# RANDOMIZED SUBSPACE NESTEROV ACCELERATED GRADIENT

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Randomized-subspace methods use only low-dimensional projected-gradient information, offering potential savings in directional-derivative queries and communication volume. Designing Nesterov-accelerated methods for general randomized subspaces that preserve this low-dimensional information structure while allowing improvements over full-gradient Nesterov acceleration in oracle complexity is technically nontrivial. We develop randomized-subspace Nesterov accelerated gradient methods for smooth convex and smooth strongly convex optimization under matrix smoothness and generic sketch moment assumptions. A key technical ingredient is a three-sequence formulation tailored to matrix smoothness. The resulting theory establishes accelerated convergence guarantees and makes explicit how matrix smoothness and the sketch distribution jointly determine the complexity. In oracle complexity, the dimension factor $d$ of full-dimensional Nesterov acceleration is replaced by the sketch-dependent factor $r\sqrt{\omega\ell}$, where $r$ is the subspace dimension and $\omega$ and $\ell$ are sketch moment parameters, identifying regimes in which randomized-subspace acceleration improves over full-dimensional Nesterov acceleration. For standard sketch families, we derive explicit complexity factors that enable direct comparison across sketch distributions.

## 1 INTRODUCTION

We consider the unconstrained optimization problem:

$$\min_{x\in\mathbb{R}^d} f(x),$$

where $f:\mathbb{R}^d\to\mathbb{R}$ is differentiable, and focus on both the smooth convex and smooth strongly convex settings. Whenever a minimizer exists, we denote by $x^\star$ an optimal solution and set $f^\star := f(x^\star)$. First-order methods are the workhorse of large-scale optimization due to their scalability and low per-iteration cost. Among them, acceleration techniques—most notably Nesterov's accelerated gradient (NAG) method (Nesterov, 1983; 2004)—play a fundamental role. They achieve optimal convergence rates, improving from $\mathcal{O}(1/N)$ to $\mathcal{O}(1/N^2)$ in the convex setting, and from $L/\mu$ to $\sqrt{L/\mu}$ dependence in the strongly convex case. As a result, acceleration has become an indispensable component in modern optimization algorithms.

**High-dimensional optimization and randomized subspace methods.** Many machine learning applications give rise to large-scale, high-dimensional optimization problems. Randomized subspace methods address this setting by using gradient information only within a randomly chosen low-dimensional subspace at each iteration. A representative update, introduced by Kozak et al. (2021), is

$$x_{k+1} = x_k - \alpha_k P_k P_k^\top \nabla f(x_k),$$

where $P_k\in\mathbb{R}^{d\times r}$ is a random sketch matrix with $r\le d$, and $r$ is the dimension of the subspace. When $P_k$ is chosen as a coordinate-sampling matrix, this framework reduces to randomized coordinate or block-coordinate gradient descent. Randomized coordinate and block-coordinate methods have been extensively studied in the optimization literature (Nesterov, 2012; Richtárik & Takáč, 2014; Lu & Xiao, 2015; Wright, 2015). It also recovers standard gradient descent when $r=d$ and $P_kP_k^\top = I_d$.



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

Table 1: Comparison of oracle complexity under a directional-derivative oracle model, where each directional-derivative evaluation counts as one oracle call. Here $R_0 := \|x_0 - x^\star\|$, and the sketch parameters $\ell$ and $\omega$ are defined in Assumption 2.2. The RS-GD bounds of Kozak et al. (2021) are re-evaluated under the same assumptions as our proposed method; see Section H. After instantiating $\omega$ and $\ell$ for standard sketches, the resulting oracle factors are compared in Section 5.

| Method | Convex | | Strongly convex | |
|---|---|---|---|---|
| GD | $\mathcal{O}\big(dR_0^2\frac{L}{\epsilon}\big)$ | | $\mathcal{O}\Big(d\frac{L}{\mu}\log\frac{1}{\epsilon}\Big)$ | |
| RS-GD (Kozak et al., 2021) | $\mathcal{O}\big(\omega r R_0^2\frac{L}{\epsilon}\big)$ | Prop. H.1 | $\mathcal{O}\Big(\ell r\frac{L}{\mu}\log\frac{1}{\epsilon}\Big)$ | Prop. H.2 |
| NAG (Nesterov, 1983; 2004) | $\mathcal{O}\Big(dR_0\sqrt{\frac{L}{\epsilon}}\Big)$ | | $\mathcal{O}\Big(d\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\Big)$ | |
| **RS-NAG (this work)** | $\mathcal{O}\Big(r\sqrt{\omega\ell}\,R_0\sqrt{\frac{L}{\epsilon}}\Big)$ | Thm. 3.3 | $\mathcal{O}\Big(r\sqrt{\omega\ell}\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\Big)$ | Thm. 4.3 |

Randomized subspace methods can be useful in at least two settings. First, they are relevant when directional derivatives are available without forming the full gradient. For instance, forward-mode automatic differentiation (AD) evaluates directional derivatives without storing the intermediate quantities typically needed by reverse-mode AD (Kozak et al., 2021). Under a directional-derivative oracle model, evaluating the full gradient requires $d$ directional-derivative queries, while evaluating $P_k^\top \nabla f(x)$ requires only $r$, corresponding to the columns of $P_k$. Second, in distributed optimization, randomized subspace methods allow the $r$-dimensional projected gradient $P_k^\top \nabla f(x)$ to be transmitted instead of the $d$-dimensional gradient $\nabla f(x)$, reducing the number of transmitted scalars from $d$ to $r$. See Section A for further discussion. These advantages have led to growing interest in randomized subspace gradient methods (Kozak et al., 2021; Nesterov & Spokoiny, 2017; Baydin et al., 2022; Nozawa et al., 2023; Omiya et al., 2026; Flynn et al., 2024).

**Limitations of existing acceleration methods.** Despite this progress, a fundamental gap remains. While acceleration is essential in first-order optimization, its integration into randomized subspace methods is still poorly understood. For coordinate descent methods, accelerated variants have been developed, including Nesterov's accelerated coordinate descent (Nesterov, 2012) and subsequent refinements such as APCG (Lin et al., 2014), APPROX (Fercoq & Richtárik, 2015), ALPHA (Qu & Richtárik, 2016), and non-uniform sampling schemes (Allen-Zhu et al., 2016; Nesterov & Stich, 2017). However, these methods rely heavily on coordinate-wise structure and do not extend naturally to general subspace directions. Beyond coordinate-based methods, SEGA (Hanzely et al., 2018) considers general linear gradient sketches, although its accelerated analysis is limited to coordinate sketches. Acceleration has also been studied for compressed gradient descent in distributed optimization (Li et al., 2020). However, directly specializing this framework to the projected-gradient operator $PP^\top$ does not yield an oracle-complexity improvement over full-dimensional Nesterov acceleration; see Section A.

A natural attempt is to directly combine Nesterov acceleration with randomized subspace gradients by replacing the full gradient with $P_k P_k^\top \nabla f$:

$$x_{k+1} = y_k - \eta P_k P_k^\top \nabla f(y_k), \qquad y_{k+1} = x_{k+1} + \beta_k(x_{k+1} - x_k).$$

However, accelerated guarantees for this direct two-sequence scheme are not obtained by a straightforward adaptation of the classical analysis. The classical two-sequence Nesterov analysis relies on delicate estimate-sequence arguments, which do not carry over directly to randomized subspace updates. In particular, the direct argument closes only under restrictive near-full-dimensional conditions; a precise discussion is deferred to Section J.

**Our approach and contributions.** These observations lead to the following question:

*Can one design accelerated methods that use only randomized subspace gradients while achieving improved oracle complexity?*

In this paper, we answer this question affirmatively. We propose randomized-subspace Nesterov accelerated gradient (RS-NAG) methods, to our knowledge the first accelerated framework for gen-

eral randomized subspace gradient methods that can achieve favorable oracle complexity compared with both non-accelerated randomized-subspace methods and standard NAG; see Table 1. Our key technical contribution is a novel three-sequence coupling tailored to the interaction between sketch moments and matrix smoothness.

**Our contributions.**

- **Accelerated randomized subspace methods.** We propose Nesterov-type randomized subspace methods for both convex and strongly convex optimization, recovering the corresponding classical Nesterov methods in the full-dimensional case.
- **Oracle complexity under matrix smoothness.** We prove convergence and oracle-complexity bounds that capture the interaction between matrix smoothness and the sketch distribution.
- **Comparison with full-dimensional acceleration.** Our bounds identify when randomized subspace acceleration can outperform full-dimensional Nesterov acceleration in oracle complexity.
- **Unified comparison of sketching strategies.** We analyze Haar, coordinate, and Gaussian sketches, revealing their relative convergence bounds and identifying optimal sketch dimensions in terms of oracle complexity.

**Additional probability guarantees.** Beyond the expectation bounds stated in Theorems 3.3 and 4.3, Section I also provides uniform-in-time high-probability bounds and almost-sure eventual rates with only mild losses. These results show that the accelerated behavior predicted by the expectation bounds is not merely an average-over-runs phenomenon, but persists with high probability uniformly over time and eventually along almost every run.

**Notation:** Unless stated otherwise, $\|\cdot\|$ denotes the Euclidean norm for vectors and the corresponding operator norm for matrices induced by the Euclidean norm. Let $I_d$ denote the $d\times d$ identity matrix. For a matrix $A \in \mathbb{R}^{d\times d}$, $\mathrm{diag}(A) \in \mathbb{R}^{d\times d}$ denotes the diagonal matrix whose diagonal entries coincide with those of $A$. For symmetric matrices $A, B \in \mathbb{R}^{d\times d}$, we write $A \preceq B$ if $B - A$ is positive semidefinite.

Throughout the paper, we assume that the ambient dimension satisfies $d \geq 2$. For sketch-based methods, we assume $1 \leq r \leq d$.

## 2 Preliminaries

### 2.1 Oracle complexity

We quantify oracle complexity by the number of directional-derivative queries. A query in direction $v \in \mathbb{R}^d$ returns $v^\top \nabla f(x)$ and counts as one oracle call. Under this model, evaluating the full gradient requires $d$ oracle calls, whereas evaluating $P^\top \nabla f(x)$, for $P \in \mathbb{R}^{d\times r}$, requires $r$ oracle calls. The same $d$-versus-$r$ counting also applies in distributed optimization: transmitting a full-dimensional gradient requires $d$ scalars, whereas transmitting an $r$-dimensional projected gradient requires $r$ scalars [1]; see Section B for details. Thus, the cost model reflects both directional-derivative query complexity and the number of transmitted scalars in distributed settings.

### 2.2 Problem setting and sketch assumptions

**Assumption 2.1** (Matrix smoothness)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable. Assume that there exists a fixed matrix $\mathbb{L} \in \mathbb{R}^{d\times d}$ with*

$$\mathbb{L} \succeq 0, \qquad \mathbb{L} \neq 0,$$

*such that, for all $x, y \in \mathbb{R}^d$,*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2}(y-x)^\top \mathbb{L}(y-x). \tag{1}$$

[1] Li et al. (2020) likewise count random-$r$ sparsification as $32r$ bits per iteration.

Throughout the paper, we write $L := \|\mathbb{L}\| > 0$. The standard scalar $L$-smooth descent lemma implies Assumption 2.1 by taking $\mathbb{L} = LI_d$. Thus, Assumption 2.1 is no stronger than standard scalar $L$-smoothness. Assumptions of this form have recently been used increasingly in matrix-smooth optimization and compression; see, e.g., Wang et al. (2022); Li et al. (2024); Maranjyan et al. (2025); Hanzely & Richtárik (2019); Flynn et al. (2024).

**Assumption 2.2** (Sketch moment conditions)**.** *Let $P \in \mathbb{R}^{d\times r}$ be a random matrix. Assume that for some constants $\omega > 0$ and $\ell > 0$,*

$$\mathbb{E}[PP^\top] = I_d, \tag{2}$$

$$\mathbb{E}[(PP^\top)^2] \preceq \omega I_d, \tag{3}$$

$$\mathbb{E}[PP^\top \mathbb{L} PP^\top] \preceq \ell L\, I_d. \tag{4}$$

The first two conditions are standard, while the third captures the interaction between the sketch and $\mathbb{L}$. Intuitively, this interaction can average curvature through the random sketch and yield a sharper bound than scalar smoothness alone. We will show in Section 5 that this assumption is satisfied by Haar, coordinate, and Gaussian sketches.

In all algorithms and analyses below, the sketch matrices $\{P_k\}_{k\geq 0}$ are assumed to be i.i.d. copies of a random matrix $P \in \mathbb{R}^{d\times r}$.

**Proposition 2.3.** *Under Assumptions 2.1 and 2.2, any constant $\omega$ satisfying* (3) *satisfies $\omega \geq d/r$, and any constant $\ell$ satisfying* (4) *satisfies $\ell \geq 1$. Consequently, every admissible pair $(\omega, \ell)$ satisfies $\ell\omega \geq 1$. Moreover, for any $\omega$ satisfying* (3)*,* (4) *holds with $\ell = \omega$.*

The proof is deferred to Section C. By the last statement of Proposition 2.3, for any admissible $\omega$, (4) also holds with $\ell = \omega$. Therefore, we assume without loss of generality throughout the rest of the paper that $\ell \leq \omega$; allowing $\ell < \omega$ is necessary for an improvement in our oracle bound, since $\ell = \omega$ gives $r\sqrt{\omega\ell} = r\omega \geq d$. The explicit constants derived in Section 5 and summarized in Table 2 satisfy this convention.

## 3 RS-NAG FOR CONVEX PROBLEMS (RS-NAG-C)

We first consider the convex, not necessarily strongly convex, setting.

**Assumption 3.1** (Convexity)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and convex. Assume that $f$ admits a minimizer.*

Under Assumption 3.1, we propose RS-NAG for convex problems (RS-NAG-C), a randomized-subspace variant of standard NAG for smooth convex optimization (Nesterov, 1983; 2004). The method is given in Algorithm 1. Compared with standard NAG, RS-NAG-C uses only the sketched gradient $P_k P_k^\top \nabla f(y_k)$ in place of the full gradient $\nabla f(y_k)$.

The corresponding two-sequence NAG recursion is recalled in Section D, (9)–(12). We prove below that RS-NAG-C reduces to this recursion when $r = d$ and $P_k P_k^\top = I_d$. Proofs for this section are deferred to Section D.

**Proposition 3.2** (Convex case: reduction to standard Nesterov)**.** *Suppose Assumptions 2.1 and 2.2 and Assumption 3.1 hold. Consider the full-sketch case $r = d$, and assume that*

$$P_k P_k^\top = I_d \qquad \text{for all } k \geq 0,$$

*which is the case, for example, for the Haar and block coordinate sketches. Then Assumption 2.2 is satisfied with $\omega = 1, \ell = 1$. With this choice, the $(x_k, y_k)$-sequence generated by Algorithm 1 coincides with classical two-sequence NAG for convex objectives.*

Therefore, the proposed randomized-subspace methods in the convex setting can be viewed as generalizations of the standard NAG.

We next state the convergence guarantee for RS-NAG-C.

**Theorem 3.3.** *Suppose Assumptions 2.1 and 2.2 and Assumption 3.1 hold, and let $\{x_k, y_k, z_k\}$ be generated by Algorithm 1. Then, for all $N \geq 1$,*

$$\mathbb{E}[f(x_N) - f^\star] \leq 2L\omega\ell\, \frac{\|x_0 - x^\star\|^2}{N^2}. \tag{5}$$

**Algorithm 1** RS-NAG-C

**Require:** $x_0 \in \mathbb{R}^d$, constants $L, \omega, \ell > 0$, and an i.i.d. sketch sequence $\{P_k\}_{k\geq 0}$

$m \leftarrow 1/(2L\ell)$
$A_0 \leftarrow 0$, $z_0 \leftarrow x_0$
**for** $k = 0, 1, 2, \ldots$ **do**
  $a_{k+1} \leftarrow \dfrac{m + \sqrt{m^2 + 2\omega m A_k}}{\omega}$
  $A_{k+1} \leftarrow A_k + a_{k+1}$
  $y_k \leftarrow \dfrac{A_k}{A_{k+1}} x_k + \dfrac{a_{k+1}}{A_{k+1}} z_k$
  $x_{k+1} \leftarrow y_k - \dfrac{1}{L\ell} P_k P_k^\top \nabla f(y_k)$
  $z_{k+1} \leftarrow z_k - a_{k+1} P_k P_k^\top \nabla f(y_k)$
**end for**

**Algorithm 2** RS-NAG-SC

**Require:** $x_0 \in \mathbb{R}^d$, constants $L, \mu, \omega, \ell > 0$, and an i.i.d. sketch sequence $\{P_k\}_{k\geq 0}$

$\theta \leftarrow \sqrt{\mu/(L\omega\ell)}$
$z_0 \leftarrow x_0$
**for** $k = 0, 1, 2, \ldots$ **do**
  $y_k \leftarrow \dfrac{1}{1+\theta} x_k + \dfrac{\theta}{1+\theta} z_k$
  $x_{k+1} \leftarrow y_k - \dfrac{1}{L\ell} P_k P_k^\top \nabla f(y_k)$
  $z_{k+1} \leftarrow (1-\theta) z_k + \theta y_k - \dfrac{\theta}{\mu} P_k P_k^\top \nabla f(y_k)$
**end for**

*In particular, for $R_0 := \|x_0 - x^\star\|$, the iteration complexity to guarantee $\mathbb{E}[f(x_N) - f^\star] \leq \epsilon$ is*

$$N = \mathcal{O}\left(\sqrt{\omega\ell}\, R_0 \sqrt{\frac{L}{\epsilon}}\right),$$

*and since one iteration uses $r$ oracle calls, the oracle complexity is*

$$\#\text{Oracle} = rN = \mathcal{O}\left(r\sqrt{\omega\ell}\, R_0 \sqrt{\frac{L}{\epsilon}}\right).$$

Compared with the convex RS-GD bound in Table 1, RS-NAG-C improves the accuracy dependence of the oracle complexity from $1/\epsilon$ to $1/\sqrt{\epsilon}$, matching the acceleration effect of NAG over GD.

## 4 RS-NAG FOR STRONGLY CONVEX PROBLEMS (RS-NAG-SC)

We next consider the strongly convex setting.

**Assumption 4.1** (Strong convexity)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and $\mu$-strongly convex with $\mu > 0$.*

Under Assumption 4.1, we propose RS-NAG for strongly convex problems (RS-NAG-SC), a randomized-subspace variant of standard NAG for smooth strongly convex optimization (Nesterov, 1983; 2004). The method is given in Algorithm 2. Compared with standard NAG, RS-NAG-SC uses only the sketched gradient $P_k P_k^\top \nabla f(y_k)$ in place of the full gradient $\nabla f(y_k)$.

The corresponding two-sequence NAG recursion is recalled in Section E, (39)–(41). We prove below that RS-NAG-SC reduces to this recursion when $r = d$ and $P_k P_k^\top = I_d$. Proofs for this section are deferred to Section E.

**Proposition 4.2** (Strongly convex case: reduction to standard Nesterov)**.** *Suppose Assumptions 2.1 and 2.2 and Assumption 4.1 hold. Consider the full-sketch case $r = d$, and assume that*

$$P_k P_k^\top = I_d \qquad \text{for all } k \geq 0,$$

*which is the case, for example, for the Haar and block coordinate sketches. Then Assumption 2.2 is satisfied with $\omega = 1, \ell = 1$. With this choice, the $(x_k, y_k)$-sequence generated by Algorithm 2 coincides with classical two-sequence NAG for strongly convex objectives.*

Therefore, the proposed randomized-subspace methods in the strongly convex setting can be viewed as generalizations of the standard NAG.

We next state the convergence guarantee for RS-NAG-SC.

Table 2: Sketch-dependent constants for Haar, coordinate, and Gaussian sketches. Here $\beta = d(d-r)/((d+2)(d-1))$.

| Sketch | $\omega$ | $\ell$ | $r\sqrt{\omega\ell}$ |
|---|---|---|---|
| Haar | $\frac{d}{r}$ | $\frac{d}{r}\left(1-\beta+\beta\frac{r_{\text{eff}}}{d}\right)$ | $d\sqrt{1-\beta+\beta\frac{r_{\text{eff}}}{d}}$ |
| Coordinate | $\frac{d}{r}$ | $\frac{d}{r}\left(\frac{r-1}{d-1}+\frac{d-r}{d-1}\delta_{\text{diag}}\right)$ | $d\sqrt{\frac{r-1}{d-1}+\frac{d-r}{d-1}\delta_{\text{diag}}}$ |
| Gaussian | $\frac{d+r+1}{r}$ | $\frac{r+1+r_{\text{eff}}}{r}$ | $\sqrt{(d+r+1)(r+1+r_{\text{eff}})}$ |

**Theorem 4.3.** *Suppose Assumptions 2.1 and 2.2 and Assumption 4.1 hold, and let $\{x_k, y_k, z_k\}$ be generated by Algorithm 2. Define $\Delta_0 \coloneqq f(x_0) - f^\star$. Then the parameter $\theta$ in Algorithm 2 satisfies $\theta \in (0, 1]$, and for all $N \geq 0$,*

$$\mathbb{E}[f(x_N) - f^\star] \leq 2(1-\theta)^N \Delta_0. \tag{6}$$

*In particular, the iteration complexity to guarantee $\mathbb{E}[f(x_N) - f^\star] \leq \epsilon$ is*

$$N = \mathcal{O}\left(\sqrt{\omega\ell}\sqrt{\frac{L}{\mu}}\,\log\frac{1}{\epsilon}\right),$$

*and since one iteration uses $r$ oracle calls, the oracle complexity is*

$$\#\text{Oracle} = rN = \mathcal{O}\left(r\sqrt{\omega\ell}\sqrt{\frac{L}{\mu}}\,\log\frac{1}{\epsilon}\right).$$

Compared with the strongly convex RS-GD bound in Table 1, RS-NAG-SC improves the condition-number dependence of the oracle complexity from $L/\mu$ to $\sqrt{L/\mu}$, matching the acceleration effect of NAG over GD.

## 5 Examples of sketches: Haar, coordinate, and Gaussian

We verify Assumption 2.2 for three standard sketches: Haar, Coordinate, and Gaussian sketches; see also Section F for a Rademacher sketch. We then compare the resulting sketch-dependent oracle-complexity factors. Proofs for this section are deferred to Section F. Throughout this section, let $L = \|\mathbb{L}\|$, where $\mathbb{L}$ is the matrix in Assumption 2.1, and define

$$r_{\text{eff}} \coloneqq \frac{\text{tr}(\mathbb{L})}{L}, \qquad \delta_{\text{diag}} \coloneqq \frac{\|\,\text{diag}(\mathbb{L})\|}{L}.$$

The oracle-complexity bounds in Theorems 3.3 and 4.3 are governed by the sketch-dependent factor $r\sqrt{\omega\ell}$. For full-dimensional Nesterov, the corresponding factor is $d$. Hence $r\sqrt{\omega\ell} < d$ means an improvement over full-dimensional Nesterov in our oracle bound.

