FROM DIMENSION TO HORIZON: KRYLOV CORESETS FOR LINEAR REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Classical coresets for linear regression preserve the objective over the full parameter space, leading to sizes that depend on the dimension d. We ask whether the optimization horizon can instead determine the amount of data needed. For any T-step gradient descent (GD) trajectory, all iterates lie in an affine Krylov space of dimension at most T. Since GD converges linearly in strongly convex settings, the required optimization horizon T can be much smaller than d. Motivated by this structure, we construct a strong Krylov coreset of size $O(\frac{T}{\epsilon^2}\log(\frac{T}{\delta}))$ that preserves the objective throughout the corresponding affine Krylov space. To further reduce the coreset size and construction cost, we develop a weak Krylov coreset based on a blockwise approximation of the GD trajectory. It has size $O(\frac{T}{\epsilon}\log(\frac{T}{\delta}))$ and returns a solution whose loss is within a factor $1+\epsilon$ of the best bounded affine combination of the exact GD iterates. With an appropriate block size, its construction requires only $O(\sqrt{T}+\log(\frac{1}{\delta}))$ full-data passes in expectation. Experiments show that Krylov coresets closely match the regression loss of full-data GD, while the weak construction further reduces coreset size and computational cost.

1 INTRODUCTION

Large-scale linear regression often requires repeated passes over the full dataset. Consider the least-squares objective

$$F(x) = \frac{1}{2}\|Ax - b\|_2^2,$$

where $A \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^n$. Starting from x_0, gradient descent (GD) performs

$$x_{t+1} = x_t - \eta A^\top (Ax_t - b),$$

so T iterations require $\Theta(T)$ full-data accesses and $O(Tnd)$ arithmetic operations when gradients are computed directly from the data. Such repeated data access can dominate the cost of large-scale optimization (Byrd et al., 2012; Friedlander & Schmidt, 2012; Bottou et al., 2018). Importantly, the optimization horizon T is governed by the target accuracy and the spectral properties of $A^\top A$, rather than directly by the full dimension d. In strongly convex settings, GD converges linearly, and the required optimization horizon T can therefore be much smaller than d (Nesterov, 2004). This raises a natural question: *if an optimization method only explores a finite number of directions, should the size of a data summary still be determined by the full dimension?*

Coresets provide a natural way to reduce this dependence on the full dataset by replacing the original problem with a small weighted subset (Feldman & Langberg, 2011; Boutsidis et al., 2013; Woodruff & Yasuda, 2023; 2024; Bansal et al., 2024; Munteanu & Omlor, 2024). For linear regression, a strong coreset preserves the objective uniformly over all $x \in \mathbb{R}^d$, while a weak coreset only requires that optimizing the reduced problem produces a near-optimal solution for the original objective (Boutsidis et al., 2013). Existing least-squares constructions achieve strong or weak guarantees with coreset sizes that scale with d, including $O(\frac{d}{\epsilon})$ rows for weak coresets (Drineas et al., 2006; Kacham & Woodruff, 2020). Sequential coresets reduce the region that must be approximated by exploiting the locality of successive iterates, but their general coreset size bounds still retain an explicit dependence on d (Huang et al., 2021). These results suggest that reducing the protected region alone may not be enough to remove the full-dimension dependence. We instead ask whether the optimization horizon itself induces a lower-dimensional structure that can control the coreset size.

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.

For least-squares GD, the finite optimization horizon has a simple algebraic structure. Let

$$H = A^\top A, \qquad g_0 = A^\top (Ax_0 - b).$$

Then the iterates generated by T GD updates lie in the affine Krylov space

$$\mathcal{K}_T = x_0 + \mathrm{span}\{g_0, Hg_0, \ldots, H^{T-1}g_0\},$$

whose dimension is at most T. Thus, when $T \ll d$, the optimization trajectory is limited to a much smaller space than the full parameter space. Moreover, the residuals generated by points in $\mathcal{K}_T$ lie in a linear subspace of dimension at most $T+1$. This suggests a different route to coreset construction: instead of preserving the regression objective over all of $\mathbb{R}^d$, preserve it only over the optimization horizon Krylov space induced by GD.

Our first result formalizes this horizon-dependent view. Since the residuals over $\mathcal{K}_T$ lie in a subspace of dimension at most $T+1$, leverage-score sampling on this residual space yields a strong Krylov coreset of size $O(\frac{T}{\epsilon^2}\log\frac{T}{\delta})$. With probability at least $1-\delta$, the coreset preserves the regression objective for every point in $\mathcal{K}_T$, rather than only at the individual GD iterates. Thus, the coreset size is controlled by the optimization horizon T instead of the full dimension d. While this establishes the desired reduction in coreset size, obtaining comparable savings in construction requires avoiding an explicit construction of the exact Krylov representation, whose direct computation still involves repeated full-data accesses.

To reduce both the coreset size and the construction cost, we next consider a weak Krylov coreset. Rather than uniformly preserving the objective over the entire Krylov space, it is sufficient to preserve the quality of the solution obtained from the GD trajectory. We therefore work with an approximate trajectory, which can be constructed using fewer full-data computations. This introduces a stability issue, since errors in the approximate iterates can be amplified when they are combined. We control this effect by restricting the solution class to bounded affine combinations of the GD iterates. This family still contains every original iterate, while remaining stable to trajectory approximation errors.

