CONVERGENCE ANALYSIS OF RANDOMIZED SUBSPACE NORMALIZED SGD UNDER HEAVY-TAILED NOISE

Anonymous authors
Paper under double-blind review

ABSTRACT

Randomized-subspace methods use low-dimensional projected stochastic gradients, offering potential savings in directional-derivative queries and communication volume. However, theoretical guarantees for such methods under heavy-tailed noise remain limited. We propose randomized subspace normalized SGD (RS-NSGD), which normalizes using only the projected stochastic gradient and thereby preserves the low-dimensional update structure. Because the random subspace also enters the normalization denominator, existing analyses of randomized subspace SGD do not directly apply. We overcome this difficulty by exploiting distributional properties specific to Haar-random projections. Under a bounded p-th-moment assumption commonly used in analyses of heavy-tailed stochastic-gradient noise, with $p \in (1, 2]$, we establish convergence guarantees both in expectation and with high probability. Our oracle-complexity bounds characterize regimes in which RS-NSGD improves over full-dimensional normalized SGD.

1 INTRODUCTION

We consider the following nonconvex stochastic optimization problem:

$$\min_{x \in \mathbb{R}^d} F(x), \qquad F(x) := \mathbb{E}_{\xi \sim \mathcal{D}}[f(x, \xi)], \tag{1}$$

where $\mathcal{D}$ denotes an underlying data distribution. Modern machine-learning problems are often high-dimensional, motivating nonconvex optimization algorithms that operate in low-dimensional subspaces.

One possible way to reduce the per-iteration dimension in (1) is to work in random low-dimensional subspaces. Let g_k denote a stochastic gradient estimator for $\nabla F(x_k)$. Given a random matrix $P_k \in \mathbb{R}^{d \times r}$ with $r \ll d$ and a stepsize $\bar{\eta} > 0$, randomized subspace SGD (RS-SGD) uses the low-dimensional projected stochastic gradient $P_k^\top g_k$ and updates

$$x_{k+1} = x_k - \bar{\eta} P_k P_k^\top g_k.$$

The underlying subspace-descent idea was developed for deterministic optimization by Kozak et al. (2021). In stochastic optimization, Flynn et al. (2024) analyze RS-SGD under the bounded-variance assumption. To handle large-scale problems, the optimization community has long relied on randomized coordinate descent methods (Nesterov, 2012; Richtárik & Takáč, 2014; Wright, 2015). These methods update the solution by randomly selecting a single coordinate (or, more generally, a small set of r coordinates) at each iteration. The random-subspace framework above provides a broader view: it includes randomized (block) coordinate descent as a special case (Kozak et al., 2021). Moreover, Kozak et al. (2021) focus on Haar-random subspaces in their analysis, reflecting the favorable concentration properties of Haar directions compared with coordinate-type sampling.

In machine learning, such randomized subspace methods can be useful in at least two settings. First, they are relevant when directional derivatives can be evaluated directly without forming the full gradient. For example, forward-mode automatic differentiation (AD) can compute directional derivatives without the storage of intermediate quantities typically required by reverse-mode AD (Kozak et al., 2021). Under a directional-derivative oracle model, computing the full gradient requires d directional-derivative queries, whereas computing $P_k^\top \nabla f(x, \xi)$ requires only r, along the

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 convergence guarantees for stochastic gradient methods with randomized subspace updates (RS) and direction normalization (Norm). Here, "with high probability" refers to single-run high-probability guarantees. Entries in the last two columns list the required noise assumptions and where the guarantees are established; p-BCM is defined in Assumption 2.4, and BV and sub-Gaussian assumptions are recalled in Appendix C.1.

Method	RS	Norm	In expectation	With high probability
SGD			BV; (Ghadimi & Lan, 2013)	sub-Gaussian; (Liu et al., 2023)
RS-SGD	✓		BV; (Flynn et al., 2024)	sub-Gaussian; ours (Thm. C.4)
NSGD		✓	p-BCM; (Hübler et al., 2025)	p-BCM; (Hübler et al., 2025)
Ours:**RS-NSGD**	✓	✓	p-BCM; ours (Thm. 4.4)	p-BCM; ours (Thm. 4.5)

columns of P_k. Second, in distributed optimization, randomized subspace methods can reduce communication volume by transmitting the r-dimensional projected gradient $P_k^\top g_k$ instead of the d-dimensional gradient g_k, reducing communication from d to r scalars. In a related but distinct line of work, low-dimensional update ideas have also been explored for memory- and communication-efficient large-scale training, but their primary focus and algorithmic setting differ from ours; see Appendix A.5. We refer the reader to Appendix A for further details and background on this setting.

Toward randomized subspace methods under heavy-tailed noise. For the RS-SGD update above, existing theory mainly provides expectation guarantees under light-tailed noise assumptions such as bounded variance (Flynn et al., 2024). We also establish a single-run high-probability guarantee under sub-Gaussian noise, as summarized in Table 1. This complements the expectation-based analysis of Flynn et al. (2024) with single-run guarantees, in line with the distinction between expectation and single-run guarantees discussed in Harvey et al. (2019); Madden et al. (2024); Liu et al. (2023). However, these light-tailed noise conditions can be overly restrictive in modern machine-learning workloads, where stochastic gradients are often heavy-tailed (Şimşekli et al., 2019); this limitation is not caused by the randomized subspace projection but rather reflects the well-known sensitivity of SGD-type updates to heavy-tailed gradient noise (Nguyen et al., 2019; Hodgkinson & Mahoney, 2021).

By contrast, *normalized* stochastic gradient methods (NSGD) admit both in-expectation and high-probability convergence guarantees under the weaker bounded p-th moment assumption stated later in Assumption 2.4 (Hübler et al., 2025), a condition commonly used in analyses of heavy-tailed gradient noise. Notably, unlike existing convergence analyses of gradient clipping under bounded p-th moment noise, which typically set the clipping thresholds and other hyperparameters using problem- or noise-dependent quantities, including the tail index p (Zhang et al., 2020; Nguyen et al., 2023), NSGD can be analyzed without knowing p.

These properties motivate the following question: *can gradient normalization be made compatible with randomized subspace methods using only projected stochastic gradients, while still obtaining convergence guarantees under heavy-tailed noise?* We answer this question affirmatively.

Randomized subspace normalized SGD. Building upon the randomized subspace framework, we propose randomized subspace normalized SGD (RS-NSGD), a subspace-compatible normalized method. A direct plug-in hybrid such as

$$x_{k+1} = x_k - \bar{\eta}\,\frac{P_k P_k^\top g_k}{\|g_k\|}$$

still requires the full-dimensional norm $\|g_k\|$, and thus falls outside the randomized-subspace model we target. In contrast, RS-NSGD normalizes using only the available projected gradient $P_k^\top g_k$ and performs the iteration

$$x_{k+1} = x_k - \bar{\eta}\,\frac{P_k P_k^\top g_k}{\|P_k^\top g_k\|},$$

thereby preserving the low-dimensional update structure of randomized subspace algorithms. Establishing convergence guarantees under heavy-tailed noise using only such low-dimensional gradient information is nontrivial, as also reflected by the limitations of existing compression-based

analyses in this regime; see Appendix A.4. Because normalization introduces a nonlinear dependence on the random quantity $\|P_k^\top g_k\|$, existing analytical tools for RS-SGD do not directly apply, substantially complicating the analysis. Nevertheless, Proposition 4.1 shows that the resulting direction remains exactly aligned in expectation with the full-dimensional normalized direction, with a factor uniformly bounded between $1/\sqrt{2}$ and 1. Moreover, Proposition 4.2 controls the corresponding normalized quadratic form while preserving the matrix-smoothness structure, revealing the effective-rank dependence needed for our sharper oracle-complexity comparison with NSGD.

Contributions.

- **Randomized subspace normalized SGD.** We propose RS-NSGD, a normalized stochastic method that preserves the low-dimensional update structure of randomized subspace methods.
- **Convergence under heavy-tailed noise.** Under a bounded p-th moment noise assumption, we establish in-expectation and high-probability convergence guarantees for RS-NSGD, covering both parameter-aware and parameter-agnostic regimes.
- **Oracle-complexity comparison with full-dimensional NSGD.** We identify asymptotic regimes in which RS-NSGD can improve oracle complexity over full-dimensional normalized SGD, clarifying when subspace normalization is theoretically beneficial.

We also provide a high-probability guarantee for RS-SGD under sub-Gaussian noise; this complementary result is summarized in Appendix C. In addition, Appendix D provides convergence analysis, both in expectation and with high probability, for randomized subspace normalized gradient descent (RS-NGD) in the deterministic setting, where $P_k^\top \nabla F(x_k)$ are used instead of $P_k^\top g_k$ as projected gradients.

Notation: We use the following notation throughout the paper. We assume $d \geq 2$ and $1 \leq r \leq d$. 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, and let $O(d)$ denote the orthogonal group consisting of all $d \times d$ real matrices U satisfying $U^\top U = I_d$.

2 PRELIMINARIES

2.1 ORACLE COMPLEXITY

We measure oracle complexity in terms of directional-derivative queries. Specifically, each directional derivative $v^\top \nabla f(x, \xi)$, with $v \in \mathbb{R}^d$, counts as one oracle call. Accordingly, computing the full gradient requires d oracle calls, whereas computing $P^\top \nabla f(x, \xi)$, for $P \in \mathbb{R}^{d\times r}$, requires r oracle calls. With minibatch size $\bar{B}$, these costs scale to $\bar{B}d$ and $\bar{B}r$ oracle calls per iteration, respectively. The same d-versus-r counting also corresponds to the number of scalars transmitted in distributed settings, where transmitting a full-dimensional gradient and an r-dimensional projected gradient requires d and r scalars, respectively [1]. Thus, this cost model captures both directional-derivative query complexity and the number of transmitted scalars in distributed settings [2].

2.2 HAAR-RANDOM SUBSPACES

Following Kozak et al. (2021), we use the scaled Haar-random subspace matrix

$$P = \sqrt{\frac{d}{r}}\, R,$$

where $R \in \mathbb{R}^{d\times r}$ consists of the first r columns of a Haar-distributed orthogonal matrix. Then

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

[1] This is consistent with Li et al. (2020), who count random-r sparsification as $32r$ bits per iteration.

[2] In stochastic optimization, *sample complexity* is often used to mean the number of samples accessed by the algorithm (e.g., $\bar{B}T$ for minibatch size $\bar{B}$ and T iterations). Thus, the oracle complexity equals the sample complexity times the number of directional-derivative queries per sample access.

VP has the same distribution as P for any $V \in \mathcal{O}(d)$, and $\|P^\top x\| > 0$ almost surely for every fixed $x \neq 0$. These properties are used throughout our analysis of the normalized projected updates. In practice, P can be generated by QR-factorizing a $d \times r$ standard Gaussian matrix and scaling by $\sqrt{d/r}$. The generation cost is $\mathcal{O}(dr^2)$, separate from the directional-derivative oracle complexity[3].

2.3 Problem Setting and Assumptions

We impose the following standard assumptions on F.

Assumption 2.1 (Lower Boundedness)**.** *F is bounded below, i.e., $F_* := \inf_{x\in\mathbb{R}^d} F(x) > -\infty$; hence, we can define $\Delta_0 := F(x_0) - F_*$.*

Assumption 2.2 (L-smoothness)**.** *F is L-smooth, i.e., $\exists L > 0, \forall x, y \in \mathbb{R}^d$,*

$$\|\nabla F(x) - \nabla F(y)\| \leq L\|x - y\|. \tag{2}$$

Under Assumption 2.2, the standard descent lemma implies that the following one-sided quadratic bound holds with the isotropic choice $\mathbb{L} = LI_d$:

$$F(y) - F(x) - \langle \nabla F(x), y - x\rangle \leq \frac{1}{2}(y - x)^\top \mathbb{L}(y - x), \qquad \forall x, y \in \mathbb{R}^d. \tag{3}$$

Throughout the paper, we fix a nonzero symmetric positive semidefinite matrix $\mathbb{L} \in \mathbb{R}^{d\times d}$ satisfying (3) and $\|\mathbb{L}\| \leq L$, and refer to it as the smoothness matrix. This allows the analysis to exploit anisotropic smoothness of F, while the isotropic choice LI_d is always valid[4].

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). This refined smoothness notion is particularly useful for *in-expectation* convergence in our Haar-based analysis; it leads to tighter bounds.

We impose the following unbiasedness assumption on the stochastic gradient.

Assumption 2.3 (Unbiasedness)**.** *For all $x \in \mathbb{R}^d$, $\mathbb{E}[\nabla f(x, \xi)] = \nabla F(x)$.*

We impose the following weak moment condition on the stochastic gradient.

Assumption 2.4 (p-BCM)**.** *For some $p \in (1, 2]$, the gradient has bounded p-th moment, i.e., for all $x \in \mathbb{R}^d$,*

$$\mathbb{E}[\|\nabla f(x, \xi) - \nabla F(x)\|^p] \leq \sigma^p.$$

Following Hübler et al. (2025), we adopt the bounded p-th moment condition in Assumption 2.4 as our main noise assumption for RS-NSGD. The light-tailed assumptions used for the SGD and RS-SGD baselines are stronger than Assumption 2.4; these assumptions and their relation to Assumption 2.4 are recalled in Appendix C.1.

3 Randomized Subspace Normalized SGD

We first recall RS-SGD, which we use as a light-tailed baseline. At iteration k, let $\{\xi_k^j\}_{j=1}^{\bar{B}}$ be i.i.d. samples from $\mathcal{D}$, and define the mini-batch stochastic gradient

$$g_k := \frac{1}{\bar{B}} \sum_{j=1}^{\bar{B}} \nabla f(x_k, \xi_k^j).$$

Let $P_k \in \mathbb{R}^{d\times r}$ be a Haar-random subspace matrix. RS-SGD performs the update

$$x_{k+1} = x_k - \bar{\eta} P_k P_k^\top g_k.$$

Note that if we choose $r = d$, then P_k becomes an orthogonal matrix and RS-SGD reduces to standard SGD. Hence, RS-SGD can be regarded as a generalization of SGD. Prior work (Flynn et al.,

[3] In Appendix F.3, Haar-sketch generation takes less than 1% of the total step time.

[4] For example, for $F(x_1, \ldots, x_d) = x_1^2/2$, the function is L-smooth with $L = 1$, and the choice $\mathbb{L} = \mathrm{diag}(1, 0, \ldots, 0)$ satisfies (3). This example shows that $\mathbb{L}$ can capture anisotropic smoothness.

Algorithm 1 Randomized Subspace Normalized SGD (RS-NSGD)

Require: Initial point $x_0 \in \mathbb{R}^d$, stepsize $\bar{\eta}$, mini-batch size $\bar{B}$

1: **for** $k = 0, 1, \ldots, T-1$ **do**
2: Sample a Haar-random matrix $P_k \in \mathbb{R}^{d\times r}$
3: Sample i.i.d. data $\{\xi_k^j\}_{j=1}^{\bar{B}} \sim \mathcal{D}$
4: Compute the projected mini-batch gradient $u_k \leftarrow \frac{1}{\bar{B}} \sum_{j=1}^{\bar{B}} P_k^\top \nabla f(x_k, \xi_k^j)$
5: **if** $\|u_k\| > 0$ **then**
6: $x_{k+1} \leftarrow x_k - \bar{\eta} \frac{P_k u_k}{\|u_k\|}$
7: **else**
8: $x_{k+1} \leftarrow x_k$

2024) establishes convergence guarantees for RS-SGD in expectation under Assumption C.2. We show in Appendix C that RS-SGD also admits a high-probability guarantee under Assumption C.1.

However, these noise conditions can be overly restrictive when stochastic gradients are heavy-tailed. For $p < 2$, Assumption 2.4 need not imply a finite second moment, so the standard smoothness analysis of SGD and RS-SGD does not close under this assumption. To address this limitation, we propose randomized subspace normalized SGD (RS-NSGD), summarized in Algorithm 1. The method uses only the projected stochastic gradient u_k and performs normalization using this r-dimensional vector. As in the discussion of RS-SGD, when $r = d$, RS-NSGD reduces to the standard NSGD; hence, RS-NSGD can be viewed as a generalization of NSGD.

The projected per-sample gradient $P_k^\top \nabla f(x_k, \xi_k^j)$ can be computed in different ways depending on the setting. In distributed optimization, the stochastic gradient $\nabla f(x_k, \xi_k^j)$ may be computed locally and projected before communication, reducing the transmitted vector from d to r dimensions. Alternatively, with access to directional-derivative queries, $P_k^\top \nabla f(x_k, \xi_k^j)$ is obtained from r directional derivatives along the columns of P_k, without forming the full stochastic gradient.

4 Convergence Results for RS-NSGD

In this section, we establish in-expectation and high-probability convergence guarantees for RS-NSGD. Proofs of the results in this section are deferred to Appendix B. First, we obtain the following important proposition.

Proposition 4.1. *Let $x \neq 0$, and let P be a Haar matrix. Then $\|P^\top x\| \neq 0$ almost surely, and*

$$\mathbb{E}\left[\frac{PP^\top x}{\|P^\top x\|}\right] = \tau \frac{x}{\|x\|}, \qquad \text{with } \tau = \sqrt{\frac{d}{r}} \frac{\Gamma\left(\frac{r+1}{2}\right)\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{r}{2}\right)\Gamma\left(\frac{d+1}{2}\right)}.$$

Here, $\Gamma(\cdot)$ denotes the Gamma function. Furthermore, $\frac{1}{\sqrt{2}} \leq \tau \leq 1$.

When $r = d$, the coefficient τ becomes 1. Moreover, as illustrated in Figure 1, even for small r, the value of τ quickly approaches 1.

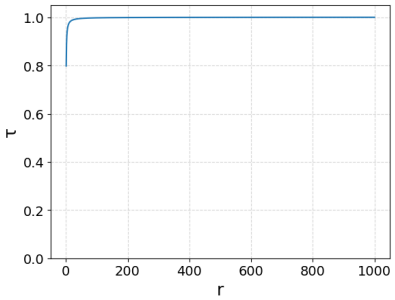


Figure 1: The value of τ as a function of r for $d = 1000$.

4.1 In-expectation Bound

To obtain tight bounds on the quadratic terms in the in-expectation convergence analysis, we derive the following elementary proposition from properties of the Haar matrix and basic linear algebra.

Proposition 4.2. *Let S be a $d \times d$ square matrix, $v \neq 0$, and P be a Haar matrix. Then,*

$$\mathbb{E}\left[\frac{v^\top PP^\top SPP^\top v}{\|P^\top v\|^2}\right] = \frac{d(r-1)}{r(d-1)} \frac{v^\top Sv}{\|v\|^2} + \frac{d-r}{r(d-1)} \operatorname{tr}(S).$$

When bounding the first term in Proposition 4.2 by $\|S\|$,
the relationship between $\mathrm{tr}(S)$ and $\|S\|$ becomes important. To this end, we introduce the following quantity.

Definition 4.3. *Let S be a nonzero positive semidefinite matrix. We define the effective rank $r_{\mathrm{eff}}(S)$ and the associated quantity $\ell(S)$ by*

$$r_{\mathrm{eff}}(S) := \frac{\mathrm{tr}(S)}{\|S\|}, \qquad \ell(S) := \frac{d(r-1) + r_{\mathrm{eff}}(S)(d-r)}{r(d-1)}.$$

Since $\ell(S)$ is monotonically increasing with respect to $r_{\mathrm{eff}}(S)$ and $r_{\mathrm{eff}}(S) \in [1, d]$, we have $\ell(S) \in [1, d/r]$. Furthermore, when $r \neq d$, $\ell(S) = d/r$ holds if and only if $r_{\mathrm{eff}}(S) = d$, which occurs when all eigenvalues of S are equal, that is, when S is a scalar multiple of the identity matrix. Moreover, when $r = d$, we simply have $\ell(S) = 1$ for all S.

We are now ready to state an in-expectation convergence bound for RS-NSGD.

Theorem 4.4. *Suppose that Assumptions 2.1 to 2.4 hold. Let $\{x_k\}_{k\geq 0}$ be the sequence generated by RS-NSGD. We set*

$$\bar{B} = \lceil \max\{1,\, BT^q\} \rceil, \qquad \bar{\eta} = \eta T^{-u},$$

for parameters $B > 0$, $q > 0$, $\eta > 0$, and $u \in (0, 1)$. Then it holds that

$$\frac{1}{T}\sum_{k=0}^{T-1} \mathbb{E}[\|\nabla F(x_k)\|] \leq \frac{\Delta_0}{\tau\eta T^{1-u}} + \frac{\eta \ell(\mathbb{L})\|\mathbb{L}\|}{2\tau T^u} + \frac{4\sigma}{\max\{1, BT^q\}^{(p-1)/p}}.$$

As noted earlier, when $r = d$, RS-NSGD reduces to NSGD. Indeed, substituting $r = d$ and $\|\mathbb{L}\| \leq L$ into Theorem 4.4 yields a convergence bound that exactly matches the known result for NSGD (Hübler et al., 2025).

From Theorem 4.4, we can derive the corresponding oracle complexity bounds.

We first consider the setting where the exact value of p and other problem-parameters are unknown. In this case, we take $u = \frac{1}{2}$ and $q = 1$ [5] . With this choice, taking η and B to be constants, the resulting oracle complexity[6] is given by

$$\mathcal{O}\left(r\left(\frac{\Delta_0}{\tau\epsilon}\right)^4 + r\left(\frac{\ell(\mathbb{L})\,\|\mathbb{L}\|}{\tau\epsilon}\right)^4 + r\left(\frac{\sigma}{\epsilon}\right)^{\frac{2p}{p-1}}\right).$$

For comparison, the corresponding bound for the full-dimensional method is obtained by substituting $r = d$ into the above expression:

$$\mathcal{O}\left(d\left(\frac{\Delta_0}{\epsilon}\right)^4 + d\left(\frac{\|\mathbb{L}\|}{\epsilon}\right)^4 + d\left(\frac{\sigma}{\epsilon}\right)^{\frac{2p}{p-1}}\right).$$

In particular, when $p = 2$ (so that $\frac{2p}{p-1} = 4$), the RS-NSGD bound can be written as

$$\mathcal{O}\Big(\epsilon^{-4}\, r\Big(\tfrac{\Delta_0^4}{\tau^4} + \tfrac{(\ell(\mathbb{L})\|\mathbb{L}\|)^4}{\tau^4} + \sigma^4\Big)\Big),$$

whereas the full-dimensional bound becomes

$$\mathcal{O}\Big(\epsilon^{-4}\, d\big(\Delta_0^4 + \|\mathbb{L}\|^4 + \sigma^4\big)\Big).$$

Thus, for instance when $\ell(\mathbb{L})$ is small, some subspace dimensions $r < d$ can reduce the leading ϵ^{-4} coefficient and yield a smaller oracle complexity than the full-dimensional method.