**Proposition 5.1.** *Let $d \geq 2$, $1 \leq r \leq d$, $\mathbb{L} \succeq 0$, and $L = \|\mathbb{L}\| > 0$. Then Assumption 2.2 holds with the constants in Table 2 for the following sketches: Haar $P = \sqrt{d/r}\,R$, where $R$ consists of the first $r$ columns of a Haar-distributed orthogonal matrix; Coordinate $P = \sqrt{d/r}\,S$, where $S$ consists of $r$ uniformly sampled distinct columns of $I_d$; and Gaussian $P_{ij} \sim \mathcal{N}(0, 1/r)$ i.i.d. Moreover,*

$$1 \leq r_{\text{eff}} \leq d, \qquad \frac{1}{d} \leq \delta_{\text{diag}} \leq 1, \qquad \delta_{\text{diag}} \geq \frac{r_{\text{eff}}}{d}.$$

We next minimize the sketch-dependent oracle factor $r\sqrt{\omega\ell}$ over the sketch dimension $r$. Here $\omega$ and $\ell$ are the values in Table 2, which generally depend on $r$.

**Proposition 5.2** (Optimal sketch dimension and comparison of sketch constants)**.** *For each of the Haar, Coordinate, and Gaussian sketches, with $\omega$ and $\ell$ chosen as in Table 2, the factor $r\sqrt{\omega\ell}$ is minimized over $r \in \{1, \dots, d\}$ at $r = 1$. At $r = 1$, the Haar, Coordinate, and Gaussian factors, denoted by $Q_{\mathrm{H}}$, $Q_{\mathrm{C}}$, and $Q_{\mathrm{G}}$, respectively, are as follows.*

**Values and ranges.**

$$\begin{aligned} Q_{\mathrm{H}} &= d\sqrt{\frac{r_{\mathrm{eff}}+2}{d+2}}, & d\sqrt{\frac{3}{d+2}} \le Q_{\mathrm{H}} \le d, \\ Q_{\mathrm{C}} &= d\sqrt{\delta_{\mathrm{diag}}}, & \sqrt{d} \le Q_{\mathrm{C}} \le d, \\ Q_{\mathrm{G}} &= \sqrt{(d+2)(r_{\mathrm{eff}}+2)}, & \sqrt{3(d+2)} \le Q_{\mathrm{G}} \le d+2. \end{aligned}$$

**Relations between sketch factors.**

$$Q_{\mathrm{G}} = \left(1 + \frac{2}{d}\right) Q_{\mathrm{H}}, \qquad Q_{\mathrm{H}} \le \sqrt{3}\, Q_{\mathrm{C}}.$$

Moreover, for $\mathbb{L} = e_1 e_1^\top$,

$$Q_{\mathrm{H}} = \sqrt{\frac{3}{d+2}}\, Q_{\mathrm{C}}.$$

Under our oracle model, Proposition 5.2 shows that $r = 1$ minimizes the oracle factor $Q$ within each sketch family[2]. Since the full-dimensional Nesterov factor is $d$, the ranges above show that the factor $Q$ can decrease to about $\sqrt{d}$ in favorable cases. Haar and Coordinate are always no worse than full-dimensional Nesterov, and Haar is always better than Gaussian, although the two are nearly identical for large $d$. Compared with Coordinate, Haar is never worse by more than $\sqrt{3}$, while it can be much better, as shown by the final example.

As shown in Table 3, all three $Q$-values are much smaller than $d$ on all six real-world datasets, suggesting that the favorable regime $Q < d$ can arise naturally in practice.

## 6 Numerical Experiments

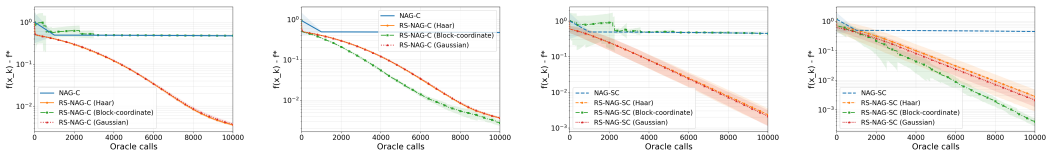
We evaluate the convex and strongly convex versions of RS-NAG, namely RS-NAG-C and RS-NAG-SC. Unless otherwise stated, each curve is the mean over independent random seeds, and the shaded region denotes mean $\pm$ one standard deviation across runs. The seeds determine the Gaussian initialization; for randomized-subspace methods, they also determine the sampled sketch sequence.

### 6.1 Quadratic objectives

We first consider four quadratic objectives $f(x) = \frac{1}{2} x^\top \mathbb{L} x$ on $\mathbb{R}^d$, with $d = 1000$ and $f^\star = 0$. These instances are designed to illustrate the sketch-dependent behavior associated with $r_{\mathrm{eff}}$ and $\delta_{\mathrm{diag}}$. The first two instances are convex but not strongly convex, while the last two are strongly convex. The diagonal instances have small $r_{\mathrm{eff}}$ and large $\delta_{\mathrm{diag}}$, whereas the dense instances have small $\delta_{\mathrm{diag}}$. The four matrices $\mathbb{L}$, together with the corresponding values of $L, \mu, r_{\mathrm{eff}}$, and $\delta_{\mathrm{diag}}$, are specified in Section G.1.

We use an oracle budget $10{,}000$. For randomized-subspace methods, we set $r = 1$ in the main experiments, which is optimal under our oracle model. Section G.2 reports results for several values of the sketch dimension $r$. We use $10$ independent random seeds with $x_0 \sim \mathcal{N}(0, I_d)$ for each seed. For randomized-subspace methods, we consider Haar, Block-coordinate, and Gaussian sketches. We compare NAG with *RS-NAG-C* in the convex setting and *RS-NAG-SC* in the strongly convex setting, and plot $f(x_k) - f^\star$ vs. oracle calls.

[2]For the Coordinate sketch with $r = 1$, taking the coordinate-wise smoothness constants as $L_i = \mathbb{L}_{ii}$, our bounds achieve the accelerated uniform coordinate-descent orders (Nesterov, 2012; Wright, 2015). Non-uniform accelerated coordinate methods exploit the individual constants $\{L_i\}_{i=1}^d$, including in the sampling distribution (Allen-Zhu et al., 2016), whereas our uniform Coordinate specialization depends only on $\max_i \mathbb{L}_{ii}$, while the same framework also covers non-coordinate sketches such as Haar and Gaussian.




(a) Convex diagonal (b) Convex dense (c) Strongly convex diagonal (d) Strongly convex dense

Figure 1: Oracle-axis convergence on the four quadratic problems. The horizontal axis shows the number of oracle calls, and the vertical axis shows the objective gap $f(x_k) - f^\star$ on a logarithmic scale.

**Discussion.** The results are consistent with the theoretical predictions in Proposition 5.2. On the diagonal instances, where $r_{\text{eff}}$ is small and $\delta_{\text{diag}}$ is large, Haar and Gaussian sketches outperform both the Block-coordinate sketch and full-dimensional Nesterov acceleration. On the dense instances, where $\delta_{\text{diag}}$ is small, the Block-coordinate sketch becomes the most effective, while Haar and Gaussian remain competitive. Thus, the observed oracle-axis behavior reflects the sketch-dependent quantities $Q_{\text{H}}, Q_{\text{G}}, Q_{\text{C}}$.

### 6.2 Logistic Regression

We next evaluate RS-NAG-SC on $\ell_2$-regularized logistic regression. Given binary classification data $\{(a_i, y_i)\}_{i=1}^n$, where $a_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$, we consider

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} \log\bigl(1 + \exp(-y_i a_i^\top x)\bigr) + \frac{\mu}{2}\|x\|_2^2, \qquad \mu > 0. \tag{7}$$

Since $\mu > 0$, the objective is $\mu$-strongly convex. For the matrix smoothness constant, let $A \in \mathbb{R}^{n\times d}$ be the data matrix whose $i$-th row is $a_i^\top$, and use

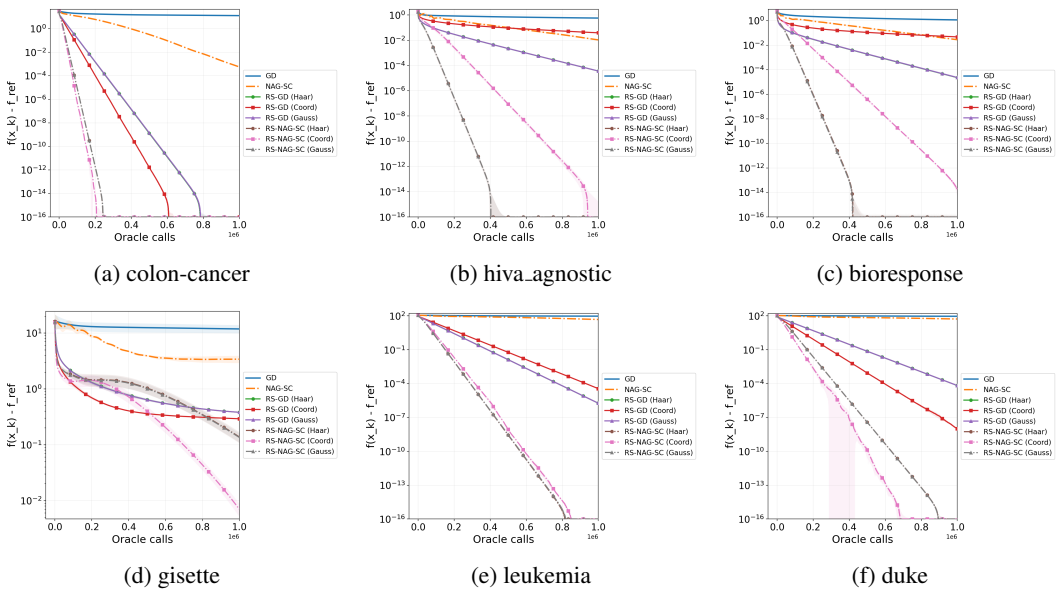
$$\mathbb{L} = \frac{1}{4n} A^\top A + \mu I_d, \tag{8}$$

see Section G.3 for the derivation.

We evaluate six real-world binary-classification benchmarks: colon-cancer (Alon et al., 1999), hiva_agnostic (Guyon et al., 2007), bioresponse (Hamner et al., 2012), gisette (Guyon et al., 2005), leukemia (Golub et al., 1999), and duke (Duke breast-cancer) (West et al., 2001). For each dataset, we set $\mu = 1/n$, use $r = 1$, and set $L = \|\mathbb{L}\|$, computed numerically as the largest eigenvalue of the matrix $\mathbb{L}$ in (8). We compare GD, NAG-SC, RS-GD, and RS-NAG-SC with Haar, coordinate, and Gaussian sketches. We initialize each run from a Gaussian random vector and plot the objective gap $f(x_k) - f_{\text{ref}}$ against oracle calls, where $f_{\text{ref}}$ is computed by L-BFGS-B (Byrd et al., 1995; Zhu et al., 1997). The oracle axis can also be read as communication bits under fixed-precision distributed implementations.[3] The results are shown in Figure 2, and the corresponding dataset-dependent quantities are summarized in Table 3. Additional implementation details, dataset sources, and reference-solver details are provided in Section G. Section G.4 further reports experiments on six additional real-world datasets.

**Discussion.** Overall, RS-NAG-SC performs strongly, and the empirical trends align with the dataset-dependent quantities in Table 3. For hiva_agnostic, bioresponse, and leukemia, $Q_{\text{H}}$ and $Q_{\text{G}}$ are smaller than $Q_{\text{C}}$, and the Haar and Gaussian sketches indeed perform better. By contrast, for colon-cancer, gisette, and duke, $Q_{\text{C}}$ is smaller than $Q_{\text{H}}$ and $Q_{\text{G}}$, which is consistent with the relatively strong performance of the coordinate sketch. In particular, for colon-cancer and leukemia, where the corresponding $Q$-values are relatively close, the empirical performance of the three sketches is also broadly comparable. Overall, these results suggest that the $Q$ values can serve as a useful practical guide when choosing the sketch distribution before running the method.

[3] This follows the communication-bit accounting in the experiments of Li et al. (2020), where an $r$-dimensional sparse message is counted as $32r$ bits. The essential comparison is $d$ versus $r$ transmitted scalars per iteration; with $b$-bit scalars, these correspond to $bd$ and $br$ bits.




(a) colon-cancer (b) hiva_agnostic (c) bioresponse

(d) gisette (e) leukemia (f) duke

Figure 2: Oracle-axis comparison for $\ell_2$-regularized logistic regression on six real-world datasets. The horizontal axis shows oracle calls, and the vertical axis shows $f(x_k) - f_{\mathrm{ref}}$ on a logarithmic scale, where $f_{\mathrm{ref}}$ is computed by L-BFGS-B. We compare GD, NAG-SC, RS-GD, and RS-NAG-SC with Haar, coordinate, and Gaussian sketches. Each curve is the mean over 3 random seeds, and the shaded region shows one standard deviation. For each dataset, $\mu = 1/n$, $r = 1$, and the oracle budget is $1{,}000{,}000$.

Table 3: Dataset-dependent quantities for the datasets. Here $d$ is the ambient dimension, $n$ is the number of training samples, and $Q_{\mathrm{H}}, Q_{\mathrm{G}}, Q_{\mathrm{C}}$ denote the $r = 1$ constants defined in Proposition 5.2 for the Haar, Gaussian, and Coordinate sketches, respectively.

| Dataset | $d$ | $Q_{\mathrm{H}}$ | $Q_{\mathrm{G}}$ | $Q_{\mathrm{C}}$ | $r_{\mathrm{eff}}$ | $\delta_{\mathrm{diag}}$ | $n$ |
|---|---|---|---|---|---|---|---|
| colon-cancer | 2000 | 132.1725 | 132.3047 | 116.1338 | 6.7435 | 0.0034 | 62 |
| hiva_agnostic | 1617 | 92.3197 | 92.4338 | 226.1505 | 3.2773 | 0.0196 | 4229 |
| bioresponse | 1776 | 88.0942 | 88.1934 | 231.3375 | 2.3746 | 0.0170 | 3751 |
| gisette | 5000 | 129.9973 | 130.0493 | 86.1400 | 1.3812 | 0.0003 | 6000 |
| leukemia | 7129 | 215.7402 | 215.8007 | 237.3443 | 4.5306 | 0.0011 | 38 |
| duke | 7129 | 189.2243 | 189.2773 | 146.8262 | 3.0240 | 0.0004 | 44 |

## 7 Conclusion

We introduced randomized-subspace Nesterov accelerated gradient methods for smooth convex and strongly convex optimization. The methods use projected gradients, recover standard full-dimensional Nesterov acceleration when $r = d$, and achieve oracle bounds in which the full-dimensional Nesterov factor $d$ is replaced by $r\sqrt{\omega\ell}$, allowing an improvement when $r\sqrt{\omega\ell} < d$. For three standard sketch distributions—Haar, coordinate, and Gaussian sketches—we derived explicit rates, compared the resulting sketch-dependent constants, and identified the theoretically preferred subspace dimension under our oracle model.

Future work includes designing sketch distributions beyond the canonical choices considered here. In particular, inspired by non-uniform sampling in accelerated coordinate descent (Allen-Zhu et al., 2016), it would be interesting to develop $\mathbb{L}$-aware sketch distributions that lead to faster convergence in oracle calls.

## AI USE STATEMENT

Generative AI tools were used to aid or polish the writing of this paper. The authors reviewed all AI-assisted content and take responsibility for the final content of the paper.

## REFERENCES


Zeyuan Allen-Zhu, Zheng Qu, Peter Richtárik, and Yang Yuan. Even faster accelerated coordinate descent using non-uniform sampling. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1110–1119, 2016.

Uri Alon, Naama Barkai, Daniel A. Notterman, Kurt Gish, Suzanne Ybarra, Daniel Mack, and Arnold J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745–6750, 1999.

Atılım Güneş Baydin, Barak A. Pearlmutter, Don Syme, Frank Wood, and Philip Torr. Gradients without backpropagation, 2022. arXiv:2202.08587.

Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

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

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

Olivier Fercoq and Peter Richtárik. Accelerated, parallel, and proximal coordinate descent. *SIAM Journal on Optimization*, 25(4):1997–2023, 2015.

Thomas Flynn, Patrick Johnstone, and Shinjae Yoo. Problem-dependent convergence bounds for randomized linear gradient compression, 2024. arXiv:2411.12898.

Louis Fournier, Stephane Rivaud, Eugene Belilovsky, Michael Eickenberg, and Edouard Oyallon. Can forward gradient match backpropagation? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10249–10264. PMLR, 2023.

Todd R. Golub, Donna K. Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P. Mesirov, Hilary Coller, Mignon L. Loh, James R. Downing, Mark A. Caligiuri, Clara D. Bloomfield, and Eric S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

Isabelle Guyon, Steve Gunn, Asa Ben-Hur, and Gideon Dror. Result analysis of the NIPS 2003 feature selection challenge. In *Advances in Neural Information Processing Systems*, volume 17, pp. 545–552, 2005.

Isabelle Guyon, Amir Saffari, Gideon Dror, and Gavin C. Cawley. Agnostic learning vs. prior knowledge challenge. In *Proceedings of the International Joint Conference on Neural Networks*, pp. 829–834, 2007.

Ben Hamner, dcthompson, and Jorg. Predicting a biological response. Kaggle competition, 2012. URL `https://kaggle.com/competitions/bioresponse`.

Filip Hanzely and Peter Richtárik. Accelerated coordinate descent with arbitrary sampling and best rates for minibatches. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 304–312, 2019.

Filip Hanzely, Konstantin Mishchenko, and Peter Richtárik. SEGA: Variance reduction via gradient sketching. In *Advances in Neural Information Processing Systems*, volume 31, pp. 2082–2093, 2018.

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

Mingyi Hong, Davood Hajinezhad, and Ming-Min Zhao. Prox-PDA: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1529–1538, 2017.

Markelle Kelly, Rachel Longjohn, and Kolby Nottingham. The UCI machine learning repository. `https://archive.ics.uci.edu`, 2023.

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

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

Hanmin Li, Avetik Karagulyan, and Peter Richtárik. Det-CGD: Compressed gradient descent with matrix stepsizes for non-convex optimization. In *Proceedings of the 12th International Conference on Learning Representations*, 2024.

Zhize Li, Dmitry Kovalev, Xun Qian, and Peter Richtárik. Acceleration for compressed gradient descent in distributed and federated optimization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5895–5904, 2020.

Qihang Lin, Zhaosong Lu, and Lin Xiao. An accelerated proximal coordinate gradient method. In *Advances in Neural Information Processing Systems*, volume 27, pp. 3059–3067, 2014.

Zhaosong Lu and Lin Xiao. On the complexity analysis of randomized block-coordinate descent methods. *Mathematical Programming*, 152(1–2):615–642, 2015.

Artavazd Maranjyan, Mher Safaryan, and Peter Richtárik. Gradskip: Communication-accelerated local gradient methods with better computational complexity. *Transactions on Machine Learning Research*, 2025.

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

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87 of *Applied Optimization*. Springer, 2004.

Yurii Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

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

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

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

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

Kunjal Panchal, Nisarg Parikh, Sunav Choudhary, Lijun Zhang, Yuriy Brun, and Hui Guan. Thinking forward: Memory-efficient federated finetuning of language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 69069–69119, 2024.

John C. Platt. Fast training of support vector machines using sequential minimal optimization. In Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola (eds.), *Advances in Kernel Methods: Support Vector Learning*, pp. 185–208. MIT Press, Cambridge, MA, 1999.

Danil Prokhorov. Ijcnn 2001 neural network competition. Slide presentation in IJCNN'01, Ford Research Laboratory, 2001.

Zheng Qu and Peter Richtárik. Coordinate descent with arbitrary sampling I: Algorithms and complexity. *Optimization Methods and Software*, 31(5):829–857, 2016.

Peter Richtárik and Martin Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1–2):1–38, 2014.

Lindon Roberts and Clément W. Royer. Direct search based on probabilistic descent in reduced spaces. *SIAM Journal on Optimization*, 33(4):3057–3082, 2023.

Mher Safaryan, Filip Hanzely, and Peter Richtárik. Smoothness matrices beat smoothness constants: Better communication compression techniques for distributed optimization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 25688–25702, 2021.

Suryanarayanan Manoj Sanu, Miguel A. Bessa, and Alejandro M. Aragón. Leveraging automatic differentiation in modern machine learning frameworks for (neural) topology optimization. *Structural and Multidisciplinary Optimization*, 69(5):126, 2026.

Luca Scalia, Rauno Cavallaro, and Andrea Cini. Structural sensitivity analysis assisted by automatic differentiation. *Structural and Multidisciplinary Optimization*, 69(7):177, 2026.

Yalan Song, Wouter J. M. Knoben, Martyn P. Clark, Dapeng Feng, Kathryn Lawson, Kamlesh Sawadekar, and Chaopeng Shen. When ancient numerical demons meet physics-informed machine learning: adjoint-based gradients for implicit differentiable modeling. *Hydrology and Earth System Sciences*, 28(13):3051–3077, 2024.

Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: Networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2013.

Bokun Wang, Mher Safaryan, and Peter Richtárik. Theoretically better and numerically faster distributed optimization with smoothness-aware quantization techniques. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9841–9852, 2022.

Mike West, Carrie Blanchette, Holly Dressman, Erich Huang, Seiichi Ishida, Rainer Spang, Harry Zuzan, John A. Olson, Jeffrey R. Marks, and Joseph R. Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 98(20):11462–11467, 2001.

Stephen J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015.

Tao Yang, Xinlei Yi, Junfeng Wu, Ye Yuan, Di Wu, Ziyang Meng, Yiguang Hong, Hong Wang, Zongli Lin, and Karl H. Johansson. A survey of distributed optimization. *Annual Reviews in Control*, 47:278–305, 2019.

Youjing Yu, Rui Xia, Qingxi Ma, Máté Lengyel, and Guillaume Hennequin. Second-order forward-mode optimization of recurrent neural networks for neuroscience. In *Advances in Neural Information Processing Systems*, volume 37, pp. 92101–92129, 2024.

Ciyou Zhu, Richard H. Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software*, 23(4):550–560, 1997.

