting

We now connect global convergence to a local initialization that satisfies the required radius and subspace conditions. The gradient test certifies the radius condition using known bounds. The subsequent restart must separately satisfy the prescribed subspace overlap; proximity to the minimizer alone does not provide that overlap.

Proposition E.2 (Local entry and restart)**.** *Under the conditions of Theorem 4.7 and Assumptions 4.1 and 4.2, choose the local parameters for fresh PSNG (Proposition 4.4) or lazy PSNG (Theorem 4.5) so that $q_{\mathrm{tar}} < 1 - 2c_A$ and $L_H M_{\max} c_s R \le 3(1 - 2c_A - q_{\mathrm{tar}})$. The first index T_{ent} satisfying $\|\mathbf{g}_t\|_2 \le \mu R/\sqrt{L}$ is finite and obeys* (4.12)*. At this index, $\mathbf{x}_{T_{\mathrm{ent}}} \in \mathcal{N}$ and $V_{T_{\mathrm{ent}}} \le R$. Restart Algorithm 1 there with the prescribed local budgets, a basis satisfying Assumption 4.3 at the entry point, and a reset accumulator. Then, the Armijo line search accepts $\alpha_t = 1$ at every subsequent iteration, and the corresponding local convergence and HVP bounds apply from this new initial iterate.*

Proof of Proposition E.2. Strong monotonicity of the gradient and $\nabla f(\mathbf{x}^\star) = 0$ give

$$\mu \|\mathbf{x} - \mathbf{x}^\star\|_2^2 \le \nabla f(\mathbf{x})^\top (\mathbf{x} - \mathbf{x}^\star) \le \|\nabla f(\mathbf{x})\|_2 \|\mathbf{x} - \mathbf{x}^\star\|_2 .$$

Thus, the gradient test implies $\|\mathbf{x}_t - \mathbf{x}^\star\|_2 \le R/\sqrt{L}$. The ball in (4.4) contains $\overline{\mathbb{B}}_2(\mathbf{x}^\star, R/\sqrt{L})$, so $\mathbf{x}_t \in \mathcal{N}$ and $V_t \le \sqrt{L}\, \|\mathbf{x}_t - \mathbf{x}^\star\|_2 \le R$. The global descent lemma applied at $\mathbf{x} - \nabla f(\mathbf{x})/L$ yields

$$f(\mathbf{x}^\star) \le f(\mathbf{x} - \nabla f(\mathbf{x})/L) \le f(\mathbf{x}) - \frac{\|\nabla f(\mathbf{x})\|_2^2}{2L}.$$

Consequently, $\|\mathbf{g}_t\|_2^2 \le 2L q_{\mathrm{G}}^t \Delta_0$. Solving for the first iteration at which this bound is at most $\mu^2 R^2/L$ proves (4.12). The maximum with one inside the logarithm covers entry at $t = 0$, including $\Delta_0 = 0$. The bound depends on $f(\mathbf{x}^\star)$ through Δ_0, whereas the entry test uses only the gradient and known parameters.

Restart at the entry iterate and reset the path accumulator. The stipulated overlap and budget give subspace error at most δ_{tol} at the restart. For fresh mode, Proposition 4.4 certifies each linearized contraction factor; for lazy mode, Theorem 4.5 does so. In either case, local invariance bounds the length of the proposed update $\mathbf{d}_t$ by $c_s R$ and places its segment in $\mathcal{N}$. The imposed margin and Proposition E.1 ensure that the Armijo line search accepts $\alpha_t = 1$. Induction makes the Armijo and local trajectories coincide after the restart and proves the stated rate and cost conclusions. Since $V_{T_{\mathrm{ent}}+k} \le q_{\mathrm{loc}}^k R$ for $k \ge 0$, adding T_{ent} to $\lceil \log(R/\varepsilon)/\log(1/q_{\mathrm{loc}}) \rceil$ gives the total iteration bound for $0 < \varepsilon < R$. □

Entry conditions and HVP accounting. Choosing $0 < c_A < (1 - \bar{\rho})/2$ permits $\bar{\rho} < q_{\mathrm{tar}} < 1 - 2c_A$; reducing R then ensures the additional condition for $\alpha_t = 1$. The parameter choices in Appendices C.4 and D.4 remain valid after this reduction. The entry test certifies proximity to the minimizer; the restart separately requires the prescribed subspace overlap.

For a restart into lazy PSNG, let N_{pre} count block HVPs before the local restart, including global initialization. For finite HVP work per iteration, finite entry gives $N_{\mathrm{pre}} < \infty$. Counting the local initialization once gives the total bound

$$N_{\mathrm{pre}} + K_{\mathrm{boot}} + 1 + (K_{\mathrm{ref}} + 1)(N_{\mathrm{ref}}^{\mathrm{local}} - 1).$$

Here, $N_{\mathrm{ref}}^{\mathrm{local}}$ counts all local refreshes, including initialization. By Theorem 4.5, $N_{\mathrm{ref}}^{\mathrm{local}} - 1 \le c_s R/[(1 - q_{\mathrm{loc}})\omega]$ because the local initial error is at most R. Thus, the total block-HVP count is at most

$$N_{\mathrm{pre}} + K_{\mathrm{boot}} + 1 + \frac{(K_{\mathrm{ref}} + 1) c_s R}{(1 - q_{\mathrm{loc}})\omega}.$$

The three terms account for the global phase, local initialization, and subsequent local refreshes, respectively. Fixed entry radius, local parameters, and initialization bounds, together with a global curvature update policy independent of ε, make this bound independent of ε.

A fixed curvature pair during the global phase. The global phase itself can use a simple policy with bounded curvature cost. Choose any orthonormal $\mathbf{P} \in \mathbb{R}^{d\times r}$ and form $\widetilde{\mathbf{B}} \coloneqq \mathbf{P}^\top \nabla^2 f(\mathbf{x}_0)\mathbf{P}$ using one block HVP. Theorem 4.7 applies to the stored pair, so $N_{\mathrm{pre}} \le 1$, allowing zero if the global phase is skipped. The local restart still requires its own overlap and budget conditions.

F Experimental Details and Additional Results

This appendix groups the reproducibility details by their role in the experiments. We first specify the objectives and data, then the algorithms and cost measurements. Additional convergence and spectral diagnostics follow, then comparisons of refresh policies under different curvature changes and a study of subspace dimension on fixed data.

F.1 Objectives and data generation

The quadratic problems isolate subspace accuracy and spectral targeting. The logistic problems retain a small leading eigenspace while allowing the Hessian to change during optimization.

Two-dimensional accuracy example. The objective is $f(\mathbf{x}) \coloneqq \frac{1}{2}\mathbf{x}^\top \mathbf{H}\mathbf{x}$ with $\mathbf{H} = \mathrm{diag}(4, 1)$. We prescribe $\mathbf{P}_\delta \coloneqq (\sqrt{1-\delta^2}, \delta)^\top$ so that the projector error is exactly δ. This example separates the effect of subspace accuracy from the procedure used to compute the subspace.

Quadratic for the power-iteration budget. The budget sweep uses $d = 100$ and $r = 5$ with a randomly rotated Hessian (rotation seed 2026). The first five eigenvalues are logarithmically spaced from 20 to 10, and the remaining 95 from 5 to 1. Thus, the spectral ratio governing block-power convergence is $\theta = 1/2$.