[5] As in Hübler et al. (2025), when p is unknown in $(1, 2]$, no fixed $q \neq 1$ can uniformly improve the ϵ-exponent for all $p \in (1, 2]$. Moreover, when $p = 2$, the ϵ-exponent is uniquely minimized at $q = 1$. By continuity in p, the same holds when p is unknown in $(1, 2)$. Therefore, even if one only knows that $p < 2$, choosing $(u, q) = (\frac{1}{2}, 1)$ remains a natural default.

[6] For oracle- and iteration-complexity bounds, $\mathcal{O}$ notation hides constants; statements hold for all sufficiently small ϵ, with thresholds possibly depending on parameters. We use $\tilde{\mathcal{O}}(\cdot)$ to additionally hide polylogarithmic factors in $1/\delta$.

When $p < 2$, the noise term dominates and the bounds simplify to

$$\text{RS-NSGD: } \mathcal{O}\left(r\left(\frac{\sigma}{\epsilon}\right)^{\frac{2p}{p-1}}\right) \qquad \text{NSGD: } \mathcal{O}\left(d\left(\frac{\sigma}{\epsilon}\right)^{\frac{2p}{p-1}}\right),$$

which shows that RS-NSGD improves the leading dependence by a factor r/d.

Next, when p and the other problem-parameters are known, take $u = \frac{1}{2}$ and $q = \frac{p}{2(p-1)}$. By setting

$$\eta \coloneqq \sqrt{\frac{\Delta_0}{\ell(\mathbb{L})\,\|\mathbb{L}\|}}, \qquad B \coloneqq \left(\frac{\sigma\,\tau}{\sqrt{\Delta_0 \ell(\mathbb{L})\|\mathbb{L}\|}}\right)^{\frac{p}{p-1}},$$

we obtain the oracle complexity bound

$$\mathcal{O}\left(r\,\frac{\ell(\mathbb{L})}{\tau^2}\,\frac{\Delta_0\,\|\mathbb{L}\|}{\epsilon^2}\left(\frac{\sigma}{\epsilon}\right)^{\frac{p}{p-1}}\right).$$

For comparison, the full-dimensional case (NSGD, i.e., $r = d$) is obtained by simply setting $r = d$ in the above bound:

$$\mathcal{O}\left(d\,\frac{\Delta_0\,\|\mathbb{L}\|}{\epsilon^2}\left(\frac{\sigma}{\epsilon}\right)^{\frac{p}{p-1}}\right).$$

Hence, the RS-NSGD rate is a factor $\frac{r}{d}\frac{\ell(\mathbb{L})}{\tau^2}$ times the NSGD rate. Since $\ell(\mathbb{L}) \le d/r$, this factor satisfies $\frac{r}{d}\frac{\ell(\mathbb{L})}{\tau^2} \le 1/\tau^2$; moreover, τ stays close to 1 even when r is small as shown in Figure 1, so the factor $1/\tau^2$ is roughly 1. Therefore, except when $r_{\text{eff}}(\mathbb{L})$ is close to d, where $\ell(\mathbb{L})$ nearly saturates its upper bound d/r, RS-NSGD can reduce the oracle cost.

4.2 High-probability Bound

We now derive the following high-probability bound.

Theorem 4.5. *Assume the same setting as in Theorem 4.4 for F, $\bar{B}$, and $\bar{\eta}$, with parameters $B > 0$, $q > 0$, $\eta > 0$, and $u \in (0,1)$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, the sequence $\{x_k\}_{k\ge 0}$ generated by RS-NSGD satisfies*

$$\frac{1}{T}\sum_{k=0}^{T-1}\|\nabla F(x_k)\| \le \frac{2\Delta_0}{\tau\eta T^{1-u}} + 17\frac{\frac{d}{r}\sqrt{\Delta_0 L}}{\tau^2 T}\log(1/\delta)$$
$$+\frac{\frac{d}{r}\eta L}{\tau T^u}\left(1 + 12\frac{\sqrt{\frac{d}{r}}}{\tau}\log(1/\delta)\right) + \frac{8\sigma}{\max\{1, BT^q\}^{\frac{p-1}{p}}}.$$

Similarly to Theorem 4.4, substituting $r = d$ into Theorem 4.5 yields a bound that exactly matches the known result for NSGD (Hübler et al., 2025).

From Theorem 4.5, we can derive the oracle complexity. When the problem-parameters (including p) are unknown, we adopt, as a natural default, the same choice used in the in-expectation setting and set $u = \frac{1}{2}$ and $q = 1$. Taking η and B to be constants, the resulting oracle complexity is given by

$$\tilde{\mathcal{O}}\left(r\left(\frac{\Delta_0}{\tau\epsilon}\right)^4 + \frac{d^6}{r^5}\left(\frac{L}{\tau^2\epsilon}\right)^4 + r\left(\frac{\sigma}{\epsilon}\right)^{\frac{2p}{p-1}}\right).$$

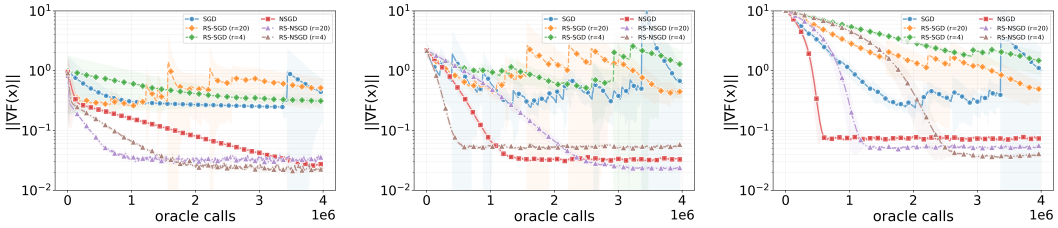
For comparison, the full-dimensional method corresponds to NSGD (i.e., $r = d$):

$$\tilde{\mathcal{O}}\left(d\left(\frac{\Delta_0}{\epsilon}\right)^4 + d\left(\frac{L}{\epsilon}\right)^4 + d\left(\frac{\sigma}{\epsilon}\right)^{\frac{2p}{p-1}}\right).$$

In particular, when $p = 2$, both bounds scale as ϵ^{-4}, and RS-NSGD can yield a smaller oracle complexity than NSGD, depending on the problem constants, for some $r < d$. When $p < 2$, similarly to the in-expectation case, RS-NSGD improves the leading term by a factor r/d compared to NSGD.

Table 2: Asymptotic oracle-complexity comparison between RS-NSGD and NSGD.

Setting	**In expectation**	**High probability**
Parameters known	RS-NSGD favorable if $\frac{r}{d}\frac{\ell}{\tau^2} < 1$	NSGD better in this bound
$p < 2$**, parameters unknown**	RS-NSGD better by $\frac{r}{d}$	RS-NSGD better by $\frac{r}{d}$
$p = 2$**, parameters unknown**	Depends on problem parameters	Depends on problem parameters



(a) $\rho = 4$ (b) $\rho = 20$ (c) $\rho = 100$

Figure 2: Synthetic quadratic with heavy-tailed noise ($d = 100$, $\bar{B} = 4$): oracle calls vs. $\|\nabla F(x)\| = \|\Lambda x\|$ (mean $\pm$ 1 std., 5 seeds).

Next, when the value of p as well as the other problem-parameters are known, we adopt the same choice used in the in-expectation setting and take $u = \frac{1}{2}$ and $q = \frac{p}{2(p-1)}$. Set

$$\eta := \left(\frac{r}{d}\right)^{\frac{3}{4}} \sqrt{\frac{\Delta_0 \tau}{L}}, \qquad B := \left(\frac{\sigma \tau^{\frac{3}{2}}}{\sqrt{\Delta_0 L}} \left(\frac{r}{d}\right)^{\frac{3}{4}}\right)^{\frac{p}{p-1}}.$$

Then the oracle complexity becomes

$$\tilde{\mathcal{O}}\left(\left(\frac{d}{r^{1/3}\tau^2}\right)^{3/2} \frac{\Delta_0 L}{\epsilon^2} \left(\frac{\sigma}{\epsilon}\right)^{\frac{p}{p-1}}\right).$$

In this case, the bound is minimized by the choice $r = d$ (i.e., NSGD).

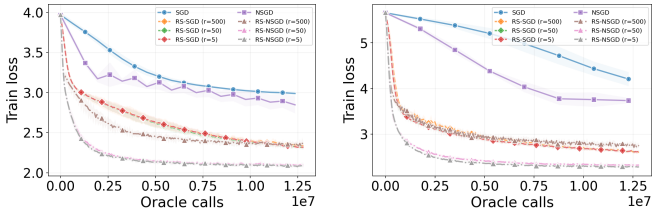
Table 2 summarizes these asymptotic comparisons. Overall, these results yield a theoretical guideline for when to use randomized subspace projection: the answer depends on the available problem information—including whether p is unknown, only known to satisfy $p \neq 2$, or known—and on whether one needs in-expectation or high-probability guarantees.

5 Experiments

We empirically validate our theoretical findings on controlled synthetic problems and real datasets. Further experimental details (including tuning details and additional results) are provided in Appendix E. Additional experiments with extra baselines and datasets are reported in Appendix F, together with a forward-mode AD timing experiment.

5.1 Synthetic quadratic with heavy-tailed noise

We consider a stochastic quadratic objective in dimension $d = 100$, $f(x, \xi) := \frac{1}{2}x^\top \Lambda x + \langle \xi, x \rangle$, where $\xi \in \mathbb{R}^d$ is sampled from a *symmetrized Pareto* distribution with tail index $\hat{p} = 1.2$. In particular, ξ has a bounded p-th moment for any $p < \hat{p}$. We set $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ as follows: $\Lambda = \mathrm{diag}\left(1, \frac{\rho-1}{d-1}, \ldots, \frac{\rho-1}{d-1}\right)$, where $\rho \leq d$. This choice ensures that $r_{\mathrm{eff}}(\Lambda) = \mathrm{tr}(\Lambda)/\|\Lambda\| = \rho$. We report results for $\rho \in \{4, 20, 100\}$. We compare six methods: SGD, NSGD, RS-SGD with $r \in \{4, 20\}$, and RS-NSGD with $r \in \{4, 20\}$, using minibatch size $\bar{B} = 4$. We tune the stepsize $\bar{\eta}$ for each method and report trajectories under a fixed oracle-call budget; details are given in Appendix E.



(a) PTB **(b)** WikiText-2

Figure 3: Character-level language modeling on PTB ($d = 5079, \bar{B} = 128$) and WikiText-2 ($d = 13700, \bar{B} = 128$): oracle calls vs. training loss (mean $\pm$ 1 std., 5 seeds).

Figure 2 summarizes the results for $\rho \in \{4, 20, 100\}$, plotting the number of oracle calls (x-axis) versus $\|\nabla F(x)\|$ (y-axis). Each curve reports the mean over 5 evaluation seeds, with a shaded band indicating ± 1 standard deviation. Note that NSGD can be interpreted as a special case of RS-NSGD with $r = d = 100$ (and similarly, SGD is RS-SGD with $r = d = 100$). Across all settings, SGD and RS-SGD ($r \in \{4, 20\}$) exhibit limited progress under heavy-tailed noise and often fail to consistently decrease $\|\nabla F(x)\|$. Moreover, the wider standard-deviation bands of SGD and RS-SGD suggest less stable optimization behavior, plausibly due to their stronger sensitivity to heavy-tailed noise. In contrast, NSGD and RS-NSGD ($r \in \{4, 20\}$) show a clear decrease of $\|\nabla F(x)\|$. Moreover, when the effective rank is largest ($r_{\text{eff}}(\Lambda) = \rho = 100$), NSGD achieves the fastest reduction per oracle call, whereas for smaller effective ranks ($\rho \in \{4, 20\}$), RS-NSGD is more oracle-efficient than NSGD. This behavior is compatible with our theory.

5.2 Character-level language modeling

Motivated by prior evidence of heavy-tailed stochastic gradients in language modeling (Zhang et al., 2020), we follow Hübler et al. (2025) and use language modeling as a real-data testbed. We evaluate on the character-level Penn Treebank (PTB) (Marcus et al., 1993) and WikiText-2 (v1) (Merity et al., 2017) benchmarks. We train a single-layer GRU language model (Cho et al., 2014) with embedding dimension $E = 8$ and hidden size $H = 28$ at the character level. The character vocabulary is built from the training split (with an `<unk>` token for unseen characters in validation/test). At each optimization step, we sample $\bar{B} = 128$ length-35 subsequences i.i.d. from the training corpus (by choosing start positions uniformly with replacement). The resulting parameter dimension is $d = 5079$ on PTB and $d = 13700$ on WikiText-2. For plotting, we compute the training loss as the average mini-batch loss over five independent mini-batches. We consider subspace dimensions $r \in \{5, 50, 500\}$ for RS-SGD/RS-NSGD. We tune the stepsize $\bar{\eta}$ for each method; the tuning and evaluation protocol is described in Appendix E.

Figure 3 reports oracle calls versus training loss on PTB and WikiText-2. Each curve shows the mean over five seeds, with shaded regions indicating ± 1 standard deviation. Across both datasets, RS-NSGD with $r \in \{5, 50\}$ achieves the strongest performance under the oracle budget. Similar trends are observed for validation and test losses; see Figure 4.

6 Conclusion

We proposed RS-NSGD and established in-expectation and high-probability guarantees under a bounded p-th moment assumption, characterizing when it can improve oracle complexity over NSGD. Our guarantees are stated for Haar-random subspaces. Extensions to block-coordinate sketches require additional care, since one can have $\mathbb{P}(P^\top x = 0) > 0$ for some fixed $x \neq 0$. By contrast, Gaussian sketches may allow an analogous analysis if the update is normalized by $\|PP^\top g_k\|$, which can remove the scale variability of the projected direction. We leave this question to future work.

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

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

Chuyan Chen, Yutong He, Pengrui Li, Weichen Jia, and Kun Yuan. Greedy low-rank gradient compression for distributed learning with convergence guarantees, 2025a. arXiv:2507.08784v4.

Yiming Chen, Yuan Zhang, Yin Liu, Kun Yuan, and Zaiwen Wen. A memory efficient randomized subspace optimization method for training large language models, 2025b. arXiv:2502.07222v1.

Zhijie Chen, Qiaobo Li, and Arindam Banerjee. Sketched adaptive federated deep learning: A sharp convergence analysis, 2024. arXiv:2411.06770v1.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, 2014.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011.

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.

S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

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.

Nicholas J. A. Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Proceedings of the 32nd Conference on Learning Theory (COLT)*, volume 99, pp. 1579–1613, 2019.

Yutong He, Pengrui Li, Yipeng Hu, Chuyan Chen, and Kun Yuan. Subspace optimization for large language models with convergence guarantees. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267, pp. 22468–22522, 2025.

Liam Hodgkinson and Michael W. Mahoney. Multiplicative noise and heavy tails in stochastic optimization. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pp. 4262–4274, 2021.

F. Hübler, I. Fatkhullin, and N. He. From gradient clipping to normalization for heavy tailed SGD. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 258, pp. 2413–2421, 2025.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. `https://kellerjordan.github.io/posts/muon/`, 2024.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian U. Stich, and Martin Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97, pp. 3252–3261, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

Nikita Kornilov, Philip Zmushko, Andrei Semenov, Alexander Gasnikov, and Alexander Beznosikov. Sign operator for coping with heavy-tailed noise: High probability convergence bounds with extensions to distributed optimization and comparison oracle, 2025. arXiv:2502.07923v1.

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

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 (ICLR)*, 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.

Z. Liu and Z. Zhou. Nonconvex stochastic optimization under heavy-tailed noises: Optimal convergence without gradient clipping. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 2025.

Z. Liu, T. D. Nguyen, T. H. Nguyen, A. Ene, and H. L. Nguyen. High probability convergence of stochastic gradient methods. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pp. 21884–21914, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations*, 2019.

Liam Madden, Emiliano Dall'Anese, and Stephen Becker. High probability convergence bounds for non-convex stochastic gradient descent with sub-weibull noise. *Journal of Machine Learning Research*, 25(241):1–36, 2024.

Matt Mahoney. Large text compression benchmark. `https://www.mattmahoney.net/dc/textdata.html`, 2011.

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

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017.

Y. E. Nesterov. Minimization methods for nonsmooth convex and quasiconvex functions. *Matekon*, 29(3):519–531, 1984.

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.

T. D. Nguyen, T. H. Nguyen, A. Ene, and H. L. Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tailed noise. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 24191–24222, 2023.

Thanh Huy Nguyen, Umut Simsekli, Mert Gurbuzbalaban, and Gaël Richard. First exit time analysis of stochastic gradient descent under heavy-tailed gradient noise. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

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

R. E. A. C. Paley and A. Zygmund. A note on analytic functions in the unit circle. *Proceedings of the Cambridge Philosophical Society*, 28(3):266–272, 1932.

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.

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.

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.

F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs. In *Interspeech*, pp. 1058–1062, 2014.

U. Şimşekli, L. Sagun, and M. Gürbüzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97, pp. 5827–5837, 2019.

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.

Tao Sun, Xinwang Liu, and Kun Yuan. Revisiting gradient normalization and clipping for nonconvex SGD under heavy-tailed noise: Necessity, sufficiency, and acceleration, 2025. arXiv:2410.16561v4.

Roman Vershynin. High-dimensional probability: An introduction with applications in data science (second edition). Pre-publication PDF, November 2025. URL `https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-2.pdf`. Pre-publication version dated Nov 28, 2025.

T. Vogels, S. P. Karimireddy, and M. Jaggi. PowerSGD: Practical low-rank gradient compression for distributed optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

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.

Ziqing Wen, Yanqi Shi, Jiahuan Wang, Ping Luo, Linbo Qiao, Dongsheng Li, and Tao Sun. SRON: State-free LLM training via row-wise gradient normalization. OpenReview: ICLR 2026 Conference (Withdrawn Submission), 2025. OpenReview:BtQLBWr6zI.

J. G. Wendel. Note on the gamma function. *American Mathematical Monthly*, 55(9):563–564, 1948.

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

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.

J. Zhang, S. P. Karimireddy, A. Veit, S. Kim, S. Reddi, S. Kumar, and S. Sra. Why are adaptive methods good for attention models? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 15383–15393, 2020.

A Related Work

A.1 Optimization via Random Subspaces

Randomized subspace methods restrict each update to a randomly selected low-dimensional subspace. When directional-derivative queries are available, an r-dimensional projected gradient can be obtained from r directional derivatives along the columns of $P \in \mathbb{R}^{d\times r}$, without forming the full gradient. This subspace-descent approach was developed for deterministic optimization by Kozak et al. (2021). Related randomized subspace approaches have also been studied in (Nesterov & Spokoiny, 2017; Baydin et al., 2022; Nozawa et al., 2023).

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 $P = [p_1, \dots, p_r] \in \mathbb{R}^{d\times r}$, the projected gradient can be written as

$$P^\top \nabla f(x,\xi) = \big(p_1^\top \nabla f(x,\xi), \dots, p_r^\top \nabla f(x,\xi)\big)^\top,$$

so it can be obtained from r directional derivatives along the columns of P. 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 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.

Gradient compression techniques have been widely studied to alleviate communication bottlenecks in distributed SGD. Classical approaches include quantization or error-feedback variants, along with low-rank projection methods such as PowerSGD (Seide et al., 2014; Karimireddy et al., 2019; Vogels et al., 2019). More recently, randomized linear compressors—including those based on Haar matrices—have been shown to provide convergence in non-convex optimization (Flynn et al., 2024). This provides theoretical support for using Haar matrices as randomized compressors in optimization. Rather than communicating the full stochastic gradient $g_k \in \mathbb{R}^d$ at the current iterate, we draw a random r-dimensional subspace represented by a matrix $P_k \in \mathbb{R}^{d\times r}$ and perform the projected update

$$x_{k+1} = x_k - \bar{\eta}\, P_k P_k^\top g_k \;=\; x_k - \bar{\eta}\, P_k\big(P_k^\top g_k\big).$$

In a minibatch setting where

$$g_k \;=\; \frac{1}{\bar{B}} \sum_{j=1}^{\bar{B}} \nabla f(x_k, \xi_k^j),$$

the communicated quantity is the projected per-sample (or per-worker) gradient

$$u_k^j \;:=\; P_k^\top \nabla f(x_k, \xi_k^j) \in \mathbb{R}^r,$$

and the aggregator forms

$$P_k^\top g_k \;=\; \frac{1}{\bar{B}} \sum_{j=1}^{\bar{B}} u_k^j.$$

Thus, each worker transmits an r-dimensional vector per iteration (instead of d dimensions), and the server applies the update using the projected gradient $P_k P_k^\top g_k \in \mathbb{R}^d$. Flynn et al. (2024) analyze the convergence penalty introduced by compression and show that leveraging matrix-smoothness leads to sharp convergence guarantees in expectation under the bounded-variance assumption.

A.2 Stochastic Optimization under Heavy-Tailed Noise

Classical analyses have established in-expectation convergence for standard SGD under the bounded-variance assumption (Ghadimi & Lan, 2013). More recently, high-probability convergence has been shown under the sub-Gaussian gradient noise assumption (Liu et al., 2023). However, growing empirical and theoretical evidence indicates that stochastic gradients arising in modern large-scale optimization can exhibit heavy-tailed behavior, which may fall outside the classical bounded-variance assumption. This phenomenon has been observed in a variety of settings, including deep neural network training (Şimşekli et al., 2019; Nguyen et al., 2019), motivating a growing body of theoretical work on optimization under heavy-tailed noise. In particular, clipped SGD and its variants have been extensively analyzed in this context, leading to convergence guarantees both in expectation and with high probability; see, e.g., Zhang et al. (2020); Nguyen et al. (2023). However, convergence analyses often rely on carefully tuned clipping thresholds and assume knowledge of problem-dependent parameters (e.g., noise moments or smoothness constants).

