THE ADVANTAGES OF FRESH SKETCHING FOR RIDGE REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Over the past 25 years, sketching and sampling have become widely used tools for accelerating large-scale regression. In iterative randomized solvers, a basic design choice is whether to *reuse* the same sketch or draw *fresh* randomness at every step. For (under-constrained) iterative ridge regression with column sampling, whether fresh sketches offer provable advantages has remained open: *We show that they do.* Fresh sketching lets us analyze error only along the current residual solution, rather than uniformly over the entire Gram matrix. This directional view yields sharper convergence guarantees for leverage score and ridge leverage score sampling and, more importantly, leads to residual-aware sampling rules. By minimizing the variance of the relevant sketched matrix-vector product, we derive an oracle distribution and practical approximations to the oracle distribution, including a mixture sampling distribution with (somewhat weaker) convergence guarantees. Experiments on synthetic and real data, including ridge probes on Qwen2.5 representations, support our theory, showing substantially faster convergence.

1 INTRODUCTION

Tools from Randomized Numerical Linear Algebra (RandNLA), such as sketching and sampling, are of paramount importance for accelerating large-scale regression and other matrix computations (Halko et al., 2011; Woodruff, 2014; Drineas & Mahoney, 2017; Martinsson & Tropp, 2020; Dereziński & Mahoney, 2021; Murray et al., 2023; Mahoney, 2011; Drineas & Mahoney, 2016). By replacing expensive matrix operations with smaller randomized approximations, these methods can substantially reduce computational and memory costs while retaining provable accuracy guarantees.

A basic design choice in an iterative randomized solver is whether to draw a sketch once and reuse the same randomness throughout the iteration, or to draw fresh randomness at every step. While reusing a sketch is natural computationally, fresh sketches may better capture the changing directions encountered by the iterative method. This raises a basic question:

When, and why, does fresh randomness help an iterative sketched solver?

We study this question for iterative ridge regression with column sampling. Given a design matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, a response vector $\mathbf{b} \in \mathbb{R}^n$, and a regularization parameter $\lambda > 0$, ridge regression solves

$$\mathbf{x}^\star = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^d} \left\{ \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_2^2 \right\} = \mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top + \lambda\mathbf{I}_n)^{-1}\mathbf{b}. \tag{1}$$

(Hoerl & Kennard, 1970; Saunders et al., 1998). We focus on the wide regime $n \ll d$, where the number of features is much larger than the number of observations. Directly forming the Gram matrix $\mathbf{A}\mathbf{A}^\top$ costs $O(n^2 d)$, which can become prohibitive when d is very large.

Column sampling reduces this cost by sampling and rescaling features to approximate the Gram matrix using a much smaller number of columns (Drineas et al., 2006a). Chowdhury et al. (2018) use this approximation in an iterative solver that repeatedly solves a sketched ridge regression problem and corrects the remaining residual. For sufficiently accurate sketches, these residual corrections yield geometric decay of the solution error.

The analysis of Chowdhury et al. (2018) uses a fixed initial sketch that is reused across iterations. Their experiments, however, indicate that drawing a new sample at every iteration can accelerate

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.

convergence and reduce the sketch size needed to converge, and they explicitly ask whether this additional randomness leads to stronger theoretical guarantees. Kacham & Woodruff (2022) subsequently analyze the same iterative solver with refreshed oblivious sketches and obtain improved sample complexity guarantees. The proof of their Lemma 3.3 analyzes the same matrix-vector product as our Lemma 3.2. More recently, Li & Niu (2024) propose a sampling probability similar to our oracle proposal: their ideal rule is $p_i^{\mathrm{LN},(j)} \propto |\mathbf{x}_i^{\star(j)}| \|\mathbf{A}_{*i}\|_2$, while ours is $p_i^{\star(j)} \propto |\mathbf{x}_i^{\star(j)}| \sqrt{w_i}$. We return to the column sampling setting and study fresh leverage score and ridge leverage score sketches, which exploit the structure of the design matrix and are natural sampling distributions for ridge regression. For a more detailed discussion of related works, we refer the readers to Section C.

Our main observation is that *fresh randomness changes what must be approximated*. Existing fixed-sketch analyses seek sufficiently accurate uniform approximation of the relevant matrix, for example by controlling the norm of the matrix $\mathbf{V}^\top \mathbf{S}\mathbf{S}^\top \mathbf{V} - \mathbf{I}$, where $\mathbf{V} \in \mathbb{R}^{d \times n}$ is the tall-and-thin matrix of the right singular vectors of $\mathbf{A}$ and the $d \times s$ matrix $\mathbf{S}$ (with $s \ll d$) is the sketching matrix (see Section A and D.3 for details). With an independently refreshed sketch at each iteration, this is stronger than necessary. At iteration j, the solver only needs the sketch to be accurate in the direction of the current residual ridge solution. The key quantity is therefore a sketched *matrix-vector* product rather than a uniform matrix-matrix approximation. Conditional independence of the fresh sketch from the previous iterations allows us to analyze the variance of this directional error and propagate it through the iteration.

This directional viewpoint has a second consequence: it tells us how the sampling probabilities themselves should change during the iteration. Once convergence is governed by the variance of a particular sketched matrix-vector product, we can choose the sampling distribution to minimize that variance. This produces an oracle distribution that combines ridge leverage information with the current residual solution. The oracle solution is unavailable in practice, but its structure suggests adaptive sampling probabilities based on an inexpensive approximation. Thus, residual-aware sampling is not a separate heuristic: it emerges directly from the same directional analysis that explains the advantages of fresh sketching.

Our contributions. We establish both analytical and algorithmic advantages of fresh sketching for iterative ridge regression with column sampling.

1. **Fresh sketches yield sharper convergence through directional error.** We derive solution error bounds for fresh leverage score and ridge leverage score sampling in Theorems 3.3 and 4.3. When $\|\mathbf{A}\|_2^2/\lambda$ is moderate, these results achieve geometric convergence at rate ε^t (where ε is the desired relative error and t is the number of iterations) using $O(n/\varepsilon^2)$ and $O(d_\lambda/\varepsilon^2)$ sampled columns per iteration, respectively, where d_λ is the effective degrees of freedom. Compared with Chowdhury et al. (2018), these bounds save factors of $\log n$ and $\log(1 + d_\lambda)$ while retaining the same geometric rate. Secondly, at the same sample sizes, Theorems E.2 and F.3 give faster guaranteed convergence: the contraction factor gains an additional inverse square root of the corresponding logarithm. Finally, for leverage score sampling, Theorem E.3 shows that independence and cancellation of the linear sketch error yield the algorithmic bias bound $\|\mathbb{E}[\widehat{\mathbf{x}}^\star - \mathbf{x}^\star]\|_2 \le \varepsilon^{2t} \|\mathbf{x}^\star\|_2$ at the stated sample size.

   Remark. We note that, in general, independent column sampling methods with replacement cannot avoid logarithmic factors; this has been known since the early days of random sampling and follows from coupon collector considerations (see Drineas & Mahoney, 2017, Remark 42). Recently, celebrated results use modern random matrix theory (RMT) tools to remove the logarithmic factor from the embedding dimension for sparse oblivious sketches, resolving the conjecture of Nelson & Nguyên (2013) up to sub-polylogarithmic factors (Chenakkod et al., 2026; 2025; 2024; Brailovskaya & van Handel, 2024). Our result, to the best of our knowledge, is the first one to avoid logarithmic factors in the sampling size complexity for the iterative ridge regression framework of Chowdhury et al. (2018), by using fresh randomness. Our proofs leverage conditional probability and standard measure concentration inequalities.

2. **Directional error leads to residual-aware sampling.** We minimize the variance of the matrix-vector product that governs the one-step error and derive an oracle sampling distribution that combines ridge leverage scores with the current residual solution. We additionally introduce practical approximations of the oracle sampling distribution. Our proposed

adaptive method uses an initial ridge pilot sketch to approximate the residual solution; mixture methods further mix this probability distribution with ridge leverage scores to improve practicality and performance. (See Section 5 for details.)

3. **Experiments demonstrate the advantages of fresh and residual-aware sketches.** We compare uniform, leverage score, ridge leverage score, oracle, adaptive, and mixture sampling on ARCENE and synthetic data (Figures 1 and 3), and separately compare fixed and fresh sketches on ARCENE (Figure 4), extending the experiments of Chowdhury et al. (2018). We also evaluate the method on semantic similarity ridge probing for the STS Benchmark, using concatenated representations from multiple layers of Qwen2.5 to produce a wide regression problem (Cer et al., 2017; Qwen Team et al., 2025). Across iteration counts, sketching budgets, and regularization parameters, the experiments show the convergence advantages of the residual-aware sampling rules.

2 Problem setup and iterative solver

Scalars are denoted by lowercase letters (e.g., x, λ); vectors by bold lowercase letters (e.g., $\mathbf{x}$, $\mathbf{g}$); matrices by bold uppercase letters (e.g., $\mathbf{W}$, $\mathbf{G}$). For a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ we write $\|\mathbf{A}\|_F$ for the Frobenius norm, $\|\mathbf{A}\|_2$ for the spectral norm. The notation $\mathbf{A}_{i*}$ and $\mathbf{A}_{*i}$ denotes the ith row and column, respectively. The order $\succeq$ denotes positive semidefinite ordering, and $\mathbf{1}_B$ denotes the indicator of an event B. Without loss of generality, throughout the analysis we assume that all sampling probabilities are strictly positive. (We will always skip outcomes with zero sampling probabilities.) Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ have full row rank, with $d > n > 1$, and let $\mathbf{b} \in \mathbb{R}^n$ and $\lambda > 0$. Write $\mathbf{I}_m$ for the $m \times m$ identity and $\|\cdot\|_2$ for the Euclidean vector norm.

We consider the ridge regression problem in (1), write the thin singular value decomposition as $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{n\times n}$ is orthogonal, $\mathbf{V} \in \mathbb{R}^{d\times n}$ has orthonormal columns, and $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$ with $\sigma_{\max} = \sigma_1 \geq \cdots \geq \sigma_n = \sigma_{\min} > 0$. Define $\mathbf{\Sigma}_\lambda = \mathbf{\Sigma}(\mathbf{\Sigma}^2 + \lambda\mathbf{I}_n)^{-1/2}, d_\lambda = \|\mathbf{\Sigma}_\lambda\|_F^2 = \sum_{i=1}^n \frac{\sigma_i^2}{\sigma_i^2+\lambda}$. Thus $\|\mathbf{\Sigma}_\lambda\|_2 \leq 1$ and $0 < d_\lambda < n$. The quantity d_λ is the effective degrees of freedom (Alaoui & Mahoney, 2015).

Column sampling. A sketch of size s is a matrix $\mathbf{S} \in \mathbb{R}^{d\times s}$ that independently samples indices $i_1, \ldots, i_s$ from $\{1, 2, \ldots, d\}$ using a probability distribution $p = (p_1, \ldots, p_d)$ and has columns $\mathbf{S}_{*k} = \mathbf{e}_{i_k}/\sqrt{s p_{i_k}}$ (Drineas et al., 2006a). The two baseline distributions use leverage scores (LS) (Drineas et al., 2012) and ridge leverage scores (RLS) (Alaoui & Mahoney, 2015; Cohen et al., 2017):

$$p_i^{\mathrm{ls}} = \frac{\|\mathbf{V}_{i*}\|_2^2}{n}, \quad p_i^{\mathrm{rls}} = \frac{w_i}{d_\lambda}, w_i := \|(\mathbf{V}\mathbf{\Sigma}_\lambda)_{i*}\|_2^2.$$

Residual correction. We use the iteration of Chowdhury et al. (2018), with a new sketch at every step. Fix an iteration count t and initialize $\mathbf{b}^{(0)} = \mathbf{b}$, $\widetilde{\mathbf{x}}^{(0)} = \mathbf{0}_d$, and $\mathbf{y}^{(0)} = \mathbf{0}_n$. At iteration $j = 1, \ldots, t$, update the residual, draw a fresh sketch $\mathbf{S}^{(j)}$ with size $s^{(j)}$ and probabilities $p^{(j)}$, and compute

$$\begin{aligned} \mathbf{b}^{(j)} &= \mathbf{b}^{(j-1)} - \lambda\mathbf{y}^{(j-1)} - \mathbf{A}\widetilde{\mathbf{x}}^{(j-1)}, \\ \mathbf{H}_j &= \mathbf{A}\mathbf{S}^{(j)}\mathbf{S}^{(j)\top}\mathbf{A}^\top + \lambda\mathbf{I}_n, \qquad \mathbf{y}^{(j)} = \mathbf{H}_j^{-1}\mathbf{b}^{(j)}, \\ \widetilde{\mathbf{x}}^{(j)} &= \mathbf{A}^\top\mathbf{y}^{(j)}. \end{aligned} \tag{2}$$

After t steps, the output is $\widehat{\mathbf{x}}^\star = \sum_{j=1}^t \widetilde{\mathbf{x}}^{(j)}$. The exact solution of residual problem j is $\mathbf{x}^{\star(j)} = \mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top + \lambda\mathbf{I}_n)^{-1}\mathbf{b}^{(j)}$. In particular, $\mathbf{b}^{(1)} = \mathbf{b}^{(0)} = \mathbf{b}$ and $\mathbf{x}^{\star(1)} = \mathbf{x}^{\star(0)} = \mathbf{x}^\star$. The history $\mathcal{F}_j = \sigma(\mathbf{S}^{(1)}, \ldots, \mathbf{S}^{(j)})$ is the sigma algebra generated by the first j sketches, with $\mathcal{F}_0 = \{\emptyset, \Omega\}$, where Ω is the sample space. Thus $\mathbf{x}^{\star(j)}$ is $\mathcal{F}_{j-1}$-measurable. For the baseline distributions, sketches are independent across iterations. For residual-aware distributions, each sketch depends on past randomness. All expectations refer to sketch randomness, with $\mathbf{A}$ and $\mathbf{b}$ fixed.

The identities $\mathbf{x}^{\star(j)} = \mathbf{x}^{\star(j-1)} - \widetilde{\mathbf{x}}^{(j-1)}$ and $\widehat{\mathbf{x}}^\star - \mathbf{x}^\star = \widetilde{\mathbf{x}}^{(t)} - \mathbf{x}^{\star(t)}$ reduce the final error to successive residual errors (Lemma D.2, proved in Appendix D.2). Algorithm 1 states the fresh

Algorithm 1 Iterative ridge regression with fresh sketches

Input: $\mathbf{A}, \mathbf{b}$; $\lambda > 0$; iteration count $t \geq 1$; sketch sizes $s^{(1)}, \ldots, s^{(t)}$; rule LS, RLS, or Oracle.
Initialize $\mathbf{b}^{(0)} = \mathbf{b}$, $\widetilde{\mathbf{x}}^{(0)} = \mathbf{0}_d$, and $\mathbf{y}^{(0)} = \mathbf{0}_n$.
for $j = 1, \ldots, t$ **do**
  Set $\mathbf{b}^{(j)} = \mathbf{b}^{(j-1)} - \lambda \mathbf{y}^{(j-1)} - \mathbf{A}\widetilde{\mathbf{x}}^{(j-1)}$.
  if rule is LS **then**
    Set $p_i^{(j)} = \|\mathbf{V}_{i*}\|_2^2/n$.
  end if
  if rule is RLS **then**
    Set $p_i^{(j)} = w_i/d_\lambda$.
  end if
  if rule is Oracle **then**
    Compute $\mathbf{x}^{\star(j)} = \mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top + \lambda \mathbf{I}_n)^{-1}\mathbf{b}^{(j)}$.
    Set $p_i^{(j)} \propto |\mathbf{x}_i^{\star(j)}|\sqrt{w_i}$.
  end if
  Draw a fresh $\mathbf{S}^{(j)}$ of size $s^{(j)}$ using Algorithm 2.
  Form $\mathbf{H}_j = \mathbf{A}\mathbf{S}^{(j)}\mathbf{S}^{(j)\top}\mathbf{A}^\top + \lambda \mathbf{I}_n$ and solve $\mathbf{H}_j \mathbf{y}^{(j)} = \mathbf{b}^{(j)}$.
  Set $\widetilde{\mathbf{x}}^{(j)} = \mathbf{A}^\top \mathbf{y}^{(j)}$.
end for
Return: $\widehat{\mathbf{x}}^\star = \sum_{j=1}^t \widetilde{\mathbf{x}}^{(j)}$.

sketching iterations, including leverage score, ridge leverage score and orcale sampling. Adaptive and mixture sampling algorithms appear in Appendix A. For completeness, the exact condition and convergence results of Chowdhury et al. (2018) are summarized in Appendix D.3.

3 Leverage Sampling

We first consider fresh leverage score sketches, with $p_i = \|\mathbf{V}_{i*}\|_2^2/n$ and deterministic sketch sizes $s^{(1)}, \ldots, s^{(t)}$. Define

$$\widehat{\mathbf{E}}^{(j)} = \mathbf{V}^\top \mathbf{S}^{(j)}\mathbf{S}^{(j)\top}\mathbf{V} - \mathbf{I}_n, \quad \mathbf{G} = \mathbf{I}_n + \lambda \mathbf{\Sigma}^{-2},$$
$$\widehat{\mathbf{R}}^{(j)} = (\mathbf{I}_n + \mathbf{G}^{-1}\widehat{\mathbf{E}}^{(j)})^{-1} - \mathbf{I}_n = -(\mathbf{G} + \widehat{\mathbf{E}}^{(j)})^{-1}\widehat{\mathbf{E}}^{(j)}. \tag{3}$$

The inverse exists for every sketch because $\mathbf{G} + \widehat{\mathbf{E}}^{(j)} \succeq \lambda \mathbf{\Sigma}^{-2} \succ \mathbf{0}$. The next two lemmas give the error identity and the variance along one vector.

Lemma 3.1. *For any sequence of sketches in* (2)*,*

$$\widehat{\mathbf{x}}^\star - \mathbf{x}^\star = (-1)^{t-1}\mathbf{V}\widehat{\mathbf{R}}^{(t)} \cdots \widehat{\mathbf{R}}^{(1)}\mathbf{V}^\top \mathbf{x}^\star. \tag{4}$$

Lemma 3.2. *For independent leverage score sketches and every $\mathcal{F}_{j-1}$-measurable vector $\mathbf{z} \in \mathbb{R}^n$ with finite second moment,*

$$\mathbb{E}\Big[\|\widehat{\mathbf{E}}^{(j)}\mathbf{z}\|_2^2 \mid \mathcal{F}_{j-1}\Big] = \frac{n-1}{s^{(j)}}\|\mathbf{z}\|_2^2. \tag{5}$$

Appendix E.1 contains the proof of (4) and the proof of (5). The above product identifies the key matrix-vector product that quantifies the final solution quality of the iterative ridge regression solver. Instead of trying to obtain uniform control over $\|\mathbf{V}^\top \mathbf{S}\mathbf{S}^\top \mathbf{V} - \mathbf{I}_n\|_2$ as in Chowdhury et al. (2018), we only need to control $\|\widehat{\mathbf{E}}^{(j)}\mathbf{z}\|_2 = \|\mathbf{V}^\top \mathbf{S}\mathbf{S}^\top \mathbf{V}\mathbf{z} - \mathbf{z}\|_2$. We now present our improved convergence result for leverage score sampling.