Quadratic for subspace selection. We use $\mathbf{H} \coloneqq \mathbf{Q}\,\mathrm{diag}(\lambda_1, \ldots, \lambda_d)\mathbf{Q}^\top$ with a dense random orthogonal $\mathbf{Q}$, $d = 2000$, and $r = 50$. The top eigenvalues equal 1000 and the remaining eigenvalues are logarithmically spaced from 10 to 1. The rotation uses seed 42. This gives $\kappa = 1000$ and $\kappa_\perp = 10$, allowing the effect of subspace selection to be compared at a fixed spectrum.

Logistic regression. We minimize the regularized logistic objective

$$f(\mathbf{w}) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \log\bigl(1 + \exp(-y_i \mathbf{a}_i^\top \mathbf{w})\bigr) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 \tag{F.1}$$

with $n = 4000$, $d = 120$, and $\lambda = 10^{-3}$. The following construction changes only the leading feature scales across problem instances with a common seed. For each seed, generate a random orthogonal matrix $\mathbf{U}$ by QR factorization of a Gaussian matrix, independent $\mathbf{z}_i \sim \mathcal{N}(0, \mathbf{I})$, and an independent Gaussian vector normalized to $\|\mathbf{b}\|_2 = 1$. Define

$$s_{1:6} \coloneqq a\ \mathrm{geomspace}(80, 20, 6), \qquad s_{7:120} \coloneqq \mathrm{geomspace}(0.1, 0.02, 114),$$
$$\mathbf{a}_i \coloneqq \mathbf{U}\,\mathrm{diag}(\sqrt{s_1}, \ldots, \sqrt{s_d})\mathbf{z}_i,$$

where geomspace denotes logarithmically spaced values. Sample $y_i \in \{-1, 1\}$ with $\Pr(y_i = 1 \mid \mathbf{z}_i) = \sigma(\mathbf{z}_i^\top \mathbf{b})$, where $\sigma(u) \coloneqq (1 + e^{-u})^{-1}$. For a fixed seed, $\mathbf{U}$, the latent samples, $\mathbf{b}$, and labels are identical across $a \in \{0.1, 0.3, 1\}$. Only the six leading feature scales change; we apply no row normalization or column standardization. The fixed regularizer means the three optima need not coincide.

Writing $\mathbf{A}$ for the feature matrix, the objective satisfies

$$\lambda\mathbf{I} \preceq \nabla^2 f(\mathbf{w}) \preceq \mathbf{H}_0 = \lambda\mathbf{I} + \frac{\mathbf{A}^\top\mathbf{A}}{4n}, \qquad L_H \le \frac{1}{6\sqrt{3}\,n}\sum_{i=1}^{n}\|\mathbf{a}_i\|_2^3.$$

F.2 ALGORITHMS, HYPERPARAMETERS, AND COST ACCOUNTING

We give the computation and fitting procedures for the quadratic experiments first, then specify the common protocol used by the logistic benchmarks. Offline spectral calculations are distinguished from the computations charged to each optimization run.

Accuracy measurement on the two-dimensional example. For the prescribed subspaces above, compute $q(\delta) = \left\|\mathbf{I} - \mathbf{M}_\delta^{1/2}\mathbf{H}\mathbf{M}_\delta^{1/2}\right\|_2$ directly. The choices $\eta = 1/2$ and $\eta = 1$ give exact-subspace factors $1/2$ and zero, respectively. The expansions in Appendix B.5 give

$$q(\delta) - \tfrac{1}{2} = \tfrac{3}{4}\delta^2 + O(\delta^4) \quad (\eta = \tfrac{1}{2}), \qquad q(\delta) = \tfrac{3}{2}\delta + O(\delta^2) \quad (\eta = 1).$$

We use 30 logarithmically spaced values of δ from 10^{-4} to $10^{-1/2}$ and fit over the 23 values with $\delta \le 0.05$. Figure 3 reports least-squares fits of the log excess factor (for $\eta = 1/2$) or log factor (for $\eta = 1$) against $\log \delta$.

Power-iteration budgets. Initial bases are reduced QR factors of Gaussian matrices with seeds $0, \dots, 9$. Each basis is shared across $K = 0, \dots, 20$. Use $\eta = 1/3$, so $\rho_\perp = 2/3$. For each K, compute the block-power basis and projected Hessian, then evaluate δ_K and $q_K = \left\|\mathbf{I} - \mathbf{H}^{1/2}\mathbf{M}_K\mathbf{H}^{1/2}\right\|_2$. The construction costs $5(K+1)$ scalar HVPs; dense spectral diagnostics and QR are excluded from this curvature count.

Subspace-selection comparison. All methods start from the same Gaussian vector generated with seed 43 and run for 200 iterations. Exact PSNG is supplied with the leading eigenspace and uses $\eta = 2/11$, giving energy-norm factor $9/11$ and objective-gap factor $(9/11)^2$. GD uses $2/(\lambda_1 + \lambda_d)$, giving objective-gap factor $(999/1001)^2$; NAG uses $1/\lambda_1$ and momentum $(\sqrt{\kappa} - 1)/(\sqrt{\kappa} + 1)$ (Nesterov, 2004). The random-subspace PSNG variant is an ablation constructed for this comparison: it replaces the principal basis by a randomly chosen orthonormal basis of the same rank, keeps that basis fixed, and applies the same hybrid update (2.2). The gradient stepsize in the orthogonal complement is independently grid-tuned for each sampled basis using 120 pilot iterations: candidates with a relative objective increase exceeding 10^{-9} are discarded, and the stable candidate with lowest pilot final objective is selected. The candidate stepsizes are $10^{-5}\{1, 2, 5, 10, 20, 50, 80, 100, 120, 150, 200, 300, 500\}$. Results average subspace seeds $0, 1, 2$. For numerical stability, we use $\mathbf{B} + 10^{-12}\mathbf{I}$ as the coefficient matrix when solving the projected linear systems. Figure 1 (left) isolates the effect of subspace selection; it does not include the cost of obtaining the exact eigenspace or tuning the random-subspace stepsize. Objective-gap values below 10^{-14} are clipped in the plot.

Logistic-regression algorithms and parameter selection. The main comparisons fix $r = 6$ before running the algorithms; Appendix F.5 varies it on fixed data. For PSNG, a matrix-free Lanczos eigensolve at $\mathbf{w}_0 = 0$ estimates the largest seven eigenvalues. With estimated eigenvalues denoted by $\widehat{\lambda}_i$, use

$$\eta = \frac{2}{\lambda + 1.000001\widehat{\lambda}_7(\mathbf{H}_0)}.$$

This fixed rule uses a lower bound on the complementary minimum curvature; it is not the quadratic-optimal stepsize based on both exact endpoints. All scales use $K_{\mathrm{boot}} = 5$ and $K_{\mathrm{ref}} = 2$. The main lazy threshold is $\omega = 1$. An additional block HVP forms each projected Hessian. Each mode uses the same random initial basis.

For GD and NAG, estimate only the largest initial eigenvalue and use $L_0 \coloneqq 1.000001\widehat{\lambda}_1(\mathbf{H}_0)$. GD uses stepsize $1/L_0$ with Armijo checking; NAG uses $1/L_0$, strong-convexity bound λ, and momentum $(\sqrt{L_0} - \sqrt{\lambda})/(\sqrt{L_0} + \sqrt{\lambda})$. Lanczos uses tolerance 10^{-10} and a reproducible initial vector. Offline full spectra confirm that the fixed numerical safety factor covers the eigenvalues used in both rules. The costs of spectral computations and residual checks are included for each optimization run.

FOSI-GD applies FOSI to gradient descent (Sivan et al., 2024). It uses a float64 NumPy implementation of the momentum-free update and fully reorthogonalized Lanczos in the official FOSI code[1].