# A Related work

## A.1 Directional-derivative access and automatic differentiation

Randomized-subspace methods can be particularly useful when directional derivatives are available without explicitly forming the full gradient. Forward-mode automatic differentiation (AD) provides one concrete way to realize such directional-derivative access. Although reverse-mode AD is computationally attractive for scalar-valued objectives, its memory footprint can become a limiting factor in large-scale differentiable optimization when evaluating the objective involves large intermediate states or iterative nonlinear computations (Song et al., 2024; Scalia et al., 2026); see also Sanu et al. (2026) for a recent discussion of the associated storage–recomputation trade-offs. Techniques such as checkpointing, selective recording, and problem-specific adjoint or implicit-differentiation rules can substantially mitigate this overhead, but typically introduce additional computation or rely on problem- and implementation-specific structure (Song et al., 2024; Sanu et al., 2026; Scalia et al., 2026). For an $r$-dimensional sketch $P_k = [p_{k,1}, \dots, p_{k,r}] \in \mathbb{R}^{d\times r}$, the projected gradient can be written as

$$P_k^\top \nabla f(x_k) = \left(p_{k,1}^\top \nabla f(x_k), \dots, p_{k,r}^\top \nabla f(x_k)\right)^\top,$$

so it can be obtained from $r$ directional derivatives along the columns of $P_k$. These directional derivatives can be evaluated directly using forward-mode AD, without first forming the full gradient or performing a full reverse sweep and incurring its associated storage requirements (Fournier et al., 2023; Panchal et al., 2024). Recent work has also demonstrated the use of forward-mode differentiation together with low-dimensional random subspaces to alleviate the memory requirements associated with backpropagation (Yu et al., 2024). Thus, memory-constrained differentiable optimization provides one setting in which such directional-derivative access can be useful.

## A.2 Other random-subspace derivative-free and zeroth-order methods

Random-subspace ideas have also been studied in related but different derivative-free and zeroth-order optimization settings. Examples include model-based derivative-free methods in random subspaces (Cartis & Roberts, 2023), direct-search methods using random subspaces (Roberts & Royer, 2023), zeroth-order methods based on orthogonal random directions (Kozak et al., 2023), and expected-decrease analyses for derivative-free algorithms using random subspaces (Hare et al., 2025). These works concern derivative-free and zeroth-order optimization, whereas our focus is on Nesterov-type acceleration under a directional-derivative oracle.

## A.3 Relation to accelerated compressed-gradient methods

Several works have studied acceleration in compressed-gradient methods for distributed optimization. These works are related to our goal of reducing the amount of first-order information used per iteration, but their models and resulting oracle-complexity implications are different from ours.

First, consider the generic-compressor framework of Li et al. (2020). If one takes

$$C(v) = PP^\top v$$

as a compressor and applies their algorithm, then their compression parameter becomes $\omega_{\mathrm{Li}} = \omega - 1$ under our notation, since $\mathbb{E}[PP^\top] = I_d$ and $\mathbb{E}[(PP^\top)^2] \preceq \omega I_d$. Accordingly, their bounds yield oracle complexities

$$\mathcal{O}\left(R_0\, r\omega \sqrt{\frac{L}{\epsilon}}\right) \qquad \text{and} \qquad \mathcal{O}\left(r\omega \sqrt{\frac{L}{\mu}} \log \frac{1}{\epsilon}\right)$$

in the convex and strongly convex cases, respectively, under our oracle model. By Proposition 2.3, we have $r\omega \geq d$. Thus, this generic-compressor route is not better than standard full-dimensional Nesterov acceleration in our oracle model. In this sense, if $PP^\top$ is used only through a generic unbiased-compressor framework, the benefit of randomized-subspace structure does not appear in the resulting oracle bounds.

The work of Safaryan et al. (2021) is also related, as it studies smoothness-aware compression and accelerated variants under matrix smoothness. Their setting is distributed: each local loss $f_i$ is

equipped with a local smoothness matrix $\mathbb{L}_i$, which is used together with a random diagonal sketch matrix to define a smoothness-aware compression mechanism. As discussed in their limitations section, this approach requires the server to store $\mathbb{L}_i^{1/2}$ for all workers; hence it is not expected to be practical for large $d$ unless the matrices $\mathbb{L}_i$ have special structure, such as low-rank or diagonal structure. This differs from our setting, where we allow general randomized subspace sketches $P_k$ and assume only a global matrix smoothness condition for the objective $f$. In particular, our framework does not require local smoothness matrices for individual workers or the storage of such matrices at the server. Thus, while both approaches exploit matrix smoothness, the optimization model and sketching mechanism are different.

## B Communication perspective on randomized subspace methods

Consider the distributed optimization problem

$$\min_{x\in\mathbb{R}^d} f(x), \qquad f(x) \coloneqq \frac{1}{n}\sum_{i=1}^n f_i(x),$$

where $n$ is the number of workers and $f_i$ is the local loss on worker $i$. This is a standard objective form in distributed optimization; see, e.g., Hong et al. (2017); Yang et al. (2019).

Each worker sends $\nabla f_i(y_k) \in \mathbb{R}^d$ to the server, which forms

$$\nabla f(y_k) = \frac{1}{n}\sum_{i=1}^n \nabla f_i(y_k).$$

To reduce the communication cost per round, we instead use a shared sketch of dimension $r$.

At iteration $k$, the server broadcasts the random seed defining $P_k \in \mathbb{R}^{d\times r}$, and each worker $i$ computes and sends

$$s_{i,k} \coloneqq P_k^\top \nabla f_i(y_k) \in \mathbb{R}^r.$$

The server then averages

$$\bar{s}_k \coloneqq \frac{1}{n}\sum_{i=1}^n s_{i,k}$$

and reconstructs

$$g_k \coloneqq P_k \bar{s}_k.$$

By linearity,

$$g_k = P_k\left(\frac{1}{n}\sum_{i=1}^n P_k^\top \nabla f_i(y_k)\right) = P_k P_k^\top \nabla f(y_k).$$

Hence the iterates coincide with those of the corresponding single-machine version of the proposed method applied to $f$. Therefore, all convergence guarantees proved for the single-machine method apply also to the distributed shared-sketch implementation. In particular, the corresponding convex and strongly convex convergence rates carry over to the distributed setting.

The matrix smoothness assumption is also natural in this setting. Indeed, if each local loss $f_i$ satisfies

$$f_i(y) \le f_i(x) + \langle \nabla f_i(x), y-x\rangle + \frac{1}{2}(y-x)^\top \mathbb{L}_i (y-x), \qquad \mathbb{L}_i \succeq 0,$$

then averaging over $i = 1,\dots,n$ yields

$$f(y) \le f(x) + \langle \nabla f(x), y-x\rangle + \frac{1}{2}(y-x)^\top \mathbb{L} (y-x), \qquad \mathbb{L} \coloneqq \frac{1}{n}\sum_{i=1}^n \mathbb{L}_i.$$

Such local smoothness matrices $\mathbb{L}_i$ arise naturally in the matrix-smoothness framework for distributed optimization; see Safaryan et al. (2021).

## C Missing Proofs for Section 2

*Proof of Proposition 2.3.* Set $A \coloneqq PP^\top$. Then $A$ is symmetric positive semidefinite and $\operatorname{rank}(A) \le r$. Also, by Assumption 2.1, $\mathbb{L} \succeq 0$, $\mathbb{L} \neq 0$, and hence $L = \|\mathbb{L}\| > 0$.

We prove in order the lower bounds $\omega \ge d/r$ and $\ell \ge 1$ for constants satisfying (3) and (4), respectively, the product bound $\ell\omega \ge 1$, and finally that (4) holds with $\ell = \omega$ whenever $\omega$ satisfies (3).

**(i) Lower bound on $\omega$.**

Let $\lambda_1, \dots, \lambda_d \ge 0$ be the eigenvalues of $A$. Since at least $d - r$ of them are zero, the Cauchy–Schwarz inequality gives

$$\big(\operatorname{tr}(A)\big)^2 = \Big(\sum_{i=1}^{d} \lambda_i\Big)^2 \le r \sum_{i=1}^{d} \lambda_i^2 = r\ \operatorname{tr}(A^2).$$

Hence

$$\operatorname{tr}(A^2) \ge \frac{(\operatorname{tr}(A))^2}{r}.$$

Taking expectations, we obtain

$$\mathbb{E}[\operatorname{tr}(A^2)] \ge \frac{1}{r}\mathbb{E}\big[(\operatorname{tr}(A))^2\big].$$

Now apply Jensen's inequality to the convex function $t \mapsto t^2$:

$$\mathbb{E}\big[(\operatorname{tr}(A))^2\big] \ge \big(\mathbb{E}[\operatorname{tr}(A)]\big)^2.$$

Using the unbiasedness condition (2),

$$\mathbb{E}[A] = I_d,$$

we obtain

$$\mathbb{E}[\operatorname{tr}(A)] = \operatorname{tr}(\mathbb{E}[A]) = \operatorname{tr}(I_d) = d.$$

Therefore,

$$\mathbb{E}[\operatorname{tr}(A^2)] \ge \frac{d^2}{r}.$$

On the other hand, taking the trace in (3) gives

$$\mathbb{E}[\operatorname{tr}(A^2)] = \operatorname{tr}(\mathbb{E}[A^2]) \le \operatorname{tr}(\omega I_d) = \omega d.$$

Combining the lower and upper bounds yields

$$\omega d \ge \frac{d^2}{r},$$

and hence

$$\omega \ge \frac{d}{r}.$$

**(ii) Lower bound on $\ell$.**

Let $u \in \mathbb{R}^d$ be a unit eigenvector corresponding to the largest eigenvalue $L$ of $\mathbb{L}$, that is,

$$\|u\| = 1, \qquad \mathbb{L}u = Lu.$$

Since $\mathbb{L} \succeq 0$, we may write its spectral decomposition as

$$\mathbb{L} = Luu^\top + \sum_{i=2}^{d} \lambda_i v_i v_i^\top, \qquad \lambda_i \ge 0.$$

Therefore,

$$\mathbb{L} - Luu^\top = \sum_{i=2}^{d} \lambda_i v_i v_i^\top \succeq 0,$$

so

$$\mathbb{L} \succeq L u u^\top.$$

Therefore, we have

$$A\mathbb{L}A \succeq L\, A u u^\top A.$$

Taking the quadratic form with $u$, we obtain

$$u^\top A\mathbb{L}Au \geq L\, u^\top A u u^\top A u = L\,(u^\top A u)^2.$$

Taking expectation, we get

$$u^\top \mathbb{E}[A\mathbb{L}A]u = \mathbb{E}[u^\top A\mathbb{L}Au] \geq L\,\mathbb{E}[(u^\top Au)^2].$$

Again by Jensen's inequality for the convex function $t \mapsto t^2$,

$$\mathbb{E}[(u^\top Au)^2] \geq \big(\mathbb{E}[u^\top Au]\big)^2.$$

Using once more the unbiasedness condition $\mathbb{E}[A] = I_d$, we obtain

$$\big(\mathbb{E}[u^\top Au]\big)^2 = (u^\top \mathbb{E}[A]u)^2 = (u^\top I_d u)^2 = 1.$$

Therefore,

$$u^\top \mathbb{E}[A\mathbb{L}A]u \geq L.$$

Now let $\ell$ be any admissible constant in (4). Then

$$\mathbb{E}[A\mathbb{L}A] \preceq \ell L\, I_d.$$

Taking again the quadratic form with $u$, we get

$$u^\top \mathbb{E}[A\mathbb{L}A]u \leq u^\top(\ell L I_d)u = \ell L\|u\|^2 = \ell L.$$

Combining the lower and upper bounds yields

$$L \leq \ell L.$$

Since $L > 0$, we conclude that

$$\ell \geq 1.$$

**(iii) Product bound.**

We have

$$\omega \geq \frac{d}{r} \geq 1 \qquad \text{since } 1 \leq r \leq d.$$

Together with the lower bound $\ell \geq 1$, this implies that for every admissible pair $(\omega, \ell)$,

$$\ell\omega \geq 1.$$

**(iv) Admissibility of $\ell = \omega$.**

Since $\mathbb{L} \succeq 0$ and $L = \|\mathbb{L}\|$, we have

$$\mathbb{L} \preceq L I_d.$$

Hence

$$A\mathbb{L}A \preceq L A^2.$$

Taking expectations and using (3), we obtain

$$\mathbb{E}[A\mathbb{L}A] \preceq L\,\mathbb{E}[A^2] \preceq \omega L\, I_d.$$

That is,

$$\mathbb{E}[PP^\top \mathbb{L} PP^\top] \preceq \omega L\, I_d.$$

Therefore, the admissible constants $\ell$ in (4) contain $\omega$. Hence one may choose $\ell$ so that

$$\ell \leq \omega.$$

This completes the proof. □

## D MISSING PROOFS FOR SECTION 3

*Proof of Proposition 3.2.* We show that the resulting $(x_k, y_k)$-sequence coincides with the classical two-sequence Nesterov accelerated gradient method

$$x_{k+1}^{\mathrm{N}} \coloneqq y_k^{\mathrm{N}} - \frac{1}{L}\nabla f(y_k^{\mathrm{N}}), \tag{9}$$

$$t_{k+1} \coloneqq \frac{1+\sqrt{1+4t_k^2}}{2}, \tag{10}$$

$$\beta_k \coloneqq \frac{t_k - 1}{t_{k+1}}, \tag{11}$$

$$y_{k+1}^{\mathrm{N}} \coloneqq x_{k+1}^{\mathrm{N}} + \beta_k\big(x_{k+1}^{\mathrm{N}} - x_k^{\mathrm{N}}\big), \tag{12}$$

with initialization $x_0^{\mathrm{N}} = x_0$, $y_0^{\mathrm{N}} = x_0$, and $t_0 = 1$.

Since $r = d$ and $P_k P_k^\top = I_d$ for all $k$, we have almost surely

$$(P_k P_k^\top)^2 = I_d, \qquad P_k P_k^\top \mathbb{L} P_k P_k^\top = \mathbb{L}.$$

Taking expectations,

$$\begin{aligned}\mathbb{E}[PP^\top] &= I_d,\\ \mathbb{E}[(PP^\top)^2] &= I_d,\\ \mathbb{E}[PP^\top \mathbb{L} PP^\top] &= \mathbb{L}.\end{aligned}$$

Hence Assumption 2.2 is satisfied with the admissible choice

$$\omega = 1, \qquad \ell = 1.$$

By Algorithm 1, the iterates satisfy

$$m = \frac{1}{2L\ell}, \tag{13}$$

$$A_{k+1} = A_k + a_{k+1}, \tag{14}$$

$$y_k = \frac{A_k}{A_{k+1}} x_k + \frac{a_{k+1}}{A_{k+1}} z_k, \tag{15}$$

$$x_{k+1} = y_k - \frac{1}{L\ell} P_k P_k^\top \nabla f(y_k), \tag{16}$$

$$z_{k+1} = z_k - a_{k+1} P_k P_k^\top \nabla f(y_k). \tag{17}$$

We now fix $\omega = \ell = 1$. By (13),

$$m = \frac{1}{2L}.$$

By the scalar update in Algorithm 1, $a_{k+1} > 0$ satisfies

$$m(A_k + a_{k+1}) = \frac{\omega}{2} a_{k+1}^2.$$

Hence, since $\omega = 1$,

$$\frac{1}{2L}(A_k + a_{k+1}) = \frac{1}{2} a_{k+1}^2.$$

By (14), this is equivalent to

$$\frac{1}{2L} A_{k+1} = \frac{1}{2} a_{k+1}^2 \quad \Longleftrightarrow \quad A_{k+1} = L a_{k+1}^2.$$

By (15),

$$y_k = \frac{A_k}{A_{k+1}} x_k + \frac{a_{k+1}}{A_{k+1}} z_k.$$

By (16), together with $P_k P_k^\top = I_d$ and $\ell = 1$,

$$x_{k+1} = y_k - \frac{1}{L\ell} P_k P_k^\top \nabla f(y_k) = y_k - \frac{1}{L} \nabla f(y_k),$$

and (17) becomes

$$z_{k+1} = z_k - a_{k+1} P_k P_k^\top \nabla f(y_k) = z_k - a_{k+1} \nabla f(y_k).$$

Thus, in the case $r = d$ with $P_k P_k^\top = I_d$, the updates become

$$y_k = \frac{A_k}{A_{k+1}} x_k + \frac{a_{k+1}}{A_{k+1}} z_k, \tag{18}$$

$$x_{k+1} = y_k - \frac{1}{L} \nabla f(y_k), \tag{19}$$

$$z_{k+1} = z_k - a_{k+1} \nabla f(y_k), \tag{20}$$

$$A_{k+1} = A_k + a_{k+1}, \quad A_{k+1} = L a_{k+1}^2. \tag{21}$$

We next analyze the scalar sequence $\{A_k\}$. Define

$$t_k := \sqrt{L A_{k+1}}, \qquad k \geq 0.$$

Since $A_0 = 0$, at $k = 0$ we have from (21)

$$A_1 = A_0 + a_1 = a_1, \qquad A_1 = L a_1^2.$$

Hence

$$a_1 = L a_1^2,$$

so the nonzero solution is $a_1 = 1/L$. Therefore

$$A_1 = A_0 + a_1 = \frac{1}{L}, \qquad t_0 = \sqrt{L A_1} = 1.$$

For general $k \geq 0$, combining $A_{k+2} = L a_{k+2}^2$ from (21) with $A_{k+2} = A_{k+1} + a_{k+2}$ yields

$$L a_{k+2}^2 = A_{k+1} + a_{k+2}.$$

Multiplying both sides by $L$ and using

$$t_k^2 = L A_{k+1}, \qquad t_{k+1}^2 = L A_{k+2} = L^2 a_{k+2}^2,$$

we obtain

$$t_{k+1}^2 = t_k^2 + L a_{k+2}.$$

On the other hand, since

$$A_{k+2} = L a_{k+2}^2,$$

we have

$$L a_{k+2} = \sqrt{L^2 a_{k+2}^2} = \sqrt{L A_{k+2}} = t_{k+1},$$

where in the first equality we use $a_{k+2} > 0$ to select the positive root. Therefore

$$t_{k+1}^2 = t_k^2 + t_{k+1},$$

that is,

$$t_{k+1}^2 - t_{k+1} - t_k^2 = 0.$$

For each fixed $t_k$, the above quadratic equation has the positive root

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}, \qquad k \geq 0,$$

since $t_{k+1} > 0$. This is exactly the recursion (10).

We now eliminate $z_k$ and derive the two-sequence recursion for $(x_k, y_k)$. From (18) we have

$$z_k = \frac{A_{k+1} y_k - A_k x_k}{a_{k+1}} \tag{22}$$

for all $k \geq 0$. Using (19),

$$\nabla f(y_k) = L\big(y_k - x_{k+1}\big),$$

and substituting into the update (20) together with (22) gives

$$\begin{aligned} z_{k+1} &= z_k - a_{k+1}\nabla f(y_k) \\ &= \frac{A_{k+1}y_k - A_k x_k}{a_{k+1}} - a_{k+1}L\big(y_k - x_{k+1}\big). \end{aligned}$$

Since $A_{k+1} = La_{k+1}^2$, we have $La_{k+1} = A_{k+1}/a_{k+1}$, hence

$$z_{k+1} = \frac{A_{k+1}y_k - A_k x_k}{a_{k+1}} - \frac{A_{k+1}}{a_{k+1}}\big(y_k - x_{k+1}\big) = \frac{A_{k+1}x_{k+1} - A_k x_k}{a_{k+1}}. \tag{23}$$

Next, from (18) with $k$ replaced by $k+1$ we obtain

$$y_{k+1} = \frac{A_{k+1}}{A_{k+2}}x_{k+1} + \frac{a_{k+2}}{A_{k+2}}z_{k+1}.$$

Substituting the expression for $z_{k+1}$ from (23) yields

$$\begin{aligned} y_{k+1} &= \frac{A_{k+1}}{A_{k+2}}x_{k+1} + \frac{a_{k+2}}{A_{k+2}} \cdot \frac{A_{k+1}x_{k+1} - A_k x_k}{a_{k+1}} \\ &= \frac{A_{k+1}}{A_{k+2}}\Big(1 + \frac{a_{k+2}}{a_{k+1}}\Big)x_{k+1} - \frac{a_{k+2}A_k}{a_{k+1}A_{k+2}}x_k. \end{aligned}$$

Define

$$\beta_k \coloneqq \frac{a_{k+2}A_k}{a_{k+1}A_{k+2}}.$$

Then the coefficient of $x_k$ is $-\beta_k$, and the coefficient of $x_{k+1}$ is

$$\begin{aligned} \frac{A_{k+1}}{A_{k+2}}\Big(1 + \frac{a_{k+2}}{a_{k+1}}\Big) &= \frac{a_{k+1}A_{k+1} + a_{k+2}A_{k+1}}{a_{k+1}A_{k+2}} \\ &= \frac{a_{k+1}(A_{k+2} - a_{k+2}) + a_{k+2}(A_k + a_{k+1})}{a_{k+1}A_{k+2}} = 1 + \beta_k, \end{aligned}$$

where the second equality is derived from $A_{k+2} = A_{k+1} + a_{k+2}$ and $A_{k+1} = A_k + a_{k+1}$. Thus, we have

$$y_{k+1} = x_{k+1} + \beta_k\big(x_{k+1} - x_k\big).$$

It remains to express $\beta_k$ in terms of $\{t_k\}$ and identify it with (11). By the definition

$$t_k \coloneqq \sqrt{LA_{k+1}} \qquad (k \geq 0),$$

we have

$$A_k = \frac{t_{k-1}^2}{L}, \qquad A_{k+2} = \frac{t_{k+1}^2}{L} \qquad (k \geq 1).$$

Moreover, since

$$A_{k+1} = La_{k+1}^2, \qquad a_{k+1} > 0,$$

it follows that

$$t_k = \sqrt{LA_{k+1}} = \sqrt{L^2 a_{k+1}^2} = La_{k+1}, \qquad t_{k+1} = La_{k+2},$$

and hence

$$a_{k+1} = \frac{t_k}{L}, \qquad a_{k+2} = \frac{t_{k+1}}{L}.$$

Therefore, for $k \geq 1$,

$$\beta_k = \frac{a_{k+2}A_k}{a_{k+1}A_{k+2}} = \frac{\frac{t_{k+1}}{L} \cdot \frac{t_{k-1}^2}{L}}{\frac{t_k}{L} \cdot \frac{t_{k+1}^2}{L}} = \frac{t_{k-1}^2}{t_k t_{k+1}}.$$

Under the positivity condition, the recursion (10) is equivalent to

$$t_{j+1}^2 - t_{j+1} - t_j^2 = 0, \qquad j \geq 0.$$

In particular, for $k \geq 1$, applying this with $j = k-1$ gives

$$t_{k-1}^2 = t_k^2 - t_k.$$

Substituting into the expression for $\beta_k$ yields

$$\beta_k = \frac{t_k^2 - t_k}{t_k t_{k+1}} = \frac{t_k - 1}{t_{k+1}},$$

which coincides with (11) for $k \geq 1$. For $k = 0$ we have $A_0 = 0$, so $\beta_0 = 0$, and also $\beta_0 = (t_0 - 1)/t_1 = 0$, so the formula holds for all $k \geq 0$.