A.3 Normalized Gradient Methods

Motivated by these limitations, there has been renewed interest in *normalized* gradient methods, going back to Nesterov's normalized gradient schemes for stochastic optimization (Nesterov, 1984). Hübler et al. (2025) study NSGD under heavy-tailed noise with bounded p-th moments, and establish both in-expectation and high-probability convergence guarantees in two regimes: (i) a fully parameter-free setting where the tail index p is unknown, and (ii) a parameter-aware setting where p is known. Relatedly, Liu & Zhou (2025); Sun et al. (2025) analyze momentum-based normalized methods under heavy-tailed noise and establish convergence guarantees in expectation; however, corresponding high-probability guarantees remain unavailable. In fact, for momentum variants, Hübler et al. (2025) empirically suggest that obtaining a clean high-probability convergence bound with a simple $\log(1/\delta)$ dependence may be challenging.

A.4 Compression Meets Heavy Tails: Limitations of Existing Work

Recently, the interaction between gradient compression and heavy-tailed noise has attracted attention. Chen et al. (2024) analyze clipped SGD with gradient compression under heavy-tailed noise. However, their high-probability convergence guarantees apply only in the parameter-aware regime where the tail index p is known. Moreover, their guarantees require choosing the subspace dimension r to satisfy a lower bound that depends on d, T, and δ. In other words, for a fixed r, the condition can be rearranged to yield a restriction on the allowable horizon T, so the result does not directly provide an ϵ-dependent iteration complexity guarantee.

As another line of work, Kornilov et al. (2025) analyze sign-based stochastic methods (SignSGD and its variants) under heavy-tailed noise. They derive high-probability convergence bounds, but these bounds exhibit an unfavourable dependence on the dimension d. In particular, measured in the Euclidean norm, their bounds include terms of the form $(\sqrt{d}/\epsilon)^{(3p-2)/(p-1)}$ when p is known and $(\sqrt{d}/\epsilon)^{2p/(p-1)}$ when p is unknown; thus, the dependence on d can become significantly worse for small p.

A.5 Randomized Subspaces in Modern Large-Scale Training

We cite the following large-scale training works only as context: although they also reduce the update dimension, their primary focus and algorithmic setting differ from ours. In a related but distinct

line of work, low-dimensional update ideas have recently gained visibility in large-scale training—especially in LLM optimization—as practical tools for improving memory and/or communication efficiency. Much of this literature is experiment-driven, but several works also develop stochastic convergence analyses (Chen et al., 2025a;b; Wen et al., 2025; He et al., 2025). Existing analyses in this literature often focus on in-expectation guarantees under standard assumptions such as bounded variance; high-probability bounds and heavy-tailed noise are typically not their primary focus. Rather, a central motivation in this line is system-level efficiency in large-scale training, whereas our motivation is theory-first convergence analysis.

B Missing Proofs for RS-NSGD

Proposition B.1 (Restatement of Proposition 4.1)**.** *Let $x \neq 0$. Suppose P is constructed from a Haar-random orthogonal matrix. Then $\|P^\top x\| \neq 0$ almost surely, and there exists a constant $\tau > 0$ such that*

$$\mathbb{E}\left[\frac{PP^\top x}{\|P^\top x\|}\right] = \tau \frac{x}{\|x\|}.$$

Specifically, τ is given by

$$\tau = \sqrt{\frac{d}{r}} \frac{\Gamma\left(\frac{r+1}{2}\right)\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{r}{2}\right)\Gamma\left(\frac{d+1}{2}\right)}.$$

Here, $\Gamma(\cdot)$ denotes the Gamma function.

Moreover, $\frac{1}{\sqrt{2}} \leq \tau \leq 1$.

Proof. Let $g : \mathbb{R}^d \to \mathbb{R}$ be defined by $g(x) := \|P^\top x\|$. For $x \neq 0$, we have $\|P^\top x\| \neq 0$ almost surely, and thus g is differentiable at x almost surely with

$$\nabla g(x) = \frac{PP^\top x}{\|P^\top x\|}. \tag{4}$$

Moreover,

$$\|\nabla g(x)\| = \left\| \frac{PP^\top x}{\|P^\top x\|} \right\| \leq \|P\| \leq \sqrt{\frac{d}{r}}. \tag{5}$$

Hence we may interchange gradient and expectation, obtaining

$$\mathbb{E}\left[\frac{PP^\top x}{\|P^\top x\|}\right] = \mathbb{E}[\nabla g(x)] = \nabla \mathbb{E}[g(x)]. \tag{6}$$

We next compute $\mathbb{E}[g(x)]$, which can be written as

$$\mathbb{E}[g(x)] = \mathbb{E}[\|P^\top x\|] = \|x\| \, \mathbb{E}\left[\left\| P^\top \frac{x}{\|x\|} \right\|\right]. \tag{7}$$

We use the following fact (see the proof of Lemma 1 in Kozak et al. (2021)): the random variable

$$Y := \frac{r}{d} \left\| P^\top \frac{x}{\|x\|} \right\|^2 \tag{8}$$

follows a Beta distribution,

$$Y \sim \mathrm{Beta}\left(\frac{r}{2}, \frac{d-r}{2}\right). \tag{9}$$

Let $\alpha := \frac{r}{2}$ and $\beta := \frac{d-r}{2}$, we have

$$\begin{aligned}
\mathbb{E}[\sqrt{Y}] &= \int_0^1 \sqrt{y} \, \frac{y^{\alpha-1}(1-y)^{\beta-1}}{\mathrm{B}(\alpha,\beta)} \, dy = \int_0^1 \frac{y^{\alpha-\frac{1}{2}}(1-y)^{\beta-1}}{\mathrm{B}(\alpha,\beta)} \, dy \\
&= \frac{\mathrm{B}\left(\alpha+\frac{1}{2},\beta\right)}{\mathrm{B}(\alpha,\beta)} = \frac{\Gamma\left(\alpha+\frac{1}{2}\right)\Gamma(\alpha+\beta)}{\Gamma\left(\alpha+\beta+\frac{1}{2}\right)\Gamma(\alpha)} \\
&= \frac{\Gamma\left(\frac{r+1}{2}\right)\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d+1}{2}\right)\Gamma\left(\frac{r}{2}\right)}.
\end{aligned} \tag{10}$$

By (8), $\left\|P^\top \frac{x}{\|x\|}\right\| = \sqrt{\frac{d}{r}}\sqrt{Y}$, hence

$$\mathbb{E}\left[\left\|P^\top \frac{x}{\|x\|}\right\|\right] = \sqrt{\frac{d}{r}}\,\mathbb{E}[\sqrt{Y}] = \sqrt{\frac{d}{r}}\,\frac{\Gamma\left(\frac{r+1}{2}\right)\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d+1}{2}\right)\Gamma\left(\frac{r}{2}\right)}. \tag{11}$$

Substituting (11) into (7) yields

$$\mathbb{E}[g(x)] = \tau\,\|x\|, \qquad \tau \coloneqq \sqrt{\frac{d}{r}}\,\frac{\Gamma\left(\frac{r+1}{2}\right)\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{r}{2}\right)\Gamma\left(\frac{d+1}{2}\right)}. \tag{12}$$

Taking the gradient of (12) and using (6), we conclude that

$$\mathbb{E}\left[\frac{PP^\top x}{\|P^\top x\|}\right] = \nabla\mathbb{E}[g(x)] = \tau\,\frac{x}{\|x\|}.$$

It remains to show $\frac{1}{\sqrt{2}} \le \tau \le 1$. For the upper bound, since $\sqrt{\cdot}$ is concave and Y is Beta with mean $\mathbb{E}[Y] = \frac{r}{d}$,

$$\tau = \sqrt{\frac{d}{r}}\mathbb{E}[\sqrt{Y}] \overset{(a)}{\le} \sqrt{\frac{d}{r}}\sqrt{\mathbb{E}[Y]} = \sqrt{\frac{d}{r}}\sqrt{\frac{r}{d}} = 1. \tag{13}$$

Here, (a) follows from Jensen's inequality. Therefore, $\tau \le 1$. To derive the lower bound on τ, we invoke Wendel's inequality (Wendel, 1948). By Wendel's inequality, for any $x > 0$ and $0 < s < 1$,

$$\left(\frac{x}{x+s}\right)^{1-s} \le \frac{\Gamma(x+s)}{x^s\Gamma(x)} \le 1.$$

Setting $s = \frac{1}{2}$ and multiplying by $\sqrt{x}$, we obtain

$$\sqrt{x}\left(\frac{x}{x+\frac{1}{2}}\right)^{1/2} \le \frac{\Gamma\left(x+\frac{1}{2}\right)}{\Gamma(x)} \le \sqrt{x}. \tag{14}$$

For $x \ge \frac{1}{2}$, the ratio $\frac{x}{x+\frac{1}{2}}$ is minimized at $x = \frac{1}{2}$, which gives $\frac{x}{x+\frac{1}{2}} \ge \frac{1}{2}$. Taking square roots on both sides yields

$$\left(\frac{x}{x+\frac{1}{2}}\right)^{1/2} \ge \frac{1}{\sqrt{2}}. \tag{15}$$

Combining (14) (left inequality) with (15) and taking $x = \frac{r}{2} \ge \frac{1}{2}$ gives

$$\frac{\Gamma\left(\frac{r}{2}+\frac{1}{2}\right)}{\Gamma\left(\frac{r}{2}\right)} \ge \frac{1}{\sqrt{2}}\sqrt{\frac{r}{2}}. \tag{16}$$

Moreover, from (14) (right inequality) we obtain $\Gamma(x)/\Gamma(x+\frac{1}{2}) \ge 1/\sqrt{x}$ for all $x > 0$, hence for $x = \frac{d}{2}$,

$$\frac{\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d}{2}+\frac{1}{2}\right)} \ge \sqrt{\frac{2}{d}}. \tag{17}$$

Using (16) and (17) in the definition of τ in (12), we obtain

$$\begin{aligned}\tau &= \sqrt{\frac{d}{r}}\,\frac{\Gamma\left(\frac{r}{2}+\frac{1}{2}\right)}{\Gamma\left(\frac{r}{2}\right)} \cdot \frac{\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d}{2}+\frac{1}{2}\right)} \\ &\ge \sqrt{\frac{d}{r}} \cdot \frac{1}{\sqrt{2}}\sqrt{\frac{r}{2}} \cdot \sqrt{\frac{2}{d}} = \frac{1}{\sqrt{2}}.\end{aligned} \tag{18}$$

This completes the proof. □

Proposition B.2 (Restatement of Proposition 4.2)**.** *Let S be a $d \times d$ square matrix, $v \neq 0$, and P be a Haar matrix. Then,*

$$\mathbb{E}\left[\frac{v^\top PP^\top SPP^\top v}{\|P^\top v\|^2}\right] = \frac{d(r-1)}{r(d-1)}\,\frac{v^\top Sv}{\|v\|^2} + \frac{d-r}{r(d-1)}\,\mathrm{tr}(S).$$

Proof. Fix a vector $v \in \mathbb{R}^d$. There exists an orthogonal matrix $\hat{Q} \in O(d)$ such that

$$v = \|v\|\hat{Q}e_1,$$

where $e_1 \in \mathbb{R}^d$ denotes the vector whose first component is 1 and whose remaining components are 0.

By the invariance of the Haar matrix, $\hat{Q}P$ has the same distribution as P. Define

$$\hat{S} \coloneqq \hat{Q}^\top S \hat{Q}.$$

Then we have

$$\begin{aligned}
\mathbb{E}\left[\frac{v^\top PP^\top SPP^\top v}{\|P^\top v\|^2}\right] &= \mathbb{E}\left[\frac{v^\top \hat{Q}PP^\top \hat{Q}^\top S\hat{Q}PP^\top \hat{Q}^\top v}{\|P^\top \hat{Q}^\top v\|^2}\right] \\
&= \mathbb{E}\left[\frac{\|v\|e_1^\top PP^\top \hat{Q}^\top S\hat{Q}PP^\top \|v\|e_1}{\|P^\top \|v\|e_1\|^2}\right] \\
&= \mathbb{E}\left[\frac{e_1^\top PP^\top \hat{S}PP^\top e_1}{\|P^\top e_1\|^2}\right]. \qquad (19)
\end{aligned}$$

Define the random vector

$$u \coloneqq \frac{PP^\top e_1}{\|P^\top e_1\|}.$$

Then

$$\begin{aligned}
\mathbb{E}\left[\frac{e_1^\top PP^\top \hat{S}PP^\top e_1}{\|P^\top e_1\|^2}\right] &= \mathbb{E}[u^\top \hat{S}u] \\
&= \mathrm{tr}\Big(\mathbb{E}[uu^\top]\hat{S}\Big), \qquad (20)
\end{aligned}$$

where we used the identity $\mathbb{E}[u^\top \hat{S}u] = \mathbb{E}[\mathrm{tr}(u^\top \hat{S}u)] = \mathbb{E}[\mathrm{tr}(uu^\top \hat{S})] = \mathrm{tr}(\mathbb{E}[uu^\top]\hat{S})$.

Combining (19) and (20), we obtain

$$\mathbb{E}\left[\frac{v^\top PP^\top SPP^\top v}{\|P^\top v\|^2}\right] = \mathrm{tr}(M\hat{S}), \quad M \coloneqq \mathbb{E}[uu^\top].$$

We now characterize the structure of M. Consider the subset

$$\mathcal{S} \coloneqq \{Q \in O(d) \mid Qe_1 = e_1\}.$$

For any $Q \in \mathcal{S}$, by the invariance of the Haar matrix,

$$\begin{aligned}
Q^\top MQ &= \mathbb{E}[Q^\top uu^\top Q] \\
&= \mathbb{E}\left[\frac{Q^\top PP^\top e_1 e_1^\top PP^\top Q}{\|P^\top e_1\|^2}\right] \\
&= \mathbb{E}\left[\frac{Q^\top PP^\top Qe_1 e_1^\top Q^\top PP^\top Q}{\|P^\top Qe_1\|^2}\right] \\
&= \mathbb{E}\left[\frac{(Q^\top P)(Q^\top P)^\top e_1 e_1^\top (Q^\top P)(Q^\top P)^\top}{\|(Q^\top P)^\top e_1\|^2}\right] \\
&= \mathbb{E}\left[\frac{PP^\top e_1 e_1^\top PP^\top}{\|P^\top e_1\|^2}\right] \\
&= \mathbb{E}[uu^\top] \\
&= M.
\end{aligned}$$

This invariance implies strong structural constraints on M. First, fix $i \in \{2, \ldots, d\}$ and define the matrix

$$A_i \coloneqq \mathrm{diag}(1, \ldots, 1, -1, 1, \ldots, 1) \in O(d-1),$$

where the $(i-1)$-th diagonal entry is -1 and the remaining $d-2$ diagonal entries are equal to 1. Define

$$Q \coloneqq \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & A_i \end{pmatrix}.$$

Then $Q \in \mathcal{S}$ by construction.

Since Q has -1 only at the i-th diagonal entry and 1 at all other diagonal entries, we compute $Q^\top M Q$ explicitly. Writing $M = (m_{jk})_{1 \le j,k \le d}$, we have

$$M = Q^\top M Q = \begin{pmatrix} m_{1,1} & \cdots & m_{1,i-1} & -m_{1,i} & m_{1,i+1} & \cdots & m_{1,d} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ m_{i-1,1} & \cdots & m_{i-1,i-1} & -m_{i-1,i} & m_{i-1,i+1} & \cdots & m_{i-1,d} \\ -m_{i,1} & \cdots & -m_{i,i-1} & m_{i,i} & -m_{i,i+1} & \cdots & -m_{i,d} \\ m_{i+1,1} & \cdots & m_{i+1,i-1} & -m_{i+1,i} & m_{i+1,i+1} & \cdots & m_{i+1,d} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ m_{d,1} & \cdots & m_{d,i-1} & -m_{d,i} & m_{d,i+1} & \cdots & m_{d,d} \end{pmatrix}.$$

Since this equality holds for all $i \in \{2, \dots, d\}$, it follows that

$$m_{jk} = 0 \quad \text{for all } j \neq k,$$

and hence M must be a diagonal matrix. Therefore, M can be written as

$$M = \begin{pmatrix} m_{1,1} & 0 & \cdots & 0 \\ 0 & m_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & m_{d,d} \end{pmatrix}.$$

Next, take $i, j \in \{2, \dots, d\}$ with $i \le j$, and let Q be the permutation matrix corresponding to the transposition (i, j), i.e.,

$$Q \coloneqq \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 0 & & 1 & & \\ & & & \ddots & & & \\ & & 1 & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{pmatrix}.$$

The lower-right block of Q is a $(d-1) \times (d-1)$ permutation matrix, and hence belongs to $O(d-1)$. Therefore, $Q \in \mathcal{S}$.

Using again the invariance $Q^\top M Q = M$, we obtain

$$\begin{aligned} M &= Q^\top M Q \\ &= \operatorname{diag}(m_{1,1}, \dots, m_{i-1,i-1}, m_{j,j}, m_{i+1,i+1}, \dots, m_{j-1,j-1}, m_{i,i}, m_{j+1,j+1}, \dots, m_{d,d}). \end{aligned}$$

Since this holds for arbitrary $i, j \in \{2, \dots, d\}$ with $i \le j$, we conclude that

$$m_{2,2} = m_{3,3} = \cdots = m_{d,d}.$$

Defining $\alpha \coloneqq m_{1,1}$ and $\beta \coloneqq m_{2,2}$, we finally obtain

$$M = \begin{pmatrix} \alpha & 0 & \cdots & 0 \\ 0 & \beta & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \beta \end{pmatrix} = (\alpha - \beta) e_1 e_1^\top + \beta I_d.$$

We now determine α and β. First,

$$\begin{aligned} e_1^\top M e_1 &= \mathbb{E}[e_1^\top u u^\top e_1] \\ &= \mathbb{E}\left[e_1^\top \frac{PP^\top e_1 e_1^\top PP^\top}{\|P^\top e_1\|^2} e_1\right] \\ &= \mathbb{E}[\|P^\top e_1\|^2] \\ &= e_1^\top \mathbb{E}[PP^\top] e_1 \\ &= 1. \end{aligned}$$

On the other hand,

$$e_1^\top M e_1 = \alpha,$$

and hence $\alpha = 1$.

Next, we compute the trace of M:

$$\begin{aligned} \mathrm{tr}(M) &= \mathrm{tr}(\mathbb{E}[uu^\top]) \\ &= \mathbb{E}[\mathrm{tr}(uu^\top)] \\ &= \mathbb{E}[\|u\|^2] \\ &= \mathbb{E}\left[\frac{e_1^\top PP^\top PP^\top e_1}{\|P^\top e_1\|^2}\right] \\ &= \frac{d}{r}. \end{aligned}$$

On the other hand,

$$\begin{aligned} \mathrm{tr}(M) &= \mathrm{tr}((1-\beta)e_1 e_1^\top + \beta I_d) \\ &= (1-\beta) + \beta d. \end{aligned}$$

Equating the two expressions yields

$$(d-1)\beta = \frac{d}{r} - 1, \quad \beta = \frac{d-r}{r(d-1)}.$$

Therefore,

$$M = \frac{d(r-1)}{r(d-1)} e_1 e_1^\top + \frac{d-r}{r(d-1)} I_d.$$

Finally,

$$\begin{aligned} \mathbb{E}\left[\frac{v^\top PP^\top SPP^\top v}{\|P^\top v\|^2}\right] &= \mathrm{tr}(M\hat{S}) \\ &= \mathrm{tr}\left(\frac{d(r-1)}{r(d-1)} e_1 e_1^\top \hat{S} + \frac{d-r}{r(d-1)} \hat{S}\right) \\ &= \frac{d(r-1)}{r(d-1)} e_1^\top \hat{S} e_1 + \frac{d-r}{r(d-1)} \mathrm{tr}(\hat{S}) \\ &= \frac{d(r-1)}{r(d-1)} \frac{v^\top S v}{\|v\|^2} + \frac{d-r}{r(d-1)} \mathrm{tr}(S), \end{aligned}$$

which completes the proof. □

Theorem B.3 (Restatement of Theorem 4.4)**.** *Suppose that F satisfies Assumptions 2.1 and 2.2, and that the gradient satisfies Assumptions 2.3 and 2.4. Let $\{x_k\}_{k\geq 0}$ be the sequence generated by the RS-NSGD iteration. We set*

$$\bar{B} = \lceil \max\{1,\, BT^q\} \rceil, \qquad \bar{\eta} = \eta T^{-u},$$

for parameters $B > 0$, $q > 0$, $\eta > 0$, and $u \in (0,1)$. Then it holds that

$$\begin{aligned} \frac{1}{T}\sum_{k=0}^{T-1} \mathbb{E}[\|\nabla F(x_k)\|] &\leq \frac{\Delta_0}{\tau \eta T^{1-u}} \\ &+ \frac{\eta \ell(\mathbb{L}) \|\mathbb{L}\|}{2\tau T^u} + \frac{4\sigma}{\max\{1, BT^q\}^{(p-1)/p}}. \end{aligned}$$

Proof. Define

$$G_k \coloneqq \begin{cases} 0, & \|P_k^\top g_k\| = 0, \\ \dfrac{P_k P_k^\top g_k}{\|P_k^\top g_k\|}, & \|P_k^\top g_k\| \neq 0, \end{cases} \qquad x_{k+1} = x_k - \bar{\eta} G_k.$$

We use the almost sure implications

$$\|g_k\| = 0 \;\Rightarrow\; \|P_k^\top g_k\| = 0 \quad \text{a.s.}, \qquad \|g_k\| \neq 0 \;\Rightarrow\; \|P_k^\top g_k\| \neq 0 \quad \text{a.s.}$$

so that on $\{\|g_k\| \neq 0\}$ we have $G_k = \frac{P_k P_k^\top g_k}{\|P_k^\top g_k\|}$ a.s., and on $\{\|g_k\| = 0\}$ we have $G_k = 0$ a.s.