[1] `https://github.com/hsivan/fosi`, commit `cf040ce`.

It selects the largest $r = 6$ eigenpairs, with 24 scalar HVPs per refresh, before the first update and every 10 updates thereafter. We use Newton scale 1, no learning-rate clipping, and no line search. To match the initial complementary step to PSNG, set the fixed base GD rate to $\eta\widehat{\lambda}_6(\mathbf{H}_0)/\widehat{\lambda}_1(\mathbf{H}_0)$. FOSI multiplies this rate by its current estimate of λ_1/λ_6 at each refresh. This is a calibrated configuration for the present strongly convex problem, rather than the official DNN defaults. Its initial spectral calibration is charged in full, as for PSNG. The port's update agrees with the pinned official update to floating-point precision when supplied with the same eigenpairs.

L-BFGS (Liu & Nocedal, 1989) uses SciPy's unconstrained L-BFGS-B implementation (Virtanen et al., 2020) with memory 10, a maximum of 50 line-search trials, and objective-change and infinity-norm gradient tolerances set to zero. A callback checks the same relative ℓ_2 gradient as the other methods, reusing the function/gradient cache. Newton–CG uses unpreconditioned, matrix-free CG from zero, with relative residual tolerance 10^{-2} and at most d HVPs per direction. Dense Newton forms the analytic Hessian and solves the Newton system by Cholesky factorization, without forming an inverse. Its online Hessian evaluations are counted separately from HVPs. PSNG and both Newton solvers use Armijo with initial step 1, $c_A = 10^{-4}$, and shrinkage $1/2$. Our Armijo checks allow slack $32\epsilon_{\mathrm{mach}} \max\{1, |f|\}$ and at most 50 backtracks.

Logistic-regression stopping, repetitions, and costs. The eight-method comparison uses the fixed settings above; the separate parameter search is reported in Appendix F.4. Each method starts at zero and targets relative gradient 10^{-8}, with iteration caps 100,000 for GD and 3000 otherwise. Costs are recorded at the first iterate reaching each of $10^{-2}, 10^{-4}, 10^{-6}, 10^{-8}$. In the three-scale benchmark, all 216 runs reach 10^{-6}. All 27 L-BFGS runs stop under their objective-change test at relative gradients between 2.30×10^{-8} and 1.31×10^{-7}; these runs are marked as not reaching 10^{-8}. The remaining 189 runs reach 10^{-8}.

This benchmark uses data seeds $0, 1, 2$, each with three timing repeats and eight methods. The random seed for PSNG initialization and FOSI Lanczos vectors is 1000 plus the data seed, fixed across scales and repeats; fresh and lazy PSNG share their initial basis. Runs are serial, with method order rotated, on an Intel Core i5-14400 using one numerical thread, Python 3.12.14, NumPy 2.5.3, and SciPy 1.18.1. Summaries first take the median over timing repeats, then the median and interquartile range over data seeds. Repeats of one dataset are not treated as independent data instances.

All methods share the same dense objective implementation. Optimization timing includes random-basis generation, QR, solving projected linear systems, Lanczos refreshes, dense Hessian construction and Cholesky factorization where applicable, online gradient tests, line searches, and in-memory logging. Spectral parameter-selection time is measured separately and included in full for each independent optimization run in the main comparison. The reported times exclude data generation, shared data preparation, file writing, computation of reference solutions, and offline diagnostics. Checkpoints are spaced by 50 iterations, with the terminal point also saved. At $a = 1$, median parameter-selection costs are 63 scalar HVPs for PSNG/FOSI-GD and 22 for GD/NAG; L-BFGS and both Newton solvers require none.

We count gradient and loss evaluations separately from HVPs; NAG's extrapolated-point and stopping-check gradients are both included. A rank-r block product counts as r scalar HVPs, giving PSNG online cost $r[6 + 3(N_{\mathrm{ref}} - 1)]$, where N_{ref} includes initialization. Costs at an iterate exclude refreshes for its next step. FOSI uses 24 scalar HVPs per refresh. Dense Newton uses one complete Hessian per update; its curvature cost is reported separately. Figure 6 gives the iteration curves and total HVP counts across tolerances. Figure 7 gives the timing curves; Figure 8 separates curvature costs.

F.3 Additional convergence and spectral diagnostics

The following results complement the main-text comparisons. The quadratic diagnostics examine the dependence on subspace error and compare observed and sufficient power budgets; the logistic diagnostics examine the actual spectra, convergence times, and linearized contraction factors.

Subspace accuracy. For the two-dimensional problem in Appendix F.1, Figure 3 plots the excess contraction factor $q(\delta) - 1/2$ at $\eta = 1/2$ and the contraction factor $q(\delta)$ at $\eta = 1$. The fitted log–log slopes 1.999 and 1.005 match the second-order bound in Section 3 and the first-order behavior

derived in Appendix B.5, respectively. The contrast illustrates why the positive- and zero-factor cases require separate accuracy bounds. Appendix F.2 gives the fitting protocol.

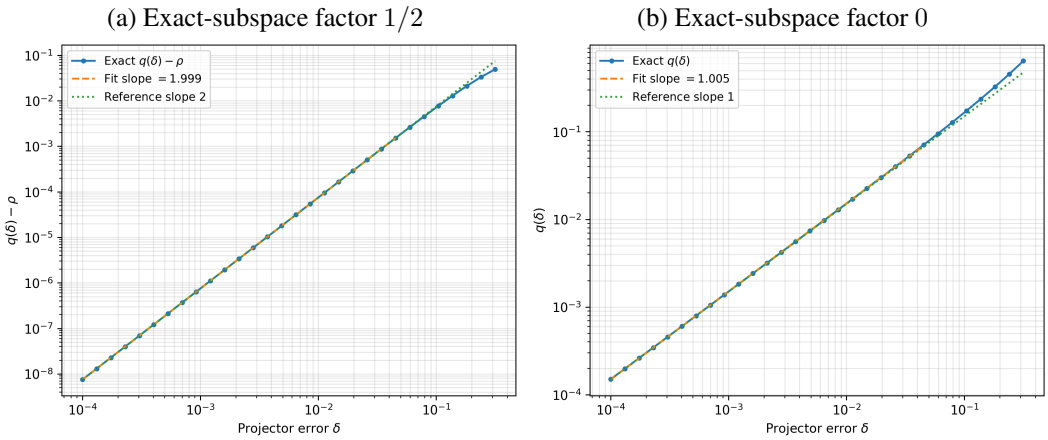

Figure 3: Contraction factor versus subspace error. For $\mathbf{H} = \mathrm{diag}(4, 1)$, $q(\delta)$ is the exact energy-norm factor and ρ is the exact-subspace factor $\rho_\perp$. Left: excess factor at $\eta = 1/2$; right: factor at $\eta = 1$. Fits over $\delta \leq 0.05$ give slopes close to two and one.

Observed and sufficient power budgets. All 210 seed–budget combinations are retained, including those with $q_K \geq 1$. Contraction first occurs at $K = 3$–7 (20–40 scalar HVPs) across ten initial bases, whereas the sufficient upper bound in (B.13) first falls below one at $K = 9$–14. The measured errors and factors satisfy the subspace and contraction bounds within numerical precision. The θ^{2K} curves in Figure 4 show the reference decay rate.