Theorem 3.3. *Running Algorithm 1 with LS gives*

$$\mathbb{E}\big[\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2^2\big] \leq \left(\frac{\sigma_{\max}^2}{\lambda}\right)^{2t} \prod_{j=1}^t \frac{n-1}{s^{(j)}}\|\mathbf{x}^\star\|_2^2. \tag{6}$$

For any fixed t and $\delta_0 \in (0,1)$, with probability at least $1-\delta_0$,

$$\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2 \le \delta_0^{-1/2} \left(\frac{\sigma_{\max}^2}{\lambda}\right)^t \prod_{j=1}^{t} \sqrt{\frac{n-1}{s^{(j)}}} \|\mathbf{x}^\star\|_2. \tag{7}$$

In particular, for $\varepsilon \in (0,1)$, a common sketch size

$$s \ge \left(\frac{\sigma_{\max}^2}{\lambda}\right)^2 \frac{n-1}{\varepsilon^2} \tag{8}$$

gives $\mathbb{E}[\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2^2] \le \varepsilon^{2t}\|\mathbf{x}^\star\|_2^2$ and $\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2 \le \delta_0^{-1/2}\varepsilon^t\|\mathbf{x}^\star\|_2$ with probability at least $1-\delta_0$.

The proof in Appendix E.2 uses $\|(\mathbf{G} + \widehat{\mathbf{E}}^{(j)})^{-1}\|_2 \le \sigma_{\max}^2/\lambda$ in (3), then applies (5) conditionally at each iteration. One application of Markov's inequality to the final error gives (7); substituting (8) gives the common-size bounds. The required size is $O((\sigma_{\max}^2/\lambda)^2 n/\varepsilon^2)$, where Chowdhury et al. (2018) requires $O(n \log n/\varepsilon^2)$. When $\sigma_{\max}^2/\lambda$ is controlled, fresh sketches can save a factor of $\log n$ for the required sketch size.

A separate bound requires larger sketch sizes and replaces $\sigma_{\max}^2/\lambda$ by $(1-\varepsilon)^{-1}$. It yields an additional contraction factor of size $O(\frac{1}{\sqrt{\log n}})$ (Theorem E.2, proved in Appendix E.3).

Bias and sharper contraction bounds. We additionally analyze the algorithmic bias, i.e., the expected error vector $\mathbb{E}[\widehat{\mathbf{x}}^\star - \mathbf{x}^\star]$. Fix $0 < \varepsilon < 1$ and let $\mathbf{M}_j = \mathbf{G}^{-1}\widehat{\mathbf{E}}^{(j)}$, $C = \frac{\sigma_{\max}^2}{\lambda} + \frac{\sigma_{\max}^2}{\sigma_{\min}^2}$, and $\delta(s) = 4(n+1)\exp(-3\varepsilon^2 s/(32n))$. For $s^{(j)} \ge 8n/\varepsilon^2$, unbiasedness and an inverse perturbation bound give

$$\mathbb{E}[\widehat{\mathbf{R}}^{(j)}] = \mathbb{E}[(\mathbf{I}_n + \mathbf{M}_j)^{-1}\mathbf{M}_j^2], \qquad \|\mathbb{E}[\widehat{\mathbf{R}}^{(j)}]\|_2 \le \frac{\varepsilon^2}{2} + C(n-1)^2\delta(s^{(j)}).$$

Independence multiplies these factors in the algorithmic bias bound. When $C(n-1)^2\delta(s^{(j)}) \le \varepsilon^2/2$, this gives $\|\mathbb{E}[\widehat{\mathbf{x}}^\star - \mathbf{x}^\star]\|_2 \le \varepsilon^{2t}\|\mathbf{x}^\star\|_2$; Theorem E.3 states a sufficient sketch size. Its proof is deferred to Appendix E.4.

4 Ridge Leverage Sampling

We now consider independent ridge leverage score sketches, with $p_i = w_i/d_\lambda$ and sketch sizes $s^{(1)}, \dots, s^{(t)}$. Set $\mathbf{X} = \mathbf{V}\mathbf{\Sigma}_\lambda$, $w_i = \|\mathbf{X}_{i*}\|_2^2$, and

$$\mathbf{E}^{(j)} = \mathbf{X}^\top(\mathbf{S}^{(j)}\mathbf{S}^{(j)\top} - \mathbf{I}_d)\mathbf{X}, \qquad c_\lambda = \frac{\sigma_{\max}^2 + \lambda}{\lambda}. \tag{9}$$

The next two lemmas relate the error in one residual solve to a sketched matrix-vector product and bound its expected squared 2-norm.

Lemma 4.1. *For every sketch, $\mathbf{I}_n + \mathbf{E}^{(j)} \succ 0$ and $\|(\mathbf{I}_n + \mathbf{E}^{(j)})^{-1}\|_2 \le c_\lambda$. Moreover,*

$$\mathbf{x}^{\star(j)} - \widetilde{\mathbf{x}}^{(j)} = \mathbf{X}(\mathbf{I}_n + \mathbf{E}^{(j)})^{-1}\mathbf{X}^\top(\mathbf{S}^{(j)}\mathbf{S}^{(j)\top} - \mathbf{I}_d)\mathbf{x}^{\star(j)}.$$

Consequently, the norm of this error is at most $c_\lambda\|\mathbf{X}^\top(\mathbf{S}^{(j)}\mathbf{S}^{(j)\top} - \mathbf{I}_d)\mathbf{x}^{\star(j)}\|_2$. If $\|\mathbf{E}^{(j)}\|_2 \le \varepsilon < 1$, we have the sharper bound

$$\|\mathbf{x}^{\star(j)} - \widetilde{\mathbf{x}}^{(j)}\|_2 \le \frac{1}{1-\varepsilon}\|\mathbf{X}^\top(\mathbf{S}^{(j)}\mathbf{S}^{(j)\top} - \mathbf{I}_d)\mathbf{x}^{\star(j)}\|_2. \tag{10}$$

Lemma 4.2. *For independent ridge leverage score sketches and every $\mathcal{F}_{j-1}$-measurable vector $\mathbf{x} \in \mathbb{R}^d$ with finite second moment,*

$$\mathbb{E}\Big[\|\mathbf{X}^\top(\mathbf{S}^{(j)}\mathbf{S}^{(j)\top} - \mathbf{I}_d)\mathbf{x}\|_2^2 \mid \mathcal{F}_{j-1}\Big] \le \frac{d_\lambda}{s^{(j)}}\|\mathbf{x}\|_2^2. \tag{11}$$

Appendix F.1 contains the proof of Lemma 4.1 and the proof of Lemma 4.2. The first lemma isolates the sketch error applied to the current residual solution. The second controls that product conditionally on previous sketches. Together they give the following convergence result.

Theorem 4.3 (Fresh ridge leverage sampling)**.** *Let* $c_\lambda = (\sigma_{\max}^2 + \lambda)/\lambda$ *as in* (9)*. Running Algorithm 1 with RLS gives*

$$\mathbb{E}\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2^2 \le c_\lambda^{2t} \prod_{j=1}^{t} \frac{d_\lambda}{s^{(j)}} \|\mathbf{x}^\star\|_2^2. \tag{12}$$

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

$$\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2 \le \delta^{-1/2} c_\lambda^{t} \prod_{j=1}^{t} \sqrt{\frac{d_\lambda}{s^{(j)}}}\, \|\mathbf{x}^\star\|_2. \tag{13}$$

In particular, for $\varepsilon \in (0,1)$*, a common sketch size*

$$s \ge \frac{c_\lambda^2 d_\lambda}{\varepsilon^2} \tag{14}$$

gives $\mathbb{E}\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2^2 \le \varepsilon^{2t}\|\mathbf{x}^\star\|_2^2$ *and* $\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2 \le \delta^{-1/2}\varepsilon^t \|\mathbf{x}^\star\|_2$ *with probability at least* $1-\delta$*.*

Set $\mathbf{e}_j = \mathbf{x}^{\star(j)} - \widetilde{\mathbf{x}}^{(j)}$ for $j = 0, \dots, t$. The proof in Appendix F.2 combines Lemmas 4.1 and 4.2 with $\mathbf{x}^{\star(j)} = \mathbf{e}_{j-1}$ to obtain $\mathbb{E}\big[\|\mathbf{e}_j\|_2^2 \mid \mathcal{F}_{j-1}\big] \le c_\lambda^2 \frac{d_\lambda}{s^{(j)}} \|\mathbf{e}_{j-1}\|_2^2$.

Taking expectations and iterating this inequality with the tower property proves (12), since $\mathbf{e}_0 = \mathbf{x}^\star$ and $\mathbf{e}_t = \mathbf{x}^\star - \widehat{\mathbf{x}}^\star$. One application of Markov's inequality gives (13). Comparing to Chowdhury et al. (2018), we again save a factor of $\log d_\lambda$ on the sketch size when c_λ is moderate.

Similarly, we provide a separate bound that replaces c_λ by $(1-\varepsilon)^{-1}$ at the cost of a larger sketch size, which also yields an additional contraction factor of size $O(\frac{1}{\sqrt{\log d_\lambda}})$. Theorem F.3 gives the full statement and conditions; its proof is in Appendix F.3.

5 Residual-Aware Sampling

The ridge leverage scores used by Chowdhury et al. (2018) depend only on the design matrix. We develop probabilities that also use the current residual solution. For a fixed $\mathbf{x} \in \mathbb{R}^d$ with nonzero coordinates, write $Z(\mathbf{x}) = \sum_i |\mathbf{x}_i|\sqrt{w_i}$.

Proposition 5.1 (Optimal sampling for the matrix product)**.** *The distribution*

$$p_i^\star(\mathbf{x}) = \frac{|\mathbf{x}_i|\sqrt{w_i}}{Z(\mathbf{x})}$$

minimizes $\mathbb{E}\|\mathbf{X}^\top(\mathbf{S}\mathbf{S}^\top - \mathbf{I}_d)\mathbf{x}\|_2^2$*. Furthermore, at* $p_i^\star(\mathbf{x})$*,* $\mathbb{E}\|\mathbf{X}^\top(\mathbf{S}\mathbf{S}^\top - \mathbf{I}_d)\mathbf{x}\|_2^2 \le s^{-1} Z(\mathbf{x})^2$*.*

The proof is a direct application of Theorem 22 from Drineas & Mahoney (2017). Now recall from (11) that for ridge leverage scores, $\mathbb{E}\|\mathbf{X}^\top(\mathbf{S}\mathbf{S}^\top - \mathbf{I}_d)\mathbf{x}\|_2^2 \le s^{-1}\left(d_\lambda \|\mathbf{x}\|_2^2\right)$. We introduce the following factor to quantify the advantage of oracle sampling over ridge leverage score sampling.

$$\rho(\mathbf{x}) = \frac{Z(\mathbf{x})^2}{d_\lambda \|\mathbf{x}\|_2^2} \in (0,1]. \tag{15}$$

Cauchy–Schwarz gives $\rho(\mathbf{x}) \le 1$, with equality exactly when $|\mathbf{x}_i|$ is proportional to $\sqrt{w_i}$.

Surprisingly, our numerical experiments suggest that the value of $\rho(\mathbf{x})$ is very close to $2/\pi$. We provide some theoretical justifications below.

Fix a nonzero $\mathbf{x} \in \mathrm{range}(\mathbf{X})$, and write $\mathbf{x} = \mathbf{X}\mathbf{h}$. Draw an index i with probability w_i/d_λ over indices with $w_i > 0$, and define $\zeta = \frac{\sqrt{d_\lambda}\,\mathbf{x}_i}{\sqrt{w_i}\,\|\mathbf{x}\|_2}$. Since $\mathbf{x}_i/\sqrt{w_i} = \langle \mathbf{X}_{i*}/\|\mathbf{X}_{i*}\|_2,\, \mathbf{h}\rangle$, ζ is a scaled projection of $\mathbf{h}$ onto a randomly sampled row direction of $\mathbf{X}$. The following proposition identifies $2/\pi$ as the Gaussian reference value of $\rho(\mathbf{x})$ and quantifies the effect of departures from normality.

Proposition 5.2 (Gaussian reference for ρ)**.** *Let* P_ζ *denote the distribution of* ζ*, and define*

$$\Delta = W_1(P_\zeta, \mathcal{N}(0,1)),$$

where W_1 *denotes the* 1*-Wasserstein distance. Then*

$$\left|\rho(\mathbf{x}) - \frac{2}{\pi}\right| \le 2\sqrt{\frac{2}{\pi}}\,\Delta + \Delta^2.$$

The proof is provided in Appendix H.5, which also reports empirical evidence that Δ is small and generally decreases with scale for the language model application presented in Section 6.

Oracle convergence. Although $p^\star(\mathbf{x}^{\star(j)})$ depends on previous sketches, it is fixed before the next sketch is drawn. Let $\bar{\rho}_t = (\prod_{j=1}^t \rho(\mathbf{x}^{\star(j)}))^{1/t}$ be the geometric mean of the oracle factors. Corollary G.1, proved in Appendix G.1, therefore gives, for a common size s and $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2 \le \delta^{-1/2} \left(c_\lambda \sqrt{\frac{\bar{\rho}_t d_\lambda}{s}} \right)^t \|\mathbf{x}^\star\|_2. \tag{16}$$

where $c_\lambda = (\sigma_{\max}^2 + \lambda)/\lambda$ is defined in (9). If a fixed $\bar{\rho} \in (0,1]$ bounds every oracle factor at iterations with nonzero residual, then for $\varepsilon \in (0,1)$, $s \ge c_\lambda^2 \bar{\rho} d_\lambda / \varepsilon^2$ yields the same $\delta^{-1/2}\varepsilon^t$ relative error guarantee as Theorem 4.3. We note that $\bar{\rho}$ is unknown before running the algorithm. An upper bound on $\bar{\rho}$ would be interesting future work.

A practical approximation. The oracle requires the unknown residual solution. Algorithm 3 gives a practical approximation. It draws an initial ridge leverage sketch $\mathbf{S}^{(0)}$ of size s_0 and caches a factorization of $\mathbf{H}_0 = \mathbf{A}\mathbf{S}^{(0)}\mathbf{S}^{(0)\top}\mathbf{A}^\top + \lambda \mathbf{I}_n$. At iteration j, it forms

$$\widehat{\mathbf{x}}^{(j)} = \mathbf{A}^\top \mathbf{H}_0^{-1} \mathbf{b}^{(j)}, \qquad \widehat{p}_i^{(j)} = \frac{|\widehat{\mathbf{x}}_i^{(j)}| \sqrt{w_i}}{\sum_k |\widehat{\mathbf{x}}_k^{(j)}| \sqrt{w_k}}. \tag{17}$$

This requires a solve with the cached factorization and a multiplication by $\mathbf{A}^\top$, without a new pilot factorization. The proposal and residual are measurable with respect to $\mathcal{F}_{j-1}^{\mathrm{ad}} = \sigma(\mathbf{S}^{(0)}, \mathbf{S}^{(1)}, \ldots, \mathbf{S}^{(j-1)})$. For a proposal $p^{(j)}$, define its variance factor by

$$R_p^{(j)} = \frac{\sum_i |\mathbf{x}_i^{\star(j)}|^2 w_i / p_i^{(j)}}{d_\lambda \|\mathbf{x}^{\star(j)}\|_2^2}. \tag{18}$$

Ridge leverage gives $R_p^{(j)} = 1$ and the oracle gives $R_p^{(j)} = \rho(\mathbf{x}^{\star(j)})$. With $\bar{R}_{p,t} = (\prod_{j=1}^t R_p^{(j)})^{1/t}$, the adaptive guarantee replaces $\bar{\rho}_t$ by $\bar{R}_{p,t}$ in (16). Theorem F.2 proves the general bound in Appendix F.2; Corollary G.2 applies it to the pilot proposal in Appendix G.2. These factors can exceed one; approximation by a pilot sketch alone does not establish contraction.

A defensive mixture with weight $\alpha \in (0,1)$ controls the worst-case performance:

$$\widetilde{p}_i^{(j)} = (1-\alpha)\widehat{p}_i^{(j)} + \alpha \frac{w_i}{d_\lambda}, \qquad 0 < \alpha < 1. \tag{19}$$

Its sketching error is bounded by $1/\alpha$ times the error by ridge leverage score sampling, regardless of pilot accuracy. Consequently, a common iteration size $s \ge c_\lambda^2 d_\lambda / (\alpha \varepsilon^2)$ guarantees mean square contraction by ε^2 per iteration. Corollary G.3, proved in Appendix G.2, gives this guarantee and the comparison with the pure proposal.

6 Experiments

We compare uniform, leverage, ridge leverage, oracle, adaptive, and mixture sampling on ARCENE, synthetic data, and a ridge probe built from saved Qwen2.5 representations. The ARCENE and synthetic experiments extend the experiments of Chowdhury et al. (2018). All six methods use the same residual update with fresh sketches; a separate experiment compares fixed and fresh sketches (Appendix H.2). Each dataset uses the same six panels: (a) solution error, (b) objective excess, (c) a sketch size sweep, (d) a regularization sweep, (e) the oracle factor along each method's residual trajectory, and (f) the actual proposal factor for adaptive sampling and its mixture. Panel (d) displays regularization through its effective dimension d_λ.

Curves show medians and interquartile ranges across independent sketch trials. We report relative error $\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2 / \|\mathbf{x}^\star\|_2$ and objective excess $(f(\widehat{\mathbf{x}}^\star) - f(\mathbf{x}^\star))/f(\mathbf{x}^\star)$, evaluated stably as described in Appendix H.1. The displayed budget s is the baseline number of columns sketched per update. For a run of t updates, adaptive methods spend s columns on an initial pilot and $\lfloor s(t-1)/t \rfloor$ per update, so their total is at most st. The mixture uses ridge weight 0.1.

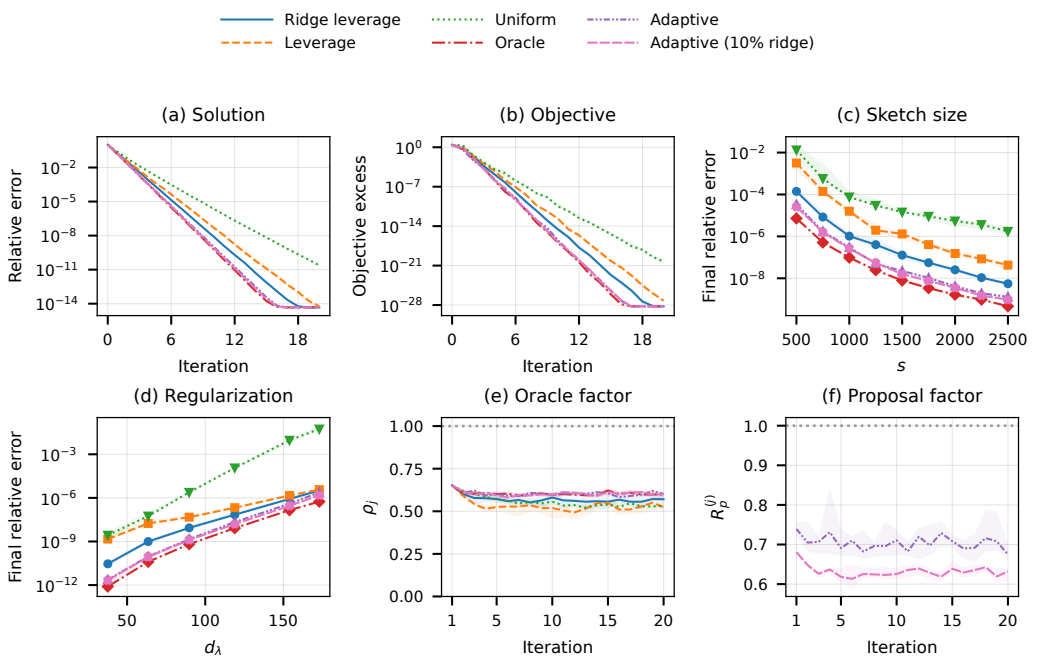


Figure 1: ARCENE, 15 trials. Panels (a,b,e,f) use 20 updates at $s = 2400$, $\lambda = 10$. Panels (c,d) use 10 updates: (c) varies s at $\lambda = 10$; (d) varies λ at $s = 2400$. Bands show interquartile ranges. Adaptive budgets include the pilot. Panel (e) shows all six methods; (f) shows adaptive sampling and its mixture.