Summarizing, the pair $(x_k, y_k)$ generated by (18)–(21) satisfies

$$x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k), \qquad y_{k+1} = x_{k+1} + \frac{t_k - 1}{t_{k+1}}\big(x_{k+1} - x_k\big),$$

with the same initialization $x_0 = y_0 = x_0^{\mathrm{N}} = y_0^{\mathrm{N}}$ and the same sequence $\{t_k\}$ as in (9)–(12). Thus $(x_k, y_k)$ obeys exactly the same recursion and initial conditions as $(x_k^{\mathrm{N}}, y_k^{\mathrm{N}})$, and hence

$$x_k = x_k^{\mathrm{N}}, \qquad y_k = y_k^{\mathrm{N}} \quad \text{for all } k \geq 0.$$

This completes the proof. □

*Proof of Theorem 3.3.* Define

$$\Psi_k := \mathbb{E}\left[A_k\big(f(x_k) - f^\star\big) + \frac{1}{2}\left\|z_k - x^\star\right\|^2\right].$$

Fix $k \geq 0$, and let

$$\mathcal{F}_k := \sigma(P_0, \ldots, P_{k-1}), \qquad g_k := \nabla f(y_k).$$

By Algorithm 1, the iterates satisfy

$$m = \frac{1}{2L\ell},$$
$$A_{k+1} = A_k + a_{k+1}, \tag{24}$$
$$y_k = \frac{A_k}{A_{k+1}}x_k + \frac{a_{k+1}}{A_{k+1}}z_k, \tag{25}$$
$$x_{k+1} = y_k - \frac{1}{L\ell}P_k P_k^\top \nabla f(y_k), \tag{26}$$
$$z_{k+1} = z_k - a_{k+1}P_k P_k^\top \nabla f(y_k). \tag{27}$$

Moreover, by the scalar update in Algorithm 1, $a_{k+1} > 0$ satisfies

$$m(A_k + a_{k+1}) = \frac{\omega}{2}a_{k+1}^2.$$

Equivalently, by (24),

$$mA_{k+1} = \frac{\omega}{2}a_{k+1}^2. \tag{28}$$

By (26) and (1) applied with $x = y_k$ and $y = x_{k+1}$, we have

$$\begin{aligned} f(x_{k+1}) &\leq f(y_k) + \langle \nabla f(y_k), x_{k+1} - y_k\rangle + \frac{1}{2}(x_{k+1} - y_k)^\top \mathbb{L}(x_{k+1} - y_k) \\ &= f(y_k) - \frac{1}{L\ell}\left\langle g_k, P_k P_k^\top g_k\right\rangle + \frac{1}{2L^2\ell^2} g_k^\top P_k P_k^\top \mathbb{L} P_k P_k^\top g_k. \end{aligned}$$

Taking conditional expectation and using (2) and (4),

$$\begin{aligned}\mathbb{E}[f(x_{k+1}) \mid \mathcal{F}_k] &\le f(y_k) - \frac{1}{L\ell}\|g_k\|^2 + \frac{1}{2L^2\ell^2} g_k^\top \mathbb{E}[P_k P_k^\top \mathbb{L} P_k P_k^\top \mid \mathcal{F}_k] g_k \\ &\le f(y_k) - \frac{1}{L\ell}\|g_k\|^2 + \frac{1}{2L\ell}\|g_k\|^2 \\ &= f(y_k) - \frac{1}{2L\ell}\|g_k\|^2 \\ &= f(y_k) - m\|g_k\|^2 . \end{aligned} \tag{29}$$

By (27),

$$z_{k+1} - x^\star = (z_k - x^\star) - a_{k+1} P_k P_k^\top g_k.$$

Hence

$$\frac{1}{2}\|z_{k+1} - x^\star\|^2 = \frac{1}{2}\|z_k - x^\star\|^2 - a_{k+1}\left\langle z_k - x^\star, P_k P_k^\top g_k\right\rangle + \frac{a_{k+1}^2}{2}\left\|P_k P_k^\top g_k\right\|^2 .$$

Taking conditional expectation and using (2), (3),

$$\mathbb{E}\left[\frac{1}{2}\|z_{k+1} - x^\star\|^2 \middle| \mathcal{F}_k\right] \le \frac{1}{2}\|z_k - x^\star\|^2 - a_{k+1}\langle z_k - x^\star, g_k\rangle + \frac{\omega a_{k+1}^2}{2}\|g_k\|^2 . \tag{30}$$

Subtract $f^\star$ from both sides of (29), multiply by $A_{k+1}$, and then add (30). This gives

$$\begin{aligned}&\mathbb{E}\left[A_{k+1}\big(f(x_{k+1}) - f^\star\big) + \frac{1}{2}\|z_{k+1} - x^\star\|^2 \middle| \mathcal{F}_k\right] \\ &\le A_{k+1}\big(f(y_k) - f^\star\big) + \frac{1}{2}\|z_k - x^\star\|^2 - a_{k+1}\langle z_k - x^\star, g_k\rangle \\ &\quad - \left(A_{k+1}m - \frac{\omega a_{k+1}^2}{2}\right)\|g_k\|^2 .\end{aligned}$$

By (28), the last coefficient is zero, hence

$$\begin{aligned}&\mathbb{E}\left[A_{k+1}\big(f(x_{k+1}) - f^\star\big) + \frac{1}{2}\|z_{k+1} - x^\star\|^2 \middle| \mathcal{F}_k\right] \\ &\le A_{k+1}\big(f(y_k) - f^\star\big) + \frac{1}{2}\|z_k - x^\star\|^2 - a_{k+1}\langle z_k - x^\star, g_k\rangle .\end{aligned} \tag{31}$$

From (25),

$$A_{k+1} y_k = A_k x_k + a_{k+1} z_k.$$

Therefore

$$A_k(x_k - y_k) + a_{k+1}(x^\star - y_k) = a_{k+1}(x^\star - z_k).$$

Taking inner product with $g_k$,

$$-a_{k+1}\langle z_k - x^\star, g_k\rangle = A_k\langle x_k - y_k, g_k\rangle + a_{k+1}\langle x^\star - y_k, g_k\rangle . \tag{32}$$

By convexity of $f$,

$$f(x_k) \ge f(y_k) + \langle g_k, x_k - y_k\rangle , \tag{33}$$

$$f^\star = f(x^\star) \ge f(y_k) + \langle g_k, x^\star - y_k\rangle . \tag{34}$$

Multiply (33) by $A_k$, (34) by $a_{k+1}$, and add:

$$A_k\langle x_k - y_k, g_k\rangle + a_{k+1}\langle x^\star - y_k, g_k\rangle \le A_k\big(f(x_k) - f(y_k)\big) + a_{k+1}\big(f^\star - f(y_k)\big).$$

Using (32), we obtain

$$-a_{k+1}\langle z_k - x^\star, g_k\rangle \le A_k\big(f(x_k) - f(y_k)\big) + a_{k+1}\big(f^\star - f(y_k)\big). \tag{35}$$

Substitute (35) into (31):

$$\begin{aligned}
&\mathbb{E}\left[A_{k+1}\big(f(x_{k+1})-f^\star\big)+\frac{1}{2}\left\|z_{k+1}-x^\star\right\|^2\middle|\mathcal{F}_k\right]\\
&\le A_{k+1}\big(f(y_k)-f^\star\big)+\frac{1}{2}\left\|z_k-x^\star\right\|^2+A_k\big(f(x_k)-f(y_k)\big)+a_{k+1}\big(f^\star-f(y_k)\big)\\
&=A_k\big(f(x_k)-f(y_k)\big)+(A_{k+1}-a_{k+1})\big(f(y_k)-f^\star\big)+\frac{1}{2}\left\|z_k-x^\star\right\|^2\\
&=A_k\big(f(x_k)-f^\star\big)+\frac{1}{2}\left\|z_k-x^\star\right\|^2. \qquad (36)
\end{aligned}$$

By nonnegativity, the tower property applies to (36); hence, taking expectation yields

$$\Psi_{k+1}\le\Psi_k.$$

Iterating gives $\Psi_N\le\Psi_0$. Since $A_0=0$ and $z_0=x_0$,

$$\Psi_0=\frac{1}{2}\left\|x_0-x^\star\right\|^2.$$

Therefore,

$$A_N\,\mathbb{E}[f(x_N)-f^\star]\le\Psi_N\le\Psi_0=\frac{1}{2}\left\|x_0-x^\star\right\|^2. \qquad (37)$$

It remains to lower bound $A_N$. Define

$$\rho\coloneqq\frac{2m}{\omega}>0.$$

From (28),

$$mA_{k+1}=\frac{\omega}{2}a_{k+1}^2,$$

or equivalently,

$$A_{k+1}=\frac{a_{k+1}^2}{\rho}.$$

Since $A_{k+1}=A_k+a_{k+1}$, we have

$$A_{k+1}-A_k=\sqrt{\rho A_{k+1}}.$$

Let

$$B_k\coloneqq\sqrt{\frac{A_k}{\rho}}.$$

Then

$$B_{k+1}^2-B_k^2=B_{k+1}.$$

Hence

$$(B_{k+1}-B_k)(B_{k+1}+B_k)=B_{k+1}.$$

Since $A_{k+1}=A_k+a_{k+1}$ with $a_{k+1}>0$, we have $A_{k+1}>A_k\ge0$, so $B_{k+1}>B_k\ge0$ and in particular $B_{k+1}+B_k>0$. Thus,

$$B_{k+1}-B_k=\frac{B_{k+1}}{B_{k+1}+B_k}\ge\frac{1}{2}.$$

Since $B_0=0$, summing gives

$$B_N\ge\frac{N}{2}.$$

Therefore,

$$A_N=\rho B_N^2\ge\frac{\rho N^2}{4}=\frac{m}{2\omega}N^2. \qquad (38)$$

Combining (37) and (38), for all $N\ge1$,

$$\mathbb{E}[f(x_N)-f^\star]\le\frac{1}{2A_N}\left\|x_0-x^\star\right\|^2\le\frac{\omega}{m}\frac{\left\|x_0-x^\star\right\|^2}{N^2}.$$

Since $m = 1/(2L\ell)$, this gives

$$\mathbb{E}[f(x_N) - f^\star] \le 2L\omega\ell \, \frac{\|x_0 - x^\star\|^2}{N^2},$$

which is (5).

Finally, let $R_0 \coloneqq \|x_0 - x^\star\|$. Solving

$$2L\omega\ell \, \frac{R_0^2}{N^2} \le \epsilon$$

gives

$$N \;\ge\; R_0 \sqrt{\frac{2L\omega\ell}{\epsilon}}.$$

This completes the proof.

□

# E Missing Proofs for Section 4

*Proof of Proposition 4.2.* We show that the resulting $(x_k, y_k)$-sequence coincides with the classical two-sequence accelerated gradient method

$$x_{k+1}^{\mathrm{N}} \coloneqq y_k^{\mathrm{N}} - \frac{1}{L}\nabla f(y_k^{\mathrm{N}}), \tag{39}$$

$$y_{k+1}^{\mathrm{N}} \coloneqq x_{k+1}^{\mathrm{N}} + \beta\big(x_{k+1}^{\mathrm{N}} - x_k^{\mathrm{N}}\big), \tag{40}$$

where

$$\beta \coloneqq \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \tag{41}$$

with initialization $x_0^{\mathrm{N}} = x_0$ and $y_0^{\mathrm{N}} = x_0$.

As in the convex case, the assumptions $r = d$ and $P_k P_k^\top = I_d$ imply

$$(P_k P_k^\top)^2 = I_d, \qquad P_k P_k^\top \mathbb{L} P_k P_k^\top = \mathbb{L}$$

almost surely, so Assumption 2.2 is satisfied with

$$\omega = 1, \qquad \ell = 1.$$

By Algorithm 2, the iterates satisfy

$$\theta = \sqrt{\frac{\mu}{L\omega\ell}}, \tag{42}$$

$$y_k = \frac{1}{1+\theta} x_k + \frac{\theta}{1+\theta} z_k, \tag{43}$$

$$x_{k+1} = y_k - \frac{1}{L\ell} P_k P_k^\top \nabla f(y_k),$$

$$z_{k+1} = (1-\theta) z_k + \theta y_k - \frac{\theta}{\mu} P_k P_k^\top \nabla f(y_k). \tag{44}$$

We now fix $\omega = \ell = 1$. By (42),

$$\theta = \sqrt{\frac{\mu}{L}}.$$

Using (43)–(44) together with $P_k P_k^\top = I_d$, we obtain

$$y_k = \frac{1}{1+\theta} x_k + \frac{\theta}{1+\theta} z_k, \tag{45}$$

$$x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k), \tag{46}$$

$$z_{k+1} = (1-\theta) z_k + \theta y_k - \frac{\theta}{\mu}\nabla f(y_k). \tag{47}$$

We first rewrite the $z$-update. From (46),

$$x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k) \quad \Longrightarrow \quad \nabla f(y_k) = L\big(y_k - x_{k+1}\big).$$

Since $\mu = \theta^2 L$, it follows that

$$\frac{\theta}{\mu}\nabla f(y_k) = \frac{\theta}{\theta^2 L} L\big(y_k - x_{k+1}\big) = \frac{1}{\theta}\big(y_k - x_{k+1}\big).$$

Substituting this into the update for $z_{k+1}$ yields

$$\begin{aligned} z_{k+1} &= (1-\theta)z_k + \theta y_k - \frac{1}{\theta}\big(y_k - x_{k+1}\big) \\ &= (1-\theta)z_k + \Big(\theta - \frac{1}{\theta}\Big) y_k + \frac{1}{\theta} x_{k+1}. \end{aligned} \tag{48}$$

We now derive a convenient representation of $z_k$ in terms of $\{x_j\}_{j\le k}$. From (45),

$$(1+\theta)y_k = x_k + \theta z_k.$$

Thus

$$\begin{aligned} \theta(1-\theta)z_k &= (1-\theta)\big((1+\theta)y_k - x_k\big) \\ &= (1-\theta^2)y_k - (1-\theta)x_k. \end{aligned}$$

Equivalently,

$$\theta(1-\theta)z_k + (\theta^2 - 1)y_k = -(1-\theta)x_k. \tag{49}$$

Dividing (49) by $\theta > 0$ gives

$$(1-\theta)z_k + \Big(\theta - \frac{1}{\theta}\Big) y_k = -\frac{1-\theta}{\theta}\, x_k. \tag{50}$$

On the other hand, from (48) we have

$$z_{k+1} = (1-\theta)z_k + \Big(\theta - \frac{1}{\theta}\Big) y_k + \frac{1}{\theta} x_{k+1}. \tag{51}$$

Substituting (50) into (51) yields

$$\begin{aligned} z_{k+1} &= -\frac{1-\theta}{\theta} x_k + \frac{1}{\theta} x_{k+1} \\ &= x_{k+1} + \frac{1-\theta}{\theta}\big(x_{k+1} - x_k\big). \end{aligned}$$

Using this representation of $z_{k+1}$ in (45) at step $k+1$, we obtain

$$y_{k+1} = \frac{1}{1+\theta} x_{k+1} + \frac{\theta}{1+\theta} z_{k+1} = \frac{1}{1+\theta} x_{k+1} + \frac{\theta}{1+\theta}\left(x_{k+1} + \frac{1-\theta}{\theta}(x_{k+1} - x_k)\right),$$

and therefore

$$\begin{aligned} y_{k+1} &= \left(\frac{1}{1+\theta} + \frac{\theta}{1+\theta}\right) x_{k+1} + \frac{1-\theta}{1+\theta}(x_{k+1} - x_k) \\ &= x_{k+1} + \frac{1-\theta}{1+\theta}(x_{k+1} - x_k). \end{aligned}$$

Using $\theta = \sqrt{\mu/L}$, we can rewrite

$$\beta = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} = \frac{1 - \sqrt{\mu/L}}{1 + \sqrt{\mu/L}} = \frac{1-\theta}{1+\theta}.$$

Hence

$$y_{k+1} = x_{k+1} + \beta(x_{k+1} - x_k),$$

and the $x$-update (46) is exactly (39) with $y_k^{\mathrm{N}} = y_k$.

Thus the pair $(x_k, y_k)$ generated by (45)–(47) satisfies the same recursion (39)–(40) and the same initialization $x_0 = y_0 = x_0^{\mathrm{N}} = y_0^{\mathrm{N}}$ as the classical accelerated gradient method. Combining the base case $k = 0$ with the inductive step, we have shown that

$$x_k^{\mathrm{N}} = x_k, \qquad y_k^{\mathrm{N}} = y_k \quad \text{for all } k \ge 0.$$

This completes the proof of the proposition. □

*Proof of Theorem 4.3.* First, we show that the parameter $\theta$ in Algorithm 2 satisfies $\theta \in (0, 1]$. Since $\mathbb{L} \succeq 0$ and $L = \|\mathbb{L}\|$, we have

$$\mathbb{L} \preceq L I_d.$$

Hence Assumption 2.1 implies

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2.$$

Since $f$ is also $\mu$-strongly convex by Assumption 4.1, it follows that

$$\mu \leq L.$$

Moreover, by Proposition 2.3,

$$\omega \ell \geq 1.$$

Therefore, using the definition of $\theta$ in Algorithm 2,

$$0 < \theta^2 = \frac{\mu}{L\omega\ell} \leq \frac{\mu}{L} \leq 1,$$

and hence $\theta \in (0, 1]$.

Define

$$\Phi_k \coloneqq \mathbb{E}\Big[ f(x_k) - f^\star + \frac{\mu}{2} \|z_k - x^\star\|^2 \Big].$$

Fix $k \geq 0$, and let

$$\mathcal{F}_k \coloneqq \sigma(P_0, \ldots, P_{k-1}), \qquad g_k \coloneqq \nabla f(y_k).$$

By Algorithm 2, the iterates satisfy

$$\theta = \sqrt{\frac{\mu}{L\omega\ell}}, \tag{52}$$

$$y_k = \frac{1}{1+\theta} x_k + \frac{\theta}{1+\theta} z_k, \tag{53}$$

$$x_{k+1} = y_k - \frac{1}{L\ell} P_k P_k^\top \nabla f(y_k), \tag{54}$$

$$z_{k+1} = (1-\theta) z_k + \theta y_k - \frac{\theta}{\mu} P_k P_k^\top \nabla f(y_k). \tag{55}$$

As shown above, $\theta \in (0, 1]$. In particular, $\theta > 0$, so division by $\theta$ below is valid, and $1 - \theta \geq 0$.

By (54) and (1) applied with $x = y_k$ and $y = x_{k+1}$, we get

$$\begin{aligned} f(x_{k+1}) &\leq f(y_k) + \langle \nabla f(y_k), x_{k+1} - y_k \rangle + \frac{1}{2} (x_{k+1} - y_k)^\top \mathbb{L} (x_{k+1} - y_k) \\ &= f(y_k) - \frac{1}{L\ell} \left\langle g_k, P_k P_k^\top g_k \right\rangle + \frac{1}{2L^2\ell^2} g_k^\top P_k P_k^\top \mathbb{L} P_k P_k^\top g_k. \end{aligned}$$

Taking conditional expectation and using (2) and (4),

$$\begin{aligned} \mathbb{E}[f(x_{k+1}) \mid \mathcal{F}_k] &\leq f(y_k) - \frac{1}{L\ell} \|g_k\|^2 + \frac{1}{2L^2\ell^2} g_k^\top \mathbb{E}[P_k P_k^\top \mathbb{L} P_k P_k^\top \mid \mathcal{F}_k] g_k \\ &\leq f(y_k) - \frac{1}{2L\ell} \|g_k\|^2. \end{aligned}$$

Define

$$m \coloneqq \frac{1}{2L\ell}. \tag{56}$$

Then

$$\mathbb{E}[f(x_{k+1}) \mid \mathcal{F}_k] \leq f(y_k) - m \|g_k\|^2. \tag{57}$$

Define

$$w_k \coloneqq (1-\theta)(z_k - x^\star) + \theta(y_k - x^\star). \tag{58}$$

From (55),

$$\begin{aligned} z_{k+1} - x^\star &= (1-\theta)z_k + \theta y_k - \frac{\theta}{\mu} P_k P_k^\top g_k - x^\star \\ &= (1-\theta)(z_k - x^\star) + \theta(y_k - x^\star) - \frac{\theta}{\mu} P_k P_k^\top g_k \\ &= w_k - \frac{\theta}{\mu} P_k P_k^\top g_k. \end{aligned}$$

Hence

$$\frac{\mu}{2}\left\|z_{k+1} - x^\star\right\|^2 = \frac{\mu}{2}\left\|w_k\right\|^2 - \theta\left\langle w_k, P_k P_k^\top g_k\right\rangle + \frac{\theta^2}{2\mu}\left\|P_k P_k^\top g_k\right\|^2.$$

Taking conditional expectation and using (2), (3),

$$\mathbb{E}\left[\frac{\mu}{2}\left\|z_{k+1} - x^\star\right\|^2 \middle| \mathcal{F}_k\right] \le \frac{\mu}{2}\left\|w_k\right\|^2 - \theta\left\langle w_k, g_k\right\rangle + \frac{\omega\theta^2}{2\mu}\left\|g_k\right\|^2. \tag{59}$$

Summing (57) and (59), and subtracting $f^\star$,

$$\begin{aligned} &\mathbb{E}\left[f(x_{k+1}) - f^\star + \frac{\mu}{2}\left\|z_{k+1} - x^\star\right\|^2 \middle| \mathcal{F}_k\right] \\ &\quad \le f(y_k) - f^\star + \frac{\mu}{2}\left\|w_k\right\|^2 - \theta\left\langle w_k, g_k\right\rangle - \left(m - \frac{\omega\theta^2}{2\mu}\right)\left\|g_k\right\|^2. \end{aligned}$$

By (52) and (56),

$$m = \frac{\omega\theta^2}{2\mu},$$

so the $\left\|g_k\right\|^2$ term vanishes. Thus

$$\begin{aligned} &\mathbb{E}\left[f(x_{k+1}) - f^\star + \frac{\mu}{2}\left\|z_{k+1} - x^\star\right\|^2 \middle| \mathcal{F}_k\right] \\ &\quad \le f(y_k) - f^\star + \frac{\mu}{2}\left\|w_k\right\|^2 - \theta\left\langle w_k, g_k\right\rangle. \end{aligned} \tag{60}$$

From (53),

$$(1+\theta)y_k = x_k + \theta z_k \quad \Longrightarrow \quad \theta(z_k - y_k) = y_k - x_k.$$

Let

$$a_k \coloneqq z_k - x^\star, \qquad b_k \coloneqq y_k - x^\star.$$

Then, recalling the definition (58), we have

$$w_k = (1-\theta)a_k + \theta b_k.$$

Also,

$$\begin{aligned} \frac{\mu}{2}\left\|w_k\right\|^2 &= \frac{\mu}{2}\left\|(1-\theta)a_k + \theta b_k\right\|^2 \\ &= (1-\theta)\frac{\mu}{2}\left\|a_k\right\|^2 + \theta\frac{\mu}{2}\left\|b_k\right\|^2 - \theta(1-\theta)\frac{\mu}{2}\left\|a_k - b_k\right\|^2. \end{aligned} \tag{61}$$

Also,

$$\begin{aligned} -\theta\left\langle w_k, g_k\right\rangle &= -\theta\left\langle (1-\theta)a_k + \theta b_k, g_k\right\rangle \\ &= -\theta(1-\theta)\left\langle a_k, g_k\right\rangle - \theta^2\left\langle b_k, g_k\right\rangle \\ &= -\theta\left\langle b_k, g_k\right\rangle - \theta(1-\theta)\left\langle a_k - b_k, g_k\right\rangle \\ &= -\theta\left\langle b_k, g_k\right\rangle + (1-\theta)\left\langle x_k - y_k, g_k\right\rangle, \end{aligned} \tag{62}$$

where in the last step we used $a_k - b_k = z_k - y_k = (y_k - x_k)/\theta$.