By (3), we have

$$\begin{aligned} F(x_{k+1}) &\le F(x_k) + \nabla F(x_k)^\top (x_{k+1} - x_k) + \frac{1}{2}(x_{k+1} - x_k)^\top \mathbb{L}(x_{k+1} - x_k) \\ &= F(x_k) - \bar{\eta}\, \nabla F(x_k)^\top G_k + \frac{\bar{\eta}^2}{2}\, G_k^\top \mathbb{L} G_k. \end{aligned}$$

Rearranging both sides and summing over $k = 0, \dots, T-1$, we obtain

$$\begin{aligned} \bar{\eta} \sum_{k=0}^{T-1} \nabla F(x_k)^\top G_k &\le \sum_{k=0}^{T-1} \big(F(x_k) - F(x_{k+1})\big) + \sum_{k=0}^{T-1} \frac{\bar{\eta}^2}{2}\, G_k^\top \mathbb{L} G_k \\ &= F(x_0) - F(x_T) + \sum_{k=0}^{T-1} \frac{\bar{\eta}^2}{2}\, G_k^\top \mathbb{L} G_k \\ &\le F(x_0) - F_* + \sum_{k=0}^{T-1} \frac{\bar{\eta}^2}{2}\, G_k^\top \mathbb{L} G_k \\ &\le \Delta_0 + \sum_{k=0}^{T-1} \frac{\bar{\eta}^2}{2}\, G_k^\top \mathbb{L} G_k. \end{aligned} \tag{21}$$

Let

$$\mathcal{F}_k \coloneqq \sigma(g_0, P_0, \dots, g_{k-1}, P_{k-1}), \qquad \mathcal{F}_k^+ \coloneqq \sigma(g_0, P_0, \dots, g_{k-1}, P_{k-1}, g_k),$$

so that $\mathcal{F}_k \subset \mathcal{F}_k^+$. By the tower property,

$$\begin{aligned} \mathbb{E}\big[G_k^\top \mathbb{L} G_k\big] &= \mathbb{E}\big[\mathbb{E}\big[G_k^\top \mathbb{L} G_k \;\big|\; \mathcal{F}_k^+\big]\big] \\ &\overset{(a)}{=} \mathbb{E}\left[\mathbf{1}_{\{\|g_k\| \neq 0\}} \left(\frac{d(r-1)}{r(d-1)} \frac{g_k^\top \mathbb{L} g_k}{\|g_k\|^2} + \frac{d-r}{r(d-1)} \operatorname{tr}(\mathbb{L})\right)\right] \\ &\overset{(b)}{\le} \ell(\mathbb{L}) \|\mathbb{L}\|. \end{aligned} \tag{22}$$

Here, (a) follows from Proposition B.2 on the event $\{\|g_k\| \neq 0\}$ (using $\|g_k\| \neq 0 \Rightarrow \|P_k^\top g_k\| \neq 0$ a.s., hence $G_k = \frac{P_k P_k^\top g_k}{\|P_k^\top g_k\|}$ a.s.), and from $G_k = 0$ a.s. on $\{\|g_k\| = 0\}$. For (b), note that on $\{\|g_k\| \neq 0\}$,

$$\frac{g_k^\top \mathbb{L} g_k}{\|g_k\|^2} \le \|\mathbb{L}\|,$$

and the term is multiplied by $\mathbf{1}_{\{\|g_k\| \neq 0\}} \le 1$; together with the definition of $\ell(\mathbb{L})$ this yields (22).

To justify the application of the tower property, we verify integrability. We have

$$\big|\nabla F(x_k)^\top G_k\big| \le \|\nabla F(x_k)\|\, \|G_k\|.$$

Moreover, if $\|P_k^\top g_k\| = 0$ then $G_k = 0$ and $\|G_k\| = 0$. If $\|P_k^\top g_k\| \neq 0$, then

$$\|G_k\| = \frac{\|P_k P_k^\top g_k\|}{\|P_k^\top g_k\|} \le \|P_k\| \le \sqrt{\frac{d}{r}},$$

and hence in all cases,

$$\left|\nabla F(x_k)^\top G_k\right| \le \sqrt{\frac{d}{r}}\|\nabla F(x_k)\|.$$

Moreover, by smoothness,

$$\begin{aligned}\|\nabla F(x_{k+1})\| &\le \|\nabla F(x_k)\| + L\|x_{k+1} - x_k\| \\ &= \|\nabla F(x_k)\| + \bar{\eta}L\|G_k\| \le \|\nabla F(x_k)\| + \bar{\eta}L\sqrt{\frac{d}{r}},\end{aligned}$$

and hence by induction,

$$\|\nabla F(x_k)\| \le \|\nabla F(x_0)\| + k\bar{\eta}L\sqrt{\frac{d}{r}}.$$

This implies integrability.

We may therefore apply the tower property to obtain

$$\mathbb{E}\big[\nabla F(x_k)^\top G_k \mid \mathcal{F}_k\big] = \mathbb{E}\big[\mathbb{E}\big[\nabla F(x_k)^\top G_k \mid \mathcal{F}_k^+\big] \;\big|\; \mathcal{F}_k\big]. \tag{23}$$

We claim that,

$$\mathbb{E}\big[\nabla F(x_k)^\top G_k \mid \mathcal{F}_k^+\big] \ge \tau\Big(\|\nabla F(x_k)\| - 2\|g_k - \nabla F(x_k)\|\Big). \tag{24}$$

To prove this, we consider two cases.

Case 1: $\|g_k\| = 0$**.** Then $\|P_k^\top g_k\| = 0$ almost surely, hence $G_k = 0$ and thus

$$\mathbb{E}\big[\nabla F(x_k)^\top G_k \mid \mathcal{F}_k^+\big] = 0.$$

On the other hand, since $g_k = 0$,

$$\|\nabla F(x_k)\| - 2\|g_k - \nabla F(x_k)\| = \|\nabla F(x_k)\| - 2\|\nabla F(x_k)\| = -\|\nabla F(x_k)\|.$$

Therefore the right-hand side of (24) equals $-\tau\|\nabla F(x_k)\| \le 0$, and the inequality holds.

Case 2: $\|g_k\| \neq 0$**.** Then $\|P_k^\top g_k\| \neq 0$ almost surely, so $G_k = \frac{P_k P_k^\top g_k}{\|P_k^\top g_k\|}$ almost surely, and by Proposition B.1,

$$\mathbb{E}\big[G_k \mid \mathcal{F}_k^+\big] = \tau\,\frac{g_k}{\|g_k\|}.$$

Hence,

$$\mathbb{E}\big[\nabla F(x_k)^\top G_k \mid \mathcal{F}_k^+\big] = \nabla F(x_k)^\top \mathbb{E}\big[G_k \mid \mathcal{F}_k^+\big] = \tau\,\frac{\nabla F(x_k)^\top g_k}{\|g_k\|}.$$

Moreover, for any $a, b \in \mathbb{R}^d$ with $b \neq 0$, the elementary inequality (see, e.g., Hübler et al. (2025))

$$\frac{a^\top b}{\|b\|} \ge \|a\| - 2\|a - b\|$$

holds. Applying it with $a = \nabla F(x_k)$ and $b = g_k$ yields

$$\frac{\nabla F(x_k)^\top g_k}{\|g_k\|} \ge \|\nabla F(x_k)\| - 2\|g_k - \nabla F(x_k)\|.$$

Multiplying by τ proves (24) in this case.

Therefore, combining (23) and (24), we obtain

$$\begin{aligned}\mathbb{E}\big[\nabla F(x_k)^\top G_k \mid \mathcal{F}_k\big] &= \mathbb{E}\big[\mathbb{E}\big[\nabla F(x_k)^\top G_k \mid \mathcal{F}_k^+\big] \;\big|\; \mathcal{F}_k\big] \\ &\ge \tau\Big(\|\nabla F(x_k)\| - 2\,\mathbb{E}[\|g_k - \nabla F(x_k)\| \mid \mathcal{F}_k]\Big).\end{aligned} \tag{25}$$

Following the same argument as in Hübler et al. (2025), Assumption 2.4 implies

$$\mathbb{E}[\|g_k - \nabla F(x_k)\| \mid \mathcal{F}_k] \leq \frac{2\sigma}{\bar{B}^{(p-1)/p}}.$$

Substituting this bound into (25), we obtain

$$\mathbb{E}\big[\nabla F(x_k)^\top G_k \mid \mathcal{F}_k\big] \geq \tau \left(\|\nabla F(x_k)\| - \frac{4\sigma}{\bar{B}^{(p-1)/p}} \right). \tag{26}$$

Taking expectations of both sides and using the tower property, we obtain

$$\mathbb{E}\big[\nabla F(x_k)^\top G_k\big] \geq \tau \left(\mathbb{E}[\|\nabla F(x_k)\|] - \frac{4\sigma}{\bar{B}^{(p-1)/p}} \right). \tag{27}$$

Combining (21) and (22), taking expectations, and using (27), we obtain

$$\bar{\eta}\tau \sum_{k=0}^{T-1} \left(\mathbb{E}[\|\nabla F(x_k)\|] - \frac{4\sigma}{\bar{B}^{(p-1)/p}} \right) \leq \Delta_0 + \frac{\bar{\eta}^2 \ell(\mathbb{L}) \|\mathbb{L}\| T}{2}.$$

Dividing both sides by $\bar{\eta}\tau T$, we obtain

$$\frac{1}{T} \sum_{k=0}^{T-1} \mathbb{E}[\|\nabla F(x_k)\|] \leq \frac{\Delta_0}{\bar{\eta}\tau T} + \frac{\bar{\eta}\, \ell(\mathbb{L}) \|\mathbb{L}\|}{2\tau} + \frac{4\sigma}{\bar{B}^{(p-1)/p}}. \tag{28}$$

Recall that $\bar{\eta} = \eta T^{-u}$ and $\bar{B} = \lceil \max\{1, BT^q\} \rceil$. Moreover, since

$$\bar{B} \geq \max\{1, BT^q\},$$

it follows that

$$\bar{B}^{(p-1)/p} \geq \max\{1, BT^q\}^{(p-1)/p}.$$

Substituting these relations into (28), we obtain

$$\frac{1}{T} \sum_{k=0}^{T-1} \mathbb{E}[\|\nabla F(x_k)\|] \leq \frac{\Delta_0}{\tau \eta T^{1-u}} + \frac{\eta\, \ell(\mathbb{L}) \|\mathbb{L}\|}{2\tau T^u} + \frac{4\sigma}{\max\{1, BT^q\}^{(p-1)/p}}.$$

□

Oracle complexity (parameters, including p, are unknown). From Theorem 4.4, taking $(u, q) = \left(\frac{1}{2}, 1\right)$ and treating η and B as constants, the iteration complexity satisfies

$$\mathcal{O}\left(\left(\frac{\Delta_0}{\tau\epsilon} \right)^2 + \left(\frac{\ell(\mathbb{L})\, \|\mathbb{L}\|}{\tau\epsilon} \right)^2 + \left(\frac{\sigma}{\epsilon} \right)^{\frac{p}{p-1}} \right).$$

Recalling that the oracle complexity is given by $r\bar{B}T$, we obtain

$$\mathcal{O}\left(r \left(\frac{\Delta_0}{\tau\epsilon} \right)^4 + r \left(\frac{\ell(\mathbb{L})\, \|\mathbb{L}\|}{\tau\epsilon} \right)^4 + r \left(\frac{\sigma}{\epsilon} \right)^{\frac{2p}{p-1}} \right).$$

Oracle complexity (parameters, including p, are known). From Theorem 4.4, we set $u = \frac{1}{2}$ and choose

$$q = \frac{p}{2(p-1)}, \qquad \eta = \sqrt{\frac{\Delta_0}{\ell(\mathbb{L})\, \|\mathbb{L}\|}}, \qquad B = \left(\frac{\sigma\, \tau}{\sqrt{\Delta_0\, \ell(\mathbb{L})\, \|\mathbb{L}\|}} \right)^{\frac{p}{p-1}}.$$

With these choices, the iteration complexity satisfies

$$\mathcal{O}\left(\left(\frac{\sqrt{\ell(\mathbb{L})\, \Delta_0\, \|\mathbb{L}\|}}{\tau\epsilon} \right)^2 \right).$$

Recalling that the oracle complexity is given by $r\bar{B}T$, we obtain

$$\mathcal{O}\left(r\, \frac{\ell(\mathbb{L})}{\tau^2}\, \frac{\Delta_0\, \|\mathbb{L}\|}{\epsilon^2}\, \left(\frac{\sigma}{\epsilon} \right)^{\frac{p}{p-1}} \right).$$

Lemma B.4. *Suppose that F satisfies Assumptions 2.1 and 2.2. Then*

$$\|\nabla F(x_0)\| \;\le\; \sqrt{2\Delta_0\, L}.$$

Proof. By L-smoothness, for any x and v,

$$F(x+v) \le F(x) + \langle \nabla F(x), v\rangle + \frac{L}{2}\|v\|^2.$$

Choosing $x = x_0$ and $v = -L^{-1}\nabla F(x_0)$ yields

$$F_* \le F\left(x_0 - \frac{1}{L}\nabla F(x_0)\right) \le F(x_0) - \frac{1}{2L}\|\nabla F(x_0)\|^2.$$

Rearranging proves the claim. □

Theorem B.5 (Restatement of Theorem 4.5)**.** *Suppose that F satisfies Assumptions 2.1 and 2.2, and that the gradient satisfies Assumptions 2.3 and 2.4. Let $\{x_k\}_{k\ge 0}$ be the sequence generated by the RS-NSGD iteration. We set*

$$\bar{B} = \lceil \max\{1,\, BT^q\} \rceil, \qquad \bar{\eta} = \eta T^{-u},$$

for parameters $B > 0$, $q > 0$, $\eta > 0$, and $u \in (0,1)$. Then, with probability at least $1-\delta$, the following holds:

$$\begin{aligned}\frac{1}{T}\sum_{k=0}^{T-1}\|\nabla F(x_k)\| \le \frac{2\Delta_0}{\tau\eta T^{1-u}} + \frac{\frac{d}{r}\eta L}{\tau T^u}\big(1 + 12\frac{\sqrt{\frac{d}{r}}}{\tau}\log(1/\delta)\big) \\ +17\frac{\frac{d}{r}\sqrt{\Delta_0 L}}{\tau^2 T}\log(1/\delta) + \frac{8\sigma}{\max\{1, BT^q\}^{\frac{p-1}{p}}}.\end{aligned}$$

Proof. Define the filtrations

$$\mathcal{F}_k \coloneqq \sigma(g_0, P_0, \ldots, g_{k-1}, P_{k-1}), \qquad \mathcal{F}_k^+ \coloneqq \sigma(g_0, P_0, \ldots, g_{k-1}, P_{k-1}, g_k).$$

Define the normalized projected direction

$$G_k \coloneqq \begin{cases} 0, & \|P_k^\top g_k\| = 0, \\ \dfrac{P_k P_k^\top g_k}{\|P_k^\top g_k\|}, & \|P_k^\top g_k\| \ne 0, \end{cases}$$

and the RS-NSGD update

$$x_{k+1} = x_k - \bar{\eta} G_k.$$

Define

$$\psi_k \coloneqq \nabla F(x_k)^\top G_k, \qquad \phi_k \coloneqq \mathbb{E}[\psi_k \mid \mathcal{F}_k].$$

We use the following almost sure implications:

$$\|g_k\| = 0 \;\Rightarrow\; \|P_k^\top g_k\| = 0 \quad \text{a.s.}, \qquad \|g_k\| \ne 0 \;\Rightarrow\; \|P_k^\top g_k\| \ne 0 \quad \text{a.s.}$$

By L-smoothness,

$$F(x_{k+1}) \le F(x_k) - \bar{\eta}\nabla F(x_k)^\top G_k + \frac{L}{2}\bar{\eta}^2\|G_k\|^2.$$

Summing this inequality for $k = 0, \ldots, T-1$ and using $F(x_T) \ge F_*$ yields

$$\bar{\eta}\sum_{k=0}^{T-1}\psi_k \le \Delta_0 + \frac{L}{2}\bar{\eta}^2\sum_{k=0}^{T-1}\|G_k\|^2. \tag{29}$$

If $\|P_k^\top g_k\| \ne 0$, then

$$\|G_k\| = \frac{\|P_k P_k^\top g_k\|}{\|P_k^\top g_k\|} \le \|P_k\| \le \sqrt{\frac{d}{r}}.$$

If $\|P_k^\top g_k\| = 0$, then $G_k = 0 \le \sqrt{\frac{d}{r}}$. Therefore $\|G_k\|^2 \le \frac{d}{r}$ for all k, and

$$\bar{\eta} \sum_{k=0}^{T-1} \psi_k \le \Delta_0 + \frac{\bar{\eta}^2 L}{2} \frac{d}{r} T. \tag{30}$$

Define $D_0 \coloneqq 0$ and, for $k \ge 1$,

$$D_k \coloneqq -\bar{\eta}(\psi_{k-1} - \phi_{k-1}).$$

Then $(D_k, \mathcal{F}_k)$ is a martingale difference sequence. Let

$$S_k \coloneqq 2\sqrt{\frac{d}{r}}\, \bar{\eta}\, \|\nabla F(x_{k-1})\|, \qquad k \ge 1.$$

Then S_k is $\mathcal{F}_{k-1}$-measurable. Moreover,

$$|\psi_{k-1}| \le \|\nabla F(x_{k-1})\|\, \|G_{k-1}\| \le \sqrt{\frac{d}{r}} \|\nabla F(x_{k-1})\|,$$

which also implies $|\phi_{k-1}| \le \sqrt{\frac{d}{r}} \|\nabla F(x_{k-1})\|$. Hence

$$\exp\left(\frac{D_k^2}{S_k^2}\right) = \exp\left(\frac{(\psi_{k-1} - \phi_{k-1})^2}{4\frac{d}{r}\|\nabla F(x_{k-1})\|^2}\right) \le \exp\left(\frac{(|\psi_{k-1}| + |\phi_{k-1}|)^2}{4\frac{d}{r}\|\nabla F(x_{k-1})\|^2}\right) \le e.$$

By Lemma 11 of Hübler et al. (2025)[7], for any $\lambda > 0$ and any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\sum_{k=1}^{T} D_k \le \frac{3}{4}\lambda \sum_{k=1}^{T} S_k^2 + \frac{1}{\lambda} \log \frac{1}{\delta}. \tag{31}$$

Combining (31) with (30) yields

$$\bar{\eta} \sum_{k=0}^{T-1} \Big(\phi_k - 3\frac{d}{r}\lambda\bar{\eta}\|\nabla F(x_k)\|^2\Big) \le \Delta_0 + \frac{\bar{\eta}^2 L}{2}\frac{d}{r}T + \frac{1}{\lambda}\log\frac{1}{\delta}. \tag{32}$$

Let $C_T \coloneqq \bar{\eta}^2 T$ and choose

$$\lambda \coloneqq \frac{\tau}{6\big(\bar{\eta}\frac{d}{r}\|\nabla F(x_0)\| + C_T L \big(\frac{d}{r}\big)^{3/2}\big)}.$$

Using the gradient growth bound

$$\|\nabla F(x_k)\| \le \|\nabla F(x_0)\| + k\bar{\eta}L\sqrt{\frac{d}{r}},$$

we obtain

$$\begin{aligned} 3\frac{d}{r}\lambda\bar{\eta}\|\nabla F(x_k)\|^2 &= \frac{\tau 3\frac{d}{r}\bar{\eta}\|\nabla F(x_k)\|^2}{6\big(\bar{\eta}\frac{d}{r}\|\nabla F(x_0)\| + C_T L(\frac{d}{r})^{3/2}\big)} \\ &\overset{(b)}{\le} \frac{\tau\frac{d}{r}\bar{\eta}(\|\nabla F(x_0)\| + k\bar{\eta}L\sqrt{\frac{d}{r}})}{2\big(\bar{\eta}\frac{d}{r}\|\nabla F(x_0)\| + C_T L(\frac{d}{r})^{3/2}\big)}\|\nabla F(x_k)\| \\ &\overset{(c)}{\le} \frac{\tau}{2}\|\nabla F(x_k)\|. \end{aligned}$$

[7] To avoid division by zero and to make the argument rigorous, one may regularize the denominator by replacing S_k with $S_k + \frac{1}{n}$ for $n \in \mathbb{N}$. This yields a monotone decreasing sequence of events indexed by n, to which Lemma 11 applies for each fixed n. Letting $n \to \infty$ and using the continuity of probability yields the stated inequality.

Here, (b) uses

$$\|\nabla F(x_k)\|^2 \le \big(\|\nabla F(x_0)\| + k\bar{\eta}L\sqrt{\tfrac{d}{r}}\big)\|\nabla F(x_k)\|,$$

and (c) follows from

$$\frac{d}{r}\bar{\eta}k\bar{\eta}L\sqrt{\tfrac{d}{r}} \le C_T L\big(\tfrac{d}{r}\big)^{3/2}.$$

As shown in (26) of Theorem B.3,

$$\phi_k \ge \tau\left(\|\nabla F(x_k)\| - \frac{4\sigma}{\bar{B}^{\frac{p-1}{p}}}\right).$$

Substituting these bounds into (32) and multiplying both sides by $\frac{2}{\tau\bar{\eta}T}$ yield

$$\frac{1}{T}\sum_{k=0}^{T-1}\|\nabla F(x_k)\| \le \frac{2\Delta_0}{\tau\bar{\eta}T} + \frac{\bar{\eta}L}{\tau}\frac{d}{r} + \frac{12\frac{d}{r}\|\nabla F(x_0)\|}{\tau^2 T}\log\frac{1}{\delta} + \frac{12\bar{\eta}L}{\tau^2}\Big(\frac{d}{r}\Big)^{3/2}\log\frac{1}{\delta} + \frac{8\sigma}{\bar{B}^{\frac{p-1}{p}}}.$$

Finally, using Lemma B.4, $\|\nabla F(x_0)\| \le \sqrt{2\Delta_0 L}$, and $12\sqrt{2} \le 17$, together with the definitions of $\bar{\eta}$ and $\bar{B}$, the stated bound follows. □

Oracle complexity (parameters, including p, are unknown). From Theorem 4.5, taking $(u, q) = \big(\frac{1}{2}, 1\big)$ and treating η and B as constants, the iteration complexity satisfies

$$\tilde{\mathcal{O}}\left(\left(\frac{\Delta_0}{\tau\epsilon}\right)^2 + \left(\frac{\big(\frac{d}{r}\big)^{\frac{3}{2}}L}{\tau^2\epsilon}\right)^2 + \Big(\frac{\sigma}{\epsilon}\Big)^{\frac{p}{p-1}}\right).$$

Recalling that the oracle complexity is given by $r\bar{B}T$, we obtain

$$\tilde{\mathcal{O}}\left(r\left(\frac{\Delta_0}{\tau\epsilon}\right)^4 + \frac{d^6}{r^5}\left(\frac{L}{\tau^2\epsilon}\right)^4 + r\left(\frac{\sigma}{\epsilon}\right)^{\frac{2p}{p-1}}\right).$$

Oracle complexity (parameters, including p, are known). From Theorem 4.5, we set $u = \frac{1}{2}$ and choose

$$q = \frac{p}{2(p-1)}, \qquad \eta = \Big(\frac{r}{d}\Big)^{\frac{3}{4}}\sqrt{\frac{\Delta_0\tau}{L}}, \qquad B = \left(\frac{\sigma\tau^{\frac{3}{2}}}{\sqrt{\Delta_0 L}}\Big(\frac{r}{d}\Big)^{\frac{3}{4}}\right)^{\frac{p}{p-1}}.$$

With these choices, the iteration complexity satisfies

$$\tilde{\mathcal{O}}\left(\left(\frac{\big(\frac{d}{r}\big)^{\frac{3}{4}}\sqrt{\Delta_0 L}}{\tau^{\frac{3}{2}}\epsilon}\right)^2\right).$$

Recalling that the oracle complexity is given by $r\bar{B}T$, we obtain

$$\tilde{\mathcal{O}}\left(\left(\frac{d}{r^{1/3}\tau^2}\right)^{3/2}\frac{\Delta_0 L}{\epsilon^2}\Big(\frac{\sigma}{\epsilon}\Big)^{\frac{p}{p-1}}\right).$$

C HIGH-PROBABILITY CONVERGENCE OF RS-SGD

This appendix provides a single-run high-probability guarantee for RS-SGD under sub-Gaussian noise. The result complements the in-expectation analysis of Flynn et al. (2024) and is based on a mini-batch extension of the high-probability framework of Liu et al. (2023).