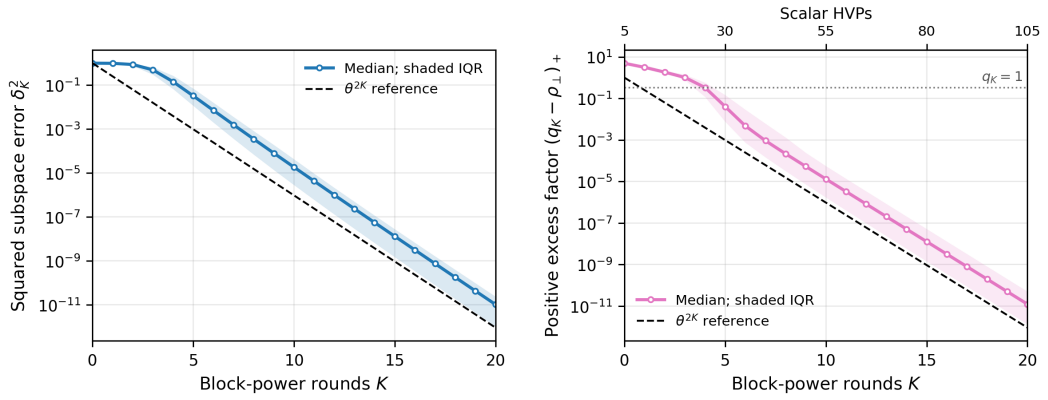

Figure 4: Block-power budget and quadratic contraction. Squared subspace error (left) and positive excess factor $(q_K - \rho_\perp)_+ := \max\{q_K - \rho_\perp, 0\}$ (right), with medians and interquartile ranges over ten initial bases. Here, $\rho_\perp = 2/3$ and $\theta = 1/2$; dashed curves give θ^{2K} as a reference. The horizontal dotted line marks $q_K = 1$. The upper axis includes the final HVP block for the projected Hessian.

Logistic-regression reference spectra. Reference solutions computed offline using Newton's method with dense Hessians have gradient norms below 10^{-12}. Table 1 reports medians of the actual reference Hessian spectra, using its smallest eigenvalue rather than λ to compute κ and $\kappa_\perp$. Reference solutions, their spectra, and generating eigenvectors are not used for online initialization or parameter selection.

Figure 5 shows how increasing the leading feature scale raises the full condition number while leaving the complementary condition number nearly unchanged. The eigengap also widens across these scales (Table 1); the main-text iteration comparison uses the same instances.

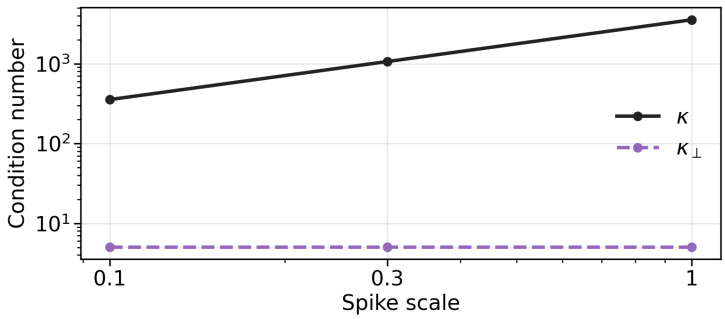

Figure 5: Full and complementary spectra. Reference-Hessian condition numbers versus the leading feature scale a, with medians and interquartile ranges over three data seeds.

Table 1: Median reference-Hessian diagnostics over three data seeds. Increasing a changes both the full condition number and the eigengap.

a	κ	$\kappa_\perp$	$\lambda_6 - \lambda_7$	λ_7/λ_6
0.1	354.6	5.025	0.402	0.05642
0.3	1063.3	5.025	1.251	0.01885
1	3543.5	5.025	4.222	0.00566

Convergence and computational costs. Table 2 compares all eight methods at relative gradient 10^{-6}; Figure 6 shows their convergence and the HVP costs across tolerances. The three-seed benchmark uses the names and abbreviations introduced in Section 5.

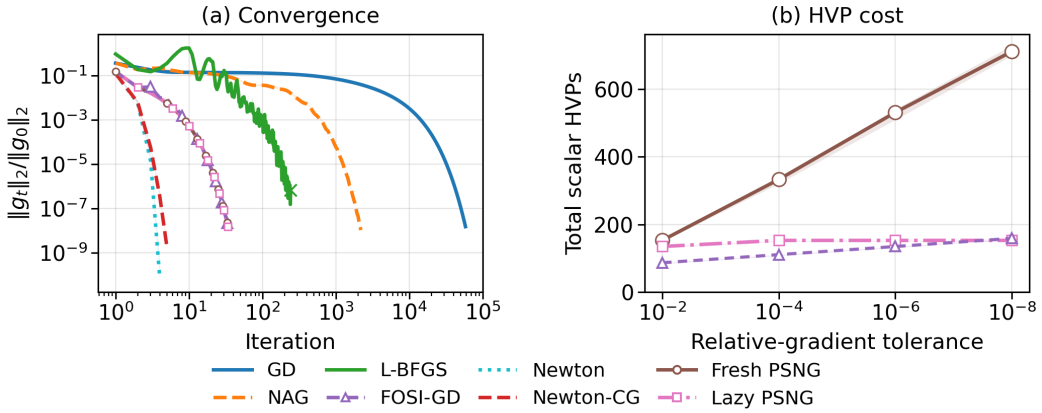

Figure 6: Convergence and curvature reuse at $a = 1$. Left: relative gradient versus iteration, using only iterations recorded for every run of each method; $\times$ marks a method with runs not reaching 10^{-8}. Right: total scalar HVPs for fresh, lazy, and FOSI-GD to each tolerance, including spectral parameter selection. Curves and shading show medians and interquartile ranges over three data seeds.

Figures 7 and 8 complement the iteration and total-HVP comparison in Figure 6. The timing curves report the time to first reach each relative-gradient target, including spectral parameter selection; the cost breakdown separates that contribution from online curvature computation. At $a = 1$ and

Table 2: Eight-method comparison: median costs to reach a relative-gradient tolerance of 10^{-6} at $a = 1$. Total scalar HVPs and time include spectral parameter selection; full-Hessian evaluations are counted separately.

Method	Iterations	Total HVPs	Full-Hessian evals.	Total time (s)
GD	41,338	22	0	11.0995
NAG	1,573	22	0	0.5863
L-BFGS	207	0	0	0.0605
Newton–CG	4	41	0	0.0083
Newton	4	0	4	0.0150
FOSI-GD	25	135	0	0.0303
Fresh PSNG	25	531	0	0.0795
Lazy PSNG	25	153	0	0.0330

relative gradient 10^{-6}, the median online scalar HVP counts for fresh PSNG, lazy PSNG, and FOSI-GD are 468, 90, and 72, respectively, with median parameter-selection cost 63 each. Newton–CG uses a median of 41 online HVPs; GD, NAG, L-BFGS, and dense Newton use none. GD and NAG each have median parameter-selection cost 22 HVPs, while dense Newton forms a median of four full Hessians.