Comparison with the leverage baselines. ARCENE (Guyon et al., 2004) gives a 200×10000 design, scaled globally to $[0, 1]$, with labels ± 1. We use 15 trials and $\lambda = 10$ except in the regularization sweep. In Figure 1(a,b), the oracle reaches numerical precision significantly faster than the 3 baselines. Adaptive and mixture methods behave very similarly to the orcale method. Their lower errors also persist across the sketch size and regularization sweeps in panels (c,d). We additionally perform the same experiments on the synthetic data used in Chen et al. (2015), which has $n = 500$, $d = 50000$, and we use 5 trials. The results are similar to ARCENE except that the gap between the 3 baselines becomes smaller. See the full result at Figure 3.

A separate fixed versus fresh comparison shows the benefit of refreshing sketches (Figure 4). At the two smaller budgets, median errors eventually grow for every reused-sketch baseline. At the smallest budget, fresh uniform and leverage sampling also show growing median errors, and fresh ridge leverage makes little progress. The residual-aware methods reduce errors at every tested budget. Appendix H.2 gives the full settings and additional plots.

Ridge probing of language model representations. We fit a ridge probe to frozen Qwen2.5-32B-Instruct representations (Qwen Team et al., 2025) for semantic textual similarity (Cer et al., 2017), using 100 training examples. Concatenating representations from 20 layers at the final retained non-padding token gives 102,400 features per example. We account for an unpenalized intercept by centering features and labels using training means. The regularization parameter λ is selected by 5-fold cross-validation on the training set. Appendix H.3 provides details of feature construction and parameter selection.

The results of Figure 2 resembel Figure 1 except that uniform sampling converges significantly slower. We again observe the superior performance of residual-aware sampling over the baselines.

Observed variance factors. Before every update, we record the oracle factor $\rho_j = \rho(\mathbf{x}^{\star(j)})$ along each method's residual trajectory and the actual proposal factor $R_p^{(j)}$ from (18). Panels (e,f) use the same trials and all 20 updates as (a,b). Panel (e) shows the best attainable directional factor for each method's current residual; panel (f) evaluates the probabilities used by adaptive sampling and

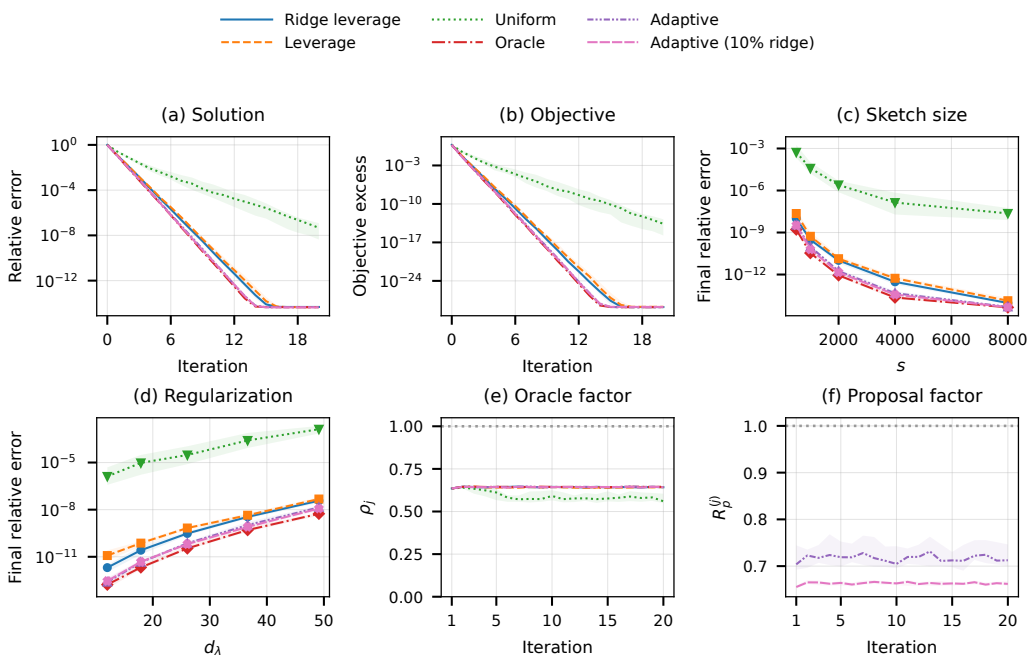


Figure 2: Qwen2.5-32B probe with 102400 features, 30 trials. Panels (a,b,e,f) use $s = 1000$ and 20 updates; (c) varies s with 10 updates. These panels use the cross-validated $\lambda = 10^{4.5}$. Panel (d) varies λ at $s = 1000$ with 10 updates. Bands show interquartile ranges. Adaptive budgets include the pilot. Panel (e) shows all six methods; (f) shows adaptive sampling and its mixture.

its mixture. Across the three datasets, the mixture has lower median proposal factors and generally narrower interquartile bands than pure adaptive sampling. The ridge leverage reference $R_p = 1$ is shown as a horizontal line.

7 Conclusion

We establish sharper convergence bounds for fresh sampling in the iterative column sampling framework of Chowdhury et al. (2018) for ridge regression. Building on the analysis of Kacham & Woodruff (2022), we analyze the variance of the sketched matrix-vector product and exploit conditional independence across iterations. This analysis yields an oracle sampling rule that minimizes the variance, along with practical adaptive and mixture strategies that approximate the oracle. Experiments on synthetic and real data, including ridge probes on pretrained Qwen2.5 hidden layers, demonstrate gains in solution accuracy.

A limitation of our analysis is that the sharper convergence guarantees for oracle, adaptive, and mixture sampling depend on factors determined by the residual trajectory and, for the practical methods, the quality of the sampling proposals. These factors are unknown before the algorithm runs. Although the oracle factor is at most one and the defensive mixture bounds its proposal factor by $1/\alpha$, these worst-case bounds do not quantify the improvements observed in our experiments. An open question is to identify conditions on the problem and the pilot approximation that yield sharper a priori bounds on these factors or their products across iterations. Such bounds could clarify when these sampling rules provably improve upon ridge leverage sampling and guide the choice of sketch sizes and mixture weights.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, code development and debugging, and literature search. All technical content, mathematical arguments, experimental design, and conclusions were developed and verified by the authors. The authors take full responsibility for the accuracy and integrity of the final manuscript.

REPRODUCIBILITY STATEMENT

The assumptions underlying our theoretical results are stated in Section 2 and the corresponding theorem statements. Proofs and supporting lemmas are provided in Appendices D–G, with the Gaussian reference result proved in Appendix H.5. Algorithm 1 and Appendix A specify the iterative solver and the adaptive and mixture sampling procedures. Appendix H.1 describes the numerical implementation, sampling budgets, evaluation metrics, and aggregation across trials. Appendix H.2 documents ARCENE preprocessing, synthetic data generation, and experimental configurations, including random seeds and trial counts. Appendix H.3 specifies the STS-B example selection, Qwen2.5 feature extraction, centering, regularization selection, and sampling configurations. Finally, Appendix H.4 details the measurement of the oracle and proposal factors.

REFERENCES

Ahmed Alaoui and Michael Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/f3f27a324736617f20abbf2ffd806f6d-Paper.pdf.

Haim Avron, Kenneth L. Clarkson, and David P. Woodruff. Faster kernel ridge regression using sketching and preconditioning. *SIAM Journal on Matrix Analysis and Applications*, 38 (4):1116–1138, 2017a. doi: 10.1137/16M1105396. URL https://doi.org/10.1137/16M1105396.

Haim Avron, Kenneth L. Clarkson, and David P. Woodruff. Sharper bounds for regularized data fitting. In Klaus Jansen, José D. P. Rolim, David P. Williamson, and Santosh S. Vempala (eds.), *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2017)*, volume 81 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 27:1–27:22, Dagstuhl, Germany, 2017b. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-044-6. doi: 10.4230/LIPIcs.APPROX-RANDOM.2017.27. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.APPROX-RANDOM.2017.27.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, March 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Tatiana Brailovskaya and Ramon van Handel. Universality and sharp matrix concentration inequalities. *Geometric and Functional Analysis*, 34(6):1734–1838, 2024.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel Cer, and David Jurgens (eds.), *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 1–14, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2001. URL https://aclanthology.org/S17-2001/.

Shouyuan Chen, Yang Liu, Michael R. Lyu, Irwin King, and Shengyu Zhang. Fast relative-error approximation algorithm for ridge regression. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, UAI'15, pp. 201–210, Arlington, Virginia, USA, 2015. AUAI Press. ISBN 9780996643108.

Shabarish Chenakkod, Michał Dereziński, Xiaoyu Dong, and Mark Rudelson. Optimal embedding dimension for sparse subspace embeddings. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, STOC 2024, pp. 1106–1117, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703836. doi: 10.1145/3618260.3649762. URL https://doi.org/10.1145/3618260.3649762.

Shabarish Chenakkod, Michał Dereziński, and Xiaoyu Dong. Optimal Oblivious Subspace Embeddings with Near-Optimal Sparsity. In Keren Censor-Hillel, Fabrizio Grandoni, Joël Ouaknine, and Gabriele Puppis (eds.), *52nd International Colloquium on Automata, Languages, and Programming (ICALP 2025)*, volume 334 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 55:1–55:20, Dagstuhl, Germany, 2025. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-372-0. doi: 10.4230/LIPIcs.ICALP.2025.55. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ICALP.2025.55.

Shabarish Chenakkod, Michał Dereziński, and Xiaoyu Dong. Optimal subspace embeddings: Resolving nelson-nguyen conjecture up to sub-polylogarithmic factors. In *Proceedings of the 2026 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 4501–4559, 2026. doi: 10.1137/1.9781611978971.163. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611978971.163.

Agniva Chowdhury, Jiasen Yang, and Petros Drineas. An iterative, sketching-based framework for ridge regression. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 989–998. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/chowdhury18a.html.

Kenneth L. Clarkson and David P. Woodruff. Low rank approximation and regression in input sparsity time. In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, STOC '13, pp. 81–90, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450320290. doi: 10.1145/2488608.2488620. URL https://doi.org/10.1145/2488608.2488620.

Michael B. Cohen, Jelani Nelson, and David P. Woodruff. Optimal approximate matrix product in terms of stable rank. In Ioannis Chatzigiannakis, Michael Mitzenmacher, Yuval Rabani, and Davide Sangiorgi (eds.), *43rd International Colloquium on Automata, Languages, and Programming (ICALP 2016)*, volume 55 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 11:1–11:14, Dagstuhl, Germany, 2016. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-013-2. doi: 10.4230/LIPIcs.ICALP.2016.11. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ICALP.2016.11.

Michael B. Cohen, Cameron Musco, and Christopher Musco. Input sparsity time low-rank approximation via ridge leverage score sampling. In *Proceedings of the 2017 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 1758–1777, 2017. doi: 10.1137/1.9781611974782.115. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611974782.115.

Michał Dereziński and Michael W. Mahoney. Determinantal point processes in randomized numerical linear algebra. *Notices of the American Mathematical Society*, 2021. URL https://www.ams.org/journals/notices/202101/noti2202/noti2202.html?adat=January%202021&trk=2202&pdfissue=202101&pdffile=rnoti-p34.pdf&cat=none&type=.html&cover=noti-jan-21-cov.jpg.

Michał Dereziński, Zhenyu Liao, Edgar Dobriban, and Michael Mahoney. Sparse sketches with small inversion bias. In Mikhail Belkin and Samory Kpotufe (eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 1467–1510. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/derezinski21a.html.

Petros Drineas and Michael W. Mahoney. Randnla: randomized numerical linear algebra. *Commun. ACM*, 59(6):80–90, May 2016. ISSN 0001-0782. doi: 10.1145/2842602. URL https://doi.org/10.1145/2842602.

Petros Drineas and Michael W. Mahoney. Lectures on randomized numerical linear algebra, 2017. URL https://arxiv.org/abs/1712.08880.

Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast monte carlo algorithms for matrices i: Approximating matrix multiplication. *SIAM Journal on Computing*, 36(1):132–157, 2006a. doi: 10.1137/S0097539704442684. URL https://doi.org/10.1137/S0097539704442684.

Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Sampling algorithms for ℓ_2 regression and applications. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1127–1136, 2006b. doi: 10.1145/1109557.1109682. URL https://doi.org/10.1145/1109557.1109682.

Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Subspace sampling and relative-error matrix approximation: Column-row-based methods. In Yossi Azar and Thomas Erlebach (eds.), *Algorithms – ESA 2006*, pp. 304–314, Berlin, Heidelberg, 2006c. Springer Berlin Heidelberg. ISBN 978-3-540-38876-0.

Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Relative-error CUR matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30(2):844–881, 2008. doi: 10.1137/07070471X. URL https://doi.org/10.1137/07070471X.

Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13 (111):3475–3506, 2012. URL http://jmlr.org/papers/v13/drineas12a.html.

Alon Gonen, Francesco Orabona, and Shai Shalev-Shwartz. Solving ridge regression using sketched preconditioned SVRG. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1397–1405, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/gonen16.html.

Wes Gurnee and Max Tegmark. Language models represent space and time. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 2483–2503, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/0a6059857ae5c82ea9726ee9282a7145-Paper-Conference.pdf.

Isabelle Guyon, Steve Gunn, Asa Ben-Hur, and Gideon Dror. Result analysis of the NIPS 2003 feature selection challenge. In L. Saul, Y. Weiss, and L. Bottou (eds.), *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004. URL https://proceedings.neurips.cc/paper_files/paper/2004/file/5e751896e527c862bf67251a474b3819-Paper.pdf.

N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011. doi: 10.1137/090771806. URL https://doi.org/10.1137/090771806.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634. URL https://doi.org/10.1080/00401706.1970.10488634.

Praneeth Kacham and David Woodruff. Optimal deterministic coresets for ridge regression. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 4141–4150. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/kacham20a.html.

Praneeth Kacham and David Woodruff. Sketching algorithms and lower bounds for ridge regression. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 10539–10556. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/kacham22a.html.

Jonathan Lacotte and Mert Pilanci. Faster least squares optimization, 2021. URL https://arxiv.org/abs/1911.02675.

Jonathan Lacotte, Yifei Wang, and Mert Pilanci. Adaptive Newton sketch: Linear-time optimization with quadratic convergence and effective hessian dimensionality. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5926–5936. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/lacotte21a.html.

Hanyu Li and Chengmei Niu. Optimal subsampling for high-dimensional ridge regression. *Knowledge-Based Systems*, 286:111426, 2024. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2024.111426. URL https://www.sciencedirect.com/science/article/pii/S0950705124000613.

Michael W. Mahoney. Randomized algorithms for matrices and data. *Found. Trends Mach. Learn.*, 3(2):123–224, February 2011. ISSN 1935-8237. doi: 10.1561/2200000035. URL https://doi.org/10.1561/2200000035.

Michael W. Mahoney and Petros Drineas. CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697–702, 2009. doi: 10.1073/pnas.0803205106. URL https://www.pnas.org/doi/abs/10.1073/pnas.0803205106.

Per-Gunnar Martinsson and Joel A. Tropp. Randomized numerical linear algebra: Foundations and algorithms. *Acta Numerica*, 29:403–572, 2020. doi: 10.1017/S0962492920000021.

Riley Murray, James Demmel, Michael W. Mahoney, N. Benjamin Erichson, Maksim Melnichenko, Osman Asif Malik, Laura Grigori, Piotr Luszczek, Michał Dereziński, Miles E. Lopes, Tianyu Liang, Hengrui Luo, and Jack Dongarra. Randomized numerical linear algebra : A perspective on the field with an eye to software, 2023. URL https://arxiv.org/abs/2302.11474.

Cameron Musco and Christopher Musco. Recursive sampling for the nystrom method. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/a03fa30821986dff10fc66647c84c9c3-Paper.pdf.

Jelani Nelson and Huy L. Nguyên. Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 117–126, 2013. doi: 10.1109/FOCS.2013.21.

Chengmei Niu, Zhenyu Liao, Zenan Ling, and Michael W. Mahoney. Fundamental bias in inverting random sampling matrices with application to sub-sampled Newton. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org, 2025.

Mert Pilanci and Martin J. Wainwright. Iterative Hessian sketch: fast and accurate solution approximation for constrained least-squares. *J. Mach. Learn. Res.*, 17(1):1842–1879, January 2016. ISSN 1532-4435.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Craig Saunders, Alexander Gammerman, and Volodya Vovk. Ridge regression learning algorithm in dual variables. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pp. 515–521, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 1558605568.

Joel A. Tropp. An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.*, 8(1–2):1–230, May 2015. ISSN 1935-8237. doi: 10.1561/2200000048. URL https://doi.org/10.1561/2200000048.

Jialei Wang, Jason Lee, Mehrdad Mahdavi, Mladen Kolar, and Nati Srebro. Sketching meets random projection in the dual: A provable recovery algorithm for big and high-dimensional data. In Aarti Singh and Jerry Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1150–1158. PMLR, 20–22 Apr 2017a. URL https://proceedings.mlr.press/v54/wang17d.html.

Shusen Wang, Alex Gittens, and Michael W. Mahoney. Sketched ridge regression: Optimization perspective, statistical perspective, and model averaging. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3608–3616. PMLR, 06–11 Aug 2017b. URL https://proceedings.mlr.press/v70/wang17c.html.

David P. Woodruff. Sketching as a tool for numerical linear algebra. *Found. Trends Theor. Comput. Sci.*, 10(1–2):1–157, October 2014. ISSN 1551-305X. doi: 10.1561/0400000060. URL https://doi.org/10.1561/0400000060.

Fangwei Zhu, Damai Dai, and Zhifang Sui. Language models encode the value of numbers linearly. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 693–709, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.47/.

A ALGORITHMS

Algorithm 2 Sampling and rescaling (Chowdhury et al., 2018)

Input: Distribution p on $\{1, \dots, d\}$; integer $s \geq 1$.
Initialize $\mathbf{S} = \mathbf{0}_{d \times s}$.
for $k = 1, \dots, s$ **do**
  Independently draw i_k with probability p_{i_k}.
  Set $\mathbf{S}_{i_k,k} = 1/\sqrt{s p_{i_k}}$.
end for
Return: $\mathbf{S}$.