Since $f$ is $\mu$-strongly convex, for any $u, v$,

$$f(u) \ge f(v) + \left\langle \nabla f(v), u - v\right\rangle + \frac{\mu}{2}\left\|u - v\right\|^2.$$

Applying this with $(u, v) = (x^\star, y_k)$ gives

$$\langle g_k, b_k \rangle = \langle g_k, y_k - x^\star \rangle \geq f(y_k) - f^\star + \frac{\mu}{2} \|y_k - x^\star\|^2 = f(y_k) - f^\star + \frac{\mu}{2} \|b_k\|^2 . \tag{63}$$

Applying it with $(u, v) = (x_k, y_k)$ gives

$$\langle g_k, x_k - y_k \rangle \leq f(x_k) - f(y_k) - \frac{\mu}{2} \|x_k - y_k\|^2 . \tag{64}$$

Substitute (61), (62), (63), (64) into (60):

$$\begin{aligned}
&\mathbb{E}\Big[f(x_{k+1}) - f^\star + \frac{\mu}{2} \|z_{k+1} - x^\star\|^2 \Big| \mathcal{F}_k\Big] \\
&\leq f(y_k) - f^\star + (1-\theta)\frac{\mu}{2} \|a_k\|^2 + \theta\frac{\mu}{2} \|b_k\|^2 - \theta(1-\theta)\frac{\mu}{2} \|a_k - b_k\|^2 \\
&\quad - \theta\Big(f(y_k) - f^\star + \frac{\mu}{2} \|b_k\|^2\Big) + (1-\theta)\Big(f(x_k) - f(y_k) - \frac{\mu}{2} \|x_k - y_k\|^2\Big) \\
&\leq f(y_k) - f^\star + (1-\theta)\frac{\mu}{2} \|a_k\|^2 + \theta\frac{\mu}{2} \|b_k\|^2 \\
&\quad - \theta\Big(f(y_k) - f^\star + \frac{\mu}{2} \|b_k\|^2\Big) + (1-\theta)(f(x_k) - f(y_k)) .
\end{aligned}$$

The coefficients of $f(y_k)$ and $\|b_k\|^2$ cancel. Using also $a_k = z_k - x^\star$, we obtain

$$\begin{aligned}
&\mathbb{E}\Big[f(x_{k+1}) - f^\star + \frac{\mu}{2} \|z_{k+1} - x^\star\|^2 \Big| \mathcal{F}_k\Big] \\
&\leq (1-\theta)\big(f(x_k) - f^\star\big) + (1-\theta)\frac{\mu}{2} \|z_k - x^\star\|^2 \\
&= (1-\theta)\left(f(x_k) - f^\star + \frac{\mu}{2} \|z_k - x^\star\|^2\right) .
\end{aligned} \tag{65}$$

By nonnegativity, the tower property applies to (65); hence, taking expectation gives

$$\Phi_{k+1} \leq (1-\theta)\Phi_k.$$

Iterating yields

$$\Phi_N \leq (1-\theta)^N \Phi_0.$$

Moreover,

$$\mathbb{E}[f(x_N) - f^\star] \leq \Phi_N.$$

By $\mu$-strong convexity,

$$f(x_0) - f^\star \geq \frac{\mu}{2} \|x_0 - x^\star\|^2 .$$

Since $z_0 = x_0$, this implies

$$\Phi_0 = f(x_0) - f^\star + \frac{\mu}{2} \|z_0 - x^\star\|^2 \leq 2\big(f(x_0) - f^\star\big) = 2\Delta_0.$$

Therefore,

$$\mathbb{E}[f(x_N) - f^\star] \leq \Phi_N \leq (1-\theta)^N \Phi_0 \leq 2(1-\theta)^N \Delta_0,$$

which is (6).

It remains to derive the complexity bound. Since $1 - \theta \leq e^{-\theta}$,

$$\mathbb{E}[f(x_N) - f^\star] \leq 2e^{-N\theta}\Delta_0.$$

Thus, to guarantee $\mathbb{E}[f(x_N) - f^\star] \leq \epsilon$, it suffices that

$$N \geq \frac{1}{\theta} \log\left(\frac{2\Delta_0}{\epsilon}\right) .$$

Using the definition of $\theta$ in Algorithm 2, this is

$$N \geq \sqrt{\frac{L\omega\ell}{\mu}} \log\left(\frac{2\Delta_0}{\epsilon}\right) .$$

This completes the proof. □

## F MISSING PROOFS FOR SECTION 5

*Proof of Proposition 5.1.* We verify the three parts of Assumption 2.2 for each sketch: (i) unbiasedness $\mathbb{E}[PP^\top] = I_d$, (ii) the second-moment bound $\mathbb{E}[(PP^\top)^2] \preceq \omega I_d$, and (iii) the matrix-smoothness interaction $\mathbb{E}[PP^\top \mathbb{L} PP^\top] \preceq \ell L I_d$.

**Haar sketch.** Let $U$ be Haar-distributed on $O(d)$, let $R \in \mathbb{R}^{d\times r}$ be the matrix formed by the first $r$ columns of $U$, and set $P = \sqrt{d/r}\, R$. It is standard that $\mathbb{E}[RR^\top] = \frac{r}{d} I_d$, so $\mathbb{E}[PP^\top] = I_d$, verifying unbiasedness.

For the second moment,

$$(PP^\top)^2 = PP^\top PP^\top = \frac{d}{r}\, PP^\top.$$

Taking expectations and using $\mathbb{E}[PP^\top] = I_d$ gives $\mathbb{E}[(PP^\top)^2] = \frac{d}{r} I_d$, so we may take $\omega_{\text{Haar}} = \frac{d}{r}$.

For the interaction with $\mathbb{L}$, write

$$L \coloneqq \|\mathbb{L}\|, \qquad r_{\text{eff}} \coloneqq \frac{\operatorname{tr}(\mathbb{L})}{\|\mathbb{L}\|}.$$

By results in Flynn et al. (2024),

$$\mathbb{E}[PP^\top \mathbb{L} PP^\top] = \frac{d}{r}(1-\beta)\mathbb{L} + \frac{\beta}{r}\operatorname{tr}(\mathbb{L})\, I_d,$$

where $\beta = \frac{d(d-r)}{(d+2)(d-1)}$. Using $\mathbb{L} \succeq 0$ and $\|\mathbb{L}\| = L$,

$$\mathbb{E}[PP^\top \mathbb{L} PP^\top] \preceq \left[\frac{d}{r}(1-\beta) + \frac{\beta}{r}\frac{\operatorname{tr}(\mathbb{L})}{L}\right] L\, I_d = \frac{d}{r}\left(1 - \beta + \beta\frac{r_{\text{eff}}}{d}\right) L\, I_d.$$

Thus (4) in Assumption 2.2 holds with

$$\ell_{\text{Haar}} = \frac{d}{r}\left(1 - \beta + \beta\frac{r_{\text{eff}}}{d}\right).$$

**Coordinate sketch.** Let $S \in \mathbb{R}^{d\times r}$ consist of $r$ distinct columns sampled uniformly from the identity matrix $I_d$, and let $P \coloneqq \sqrt{\frac{d}{r}} S$. Then $SS^\top$ is a coordinate projection, and it is straightforward to check that $\mathbb{E}[SS^\top] = \frac{r}{d} I_d$, so $\mathbb{E}[PP^\top] = I_d$.

For the second moment,

$$(PP^\top)^2 = PP^\top PP^\top = \frac{d}{r}\, PP^\top.$$

Taking expectations, we obtain $\mathbb{E}[(PP^\top)^2] = \frac{d}{r} I_d$, so we may take $\omega_{\text{Coord}} = \frac{d}{r}$.

For the interaction with $\mathbb{L}$, define

$$\delta_{\text{diag}} \coloneqq \frac{\|\operatorname{diag}(\mathbb{L})\|}{\|\mathbb{L}\|}.$$

By a result in Flynn et al. (2024),

$$\mathbb{E}[PP^\top \mathbb{L} PP^\top] = \frac{d}{r}\left(\frac{r-1}{d-1}\mathbb{L} + \frac{d-r}{d-1}\operatorname{diag}(\mathbb{L})\right).$$

Using $\|\mathbb{L}\| = L$ and $\|\operatorname{diag}(\mathbb{L})\| = \delta_{\text{diag}} L$,

$$\mathbb{E}[PP^\top \mathbb{L} PP^\top] \preceq \frac{d}{r}\left(\frac{r-1}{d-1}L + \frac{d-r}{d-1}\delta_{\text{diag}} L\right) I_d = \frac{d}{r}\left(\frac{r-1}{d-1} + \frac{d-r}{d-1}\delta_{\text{diag}}\right) L\, I_d.$$

Thus we may take

$$\ell_{\text{Coord}} = \frac{d}{r}\left(\frac{r-1}{d-1} + \frac{d-r}{d-1}\delta_{\text{diag}}\right).$$

**Gaussian sketch.** Finally, assume $P \in \mathbb{R}^{d\times r}$ has i.i.d. entries $P_{ij} \sim \mathcal{N}(0, 1/r)$. Then $\mathbb{E}[PP^\top] = I_d$ by construction. Let $G \coloneqq \sqrt{r}\, P$, so that $G_{ij} \sim \mathcal{N}(0,1)$ i.i.d. and $PP^\top = \frac{1}{r}GG^\top$.

For the second moment, a result from Flynn et al. (2024) implies

$$\mathbb{E}[GG^\top I_d GG^\top] = r(r+1)I_d + r\ \mathrm{tr}(I_d)\, I_d = r(r+1)I_d + rd\, I_d.$$

Hence

$$\mathbb{E}[(PP^\top)^2] = \frac{1}{r^2}\mathbb{E}[(GG^\top)^2] = \frac{1}{r^2}\big(r(r+1) + rd\big) I_d = \frac{d+r+1}{r}\, I_d,$$

so $\omega_{\mathrm{Gauss}} = \frac{d+r+1}{r}$.

For the interaction with $\mathbb{L}$, a result in Flynn et al. (2024) implies

$$\mathbb{E}[GG^\top \mathbb{L} GG^\top] = r(r+1)\mathbb{L} + r\ \mathrm{tr}(\mathbb{L})\, I_d.$$

Recalling that $P = G/\sqrt{r}$ and hence $PP^\top = \frac{1}{r}GG^\top$, we obtain

$$\mathbb{E}[PP^\top \mathbb{L} PP^\top] = \frac{1}{r^2}\,\mathbb{E}[GG^\top \mathbb{L} GG^\top] = \frac{r+1}{r}\,\mathbb{L} + \frac{\mathrm{tr}(\mathbb{L})}{r}\, I_d.$$

Writing

$$L \coloneqq \|\mathbb{L}\|, \qquad r_{\mathrm{eff}} \coloneqq \frac{\mathrm{tr}(\mathbb{L})}{\|\mathbb{L}\|},$$

and using $\mathbb{L} \succeq 0$ and $\|\mathbb{L}\| = L$, we obtain

$$\begin{aligned}\mathbb{E}[PP^\top \mathbb{L} PP^\top] &\preceq \frac{r+1}{r}\, L I_d + \frac{r_{\mathrm{eff}} L}{r}\, I_d \\ &= \frac{r+1+r_{\mathrm{eff}}}{r}\, L\, I_d.\end{aligned}$$

Thus,

$$\ell_{\mathrm{Gauss}} = \frac{r+1+r_{\mathrm{eff}}}{r}.$$

It remains to prove the stated basic bounds on $r_{\mathrm{eff}}$ and $\delta_{\mathrm{diag}}$.

Let $\lambda_1, \ldots, \lambda_d$ be the eigenvalues of $\mathbb{L}$. Since $\mathbb{L} \succeq 0$ and $L = \|\mathbb{L}\|$, we have

$$0 \le \lambda_i \le L \qquad (i = 1, \ldots, d),$$

and since $\mathbb{L} \neq 0$, at least one eigenvalue equals $L$. Therefore

$$L \le \mathrm{tr}(\mathbb{L}) = \sum_{i=1}^{d} \lambda_i \le dL,$$

which gives

$$1 \le r_{\mathrm{eff}} = \frac{\mathrm{tr}(\mathbb{L})}{L} \le d.$$

Next, since $\delta_{\mathrm{diag}} = \|\,\mathrm{diag}(\mathbb{L})\|/L$, we have

$$\delta_{\mathrm{diag}} = \frac{\max_{1\le i\le d} \mathbb{L}_{ii}}{L}.$$

Because $\mathbb{L} \succeq 0$, for each $i$,

$$\mathbb{L}_{ii} = e_i^\top \mathbb{L} e_i \le \|\mathbb{L}\| = L,$$

so $\delta_{\mathrm{diag}} \le 1$.

Also,

$$\max_{1\le i\le d} \mathbb{L}_{ii} \ge \frac{1}{d}\sum_{i=1}^{d} \mathbb{L}_{ii} = \frac{\mathrm{tr}(\mathbb{L})}{d} = \frac{L\, r_{\mathrm{eff}}}{d}.$$

Dividing by $L$ gives

$$\delta_{\mathrm{diag}} \ge \frac{r_{\mathrm{eff}}}{d}.$$

Since $r_{\mathrm{eff}} \ge 1$, this also implies

$$\delta_{\mathrm{diag}} \ge \frac{1}{d}.$$

This completes the proof. □

*Proof of Proposition 5.2.* We first prove that $\sqrt{\omega \ell r^2}$ is minimized at $r = 1$ for each sketch.

For the Haar sketch, by Table 2,

$$\sqrt{\omega_{\mathrm{Haar}} \ell_{\mathrm{Haar}} r^2} = d\sqrt{1 - \beta + \beta \frac{r_{\mathrm{eff}}}{d}}, \qquad \beta = \frac{d(d-r)}{(d+2)(d-1)}.$$

Since

$$1 - \beta + \beta \frac{r_{\mathrm{eff}}}{d} = 1 - \frac{(d-r)(d-r_{\mathrm{eff}})}{(d+2)(d-1)},$$

and $d - r_{\mathrm{eff}} \geq 0$ by Proposition 5.1, the quantity inside the square root is nondecreasing in $r$. Hence

$$\sqrt{\omega_{\mathrm{Haar}} \ell_{\mathrm{Haar}} r^2}$$

is minimized at $r = 1$.

For the Coordinate sketch,

$$\sqrt{\omega_{\mathrm{Coord}} \ell_{\mathrm{Coord}} r^2} = d\sqrt{\frac{r-1}{d-1} + \frac{d-r}{d-1}\delta_{\mathrm{diag}}} = d\sqrt{\delta_{\mathrm{diag}} + \frac{r-1}{d-1}(1 - \delta_{\mathrm{diag}})}.$$

Since $\delta_{\mathrm{diag}} \leq 1$ by Proposition 5.1, the quantity inside the square root is nondecreasing in $r$. Hence

$$\sqrt{\omega_{\mathrm{Coord}} \ell_{\mathrm{Coord}} r^2}$$

is minimized at $r = 1$.

For the Gaussian sketch,

$$\sqrt{\omega_{\mathrm{Gauss}} \ell_{\mathrm{Gauss}} r^2} = \sqrt{(d + r + 1)(r + 1 + r_{\mathrm{eff}})}.$$

Both factors are increasing in $r$, so this quantity is increasing in $r$, and therefore it is minimized at $r = 1$.

Substituting $r = 1$ into the three expressions gives

$$Q_{\mathrm{H}} = d\sqrt{\frac{r_{\mathrm{eff}} + 2}{d+2}}, \qquad Q_{\mathrm{C}} = d\sqrt{\delta_{\mathrm{diag}}}, \qquad Q_{\mathrm{G}} = \sqrt{(d+2)(r_{\mathrm{eff}} + 2)}.$$

By Proposition 5.1,

$$1 \leq r_{\mathrm{eff}} \leq d, \qquad \frac{1}{d} \leq \delta_{\mathrm{diag}} \leq 1, \qquad \delta_{\mathrm{diag}} \geq \frac{r_{\mathrm{eff}}}{d}.$$

Therefore,

$$d\sqrt{\frac{3}{d+2}} \leq Q_{\mathrm{H}} \leq d, \qquad \sqrt{d} \leq Q_{\mathrm{C}} \leq d, \qquad \sqrt{3(d+2)} \leq Q_{\mathrm{G}} \leq d + 2.$$

Next, the Haar and Gaussian factors satisfy

$$\frac{Q_{\mathrm{G}}}{Q_{\mathrm{H}}} = \frac{\sqrt{(d+2)(r_{\mathrm{eff}} + 2)}}{d\sqrt{(r_{\mathrm{eff}} + 2)/(d+2)}} = 1 + \frac{2}{d},$$

and hence

$$Q_{\mathrm{G}} = \left(1 + \frac{2}{d}\right) Q_{\mathrm{H}}.$$

For the Haar–Coordinate comparison, using $\delta_{\mathrm{diag}} \geq r_{\mathrm{eff}}/d$, we obtain

$$Q_{\mathrm{C}}^2 = d^2 \delta_{\mathrm{diag}} \geq d\, r_{\mathrm{eff}}.$$

On the other hand,

$$Q_{\mathrm{H}}^2 = d^2 \frac{r_{\mathrm{eff}} + 2}{d+2} \leq d(r_{\mathrm{eff}} + 2).$$

Thus

$$\frac{Q_{\mathrm{H}}^2}{Q_{\mathrm{C}}^2} \le \frac{r_{\mathrm{eff}} + 2}{r_{\mathrm{eff}}} \le 3,$$

where the last inequality follows from $r_{\mathrm{eff}} \ge 1$. Hence

$$Q_{\mathrm{H}} \le \sqrt{3}\, Q_{\mathrm{C}}.$$

Finally, if $\mathbb{L} = e_1 e_1^\top$, then

$$L = 1, \qquad r_{\mathrm{eff}} = \frac{\mathrm{tr}(\mathbb{L})}{L} = 1, \qquad \delta_{\mathrm{diag}} = \frac{\|\,\mathrm{diag}(\mathbb{L})\|}{L} = 1.$$

Therefore,

$$Q_{\mathrm{C}} = d, \qquad Q_{\mathrm{H}} = d\sqrt{\frac{3}{d+2}} = \sqrt{\frac{3}{d+2}}\, Q_{\mathrm{C}}.$$

This completes the proof. □

**Rademacher sketch.** Let $P \in \mathbb{R}^{d\times r}$ have i.i.d. entries

$$P_{ij} = \frac{\varepsilon_{ij}}{\sqrt{r}}, \qquad \Pr(\varepsilon_{ij} = 1) = \Pr(\varepsilon_{ij} = -1) = \frac{1}{2}.$$

We verify that Assumption 2.2 holds for this sketch.

Write

$$P = \frac{1}{\sqrt{r}}[\,\varepsilon_1\ \cdots\ \varepsilon_r\,],$$

where the $\varepsilon_j$ are independent Rademacher vectors. Since $\mathbb{E}[\varepsilon_j \varepsilon_j^\top] = I_d$, we have $\mathbb{E}[PP^\top] = I_d$.

For any symmetric $A$, the identity

$$\mathbb{E}[\varepsilon_i \varepsilon_j \varepsilon_k \varepsilon_\ell] = \delta_{ij}\delta_{k\ell} + \delta_{ik}\delta_{j\ell} + \delta_{i\ell}\delta_{jk} - 2\delta_{ij}\delta_{ik}\delta_{i\ell}$$

gives

$$\begin{aligned}\big[\mathbb{E}[\varepsilon\varepsilon^\top A \varepsilon\varepsilon^\top]\big]_{ij} &= \sum_{k,\ell} A_{k\ell}\mathbb{E}[\varepsilon_i\varepsilon_j\varepsilon_k\varepsilon_\ell] \\ &= 2A_{ij} + \mathrm{tr}(A)\delta_{ij} - 2A_{ii}\delta_{ij}.\end{aligned}$$

Hence

$$\mathbb{E}[\varepsilon\varepsilon^\top A\varepsilon\varepsilon^\top] = 2A + \mathrm{tr}(A) I_d - 2\,\mathrm{diag}(A).$$

Since

$$PP^\top = \frac{1}{r}\sum_{a=1}^{r} \varepsilon_a \varepsilon_a^\top,$$

we have

$$\begin{aligned}\mathbb{E}[PP^\top \mathbb{L} PP^\top] &= \frac{1}{r^2}\left[r(r-1)\mathbb{L} + r\big(2\mathbb{L} + \mathrm{tr}(\mathbb{L})I_d - 2\,\mathrm{diag}(\mathbb{L})\big)\right] \\ &= \frac{r+1}{r}\mathbb{L} + \frac{\mathrm{tr}(\mathbb{L})}{r} I_d - \frac{2}{r}\,\mathrm{diag}(\mathbb{L}).\end{aligned}$$

Taking $\mathbb{L} = I_d$ in the identity above gives

$$\mathbb{E}[(PP^\top)^2] = \frac{d + r - 1}{r} I_d,$$

so we may take $\omega = (d + r - 1)/r$. Moreover, for general $\mathbb{L}$,

$$\ell = \frac{\lambda_{\max}((r+1)\mathbb{L} + \mathrm{tr}(\mathbb{L})I_d - 2\,\mathrm{diag}(\mathbb{L}))}{r\,\|\mathbb{L}\|}.$$

Thus, Assumption 2.2 holds for the Rademacher sketch.