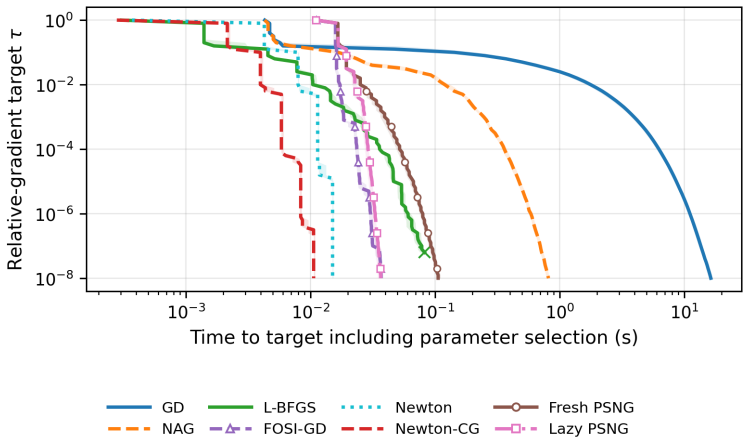

Figure 7: Time to target accuracy at $a = 1$. Time to first reach each relative-gradient target, including spectral parameter selection. We use the first qualifying recorded iterate without interpolation, taking timing medians within each seed before reporting medians and horizontal interquartile bands across three seeds. Targets are shown only when reached by every run; an endpoint $\times$ marks a method with runs not reaching 10^{-8}.

Dense and Hessian-free Newton. Table 3 compares the two Newton implementations under the same stopping test and Armijo rule. Both reach 10^{-6} in a median of four updates. Newton–CG is faster here because the Newton systems require only about ten HVPs per update, while dense Newton constructs and factors a full Hessian. At $d = 120$, both are effective baselines for this problem.

Offline diagnostics and scope. We replay fresh and lazy trajectories, verify their agreement with timed runs, and inspect the initial update and every fifth nonterminal update. For the stored $\mathbf{M}_t$ and current $\mathbf{H}_t$, compute the condition number of $\mathbf{H}_t^{1/2}\mathbf{M}_t\mathbf{H}_t^{1/2}$ and $q_t := \|\mathbf{I} - \mathbf{M}_t\mathbf{H}_t\|_{\mathbf{H}_t}$. Across the sampled points, these lie in $[4.85, 5.29]$ and $[0.659, 0.932]$, respectively (Figure 9). These describe the linearized update with $\alpha_t = 1$, not the observed nonlinear error ratio or a bound at unsampled iterates. In this three-scale comparison, every lazy run uses 90 online HVPs at each tolerance from 10^{-4} to 10^{-8}. All PSNG steps in this three-scale comparison have $\alpha_t = 1$. No certified local radius

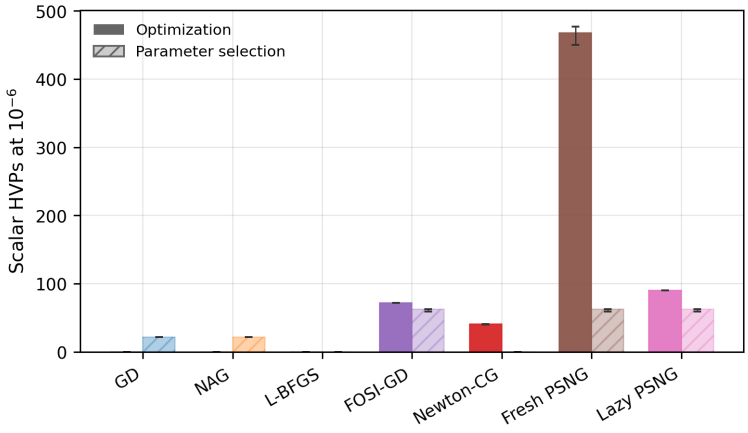

Figure 8: Curvature costs at $a = 1$. Separate online and parameter-selection scalar HVP counts to relative gradient 10^{-6}, with medians and interquartile ranges over three data seeds. Dense Newton uses full-Hessian evaluations, reported in Tables 2 and 3.

Table 3: Second-order reference costs to relative gradient 10^{-6}, using the same three-seed and timing-repeat aggregation as Table 2. Full Hessians are counted separately from scalar HVPs; neither method requires spectral parameter selection.

a	Method	Iterations	HVPs	Full Hessians	Time (s)
0.1	Newton–CG	4	42	0	0.0093
0.1	Newton	4	0	4	0.0153
0.3	Newton–CG	4	41	0	0.0081
0.3	Newton	4	0	4	0.0150
1	Newton–CG	4	41	0	0.0083
1	Newton	4	0	4	0.0150

or explicit restart test is used; the practical power budgets and path threshold are fixed across the scales.

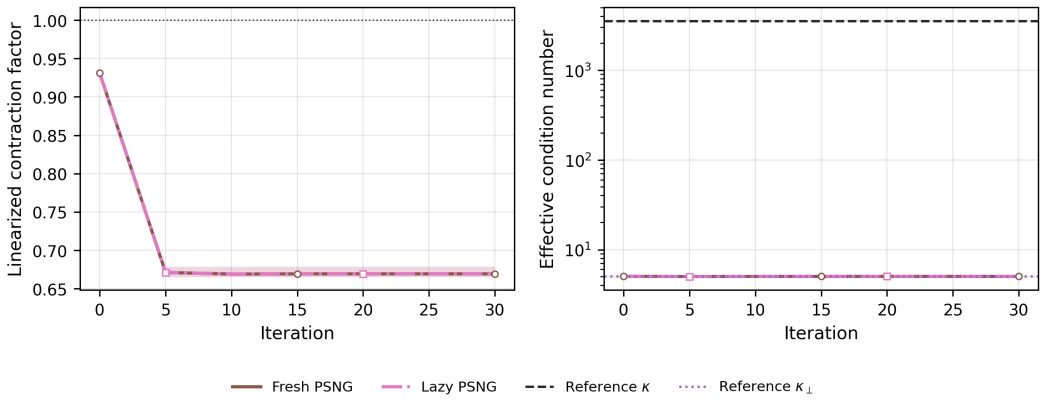

Figure 9: Linearized contraction and preconditioned curvature at $a = 1$. Left: the sampled linearized contraction factors q_t for fresh and lazy PSNG. Right: the actual preconditioned Hessian condition numbers, compared with the reference Hessian's full and complementary condition numbers. All computations in this figure are offline and excluded from costs.

F.4 Refresh policies and changing curvature

Reuse on the original data. To examine the effect of refresh frequency, we fix $a = 1$ and use data seeds $0, \ldots, 4$, with one run per policy and seed. We compare fresh, lazy with $\omega \in \{0.01, 0.1, 1\}$, and a baseline that keeps both the initial subspace and projected Hessian fixed. All policies share the data, zero initial point, basis seed, complementary stepsize, and power budgets specified above; gradients and Armijo checks continue normally for the fixed pair. We record the first iterate reaching each of $10^{-4}, 10^{-6}, 10^{-8}$, with the same 3000-iteration cap. The full cost of spectral parameter selection is included for each optimization run. This comparison reports iteration and HVP counts, without a new timing comparison.

Table 4: Refresh-policy comparison at $a = 1$. Costs to relative gradient 10^{-8}, reported as medians [first, third quartiles] over five data seeds. All 25 displayed runs reach the target. Refreshes include initialization; total HVPs include parameter selection, with totals formed per run before aggregation.

Policy	Iterations	Refreshes	Online HVPs	Total HVPs
Fresh	35 [34, 37]	35 [34, 37]	648 [630, 684]	711 [693, 747]
Lazy, $\omega = 0.01$	35 [34, 37]	13 [12, 13]	252 [234, 252]	315 [297, 315]
Lazy, $\omega = 0.1$	35 [34, 37]	7 [7, 8]	144 [144, 162]	207 [207, 225]
Lazy, $\omega = 1$	35 [34, 37]	4 [4, 4]	90 [90, 90]	153 [146, 153]
Frozen after initialization	36 [35, 37]	1 [1, 1]	36 [36, 36]	99 [92, 99]

More frequent refreshes do not reduce the median iteration count among the updating policies (Table 4). The fixed pair takes only zero or one additional iteration relative to lazy PSNG with $\omega = 1$ on each seed, while requiring just 36 online HVPs. For $\omega = 1$, total HVPs remain unchanged from 10^{-4} to 10^{-8} on every seed; the median is 153.