Algorithm 3 Adaptive sampling with a cached pilot factorization

Input: $\mathbf{A}, \mathbf{b}$; regularization $\lambda > 0$; $t \geq 1$; positive integer sizes $s_0, s^{(1)}, \dots, s^{(t)}$; mixture weight $\alpha \in [0, 1)$.
Draw a pilot sketch $\mathbf{S}^{(0)}$ of size s_0 using Algorithm 2 with $p_i = w_i/d_\lambda$. Factor $\mathbf{H}_0 = \mathbf{A}\mathbf{S}^{(0)}\mathbf{S}^{(0)\top}\mathbf{A}^\top + \lambda \mathbf{I}_n$.
Initialize $\mathbf{b}^{(0)} = \mathbf{b}$, $\widetilde{\mathbf{x}}^{(0)} = \mathbf{0}_d$, and $\mathbf{y}^{(0)} = \mathbf{0}_n$.
for $j = 1, \dots, t$ **do**
  Set $\mathbf{b}^{(j)} = \mathbf{b}^{(j-1)} - \lambda \mathbf{y}^{(j-1)} - \mathbf{A}\widetilde{\mathbf{x}}^{(j-1)}$.
  Solve $\mathbf{H}_0 \mathbf{u} = \mathbf{b}^{(j)}$ using the cached factorization; set $\widehat{\mathbf{x}}^{(j)} = \mathbf{A}^\top \mathbf{u}$.
  Let $z = \sum_i |\widehat{\mathbf{x}}_i^{(j)}| \sqrt{w_i}$.
  Set $\widehat{p}_i^{(j)} = |\widehat{\mathbf{x}}_i^{(j)}| \sqrt{w_i}/z$.
  Set $p_i^{(j)} = (1 - \alpha)\widehat{p}_i^{(j)} + \alpha w_i/d_\lambda$.
  Draw a fresh $\mathbf{S}^{(j)}$ of size $s^{(j)}$ using Algorithm 2 with $p^{(j)}$.
  Form $\mathbf{H}_j = \mathbf{A}\mathbf{S}^{(j)}\mathbf{S}^{(j)\top}\mathbf{A}^\top + \lambda \mathbf{I}_n$ and solve $\mathbf{H}_j \mathbf{y}^{(j)} = \mathbf{b}^{(j)}$.
  Set $\widetilde{\mathbf{x}}^{(j)} = \mathbf{A}^\top \mathbf{y}^{(j)}$.
end for
Return: $\widehat{\mathbf{x}}^\star = \sum_{j=1}^t \widetilde{\mathbf{x}}^{(j)}$.

B RUNNING TIME ANALYSIS

We discuss the running time of Algorithm 1 with LS or RLS, assuming the exact leverage scores or ridge leverage scores are provided. Approximations can be computed efficiently (Drineas et al., 2012; Clarkson & Woodruff, 2013; Cohen et al., 2017). Within one iteration of (2), the two products $\mathbf{A}\widetilde{\mathbf{x}}^{(j-1)}$ and $\mathbf{A}^\top \mathbf{y}^{(j)}$ cost $O(\mathsf{nnz}(\mathbf{A}))$, constructing $\mathbf{A}\mathbf{S}^{(j)}$ costs $O(ns)$, forming $\mathbf{H}_j$ costs $O(n^2 s)$, and factoring and solving the resulting $n \times n$ system costs $O(n^3)$. The remaining vector operations are dominated by these costs. Thus each iteration costs $O(\mathsf{nnz}(\mathbf{A}) + n^2 \max(n, s))$.

Relative solution error. Fix $\varepsilon, \varepsilon_{\mathtt{alg}}, \delta \in (0, 1)$, where ε is the per-iteration contraction parameter. By Theorems 3.3 and 4.3, choosing the common sketch sizes

$$s_{\mathrm{LS}} = \left\lceil \left(\frac{\sigma_{\max}^2}{\lambda} \right)^2 \frac{n-1}{\varepsilon^2} \right\rceil, \qquad s_{\mathrm{RLS}} = \left\lceil \frac{c_\lambda^2 d_\lambda}{\varepsilon^2} \right\rceil \tag{20}$$

with $c_\lambda = (\sigma_{\max}^2 + \lambda)/\lambda$ from (9) gives $\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2 \leq \delta^{-1/2} \varepsilon^t \|\mathbf{x}^\star\|_2$ with probability at least $1 - \delta$. Consequently, $\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2 \leq \varepsilon_{\mathtt{alg}} \|\mathbf{x}^\star\|_2$ is guaranteed after

$$t_{\mathrm{sol}} = \left\lceil \frac{\log(1/\varepsilon_{\mathtt{alg}}) + \frac{1}{2}\log(1/\delta)}{\log(1/\varepsilon)} \right\rceil$$

iterations. With the sketch sizes in (20), the running times for t iterations are

$$T_{\mathrm{LS}}(t) = O\left(\mathsf{nnz}(\mathbf{A}) + \frac{n^3}{\varepsilon^2}\left(\frac{\sigma_{\max}^2}{\lambda}\right)^2\right) t, \tag{21}$$

$$T_{\mathrm{RLS}}(t) = O\left(\mathsf{nnz}(\mathbf{A}) + n^3 + \frac{n^2 c_\lambda^2 d_\lambda}{\varepsilon^2}\right) t. \tag{22}$$

Here the LS simplification assumes $\lambda \le \sigma_{\max}^2$.

Relative objective error. Let $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_2^2$ and $\mathbf{e} = \widehat{\mathbf{x}}^\star - \mathbf{x}^\star$. The optimality condition $\mathbf{A}^\top(\mathbf{A}\mathbf{x}^\star - \mathbf{b}) + \lambda\mathbf{x}^\star = \mathbf{0}_d$ cancels the cross terms, giving

$$f(\widehat{\mathbf{x}}^\star) - f(\mathbf{x}^\star) = \|\mathbf{A}\mathbf{e}\|_2^2 + \lambda\|\mathbf{e}\|_2^2 \le (\sigma_{\max}^2 + \lambda)\|\mathbf{e}\|_2^2.$$

Since $f(\mathbf{x}^\star) \ge \lambda\|\mathbf{x}^\star\|_2^2$ and $c_\lambda = (\sigma_{\max}^2 + \lambda)/\lambda$, for $\mathbf{b} \ne \mathbf{0}_n$ we obtain

$$\frac{f(\widehat{\mathbf{x}}^\star) - f(\mathbf{x}^\star)}{f(\mathbf{x}^\star)} \le c_\lambda \frac{\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2^2}{\|\mathbf{x}^\star\|_2^2} \le \frac{c_\lambda}{\delta}\varepsilon^{2t}$$

with probability at least $1 - \delta$, with the sketch sizes in (20). If $\mathbf{b} = \mathbf{0}_n$, the solution and all iterates are zero, so the objective guarantee is trivial. To ensure $f(\widehat{\mathbf{x}}^\star) \le (1 + \varepsilon_{\mathtt{alg}})f(\mathbf{x}^\star)$ with probability at least $1 - \delta$, it suffices to take

$$t_{\mathrm{obj}} = \left\lceil \frac{\log\big(c_\lambda/(\delta\varepsilon_{\mathtt{alg}})\big)}{2\log(1/\varepsilon)} \right\rceil.$$

The corresponding running times are $T_{\mathrm{LS}}(t_{\mathrm{obj}})$ and $T_{\mathrm{RLS}}(t_{\mathrm{obj}})$ from (21)–(22), using the same sketch sizes as in (20).

C Related work

Sketching for ridge regression. Chen et al. (2015) give a fast relative-error approximation algorithm for ridge regression. Wang et al. (2017b) study the statistical and optimization impacts of classical sketch and Hessian sketch used for matrix ridge regression problems. Avron et al. (2017b) derive bounds for regularized objective approximation governed by effective dimension rather than rank. Chowdhury et al. (2018) introduce an iterative solver using leverage and ridge leverage sampling, and ask whether fresh randomness improves convergence. Kacham & Woodruff (2022) prove geometric convergence for the refreshed solver and improve sparsity guarantees for oblivious sketches. Their Lemma 3.3 controls sketch error in the direction of the exact residual solution, the same matrix-vector product analyzed in this paper.

Iterative sketching and preconditioning. Pilanci & Wainwright (2016) introduced the Iterative Hessian Sketch (IHS) for convex-constrained least-squares problems, using fresh random sketches to approximate the Hessian at each iteration and achieving geometric convergence with sketch dimensions governed by the intrinsic complexity of the constraint set. Wang et al. (2017a) connect Hessian sketching to dual random projection and develop accelerated methods using preconditioned conjugate gradients. Sketching also provides preconditioners for ridge regression with stochastic variance reduced gradient (Gonen et al., 2016) and for kernel ridge systems through random features (Avron et al., 2017a). Lacotte & Pilanci (2021) compare fixed and refreshed sketches used in iterative methods for solving overdetermined least-squares problems, where the sketches are data-oblivious. They give a closed-form formula for the expected error of IHS. Lacotte et al. (2021) consider choosing the sketch sizes adaptively.

Feature and column sampling for ridge regression. Leverage scores are used in least-squares regression (Drineas et al., 2006b) and low-rank matrix approximation via column and row sampling (Drineas et al., 2006c; 2008; Mahoney & Drineas, 2009). Ridge leverage scores connect sampling to effective dimension in kernel ridge regression (Alaoui & Mahoney, 2015) and support low-rank and Nyström approximation (Cohen et al., 2017; Musco & Musco, 2017). Li & Niu (2024) use sampling probabilities similar to our oracle proposal: applied to the residual problem, their ideal

rule is $p_i^{\mathrm{LN},(j)} \propto |\mathbf{x}_i^{\star(j)}| \|\mathbf{A}_{*i}\|_2$, while ours is $p_i^{\star(j)} \propto |\mathbf{x}_i^{\star(j)}| \sqrt{w_i}$. Complementarily, Kacham & Woodruff (2020) study deterministic coresets for multi-response ridge regression, obtaining coresets of size $O((d_\lambda + d')/\varepsilon)$, together with corresponding lower bounds, where d' denotes the number of response vectors. Their row-sampling coreset setting, however, is not directly comparable to our column-sampling framework.

Ridge probes for language models. Probing fits simple predictors to a model's internal representations (Belinkov, 2022). Ridge probes have been used to decode spatial and temporal coordinates (Gurnee & Tegmark, 2024) and numerical values (Zhu et al., 2025). Our application concerns computing a probe on fixed hidden features, with semantic similarity as the prediction target.

D Preliminaries and embedding bounds

D.1 An explicit embedding tail

Spectral approximation of matrix products can be controlled through stable rank (Cohen et al., 2016). For the column sampling laws studied here, matrix concentration gives spectral error bounds for sampled Gram matrices (Tropp, 2015). We use the explicit specialization in the proof of Chowdhury et al. (2018, Theorem 3), retaining its constants and admissibility condition below.

Lemma D.1 (Explicit failure probability)**.** *Let $\mathbf{X} \in \mathbb{R}^{d\times n}$ be nonzero with $\|\mathbf{X}\|_2 \le 1$, and let $\mathbf{S}$ be constructed by Algorithm 2 with $p_i = \|\mathbf{X}_{i*}\|_2^2/\|\mathbf{X}\|_F^2$. Fix $\varepsilon \in (0,1]$. If*

$$s \ge \frac{2\|\mathbf{X}\|_F^2}{\varepsilon^2}, \tag{23}$$

then

$$\mathbb{P}\left(\|\mathbf{X}^\top \mathbf{S}\mathbf{S}^\top \mathbf{X} - \mathbf{X}^\top \mathbf{X}\|_2 > \varepsilon\right) \le 4(1+\|\mathbf{X}\|_F^2) \exp\left(-\frac{s\varepsilon^2}{\|\mathbf{X}\|_F^2(2+2\varepsilon/3)}\right) \tag{24}$$

$$\le 4(1+\|\mathbf{X}\|_F^2) \exp\left(-\frac{3s\varepsilon^2}{8\|\mathbf{X}\|_F^2}\right). \tag{25}$$

In the concentration argument cited below, ρ_1 denotes the uniform summand bound, and $\mathbf{P}$ denotes the positive semidefinite variance bound.

Proof. In the proof of Chowdhury et al. (2018, Theorem 3), equations (119) and (124) give $\rho_1 = \|\mathbf{X}\|_F^2/s$ and $\|\mathbf{P}\|_2 = \|\mathbf{X}\|_F^2/s$, respectively. Thus the required admissibility condition is

$$\varepsilon \ge \sqrt{\frac{\|\mathbf{X}\|_F^2}{s}} + \frac{\|\mathbf{X}\|_F^2}{3s}. \tag{26}$$

Under (23), the right-hand side of (26) is at most

$$\frac{\varepsilon}{\sqrt{2}} + \frac{\varepsilon^2}{6} \le \left(\frac{1}{\sqrt{2}} + \frac{1}{6}\right)\varepsilon < \varepsilon,$$

so the condition is satisfied. Equation (128) of that proof now gives (24). Since $\varepsilon \le 1$, we have $2 + 2\varepsilon/3 \le 8/3$, which proves (25). □

D.2 Residual identities

Lemma D.2 (Residual decomposition)**.** *For the iteration in* (2)*, any sketches, and $j = 1, \dots, t$,*

$$\mathbf{x}^{\star(j)} = \mathbf{V}\boldsymbol{\Sigma}_\lambda^2 \boldsymbol{\Sigma}^{-1}\mathbf{U}^\top \mathbf{b}^{(j)}, \tag{27}$$

$$\mathbf{x}^{\star(j)} = \mathbf{x}^{\star(j-1)} - \widetilde{\mathbf{x}}^{(j-1)}, \tag{28}$$

$$\mathbf{x}^\star = \mathbf{x}^{\star(t)} + \sum_{j=1}^{t-1} \widetilde{\mathbf{x}}^{(j)}, \qquad \widehat{\mathbf{x}}^\star - \mathbf{x}^\star = \widetilde{\mathbf{x}}^{(t)} - \mathbf{x}^{\star(t)}. \tag{29}$$

Proof. These are the identities used by Chowdhury et al. (2018). Substitution of the SVD into the exact residual solution gives (27). Apply $\mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top + \lambda \mathbf{I}_n)^{-1}$ to the residual update in (2) to obtain (28). Telescoping from $\mathbf{x}^{\star(0)} = \mathbf{x}^\star$, with $\widetilde{\mathbf{x}}^{(0)} = \mathbf{0}_d$, gives (29). □

D.3 Fixed sketch comparison

For completeness, the two embedding conditions in Chowdhury et al. (2018, Theorems 1 and 2) are

$$\|\mathbf{V}^\top \mathbf{S}\mathbf{S}^\top \mathbf{V} - \mathbf{I}_n\|_2 \le \varepsilon/2,$$
$$\|\mathbf{\Sigma}_\lambda \mathbf{V}^\top \mathbf{S}\mathbf{S}^\top \mathbf{V}\mathbf{\Sigma}_\lambda - \mathbf{\Sigma}_\lambda^2\|_2 \le \varepsilon/(4\sqrt{2}).$$

For $\varepsilon \in (0,1)$, the first gives

$$\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2 \le \varepsilon^t \|\mathbf{x}^\star\|_2,$$

and the second gives

$$\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2 \le \frac{\varepsilon^t}{2}\left(\|\mathbf{x}^\star\|_2 + \frac{\|\mathbf{U}_{k,\perp}^\top \mathbf{b}\|_2}{\sqrt{2\lambda}}\right).$$

Here $k \in \{1, \cdots, n\}$ counts singular values with $\sigma_i^2 > \lambda$, and $\mathbf{U}_{k,\perp}$ contains the remaining left singular vectors. Leverage and ridge leverage sampling respectively satisfy these events with probability at least $1-\delta$, for $\delta \in (0,1)$, with sufficient sample sizes of order $O(\varepsilon^{-2} n \ln n)$ and $O(\varepsilon^{-2} d_\lambda \ln d_\lambda)$. One sketch is reused for all iterations in these bounds.

E Additional Results for Leverage Sampling

Throughout this section, sketches use independent sampling with replacement with $p_i = \|\mathbf{V}_{i*}\|_2^2/n$ and deterministic sizes $s^{(j)}$. Rescaling gives $\mathbb{E}[\mathbf{V}^\top \mathbf{S}^{(j)} \mathbf{S}^{(j)\top} \mathbf{V}] = \mathbf{I}_n$ and therefore $\mathbb{E}[\widehat{\mathbf{E}}^{(j)}] = \mathbf{0}$.

E.1 Error Identities and Vector Variance

We first record the commutation identity used to telescope the residuals.

Lemma E.1 (Commutation identity)**.** *Define* $\widetilde{\mathbf{R}}^{(j)} = \mathbf{I}_n - (\mathbf{I}_n + \widehat{\mathbf{E}}^{(j)}\mathbf{G}^{-1})^{-1}$. *Then*

$$\mathbf{G}^{-1}\widetilde{\mathbf{R}}^{(j)} = -\widehat{\mathbf{R}}^{(j)}\mathbf{G}^{-1}. \tag{30}$$

Proof. Both $\mathbf{I}_n + \mathbf{G}^{-1}\widehat{\mathbf{E}}^{(j)}$ and $\mathbf{I}_n + \widehat{\mathbf{E}}^{(j)}\mathbf{G}^{-1}$ are invertible, and

$$\mathbf{G}(\mathbf{I}_n + \mathbf{G}^{-1}\widehat{\mathbf{E}}^{(j)}) = \mathbf{G} + \widehat{\mathbf{E}}^{(j)} = (\mathbf{I}_n + \widehat{\mathbf{E}}^{(j)}\mathbf{G}^{-1})\mathbf{G}.$$

Inverting this identity gives

$$(\mathbf{I}_n + \mathbf{G}^{-1}\widehat{\mathbf{E}}^{(j)})^{-1}\mathbf{G}^{-1} = \mathbf{G}^{-1}(\mathbf{I}_n + \widehat{\mathbf{E}}^{(j)}\mathbf{G}^{-1})^{-1}.$$