# G Experimental Details

## G.1 Quadratic instances

We describe the four quadratic instances used in Section 6.1. In all cases,

$$f(x) = \frac{1}{2} x^\top \mathbb{L} x, \qquad d = 1000, \qquad f^\star = 0.$$

**Convex diagonal.** Let

$$\mathbb{L} = \operatorname{diag}\left(1, \frac{1}{d-2}, \ldots, \frac{1}{d-2}, 0\right).$$

Then $\mathbb{L} \succeq 0$ but is not positive definite. Its eigenvalues are

$$\left\{1, \frac{1}{d-2}, \ldots, \frac{1}{d-2}, 0\right\}.$$

Hence

$$L = \|\mathbb{L}\| = 1, \qquad \operatorname{tr}(\mathbb{L}) = 1 + (d-2)\frac{1}{d-2} = 2, \qquad r_{\text{eff}} = \frac{\operatorname{tr}(\mathbb{L})}{\|\mathbb{L}\|} = 2.$$

Moreover, since the maximum diagonal entry of $\mathbb{L}$ is 1,

$$\delta_{\text{diag}} = \frac{\|\operatorname{diag}(\mathbb{L})\|}{\|\mathbb{L}\|} = 1.$$

**Convex dense.** Let

$$\mathbf{1} \coloneqq (1, \ldots, 1)^\top \in \mathbb{R}^d, \qquad u \coloneqq \frac{1}{\sqrt{d}}(1, -1, 1, -1, \ldots, 1, -1)^\top \in \mathbb{R}^d.$$

Since $d$ is even, $u^\top \mathbf{1} = 0$. Define

$$\mathbb{L} = \frac{1}{d-2} I_d + \left(1 - \frac{1}{d-2}\right) u u^\top - \frac{1}{d(d-2)} \mathbf{1}\mathbf{1}^\top.$$

Because $u$ and $\mathbf{1}$ are orthogonal, the eigenspaces split as follows. Along $u$, the eigenvalue is

$$\frac{1}{d-2} + \left(1 - \frac{1}{d-2}\right) = 1.$$

Along $\mathbf{1}$, the eigenvalue is

$$\frac{1}{d-2} - \frac{1}{d(d-2)}\, d = 0.$$

On the orthogonal complement of $\operatorname{span}\{u, \mathbf{1}\}$, the eigenvalue is

$$\frac{1}{d-2}.$$

Thus the eigenvalues are

$$\left\{1, \frac{1}{d-2}, \ldots, \frac{1}{d-2}, 0\right\},$$

and therefore

$$L = \|\mathbb{L}\| = 1, \qquad \operatorname{tr}(\mathbb{L}) = 1 + (d-2)\frac{1}{d-2} = 2, \qquad r_{\text{eff}} = \frac{\operatorname{tr}(\mathbb{L})}{\|\mathbb{L}\|} = 2.$$

Each diagonal entry equals

$$\frac{1}{d-2} + \frac{1}{d}\left(1 - \frac{1}{d-2}\right) - \frac{1}{d(d-2)} = \frac{2}{d},$$

and hence

$$\|\operatorname{diag}(\mathbb{L})\| = \frac{2}{d}, \qquad \delta_{\text{diag}} = \frac{\|\operatorname{diag}(\mathbb{L})\|}{\|\mathbb{L}\|} = \frac{2}{d}.$$

**Strongly convex diagonal.** Let

$$\mathbb{L} = \operatorname{diag}\left(1, \frac{1}{d-1}, \ldots, \frac{1}{d-1}\right).$$

Then the eigenvalues of $\mathbb{L}$ are

$$\left\{1, \frac{1}{d-1}, \ldots, \frac{1}{d-1}\right\}.$$

Therefore,

$$\mu = \lambda_{\min}(\mathbb{L}) = \frac{1}{d-1}, \qquad L = \|\mathbb{L}\| = \lambda_{\max}(\mathbb{L}) = 1.$$

Moreover,

$$\operatorname{tr}(\mathbb{L}) = 1 + (d-1)\frac{1}{d-1} = 2, \qquad r_{\text{eff}} = \frac{\operatorname{tr}(\mathbb{L})}{\|\mathbb{L}\|} = 2.$$

Since the maximum diagonal entry of $\mathbb{L}$ is 1, we have

$$\delta_{\text{diag}} = \frac{\|\operatorname{diag}(\mathbb{L})\|}{\|\mathbb{L}\|} = 1.$$

**Strongly convex dense.** Let

$$\mathbb{L} = \frac{1}{d-1} I_d + \frac{d-2}{d(d-1)} \mathbf{1}\mathbf{1}^\top.$$

The vector $\mathbf{1}$ is an eigenvector with eigenvalue

$$\frac{1}{d-1} + \frac{d-2}{d(d-1)} \cdot d = 1,$$

while every vector orthogonal to $\mathbf{1}$ is an eigenvector with eigenvalue

$$\frac{1}{d-1}.$$

Therefore, the eigenvalues of $\mathbb{L}$ are

$$\left\{1, \frac{1}{d-1}, \ldots, \frac{1}{d-1}\right\}.$$

Hence

$$\mu = \lambda_{\min}(\mathbb{L}) = \frac{1}{d-1}, \qquad L = \|\mathbb{L}\| = \lambda_{\max}(\mathbb{L}) = 1.$$

Moreover,

$$\operatorname{tr}(\mathbb{L}) = 1 + (d-1)\frac{1}{d-1} = 2, \qquad r_{\text{eff}} = \frac{\operatorname{tr}(\mathbb{L})}{\|\mathbb{L}\|} = 2.$$

Finally, each diagonal entry equals

$$\frac{1}{d-1} + \frac{d-2}{d(d-1)} = \frac{2}{d},$$

and hence

$$\|\operatorname{diag}(\mathbb{L})\| = \frac{2}{d}, \qquad \delta_{\text{diag}} = \frac{\|\operatorname{diag}(\mathbb{L})\|}{\|\mathbb{L}\|} = \frac{2}{d}.$$

### G.2 Effect of the sketch dimension

In the main quadratic experiments, we fixed the sketch dimension to $r = 1$. Here we additionally examine how the sketch dimension affects oracle-axis convergence. More precisely, for each of the four quadratic instances and for each sketch family (Haar, Block-coordinate, and Gaussian), we compare the proposed method with $r \in \{1, 10, 100\}$ against the corresponding full-dimensional accelerated method under the same oracle budget 10,000. All other settings are unchanged: $d = 1000$, 10 random seeds, and independent Gaussian initialization $x_0 \sim \mathcal{N}(0, I_d)$.

The purpose of this experiment is to examine whether the theoretical prediction that $r = 1$ is oracle-optimal is also reflected numerically in the quadratic examples. For convenience, we recall the four quadratic instances and their associated quantities $(r_{\text{eff}}, \delta_{\text{diag}})$:

- **Convex diagonal:**

$$\mathbb{L} = \operatorname{diag}\left(1, \frac{1}{d-2}, \ldots, \frac{1}{d-2}, 0\right), \qquad r_{\text{eff}} = 2, \qquad \delta_{\text{diag}} = 1.$$

- **Convex dense:**

$$\mathbb{L} = \frac{1}{d-2} I_d + \left(1 - \frac{1}{d-2}\right) uu^\top - \frac{1}{d(d-2)} \mathbf{1}\mathbf{1}^\top, \qquad r_{\text{eff}} = 2, \qquad \delta_{\text{diag}} = \frac{2}{d}.$$

- **Strongly convex diagonal:**

$$\mathbb{L} = \operatorname{diag}\left(1, \frac{1}{d-1}, \ldots, \frac{1}{d-1}\right), \qquad r_{\text{eff}} = 2, \qquad \delta_{\text{diag}} = 1.$$

- **Strongly convex dense:**

$$\mathbb{L} = \frac{1}{d-1} I_d + \frac{d-2}{d(d-1)} \mathbf{1}\mathbf{1}^\top, \qquad r_{\text{eff}} = 2, \qquad \delta_{\text{diag}} = \frac{2}{d}.$$

The results are shown in Figure 3. As summarized in Table 2, the oracle-complexity comparison is governed by the quantity $\sqrt{\omega \ell r^2}$, and Proposition 5.2 shows that this quantity is minimized at $r = 1$ for the Haar, Block-coordinate, and Gaussian sketches.

The plots are broadly consistent with this prediction. In all Haar and Gaussian panels, and also in the Block-coordinate panels for the two dense instances, the oracle-axis performance deteriorates as $r$ increases from 1 to 10 and 100. The only clear exceptions are the Block-coordinate panels for the convex diagonal and strongly convex diagonal instances; see Figures 3b and 3h.

This behavior is also explained by Table 2. For the Block-coordinate sketch,

$$\sqrt{\omega_{\text{Coord}} \ell_{\text{Coord}} r^2} = d\sqrt{\frac{r-1}{d-1} + \frac{d-r}{d-1}\delta_{\text{diag}}}.$$

For the two diagonal instances considered here, we have $\delta_{\text{diag}} = 1$, and hence

$$\sqrt{\omega_{\text{Coord}} \ell_{\text{Coord}} r^2} = d\sqrt{\frac{r-1}{d-1} + \frac{d-r}{d-1}} = d,$$

independently of $r$. Therefore, in these two cases, the theory itself does not predict any oracle-complexity improvement from taking a smaller sketch dimension, which explains why the separation among $r = 1, 10, 100$ is weak in Figures 3b and 3h.

Overall, except for these two Block-coordinate diagonal panels, the numerical results are well aligned with the theoretical prediction that $r = 1$ is oracle-optimal in the present quadratic examples.

### G.3 Matrix smoothness for logistic regression

We derive the matrix smoothness bound used in Section 6.2. Recall the $\ell_2$-regularized logistic regression objective

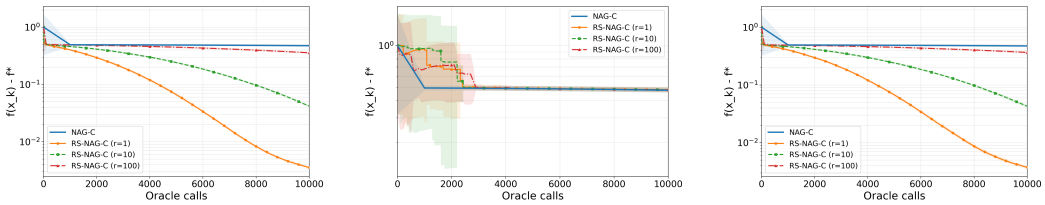
$$f(x) = \frac{1}{n} \sum_{i=1}^{n} \log\bigl(1 + \exp(-y_i a_i^\top x)\bigr) + \frac{\mu}{2}\|x\|_2^2, \qquad \mu > 0,$$

where $a_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$. Its gradient is

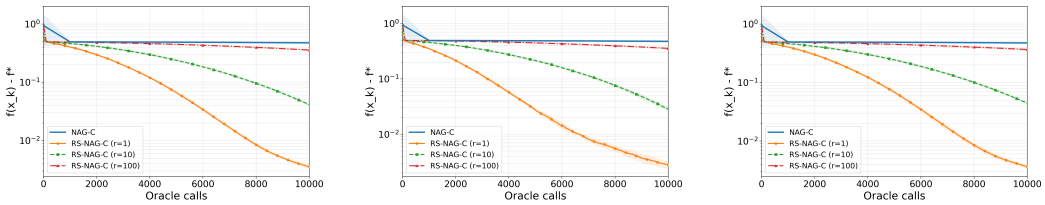
$$\nabla f(x) = -\frac{1}{n} \sum_{i=1}^{n} \frac{y_i a_i}{1 + \exp(y_i a_i^\top x)} + \mu x.$$

Let $\sigma(t) = 1/(1 + e^{-t})$. The Hessian is

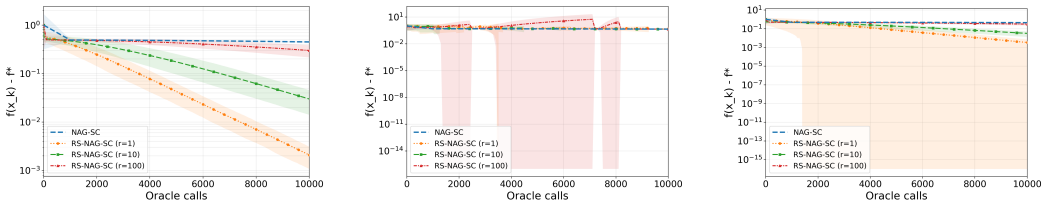
$$\nabla^2 f(x) = \frac{1}{n} \sum_{i=1}^{n} \sigma(y_i a_i^\top x)\bigl(1 - \sigma(y_i a_i^\top x)\bigr)\, a_i a_i^\top + \mu I_d.$$



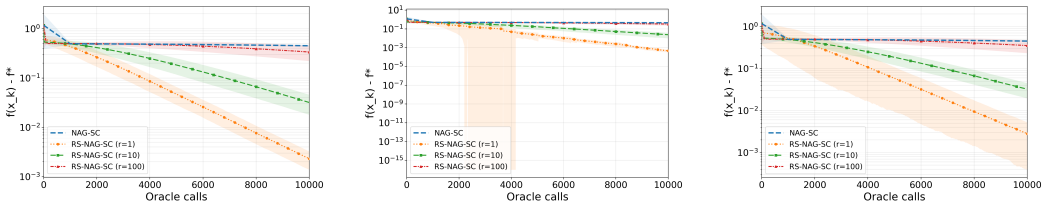

(a) Convex diagonal / Haar (b) Convex diagonal / Block-coordinate (c) Convex diagonal / Gaussian




(d) Convex dense / Haar (e) Convex dense / Block-coordinate (f) Convex dense / Gaussian

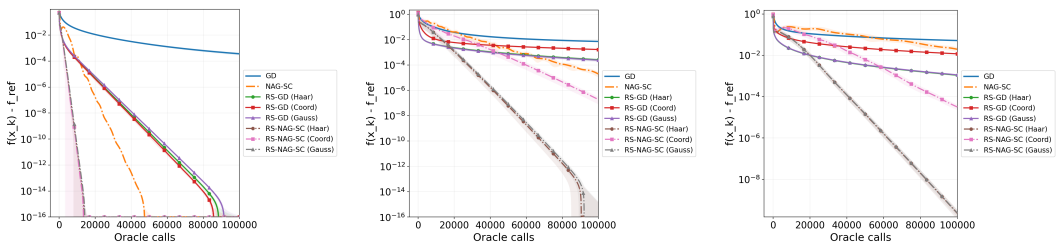



(g) Strongly convex diagonal / Haar (h) Strongly convex diagonal / Block-coordinate (i) Strongly convex diagonal / Gaussian

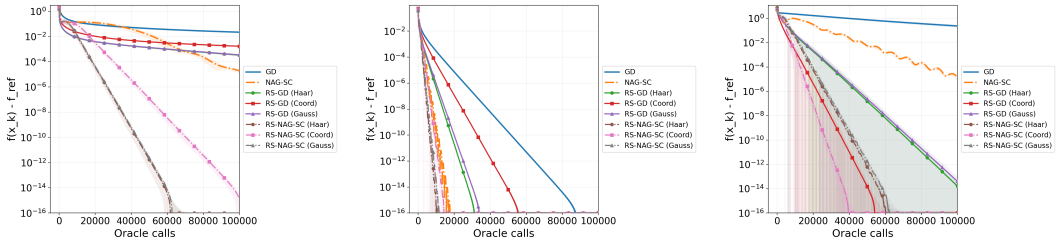



(j) Strongly convex dense / Haar (k) Strongly convex dense / Block-coordinate (l) Strongly convex dense / Gaussian

Figure 3: Oracle-axis convergence in the sketch-dimension scan. The horizontal axis shows the cumulative number of oracle calls, and the vertical axis shows the objective gap $f(x_k) - f^\star$ on a logarithmic scale. Each row corresponds to one quadratic instance, and each column corresponds to one sketch family. Each panel compares RS-NAG with $r \in \{1, 10, 100\}$ and the corresponding full-dimensional accelerated method.




(a) phishing (b) a9a (c) w8a




(d) mushroom (e) ijcnn1 (f) splice

Figure 4: Oracle-axis comparison for $\ell_2$-regularized logistic regression on six real-world datasets. The horizontal axis shows oracle calls, and the vertical axis shows $f(x_k) - f_{\mathrm{ref}}$ on a logarithmic scale, where $f_{\mathrm{ref}}$ is a reference value computed by L-BFGS-B. We compare GD, NAG-SC, RS-GD, and RS-NAG-SC with Haar, coordinate, and Gaussian sketches. Each plotted curve is the mean over 10 random seeds, and the shaded region shows one standard deviation. For each dataset, $\mu = 1/n$, and the oracle budget is 100,000.

Table 4: Dataset-dependent quantities for the logistic-regression datasets. Here $d$ is the ambient dimension, $n$ is the number of training samples, and $Q_{\mathrm{H}}, Q_{\mathrm{G}}, Q_{\mathrm{C}}$ denote the $r = 1$ constants defined in Proposition 5.2 for the Haar, Gaussian, and Coordinate sketches, respectively.

| Dataset | $d$ | $Q_{\mathrm{H}}$ | $Q_{\mathrm{G}}$ | $Q_{\mathrm{C}}$ | $r_{\mathrm{eff}}$ | $\delta_{\mathrm{diag}}$ | $n$ |
|---|---|---|---|---|---|---|---|
| phishing | 68 | 15.3670 | 15.8190 | 15.1421 | 1.5749 | 0.0496 | 11055 |
| a9a | 123 | 22.5681 | 22.9350 | 47.8971 | 2.2081 | 0.1516 | 32561 |
| w8a | 300 | 43.7200 | 44.0115 | 109.3820 | 4.4139 | 0.1329 | 49749 |
| mushroom | 112 | 21.0717 | 21.4480 | 34.8299 | 2.0352 | 0.0967 | 8124 |
| ijcnn1 | 22 | 13.1196 | 14.3123 | 17.5017 | 6.5350 | 0.6329 | 49990 |
| splice | 60 | 13.5925 | 14.0456 | 9.5503 | 1.1819 | 0.0253 | 1000 |

Since

$$0 \le \sigma(t)(1-\sigma(t)) \le \frac{1}{4} \qquad \text{for all } t \in \mathbb{R},$$

we have

$$\nabla^2 f(x) \preceq \frac{1}{4n} \sum_{i=1}^{n} a_i a_i^\top + \mu I_d.$$

If $A \in \mathbb{R}^{n \times d}$ denotes the data matrix whose $i$th row is $a_i^\top$, then $\sum_{i=1}^{n} a_i a_i^\top = A^\top A$. Thus we may take

$$\mathbb{L} = \frac{1}{4n} A^\top A + \mu I_d.$$

This gives the matrix smoothness bound used in the logistic-regression experiments.

### G.4 Additional logistic-regression results on standard benchmarks

**Experimental setup.** We evaluate the strongly convex logistic-regression setting on six real-world binary-classification benchmarks: phishing, a9a, w8a, mushroom, ijcnn1, and splice. The UCI-derived datasets are cited collectively through the UCI Machine Learning Repository (Kelly et al.,

2023); for w8a and ijcnn1, we follow the standard benchmark attributions (Platt, 1999; Prokhorov, 2001).

For each dataset, we use the $\ell_2$-regularized logistic objective in (7), with the regularization parameter set to $\mu = 1/n$, where $n$ is the number of training samples. We compare the full-gradient methods GD and NAG-SC with the randomized-subspace methods RS-GD and RS-NAG-SC. For the randomized-subspace methods, we use three sketch families: Haar, coordinate, and Gaussian. We use 10 random seeds, and initialize each run from a Gaussian random vector. For logistic regression, we use

$$\mathbb{L} = \frac{1}{4n} A^\top A + \mu I_d,$$

and set $L = \|\mathbb{L}\|$, computed numerically as the largest eigenvalue of $\mathbb{L}$. In the randomized-subspace methods, we use the $r = 1$ setting throughout, in accordance with the theoretical comparison developed above.

**Dataset-dependent quantities.** For each dataset, we also compute the matrix-smoothness-derived quantities $r_{\text{eff}}$ and $\delta_{\text{diag}}$, together with the corresponding $r = 1$ constants $Q_{\text{H}}$, $Q_{\text{G}}$, and $Q_{\text{C}}$ for the Haar, Gaussian, and coordinate sketches, respectively, as defined in Proposition 5.2. These values are reported in Table 4.

**Discussion.** We observe that RS-NAG-SC consistently achieves strong performance across all six datasets. Moreover, the relative convergence behavior among the Haar, Gaussian, and Coordinate sketches is broadly consistent with the dataset-dependent $Q$ values reported in Table 4. Indeed, on datasets such as a9a, w8a, mushroom, and ijcnn1, where $Q_{\text{H}}$ and $Q_{\text{G}}$ are relatively small, the Haar and Gaussian variants tend to converge faster than the Coordinate variant and also faster than NAG-SC. On the other hand, for phishing, where the Coordinate constant is also relatively small, the Coordinate variant performs comparably well to the Haar and Gaussian variants. For splice, where the Coordinate constant is particularly small, the Coordinate variant performs better. Overall, these results suggest that the $Q$ values can serve as a useful practical guide when choosing the sketch distribution before running the method.

### G.5 Experimental resources, dataset sources, and terms of use

**Computational resources.** All reported logistic-regression experiments were run on a single NVIDIA RTX A5000 GPU with 24564 MiB of memory and CUDA support. The machine had an AMD EPYC 7413 CPU and 503 GiB of system memory. The software environment used Python 3.10.12, NumPy 2.2.6, SciPy 1.15.3, scikit-learn 1.7.2, Matplotlib 3.10.8, and PyTorch 2.11.0 with CUDA support. The main and additional logistic-regression experiments each took about 70 hours on this machine. Thus, the reported logistic-regression experiments required about 140 GPU-hours in total. The reported quadratic experiments were run separately on a local personal computer using CPU only. The local machine had an Apple M1 processor and 8 GB of memory. The main and appendix quadratic experiments each completed within one wall-clock hour in our runs.

**Dataset sources and terms of use.** We use only previously released public benchmark datasets and do not redistribute any dataset. The datasets hiva_agnostic and bioresponse were obtained from OpenML (Vanschoren et al., 2013) with dataset IDs 1039 and 46912, respectively; their OpenML license fields are listed as `Public` and `Public Domain`, respectively.

The remaining datasets were obtained from the LIBSVM binary-classification dataset page (Chang & Lin, 2011), which provides LIBSVM-formatted versions of datasets from existing public benchmark collections and lists source and preprocessing information for each dataset. We cite the corresponding original source papers or dataset records whenever available. We checked the license or terms of use stated on the corresponding source or access pages whenever available. We use the datasets only for research evaluation and do not redistribute any dataset.

## H ORACLE COMPLEXITY OF THE BASIC RANDOMIZED-SUBSPACE GRADIENT METHOD

In Table 1, the row labeled "RS-GD (Kozak et al.)" refers to the basic randomized-subspace gradient iteration

$$x_{k+1} = x_k - \eta\, P_k P_k^\top \nabla f(x_k), \qquad k \geq 0, \tag{66}$$

introduced by Kozak et al. (2021). The exact oracle-complexity expressions shown in Table 1 are not stated in this form in Kozak et al. (2021); for convenience, we record below short derivations.