We additionally evaluate δ_t and q_t at every nonterminal iterate of the lazy ($\omega = 1$) and fixed-pair runs, using the current Hessian and the stored pair after any refresh. These diagnostics are offline and excluded from costs. Figure 10 shows the preselected seed 0: refreshing reduces subspace error, while the linearized factors remain close. Across all five seeds, both policies have $q_t < 0.934$ at every recorded update. These observations explain why a fixed pair works well here.

Here, the initial curvature pair remains effective throughout the trajectories. We next change the label-generating signal to examine a regime in which updating that pair improves convergence.

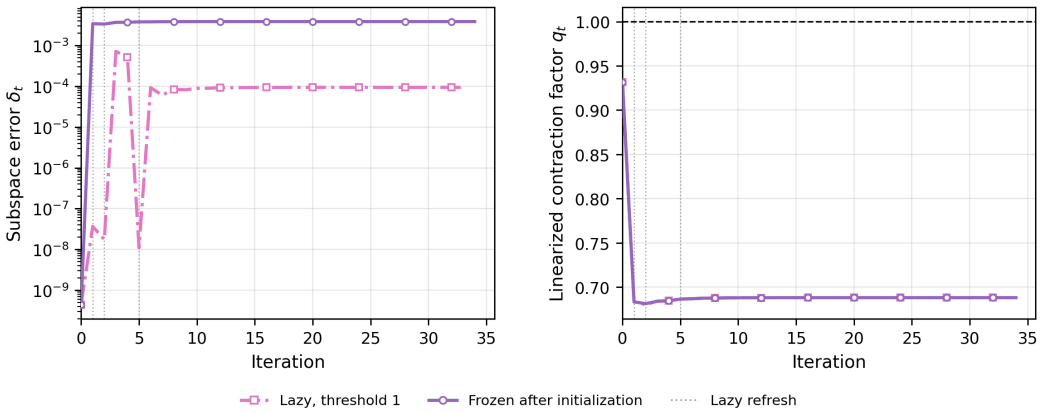

Figure 10: Lazy updates versus a fixed curvature pair. Subspace error (left) and linearized contraction factor (right) for data seed 0 at $a = 1$. Dotted vertical lines mark the lazy policy's refreshes at $\omega = 1$; the dashed horizontal line marks $q_t = 1$. Both curves use each policy's current Hessian and the pair used for its next step, and stop at its last nonterminal iterate. The factor curves nearly coincide despite different subspace errors.

Controlled changes in curvature. We retain the feature construction in Appendix F.1 at $a = 1$, including $n = 4000$, $d = 120$, and $\lambda = 10^{-3}$. Starting from the same unit vector $\mathbf{b}$, form

$$\mathbf{b}_{\text{head}} \coloneqq \begin{pmatrix} \sqrt{0.8}\,\mathbf{b}_{1:6} / \|\mathbf{b}_{1:6}\|_2 \\ \sqrt{0.2}\,\mathbf{b}_{7:120} / \|\mathbf{b}_{7:120}\|_2 \end{pmatrix}.$$

Generate labels with $\Pr(y_i = 1 \mid \mathbf{z}_i) = \sigma(\beta \mathbf{z}_i^\top \mathbf{v})$ for $\mathbf{v} \in \{\mathbf{b}, \mathbf{b}_{\text{head}}\}$ and $\beta \in \{1, 3, 5\}$. Within each seed, all settings share the features and the uniform random draws used to sample the labels. The 80% allocation concerns signal energy in latent coordinates. The initial Hessian at $\mathbf{w}_0 = 0$ is identical across these settings, so all policies share the initial curvature pair and the spectral stepsize rule.

An exploratory pilot used seeds $0, 1$. We then fixed the entire design and validated it on ten new seeds, 100–109, without further parameter selection. Each of the six data settings compares fresh PSNG, lazy PSNG with $\omega \in \{0.01, 0.1, 1\}$, and the fixed pair, giving 300 validation runs. All use $r = 6$, $K_{\text{boot}} = 5$, $K_{\text{ref}} = 2$, the preceding Armijo settings, and a 3000-iteration cap. Every run reaches 10^{-8} with $\alpha_t = 1$ throughout. This comparison measures iteration and HVP counts, with initial spectral calibration charged to each optimization run; it does not introduce a new timing benchmark.

Validation results. Figure 11 reports all six settings. At 10^{-6}, lazy PSNG with $\omega = 1$ matches fresh's iteration count on 55 of the 60 data instances and takes one extra iteration on the other five. Table 5 summarizes the prespecified stronger-signal setting $\mathbf{v} = \mathbf{b}_{\text{head}}$, $\beta = 5$. On these ten seeds, the median paired ratio of fixed-pair to lazy iteration counts is 3.69. Lazy's HVP saving relative to fresh is computed per seed before taking the median, giving 78.8%. Figure 2 shows the corresponding fresh, lazy ($\omega = 1$), and fixed-pair curves.

For lazy PSNG with $\omega = 1$, only four or five refreshes occur after initialization. Eight seeds require no new HVPs between targets 10^{-4} and 10^{-8}. The other two each add one refresh (18 scalar HVPs) before reaching 10^{-6}; thereafter their HVP counts also remain constant through 10^{-8}. More frequent refreshing gives little iteration benefit while increasing curvature cost.

Curvature diagnostics. We save every iterate and curvature pair, then compute q_t and subspace errors offline. Across all validation trajectories, the recorded eigengap is at least 0.6978 and $\lambda_7/\lambda_6 \leq 0.01535$. These observations concern the recorded iterates; as in the main benchmark, we do not certify the full local-radius and threshold conditions.

At each of the ten reference solutions for the stronger-signal setting, we evaluate the initial fixed pair and the final lazy pair ($\omega = 1$) against the same Hessian. Their median factors are 0.93137