Hence, using the closed forms (3) and the definition in Lemma E.1,

$$\begin{aligned}\mathbf{G}^{-1}\widetilde{\mathbf{R}}^{(j)} &= \mathbf{G}^{-1}\big[\mathbf{I}_n - (\mathbf{I}_n + \widehat{\mathbf{E}}^{(j)}\mathbf{G}^{-1})^{-1}\big] = \mathbf{G}^{-1} - \mathbf{G}^{-1}(\mathbf{I}_n + \widehat{\mathbf{E}}^{(j)}\mathbf{G}^{-1})^{-1} \\ &= \mathbf{G}^{-1} - (\mathbf{I}_n + \mathbf{G}^{-1}\widehat{\mathbf{E}}^{(j)})^{-1}\mathbf{G}^{-1} = -\big[(\mathbf{I}_n + \mathbf{G}^{-1}\widehat{\mathbf{E}}^{(j)})^{-1} - \mathbf{I}_n\big]\mathbf{G}^{-1} = -\widehat{\mathbf{R}}^{(j)}\mathbf{G}^{-1},\end{aligned}$$

which is (30). □

Proof of Lemma 3.1. By Lemma D.2, $\widehat{\mathbf{x}}^\star - \mathbf{x}^\star = \widetilde{\mathbf{x}}^{(t)} - \mathbf{x}^{\star(t)}$, and by Chowdhury et al. (2018, Lemma 11) this equals $\mathbf{V}\widehat{\mathbf{R}}^{(t)}\mathbf{G}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\top \mathbf{b}^{(t)}$. We now unwind $\mathbf{b}^{(t)}$. For $j = 2, \ldots, t$, the updates of Algorithm 1 give $\mathbf{b}^{(j)} = \mathbf{b}^{(j-1)} - (\mathbf{A}\mathbf{A}^\top + \lambda\mathbf{I}_n)(\mathbf{A}\mathbf{S}^{(j-1)}\mathbf{S}^{(j-1)\top}\mathbf{A}^\top + \lambda\mathbf{I}_n)^{-1}\mathbf{b}^{(j-1)}$ (collecting $\lambda\mathbf{y}^{(j-1)} + \mathbf{A}\widetilde{\mathbf{x}}^{(j-1)} = (\mathbf{A}\mathbf{A}^\top + \lambda\mathbf{I}_n)\mathbf{y}^{(j-1)}$). Using the thin SVD and $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_n$,

$$\begin{aligned}\mathbf{A}\mathbf{A}^\top + \lambda\mathbf{I}_n &= \mathbf{U}\mathbf{\Sigma}(\mathbf{I}_n + \lambda\mathbf{\Sigma}^{-2})\mathbf{\Sigma}\mathbf{U}^\top = \mathbf{U}\mathbf{\Sigma}\mathbf{G}\mathbf{\Sigma}\mathbf{U}^\top, \\ \mathbf{A}\mathbf{S}^{(j-1)}\mathbf{S}^{(j-1)\top}\mathbf{A}^\top + \lambda\mathbf{I}_n &= \mathbf{U}\mathbf{\Sigma}\big(\mathbf{V}^\top\mathbf{S}^{(j-1)}\mathbf{S}^{(j-1)\top}\mathbf{V} + \lambda\mathbf{\Sigma}^{-2}\big)\mathbf{\Sigma}\mathbf{U}^\top \\ &= \mathbf{U}\mathbf{\Sigma}(\widehat{\mathbf{E}}^{(j-1)} + \mathbf{G})\mathbf{\Sigma}\mathbf{U}^\top,\end{aligned}$$

where the last equality uses $\widehat{\mathbf{E}}^{(j-1)} + \mathbf{G} = (\mathbf{V}^\top\mathbf{S}^{(j-1)}\mathbf{S}^{(j-1)\top}\mathbf{V} - \mathbf{I}_n) + (\mathbf{I}_n + \lambda\mathbf{\Sigma}^{-2})$. Hence

$$\begin{aligned}(\mathbf{A}\mathbf{A}^\top + \lambda\mathbf{I}_n)(\mathbf{A}\mathbf{S}^{(j-1)}\mathbf{S}^{(j-1)\top}\mathbf{A}^\top + \lambda\mathbf{I}_n)^{-1} &= \mathbf{U}\mathbf{\Sigma}\mathbf{G}\mathbf{\Sigma}\mathbf{U}^\top \cdot \mathbf{U}\mathbf{\Sigma}^{-1}(\widehat{\mathbf{E}}^{(j-1)} + \mathbf{G})^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\top \\ &= \mathbf{U}\mathbf{\Sigma}\mathbf{G}(\mathbf{G} + \widehat{\mathbf{E}}^{(j-1)})^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\top.\end{aligned}$$

Factor $\mathbf{G}+\widehat{\mathbf{E}}^{(j-1)} = (\mathbf{I}_n+\widehat{\mathbf{E}}^{(j-1)}\mathbf{G}^{-1})\mathbf{G}$, so $\mathbf{G}(\mathbf{G}+\widehat{\mathbf{E}}^{(j-1)})^{-1} = \mathbf{G}\mathbf{G}^{-1}(\mathbf{I}_n+\widehat{\mathbf{E}}^{(j-1)}\mathbf{G}^{-1})^{-1} = (\mathbf{I}_n + \widehat{\mathbf{E}}^{(j-1)}\mathbf{G}^{-1})^{-1}$. Therefore

$$\begin{aligned}\mathbf{b}^{(j)} &= \mathbf{b}^{(j-1)} - \mathbf{U}\mathbf{\Sigma}(\mathbf{I}_n + \widehat{\mathbf{E}}^{(j-1)}\mathbf{G}^{-1})^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\top\mathbf{b}^{(j-1)} \\ &= \mathbf{U}\mathbf{\Sigma}\big[\mathbf{I}_n - (\mathbf{I}_n + \widehat{\mathbf{E}}^{(j-1)}\mathbf{G}^{-1})^{-1}\big]\mathbf{\Sigma}^{-1}\mathbf{U}^\top\mathbf{b}^{(j-1)} = \mathbf{U}\mathbf{\Sigma}\,\widetilde{\mathbf{R}}^{(j-1)}\mathbf{\Sigma}^{-1}\mathbf{U}^\top\mathbf{b}^{(j-1)},\end{aligned}$$

using $\mathbf{b}^{(j-1)} = \mathbf{U}\mathbf{\Sigma}\,\mathbf{\Sigma}^{-1}\mathbf{U}^\top\mathbf{b}^{(j-1)}$ and $\mathbf{I}_n - (\mathbf{I}_n + \widehat{\mathbf{E}}^{(j-1)}\mathbf{G}^{-1})^{-1} = \widetilde{\mathbf{R}}^{(j-1)}$, which is the definition in Lemma E.1. Premultiplying by $\mathbf{\Sigma}^{-1}\mathbf{U}^\top$ and using $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_n$, $\mathbf{\Sigma}^{-1}\mathbf{\Sigma} = \mathbf{I}_n$,

$$\mathbf{\Sigma}^{-1}\mathbf{U}^\top\mathbf{b}^{(j)} = \widetilde{\mathbf{R}}^{(j-1)}\mathbf{\Sigma}^{-1}\mathbf{U}^\top\mathbf{b}^{(j-1)}. \tag{31}$$

Apply (31) repeatedly, starting from $\widehat{\mathbf{x}}^\star - \mathbf{x}^\star = \mathbf{V}\widehat{\mathbf{R}}^{(t)}\mathbf{G}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\top\mathbf{b}^{(t)}$ and ending at $\mathbf{b}^{(1)} = \mathbf{b}^{(0)} = \mathbf{b}$:

$$\widehat{\mathbf{x}}^\star - \mathbf{x}^\star = \mathbf{V}\widehat{\mathbf{R}}^{(t)}\mathbf{G}^{-1}\,\widetilde{\mathbf{R}}^{(t-1)}\cdots\widetilde{\mathbf{R}}^{(1)}\,\mathbf{\Sigma}^{-1}\mathbf{U}^\top\mathbf{b}.$$

Finally, move $\mathbf{G}^{-1}$ rightward through the $(t-1)$ factors $\widetilde{\mathbf{R}}$ using $\mathbf{G}^{-1}\widetilde{\mathbf{R}}^{(j)} = -\widehat{\mathbf{R}}^{(j)}\mathbf{G}^{-1}$ from (30); each pass flips a sign, producing $(-1)^{t-1}$:

$$\mathbf{G}^{-1}\widetilde{\mathbf{R}}^{(t-1)}\cdots\widetilde{\mathbf{R}}^{(1)} = (-1)^{t-1}\widehat{\mathbf{R}}^{(t-1)}\cdots\widehat{\mathbf{R}}^{(1)}\mathbf{G}^{-1}.$$

Substituting gives $\widehat{\mathbf{x}}^\star - \mathbf{x}^\star = (-1)^{t-1}\mathbf{V}\widehat{\mathbf{R}}^{(t)}\widehat{\mathbf{R}}^{(t-1)}\cdots\widehat{\mathbf{R}}^{(1)}\mathbf{G}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\top\mathbf{b}$, which is (4). The terminal expression equals that in (4) because $\mathbf{G}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\top\mathbf{b} = \mathbf{V}^\top\mathbf{x}^\star$.

□

The terminal vector can also be written as

$$\mathbf{V}^\top\mathbf{x}^\star = \mathbf{G}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\top\mathbf{b}, \qquad \|\mathbf{V}^\top\mathbf{x}^\star\|_2 = \|\mathbf{x}^\star\|_2. \tag{32}$$

Thus, for any sketches satisfying $\|\widehat{\mathbf{R}}^{(j)}\|_2 \le \varepsilon$,

$$\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2 \le \varepsilon^t\|\mathbf{x}^\star\|_2.$$

No independence assumption is required for these identities.

Proof of Lemma 3.2. Write $\mathbf{v}_i^\top := \mathbf{V}_{i*}$. If $i_1, \ldots, i_{s^{(j)}}$ are the sampled indices for $\mathbf{S}^{(j)}$, then

$$\widehat{\mathbf{E}}^{(j)}\mathbf{z} = \frac{1}{s^{(j)}}\sum_{\ell=1}^{s^{(j)}}\frac{1}{p_{i_\ell}}\mathbf{v}_{i_\ell}\mathbf{v}_{i_\ell}^\top\mathbf{z} - \mathbf{z}.$$

Writing $\mathbf{w}_\ell^{(j)} := p_{i_\ell}^{-1}\mathbf{v}_{i_\ell}\mathbf{v}_{i_\ell}^\top\mathbf{z}$ for the summands, these are i.i.d. with mean

$$\mathbb{E}[\mathbf{w}_1^{(j)}] = \sum_{i=1}^d p_i\, p_i^{-1}\mathbf{v}_i\mathbf{v}_i^\top\mathbf{z} = \Big(\sum_{i=1}^d \mathbf{v}_i\mathbf{v}_i^\top\Big)\mathbf{z} = \mathbf{V}^\top\mathbf{V}\,\mathbf{z} = \mathbf{z},$$

where $\sum_i \mathbf{v}_i\mathbf{v}_i^\top = \mathbf{V}^\top\mathbf{V} = \mathbf{I}_n$ because $\mathbf{V}$ has orthonormal columns. Their second moment is

$$\begin{aligned}\mathbb{E}\Big[\|\mathbf{w}_1^{(j)}\|_2^2\Big] &= \sum_{i=1}^d p_i\,\frac{1}{p_i^2}\|\mathbf{v}_i\|_2^2(\mathbf{v}_i^\top\mathbf{z})^2 \\ &= \sum_{i=1}^d \frac{1}{p_i}\|\mathbf{v}_i\|_2^2(\mathbf{v}_i^\top\mathbf{z})^2 \\ &= n\sum_{i=1}^d(\mathbf{v}_i^\top\mathbf{z})^2 = n\|\mathbf{V}\mathbf{z}\|_2^2 = n\|\mathbf{z}\|_2^2.\end{aligned}$$

Since $\widehat{\mathbf{E}}^{(j)}\mathbf{z} = \frac{1}{s^{(j)}}\sum_{\ell=1}^{s^{(j)}}(\mathbf{w}_\ell^{(j)} - \mathbf{z})$ is an average of i.i.d. mean-zero vectors, the cross terms vanish, so

$$\mathbb{E}\Big[\|\widehat{\mathbf{E}}^{(j)}\mathbf{z}\|_2^2\Big] = \frac{1}{s^{(j)}}\,\mathbb{E}\Big[\|\mathbf{w}_1^{(j)} - \mathbf{z}\|_2^2\Big] = \frac{1}{s^{(j)}}\Big(\mathbb{E}\Big[\|\mathbf{w}_1^{(j)}\|_2^2\Big] - \|\mathbf{z}\|_2^2\Big) = \frac{n-1}{s^{(j)}}\|\mathbf{z}\|_2^2.$$

The conditional form (5) follows by applying the same calculation after conditioning on $\mathcal{F}_{j-1}$. □

E.2 Convergence at the Improved Sample Size

Proof of Theorem 3.3. The proof conditions on the past and uses Lemma 3.2. Let

$$\mathbf{z}_0 := \mathbf{G}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\top\mathbf{b}, \qquad \mathbf{z}_j := \widehat{\mathbf{R}}^{(j)}\mathbf{z}_{j-1}, \quad j = 1, \dots, t.$$

By (4) and $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_n$,

$$\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2 = \|\mathbf{z}_t\|_2, \qquad \|\mathbf{z}_0\|_2 = \|\mathbf{x}^\star\|_2, \tag{33}$$

where the second equality follows from $\mathbf{V}\mathbf{G}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\top\mathbf{b} = \mathbf{x}^\star$.

For an arbitrary $j \in \{1, \dots, t\}$, recall from (3) that $\widehat{\mathbf{R}}^{(j)} = -(\mathbf{G} + \widehat{\mathbf{E}}^{(j)})^{-1}\widehat{\mathbf{E}}^{(j)}$. Since

$$\mathbf{G} + \widehat{\mathbf{E}}^{(j)} = \lambda\mathbf{\Sigma}^{-2} + \mathbf{V}^\top\mathbf{S}^{(j)}\mathbf{S}^{(j)\top}\mathbf{V} \succeq \lambda\mathbf{\Sigma}^{-2},$$

we have $\|(\mathbf{G} + \widehat{\mathbf{E}}^{(j)})^{-1}\|_2 \le \sigma_{\max}^2/\lambda$. Therefore, conditionally on $\mathcal{F}_{j-1}$,

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{z}_j\|_2^2 \mid \mathcal{F}_{j-1}\big] &\le \left(\frac{\sigma_{\max}^2}{\lambda}\right)^2 \mathbb{E}\Big[\|\widehat{\mathbf{E}}^{(j)}\mathbf{z}_{j-1}\|_2^2 \mid \mathcal{F}_{j-1}\Big] \\ &= \left(\frac{\sigma_{\max}^2}{\lambda}\right)^2 \frac{n-1}{s^{(j)}}\|\mathbf{z}_{j-1}\|_2^2,\end{aligned} \tag{34}$$

where the last equality uses Lemma 3.2 with $\mathbf{z} = \mathbf{z}_{j-1}$. Iterating (34) by the tower property and then using (33) gives (6). The high-probability bound (7) follows from one terminal application of Markov's inequality.

For a common sketch size satisfying (8), the product in (6) is bounded by

$$\left[\left(\frac{\sigma_{\max}^2}{\lambda}\right)^2 \frac{n-1}{s}\right]^t \le \varepsilon^{2t}.$$

This proves the stated mean square contraction. The high-probability statement follows by substituting the same sample-size bound into (7). □

E.3 Convergence on the Embedding Event

The next result retains a uniform embedding event to avoid the factor $\sigma_{\max}^2/\lambda$.

Theorem E.2 (Leverage score sampling convergence with an embedding event)**.** *Run Algorithm 1 with LS. Fix $\varepsilon \in (0,1)$ and assume $s^{(j)} \ge 2n/\varepsilon^2$ for every j. Define*

$$\delta_{\mathrm{ls}}(s) = 4(n+1)\exp\left(-\frac{3\varepsilon^2 s}{8n}\right), \qquad J_t = \bigcap_{j=1}^{t}\{\|\widehat{\mathbf{E}}^{(j)}\|_2 \le \varepsilon\}.$$

Then

$$\mathbb{E}[\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2^2 \mathbf{1}_{J_t}] \le (1-\varepsilon)^{-2t}\left(\prod_{j=1}^{t}\frac{n-1}{s^{(j)}}\right)\|\mathbf{x}^\star\|_2^2. \tag{35}$$

For any fixed t and $\delta_{\mathrm{tail}} \in (0,1)$, with probability at least $1 - \delta_{\mathrm{tail}} - \sum_{j=1}^{t}\delta_{\mathrm{ls}}(s^{(j)})$,

$$\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2 \le \frac{(1-\varepsilon)^{-t}}{\sqrt{\delta_{\mathrm{tail}}}}\prod_{j=1}^{t}\sqrt{\frac{n-1}{s^{(j)}}}\|\mathbf{x}^\star\|_2. \tag{36}$$

In particular, fix $\delta_{\mathrm{emb}} \in (0,1)$. If the common sketch size satisfies

$$s \ge \max\left\{\frac{2n}{\varepsilon^2}, \frac{8n}{3\varepsilon^2}\log\left(\frac{4t(n+1)}{\delta_{\mathrm{emb}}}\right)\right\},$$

then, with probability at least $1 - \delta_{\mathrm{tail}} - \delta_{\mathrm{emb}}$,

$$\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2 \le \frac{1}{\sqrt{\delta_{\mathrm{tail}}}}\left(\frac{\varepsilon}{1-\varepsilon}\sqrt{\frac{3(n-1)}{8n\log(4t(n+1)/\delta_{\mathrm{emb}})}}\right)^t\|\mathbf{x}^\star\|_2.$$

Proof. If $\mathbf{x}^\star = \mathbf{0}$, then $\mathbf{b} = \mathbf{0}$ and every iterate is zero, so the conclusions hold immediately. Assume otherwise in the calculation below. Let $B_j = \{\|\widehat{\mathbf{E}}^{(j)}\|_2 \le \varepsilon\}$ and $J_j = \bigcap_{\ell=1}^{j} B_\ell$, with $J_0 = \Omega$. Use the same variables as in the proof of Theorem 3.3:

$$\mathbf{z}_0 := \mathbf{G}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\top\mathbf{b}, \qquad \mathbf{z}_j := \widehat{\mathbf{R}}^{(j)}\mathbf{z}_{j-1}, \quad j = 1, \dots, t.$$

By (4), $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_n$, and (32),

$$\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2 = \|\mathbf{z}_t\|_2, \qquad \|\mathbf{z}_0\|_2 = \|\mathbf{x}^\star\|_2.$$

On B_j, we have $\mathbf{I}_n + \widehat{\mathbf{E}}^{(j)} \succeq (1-\varepsilon)\mathbf{I}_n$, and hence

$$\mathbf{G} + \widehat{\mathbf{E}}^{(j)} = \lambda\mathbf{\Sigma}^{-2} + \mathbf{I}_n + \widehat{\mathbf{E}}^{(j)} \succeq (1-\varepsilon)\mathbf{I}_n.$$

Thus $\|(\mathbf{G} + \widehat{\mathbf{E}}^{(j)})^{-1}\|_2 \le (1-\varepsilon)^{-1}$. Since $\widehat{\mathbf{R}}^{(j)} = -(\mathbf{G} + \widehat{\mathbf{E}}^{(j)})^{-1}\widehat{\mathbf{E}}^{(j)}$ and $J_j = J_{j-1} \cap B_j$,

$$\|\mathbf{z}_j\|_2^2 \mathbf{1}_{J_j} \le (1-\varepsilon)^{-2}\|\widehat{\mathbf{E}}^{(j)}\mathbf{z}_{j-1}\|_2^2 \mathbf{1}_{J_{j-1}}.$$

Condition on $\mathcal{F}_{j-1}$. The event J_{j-1} and vector $\mathbf{z}_{j-1}$ are $\mathcal{F}_{j-1}$-measurable, while $\sigma(\mathbf{S}^{(j)})$ is independent of $\mathcal{F}_{j-1}$. Lemma 3.2 with $\mathbf{z} = \mathbf{z}_{j-1}$ therefore gives

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{z}_j\|_2^2 \mathbf{1}_{J_j}\right] &\le (1-\varepsilon)^{-2}\mathbb{E}\left[\|\widehat{\mathbf{E}}^{(j)}\mathbf{z}_{j-1}\|_2^2 \mathbf{1}_{J_{j-1}}\right] \\
&= (1-\varepsilon)^{-2}\mathbb{E}\left[\mathbb{E}\left[\|\widehat{\mathbf{E}}^{(j)}\mathbf{z}_{j-1}\|_2^2 \mathbf{1}_{J_{j-1}} \mid \mathcal{F}_{j-1}\right]\right] \\
&= (1-\varepsilon)^{-2}\mathbb{E}\left[\mathbf{1}_{J_{j-1}}\mathbb{E}\left[\|\widehat{\mathbf{E}}^{(j)}\mathbf{z}_{j-1}\|_2^2 \mid \mathcal{F}_{j-1}\right]\right] \\
&= (1-\varepsilon)^{-2}\frac{n-1}{s^{(j)}}\mathbb{E}\left[\|\mathbf{z}_{j-1}\|_2^2 \mathbf{1}_{J_{j-1}}\right]. \qquad (37)
\end{aligned}$$

Iterating (37) proves (35).

$$\begin{aligned}
&\mathbb{P}\left(\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2 \ge \frac{1}{\sqrt{\delta_{\mathrm{tail}}}}(1-\varepsilon)^{-t}\prod_{j=1}^{t}\sqrt{\frac{n-1}{s^{(j)}}}\,\|\mathbf{x}^\star\|_2,\, J_t\right) \\
&= \mathbb{P}\left(\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2^2 \ge \frac{1}{\delta_{\mathrm{tail}}}(1-\varepsilon)^{-2t}\prod_{j=1}^{t}\frac{n-1}{s^{(j)}}\,\|\mathbf{x}^\star\|_2^2,\, J_t\right) \\
&= \mathbb{P}\left(\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2^2 \mathbf{1}_{J_t} \ge \frac{1}{\delta_{\mathrm{tail}}}(1-\varepsilon)^{-2t}\prod_{j=1}^{t}\frac{n-1}{s^{(j)}}\,\|\mathbf{x}^\star\|_2^2\right) \\
&\le \delta_{\mathrm{tail}},
\end{aligned}$$

where the last inequality follows from Markov's inequality and (35). Lemma D.1, applied to $\mathbf{V}$, gives $\mathbb{P}\left(B_j^c\right) \le \delta_{\mathrm{ls}}(s^{(j)})$ for every j. Hence $\mathbb{P}\left(J_t^c\right) \le \sum_{j=1}^{t}\delta_{\mathrm{ls}}(s^{(j)})$ by a union bound. Combining the two bounds proves (36).