We first recall the light-tailed noise assumptions used for the SGD and RS-SGD baseline results, then state a mini-batch sub-Gaussian lemma, and then give the main high-probability guarantee and its oracle-complexity consequence before turning to the proof. We state these results before the proof to separate the guarantees from the technical estimates and to make the proof structure easier to follow. The proof first establishes the mini-batch sub-Gaussian bound for the stochastic gradient estimator and then combines it with an argument controlling the randomized-subspace updates in the RS-SGD iterates.

C.1 Light-tailed noise assumptions

For completeness, we recall the light-tailed noise assumptions used for the SGD and RS-SGD baseline results.

Assumption C.1 (Sub-Gaussian)**.** *The gradient satisfies a σ-sub-Gaussian tail condition, i.e., for all $x \in \mathbb{R}^d$ and all $\lambda \in \mathbb{R}$ with $|\lambda| \le 1/\sigma$,*

$$\mathbb{E}\big[\exp\big(\lambda^2\|\nabla f(x,\xi) - \nabla F(x)\|^2\big)\big] \le \exp(\lambda^2\sigma^2). \tag{33}$$

Assumption C.2 (BV)**.** *The gradient has bounded variance, i.e., for all $x \in \mathbb{R}^d$,*

$$\mathbb{E}\big[\|\nabla f(x,\xi) - \nabla F(x)\|^2\big] \le \sigma^2.$$

These assumptions are related to Assumption 2.4 as follows:

$$\text{Assumption C.1} \ \subset \ \text{Assumption C.2} \ \subset \ \text{Assumption 2.4}.$$

Indeed, Assumption C.1 implies Assumption C.2 by Jensen's inequality. Moreover, Assumption C.2 is Assumption 2.4 with $p = 2$, whereas for $p < 2$, Assumption 2.4 permits stochastic-gradient noise with infinite second moment. If Assumption 2.4 holds for some $p \in (1, 2]$, then it also holds for any $q \in (1, p]$ by Jensen's inequality. Hence, Assumption C.2 implies Assumption 2.4 for every $p \in (1, 2]$. Therefore, Assumption 2.4 is weaker than Assumption C.2, and Assumption C.2 is weaker than Assumption C.1.

C.2 Main Result

We now state the high-probability result for RS-SGD. The following lemma controls the mini-batch stochastic gradient estimator under sub-Gaussian noise.

Lemma C.3. *Assume that the stochastic gradient $\nabla f(x,\xi)$ satisfies Assumptions 2.3 and C.1. For a fixed $x \in \mathbb{R}^d$, define the mini-batch gradient estimator*

$$g(x) := \frac{1}{\bar{B}} \sum_{j=1}^{\bar{B}} \nabla f(x, \xi^j),$$

and set

$$\bar{\sigma} := 24\sqrt{\frac{e(e+1)}{2}}\,\frac{\sigma}{\sqrt{\bar{B}}}.$$

Then for all λ satisfying $|\lambda| \le 1/\bar{\sigma}$, it holds that

$$\mathbb{E}\Big[\exp\Big(\lambda^2\big\|g(x) - \nabla F(x)\big\|^2\Big)\Big] \le 5^d \exp\big(\lambda^2\bar{\sigma}^2\big).$$

By Lemma C.3, increasing the mini-batch size reduces the noise, and the following theorem establishes convergence.

Theorem C.4. *Suppose that Assumptions 2.1 to 2.3 and C.1 hold. Let $\{x_k\}_{k\ge 0}$ be the sequence generated by RS-SGD. Let $\bar{B} := \lceil \max\{1, BT^q\} \rceil$ for $B > 0$ and $q > 0$, $\alpha := 24\sqrt{\frac{e(e+1)}{2}}$, and $\bar{\eta} = \frac{r}{Ld}$. Further define*

$$\mu := \begin{cases} 1 - I_{\frac{r}{2d}}\left(\frac{r}{2}, \frac{d-r}{2}\right), & r < d, \\ 1, & r = d. \end{cases} \tag{34}$$

where $I_x(a,b)$ denotes the regularized incomplete Beta function. Then for any $\delta \in (0,1)$ and any T such that $2\log\big(\frac{2}{\delta}\big) < \mu T$, the following holds with probability at least $1 - \delta$:

$$\min_{0\le k\le T-1} \|\nabla F(x_k)\|^2 \le \frac{4L\frac{d}{r}\Delta_0}{\mu T - \sqrt{2\mu T\log\big(\frac{2}{\delta}\big)}} + \frac{2\frac{d}{r}\left(\frac{\alpha^2\sigma^2}{\bar{B}}\right)\big(d\log 5 + 1\big)}{\mu - \sqrt{2\frac{\mu}{T}\log\big(\frac{2}{\delta}\big)}} + \frac{2\frac{d}{r}\left(\frac{\alpha^2\sigma^2}{\bar{B}}\right)\log\big(\frac{2}{\delta}\big)}{\mu T - \sqrt{2\mu T\log\big(\frac{2}{\delta}\big)}}.$$

For μ defined in (34), for all $1 \le r \le d$, $\frac{1}{12} \le \mu \le 1$.[8]

From the result of Theorem C.4, we can derive the oracle complexity required to achieve an ϵ-stationary point:

$$\min_{0 \le k \le T-1} \|\nabla F(x_k)\| \le \epsilon. \tag{35}$$

Take $q = 1$ and $B = \frac{d\sigma^2}{L\Delta_0}$ for $\bar{B}$. Then the oracle complexity is

$$\tilde{\mathcal{O}}\left(\frac{d^3}{\mu^2 r}\,\Delta_0 L\sigma^2\,\epsilon^{-4}\right).$$

Expectation bounds capture average behavior but not the likelihood of large deviations in one run. Our high-probability analysis quantifies this risk. To the best of our knowledge, Theorem C.4 provides the first high-probability convergence guarantee for RS-SGD under the sub-Gaussian noise assumption. Although the constants become larger compared to in-expectation bounds, our result matches the ϵ-dependence (ϵ^{-4}) implied by the in-expectation analysis of Flynn et al. (2024) for achieving (35); it also matches the standard ϵ^{-4} high-probability rate of SGD under sub-Gaussian noise (Liu et al., 2023).

C.3 Proofs

Definition C.5. *Let X be an $\mathbb{R}^d$-valued random vector. We say that $\|X\|$ is σ-sub-Gaussian if*

$$\forall |\lambda| \le \frac{1}{\sigma}, \qquad \mathbb{E}\big[\exp\big(\lambda^2\|X\|^2\big)\big] \le \exp\big(\lambda^2\sigma^2\big).$$

We study how the parameter scales under mini-batching.

Lemma C.6. *Assume $\mathbb{E}[X] = 0$ and $\|X\|$ is σ-sub-Gaussian in the sense of Definition C.5. Then for all $u \in \mathbb{R}^d$,*

$$\mathbb{E}\big[\exp(\langle u, X\rangle)\big] \le \exp\left(\frac{e+1}{2}\,\|u\|^2\sigma^2\right).$$

Proof. We borrow ideas from the proof of Proposition 2.6.1 in Vershynin (2025) for the one-dimensional case, adapting them to our setting involving $\|X\|$.

Fix $u \in \mathbb{R}^d$. We start with

$$e^{\langle u,X\rangle} \overset{(a)}{\le} 1 + \langle u, X\rangle + \frac{\langle u, X\rangle^2}{2} e^{|\langle u,X\rangle|} \overset{(b)}{\le} 1 + \langle u, X\rangle + \frac{\|u\|^2\|X\|^2}{2} e^{\|u\|\|X\|}. \tag{36}$$

Here (a) is the inequality $e^x \le 1 + x + \frac{x^2}{2}e^{|x|}$ for all $x \in \mathbb{R}$, and (b) uses $|\langle u, X\rangle| \le \|u\|\,\|X\|$ and $\langle u, X\rangle^2 \le \|u\|^2\|X\|^2$.

Next,

$$e^{\|u\|\|X\|} \overset{(a)}{\le} \exp\left(\frac{\sigma^2\|u\|^2}{2} + \frac{\|X\|^2}{2\sigma^2}\right), \tag{37}$$

where (a) follows from $ab \le \frac{\sigma^2 a^2}{2} + \frac{b^2}{2\sigma^2}$ applied to $a = \|u\|$ and $b = \|X\|$.

From (36) and (37), we obtain

$$\begin{aligned} e^{\langle u,X\rangle} &\le 1 + \langle u, X\rangle + \frac{\|u\|^2\|X\|^2}{2} \exp\left(\frac{\sigma^2\|u\|^2}{2} + \frac{\|X\|^2}{2\sigma^2}\right) \\ &= 1 + \langle u, X\rangle + \frac{\|u\|^2\sigma^2}{2} e^{\frac{\sigma^2\|u\|^2}{2}} \cdot \frac{\|X\|^2}{\sigma^2} e^{\frac{\|X\|^2}{2\sigma^2}} \\ &\overset{(a)}{\le} 1 + \langle u, X\rangle + \frac{\|u\|^2\sigma^2}{2} e^{\frac{\sigma^2\|u\|^2}{2}} \exp\left(\frac{\|X\|^2}{\sigma^2}\right). \end{aligned}$$

[8] If $r = d$, then $\mu = 1$; otherwise let $Z \sim \mathrm{Beta}(\frac{r}{2}, \frac{d-r}{2})$. By definition of $I_x(a,b)$ and $\mathbb{E}[Z] = r/d$, $\mu = \mathbb{P}(Z \ge r/(2d)) = \mathbb{P}(Z \ge \frac{1}{2}\,\mathbb{E}[Z])$. Paley–Zygmund (Paley & Zygmund, 1932) gives $\mu \ge \frac{1}{4}\,\frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]}$, and using $\mathbb{E}[Z^2] = \frac{r(r+2)}{d(d+2)}$ yields $\mu \ge \frac{1}{12}$.

Here (a) uses $\frac{\|X\|^2}{\sigma^2} \le \exp\Big(\frac{\|X\|^2}{2\sigma^2}\Big)$, which follows from $t^2 \le e^{t^2/2}$ applied to $t = \|X\|/\sigma$.

Taking expectations and using $\mathbb{E}[X] = 0$ yields

$$\mathbb{E}\big[e^{\langle u, X\rangle}\big] \le 1 + \frac{\|u\|^2\sigma^2}{2} e^{\frac{\sigma^2\|u\|^2}{2}} \mathbb{E}\left[\exp\left(\frac{\|X\|^2}{\sigma^2}\right)\right].$$

By Definition C.5 with $\lambda = 1/\sigma$, $\mathbb{E}[\exp(\|X\|^2/\sigma^2)] \le e$. Therefore,

$$\mathbb{E}\big[e^{\langle u, X\rangle}\big] \le 1 + \frac{e}{2}\|u\|^2\sigma^2\, e^{\frac{\sigma^2\|u\|^2}{2}} \overset{(a)}{\le} e^{\frac{\sigma^2\|u\|^2}{2}} \left(1 + \frac{e}{2}\|u\|^2\sigma^2\right) \overset{(b)}{\le} \exp\left(\frac{1}{2}\sigma^2\|u\|^2 + \frac{e}{2}\sigma^2\|u\|^2\right).$$

Here (a) uses that $e^{\frac{\sigma^2\|u\|^2}{2}} \ge 1$, hence $1 + ae^b \le e^b(1 + a)$ with $a = \frac{e}{2}\|u\|^2\sigma^2$, $b = \frac{\sigma^2\|u\|^2}{2}$. Here (b) uses $1 + y \le e^y$ with $y = \frac{e}{2}\|u\|^2\sigma^2$.

□

Lemma C.7. *Assume $\mathbb{E}[X] = 0$ and there exists $\sigma' > 0$ such that*

$$\forall u \in \mathbb{R}^d, \qquad \mathbb{E}\big[\exp(\langle u, X\rangle)\big] \le \exp(\|u\|^2\sigma'^2).$$

Then with $\bar\sigma := 24\sqrt{e}\,\sigma'$, we have

$$\forall |\lambda| \le \frac{1}{\bar\sigma}, \qquad \mathbb{E}\big[\exp(\lambda^2\|X\|^2)\big] \le 5^d \exp(\lambda^2\bar\sigma^2).$$

Proof. We follow the one-dimensional argument underlying Proposition 2.6.1 in Vershynin (2025), and adapt it to our setting. We then combine this with a covering argument in $\mathbb{R}^d$ to control $\|X\|$.

Step 1 Fix $u \in \mathbb{R}^d \setminus \{0\}$ and $t > 0$. For any $\lambda > 0$,

$$\mathbb{P}(\langle u, X\rangle \ge t) = \mathbb{P}\big(e^{\lambda\langle u, X\rangle} \ge e^{\lambda t}\big) \overset{(a)}{\le} e^{-\lambda t}\,\mathbb{E}\Big[e^{\lambda\langle u, X\rangle}\Big] \overset{(b)}{\le} \exp\big(-\lambda t + \lambda^2\sigma'^2\|u\|^2\big). \tag{38}$$

Here (a) is Markov's inequality, and (b) uses the assumed bound $\mathbb{E}[\exp(\langle v, X\rangle)] \le \exp(\sigma'^2\|v\|^2)$ with $v = \lambda u$.

Optimizing the right-hand side of (38) over λ gives

$$\mathbb{P}(\langle u, X\rangle \ge t) \le \exp\left(-\frac{t^2}{4\sigma'^2\|u\|^2}\right). \tag{39}$$

Applying (39) with u and with $-u$ yields, for all $t \ge 0$,

$$\mathbb{P}\big(|\langle u, X\rangle| \ge t\big) \le 2\exp\left(-\frac{t^2}{4\sigma'^2\|u\|^2}\right). \tag{40}$$

Step 2 For $p \ge 1$, we use the identity

$$\mathbb{E}[|Y|^p] = \int_0^\infty \mathbb{P}(|Y| \ge t)\, pt^{p-1}\, dt, \tag{41}$$

and apply it with $Y = \langle u, X\rangle$. Substituting (40) into (41) gives

$$\begin{aligned}\mathbb{E}[|\langle u, X\rangle|^p] &\overset{(a)}{\le} \int_0^\infty 2\exp\left(-\frac{t^2}{4\sigma'^2\|u\|^2}\right) pt^{p-1}\, dt \\ &= p\Gamma\Big(\frac{p}{2}\Big)\,(2\sigma'\|u\|)^p. \end{aligned} \tag{42}$$

Here (a) uses the tail bound (40).

Using $\Gamma(x) \le 3x^x$ for $x \ge 1/2$, we obtain

$$\mathbb{E}[|\langle u, X\rangle|^p] \le 3p\left(\frac{p}{2}\right)^{p/2} (2\sigma'\|u\|)^p. \tag{43}$$

Taking p-th roots in (43) and using $(3p)^{1/p} \le 3$ for $p \ge 1$ yields

$$\forall p \ge 1, \qquad \big(\mathbb{E}[|\langle u, X\rangle|^p]\big)^{1/p} \le 6\sigma'\|u\|\sqrt{p}. \tag{44}$$

Step 3 Let $\hat{\sigma} := 6\sigma'$. By (44),

$$\mathbb{E}\big[(\langle u, X\rangle)^{2p}\big] \overset{(a)}{\le} \big(\hat{\sigma}\|u\|\sqrt{2p}\big)^{2p} = \big(\hat{\sigma}^2\|u\|^2 \cdot 2p\big)^p. \tag{45}$$

Here (a) is (44) applied with exponent $2p$.

Therefore,

$$\begin{aligned}\mathbb{E}\big[e^{\lambda^2(\langle u,X\rangle)^2}\big] &= 1 + \sum_{p=1}^{\infty} \frac{\lambda^{2p}\,\mathbb{E}[(\langle u, X\rangle)^{2p}]}{p!} \\ &\overset{(a)}{\le} 1 + \sum_{p=1}^{\infty} \frac{\lambda^{2p}\,(\hat{\sigma}^2\|u\|^2 \cdot 2p)^p}{p!} \overset{(b)}{\le} 1 + \sum_{p=1}^{\infty} \big(\lambda^2 \cdot 2\hat{\sigma}^2\|u\|^2 e\big)^p. \end{aligned} \tag{46}$$

Here (a) uses (45), and (b) uses $p! \ge (p/e)^p$.

Now restrict to $u \in B_d := \{u : \|u\| \le 1\}$ and define $\sigma := 12\sqrt{e}\,\sigma'$. Then $2\hat{\sigma}^2 e = \sigma^2/2$, and (46) gives, for $\|u\| \le 1$,

$$\mathbb{E}\big[e^{\lambda^2(\langle u,X\rangle)^2}\big] \le 1 + \sum_{p=1}^{\infty} \left(\frac{\lambda^2\sigma^2}{2}\right)^p = \frac{1}{1 - \lambda^2\sigma^2/2}. \tag{47}$$

For $|\lambda| \le 1/\sigma$ we have $\lambda^2\sigma^2/2 \le 1/2$, hence

$$\frac{1}{1 - \lambda^2\sigma^2/2} \overset{(a)}{\le} \exp(\lambda^2\sigma^2). \tag{48}$$

Here (a) uses $\frac{1}{1-x} \le e^{2x}$ for $x \in [0, 1/2]$ with $x = \lambda^2\sigma^2/2$.

Combining (47) and (48) yields

$$\forall u \in B_d,\ \forall |\lambda| \le \frac{1}{\sigma}, \qquad \mathbb{E}\big[e^{\lambda^2(\langle u,X\rangle)^2}\big] \le e^{\lambda^2\sigma^2}. \tag{49}$$

Step 4 Let $\mathcal{N}_{1/2}$ be a $(1/2)$-net of B_d in Euclidean norm. Then

$$|\mathcal{N}_{1/2}| \overset{(a)}{\le} \left(1 + \frac{2}{1/2}\right)^d = 5^d. \tag{50}$$

(a) standard volumetric bound for Euclidean nets.

Moreover, since $\|X\| = \max_{v \in B_d} \langle v, X\rangle$, the net property implies

$$\|X\| \overset{(a)}{\le} 2 \max_{z \in \mathcal{N}_{1/2}} \langle z, X\rangle, \qquad \text{and hence} \qquad \|X\|^2 \overset{(b)}{\le} 4 \max_{z \in \mathcal{N}_{1/2}} \langle z, X\rangle^2. \tag{51}$$

(a) approximation of the supremum over B_d by a $(1/2)$-net, (b) squaring both sides.

Therefore, for $|\lambda| \le 1/(2\sigma)$,

$$\mathbb{E}\big[e^{\lambda^2\|X\|^2}\big] \overset{(a)}{\le} \mathbb{E}\Big[e^{4\lambda^2 \max_{z \in \mathcal{N}_{1/2}} \langle z,X\rangle^2}\Big] \overset{(b)}{\le} \sum_{z \in \mathcal{N}_{1/2}} \mathbb{E}\big[e^{4\lambda^2\langle z,X\rangle^2}\big]. \tag{52}$$

(a) using (51), (b) using $e^{\max_i a_i} \le \sum_i e^{a_i}$.

For each $z \in \mathcal{N}_{1/2} \subset B_d$, applying (49) with parameter 2λ yields

$$\sum_{z \in \mathcal{N}_{1/2}} \mathbb{E}\big[e^{4\lambda^2\langle z,X\rangle^2}\big] \overset{(a)}{\le} \sum_{z \in \mathcal{N}_{1/2}} e^{4\lambda^2\sigma^2} = |\mathcal{N}_{1/2}|\, e^{4\lambda^2\sigma^2} \overset{(b)}{\le} 5^d e^{4\lambda^2\sigma^2}. \tag{53}$$

(a) applying (49) with $u = z$, (b) using (50).

Finally, define $\bar{\sigma} := 2\sigma = 24\sqrt{e}\,\sigma'$. Then for $|\lambda| \le 1/\bar{\sigma}$ we have $|\lambda| \le 1/(2\sigma)$ and $4\lambda^2\sigma^2 = \lambda^2\bar{\sigma}^2$, so

$$\mathbb{E}\big[e^{\lambda^2\|X\|^2}\big] \le 5^d\, e^{\lambda^2\bar{\sigma}^2}. \tag{54}$$

□

Proposition C.8 (Generalization of Lemma C.3)**.** *Let $X_1, \dots, X_n$ be independent $\mathbb{R}^d$-valued random vectors such that $\mathbb{E}[X_i] = 0$ and $\|X_i\|$ is σ-sub-Gaussian for all i. Let $\bar{X} := \frac{1}{n}\sum_{i=1}^n X_i$. Define*

$$\bar{\sigma} := 24\sqrt{\frac{e(e+1)}{2}} \cdot \frac{\sigma}{\sqrt{n}}.$$

Then for all $|\lambda| \le 1/\bar{\sigma}$,

$$\mathbb{E}\big[e^{\lambda^2\|\bar{X}\|^2}\big] \le 5^d \, \exp(\lambda^2\bar{\sigma}^2).$$

Proof. By Lemma C.6, for each i and all $u \in \mathbb{R}^d$,

$$\mathbb{E}\big[\exp(\langle u, X_i\rangle)\big] \le \exp\left(\frac{e+1}{2}\sigma^2\|u\|^2\right). \tag{55}$$

Using independence, we obtain

$$\begin{aligned}
\mathbb{E}\big[\exp(\langle u, \bar{X}\rangle)\big] &= \mathbb{E}\Big[\exp\Big(\frac{1}{n}\sum_{i=1}^n \langle u, X_i\rangle\Big)\Big] \\
&= \prod_{i=1}^n \mathbb{E}\Big[\exp\Big(\frac{1}{n}\langle u, X_i\rangle\Big)\Big] \\
&\overset{(a)}{\le} \prod_{i=1}^n \exp\left(\frac{1}{n^2}\cdot\frac{e+1}{2}\sigma^2\|u\|^2\right) = \exp\left(\frac{e+1}{2n}\sigma^2\|u\|^2\right).
\end{aligned} \tag{56}$$

(a) applying (55) with u/n.