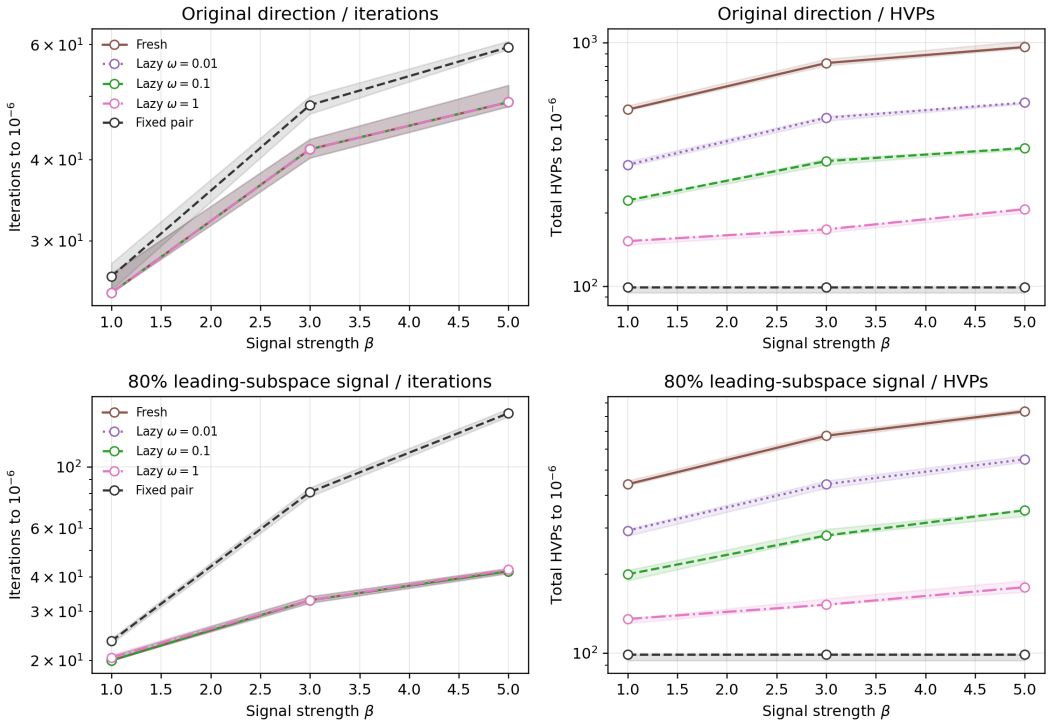

Figure 11: Refresh policies across signal settings. Iterations (left) and total scalar HVPs (right) to relative gradient 10^{-6}, for the original signal direction (top) and the direction with 80% leading-subspace signal energy (bottom). Medians and interquartile ranges are over ten validation seeds. All five policies use identical initial curvature pairs within each seed.

Table 5: Costs under stronger curvature changes. Relative-gradient target 10^{-6}, $\mathbf{v} = \mathbf{b}_{\text{head}}$, $\beta = 5$. Entries are medians [first, third quartiles] over ten validation seeds. Refresh counts include initialization; total scalar HVPs include spectral calibration.

Policy	Iterations	Refreshes	Total HVPs
Fresh	$42\,[41.25, 43]$	$42\,[41.25, 43]$	$837\,[823.5, 855]$
Lazy, $\omega = 0.01$	$42\,[41.25, 43]$	$26\,[25.25, 27]$	$549\,[535.5, 567]$
Lazy, $\omega = 0.1$	$42\,[41.25, 43]$	$15\,[14, 15]$	$351\,[333, 351]$
Lazy, $\omega = 1$	$42.5\,[41.25, 43]$	$6\,[5, 6]$	$178.5\,[171, 189]$
Fixed pair	$156\,[151.5, 161.75]$	$1\,[1, 1]$	$99\,[93.75, 99]$

and 0.82739, respectively. Keeping the initial basis and replacing only its projected Hessian by the current one gives median factor 0.82823, close to the exact-subspace value 0.82738. This identifies projected-Hessian reuse error as the main source of the fixed pair's excess factor in this setting.

In this stronger-signal setting, we also evaluate the old and new pairs at the same iterate at actual lazy refresh points. For $\omega = 1$, the decreases in q_t range from 1.09×10^{-6} to 1.73×10^{-4}. These small immediate reductions accompany sustained contraction factors close to fresh PSNG. Figure 12 shows the contraction factors and the accumulated step length used to trigger refreshes on the preselected validation seed 100.

The initial drop and subsequent rise reflect changes in the complementary spectrum under the fixed stepsize $\eta \approx 66.2382$. For the exact leading eigenspace, the linearized factor is $\max\{|1 - \eta\lambda_7(\mathbf{H}_t)|, |1 - \eta\lambda_d(\mathbf{H}_t)|\}$. Along the fresh trajectory, these two terms change from $(0.93376, 0.63620)$ at $t = 0$ to $(0.38486, 0.72791)$ at $t = 1$: the larger complementary eigenvalue initially determines the factor, but the smallest eigenvalue takes over after the first step. As the latter decreases further, $1 - \eta\lambda_d(\mathbf{H}_t)$ rises toward 0.83361. The computed fresh factors agree with this exact-subspace expression to five decimal places at these two iterates and at $t = 60$.

At zero initialization, the predicted probabilities are $p_i = 1/2$, so the logistic Hessian weights $p_i(1 - p_i)$ equal their maximum $1/4$; moving away from zero lowers these weights. With η fixed, decreasing an eigenvalue with $\eta\lambda_i > 1$ initially improves its factor $|1 - \eta\lambda_i|$, whereas decreasing one with $\eta\lambda_i < 1$ weakens contraction. This explains the observed switch between spectral endpoints. Lazy PSNG closely follows fresh, so their curves nearly overlap; the first-step drop also occurs for the fixed pair, reflecting the Hessian change. Its larger later factor is consistent with the projected-Hessian mismatch identified above. All three factors remain below one along the recorded trajectories.

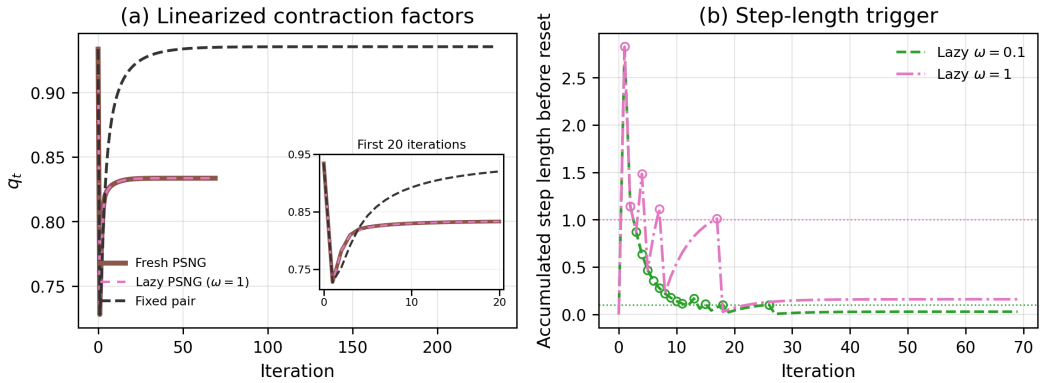

Figure 12: Refresh diagnostics on validation seed 100. The stronger-signal setting uses $\beta = 5$ and 80% leading-subspace signal energy. Left: factors for the pairs used in each update; the inset enlarges iterations 0–20. Fresh (thick solid) and lazy (thin dashed) nearly overlap. Right: accumulated step length since the last refresh, $\ell_{s,t}$, shown before checking the threshold and resetting the accumulator. Circles mark refreshes; dotted lines show the thresholds. An update may take $\ell_{s,t}$ above the threshold; the pair is refreshed before the next update. Curvature diagnostics are computed offline.

Refresh-parameter search. Separately, using the original data model in Appendix F.1, we search $\omega \in \{0.01, 0.03, 0.1, 0.3, 1, 3, 10\}$ for lazy PSNG and periods $p \in \{1, 2, 5, 10, 20, 50, 100\}$ for FOSI-GD, keeping other settings fixed. Both searches use scales $a \in \{0.1, 0.3, 1\}$, data seeds $10, 11, 12$, and five timing repeats per candidate and instance. We minimize the geometric mean time to relative gradient 10^{-6} after taking within-instance timing medians. All runs must reach the target; candidates within 5% of the fastest score are ranked by HVP count, iterations, then time. Figure 13 shows all candidates. The selected settings, $\omega = 10$ and $p = 100$, are then fixed across scales and tolerances on independent seeds 20–24, again with five repeats. Periods 50 and 100 have identical tuning trajectories.

Across the fifteen evaluation instances, these settings give geometric-mean speedups of 1.63 and 1.35 over $\omega = 0.1$ and $p = 10$, respectively. Both use only their initial curvature pair through tolerance 10^{-8}. At $a = 1$ and tolerance 10^{-6}, their median iteration counts are both 27, with 99 and 87 total HVPs and times 0.0256 and 0.0229 seconds, respectively. These are separate held-out results; the eight-method comparison uses $\omega = 1$ and $p = 10$ to display convergence with occasional refreshes. The complete search and evaluation comprise 630 and 300 runs, respectively, all reaching 10^{-8}.