Common sketch size. The sample-size assumption gives $\delta_{\mathrm{ls}}(s) \le \delta_{\mathrm{emb}}/t$, so the embedding failure probability in Theorem E.2 is at most δ_{emb}. It also implies

$$\sqrt{\frac{n-1}{s}} \le \varepsilon\sqrt{\frac{3(n-1)}{8n\log(4t(n+1)/\delta_{\mathrm{emb}})}}.$$

Substituting this bound into (36) proves the claim. □

E.4 Squared Contraction of Algorithmic Bias

Here bias refers to the mean solver error over the sketches, $\mathbb{E}[\widehat{\mathbf{x}}^\star - \mathbf{x}^\star]$, with $\mathbf{A}$ and $\mathbf{b}$ held fixed. It is distinct from the statistical bias of ridge regression. Related works study inversion bias associated with sketching (Dereziński et al., 2021; Niu et al., 2025).

Theorem E.3 (Squared contraction of algorithmic bias)**.** *Fix $\varepsilon \in (0,1)$ and define*

$$C = \frac{\sigma_{\max}^2}{\lambda} + \frac{\sigma_{\max}^2}{\sigma_{\min}^2}, \qquad \delta(s) = 4(n+1)\exp\left(-\frac{3\varepsilon^2 s}{32n}\right).$$

If $s^{(j)} \geq 8n/\varepsilon^2$ for every j, then

$$\|\mathbb{E}[\widehat{\mathbf{x}}^\star - \mathbf{x}^\star]\|_2 \leq \prod_{j=1}^{t}\left(\frac{\varepsilon^2}{2} + C(n-1)^2\delta(s^{(j)})\right)\|\mathbf{x}^\star\|_2. \tag{38}$$

In particular, suppose all iterations use a common sketch size satisfying

$$s \geq \max\left\{\frac{8n}{\varepsilon^2},\ \frac{32n}{3\varepsilon^2}\log\frac{8C(n+1)(n-1)^2}{\varepsilon^2}\right\}.$$

Then

$$\|\mathbb{E}[\widehat{\mathbf{x}}^\star - \mathbf{x}^\star]\|_2 \leq \varepsilon^{2t}\|\mathbf{x}^\star\|_2.$$

Proof. Step 1 (factorize the expectation). Take expectations in (4). The factors $\mathbf{V}$ and $\mathbf{G}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\top\mathbf{b}$ are deterministic, and the $\widehat{\mathbf{R}}^{(j)}$ are independent (each depends only on its own $\mathbf{S}^{(j)}$), so

$$\mathbb{E}[\widehat{\mathbf{x}}^\star - \mathbf{x}^\star] = (-1)^{t-1}\mathbf{V}\,\mathbb{E}[\widehat{\mathbf{R}}^{(t)}]\cdots\mathbb{E}[\widehat{\mathbf{R}}^{(1)}]\,\mathbf{G}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^\top\mathbf{b}.$$

Step 2 (an exact identity for $\mathbb{E}[\widehat{\mathbf{R}}^{(j)}]$). Fix j and write $\mathbf{M} := \mathbf{G}^{-1}\widehat{\mathbf{E}}^{(j)}$, so that by definition (3)

$$\widehat{\mathbf{R}}^{(j)} = (\mathbf{I}_n + \mathbf{M})^{-1} - \mathbf{I}_n.$$

We claim that

$$(\mathbf{I}_n + \mathbf{M})^{-1} - \mathbf{I}_n = -\mathbf{M} + (\mathbf{I}_n + \mathbf{M})^{-1}\mathbf{M}^2.$$

Indeed,

$$\begin{aligned}(\mathbf{I}_n + \mathbf{M})^{-1} - \mathbf{I}_n + \mathbf{M} &= (\mathbf{I}_n + \mathbf{M})^{-1}\big[\mathbf{I}_n - (\mathbf{I}_n + \mathbf{M}) + (\mathbf{I}_n + \mathbf{M})\mathbf{M}\big] \\ &= (\mathbf{I}_n + \mathbf{M})^{-1}\big[\mathbf{I}_n - \mathbf{I}_n - \mathbf{M} + \mathbf{M} + \mathbf{M}^2\big] = (\mathbf{I}_n + \mathbf{M})^{-1}\mathbf{M}^2,\end{aligned}$$

Therefore

$$\widehat{\mathbf{R}}^{(j)} = -\mathbf{M} + (\mathbf{I}_n + \mathbf{M})^{-1}\mathbf{M}^2,$$

and since $\mathbb{E}[\mathbf{M}] = \mathbf{0}$,

$$\mathbb{E}[\widehat{\mathbf{R}}^{(j)}] = -\mathbb{E}[\mathbf{M}] + \mathbb{E}\big[(\mathbf{I}_n + \mathbf{M})^{-1}\mathbf{M}^2\big] = \mathbb{E}\big[(\mathbf{I}_n + \mathbf{M})^{-1}\mathbf{M}^2\big]. \tag{39}$$

This is exact: it uses only $\mathbb{E}[\mathbf{M}] = \mathbf{0}$ and the invertibility of $\mathbf{I}_n + \mathbf{M}$ (guaranteed by $\lambda > 0$), with no smallness assumption on $\mathbf{M}$. The linear term has cancelled, leaving a quantity *quadratic* in $\mathbf{M}$.

Step 3 (bound the quadratic). A deterministic bound on $\|(\mathbf{I}_n + \mathbf{M})^{-1}\|_2$ holds always: $\mathbf{I}_n + \mathbf{M} = \mathbf{G}^{-1}(\mathbf{G} + \widehat{\mathbf{E}}^{(j)})$ with $\mathbf{G} + \widehat{\mathbf{E}}^{(j)} = \lambda\mathbf{\Sigma}^{-2} + \mathbf{V}^\top\mathbf{S}^{(j)}\mathbf{S}^{(j)\top}\mathbf{V} \succeq \lambda\mathbf{\Sigma}^{-2} \succ 0$, so

$$\|(\mathbf{I}_n + \mathbf{M})^{-1}\|_2 = \|(\mathbf{G} + \widehat{\mathbf{E}}^{(j)})^{-1}\mathbf{G}\|_2 \leq \|(\mathbf{G} + \widehat{\mathbf{E}}^{(j)})^{-1}\|_2\,\|\mathbf{G}\|_2 \leq \frac{\sigma_{\max}^2}{\lambda}\left(1 + \frac{\lambda}{\sigma_{\min}^2}\right) = C. \tag{40}$$

Split on the event $B_j = \{\|\widehat{\mathbf{E}}^{(j)}\|_2 \leq \varepsilon/2\}$, which by Lemma D.1 satisfies $\mathbb{P}\left(B_j^c\right) \leq \delta(s^{(j)})$. On B_j, $\|\mathbf{M}\|_2 \leq \|\widehat{\mathbf{E}}^{(j)}\|_2 \leq \varepsilon/2$ because $\|\mathbf{G}^{-1}\|_2 \leq 1$. From (39),

$$\begin{aligned}\|\mathbb{E}[\widehat{\mathbf{R}}^{(j)}]\|_2 &\leq \mathbb{E}\big[\|(\mathbf{I}_n + \mathbf{M})^{-1}\mathbf{M}^2\|_2\big] \\ &\leq \mathbb{E}\big[\|(\mathbf{I}_n + \mathbf{M})^{-1}\|_2\,\|\mathbf{M}\|_2^2\big] \\ &= \mathbb{E}\big[\|(\mathbf{I}_n + \mathbf{M})^{-1}\|_2\|\mathbf{M}\|_2^2\,\mathbf{1}_{B_j}\big] + \mathbb{E}\big[\|(\mathbf{I}_n + \mathbf{M})^{-1}\|_2\|\mathbf{M}\|_2^2\,\mathbf{1}_{B_j^c}\big],\end{aligned}$$

where in the first inequality we used Jensen's inequality and for the second we used submultiplicativity of the spectral norm. On B_j we have $\|(\mathbf{I}_n+\mathbf{M})^{-1}\|_2 \le 1/(1-\varepsilon/2)$ (since $\|\mathbf{M}\|_2 \le \varepsilon/2 < 1$, so $\|(\mathbf{I}_n+\mathbf{M})^{-1}\|_2 = \|\sum_{\ell=0}^{\infty}(-1)^\ell \mathbf{M}^\ell\|_2 \le \sum_{\ell=0}^{\infty}\|\mathbf{M}\|_2^\ell \le 1/(1-\|\mathbf{M}\|_2)$) and $\|\mathbf{M}\|_2^2 \le (\varepsilon/2)^2$, so the first term is at most $\frac{(\varepsilon/2)^2}{1-\varepsilon/2} \le \varepsilon^2/2$ for $\varepsilon \le 1$. On B_j^c we use the deterministic bound (40), giving the second term $\le C\,\mathbb{E}[\|\mathbf{M}\|_2^2 \mathbf{1}_{B_j^c}]$.

It remains to bound $\mathbb{E}[\|\mathbf{M}\|_2^2 \mathbf{1}_{B_j^c}]$. Leverage score sampling yields

$$\mathbf{0} \preceq \mathbf{V}^\top \mathbf{S}^{(j)} \mathbf{S}^{(j)\top} \mathbf{V} \preceq n\mathbf{I}_n,$$

since $\mathbf{V}^\top \mathbf{S}^{(j)} \mathbf{S}^{(j)\top} \mathbf{V} = \frac{1}{s^{(j)}} \sum_{k=1}^{s^{(j)}} p_{i_k}^{-1} (\mathbf{V}_{i_k *})^\top (\mathbf{V}_{i_k *})$ and each rank-one term obeys $p_{i_k}^{-1}(\mathbf{V}_{i_k *})^\top(\mathbf{V}_{i_k *}) \preceq (\|\mathbf{V}_{i_k *}\|_2^2/p_{i_k})\mathbf{I}_n = n\mathbf{I}_n$, using $p_i = \|\mathbf{V}_{i*}\|_2^2/n$. Hence $-\mathbf{I}_n \preceq \widehat{\mathbf{E}}^{(j)} \preceq (n-1)\mathbf{I}_n$, and since $\|\mathbf{G}^{-1}\|_2 \le 1$,

$$\|\mathbf{M}\|_2 \le \|\mathbf{G}^{-1}\|_2 \, \|\widehat{\mathbf{E}}^{(j)}\|_2 \le n-1.$$

With $\mathbb{P}\left(B_j^c\right) \le \delta(s^{(j)})$ this gives the deterministic control of the remainder,

$$\mathbb{E}\big[\|\mathbf{M}\|_2^2 \mathbf{1}_{B_j^c}\big] \le (n-1)^2\, \mathbb{P}\left(B_j^c\right) \le (n-1)^2\, \delta(s^{(j)}).$$

Altogether,

$$\|\mathbb{E}[\widehat{\mathbf{R}}^{(j)}]\|_2 \le \tfrac{\varepsilon^2}{2} + C(n-1)^2\, \delta(s^{(j)}).$$

Step 4 (conclude). By Step 1, submultiplicativity of the spectral norm, $\|\mathbf{V}\|_2 = 1$, and (32),

$$\|\mathbb{E}[\widehat{\mathbf{x}}^\star - \mathbf{x}^\star]\|_2 \le \prod_{j=1}^{t} \|\mathbb{E}[\widehat{\mathbf{R}}^{(j)}]\|_2 \, \|\mathbf{x}^\star\|_2 \le \prod_{j=1}^{t} \left(\tfrac{\varepsilon^2}{2} + C(n-1)^2\, \delta(s^{(j)})\right) \|\mathbf{x}^\star\|_2.$$

Common sketch size. The stated s gives $\delta(s) = 4(n+1)\exp\big(-\frac{3\varepsilon^2 s}{32n}\big) \le \varepsilon^2/(2C(n-1)^2)$, so $C(n-1)^2\,\delta(s) \le \varepsilon^2/2$ and every factor in (38) is at most ε^2. □

F Ridge Leverage Analysis

Throughout this appendix, $\mathbf{X} = \mathbf{V}\mathbf{\Sigma}_\lambda$ and $w_i = \|\mathbf{X}_{i*}\|_2^2$, so $\sum_i w_i = d_\lambda > 0$. We write $c_\lambda = (\sigma_{\max}^2 + \lambda)/\lambda$ as in (9).

F.1 Error identity and conditional variance

We also use the inverse error matrix

$$\mathbf{R}^{(j)} = (\mathbf{I}_n + \mathbf{E}^{(j)})^{-1} - \mathbf{I}_n = -(\mathbf{I}_n + \mathbf{E}^{(j)})^{-1}\mathbf{E}^{(j)}.$$

Proof of Lemma 4.1. From (9),

$$\mathbf{I}_n + \mathbf{E}^{(j)} = (\mathbf{I}_n - \mathbf{\Sigma}_\lambda^2) + \mathbf{X}^\top \mathbf{S}^{(j)} \mathbf{S}^{(j)\top} \mathbf{X}.$$

The second matrix is a Gram matrix, hence symmetric positive semidefinite. The first is diagonal with entries $1 - \sigma_i^2/(\sigma_i^2+\lambda) = \lambda/(\sigma_i^2+\lambda) > 0$, so $\mathbf{I}_n - \mathbf{\Sigma}_\lambda^2 \succ 0$ with smallest eigenvalue $\min_i \lambda/(\sigma_i^2+\lambda) = \lambda/(\sigma_{\max}^2+\lambda)$. Adding a PSD matrix only raises eigenvalues, so $\mathbf{I}_n + \mathbf{E}^{(j)}$ is symmetric positive definite with $\lambda_{\min}(\mathbf{I}_n + \mathbf{E}^{(j)}) \ge \lambda/(\sigma_{\max}^2+\lambda)$. For a symmetric positive-definite matrix $\mathbf{A}$, $\|\mathbf{A}^{-1}\|_2 = 1/\lambda_{\min}(\mathbf{A})$, gives

$$\|(\mathbf{I}_n + \mathbf{E}^{(j)})^{-1}\|_2 \le \frac{\sigma_{\max}^2 + \lambda}{\lambda} = c_\lambda. \tag{41}$$

By Chowdhury et al. (2018, Lemma 5, eqn. (51))[1]

$$\widetilde{\mathbf{x}}^{(j)} = \mathbf{x}^{\star(j)} + \mathbf{V}\mathbf{\Sigma}_\lambda \, \mathbf{R}^{(j)} \, \mathbf{\Sigma}_\lambda \mathbf{\Sigma}^{-1} \mathbf{U}^\top \mathbf{b}^{(j)},$$

[1] Even though the proof steps in Chowdhury et al. (2018) use the series form, the same result holds with our definition of $\mathbf{R}^{(j)}$.

so with $\mathbf{R}^{(j)} = -(\mathbf{I}_n + \mathbf{E}^{(j)})^{-1}\mathbf{E}^{(j)}$,

$$\mathbf{x}^{\star(j)} - \widetilde{\mathbf{x}}^{(j)} = \mathbf{V}\mathbf{\Sigma}_\lambda(\mathbf{I}_n + \mathbf{E}^{(j)})^{-1}\,\mathbf{E}^{(j)}\mathbf{\Sigma}_\lambda\mathbf{\Sigma}^{-1}\mathbf{U}^\top\mathbf{b}^{(j)}. \tag{42}$$

We simplify the tail. Using $\mathbf{E}^{(j)} = \mathbf{\Sigma}_\lambda\mathbf{V}^\top\mathbf{S}^{(j)}\mathbf{S}^{(j)\top}\mathbf{V}\mathbf{\Sigma}_\lambda - \mathbf{\Sigma}_\lambda^2$, $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_n$, $\mathbf{X} = \mathbf{V}\mathbf{\Sigma}_\lambda$, and Lemma D.2 ($\mathbf{x}^{\star(j)} = \mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{\Sigma}^{-1}\mathbf{U}^\top\mathbf{b}^{(j)}$),

$$\begin{aligned}\mathbf{E}^{(j)}\mathbf{\Sigma}_\lambda\mathbf{\Sigma}^{-1}\mathbf{U}^\top\mathbf{b}^{(j)} &= \mathbf{\Sigma}_\lambda\mathbf{V}^\top\mathbf{S}^{(j)}\mathbf{S}^{(j)\top}\mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{\Sigma}^{-1}\mathbf{U}^\top\mathbf{b}^{(j)} - \mathbf{\Sigma}_\lambda\mathbf{V}^\top\mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{\Sigma}^{-1}\mathbf{U}^\top\mathbf{b}^{(j)} \\ &= \mathbf{\Sigma}_\lambda\mathbf{V}^\top\mathbf{S}^{(j)}\mathbf{S}^{(j)\top}\,\mathbf{x}^{\star(j)} - \mathbf{\Sigma}_\lambda\mathbf{V}^\top\,\mathbf{x}^{\star(j)} = \mathbf{X}^\top\big(\mathbf{S}^{(j)}\mathbf{S}^{(j)\top} - \mathbf{I}_d\big)\mathbf{x}^{\star(j)},\end{aligned} \tag{43}$$

where the last line uses $\mathbf{\Sigma}_\lambda\mathbf{V}^\top = \mathbf{X}^\top$ and $\mathbf{V}\mathbf{\Sigma}_\lambda^2\mathbf{\Sigma}^{-1}\mathbf{U}^\top\mathbf{b}^{(j)} = \mathbf{x}^{\star(j)}$. Substituting (43) into (42) gives the stated identity. Finally $\|\mathbf{V}\mathbf{\Sigma}_\lambda\|_2 \le \|\mathbf{V}\|_2\|\mathbf{\Sigma}_\lambda\|_2 \le 1$ and (41) give $\|\mathbf{x}^{\star(j)} - \widetilde{\mathbf{x}}^{(j)}\|_2 \le c_\lambda\|\mathbf{X}^\top(\mathbf{S}^{(j)}\mathbf{S}^{(j)\top} - \mathbf{I}_d)\mathbf{x}^{\star(j)}\|_2$. If $\|\mathbf{E}^{(j)}\|_2 \le \varepsilon < 1$, then the Neumann series gives $\|(\mathbf{I}_n + \mathbf{E}^{(j)})^{-1}\|_2 \le 1/(1-\varepsilon)$, proving (10). □