Thus, the assumption of Lemma C.7 holds for $\bar{X}$ with

$$\sigma' = \sqrt{\frac{e+1}{2n}}\,\sigma.$$

Applying Lemma C.7 yields

$$\bar{\sigma} = 24\sqrt{e}\,\sigma' = 24\sqrt{\frac{e(e+1)}{2}}\,\frac{\sigma}{\sqrt{n}},$$

which completes the proof.

□

Lemma C.9. *Assume F is L-smooth and bounded below by F^*. Assume $\mathbb{E}[\nabla f(x,\xi)] = \nabla F(x)$ and that $\|\nabla f(x,\xi) - \nabla F(x)\|$ is σ-sub-Gaussian (in the sense of Definition C.5). Let P_k be i.i.d. Haar random matrices, and let $\{x_k\}_{k\ge 0}$ be generated by the update rule*

$$x_{k+1} = x_k - \bar{\eta} P_k P_k^\top g_k,$$

where g_k denotes a mini-batch stochastic gradient estimator at x_k. Define the events

$$A_k := \Big\{\|P_k^\top \nabla F(x_k)\|^2 \ge \tfrac{1}{2}\|\nabla F(x_k)\|^2\Big\}, \qquad X := \sum_{j=0}^{T-1} \mathbf{1}_{A_j}.$$

Let

$$\mu := \begin{cases} 1 - I_{\frac{r}{2d}}\left(\frac{r}{2}, \frac{d-r}{2}\right), & r < d, \\ 1, & r = d, \end{cases}$$

where $I_x(a,b)$ is the regularized incomplete beta function. Then for all $s \in (0,1)$,

$$\mathbb{P}\big(X \le (1-s)\mu T\big) \le \exp\left(-\frac{s^2}{2}\mu T\right).$$

Proof. Let $\mathcal{F}_k := \sigma(g_0, P_0, \ldots, g_{k-1}, P_{k-1})$. We first consider the case $r < d$. Define

$$p_k := \mathbb{P}(A_k \mid \mathcal{F}_k) = \mathbb{E}[\mathbf{1}_{A_k} \mid \mathcal{F}_k].$$

For $\|\nabla F(x_k)\| \neq 0$, the identity $p_k = \mu$ follows from rotational invariance; this argument appears in the proof of Lemma 1 in Kozak et al. (2021). If $\|\nabla F(x_k)\| = 0$, then $p_k = 1$.

Therefore,

$$\sum_{k=0}^{T-1} p_k \geq \mu T. \tag{57}$$

Let $\psi(\lambda) := e^{-\lambda} - 1 + \lambda$ and fix $\lambda > 0$. Then

$$\begin{aligned}
\mathbb{E}\Big[e^{-\lambda(\mathbf{1}_{A_k} - p_k)} \mid \mathcal{F}_k\Big] &= p_k e^{-\lambda(1-p_k)} + (1-p_k)e^{\lambda p_k} \\
&= e^{\lambda p_k}\big(p_k e^{-\lambda} + (1-p_k)\big) \\
&\overset{(a)}{\leq} e^{\lambda p_k} \exp\big(p_k(e^{-\lambda} - 1)\big) = \exp\big(\psi(\lambda) p_k\big).
\end{aligned} \tag{58}$$

(a) using $1 + x \leq e^x$ with $x = p_k(e^{-\lambda} - 1)$.

Hence,

$$\mathbb{E}\Big[e^{-\lambda(\mathbf{1}_{A_k} - p_k) - \psi(\lambda) p_k} \mid \mathcal{F}_k\Big] \leq 1. \tag{59}$$

Define

$$M_k := \exp\left(-\lambda \sum_{j=0}^{k} (\mathbf{1}_{A_j} - p_j) - \psi(\lambda) \sum_{j=0}^{k} p_j\right).$$

By (59), $\{M_k\}$ is a supermartingale, and thus

$$\mathbb{E}[M_{T-1}] \leq 1. \tag{60}$$

Fix $s \in (0, 1)$. Using (57), for any $\lambda > 0$,

$$\begin{aligned}
\mathbb{P}(X \leq (1-s)\mu T) &\leq \mathbb{P}\left(X \leq (1-s) \sum_{k=0}^{T-1} p_k\right) \\
&= \mathbb{P}\Big(e^{-\lambda X} \geq e^{-\lambda(1-s)\sum_{k=0}^{T-1} p_k}\Big) \\
&= \mathbb{P}\Big(e^{-\lambda\left(X - \sum_{k=0}^{T-1} p_k\right)} \geq e^{\lambda s \sum_{k=0}^{T-1} p_k}\Big) \\
&= \mathbb{P}\Big(e^{-\lambda\left(X - \sum_{k=0}^{T-1} p_k\right) - \psi(\lambda)\sum_{k=0}^{T-1} p_k} \geq e^{(\lambda s - \psi(\lambda))\sum_{k=0}^{T-1} p_k}\Big) \\
&\overset{(a)}{=} \mathbb{P}\left(M_{T-1} \geq \exp\left((\lambda s - \psi(\lambda)) \sum_{k=0}^{T-1} p_k\right)\right).
\end{aligned} \tag{61}$$

Here, (a) follows from the definition of M_k.

Choose $\lambda = -\log(1-s)$, so that

$$\lambda s - \psi(\lambda) = s + (1-s)\log(1-s) \overset{(a)}{\geq} \frac{s^2}{2}. \tag{62}$$

(a) the standard inequality $s + (1-s)\log(1-s) \geq s^2/2$ for $s \in (0, 1)$.

Combining (61), (57), and (62) yields

$$\begin{aligned}
\mathbb{P}(X \leq (1-s)\mu T) &\leq \mathbb{P}\Big(M_{T-1} \geq e^{\frac{s^2}{2}\sum_{k=0}^{T-1} p_k}\Big) \leq \mathbb{P}\Big(M_{T-1} \geq e^{\frac{s^2}{2}\mu T}\Big) \\
&\overset{(a)}{\leq} \mathbb{E}[M_{T-1}]\, e^{-\frac{s^2}{2}\mu T} \overset{(b)}{\leq} \exp\left(-\frac{s^2}{2}\mu T\right).
\end{aligned} \tag{63}$$

(a) Markov's inequality applied to M_{T-1}. (b) using (60). This completes the proof for the case $r < d$.

When $r = d$, we have $\mu = 1$ and $P_k P_k^\top = I_d$, hence $\|P_k^\top \nabla F(x_k)\| = \|\nabla F(x_k)\|$ for all k. Therefore A_k holds for every k, and thus $X = T$. Since $(1-s)\mu T = (1-s)T < T$, it follows that

$$\mathbb{P}(X \le (1-s)\mu T) = 0,$$

which completes the proof.

□

Theorem C.10 (Generalization of Theorem C.4)**.** *Suppose that F satisfies Assumptions 2.1 and 2.2, and that the gradient satisfies Assumptions 2.3 and C.1. Let $\{x_k\}_{k\ge 0}$ be the sequence generated by the RS-SGD iteration. Let $\bar{B} := \lceil \max\{1, BT^q\} \rceil$ for $B > 0$ and $q > 0$, $\alpha := 24\sqrt{\frac{e(e+1)}{2}}$, and $\bar{\eta} = \frac{r}{Ld}$. Further define*

$$\mu := \begin{cases} 1 - I_{\frac{r}{2d}}\left(\frac{r}{2}, \frac{d-r}{2}\right), & r < d, \\ 1, & r = d, \end{cases} \tag{64}$$

where $I_x(a,b)$ denotes the regularized incomplete Beta function. Then for any $\delta, \delta' \in (0,1)$ and T such that $2\log\left(\frac{1}{\delta'}\right) < \mu T$, with probability at least $1 - \delta - \delta'$, it holds that

$$\min_{0\le k\le T-1} \|\nabla F(x_k)\|^2 \le \frac{4L\frac{d}{r}\Delta_0}{\mu T - \sqrt{2\mu T \log\left(\frac{1}{\delta'}\right)}} + \frac{2\frac{d}{r}\left(\frac{\alpha^2\sigma^2}{\bar{B}}\right)\left(d\log 5 + 1\right)}{\mu - \sqrt{2\frac{\mu}{T}\log\left(\frac{1}{\delta'}\right)}} + \frac{2\frac{d}{r}\left(\frac{\alpha^2\sigma^2}{\bar{B}}\right)\log\left(\frac{1}{\delta}\right)}{\mu T - \sqrt{2\mu T \log\left(\frac{1}{\delta'}\right)}}.$$

Remark. Setting $\delta \leftarrow \delta/2$ and $\delta' \leftarrow \delta/2$ in Theorem C.10 yields Theorem C.4.

Proof. Let $n_k := g_k - \nabla F(x_k)$. By L-smoothness,

$$\begin{aligned} F(x_{k+1}) - F(x_k) &\le \langle \nabla F(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &\overset{(a)}{=} -\bar{\eta}\langle \nabla F(x_k), P_k P_k^\top g_k\rangle + \frac{L\bar{\eta}^2}{2}\|P_k P_k^\top g_k\|^2 \\ &\overset{(b)}{=} -\bar{\eta}\langle \nabla F(x_k), P_k P_k^\top (\nabla F(x_k) + n_k)\rangle + \frac{L\bar{\eta}^2}{2}\|P_k P_k^\top (\nabla F(x_k) + n_k)\|^2, \end{aligned} \tag{65}$$

(a) uses $x_{k+1} - x_k = -\bar{\eta} P_k P_k^\top g_k$, and (b) uses $g_k = \nabla F(x_k) + n_k$.

Using $P_k^\top P_k = \frac{d}{r} I_r$ and expanding (65) yields

$$\bar{\eta}\Big(1 - \frac{L\bar{\eta}}{2}\cdot\frac{d}{r}\Big)\|P_k^\top \nabla F(x_k)\|^2 + F(x_{k+1}) - F(x_k) \le \Big(L\bar{\eta}^2\frac{d}{r} - \bar{\eta}\Big)\langle \nabla F(x_k), P_k P_k^\top n_k\rangle + \frac{L\bar{\eta}^2}{2}\cdot\frac{d}{r}\|P_k^\top n_k\|^2. \tag{66}$$

Substituting $\bar{\eta} = \frac{r}{Ld}$ in (66) gives

$$\frac{\bar{\eta}}{2}\|P_k^\top \nabla F(x_k)\|^2 + F(x_{k+1}) - F(x_k) \le \frac{\bar{\eta}}{2}\|P_k^\top n_k\|^2. \tag{67}$$

Fix $w > 0$ and define

$$C_k := \frac{\bar{\eta}}{2}\|P_k^\top \nabla F(x_k)\|^2 + F(x_{k+1}) - F(x_k), \quad Z_k := wC_k - d\log 5, \quad S_k := \sum_{j=k}^{T-1} Z_j.$$

Let $\mathcal{F}_k := \sigma(g_0, P_0, \ldots, g_{k-1}, P_{k-1})$ and $\bar{\sigma} := \alpha \cdot \frac{\sigma}{\sqrt{\bar{B}}}$. Recall that Proposition C.8 provides, for $|\lambda| \le 1/\bar{\sigma}$,

$$\mathbb{E}\big[\exp(\lambda^2\|n_k\|^2) \mid \mathcal{F}_k\big] \le 5^d \exp(\lambda^2\bar{\sigma}^2).$$

In what follows we will apply this bound with

$$\lambda^2 := \frac{w\bar{\eta}}{2} \cdot \frac{d}{r}, \tag{68}$$

which requires the condition

$$\lambda^2 \le \frac{1}{\bar{\sigma}^2} \quad \Longleftrightarrow \quad w \le \frac{2}{\bar{\sigma}^2} \frac{r}{\bar{\eta} d}. \tag{69}$$

Assume (69). Then

$$\begin{aligned}
\mathbb{E}\big[e^{Z_k} \mid \mathcal{F}_k\big] &= \mathbb{E}\big[\exp(wC_k - d\log 5) \mid \mathcal{F}_k\big] \\
&\overset{(a)}{\le} e^{-d\log 5}\, \mathbb{E}\Big[\exp\big(\tfrac{w\bar{\eta}}{2}\|P_k^\top n_k\|^2\big) \mid \mathcal{F}_k\Big] \\
&\overset{(b)}{\le} e^{-d\log 5}\, \mathbb{E}\Big[\exp\big(\tfrac{w\bar{\eta}}{2} \cdot \tfrac{d}{r}\|n_k\|^2\big) \mid \mathcal{F}_k\Big] \\
&\overset{(c)}{\le} e^{-d\log 5} \cdot \Big(5^d \exp\big(\tfrac{w\bar{\eta}}{2} \cdot \tfrac{d}{r}\bar{\sigma}^2\big)\Big) = \exp\Big(\tfrac{w\bar{\eta}}{2} \cdot \tfrac{d}{r}\bar{\sigma}^2\Big).
\end{aligned} \tag{70}$$

(a) uses (67), (b) uses $\|P_k^\top n_k\| \le \|P_k^\top\|\|n_k\| = \sqrt{\frac{d}{r}}\|n_k\|$, and (c) applies Proposition C.8 with the choice (68) (which is valid under (69)).

We next prove by backward induction that, for all $k = 0, 1, \dots, T-1$,

$$\mathbb{E}\big[e^{S_k} \mid \mathcal{F}_k\big] \le \exp\Big(\tfrac{w\bar{\eta}}{2} \cdot \tfrac{d}{r}\bar{\sigma}^2\,(T-k)\Big). \tag{71}$$

For the base case $k = T-1$, we have $S_{T-1} = Z_{T-1}$, hence (71) follows directly from (70). Assume (71) holds for $k+1$. Then, using $S_k = Z_k + S_{k+1}$ and the tower property,

$$\begin{aligned}
\mathbb{E}\big[e^{S_k} \mid \mathcal{F}_k\big] &= \mathbb{E}\big[e^{Z_k} e^{S_{k+1}} \mid \mathcal{F}_k\big] \\
&= \mathbb{E}\Big[e^{Z_k}\, \mathbb{E}\big[e^{S_{k+1}} \mid \mathcal{F}_{k+1}\big] \mid \mathcal{F}_k\Big] \\
&\overset{(a)}{\le} \exp\Big(\tfrac{w\bar{\eta}}{2} \cdot \tfrac{d}{r}\bar{\sigma}^2\,(T-k-1)\Big)\, \mathbb{E}\big[e^{Z_k} \mid \mathcal{F}_k\big] \\
&\overset{(b)}{\le} \exp\Big(\tfrac{w\bar{\eta}}{2} \cdot \tfrac{d}{r}\bar{\sigma}^2\,(T-k)\Big),
\end{aligned}$$

where (a) uses the induction hypothesis, and (b) uses (70). This completes the induction, and in particular yields

$$\mathbb{E}[e^{S_0}] \le \exp\Big(\tfrac{w\bar{\eta}}{2} \cdot \tfrac{d}{r}\bar{\sigma}^2\,T\Big). \tag{72}$$

Applying Markov's inequality to the nonnegative random variable $\exp(S_0)$, for any $t > 0$,

$$\mathbb{P}\big(S_0 \ge t\big) = \mathbb{P}\big(e^{S_0} \ge e^t\big) \le e^{-t}\, \mathbb{E}\big[e^{S_0}\big]. \tag{73}$$

Using (72) and choosing $t := \frac{w\bar{\eta}}{2} \cdot \frac{d}{r}\bar{\sigma}^2\,T + \log\frac{1}{\delta}$, we obtain

$$\mathbb{P}\left(S_0 \ge \tfrac{w\bar{\eta}}{2} \cdot \tfrac{d}{r}\bar{\sigma}^2\,T + \log\frac{1}{\delta}\right) \le \delta.$$

Therefore, with probability at least $1-\delta$, it holds that

$$S_0 \le \tfrac{w\bar{\eta}}{2} \cdot \tfrac{d}{r}\bar{\sigma}^2\,T + \log\frac{1}{\delta}. \tag{74}$$

By definition of S_0, we have

$$\begin{aligned}
S_0 &= \sum_{k=0}^{T-1} \big(wC_k - d\log 5\big) \\
&= \sum_{k=0}^{T-1} \Big(w\Big(\tfrac{\bar{\eta}}{2}\|P_k^\top \nabla F(x_k)\|^2 + F(x_{k+1}) - F(x_k)\Big) - d\log 5\Big) \\
&= \frac{w\bar{\eta}}{2} \sum_{k=0}^{T-1} \|P_k^\top \nabla F(x_k)\|^2 + w\big(F(x_T) - F(x_0)\big) - dT\log 5.
\end{aligned}$$

Choose

$$w \coloneqq \frac{2r}{\bar{\sigma}^2 \bar{\eta} d},$$

so that (70) is satisfied. Combining this choice with (74), we conclude that with probability at least $1-\delta$,

$$\frac{w\bar{\eta}}{2} \sum_{k=0}^{T-1} \|P_k^\top \nabla F(x_k)\|^2 + w\big(F(x_T) - F(x_0)\big) - dT \log 5 \le \frac{w\bar{\eta}}{2} \cdot \frac{d}{r} \bar{\sigma}^2 T + \log \frac{1}{\delta}.$$

Using $F(x_0) - F(x_T) \le F(x_0) - F^* \le \Delta_0$, $\frac{w\bar{\eta}}{2} = \frac{1}{\bar{\sigma}^2} \cdot \frac{r}{d}$, and $\bar{\eta} = \frac{r}{Ld}$, this inequality simplifies to

$$\sum_{k=0}^{T-1} \|P_k^\top \nabla F(x_k)\|^2 \le 2L\frac{d}{r}\Delta_0 + \frac{d}{r}\bar{\sigma}^2 dT \log 5 + \frac{d}{r}\bar{\sigma}^2 T + \frac{d}{r}\bar{\sigma}^2 \log \frac{1}{\delta}. \tag{75}$$

Next, define the events

$$A_k \coloneqq \Big\{ \|P_k^\top \nabla F(x_k)\|^2 \ge \tfrac{1}{2} \|\nabla F(x_k)\|^2 \Big\}, \qquad X \coloneqq \sum_{k=0}^{T-1} \mathbf{1}_{A_k}.$$

Then

$$\sum_{k=0}^{T-1} \|P_k^\top \nabla F(x_k)\|^2 \ge \frac{1}{2} \sum_{k=0}^{T-1} \|\nabla F(x_k)\|^2 \mathbf{1}_{A_k} \ge \frac{X}{2} \min_{0 \le k \le T-1} \|\nabla F(x_k)\|^2. \tag{76}$$

By Lemma C.9, for any $s \in (0,1)$,

$$\mathbb{P}\big(X \le (1-s)\mu T\big) \le \exp\left(-\frac{s^2}{2}\mu T\right). \tag{77}$$

Fix $\delta' \in (0,1)$ and assume $2\log(1/\delta') < \mu T$. Setting

$$s \coloneqq \sqrt{\frac{2}{\mu T} \log \frac{1}{\delta'}},$$

which satisfies $s \in (0,1)$ under this condition, (77) yields

$$\mathbb{P}\left(X \le \mu T - \sqrt{2\mu T \log \frac{1}{\delta'}} \right) \le \delta'. \tag{78}$$

Therefore, with probability at least $1-\delta'$,

$$\sum_{k=0}^{T-1} \|P_k^\top \nabla F(x_k)\|^2 \ge \frac{1}{2}\Big(\mu T - \sqrt{2\mu T \log \frac{1}{\delta'}}\Big) \min_{0 \le k \le T-1} \|\nabla F(x_k)\|^2. \tag{79}$$

Since $2\log(1/\delta') < \mu T$, we have $\mu T - \sqrt{2\mu T \log(1/\delta')} > 0$. Therefore, combining (75) and (79) and applying a union bound shows that, with probability at least $1 - \delta - \delta'$,

$$\min_{0 \le k \le T-1} \|\nabla F(x_k)\|^2 \le \frac{4L\frac{d}{r}\Delta_0}{\mu T - \sqrt{2\mu T \log \frac{1}{\delta'}}} + \frac{2\frac{d}{r}\bar{\sigma}^2(d\log 5 + 1)}{\mu - \sqrt{2\frac{\mu}{T}\log\frac{1}{\delta'}}} + \frac{2\frac{d}{r}\bar{\sigma}^2 \log\frac{1}{\delta}}{\mu T - \sqrt{2\mu T \log \frac{1}{\delta'}}}.$$

This completes the proof. □

Oracle complexity. From Theorem C.4 (applied with $q = 1$ and $B = \frac{d\sigma^2}{L\Delta_0}$), the iteration complexity satisfies

$$\tilde{\mathcal{O}}\left(\frac{d}{\mu r}\, \Delta_0 L\, \epsilon^{-2} \right).$$

Recalling that the oracle complexity is given by $r\bar{B}T$, we obtain

$$\tilde{\mathcal{O}}\left(\frac{d^3}{\mu^2 r}\, \Delta_0 L \sigma^2\, \epsilon^{-4} \right).$$

D RS-NGD: Update Rule and Convergence Analysis

Randomized Subspace Normalized Gradient Descent (RS-NGD). At each iteration k, we sample $P_k \in \mathbb{R}^{d\times r}$ as an i.i.d. Haar matrix, and perform the update

$$x_{k+1} = x_k - \bar{\eta}\, G_k,$$
$$G_k \coloneqq \begin{cases} \dfrac{P_k P_k^\top \nabla F(x_k)}{\|P_k^\top \nabla F(x_k)\|}, & \text{if } \|P_k^\top \nabla F(x_k)\| \neq 0, \\ 0, & \text{if } \|P_k^\top \nabla F(x_k)\| = 0. \end{cases}$$

Theorem D.1. *Assume that F satisfies Assumptions 2.1 and 2.2. Let $\{x_k\}_{k\geq 0}$ be generated by RS-NGD from x_0. We set*

$$\bar{\eta} = \eta T^{-u},$$

for parameters $\eta > 0$, and $u \in (0,1)$. Then, for all $T \geq 1$,

$$\frac{1}{T}\sum_{k=0}^{T-1} \mathbb{E}[\|\nabla F(x_k)\|] \leq \frac{\Delta_0}{\eta\tau T^{1-u}} + \frac{\eta \ell(\mathbb{L})\|\mathbb{L}\|}{2\tau T^u}.$$

Proof. Let $\mathcal{F}_k \coloneqq \sigma(P_0, \ldots, P_{k-1})$. Note that x_k is $\mathcal{F}_k$-measurable.