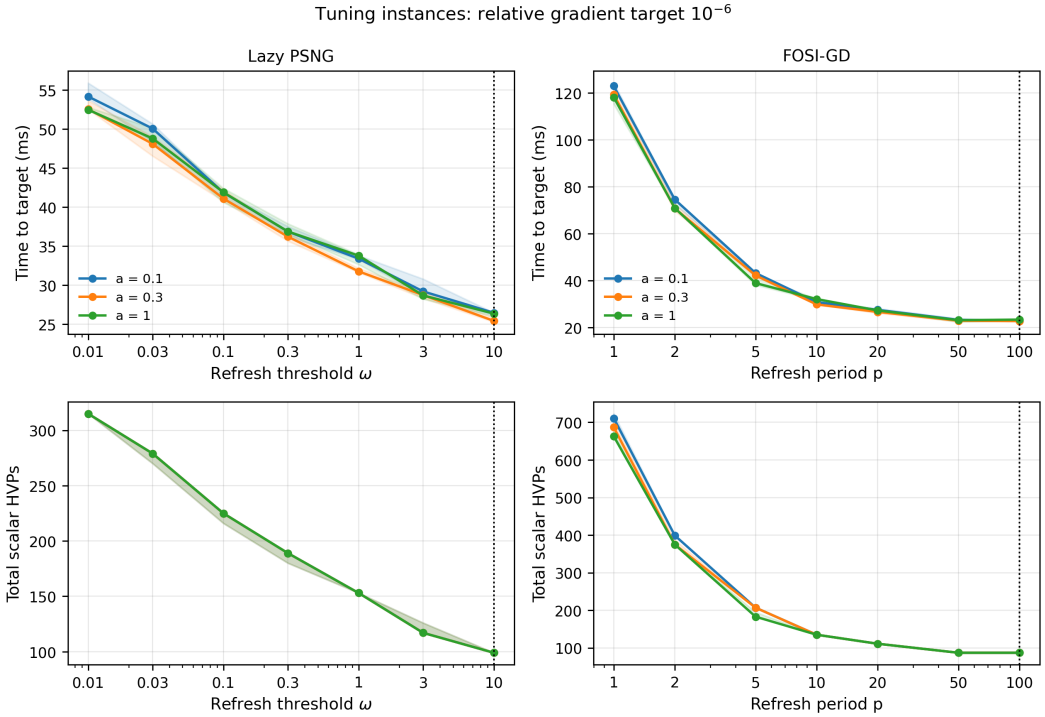

Figure 13: Refresh-parameter sensitivity. Time (top) and total scalar HVPs (bottom) to relative gradient 10^{-6} for lazy PSNG (left) and FOSI-GD (right). Curves show medians and shaded interquartile ranges over three tuning seeds after aggregating five timing repeats. Vertical dotted lines mark the selected parameters.

F.5 SUBSPACE DIMENSION ON FIXED DATA

The main comparison matches $r = 6$ to the six leading directions in the data covariance. To examine sensitivity to this choice, we keep the $a = 1$ datasets fixed and vary only the algorithm rank over $r \in \{2, 4, 6, 8, 12\}$. All ranks use the same samples, labels, regularization, and zero starting point for each seed. Their initial bases are nested prefixes of a common random orthonormal 12-column basis, shared by fresh and lazy PSNG. We retain five initial power iterations, two per refresh, $\omega = 1$, and the same Armijo settings. The initial spectral stepsize rule is applied at each rank, with its HVPs and time charged to each optimization run.

Three data seeds and three timing repeats give 90 runs, all reaching relative gradient 10^{-8} within the common 30,000-iteration cap. Rank and method order are randomized within each timing repeat. Figure 14 reports first-hit costs at 10^{-6}, using within-seed timing medians before aggregating over seeds.

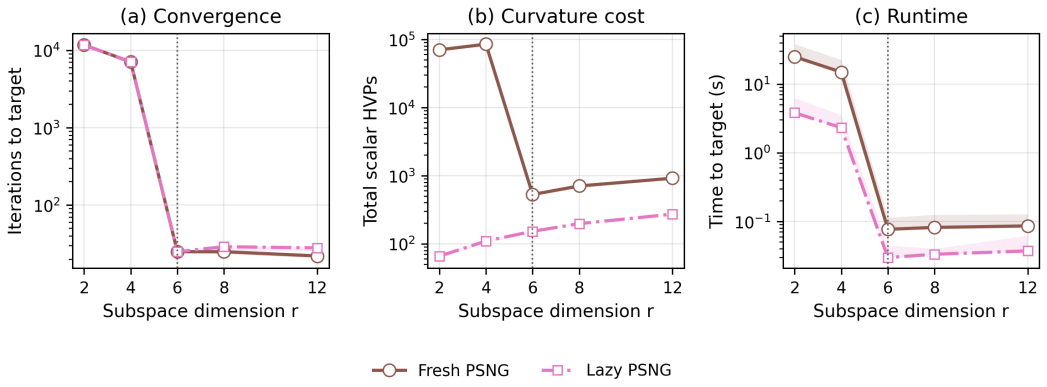

Figure 14: Rank sensitivity on fixed data. Iterations (left), total scalar HVPs (middle), and time (right) to relative gradient 10^{-6} at $a = 1$. The data always have six large-curvature directions, marked by the vertical dotted line; only the algorithm rank changes. Total HVPs and time include the initial spectral parameter selection. Curves and shaded regions show medians and interquartile ranges over three data seeds after aggregating three timing repeats per seed. All vertical axes are logarithmic.

For $r < 6$, large eigenvalues remain in the complementary subspace and restrict its gradient stepsize. Increasing r from 4 to 6 reduces the fresh and lazy median iteration counts from 7,117–7,123 to 25. The median complementary condition number at the reference solution falls from 1184 at $r = 4$ to 5.02 at $r = 6$, explaining the large iteration reduction. Increasing r further to 12 lowers it only to 4.46. Fresh PSNG changes from 25 to 22 median iterations between $r = 6$ and $r = 12$, while its total HVP count increases from 531 to 921; lazy PSNG changes from 25 to 28 iterations and from 153 to 273 HVPs. Each block product costs r scalar HVPs, so the extra directions raise curvature cost even when the iteration count decreases.

The increase in lazy iterations reflects the interaction between subspace accuracy and curvature reuse. For $r > 6$, the selected boundary lies among closely spaced tail eigenvalues, making accurate subspace estimation harder with the fixed power budget. At $r = 6, 8, 12$, every lazy run forms only four pairs, including initialization, and retains the remaining subspace error after the last refresh; fresh PSNG continues updating the subspace. An offline check at the reference solutions gives median linearized contraction factors of 0.669, 0.715, and 0.696 for the final stored lazy pairs at these ranks. Exact eigenspaces and current projected Hessians give 0.669, 0.650, and 0.645, respectively, with the same rank-specific stepsizes. Recomputing only the projected Hessian with the stored basis changes each factor by less than 1.2×10^{-4}. These diagnostics support subspace approximation as the main source of the excess contraction factor, which can offset the modest benefit of a larger rank.

At $r = 2$, lazy PSNG uses 66 total HVPs but requires 11,756 iterations. The results favor a rank near the pronounced eigengap: capturing the six large-curvature directions reduces gradient work, while additional directions offer only a small reduction in the remaining condition number.