Lemma F.1 (Variance of a sketched matrix product)**.** *For fixed $\mathbf{x} \in \mathbb{R}^d$ and draw $\mathbf{S}$ by Algorithm 2 with size s and probabilities p. Then*

$$\mathbb{E}\|\mathbf{X}^\top(\mathbf{S}\mathbf{S}^\top - \mathbf{I}_d)\mathbf{x}\|_2^2 \le \frac{1}{s}\sum_i \frac{|\mathbf{x}_i|^2 w_i}{p_i}.$$

If $\mathbf{x}$ and p are measurable with respect to a sigma algebra $\mathcal{G}$, the same calculation gives

$$\mathbb{E}\big[\|\mathbf{X}^\top(\mathbf{S}\mathbf{S}^\top - \mathbf{I}_d)\mathbf{x}\|_2^2 \mid \mathcal{G}\big] \le \frac{1}{s}\sum_i \frac{|\mathbf{x}_i|^2 w_i}{p_i}. \tag{44}$$

Proof. Use Theorem 22 from Drineas & Mahoney (2017). The conditional form (44) follows by applying the same calculation after conditioning on $\mathcal{G}$. □

Proof of Lemma 4.2. Condition on $\mathcal{F}_{j-1}$, so that $\mathbf{x}$ is fixed. Apply the exact variance identity in Lemma F.1 with $p_i = w_i/d_\lambda$ to obtain

$$\mathbb{E}\Big[\|\mathbf{X}^\top(\mathbf{S}^{(j)}\mathbf{S}^{(j)\top} - \mathbf{I}_d)\mathbf{x}\|_2^2 \mid \mathcal{F}_{j-1}\Big] \le \frac{1}{s^{(j)}}\left(\sum_i \frac{|\mathbf{x}_i|^2 w_i}{w_i/d_\lambda}\right) = \frac{d_\lambda}{s^{(j)}}\|\mathbf{x}\|_2^2.$$

□

F.2 A CONVERGENCE THEOREM FOR RESIDUAL-AWARE DISTRIBUTIONS

Theorem F.2 (Convergence from conditional variance)**.** *For the iteration in (2), let $(\mathcal{H}_j)_{j=0}^t$ be a filtration such that $\mathcal{H}_{j-1}$ contains all randomness available before drawing $\mathbf{S}^{(j)}$, for $j = 1, \dots, t$. Suppose $p^{(j)}$ and $\mathbf{x}^{\star(j)}$ are $\mathcal{H}_{j-1}$-measurable and, conditionally on $\mathcal{H}_{j-1}$, the sketch contains $s^{(j)}$ independent draws from $p^{(j)}$. For nonzero residual solutions, define*

$$R_p^{(j)} = \frac{\sum_i |\mathbf{x}_i^{\star(j)}|^2 w_i/p_i^{(j)}}{d_\lambda\|\mathbf{x}^{\star(j)}\|_2^2}, \qquad P_t = \prod_{j=1}^t R_p^{(j)}. \tag{45}$$

Set $R_p^{(j)} = 1$ when $\mathbf{x}^{\star(j)} = \mathbf{0}$. Then $P_t > 0$ and

$$\mathbb{E}\left[\frac{\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2^2}{P_t}\right] \le c_\lambda^{2t}\prod_{j=1}^t \frac{d_\lambda}{s^{(j)}}\|\mathbf{x}^\star\|_2^2. \tag{46}$$

For every $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2 \le \delta^{-1/2}c_\lambda^t\prod_{j=1}^t \sqrt{\frac{R_p^{(j)}d_\lambda}{s^{(j)}}}\|\mathbf{x}^\star\|_2. \tag{47}$$

Proof. Set $\mathbf{e}_j := \mathbf{x}^{\star(j)} - \widetilde{\mathbf{x}}^{(j)}$ and $\mathbf{e}_0 := \mathbf{x}^\star$, and define $P_j := \prod_{\ell=1}^{j} R_p^{(\ell)}$ with $P_0 := 1$.

Let

$$\mathbf{y}_j := \mathbf{X}^\top (\mathbf{S}^{(j)} \mathbf{S}^{(j)\top} - \mathbf{I}_d) \mathbf{x}^{\star(j)}.$$

Lemma 4.1 gives $\|\mathbf{e}_j\|_2 \le c_\lambda \|\mathbf{y}_j\|_2$ for every sketch realization. Since $p^{(j)}$ and $\mathbf{x}^{\star(j)}$ are $\mathcal{H}_{j-1}$-measurable, (44) and the definition of $R_p^{(j)}$ give

$$\mathbb{E}\big[\|\mathbf{e}_j\|_2^2 \mid \mathcal{H}_{j-1}\big] \le c_\lambda^2 \frac{R_p^{(j)} d_\lambda}{s^{(j)}} \|\mathbf{x}^{\star(j)}\|_2^2. \tag{48}$$

By Lemma D.2, $\|\mathbf{x}^{\star(j)}\|_2 = \|\mathbf{e}_{j-1}\|_2$. Since $P_j = P_{j-1} R_p^{(j)}$ is $\mathcal{H}_{j-1}$-measurable,

$$\mathbb{E}\left[\frac{\|\mathbf{e}_j\|_2^2}{P_j} \,\middle|\, \mathcal{H}_{j-1} \right] \le c_\lambda^2 \frac{d_\lambda}{s^{(j)}} \frac{\|\mathbf{e}_{j-1}\|_2^2}{P_{j-1}}. \tag{49}$$

Iterating (49) with the tower property and using Lemma D.2 ($\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2 = \|\mathbf{e}_t\|_2$) gives (46). Markov's inequality applied to $\|\mathbf{e}_t\|_2^2 / P_t$ gives (47). □

Proof of Theorem 4.3. Apply Theorem F.2 with $\mathcal{H}_j = \mathcal{F}_j$ and $p_i^{(j)} = w_i / d_\lambda$. The residual solution $\mathbf{x}^{\star(j)}$ is $\mathcal{F}_{j-1}$-measurable, and each fresh sketch consists of independent draws from this fixed distribution. For every nonzero residual solution,

$$R_p^{(j)} = \frac{\sum_i |\mathbf{x}_i^{\star(j)}|^2 w_i / (w_i / d_\lambda)}{d_\lambda \|\mathbf{x}^{\star(j)}\|_2^2} = 1.$$

For zero residual solutions, $R_p^{(j)} = 1$ by definition. Thus $P_t = 1$, and (46) and (47) give (12) and (13), respectively.

For a common sketch size satisfying (14), $c_\lambda^2 d_\lambda / s \le \varepsilon^2$; substituting this into the two bounds gives the final claims. □

F.3 Convergence on the embedding events

Theorem F.3 (Ridge leverage score sampling convergence with an embedding event)**.** *Run Algorithm 1 with RLS. Fix $\varepsilon \in (0,1)$ and assume $s^{(j)} \ge 2d_\lambda/\varepsilon^2$ for every j. Define*

$$\delta_{\mathrm{rls}}(s) = 4(1+d_\lambda) \exp\left(-\frac{3\varepsilon^2 s}{8 d_\lambda} \right), \qquad H_j = \bigcap_{\ell=1}^{j} \{\|\mathbf{E}^{(\ell)}\|_2 \le \varepsilon\}.$$

Then

$$\mathbb{E}[\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2^2 \mathbf{1}_{H_t}] \le (1-\varepsilon)^{-2t} \prod_{j=1}^{t} \frac{d_\lambda}{s^{(j)}} \|\mathbf{x}^\star\|_2^2. \tag{50}$$

For $\delta_{\mathrm{tail}} \in (0,1)$, with probability at least $1 - \delta_{\mathrm{tail}} - \sum_{j=1}^{t} \delta_{\mathrm{rls}}(s^{(j)})$,

$$\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2 \le \delta_{\mathrm{tail}}^{-1/2} (1-\varepsilon)^{-t} \prod_{j=1}^{t} \sqrt{\frac{d_\lambda}{s^{(j)}}} \|\mathbf{x}^\star\|_2. \tag{51}$$

In particular, fix $\delta_{\mathrm{emb}} \in (0,1)$. If every iteration uses a common size

$$s \ge \max\left\{ \frac{2d_\lambda}{\varepsilon^2}, \frac{8d_\lambda}{3\varepsilon^2} \log\left(\frac{4t(1+d_\lambda)}{\delta_{\mathrm{emb}}} \right) \right\},$$

then, with probability at least $1 - \delta_{\mathrm{tail}} - \delta_{\mathrm{emb}}$,

$$\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2 \le \delta_{\mathrm{tail}}^{-1/2} \left(\frac{\varepsilon}{1-\varepsilon} \sqrt{\frac{3}{8 \log(4t(1+d_\lambda)/\delta_{\mathrm{emb}})}} \right)^t \|\mathbf{x}^\star\|_2.$$

Proof. Let $G_j = \{\|\mathbf{E}^{(j)}\|_2 \le \varepsilon\}$, $H_j = \bigcap_{\ell=1}^{j} G_\ell$, and $H_0 = \Omega$. Set $\mathbf{e}_j := \mathbf{x}^{\star(j)} - \widetilde{\mathbf{x}}^{(j)}$ and $\mathbf{e}_0 := \mathbf{x}^\star$. Also set

$$\mathbf{y}_j := \mathbf{X}^\top(\mathbf{S}^{(j)}\mathbf{S}^{(j)\top} - \mathbf{I}_d)\mathbf{x}^{\star(j)}.$$

On G_j, (10) gives $\|\mathbf{e}_j\|_2 \le (1-\varepsilon)^{-1}\|\mathbf{y}_j\|_2$. Since $H_j = H_{j-1} \cap G_j$,

$$\|\mathbf{e}_j\|_2^2 \mathbf{1}_{H_j} \le (1-\varepsilon)^{-2}\|\mathbf{y}_j\|_2^2 \mathbf{1}_{H_{j-1}}\mathbf{1}_{G_j} \le (1-\varepsilon)^{-2}\|\mathbf{y}_j\|_2^2 \mathbf{1}_{H_{j-1}}.$$

The last inequality is the key point: we use G_j to bound the inverse, then drop $\mathbf{1}_{G_j}$ before conditioning, so the conditional sampling law is not changed.

Now take expectations and condition on $\mathcal{F}_{j-1}$. The event H_{j-1} and the vector $\mathbf{x}^{\star(j)}$ are $\mathcal{F}_{j-1}$-measurable, while $\mathbf{S}^{(j)}$ is fresh. Substituting $p_i = w_i/d_\lambda$ into (44) gives $\sum_i \left|\mathbf{x}_i^{\star(j)}\right|^2 w_i/p_i = d_\lambda\|\mathbf{x}^{\star(j)}\|_2^2$, and hence

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{e}_j\|_2^2\mathbf{1}_{H_j}\big] &\le (1-\varepsilon)^{-2}\mathbb{E}\big[\mathbf{1}_{H_{j-1}}\,\mathbb{E}\big[\|\mathbf{y}_j\|_2^2 \mid \mathcal{F}_{j-1}\big]\big] \\ &\le (1-\varepsilon)^{-2}\frac{d_\lambda}{s^{(j)}}\mathbb{E}\Big[\|\mathbf{x}^{\star(j)}\|_2^2\mathbf{1}_{H_{j-1}}\Big].\end{aligned}$$

By Lemma D.2, $\|\mathbf{x}^{\star(j)}\|_2 = \|\mathbf{e}_{j-1}\|_2$ for each j, with $\mathbf{e}_0 = \mathbf{x}^\star$. Therefore

$$\mathbb{E}\big[\|\mathbf{e}_j\|_2^2\mathbf{1}_{H_j}\big] \le (1-\varepsilon)^{-2}\frac{d_\lambda}{s^{(j)}}\mathbb{E}\big[\|\mathbf{e}_{j-1}\|_2^2\mathbf{1}_{H_{j-1}}\big]. \tag{52}$$

Iterating (52) and using Lemma D.2 ($\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2 = \|\mathbf{e}_t\|_2$) gives (50).

For the high-probability statement, Markov's inequality applied to $\|\mathbf{e}_t\|_2^2\mathbf{1}_{H_t}$ gives

$$\mathbb{P}\left(\|\mathbf{e}_t\|_2 > \frac{1}{\sqrt{\delta_{\mathrm{tail}}}}(1-\varepsilon)^{-t}\prod_{j=1}^{t}\sqrt{\frac{d_\lambda}{s^{(j)}}}\|\mathbf{x}^\star\|_2,\ H_t\right) \le \delta_{\mathrm{tail}}.$$

Lemma D.1, applied to $\mathbf{X}$, gives $\mathbb{P}\left(G_j^c\right) \le \delta_{\mathrm{rls}}(s^{(j)})$ for every j. Hence, by a union bound, $\mathbb{P}\left(H_t^c\right) \le \sum_{j=1}^{t}\delta_{\mathrm{rls}}(s^{(j)})$. Combining the two bounds proves (51).

Common sketch size. The sample size assumption gives $\delta_{\mathrm{rls}}(s) \le \delta_{\mathrm{emb}}/t$, so the embedding failure probability in Theorem F.3 is at most δ_{emb}. It also implies

$$\sqrt{\frac{d_\lambda}{s}} \le \varepsilon\sqrt{\frac{3}{8\log(4t(1+d_\lambda)/\delta_{\mathrm{emb}})}}.$$

Substituting this bound into (51) proves the claim. □

G Oracle and Adaptive Analysis

Throughout this appendix, $c_\lambda = (\sigma_{\max}^2 + \lambda)/\lambda$ as in (9).

G.1 Oracle convergence

Corollary G.1 (Oracle convergence)**.** *Run Algorithm 1 with Oracle. Set $\rho^{(j)} = \rho(\mathbf{x}^{\star(j)})$ and $\Theta_t = \prod_{j=1}^{t}\rho^{(j)}$. Then*

$$\mathbb{E}\left[\frac{\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2^2}{\Theta_t}\right] \le c_\lambda^{2t}\prod_{j=1}^{t}\frac{d_\lambda}{s^{(j)}}\|\mathbf{x}^\star\|_2^2. \tag{53}$$

For every $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2 \le \delta^{-1/2}c_\lambda^t\prod_{j=1}^{t}\sqrt{\frac{\rho^{(j)}d_\lambda}{s^{(j)}}}\|\mathbf{x}^\star\|_2. \tag{54}$$

For a common size s, this is (16)*. If a deterministic $\bar\rho \in (0,1]$ bounds $\rho^{(j)}$ at every iteration with nonzero residual, and $s \ge c_\lambda^2\bar\rho d_\lambda/\varepsilon^2$ for $\varepsilon \in (0,1)$, then*

$$\mathbb{E}\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2^2 \le \varepsilon^{2t}\|\mathbf{x}^\star\|_2^2, \qquad \|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2 \le \delta^{-1/2}\varepsilon^t\|\mathbf{x}^\star\|_2 \tag{55}$$

where the second inequality holds with probability at least $1-\delta$.

Proof. Apply Theorem F.2 with $\mathcal{H}_j = \mathcal{F}_j$ and $p^{(j)} = p^\star(\mathbf{x}^{\star(j)})$. By Proposition 5.1 and (15), the master factor (45) is $R_p^{(j)} = \rho^{(j)}$. Hence $P_t = \Theta_t$, and (46)–(47) give (53)–(54).

For common s, the product in (54) satisfies

$$\prod_{j=1}^{t} \sqrt{\frac{\rho^{(j)} d_\lambda}{s}} = \left(\sqrt{\frac{\bar{\rho}_t d_\lambda}{s}} \right)^t .$$

This proves (16). If $\rho^{(j)} \le \bar{\rho}$ for all j, then $\bar{\rho}_t \le \bar{\rho}$; substituting $s \ge c_\lambda^2 \bar{\rho} d_\lambda / \varepsilon^2$ gives (55).

□

G.2 Adaptive proposals and a defensive mixture

The pilot sketch $\mathbf{S}^{(0)}$ is part of the history:

$$\mathcal{F}_0^{\mathrm{ad}} = \sigma(\mathbf{S}^{(0)}), \qquad \mathcal{F}_j^{\mathrm{ad}} = \sigma(\mathbf{S}^{(0)}, \mathbf{S}^{(1)}, \dots, \mathbf{S}^{(j)}).$$

The proposal in (17) is therefore fixed conditionally on $\mathcal{F}_{j-1}^{\mathrm{ad}}$.

Corollary G.2 (Pure adaptive proposals)**.** *Run Algorithm 3 with the pure proposal $\widehat{p}^{(j)}$ and sketch sizes $s^{(j)}$. At a nonzero residual solution, define*

$$R_{\mathrm{pa}}^{(j)} = \frac{\sum_i |\mathbf{x}_i^{\star(j)}|^2 w_i / \widehat{p}_i^{(j)}}{d_\lambda \|\mathbf{x}^{\star(j)}\|_2^2}, \qquad P_{\mathrm{pa},t} = \prod_{j=1}^{t} R_{\mathrm{pa}}^{(j)}. \tag{56}$$

Then

$$\mathbb{E}\left[\frac{\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2^2}{P_{\mathrm{pa},t}} \right] \le c_\lambda^{2t} \prod_{j=1}^{t} \frac{d_\lambda}{s^{(j)}} \|\mathbf{x}^\star\|_2^2.$$

For every $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2 \le \delta^{-1/2} c_\lambda^t \prod_{j=1}^{t} \sqrt{\frac{R_{\mathrm{pa}}^{(j)} d_\lambda}{s^{(j)}}} \|\mathbf{x}^\star\|_2.$$

Proof. Apply Theorem F.2 with $\mathcal{H}_j = \mathcal{F}_j^{\mathrm{ad}}$ and $p^{(j)} = \widehat{p}^{(j)}$. Then $R_p^{(j)} = R_{\mathrm{pa}}^{(j)}$ by (56). □

The corollary does not bound proposal quality: small probabilities can make $R_{\mathrm{pa}}^{(j)}$ arbitrarily large. The mixture in (19) bounds this factor uniformly.

Corollary G.3 (Defensive mixture)**.** *Use* (19) *with any measurable probability proposal $\widehat{p}^{(j)}$ and a fixed $\alpha \in (0,1)$. At a nonzero residual solution, its factor $R_{\mathrm{mix}}^{(j)}$ in* (45) *satisfies*

$$R_{\mathrm{mix}}^{(j)} \le \frac{1}{\alpha}, \qquad R_{\mathrm{mix}}^{(j)} \le \frac{R_{\mathrm{pa}}^{(j)}}{1-\alpha}.$$

Consequently,

$$\mathbb{E}\|\widehat{\mathbf{x}}^\star - \mathbf{x}^\star\|_2^2 \le c_\lambda^{2t} \prod_{j=1}^{t} \frac{d_\lambda}{\alpha s^{(j)}} \|\mathbf{x}^\star\|_2^2. \tag{57}$$

For $\varepsilon, \delta \in (0,1)$ and a common size $s \ge c_\lambda^2 d_\lambda / (\alpha \varepsilon^2)$, the right side is at most $\varepsilon^{2t} \|\mathbf{x}^\star\|_2^2$; with probability at least $1-\delta$, the relative error is at most $\delta^{-1/2}\varepsilon^t$.