By (3), for any x, v,

$$F(x+v) \leq F(x) + \langle \nabla F(x), v\rangle + \frac{1}{2} v^\top \mathbb{L} v.$$

Applying this with $x = x_k$ and $v = x_{k+1} - x_k = -\bar{\eta}\, G_k$, we obtain

$$F(x_{k+1}) \leq F(x_k) - \bar{\eta}\,\langle \nabla F(x_k), G_k\rangle + \frac{\bar{\eta}^2}{2} G_k^\top \mathbb{L} G_k.$$

By the definition of G_k, on $\{\|P_k^\top \nabla F(x_k)\| \neq 0\}$ we have $G_k = \frac{P_k P_k^\top \nabla F(x_k)}{\|P_k^\top \nabla F(x_k)\|}$, hence

$$\langle \nabla F(x_k), G_k\rangle = \frac{\nabla F(x_k)^\top P_k P_k^\top \nabla F(x_k)}{\|P_k^\top \nabla F(x_k)\|} = \|P_k^\top \nabla F(x_k)\|.$$

On $\{\|P_k^\top \nabla F(x_k)\| = 0\}$ we have $G_k = 0$, and thus both sides equal 0. Therefore, for all outcomes,

$$\langle \nabla F(x_k), G_k\rangle = \|P_k^\top \nabla F(x_k)\|.$$

Substituting this identity yields

$$F(x_{k+1}) \leq F(x_k) - \bar{\eta}\,\|P_k^\top \nabla F(x_k)\| + \frac{\bar{\eta}^2}{2} G_k^\top \mathbb{L} G_k.$$

Rearranging gives

$$\bar{\eta}\,\|P_k^\top \nabla F(x_k)\| \leq F(x_k) - F(x_{k+1}) + \frac{\bar{\eta}^2}{2} G_k^\top \mathbb{L} G_k.$$

Summing over $k = 0, \ldots, T-1$ yields

$$\bar{\eta}\sum_{k=0}^{T-1} \|P_k^\top \nabla F(x_k)\| \leq \Delta_0 + \frac{\bar{\eta}^2}{2}\sum_{k=0}^{T-1} G_k^\top \mathbb{L} G_k. \tag{80}$$

Next, the Haar property in Proposition B.1, in particular (12), implies that

$$\mathbb{E}\big[\|P_k^\top \nabla F(x_k)\| \mid \mathcal{F}_k\big] = \tau\,\|\nabla F(x_k)\|.$$

Taking expectations and using the tower property yields

$$\mathbb{E}\big[\|P_k^\top \nabla F(x_k)\|\big] = \tau\,\mathbb{E}[\|\nabla F(x_k)\|]\,.$$

Moreover, by Proposition B.2 and the tower property,

$$\mathbb{E}\big[G_k^\top \mathbb{L} G_k\big] = \mathbb{E}\big[\mathbb{E}\big[G_k^\top \mathbb{L} G_k \mid \mathcal{F}_k\big]\big] \leq \ell(\mathbb{L})\,\|\mathbb{L}\|.$$

Taking expectations in (80) and combining the above bounds, we obtain

$$\bar{\eta}\tau \sum_{k=0}^{T-1} \mathbb{E}[\|\nabla F(x_k)\|] \le \Delta_0 + \frac{\bar{\eta}^2}{2}\, \ell(\mathbb{L})\, \|\mathbb{L}\|\, T.$$

Dividing both sides by $\bar{\eta}\tau T$ completes the proof. □

We next establish a high-probability convergence bound for RS-NGD.

Theorem D.2. *Assume that F satisfies Assumptions 2.1 and 2.2. Let $\{x_k\}_{k\ge 0}$ be generated by RS-NGD from x_0. Fix $T \ge 1$ and set*

$$\bar{\eta} = \eta T^{-u}, \qquad \text{for parameters } \eta > 0,\ u \in (0,1).$$

Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\frac{1}{T}\sum_{k=0}^{T-1} \|\nabla F(x_k)\| \le \frac{2\Delta_0}{\eta\tau T^{1-u}} + \frac{\eta}{\tau T^u}\frac{d}{r}L + \frac{3}{\tau^2}\left(\frac{d}{r}\frac{\sqrt{2\Delta_0 L}}{T} + \frac{\eta}{T^u}\Big(\frac{d}{r}\Big)^{3/2} L\right)\log\frac{1}{\delta}.$$

Proof. Let $\mathcal{F}_k \coloneqq \sigma(P_0, \dots, P_{k-1})$.

By Assumption 2.2, for any x, v,

$$F(x+v) \le F(x) + \langle \nabla F(x), v\rangle + \frac{1}{2}L\|v\|^2.$$

Applying this with $x = x_k$ and $v = x_{k+1} - x_k = -\bar{\eta} G_k$, we obtain

$$F(x_{k+1}) \le F(x_k) - \bar{\eta}\,\langle \nabla F(x_k), G_k\rangle + \frac{\bar{\eta}^2}{2}\, L\|G_k\|^2.$$

By the same argument as in the proof of Theorem D.1, we have

$$\langle \nabla F(x_k), G_k\rangle = \|P_k^\top \nabla F(x_k)\|.$$

Moreover, since $\|P_k\| \le \sqrt{d/r}$, it follows that $\|G_k\| \le \sqrt{d/r}$ and hence

$$L\|G_k\|^2 \le L\frac{d}{r}.$$

Substituting these bounds yields

$$F(x_{k+1}) \le F(x_k) - \bar{\eta}\,\|P_k^\top \nabla F(x_k)\| + \frac{\bar{\eta}^2}{2}\frac{d}{r}L.$$

Rearranging gives

$$\bar{\eta}\,\|P_k^\top \nabla F(x_k)\| \le F(x_k) - F(x_{k+1}) + \frac{\bar{\eta}^2}{2}\frac{d}{r}L.$$

Summing over $k = 0, \dots, T-1$ and using $F(x_T) \ge F^*$ yields

$$\bar{\eta}\sum_{k=0}^{T-1} \|P_k^\top \nabla F(x_k)\| \le \Delta_0 + \frac{\bar{\eta}^2}{2}\frac{d}{r}LT. \tag{81}$$

Define

$$\psi_k \coloneqq \|P_k^\top \nabla F(x_k)\|, \qquad \phi_k \coloneqq \mathbb{E}[\psi_k \mid \mathcal{F}_k], \qquad D_k \coloneqq \bar{\eta}(\phi_{k-1} - \psi_{k-1}), \quad k \ge 1,$$

with $D_0 \coloneqq 0$.

Then $(D_k, \mathcal{F}_k)$ is a martingale difference sequence. Moreover, since $0 \le \psi_k \le \sqrt{d/r}\|\nabla F(x_k)\|$, we have

$$|D_k| \le \bar{\eta}\sqrt{\frac{d}{r}}\,\|\nabla F(x_{k-1})\|.$$

Define

$$S_k \coloneqq \bar{\eta}\sqrt{\frac{d}{r}}\,\|\nabla F(x_{k-1})\|.$$

Then S_k is $\mathcal{F}_{k-1}$-measurable and satisfies

$$\frac{D_k^2}{S_k^2} \le 1.$$

Consequently,

$$\mathbb{E}\left[\exp\left(\frac{D_k^2}{S_k^2}\right) \,\middle|\, \mathcal{F}_{k-1}\right] \le e.$$

Applying Lemma 11 of Hübler et al. (2025), for any $\lambda > 0$, with probability at least $1-\delta$,

$$\sum_{t=1}^{T} D_t \le \frac{3}{4}\lambda \sum_{t=1}^{T} S_t^2 + \frac{1}{\lambda}\log\frac{1}{\delta}. \tag{82}$$

[9]

As in the proof of Theorem D.1,

$$\phi_k = \mathbb{E}[\psi_k \mid \mathcal{F}_k] = \tau\,\|\nabla F(x_k)\|.$$

Combining (82) with (81), we obtain, with probability at least $1-\delta$,

$$\bar{\eta}\tau \sum_{k=0}^{T-1} \|\nabla F(x_k)\| \le \Delta_0 + \frac{\bar{\eta}^2}{2}\frac{d}{r}LT + \frac{3}{4}\lambda\bar{\eta}^2\frac{d}{r}\sum_{k=0}^{T-1}\|\nabla F(x_k)\|^2 + \frac{1}{\lambda}\log\frac{1}{\delta}. \tag{83}$$

By L-smoothness,

$$\|\nabla F(x_{k+1})\| \le \|\nabla F(x_k)\| + \bar{\eta}L\sqrt{\frac{d}{r}}.$$

Iterating this bound yields, for all $k \in \{0,\dots,T\}$,

$$\|\nabla F(x_k)\| \le \|\nabla F(x_0)\| + k\,\bar{\eta}L\sqrt{\frac{d}{r}} \le \|\nabla F(x_0)\| + T\,\bar{\eta}L\sqrt{\frac{d}{r}}. \tag{84}$$

We now choose

$$\lambda \coloneqq \frac{2}{3}\,\frac{\tau}{\bar{\eta}\,\frac{d}{r}\left(\|\nabla F(x_0)\| + \bar{\eta}TL\sqrt{\frac{d}{r}}\right)}.$$

Then, using (84), for each $k = 0,\dots,T-1$ we have

$$\begin{aligned}\frac{3}{4}\lambda\,\bar{\eta}\,\frac{d}{r}\,\|\nabla F(x_k)\| &\le \frac{3}{4}\lambda\,\bar{\eta}\,\frac{d}{r}\left(\|\nabla F(x_0)\| + \bar{\eta}TL\sqrt{\frac{d}{r}}\right)\\ &= \frac{3}{4}\cdot\frac{2}{3}\,\tau = \frac{\tau}{2}.\end{aligned}$$

Substituting this pointwise bound into (83) gives

$$\bar{\eta}\frac{\tau}{2}\sum_{k=0}^{T-1}\|\nabla F(x_k)\| \le \Delta_0 + \frac{\bar{\eta}^2 TL}{2}\frac{d}{r} + \frac{1}{\lambda}\log\frac{1}{\delta}.$$

Multiplying both sides by $\frac{2}{\bar{\eta}\tau T}$, we obtain

$$\frac{1}{T}\sum_{k=0}^{T-1}\|\nabla F(x_k)\| \le \frac{2\Delta_0}{\bar{\eta}\tau T} + \frac{\bar{\eta}}{\tau}\frac{d}{r}L + \frac{2}{\bar{\eta}\tau T}\cdot\frac{1}{\lambda}\log\frac{1}{\delta}.$$

[9] Similar to Theorem B.5, to avoid division by zero and to make the argument rigorous, one may regularize the denominator by replacing S_k with $S_k + \frac{1}{n}$ for $n \in \mathbb{N}$. This yields a monotone decreasing sequence of events indexed by n, to which Lemma 11 applies for each fixed n. Letting $n \to \infty$ and using the continuity of probability yields the stated inequality.

We have

$$\frac{1}{\lambda} = \frac{3}{2}\frac{\bar{\eta}\frac{d}{r}\left(\|\nabla F(x_0)\| + \bar{\eta}TL\sqrt{\frac{d}{r}}\right)}{\tau}.$$

Substituting this into the last term yields

$$\frac{2}{\bar{\eta}\tau T}\cdot\frac{1}{\lambda}\log\frac{1}{\delta} = \frac{3}{\tau^2}\left(\frac{d}{r}\frac{\|\nabla F(x_0)\|}{T} + \bar{\eta}L\Big(\frac{d}{r}\Big)^{3/2}\right)\log\frac{1}{\delta}.$$

Therefore,

$$\frac{1}{T}\sum_{k=0}^{T-1}\|\nabla F(x_k)\| \le \frac{2\Delta_0}{\bar{\eta}\tau T} + \frac{\bar{\eta}}{\tau}\frac{d}{r}L + \frac{3}{\tau^2}\left(\frac{d}{r}\frac{\|\nabla F(x_0)\|}{T} + \bar{\eta}L\Big(\frac{d}{r}\Big)^{3/2}\right)\log\frac{1}{\delta}.$$

Substituting $\bar{\eta} = \eta T^{-u}$ gives

$$\frac{1}{T}\sum_{k=0}^{T-1}\|\nabla F(x_k)\| \le \frac{2\Delta_0}{\eta\tau T^{1-u}} + \frac{\eta}{\tau T^u}\frac{d}{r}L + \frac{3}{\tau^2}\left(\frac{d}{r}\frac{\|\nabla F(x_0)\|}{T} + \frac{\eta}{T^u}\Big(\frac{d}{r}\Big)^{3/2}L\right)\log\frac{1}{\delta}.$$

Finally, by Lemma B.4, we have

$$\|\nabla F(x_0)\| \le \sqrt{2\Delta_0 L}.$$

Substituting this bound into the last display yields

$$\frac{1}{T}\sum_{k=0}^{T-1}\|\nabla F(x_k)\| \le \frac{2\Delta_0}{\eta\tau T^{1-u}} + \frac{\eta}{\tau T^u}\frac{d}{r}L + \frac{3}{\tau^2}\left(\frac{d}{r}\frac{\sqrt{2\Delta_0 L}}{T} + \frac{\eta}{T^u}\Big(\frac{d}{r}\Big)^{3/2}L\right)\log\frac{1}{\delta}.$$

This completes the proof. □

E Experimental Details

E.1 Initialization

For each random seed, all methods start from the same initialization in both the synthetic quadratic experiment and character-level language modeling. In the synthetic quadratic experiment, the random seed controls the initialization, the stochastic noise samples, and, for randomized-subspace methods, the sampled subspaces. In character-level language modeling, the random seed controls initialization, training and evaluation mini-batch sampling, and, for randomized-subspace methods, the sampled subspaces.

E.2 Stepsize tuning

For each method (and each subspace dimension r when applicable), we tune the stepsize $\bar{\eta}$ over the prescribed grid using three independent tuning runs, and then evaluate the selected $\bar{\eta}$ on five disjoint evaluation runs under the same oracle-call budget.

For the synthetic quadratic experiment, the tuning score of a candidate step size $\bar{\eta}$ is defined as the average of $\|\nabla F(x)\| = \|\Lambda x\|$ over the last 10 iterations of the run. We then average this score over the tuning runs (across tuning seeds) and select the $\bar{\eta}$ that minimizes the resulting averaged score. The resulting best stepsizes are summarized in Table 3. The selected stepsize is then fixed and used for all five evaluation runs.

For character-level language modeling, the tuning score of a candidate $\bar{\eta}$ is defined as the final *estimated validation loss* at the end of the oracle-call budget. The validation loss is estimated by averaging the per-character negative log-likelihood over five randomly sampled minibatches. We average this score over tuning seeds and select the $\bar{\eta}$ with the smallest averaged score. The resulting best stepsizes are summarized in Table 4. The selected stepsize is fixed and used for all five evaluation runs.

Table 3: Best stepsizes for the synthetic quadratic with heavy-tailed noise ($d = 100$, $\bar{B} = 4$), selected by validation tuning.

Method	r	Best $\bar{\eta}$ ($\rho = 4$)	Best $\bar{\eta}$ ($\rho = 20$)	Best $\bar{\eta}$ ($\rho = 100$)
SGD	–	10^{-3}	10^{-2}	10^{-3}
RS-SGD	20	10^{-3}	10^{-3}	10^{-4}
RS-SGD	4	10^{-5}	10^{-4}	10^{-5}
NSGD	–	10^{-2}	10^{-2}	10^{-2}
RS-NSGD	20	10^{-2}	10^{-3}	10^{-3}
RS-NSGD	4	10^{-3}	10^{-3}	10^{-4}

Tuning grid: $\bar{\eta} \in \{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$.

Table 4: Best stepsizes for CharLM ($d = 5079$ for PTB and $d = 13700$ for WikiText-2, $\bar{B} = 128$), selected by validation tuning (minimizing the final validation loss).

Method	r	Best $\bar{\eta}$ (PTB)	Best $\bar{\eta}$ (WikiText-2)
SGD	–	1	1
RS-SGD	500	1	1
RS-SGD	50	10^{-1}	10^{-1}
RS-SGD	5	10^{-2}	10^{-2}
NSGD	–	1	1
RS-NSGD	500	1	1
RS-NSGD	50	10^{-1}	10^{-1}
RS-NSGD	5	10^{-2}	10^{-2}

Tuning grid: $\bar{\eta} \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$.

E.3 Logging and evaluation

For the synthetic quadratic experiment, we log the true gradient norm $\|\nabla F(x)\| = \|\Lambda x\|$ along the trajectory. In all plots, each curve reports the mean over five evaluation runs, and the shaded region indicates ± 1 standard deviation.

For character-level language modeling, we log training/validation/test losses at each logging point using the same evaluation procedure applied separately to each data split. To reduce evaluation overhead, we compute each reported loss as the average per-character negative log-likelihood over five randomly sampled mini-batches. In all plots, each curve reports the mean over five evaluation runs (five random seeds), and the shaded region indicates ± 1 standard deviation across seeds.

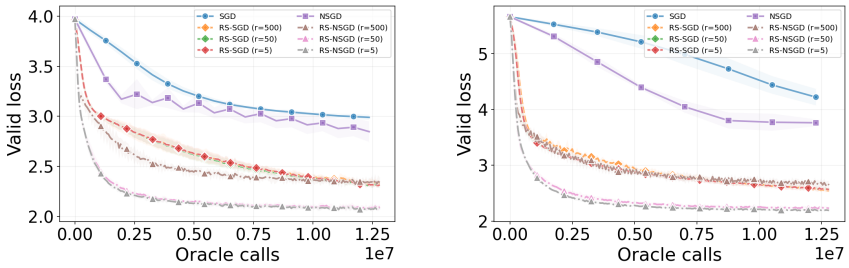
E.4 Validation/test results

At each logging point, we record the validation and test losses in addition to the training loss using the same evaluation procedure described above. Figure 4 reports the resulting validation and test trajectories. Consistent with the training results, RS-NSGD with $r \in \{5, 50\}$ achieves the lowest losses under the oracle-call budget across both datasets.

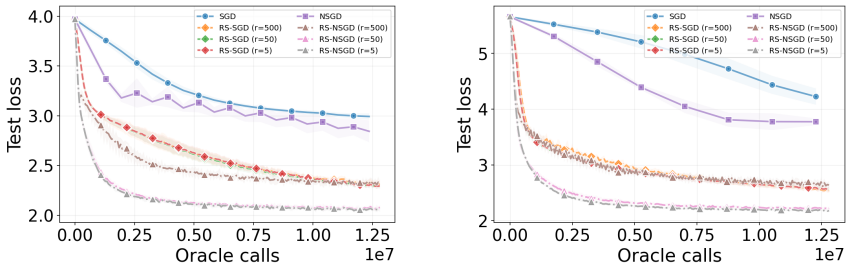
F Additional Experiments

This section reports additional experiments that complement the results in the main text. First, we add Adam (Kingma & Ba, 2015) and AdaGrad (Duchi et al., 2011) baselines and report results on WikiText-103 (Merity et al., 2017) as an additional dataset. Second, we consider additional character-level language-modeling datasets, enwik8 and text8 (Mahoney, 2011), including AdamW (Loshchilov & Hutter, 2019) and Muon (Jordan et al., 2024) baselines. [10] Finally, we provide timing results for a forward-mode AD implementation.

[10] The selected stepsize was $\bar{\eta} = 10^{-1}$ for Adam and AdaGrad on PTB, WikiText-2, and WikiText-103, and for AdamW and Muon on enwik8 and text8.



(a) PTB (validation) **(b)** WikiText-2 (validation)



(c) PTB (test) **(d)** WikiText-2 (test)

Figure 4: Character-level language modeling on PTB ($d = 5079, \bar{B} = 128$) and WikiText-2 ($d = 13700, \bar{B} = 128$): oracle calls vs. validation/test loss (5 seeds; mean $\pm$ 1 std.).

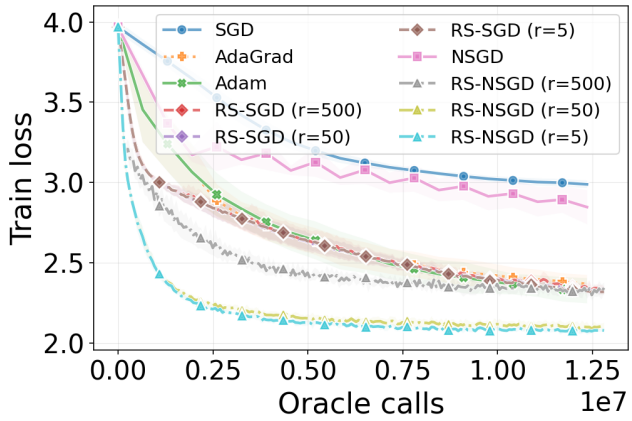


Figure 5: PTB character-level language modeling with additional Adam and AdaGrad baselines. The horizontal axis reports oracle calls, and the vertical axis reports training loss. We compare SGD, NSGD, RS-SGD, RS-NSGD, Adam, and AdaGrad under the same experimental setup as in Section 5.2. Curves show the mean over five evaluation seeds, and shaded regions indicate $\pm$ one standard deviation.

F.1 Additional Adam/AdaGrad baselines and WikiText-103 results

We additionally compare against Adam and AdaGrad using the same character-level language-modeling setup, stepsize grid, tuning protocol, and evaluation protocol as in Section 5.2 and Appendix E. We also include WikiText-103 as a larger dataset. The PTB results are shown in Figures 5 to 7, the WikiText-2 results in Figures 8 to 10, and the WikiText-103 results in Figures 11 to 13. These experiments are intended as additional robustness checks for the oracle-axis behavior observed in the main text.

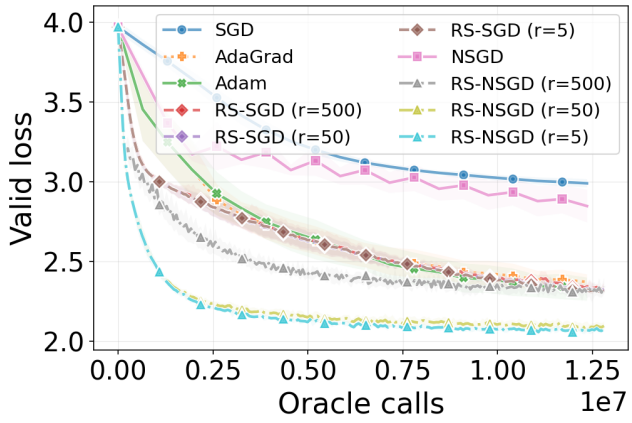


Figure 6: PTB character-level language modeling with additional Adam and AdaGrad baselines. The horizontal axis reports oracle calls, and the vertical axis reports validation loss. We compare SGD, NSGD, RS-SGD, RS-NSGD, Adam, and AdaGrad under the same experimental setup as in Section 5.2. Curves show the mean over five evaluation seeds, and shaded regions indicate $\pm$ one standard deviation.