**Proposition H.1** (Convex rate for the basic randomized-subspace gradient method)**.** *Suppose Assumptions 2.1, 2.2 and 3.1 hold, with $\ell$ chosen so that $\ell \leq \omega$, as assumed throughout the paper. Let $\{x_k\}_{k\geq 0}$ be generated by* (66) *with the constant step-size*

$$\eta = \frac{1}{2\omega L}.$$

*Then the expected objective values are nonincreasing:*

$$\mathbb{E}[f(x_{k+1})] \leq \mathbb{E}[f(x_k)] \qquad \textit{for all } k \geq 0.$$

*Moreover, for every $N \geq 1$,*

$$\mathbb{E}[f(x_N) - f^\star] \;\leq\; \frac{2\omega L\, \|x_0 - x^\star\|^2}{N}. \tag{67}$$

*Consequently, defining $R_0 := \|x_0 - x^\star\|$, it suffices to take*

$$N \geq \frac{2\omega L R_0^2}{\epsilon}$$

*to guarantee $\mathbb{E}[f(x_N) - f^\star] \leq \epsilon$. Since each iteration uses $r$ oracle calls, the oracle complexity is*

$$\#\mathrm{Oracle} = \mathcal{O}\left( r\omega R_0^2 \frac{L}{\epsilon} \right). \tag{68}$$

*Proof.* Let

$$g_k := \nabla f(x_k).$$

Since $\mathbb{L} \succeq 0$ and $L = \|\mathbb{L}\|$, we have $\mathbb{L} \preceq L I_d$. Therefore Assumption 2.1 implies the standard $L$-smoothness inequality

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2 \qquad \text{for all } x, y \in \mathbb{R}^d.$$

In particular, since $f$ is convex and attains its minimum at $x^\star$, the standard smooth-convex inequality gives

$$\|\nabla f(x)\|^2 \leq 2L\big(f(x) - f^\star\big) \qquad \text{for all } x \in \mathbb{R}^d. \tag{69}$$

We first show monotonicity of the expected objective. Applying Assumption 2.1 with $x = x_k$ and $y = x_{k+1} = x_k - \eta P_k P_k^\top g_k$, we obtain

$$f(x_{k+1}) \leq f(x_k) - \eta\, g_k^\top P_k P_k^\top g_k + \frac{\eta^2}{2}\, g_k^\top P_k P_k^\top \mathbb{L} P_k P_k^\top g_k.$$

Taking conditional expectation with respect to $\mathcal{F}_k$, and using

$$\mathbb{E}[P_k P_k^\top \mid \mathcal{F}_k] = I_d, \qquad \mathbb{E}[P_k P_k^\top \mathbb{L} P_k P_k^\top \mid \mathcal{F}_k] \preceq \ell L I_d,$$

we obtain

$$\mathbb{E}[f(x_{k+1}) \mid \mathcal{F}_k] \leq f(x_k) - \eta\, \|g_k\|^2 + \frac{\eta^2 \ell L}{2}\, \|g_k\|^2 = f(x_k) - \eta\Big(1 - \frac{\eta \ell L}{2}\Big)\, \|g_k\|^2 .$$

With $\eta = 1/(2\omega L)$ and $\ell \leq \omega$, we have

$$\eta = \frac{1}{2\omega L} \leq \frac{1}{2\ell L},$$

and hence

$$1 - \frac{\eta \ell L}{2} \geq 0.$$

Therefore,

$$\mathbb{E}[f(x_{k+1}) \mid \mathcal{F}_k] \leq f(x_k).$$

Next, expanding the squared distance to $x^\star$, we obtain

$$\|x_{k+1} - x^\star\|^2 = \|x_k - x^\star\|^2 - 2\eta \left\langle x_k - x^\star, P_k P_k^\top g_k \right\rangle + \eta^2 g_k^\top (P_k P_k^\top)^2 g_k.$$

Taking conditional expectation and using

$$\mathbb{E}[P_k P_k^\top \mid \mathcal{F}_k] = I_d, \qquad \mathbb{E}[(P_k P_k^\top)^2 \mid \mathcal{F}_k] \preceq \omega I_d,$$

we get

$$\mathbb{E}[\|x_{k+1} - x^\star\|^2 \mid \mathcal{F}_k] \leq \|x_k - x^\star\|^2 - 2\eta \left\langle x_k - x^\star, g_k \right\rangle + \eta^2 \omega \|g_k\|^2.$$

By convexity,

$$\langle x_k - x^\star, g_k \rangle \geq f(x_k) - f^\star,$$

and by (69),

$$\|g_k\|^2 \leq 2L(f(x_k) - f^\star).$$

Therefore

$$\mathbb{E}[\|x_{k+1} - x^\star\|^2 \mid \mathcal{F}_k] \leq \|x_k - x^\star\|^2 - \big(2\eta - 2\eta^2 \omega L\big)\big(f(x_k) - f^\star\big).$$

Taking expectation and using the tower property yields

$$\mathbb{E}[\|x_{k+1} - x^\star\|^2] \leq \mathbb{E}[\|x_k - x^\star\|^2] - \big(2\eta - 2\eta^2 \omega L\big)\mathbb{E}[f(x_k) - f^\star].$$

Summing for $k = 0, \ldots, N-1$, we obtain

$$\big(2\eta - 2\eta^2 \omega L\big) \sum_{k=0}^{N-1} \mathbb{E}[f(x_k) - f^\star] \leq \|x_0 - x^\star\|^2.$$

Since $\mathbb{E}[f(x_k)]$ is nonincreasing,

$$N\, \mathbb{E}[f(x_N) - f^\star] \leq \sum_{k=0}^{N-1} \mathbb{E}[f(x_k) - f^\star] \leq \frac{\|x_0 - x^\star\|^2}{2\eta - 2\eta^2 \omega L}.$$

Using $\eta = 1/(2\omega L)$, we have

$$2\eta - 2\eta^2 \omega L = \frac{1}{2\omega L}.$$

Therefore

$$\mathbb{E}[f(x_N) - f^\star] \leq \frac{2\omega L \, \|x_0 - x^\star\|^2}{N},$$

which proves (67). The oracle bound (68) follows immediately because each iteration uses $r$ oracle calls. □

**Proposition H.2** (Strongly convex rate for the basic randomized-subspace gradient method)**.** *Suppose Assumptions 2.1, 2.2 and 4.1 hold, and let $\{x_k\}_{k \geq 0}$ be generated by* (66) *with the constant step-size*

$$\eta = \frac{1}{\ell L}.$$

*Then, for every $k \geq 0$,*

$$\mathbb{E}[f(x_{k+1}) - f^\star] \leq \left(1 - \frac{\mu}{\ell L}\right) \mathbb{E}[f(x_k) - f^\star]. \tag{70}$$

*Hence, for every $N \geq 0$,*

$$\mathbb{E}[f(x_N) - f^\star] \leq \left(1 - \frac{\mu}{\ell L}\right)^N \big(f(x_0) - f^\star\big) \leq \exp\left(-\frac{\mu N}{\ell L}\right) \big(f(x_0) - f^\star\big). \tag{71}$$

*Consequently, defining $\Delta_0 \coloneqq f(x_0) - f^\star$, it suffices to take*

$$N \geq \frac{\ell L}{\mu} \log \frac{\Delta_0}{\epsilon}$$

*to guarantee $\mathbb{E}[f(x_N) - f^\star] \leq \epsilon$. Since each iteration uses $r$ oracle calls, the oracle complexity is*

$$\#\text{Oracle} = \mathcal{O}\left( r\ell \frac{L}{\mu} \log \frac{1}{\epsilon} \right). \tag{72}$$

*Proof.* Let

$$g_k \coloneqq \nabla f(x_k).$$

Applying Assumption 2.1 with $x = x_k$ and $y = x_{k+1} = x_k - \eta P_k P_k^\top g_k$, we obtain

$$f(x_{k+1}) \leq f(x_k) - \eta\, g_k^\top P_k P_k^\top g_k + \frac{\eta^2}{2}\, g_k^\top P_k P_k^\top \mathbb{L} P_k P_k^\top g_k.$$

Taking conditional expectation with respect to $\mathcal{F}_k$, and using

$$\mathbb{E}[P_k P_k^\top \mid \mathcal{F}_k] = I_d, \qquad \mathbb{E}[P_k P_k^\top \mathbb{L} P_k P_k^\top \mid \mathcal{F}_k] \preceq \ell L I_d,$$

we obtain

$$\mathbb{E}[f(x_{k+1}) \mid \mathcal{F}_k] \leq f(x_k) - \eta \left\|g_k\right\|^2 + \frac{\eta^2 \ell L}{2} \left\|g_k\right\|^2 = f(x_k) - \eta\Big(1 - \frac{\eta \ell L}{2}\Big) \left\|g_k\right\|^2.$$

With $\eta = 1/(\ell L)$, this yields

$$\mathbb{E}[f(x_{k+1}) \mid \mathcal{F}_k] \leq f(x_k) - \frac{1}{2\ell L} \left\|g_k\right\|^2.$$

Since $f$ is differentiable and $\mu$-strongly convex,

$$\left\|\nabla f(x)\right\|^2 \geq 2\mu\big(f(x) - f^\star\big) \qquad \text{for all } x \in \mathbb{R}^d,$$

we obtain

$$\mathbb{E}[f(x_{k+1}) - f^\star \mid \mathcal{F}_k] \leq f(x_k) - f^\star - \frac{\mu}{\ell L}\big(f(x_k) - f^\star\big) = \Big(1 - \frac{\mu}{\ell L}\Big)\big(f(x_k) - f^\star\big).$$

Taking expectation proves (70), and (71) follows by iteration. The oracle bound (72) is immediate because each iteration uses $r$ oracle calls. □

# I HIGH-PROBABILITY AND ALMOST-SURE GUARANTEES

We now record two consequences of the supermartingale Lyapunov bounds: one in the convex case and one in the strongly convex case.

In the convex case, the maximal inequality yields a uniform high-probability bound, and a dyadic Borel–Cantelli argument yields an almost sure eventual rate. In the strongly convex case, after normalization by the linear contraction factor, the Lyapunov process is again a nonnegative supermartingale, and the almost sure argument becomes simpler: no dyadic reduction is needed.

## I.1 CONVEX CASE

Recall the convex Lyapunov process

$$\Phi_k^{\mathrm{C}} \;\coloneqq\; A_k\big(f(x_k) - f^\star\big) + \frac{1}{2}\|z_k - x^\star\|^2.$$

**Proposition I.1** (Convex case: high-probability and almost sure convergence)**.** *Suppose Assumptions 2.1, 2.2 and 3.1 hold, and let $\{x_k, y_k, z_k\}_{k\geq 0}$ be generated by Algorithm 1. Then, by* (36)*, the process $(\Phi_k^{\mathrm{C}})_{k\geq 0}$ is a nonnegative supermartingale:*

$$\mathbb{E}[\Phi_{k+1}^{\mathrm{C}} \mid \mathcal{F}_k] \leq \Phi_k^{\mathrm{C}} \qquad \textit{for all } k \geq 0.$$

*Consequently, the following hold.*

*(i) For every $\eta \in (0,1)$, with probability at least $1-\eta$,*

$$A_k\big(f(x_k)-f^\star\big) \le \frac{\Phi_0^{\mathrm{C}}}{\eta} \qquad \text{for all } k \ge 0. \tag{73}$$

*Consequently, since $A_k > 0$ for all $k \ge 1$,*

$$f(x_k)-f^\star \le \frac{\Phi_0^{\mathrm{C}}}{\eta\, A_k} \qquad \text{for all } k \ge 1. \tag{74}$$

*(ii) Fix any $\varepsilon > 0$. Then, almost surely, there exists a finite random integer $K_\varepsilon$ such that for all integers $k \ge K_\varepsilon$,*

$$A_k\big(f(x_k)-f^\star\big) \le \Phi_0^{\mathrm{C}} \left\lceil \log_2 k \right\rceil \Big(\log\lceil \log_2 k \rceil\Big)^{1+\varepsilon}. \tag{75}$$

*Consequently, for all $k \ge K_\varepsilon$,*

$$f(x_k)-f^\star \le \frac{\Phi_0^{\mathrm{C}} \left\lceil \log_2 k \right\rceil \big(\log\lceil \log_2 k \rceil\big)^{1+\varepsilon}}{A_k}. \tag{76}$$

*Moreover, for the sequence $(A_k)$ generated by Algorithm 1, one has*

$$A_k \ge \frac{k^2}{4L\ell\omega} \qquad \text{for all } k \ge 0, \tag{77}$$

*and therefore, for all $k \ge K_\varepsilon$,*

$$f(x_k)-f^\star \le 4L\ell\omega\, \Phi_0^{\mathrm{C}}\, \frac{\left\lceil \log_2 k \right\rceil \big(\log\lceil \log_2 k \rceil\big)^{1+\varepsilon}}{k^2}. \tag{78}$$

*Proof.* We first prove part (i). Since $(\Phi_k^{\mathrm{C}})$ is a nonnegative supermartingale, Ville's maximal inequality gives, for every $b > 0$,

$$\mathbb{P}\left(\sup_{j\ge 0} \Phi_j^{\mathrm{C}} \ge b\right) \le \frac{\Phi_0^{\mathrm{C}}}{b}.$$

Taking $b = \Phi_0^{\mathrm{C}}/\eta$, we obtain

$$\mathbb{P}\left(\sup_{j\ge 0} \Phi_j^{\mathrm{C}} \le \frac{\Phi_0^{\mathrm{C}}}{\eta}\right) \ge 1-\eta.$$

Because

$$A_k\big(f(x_k)-f^\star\big) \le \Phi_k^{\mathrm{C}} \qquad \text{a.s. for every } k \ge 0,$$

this proves (73). Since $A_k > 0$ for all $k \ge 1$, (74) follows by dividing by $A_k$.

We now prove part (ii). Fix $\varepsilon > 0$, and for every integer $m \ge 2$, define

$$N_m := 2^m, \qquad b_m := \Phi_0^{\mathrm{C}}\, m(\log m)^{1+\varepsilon},$$

and the bad events

$$E_m := \left\{ \max_{0\le j\le 2^m} \Phi_j^{\mathrm{C}} > b_m \right\}.$$

Applying Ville's maximal inequality on the finite horizon $\{0,\dots,2^m\}$ gives

$$\mathbb{P}(E_m) \le \frac{\Phi_0^{\mathrm{C}}}{b_m} = \frac{1}{m(\log m)^{1+\varepsilon}}.$$

Since

$$\sum_{m=2}^{\infty} \frac{1}{m(\log m)^{1+\varepsilon}} < \infty,$$

the Borel–Cantelli lemma implies

$$\mathbb{P}(E_m \text{ i.o.}) = 0.$$

Therefore, almost surely, there exists a finite random integer $m_\varepsilon \geq 2$ such that for all $m \geq m_\varepsilon$,

$$\max_{0 \leq j \leq 2^m} \Phi_j^{\mathrm{C}} \leq \Phi_0^{\mathrm{C}}\, m(\log m)^{1+\varepsilon}. \tag{79}$$

Now let

$$K_\varepsilon \coloneqq 2^{m_\varepsilon}.$$

Fix any integer $k \geq K_\varepsilon$, and set

$$m \coloneqq \lceil \log_2 k \rceil .$$

Then $m \geq m_\varepsilon$ and $k \leq 2^m$. Hence, by (79),

$$A_k\big(f(x_k) - f^\star\big) \leq \Phi_k^{\mathrm{C}} \leq \max_{0 \leq j \leq 2^m} \Phi_j^{\mathrm{C}} \leq \Phi_0^{\mathrm{C}}\, m(\log m)^{1+\varepsilon}.$$

Since $m = \lceil \log_2 k \rceil$, this proves (75). Dividing by $A_k$ gives (76).

The lower bound (77) follows directly from (38). Finally, substituting (77) into (76) yields (78). □

**Comparison with the expectation rate** The expectation-level rate from the convex supermartingale argument is

$$\mathbb{E}[f(x_k) - f^\star] \leq \frac{\Phi_0^{\mathrm{C}}}{A_k}.$$

By contrast, the uniform high-probability bound in Proposition I.1(i) differs only by the multiplicative factor $1/\eta$:

$$f(x_k) - f^\star \leq \frac{\Phi_0^{\mathrm{C}}}{\eta\, A_k} \qquad \text{for all } k \geq 1$$

with probability at least $1 - \eta$. The almost sure eventual bound in Proposition I.1(ii) differs from the expectation rate by the explicit logarithmic factor

$$\big\lceil \log_2 k \big\rceil \big(\log \lceil \log_2 k \rceil\big)^{1+\varepsilon}.$$

No monotonicity of $f(x_k)$ is used anywhere in the proof.

### I.2 Strongly convex case

Define the strongly convex Lyapunov process by

$$\Phi_k^{\mathrm{SC}} \coloneqq f(x_k) - f^\star + \frac{\mu}{2}\|z_k - x^\star\|^2.$$

Set

$$\rho \coloneqq 1 - \theta.$$

By Theorem 4.3, $\theta \in (0, 1]$, and hence $\rho = 1 - \theta \in [0, 1)$. Moreover, by (65),

$$\mathbb{E}[\Phi_{k+1}^{\mathrm{SC}} \mid \mathcal{F}_k] \leq \rho\, \Phi_k^{\mathrm{SC}} \qquad \text{for all } k \geq 0. \tag{80}$$

Also, by the definition of $\Phi_k^{\mathrm{SC}}$,

$$f(x_k) - f^\star \leq \Phi_k^{\mathrm{SC}} \qquad \text{for all } k \geq 0. \tag{81}$$

**Proposition I.2** (Strongly convex case: high-probability and almost sure convergence)**.** *Suppose Assumptions 2.1, 2.2 and 4.1 hold, and let $\{x_k, y_k, z_k\}_{k \geq 0}$ be generated by Algorithm 2. Let $\rho \coloneqq 1 - \theta$.*

*If $\rho = 0$, then*

$$\Phi_k^{\mathrm{SC}} = 0 \qquad \textit{for all } k \geq 1$$

*almost surely. Consequently, for every $\eta \in (0, 1)$, with probability one,*

$$f(x_0) - f^\star \leq \frac{\Phi_0^{\mathrm{SC}}}{\eta}, \qquad f(x_k) - f^\star = 0 \quad \textit{for all } k \geq 1,$$

*and for every $q \in (0, 1)$,* (83) *holds with $K_q = 1$.*

*If $\rho \in (0, 1)$, then the following hold.*

*(i) For every $\eta \in (0,1)$, with probability at least $1-\eta$,*

$$f(x_k) - f^\star \le \frac{\Phi_0^{\mathrm{SC}}}{\eta}\,\rho^k \qquad \text{for all } k \ge 0. \tag{82}$$

*(ii) For every number $q \in (\rho, 1)$, almost surely there exists a finite random integer $K_q$ such that for all $k \ge K_q$,*

$$f(x_k) - f^\star \le \Phi_0^{\mathrm{SC}}\, q^k. \tag{83}$$

*Proof.* First consider the case $\rho = 0$. Since $\Phi_{k+1}^{\mathrm{SC}} \ge 0$ and (80) gives

$$\mathbb{E}[\Phi_{k+1}^{\mathrm{SC}} \mid \mathcal{F}_k] \le 0,$$

we have $\Phi_{k+1}^{\mathrm{SC}} = 0$ almost surely for every $k \ge 0$. Therefore $\Phi_k^{\mathrm{SC}} = 0$ almost surely for all $k \ge 1$. The claims in the case $\rho = 0$ follow from (81).

It remains to consider the case $\rho \in (0,1)$. Define the normalized process

$$M_k := \rho^{-k}\Phi_k^{\mathrm{SC}}.$$

Then

$$\mathbb{E}[M_{k+1} \mid \mathcal{F}_k] = \rho^{-(k+1)}\mathbb{E}[\Phi_{k+1}^{\mathrm{SC}} \mid \mathcal{F}_k] \le \rho^{-k}\Phi_k^{\mathrm{SC}} = M_k.$$

Hence $(M_k)_{k\ge 0}$ is a nonnegative supermartingale.

For part (i), Ville's maximal inequality gives

$$\mathbb{P}\left(\sup_{j\ge 0} M_j \ge \frac{\Phi_0^{\mathrm{SC}}}{\eta}\right) \le \eta.$$

Therefore, with probability at least $1-\eta$,

$$M_k \le \frac{\Phi_0^{\mathrm{SC}}}{\eta} \qquad \text{for all } k \ge 0,$$

that is,

$$\Phi_k^{\mathrm{SC}} \le \frac{\Phi_0^{\mathrm{SC}}}{\eta}\rho^k \qquad \text{for all } k \ge 0.$$

Using (81), we obtain (82).

For part (ii), fix any $q \in (\rho, 1)$, and define

$$\eta_k := \left(\frac{\rho}{q}\right)^k.$$

Since $\rho/q < 1$, the series $\sum_{k=0}^{\infty}\eta_k$ converges. Moreover, because $\mathbb{E}[M_k] \le M_0 = \Phi_0^{\mathrm{SC}}$, Markov's inequality gives

$$\mathbb{P}\left(M_k > \frac{\Phi_0^{\mathrm{SC}}}{\eta_k}\right) \le \eta_k.$$

Equivalently,

$$\mathbb{P}\big(\Phi_k^{\mathrm{SC}} > \Phi_0^{\mathrm{SC}}\, q^k\big) \le \left(\frac{\rho}{q}\right)^k.$$

Since

$$\sum_{k=0}^{\infty}\left(\frac{\rho}{q}\right)^k < \infty,$$

the Borel–Cantelli lemma implies that, almost surely, only finitely many of these events occur. Hence, almost surely, there exists a finite random integer $K_q$ such that for all $k \ge K_q$,

$$\Phi_k^{\mathrm{SC}} \le \Phi_0^{\mathrm{SC}}\, q^k.$$

Using (81), we obtain (83). □

**Comparison with the expectation rate** The expectation-level strongly convex rate is

$$\mathbb{E}[f(x_k) - f^\star] \le \Phi_0^{\mathrm{SC}} \rho^k.$$

The uniform high-probability bound in Proposition I.2(i) differs only by the multiplicative factor $1/\eta$:

$$f(x_k) - f^\star \le \frac{\Phi_0^{\mathrm{SC}}}{\eta} \rho^k \qquad \text{for all } k \ge 0$$

with probability at least $1 - \eta$. The almost sure eventual bound in Proposition I.2(ii) replaces the contraction factor $\rho$ by any prescribed factor $q \in (\rho, 1)$. Thus the almost sure linear rate can be made arbitrarily close to the expectation linear rate, but it is not exactly identical.

# J WHY THE CLASSICAL TWO-SEQUENCE TEMPLATE IS NOT DIRECTLY PORTABLE

In this appendix, we formalize a limitation of a direct transplantation of the classical two-sequence Nesterov template to randomized-subspace updates. These results should be interpreted as proof-template obstructions rather than impossibility results for all conceivable two-sequence accelerated schemes. Rather, they show that, under the sketch structure satisfied by the Haar and coordinate sketches, the classical proof mechanism is not directly compatible with the naive sketched analogue outside a near-full-dimensional regime.