Proof. Since $\widetilde{p}_i^{(j)} \ge \alpha w_i / d_\lambda$,

$$\sum_i \frac{|\mathbf{x}_i^{\star(j)}|^2 w_i}{\widetilde{p}_i^{(j)}} \le \frac{d_\lambda}{\alpha} \|\mathbf{x}^{\star(j)}\|_2^2.$$

The second comparison follows from $\widetilde{p}_i^{(j)} \ge (1-\alpha)\widehat{p}_i^{(j)}$. The first comparison inserted into (48) and iterated proves (57); it also holds after termination. The remaining statements follow by substitution and Markov's inequality. □

H Details of experiments and additional results

H.1 Implementation, metrics, and sampling budgets

The experiments use the residual correction in (2) for uniform, leverage, ridge leverage, oracle, adaptive, and mixture sampling. Sampling is with replacement, with the rescaling defined in Section 2. The mixture places weight $\alpha = 0.1$ on ridge leverage probabilities. Each trial runs for a fixed number t of updates; the zero initial iterate is included in error curves.

A thin SVD supplies the reference solution and the leverage scores, and is reused when varying λ. Each sampled system is solved by a Cholesky factorization. For a reused sketch, this factorization is cached across iterations. Both adaptive and mixture methods draw one initial ridge leverage pilot, cache its Cholesky factor, and evaluate subsequent pilot estimates by triangular solves followed by a matrix vector product. Explicit matrix inverses are not formed.

Let $\widehat{\mathbf{x}}_j^\star = \sum_{\ell=1}^{j} \widetilde{\mathbf{x}}^{(\ell)}$ denote the accumulated solution after j updates, with $\widehat{\mathbf{x}}_0^\star = \mathbf{0}_d$ and $\widehat{\mathbf{x}}_t^\star = \widehat{\mathbf{x}}^\star$. We report relative solution error and normalized objective excess,

$$\frac{\|\widehat{\mathbf{x}}_j^\star - \mathbf{x}^\star\|_2}{\|\mathbf{x}^\star\|_2}, \qquad \frac{f(\widehat{\mathbf{x}}_j^\star) - f(\mathbf{x}^\star)}{f(\mathbf{x}^\star)}.$$

The objective f is defined in (1). To avoid cancellation near convergence, for $\mathbf{e} = \widehat{\mathbf{x}}_j^\star - \mathbf{x}^\star$ the implementation evaluates its numerator as $\|\mathbf{A}\mathbf{e}\|_2^2 + \lambda\|\mathbf{e}\|_2^2$. This equals the objective excess at the exact minimizer.

The plotted parameter s is the number of columns sampled in each iteration. For adaptive and mixture methods, the initial pilot has $s_0 = s$ columns and each update has

$$s^{(j)} = \left\lfloor \frac{st - s_0}{t} \right\rfloor = \left\lfloor \frac{s(t-1)}{t} \right\rfloor, \qquad j = 1, \dots, t.$$

Thus the pilot and updates together use at most st sampled columns.

Curves show medians and interquartile ranges over independent sketch trials. Iteration curves use thin lines without markers; parameter sweeps use small filled markers at the evaluated settings. The same six-panel layout is used for all three datasets.

H.2 ARCENE and synthetic data

These experiments directly compare with the column sampling framework of Chowdhury et al. (2018). We reproduce their types of iteration, sketch size, and regularization sweeps, adding the oracle and adaptive probabilities. Panels (a–d) use fresh sketches for all six distributions; panel (e) shows the oracle factor along each of the six methods' residual trajectories, and panel (f) shows the proposal factor for adaptive sampling and its mixture. These trajectories are the same as in panels (a,b). A separate experiment compares the original reused-sketch baselines with their refreshed versions.

ARCENE. The ARCENE feature selection data (Guyon et al., 2004) combine the provided training and validation arrays into one 200×10000 design, with responses in $\{-1, 1\}$. We apply a single global min–max scaling to this combined design. This experiment studies numerical solution of that fixed ridge problem; it does not use the combined data to report held-out classification performance. Identically zero columns retain zero leverage mass and zero solution coefficients; uniform sampling still uses all 10000 columns.

Synthetic construction. Following the noisy low rank construction of Chen et al. (2015), the implementation uses $n = 500$, $d = 50000$, and rank parameter $r = 50$. Draw independent standard Gaussian matrices $\mathbf{M} \in \mathbb{R}^{n\times r}$, $\mathbf{G} \in \mathbb{R}^{d\times r}$, $\mathbf{E} \in \mathbb{R}^{n\times d}$, and independent standard Gaussian vectors $\mathbf{x}_{\mathrm{gen}} \in \mathbb{R}^d$ and $\boldsymbol{\xi} \in \mathbb{R}^n$. Let $\mathbf{Q} \in \mathbb{R}^{d\times r}$ be the orthonormal factor from a thin QR factorization of $\mathbf{G}$, and let $\mathbf{D} \in \mathbb{R}^{r\times r}$ be diagonal with $\mathbf{D}_{kk} = 1 - (k-1)/d$ for $k = 1, \dots, r$. The fixed design and response are

$$\mathbf{A} = \mathbf{M}\mathbf{D}\mathbf{Q}^\top + 0.05\mathbf{E}, \qquad \mathbf{b} = \mathbf{A}\mathbf{x}_{\mathrm{gen}} + 5\boldsymbol{\xi}.$$

A NumPy random generator with seed 42 generates these arrays separately from the sketching streams. The same design and response are used for all methods and all regularization values.

Table 1: ARCENE and synthetic configurations. Iteration curves contain the initial iterate and twenty updates. The parameter s follows the budget convention in Appendix H.1.

Setting	ARCENE	Synthetic
Design dimensions	200×10000	500×50000
Independent sketch trials	15	5
Iteration panels: s	2400	20000
Iteration panels: λ	10	10
Sketch size sweep	$500, 750, \ldots, 2500$	$2500, 5000, \ldots, 20000$
Regularization sweep: λ	$1, 2, 5, 10, 20, 50$	$10, 20, 50, 75, 100, 150$
Updates in parameter sweeps	10	10
Regularization panels: s	2400	20000
Master seed	0	42

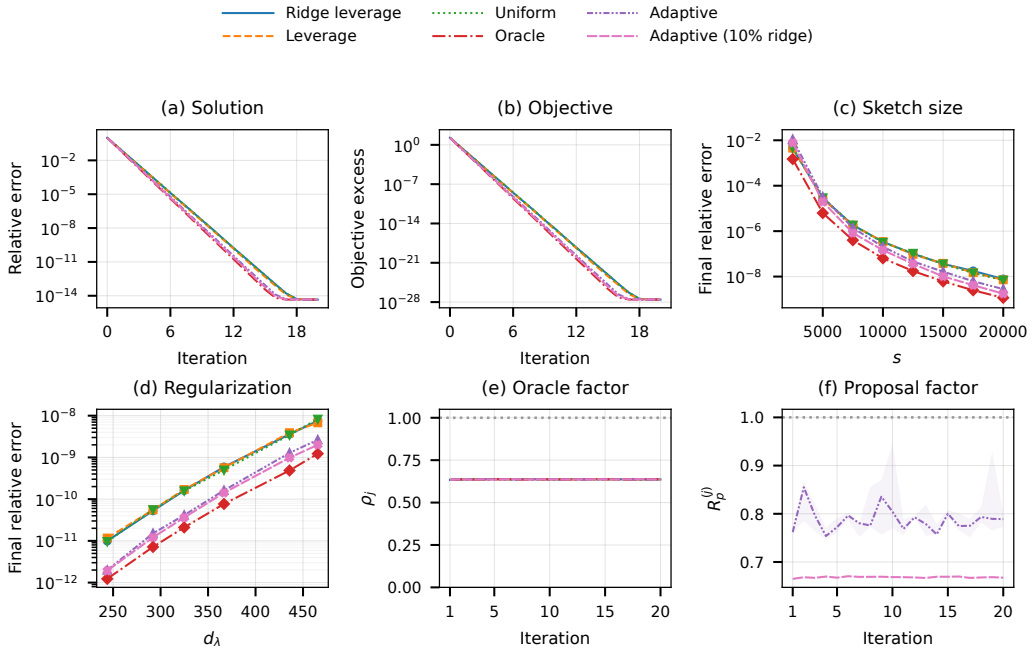


Figure 3: Synthetic comparisons with fresh sketches: solution error and normalized objective excess against iteration, final error against the column budget and effective degrees of freedom, and oracle and proposal factors along the same trajectories. Curves and bands are medians and interquartile ranges over 5 trials. Settings are given in Table 1. Panel (e) shows all six methods; (f) shows adaptive sampling and its mixture.

Fixed and fresh sketches. The separate ARCENE comparison uses $\lambda = 10$, 25 updates, and $s \in \{200, 400, 800, 1600\}$, with 15 trials and master seed 0. The reused sketch row contains uniform, leverage, and ridge leverage sampling. The fresh sketch row contains these three laws and the oracle, adaptive, and mixture proposals. These rows isolate the effect of refreshing each baseline distribution, while also showing the added probabilities.

H.3 Ridge probes on Qwen2.5 representations

We use frozen Qwen2.5-32B-Instruct representations (Qwen Team et al., 2025) of $N = 100$ STS Benchmark training examples (Cer et al., 2017), with similarity scores from 0 to 5 as labels. We concatenate representations from 20 layers at the final retained non-padding token. Each layer contributes 5120 coordinates, giving $d = 102{,}400$ features per example. The layer indices, in concate-

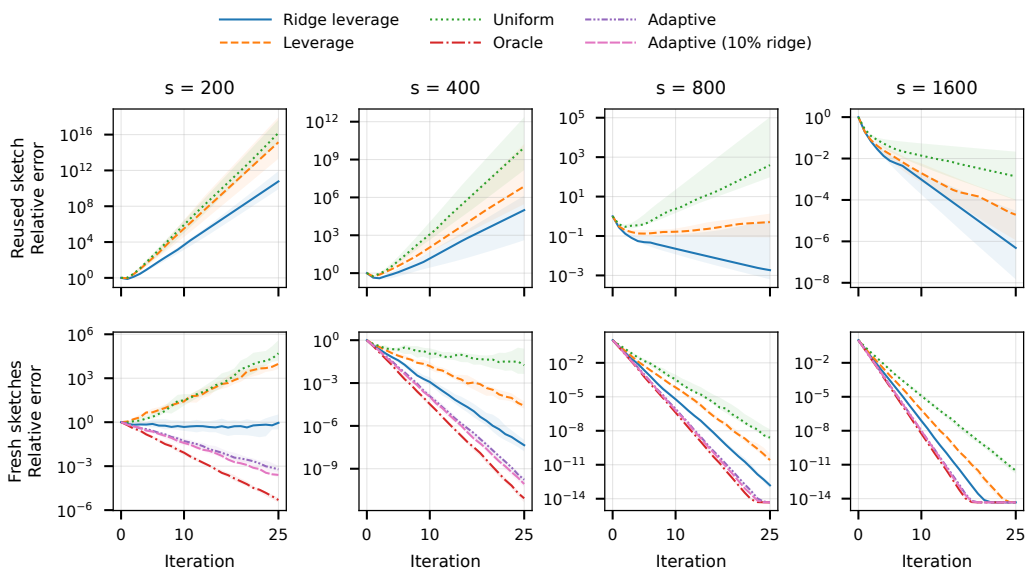


Figure 4: Reused (top) and fresh (bottom) sketches on ARCENE. Columns use $s = 200, 400, 800, 1600$; all panels use $\lambda = 10$ and 25 updates. Lines and bands are medians and interquartile ranges over 15 trials. Both rows include the three baseline laws; the fresh row also includes our oracle, adaptive and mixture proposals. Panels use separate vertical scales.

nation order, are

$$64, 32, 16, 48, 8, 24, 40, 56, 4, 12, 20, 28, 36, 44, 52, 60, 6, 22, 38, 54.$$

Centering and the fitted intercept. Let $\mathbf{F} \in \mathbb{R}^{N \times d}$ denote the raw training feature matrix, $\mathbf{y} \in \mathbb{R}^N$ its labels, and $\mathbf{1} \in \mathbb{R}^N$ the vector of ones. We convert the stored features and labels from FP32 to FP64 before computing the training means $\boldsymbol{\mu} = \mathbf{F}^\top \mathbf{1}/N$ and $\bar{y} = \mathbf{1}^\top \mathbf{y}/N$. The centered design and labels are

$$\mathbf{A} = \mathbf{F} - \mathbf{1}\boldsymbol{\mu}^\top, \qquad \mathbf{b} = \mathbf{y} - \bar{y}\mathbf{1}.$$

This accounts for an unpenalized intercept $\bar{y} - \boldsymbol{\mu}^\top \mathbf{x}$. The centered design has numerical rank 99; positive regularization keeps both the exact and sketched ridge systems positive definite.

Regularization selection. We select λ by five-fold cross-validation on the 100 training examples, using shuffled folds with seed 2026 and the candidate grid $\{10^{-4+k/2} : k = 0, \ldots, 24\}$. The penalty $\lambda\|\mathbf{x}\|_2^2$ is added to the sum of squared errors. Within each fold, features and labels are centered using training-fold means only. We choose the candidate minimizing the mean fold MSE, giving $\lambda = 10^{4.5}$, and refit on all 100 training examples.

Sampler comparisons and settings. Figure 2 compares all six sampling methods on this one design, using 30 trials and master seed 20260922. Panels (a,b,e,f) use $s = 1000$ and 20 updates at the selected regularization value. The sketch size sweep uses $s \in \{500, 1000, 2000, 4000, 8000\}$ and 10 updates; the regularization sweep uses $\lambda/4, \lambda/2, \lambda, 2\lambda, 4\lambda$ with $s = 1000$ and ten updates. It measures numerical sensitivity around the selected value without selecting a new regularization parameter.

Data and feature construction. We use the official training split of the Semantic Textual Similarity Benchmark (STS-B) Cer et al. (2017), accessed through Hugging Face (`nyu-mll/glue`, configuration `stsb`). We shuffle this split with seed 2026 and retain the first 100 examples. Each sentence pair is inserted into the following user message:

```
Consider the following two sentences.
```

```
Sentence 1:  [sentence1]
Sentence 2:  [sentence2]
Assess their semantic similarity.
```

We apply the model's chat template with an assistant generation prefix, then tokenize without adding further special tokens. Inputs are right-truncated to 128 tokens and right-padded within batches. Feature extraction uses 4-bit NF4 weights with FP16 computation; no response is generated. No feature variance standardization is applied. The code repository includes metadata files recording the selected example IDs and the dataset, model, and tokenizer revision identifiers.

H.4 Variance factor measurements

At iteration $j = 1, \dots, t$, after updating $\mathbf{b}^{(j)}$ and before drawing $\mathbf{S}^{(j)}$, we evaluate the exact solution $\mathbf{x}^{\star(j)}$ to the current residual problem using the cached SVD. With the ridge leverage weights w_i and effective dimension d_λ from Section 2, we record

$$\rho_j = \frac{\left(\sum_{i=1}^d |\mathbf{x}_i^{\star(j)}|\sqrt{w_i}\right)^2}{d_\lambda \|\mathbf{x}^{\star(j)}\|_2^2}, \qquad R_p^{(j)} = \frac{\sum_{i=1}^d |\mathbf{x}_i^{\star(j)}|^2 w_i / p_i^{(j)}}{d_\lambda \|\mathbf{x}^{\star(j)}\|_2^2},$$

where $p^{(j)}$ is the distribution actually used for that update. The first quantity is the ideal oracle factor for the current residual; the second measures the actual proposal relative to ridge proposal. In particular, $R_p^{(j)} = \rho_j$ for the oracle and $R_p^{(j)} = 1$ for ridge leverage sampling. For adaptive trajectories we measure both using the exact residual solution, rather than substituting the pilot estimate in the diagnostic. The exact solution is used only for the oracle sampler and for evaluation; the practical adaptive and mixture samplers use their pilot estimates. Different methods generally encounter different residuals after the first update.

H.5 Gaussian reference for the oracle factor

Proof of Proposition 5.2. Taking expectations over the sampled index gives

$$\mathbb{E}_i \zeta^2 = 1, \qquad \mathbb{E}_i |\zeta| = \frac{\sum_i |\mathbf{x}_i|\sqrt{w_i}}{\sqrt{d_\lambda}\,\|\mathbf{x}\|_2} = \sqrt{\rho(\mathbf{x})}.$$

Since $z \mapsto |z|$ is 1-Lipschitz and the standard normal absolute first moment is $\sqrt{2/\pi}$,

$$\left| \sqrt{\rho(\mathbf{x})} - \sqrt{\frac{2}{\pi}} \right| \le \Delta.$$

Consequently,

$$\begin{aligned} \left| \rho(\mathbf{x}) - \frac{2}{\pi} \right| &= \left| \sqrt{\rho(\mathbf{x})} - \sqrt{\frac{2}{\pi}} \right| \left(\sqrt{\rho(\mathbf{x})} + \sqrt{\frac{2}{\pi}} \right) \\ &\le \Delta \left(2\sqrt{\frac{2}{\pi}} + \Delta \right). \end{aligned}$$

□

Empirical assessment. We compute Δ for the Qwen2.5 ridge probes to assess departures of P_ζ from $\mathcal{N}(0,1)$. Figure 5 shows small distances and an overall downward trend with increasing model size and feature dimension.

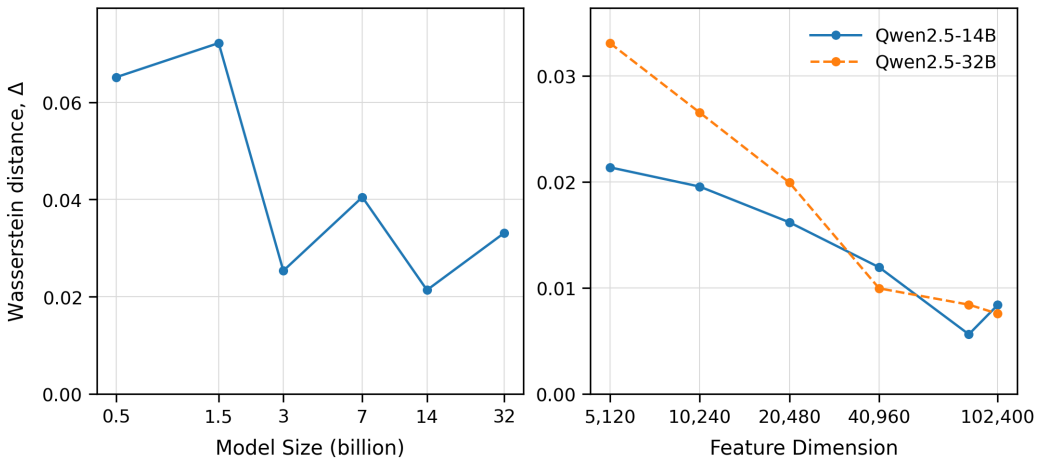


Figure 5: Wasserstein distance $\Delta = W_1(P_\zeta, \mathcal{N}(0,1))$ for Qwen2.5 ridge probes. Left: final-layer representations across model sizes. Right: representations concatenated from 1, 2, 4, 8, 16, or 20 layers of Qwen2.5-14B and Qwen2.5-32B. Both horizontal axes are logarithmic.