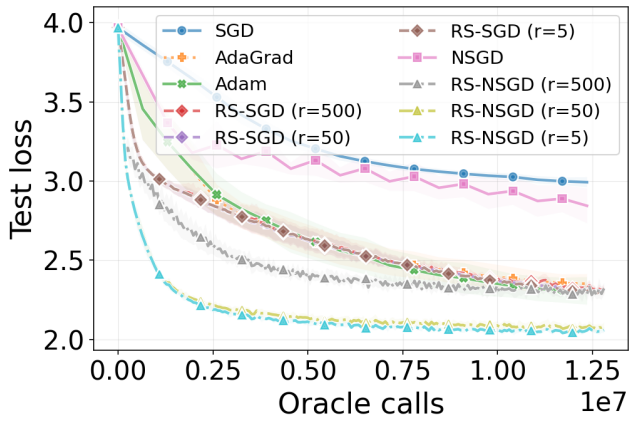


Figure 7: PTB character-level language modeling with additional Adam and AdaGrad baselines. The horizontal axis reports oracle calls, and the vertical axis reports test loss. We compare SGD, NSGD, RS-SGD, RS-NSGD, Adam, and AdaGrad under the same experimental setup as in Section 5.2. Curves show the mean over five evaluation seeds, and shaded regions indicate $\pm$ one standard deviation.

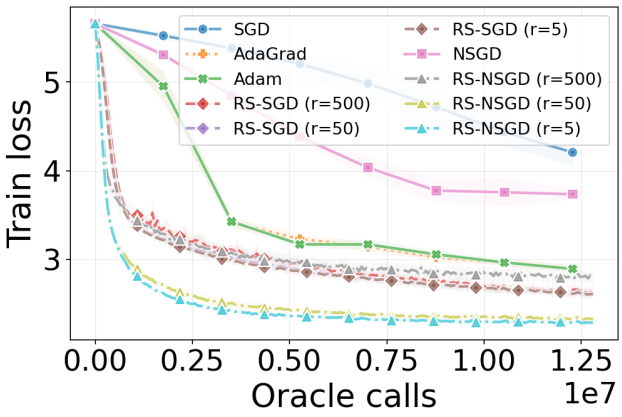


Figure 8: WikiText-2 character-level language modeling with additional Adam and AdaGrad baselines. The horizontal axis reports oracle calls, and the vertical axis reports training loss. We compare SGD, NSGD, RS-SGD, RS-NSGD, Adam, and AdaGrad under the same experimental setup as in Section 5.2. Curves show the mean over five evaluation seeds, and shaded regions indicate $\pm$ one standard deviation.

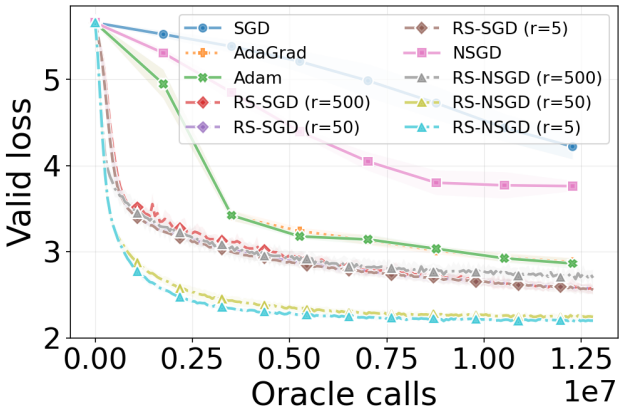


Figure 9: WikiText-2 character-level language modeling with additional Adam and AdaGrad baselines. The horizontal axis reports oracle calls, and the vertical axis reports validation loss. We compare SGD, NSGD, RS-SGD, RS-NSGD, Adam, and AdaGrad under the same experimental setup as in Section 5.2. Curves show the mean over five evaluation seeds, and shaded regions indicate $\pm$ one standard deviation.

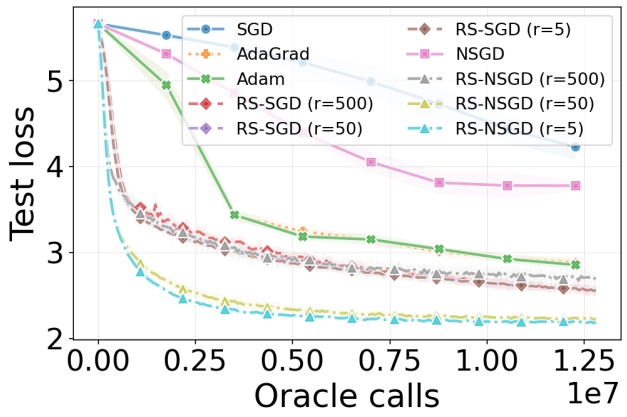


Figure 10: WikiText-2 character-level language modeling with additional Adam and AdaGrad baselines. The horizontal axis reports oracle calls, and the vertical axis reports test loss. We compare SGD, NSGD, RS-SGD, RS-NSGD, Adam, and AdaGrad under the same experimental setup as in Section 5.2. Curves show the mean over five evaluation seeds, and shaded regions indicate $\pm$ one standard deviation.

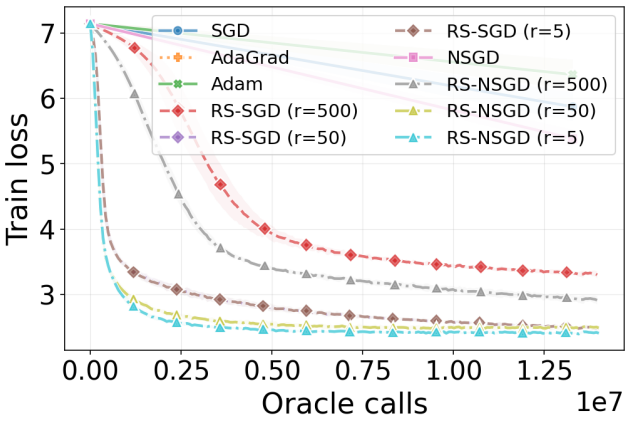


Figure 11: WikiText-103 character-level language modeling with additional Adam and AdaGrad baselines. The horizontal axis reports oracle calls, and the vertical axis reports training loss. We use a GRU with $E = 16$, $H = 56$, minibatch size $\bar{B} = 128$, and parameter dimension $d = 103{,}755$. We compare SGD, NSGD, RS-SGD, RS-NSGD, Adam, and AdaGrad. Curves show the mean over five evaluation seeds, and shaded regions indicate $\pm$ one standard deviation.

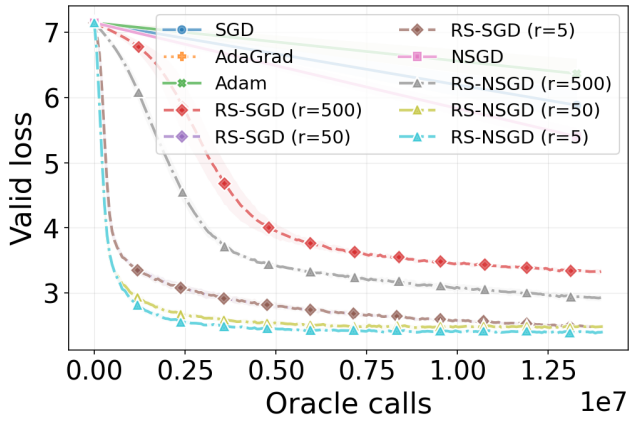


Figure 12: WikiText-103 character-level language modeling with additional Adam and AdaGrad baselines. The horizontal axis reports oracle calls, and the vertical axis reports validation loss. We use a GRU with $E = 16$, $H = 56$, minibatch size $\bar{B} = 128$, and parameter dimension $d = 103{,}755$. We compare SGD, NSGD, RS-SGD, RS-NSGD, Adam, and AdaGrad. Curves show the mean over five evaluation seeds, and shaded regions indicate $\pm$ one standard deviation.

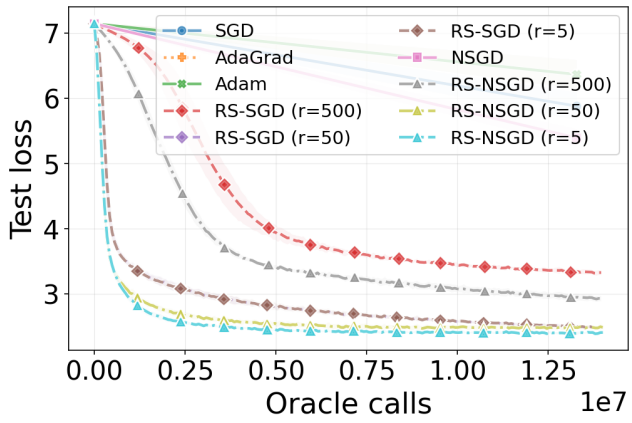


Figure 13: WikiText-103 character-level language modeling with additional Adam and AdaGrad baselines. The horizontal axis reports oracle calls, and the vertical axis reports test loss. We use a GRU with $E = 16$, $H = 56$, minibatch size $\bar{B} = 128$, and parameter dimension $d = 103{,}755$. We compare SGD, NSGD, RS-SGD, RS-NSGD, Adam, and AdaGrad. Curves show the mean over five evaluation seeds, and shaded regions indicate $\pm$ one standard deviation.

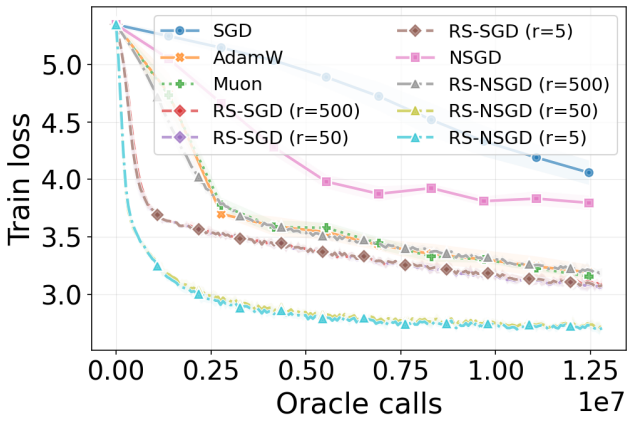


Figure 14: Oracle-axis comparison on enwik8 for character-level language modeling. The GRU size is fixed to $(E, H) = (8, 28)$, and the minibatch size is $B = 128$, and the parameter dimension is $d = 10{,}814$. The horizontal axis reports oracle calls, and the vertical axis reports training loss. We use the contiguous $90\text{M}/5\text{M}/5\text{M}$ train/validation/test split. AdamW and Muon are included as additional baselines; Muon is applied to the two-dimensional recurrent weights, while the remaining parameters are optimized by AdamW. Curves show the mean over five evaluation seeds, and shaded regions indicate $\pm$ one standard deviation.

F.2 Additional datasets and additional baselines

We also report character-level language-modeling results on enwik8 and text8. We use the same character-level language-modeling model setup, stepsize grid, tuning protocol, and evaluation protocol as in the PTB/WikiText-2 experiments. We include AdamW and Muon as additional baselines. Muon is applied to the two-dimensional recurrent weights, while the remaining parameters are optimized by AdamW. The enwik8 results are shown in Figures 14 to 16, and the text8 results are shown in Figures 17 to 19. Overall, the results are qualitatively consistent with the PTB/WikiText-2 experiments.

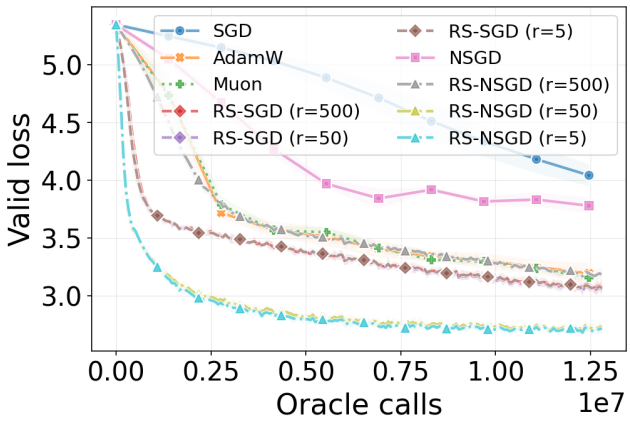


Figure 15: Oracle-axis comparison on enwik8 for character-level language modeling. The GRU size is fixed to $(E, H) = (8, 28)$, and the minibatch size is $\bar{B} = 128$, and the parameter dimension is $d = 10{,}814$. The horizontal axis reports oracle calls, and the vertical axis reports validation loss. We use the contiguous $90\mathrm{M}/5\mathrm{M}/5\mathrm{M}$ train/validation/test split. AdamW and Muon are included as additional baselines; Muon is applied to the two-dimensional recurrent weights, while the remaining parameters are optimized by AdamW. Curves show the mean over five evaluation seeds, and shaded regions indicate $\pm$ one standard deviation.

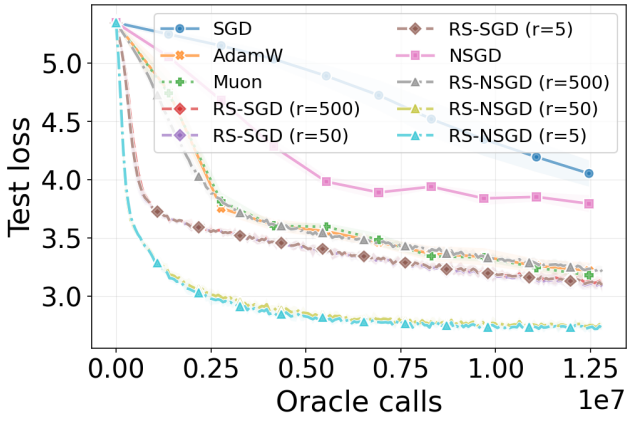


Figure 16: Oracle-axis comparison on enwik8 for character-level language modeling. The GRU size is fixed to $(E, H) = (8, 28)$, and the minibatch size is $\bar{B} = 128$, and the parameter dimension is $d = 10{,}814$. The horizontal axis reports oracle calls, and the vertical axis reports test loss. We use the contiguous $90\mathrm{M}/5\mathrm{M}/5\mathrm{M}$ train/validation/test split. AdamW and Muon are included as additional baselines; Muon is applied to the two-dimensional recurrent weights, while the remaining parameters are optimized by AdamW. Curves show the mean over five evaluation seeds, and shaded regions indicate $\pm$ one standard deviation.

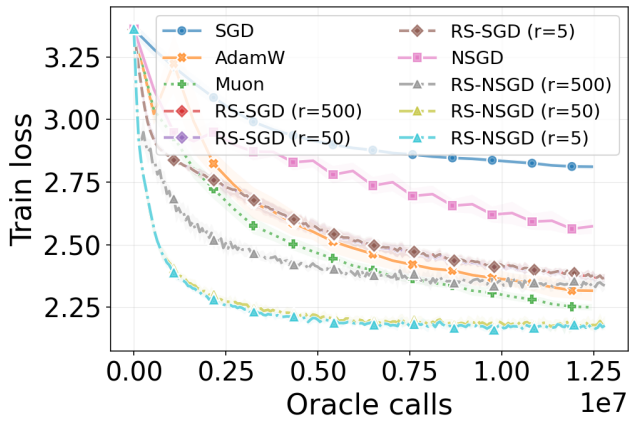


Figure 17: Oracle-axis comparison on text8 for character-level language modeling. The GRU size is fixed to $(E, H) = (8, 28)$, and the minibatch size is $\bar{B} = 128$, and the parameter dimension is $d = 4{,}228$. The horizontal axis reports oracle calls, and the vertical axis reports training loss. We use the contiguous $90\mathrm{M}/5\mathrm{M}/5\mathrm{M}$ train/validation/test split. AdamW and Muon are included as additional baselines; Muon is applied to the two-dimensional recurrent weights, while the remaining parameters are optimized by AdamW. Curves show the mean over five evaluation seeds, and shaded regions indicate $\pm$ one standard deviation.

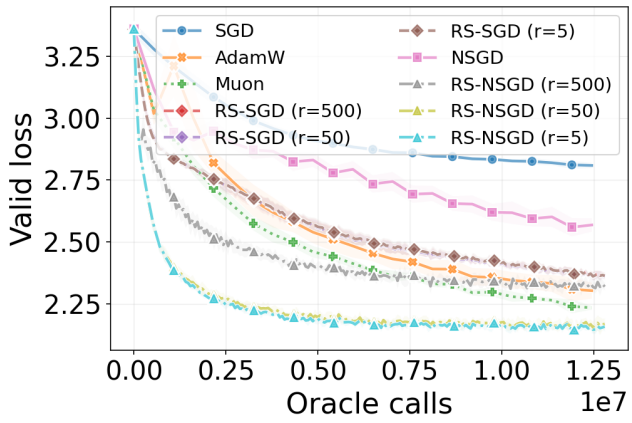


Figure 18: Oracle-axis comparison on text8 for character-level language modeling. The GRU size is fixed to $(E, H) = (8, 28)$, and the minibatch size is $\bar{B} = 128$, and the parameter dimension is $d = 4{,}228$. The horizontal axis reports oracle calls, and the vertical axis reports validation loss. We use the contiguous $90\mathrm{M}/5\mathrm{M}/5\mathrm{M}$ train/validation/test split. AdamW and Muon are included as additional baselines; Muon is applied to the two-dimensional recurrent weights, while the remaining parameters are optimized by AdamW. Curves show the mean over five evaluation seeds, and shaded regions indicate $\pm$ one standard deviation.

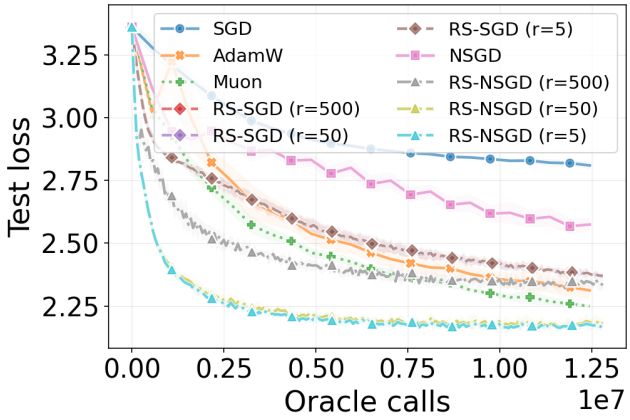


Figure 19: Oracle-axis comparison on text8 for character-level language modeling. The GRU size is fixed to $(E, H) = (8, 28)$, and the minibatch size is $\bar{B} = 128$, and the parameter dimension is $d = 4{,}228$. The horizontal axis reports oracle calls, and the vertical axis reports test loss. We use the contiguous $90\mathrm{M}/5\mathrm{M}/5\mathrm{M}$ train/validation/test split. AdamW and Muon are included as additional baselines; Muon is applied to the two-dimensional recurrent weights, while the remaining parameters are optimized by AdamW. Curves show the mean over five evaluation seeds, and shaded regions indicate $\pm$ one standard deviation.

Table 5: Time breakdown for JVP-based RS-NSGD and full NSGD on PTB ($E = 8$, $H = 28$, $\bar{B} = 128$, $d = 5079$). The column "Oracle calls" reports the oracle amount over the 100 timed steps. The column "100-step total" reports the total wall-clock time over the 100 timed steps after 20 warmup steps. `Haar (%)` denotes the share of the total step time spent in generating the Haar-random sketch matrix. `JVP (%)` denotes the share of the total step time spent in the grouped JVP block. For full NSGD, no Haar step is used, so `Haar (%)` is 0.00.

Method	r	Oracle calls	100-step total [ms]	`Haar (%)` / `JVP (%)`
RS-NSGD	5	64,000	8,788	0.22 / 98.16
RS-NSGD	50	640,000	86,105	0.07 / 99.61
RS-NSGD	500	6,400,000	869,631	0.07 / 99.74
Full NSGD	d	65,011,200	8,812,018	0.00 / 99.78

F.3 Timing results

We report timing measurements for a Jacobian-vector-product (JVP)-based forward-mode AD implementation.[11] The timing results are summarized in Table 5. These measurements clarify the relative cost of Haar sketch generation and projected-gradient computation. They also illustrate the relation between oracle calls and wall-clock time in this forward-mode AD implementation.

The timing table shows that Haar sketch generation accounts for less than 1% of the total step time in the tested settings. This table also compares JVP-based RS-NSGD with full NSGD. In this forward-mode AD implementation, wall-clock time scales closely with the oracle count in the tested settings.

F.4 Experimental resources, dataset sources, and code availability

Computational resources. The experiments were conducted using a GPU server and a local CPU machine. The GPU server had an NVIDIA RTX A5000 GPU with 24564 MiB of memory, an AMD EPYC 7413 CPU, and 503 GiB of system memory. The local CPU machine was an Apple M1 machine with 8 GiB of memory. The character-level language-modeling runs took on the order of 6 wall-clock hours per dataset on average. The synthetic quadratic experiments were smaller, taking roughly 0.5 wall-clock hours per figure.

Dataset sources and terms of use. We use previously released benchmark datasets and credit their original sources. We do not redistribute the raw datasets. PTB is credited to the Penn Treebank (Marcus et al., 1993); WikiText-2 and WikiText-103 are credited to the corresponding WikiText datasets (Merity et al., 2017); and enwik8/text8 are credited to the Large Text Compression Benchmark (Mahoney, 2011). For each dataset, the README or code specifies the source location used for reproduction. The datasets are subject to the licenses or terms of use of their original providers, and no raw dataset files are redistributed.

Code availability. We provide an anonymized supplementary code package for reproducing the experiments. The package contains scripts for running the main and additional experiments, plotting scripts for generating the reported figures from the resulting CSV logs, and a README describing the required environment and reproduction commands.

[11] The JVP implementation is used only for this timing experiment. In the main trajectory experiments, we instead form the minibatch stochastic gradient before sketching—analytically for the synthetic quadratic and via reverse-mode AD for character-level language modeling—as this was more efficient in our experimental environment. For fixed minibatches and sketches, both implementations yield the same projected gradients, and hence the same trajectories, in exact arithmetic. Projecting after gradient computation is also natural in communication settings, where only the r-dimensional projected gradient is transmitted.