**Proposition J.1.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and $L$-smooth, and let $x^\star$ be a minimizer of $f$. Let $\{P_k\}_{k\ge 0}$ be an i.i.d. sketch sequence satisfying*

$$\mathbb{E}[P_k P_k^\top] = I_d, \qquad P_k^\top P_k = \frac{d}{r} I_r$$

*for some $1 \le r \le d$. Consider the direct sketched two-sequence recursion*

$$x_{k+1} = y_k - \eta P_k P_k^\top \nabla f(y_k), \qquad y_{k+1} = x_{k+1} + \beta_k (x_{k+1} - x_k),$$

*where $\eta > 0$ is constant and $\{\beta_k\}_{k\ge 0}$ is arbitrary.*

*For each $k \ge 0$, let*

$$\mathcal{F}_k := \sigma(P_0, \dots, P_{k-1}), \qquad g_k := \nabla f(y_k), \qquad M_k := P_k P_k^\top.$$

*Suppose, as in the classical proof template, that there exist sequences $\{\lambda_k\}_{k\ge 0} \subset (0, 1]$ and $\{z_k\}_{k\ge 0} \subset \mathbb{R}^d$ such that, for every $k \ge 0$,*

$$y_k = (1 - \lambda_k) x_k + \lambda_k z_k, \qquad z_{k+1} = z_k - \frac{\eta}{\lambda_k} M_k g_k. \tag{84}$$

*Then, for every $k \ge 0$,*

$$\begin{aligned} &\mathbb{E}\left[ \frac{1}{\lambda_k^2}\big(f(x_{k+1}) - f^\star\big) + \frac{1}{2\eta}\|z_{k+1} - x^\star\|^2 \;\middle|\; \mathcal{F}_k \right] \\ &\le \frac{1 - \lambda_k}{\lambda_k^2}\big(f(x_k) - f^\star\big) + \frac{1}{2\eta}\|z_k - x^\star\|^2 + \frac{\eta}{2\lambda_k^2}\Big(\frac{d}{r}(1 + L\eta) - 2\Big)\|g_k\|^2. \end{aligned} \tag{85}$$

*In particular, if $r \le d/2$, then the coefficient*

$$\frac{\eta}{2\lambda_k^2}\Big(\frac{d}{r}(1 + L\eta) - 2\Big)$$

*is strictly positive for every $\eta > 0$. Therefore, the estimate* (85) *retains a strictly positive residual coefficient, and hence this affine hidden-sequence template does not recover the classical two-term decrease by the standard argument.*

*Proof.* Fix $k \geq 0$. Since $P_k^\top P_k = (d/r) I_r$,

$$M_k^2 = P_k (P_k^\top P_k) P_k^\top = \frac{d}{r} M_k.$$

Moreover, $P_k$ is independent of $\mathcal{F}_k$, and $\mathbb{E}[M_k] = I_d$, so

$$\mathbb{E}[M_k \mid \mathcal{F}_k] = I_d, \qquad \mathbb{E}[M_k^2 \mid \mathcal{F}_k] = \frac{d}{r} I_d.$$

By $L$-smoothness of $f$,

$$f(v) \leq f(u) + \langle \nabla f(u), v - u \rangle + \frac{L}{2} \|v - u\|^2 \qquad \text{for all } u, v \in \mathbb{R}^d.$$

Applying this with $u = y_k$ and $v = x_{k+1} = y_k - \eta M_k g_k$, we obtain

$$f(x_{k+1}) \leq f(y_k) - \eta g_k^\top M_k g_k + \frac{L\eta^2}{2} g_k^\top M_k^2 g_k.$$

Taking conditional expectation and using $\mathbb{E}[M_k \mid \mathcal{F}_k] = I_d$ and $\mathbb{E}[M_k^2 \mid \mathcal{F}_k] = \frac{d}{r} I_d$ gives

$$\mathbb{E}[f(x_{k+1}) \mid \mathcal{F}_k] \leq f(y_k) - \eta \|g_k\|^2 + \frac{L\eta^2}{2} \frac{d}{r} \|g_k\|^2. \tag{86}$$

We now estimate the $z$-term. From (84),

$$z_{k+1} - x^\star = (z_k - x^\star) - \frac{\eta}{\lambda_k} M_k g_k,$$

and hence

$$\|z_{k+1} - x^\star\|^2 = \|z_k - x^\star\|^2 - \frac{2\eta}{\lambda_k} \langle z_k - x^\star, M_k g_k \rangle + \frac{\eta^2}{\lambda_k^2} \|M_k g_k\|^2.$$

Multiplying by $1/(2\eta)$, taking conditional expectation, and using again $\mathbb{E}[M_k \mid \mathcal{F}_k] = I_d$ and $\mathbb{E}[M_k^2 \mid \mathcal{F}_k] = \frac{d}{r} I_d$, we get

$$\begin{aligned} &\mathbb{E}\left[ \frac{1}{2\eta} \|z_{k+1} - x^\star\|^2 \;\middle|\; \mathcal{F}_k \right] \\ &= \frac{1}{2\eta} \|z_k - x^\star\|^2 - \frac{1}{\lambda_k} \langle z_k - x^\star, g_k \rangle + \frac{\eta}{2\lambda_k^2} \frac{d}{r} \|g_k\|^2. \end{aligned} \tag{87}$$

Subtracting $f^\star$ from (86), multiplying by $1/\lambda_k^2$, and adding (87), we obtain

$$\begin{aligned} &\mathbb{E}\left[ \frac{1}{\lambda_k^2} \big(f(x_{k+1}) - f^\star\big) + \frac{1}{2\eta} \|z_{k+1} - x^\star\|^2 \;\middle|\; \mathcal{F}_k \right] \\ &\leq \frac{1}{\lambda_k^2} \big(f(y_k) - f^\star\big) + \frac{1}{2\eta} \|z_k - x^\star\|^2 - \frac{1}{\lambda_k} \langle z_k - x^\star, g_k \rangle \\ &\quad + \frac{\eta}{2\lambda_k^2} \Big( \frac{d}{r}(1 + L\eta) - 2 \Big) \|g_k\|^2. \end{aligned} \tag{88}$$

It remains to control the mixed term. Since

$$y_k = (1 - \lambda_k) x_k + \lambda_k z_k,$$

we have

$$\frac{1 - \lambda_k}{\lambda_k^2} (x_k - y_k) + \frac{1}{\lambda_k} (x^\star - y_k) = \frac{1}{\lambda_k} (x^\star - z_k).$$

Taking inner products with $g_k$ gives

$$-\frac{1}{\lambda_k} \langle z_k - x^\star, g_k \rangle = \frac{1 - \lambda_k}{\lambda_k^2} \langle x_k - y_k, g_k \rangle + \frac{1}{\lambda_k} \langle x^\star - y_k, g_k \rangle . \tag{89}$$

By convexity of $f$,

$$f(x_k) \geq f(y_k) + \langle g_k, x_k - y_k \rangle, \qquad f^\star = f(x^\star) \geq f(y_k) + \langle g_k, x^\star - y_k \rangle,$$

so

$$\langle x_k - y_k, g_k \rangle \leq f(x_k) - f(y_k), \qquad \langle x^\star - y_k, g_k \rangle \leq f^\star - f(y_k).$$

Substituting these bounds into (89) yields

$$-\frac{1}{\lambda_k} \langle z_k - x^\star, g_k \rangle \leq \frac{1-\lambda_k}{\lambda_k^2}\big(f(x_k) - f(y_k)\big) + \frac{1}{\lambda_k}\big(f^\star - f(y_k)\big). \tag{90}$$

Substituting (90) into (88), we get

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{\lambda_k^2}\big(f(x_{k+1}) - f^\star\big) + \frac{1}{2\eta}\|z_{k+1} - x^\star\|^2 \;\middle|\; \mathcal{F}_k\right] \\
&\leq \frac{1}{\lambda_k^2}\big(f(y_k) - f^\star\big) + \frac{1-\lambda_k}{\lambda_k^2}\big(f(x_k) - f(y_k)\big) + \frac{1}{\lambda_k}\big(f^\star - f(y_k)\big) + \frac{1}{2\eta}\|z_k - x^\star\|^2 \\
&\quad + \frac{\eta}{2\lambda_k^2}\Big(\frac{d}{r}(1 + L\eta) - 2\Big)\|g_k\|^2.
\end{aligned}$$

The $f(y_k)$-terms cancel, since

$$\frac{1}{\lambda_k^2} - \frac{1-\lambda_k}{\lambda_k^2} - \frac{1}{\lambda_k} = 0.$$

Therefore,

$$\frac{1}{\lambda_k^2}\big(f(y_k) - f^\star\big) + \frac{1-\lambda_k}{\lambda_k^2}\big(f(x_k) - f(y_k)\big) + \frac{1}{\lambda_k}\big(f^\star - f(y_k)\big) = \frac{1-\lambda_k}{\lambda_k^2}\big(f(x_k) - f^\star\big),$$

and (85) follows.

Finally, if $r \leq d/2$, then $d/r \geq 2$, and thus

$$\frac{d}{r}(1 + L\eta) - 2 \geq 2(1 + L\eta) - 2 = 2L\eta > 0 \qquad \text{for every } \eta > 0.$$

Hence the coefficient of $\|g_k\|^2$ in (85) is strictly positive. Therefore, (85) retains a strictly positive residual coefficient, and thus the estimate does not collapse to the classical two-term decrease by the standard argument. □

**Proposition J.2.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be $\mu$-strongly convex and $L$-smooth, with $\mu > 0$. Let $\{P_k\}_{k\geq 0}$ be an i.i.d. sketch sequence satisfying*

$$\mathbb{E}[P_k P_k^\top] = I_d, \qquad P_k^\top P_k = \frac{d}{r} I_r$$

*for some $1 \leq r \leq d$. Consider the direct sketched two-sequence recursion*

$$x_{k+1} = y_k - \eta P_k P_k^\top \nabla f(y_k), \qquad y_{k+1} = x_{k+1} + \beta(x_{k+1} - x_k), \tag{91}$$

*where $\eta > 0$ and $\beta \in [0, 1)$ are constants. Define*

$$\theta \coloneqq \frac{1-\beta}{1+\beta} \in (0, 1], \qquad \gamma \coloneqq \frac{\eta}{\theta}, \qquad z_k \coloneqq \frac{(1+\theta)y_k - x_k}{\theta}, \qquad u_k \coloneqq (1-\theta)z_k + \theta y_k,$$

*and*

$$\Phi_k \coloneqq f(x_k) - f^\star + \frac{\mu}{2}\|z_k - x^\star\|^2.$$

*Then*

$$y_k = \frac{1}{1+\theta} x_k + \frac{\theta}{1+\theta} z_k, \qquad z_{k+1} = u_k - \gamma P_k P_k^\top \nabla f(y_k). \tag{92}$$

*Writing $g_k \coloneqq \nabla f(y_k)$, $M_k \coloneqq P_k P_k^\top$, and $\mathcal{F}_k \coloneqq \sigma(P_0, \dots, P_{k-1})$, one has*

$$\begin{aligned}
&\mathbb{E}[\Phi_{k+1}|\mathcal{F}_k] - (1-\theta)\Phi_k \\
&\le \theta\big(f(y_k) - f^\star\big) + (1-\theta)\big(f(y_k) - f(x_k)\big) - \mu\gamma\langle g_k, y_k - x^\star\rangle + \mu\gamma\frac{1-\theta}{\theta}\langle g_k, x_k - y_k\rangle \\
&\quad + \frac{\mu\theta}{2}\|y_k - x^\star\|^2 - \frac{\mu(1-\theta)}{2\theta}\|x_k - y_k\|^2 + \left(-\eta + \frac{d}{2r}(L\eta^2 + \mu\gamma^2)\right)\|g_k\|^2. \qquad (93)
\end{aligned}$$

*If one wants the right-hand side of* (93) *to collapse into the same two brackets as in the standard strongly-convex proof, namely*

$$\begin{aligned}
&\mathbb{E}[\Phi_{k+1}|\mathcal{F}_k] - (1-\theta)\Phi_k \\
&\le \theta\Big(f(y_k) - f^\star - \langle g_k, y_k - x^\star\rangle + \frac{\mu}{2}\|y_k - x^\star\|^2\Big) \\
&\quad + (1-\theta)\Big(f(y_k) - f(x_k) + \langle g_k, x_k - y_k\rangle\Big) \\
&\quad - \frac{\mu(1-\theta)}{2\theta}\|x_k - y_k\|^2 \qquad (94)
\end{aligned}$$

*then matching the coefficients of the inner-product terms yields*

$$\mu\gamma = \theta, \qquad \textit{that is,} \qquad \gamma = \frac{\theta}{\mu}, \qquad \textit{equivalently} \qquad \eta = \frac{\theta^2}{\mu}. \qquad (95)$$

*Under* (95)*,* (93) *becomes*

$$\begin{aligned}
&\mathbb{E}[\Phi_{k+1}|\mathcal{F}_k] - (1-\theta)\Phi_k \\
&\le \theta\Big(f(y_k) - f^\star - \langle g_k, y_k - x^\star\rangle + \frac{\mu}{2}\|y_k - x^\star\|^2\Big) \\
&\quad + (1-\theta)\Big(f(y_k) - f(x_k) + \langle g_k, x_k - y_k\rangle\Big) \\
&\quad - \frac{\mu(1-\theta)}{2\theta}\|x_k - y_k\|^2 + \frac{\theta^2}{\mu}\left[-1 + \frac{d}{2r}\left(1 + \frac{L\theta^2}{\mu}\right)\right]\|g_k\|^2. \qquad (96)
\end{aligned}$$

*Therefore, in order for the remaining coefficient of $\|g_k\|^2$ in* (96) *to be nonpositive, it is necessary that*

$$\frac{d}{r}\left(1 + \frac{L\theta^2}{\mu}\right) \le 2. \qquad (97)$$

*Equivalently, since $\eta = \theta^2/\mu$,*

$$\frac{d}{r}(1 + L\eta) \le 2. \qquad (98)$$

*In particular, if $r \le d/2$, then* (97) *cannot hold for any $\beta \in [0, 1)$. Therefore, outside a near-full-dimensional regime, the direct sketched analogue of the classical strongly-convex two-sequence Nesterov template is incompatible with the standard hidden-sequence proof mechanism.*

*Proof.* Fix $k \ge 0$, and let

$$\mathcal{F}_k \coloneqq \sigma(P_0, \dots, P_{k-1}), \qquad M_k \coloneqq P_k P_k^\top, \qquad g_k \coloneqq \nabla f(y_k).$$

We first rewrite (91) in hidden-sequence form. By the definition of $z_k$,

$$z_k = \frac{(1+\theta)y_k - x_k}{\theta},$$

so that

$$y_k = \frac{1}{1+\theta}x_k + \frac{\theta}{1+\theta}z_k.$$

Also,

$$z_{k+1} = \frac{(1+\theta)y_{k+1} - x_{k+1}}{\theta}.$$

Using

$$y_{k+1} = x_{k+1} + \frac{1-\theta}{1+\theta}(x_{k+1} - x_k),$$

we obtain

$$\begin{aligned} z_{k+1} &= \frac{(1+\theta)y_{k+1} - x_{k+1}}{\theta} \\ &= \frac{1}{\theta}x_{k+1} - \frac{1-\theta}{\theta}x_k \\ &= \frac{1}{\theta}(y_k - \eta M_k g_k) - \frac{1-\theta}{\theta}x_k. \end{aligned}$$

On the other hand,

$$(1-\theta)z_k + \theta y_k = \frac{1-\theta}{\theta}\big((1+\theta)y_k - x_k\big) + \theta y_k = \frac{1}{\theta}y_k - \frac{1-\theta}{\theta}x_k.$$

Hence

$$z_{k+1} = (1-\theta)z_k + \theta y_k - \frac{\eta}{\theta}M_k g_k = u_k - \gamma M_k g_k,$$

which proves (92).

We next derive the one-step Lyapunov estimate for a general $\gamma$, and only then identify the value of $\gamma$ for which the same collapse pattern as in the standard proof occurs.

From

$$x_k - y_k = \theta(y_k - z_k),$$

we obtain

$$z_k - x^\star = y_k - x^\star - \frac{1}{\theta}(x_k - y_k).$$

Also, since

$$u_k = (1-\theta)z_k + \theta y_k = y_k - \frac{1-\theta}{\theta}(x_k - y_k),$$

the update $z_{k+1} = u_k - \gamma M_k g_k$ yields

$$z_{k+1} - x^\star = y_k - x^\star - \frac{1-\theta}{\theta}(x_k - y_k) - \gamma M_k g_k.$$

Therefore, a direct expansion gives

$$\begin{aligned} &\|z_{k+1} - x^\star\|^2 - (1-\theta)\|z_k - x^\star\|^2 \\ &= \theta\|y_k - x^\star\|^2 - \frac{1-\theta}{\theta}\|x_k - y_k\|^2 - 2\gamma\langle M_k g_k, y_k - x^\star\rangle + \frac{2\gamma(1-\theta)}{\theta}\langle M_k g_k, x_k - y_k\rangle + \gamma^2 g_k^\top M_k^2 g_k. \end{aligned} \tag{99}$$

Now

$$M_k^2 = P_k(P_k^\top P_k)P_k^\top = \frac{d}{r}M_k.$$

Since $P_k$ is independent of $\mathcal{F}_k$ and $\mathbb{E}[M_k] = I_d$, we have

$$\mathbb{E}[M_k \mid \mathcal{F}_k] = I_d, \qquad \mathbb{E}[M_k^2 \mid \mathcal{F}_k] = \frac{d}{r}I_d.$$

Taking conditional expectation in (99), we obtain

$$\begin{aligned} &\mathbb{E}\Big[\frac{\mu}{2}\|z_{k+1} - x^\star\|^2\Big|\mathcal{F}_k\Big] - \frac{(1-\theta)\mu}{2}\|z_k - x^\star\|^2 \\ &= \frac{\mu\theta}{2}\|y_k - x^\star\|^2 - \mu\gamma\langle g_k, y_k - x^\star\rangle + \mu\gamma\frac{1-\theta}{\theta}\langle g_k, x_k - y_k\rangle - \frac{\mu(1-\theta)}{2\theta}\|x_k - y_k\|^2 + \frac{d}{2r}\mu\gamma^2\|g_k\|^2. \end{aligned} \tag{100}$$

On the function-value side, by $L$-smoothness and

$$x_{k+1} = y_k - \eta M_k g_k,$$

we have

$$f(x_{k+1}) \le f(y_k) - \eta g_k^\top M_k g_k + \frac{L\eta^2}{2} g_k^\top M_k^2 g_k.$$

Taking conditional expectation gives

$$\mathbb{E}[f(x_{k+1}) \mid \mathcal{F}_k] \le f(y_k) - \eta \|g_k\|^2 + \frac{d}{2r} L\eta^2 \|g_k\|^2. \tag{101}$$

We now expand $\mathbb{E}[\Phi_{k+1} \mid \mathcal{F}_k] - (1-\theta)\Phi_k$:

$$\begin{aligned}
&\mathbb{E}[\Phi_{k+1}|\mathcal{F}_k] - (1-\theta)\Phi_k \\
&= \mathbb{E}[f(x_{k+1}) \mid \mathcal{F}_k] - f^\star - (1-\theta)(f(x_k) - f^\star) \\
&\quad + \mathbb{E}\Big[\frac{\mu}{2}\|z_{k+1} - x^\star\|^2 \Big| \mathcal{F}_k\Big] - \frac{(1-\theta)\mu}{2}\|z_k - x^\star\|^2 \\
&= \big(\mathbb{E}[f(x_{k+1}) \mid \mathcal{F}_k] - f(y_k)\big) + \theta(f(y_k) - f^\star) + (1-\theta)(f(y_k) - f(x_k)) \\
&\quad + \mathbb{E}\Big[\frac{\mu}{2}\|z_{k+1} - x^\star\|^2 \Big| \mathcal{F}_k\Big] - \frac{(1-\theta)\mu}{2}\|z_k - x^\star\|^2.
\end{aligned} \tag{102}$$

Substituting (100) and (101) into (102), we obtain

$$\begin{aligned}
&\mathbb{E}[\Phi_{k+1}|\mathcal{F}_k] - (1-\theta)\Phi_k \\
&\le \theta(f(y_k) - f^\star) + (1-\theta)(f(y_k) - f(x_k)) - \mu\gamma\langle g_k, y_k - x^\star\rangle + \mu\gamma\frac{1-\theta}{\theta}\langle g_k, x_k - y_k\rangle \\
&\quad + \frac{\mu\theta}{2}\|y_k - x^\star\|^2 - \frac{\mu(1-\theta)}{2\theta}\|x_k - y_k\|^2 + \left(-\eta + \frac{d}{2r}(L\eta^2 + \mu\gamma^2)\right)\|g_k\|^2,
\end{aligned}$$

which is (93).

We now derive the value of $\gamma$ for which the same collapse pattern as in the standard proof occurs. In (93), the inner-product terms are

$$-\mu\gamma\langle g_k, y_k - x^\star\rangle + \mu\gamma\frac{1-\theta}{\theta}\langle g_k, x_k - y_k\rangle.$$

In order to rewrite the inner-product terms in (93) with the same coefficients as in (94), one must impose

$$\mu\gamma = \theta, \qquad \mu\gamma\frac{1-\theta}{\theta} = 1-\theta.$$

These two identities are equivalent, and therefore

$$\gamma = \frac{\theta}{\mu}.$$

Since $\gamma = \eta/\theta$, this is equivalent to

$$\eta = \frac{\theta^2}{\mu},$$

which proves (95).

Substituting $\gamma = \theta/\mu$ and $\eta = \theta^2/\mu$ into (93), we obtain

$$\begin{aligned}
&\mathbb{E}[\Phi_{k+1}|\mathcal{F}_k] - (1-\theta)\Phi_k \\
&\le \theta\Big(f(y_k) - f^\star - \langle g_k, y_k - x^\star\rangle + \frac{\mu}{2}\|y_k - x^\star\|^2\Big) \\
&\quad + (1-\theta)\Big(f(y_k) - f(x_k) + \langle g_k, x_k - y_k\rangle\Big) \\
&\quad - \frac{\mu(1-\theta)}{2\theta}\|x_k - y_k\|^2 + \left(-\eta + \frac{d}{2r}(L\eta^2 + \mu\gamma^2)\right)\|g_k\|^2.
\end{aligned}$$

Since

$$-\eta + \frac{d}{2r}(L\eta^2 + \mu\gamma^2) = -\frac{\theta^2}{\mu} + \frac{d}{2r}\left(L\frac{\theta^4}{\mu^2} + \frac{\theta^2}{\mu}\right) = \frac{\theta^2}{\mu}\left[-1 + \frac{d}{2r}\left(1 + \frac{L\theta^2}{\mu}\right)\right],$$

this yields (96).

Finally, by $\mu$-strong convexity,

$$f(y_k) - f^\star - \langle g_k, y_k - x^\star \rangle + \frac{\mu}{2}\|y_k - x^\star\|^2 \leq 0,$$

and by convexity,

$$f(y_k) - f(x_k) + \langle g_k, x_k - y_k \rangle \leq 0.$$

Thus, in order for the remaining coefficient of $\|g_k\|^2$ in (96) to be nonpositive, it is necessary that

$$\frac{d}{r}\left(1 + \frac{L\theta^2}{\mu}\right) \leq 2,$$

which is (97). The equivalent form (98) follows from $\eta = \theta^2/\mu$.

Finally, if $r \leq d/2$, then $d/r \geq 2$, and hence for every $\theta \in (0, 1]$,

$$\frac{d}{r}\left(1 + \frac{L\theta^2}{\mu}\right) \geq 2\left(1 + \frac{L\theta^2}{\mu}\right) > 2.$$

Therefore (97) cannot hold for any $\theta \in (0, 1]$, and hence for any $\beta \in [0, 1)$. This completes the proof.

□