Building on this stable formulation, we construct the approximate trajectory blockwise. At each block anchor, we compute an exact residual and gradient, while the remaining updates use randomized approximations of the Hessian actions. A spectral telescoping argument controls how these approximation errors accumulate along the trajectory. We further reuse the information from the block anchors to construct a candidate row set, so that the residual basis and leverage scores need only be computed on a reduced set of rows. The resulting weak Krylov coreset has size $O(\frac{T}{\epsilon}\log\frac{T}{\delta})$ and returns a solution whose loss is within a factor $1+\epsilon$ of the best bounded affine combination of the exact GD iterates. With block size $B \le \sqrt{T}$, the construction uses $O(T/B + \log(\frac{1}{\delta}))$ full-data accesses in expectation, which becomes $O(\sqrt{T} + \log(\frac{1}{\delta}))$ for $B = \Theta(\sqrt{T})$.

1.1 Our Contributions

Our main contributions are summarized as follows.

- **Horizon-dependent strong Krylov coresets.** We show that any T-step GD trajectory is contained in an affine Krylov space of dimension at most T, and construct a strong Krylov coreset that preserves the regression objective throughout this space. With probability at least $1-\delta$, the coreset has size $O(\frac{T}{\epsilon^2}\log\frac{T}{\delta})$, governed by the optimization horizon rather than the full dimension d. This can yield a substantially smaller coreset when $T \ll d$.
- **Weak Krylov coresets.** We further develop a weak Krylov coreset that avoids first constructing the exact GD trajectory. The method combines a blockwise approximate trajectory with bounded affine combinations and candidate-row reduction. With probability at least $1-\delta$, the resulting coreset has size $O(\frac{T}{\epsilon}\log\frac{T}{\delta})$, and the returned solution has loss within a factor $1+\epsilon$ of the best bounded affine combination of the exact GD iterates. With block size $B \le \sqrt{T}$, the construction uses $O(\frac{T}{B} + \log(\frac{1}{\delta}))$ full-data accesses in expectation, yielding $O(\sqrt{T} + \log(\frac{1}{\delta}))$ for $B = \Theta(\sqrt{T})$.
- **Empirical evaluation.** Experiments on synthetic and real datasets show that the proposed Krylov coresets closely match the regression loss of full-data GD while reducing computational cost. On MNIST8M, the weak Krylov coreset achieves a relative loss difference below 0.01% and runs more than $2\times$ faster than full-data GD in the reported setting.

1.2 Other Related Work

Sequential and local coresets. A closely related approach is the sequential coreset of Huang et al. (2021), which preserves the objective over a local ball around an optimization anchor and rebuilds the coreset as the trajectory moves. For least squares, its general coreset-size bound has the form $\widetilde{O}(\Gamma_r^2 d/\epsilon^2)$, where Γ_r is a local regularity term depending on the protected radius. Thus, restricting the approximation to a local region can reduce the domain that must be preserved, but does not remove the explicit dependence on d. A related local approach was developed by Raj et al. (2020), who estimate sensitivities within a neighborhood using a quadratic approximation. These methods exploit locality along the optimization path, whereas our Krylov coresets exploit the algebraic structure induced by a finite optimization horizon, yielding a single affine subspace whose dimension is controlled by T.

Randomized least-squares methods. Randomized sketching has been widely used to accelerate least-squares optimization through preconditioning or approximate curvature information. Blendenpik (Avron et al., 2010) and LSRN (Meng et al., 2014) construct randomized preconditioners and then solve the original problem iteratively. The iterative Hessian sketch (Pilanci & Wainwright, 2016) instead combines a sketched Hessian with exact gradient information. These methods aim to accelerate the computation of an accurate regression solution, while our goal is to construct a small weighted subset of the original rows with guarantees tied to a finite GD horizon.

Randomized Krylov methods. Randomization has also been used to reduce the cost of Krylov-basis construction and residual minimization. Randomized Gram–Schmidt, sketch-and-select Arnoldi, and related block methods reduce orthogonalization and basis-construction costs (Balabanov & Grigori, 2022; 2025; Güttel & Simunec, 2024), while sketched GMRES and randomized Golub–Kahan methods accelerate optimization within Krylov spaces (Nakatsukasa & Tropp, 2024; Chung & Gazzola, 2025; Burke et al., 2025). These methods typically continue to apply the original operator while constructing or optimizing within the Krylov space. In contrast, we use approximate trajectory information to construct a weighted subset of the original data rows and explicitly bound the number of full-data accesses required for the construction.

2 Preliminaries

We write $[n] = \{1, \dots, n\}$ and use $\|\cdot\|_2$ for the Euclidean norm of a vector and the spectral norm of a matrix. We denote by e_i the ith standard basis vector. Given $A \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^n$, we consider the least squares objective $F(x) = \frac{1}{2}\|Ax - b\|_2^2$, where $a_i^\top$ denotes the ith row of A. A weighted subset of size $|S|$ consists of indices $S \subseteq [n]$ and positive weights $w = (w_i)_{i\in S}$, and defines the objective $F_S(x) = \frac{1}{2}\sum_{i\in S} w_i(a_i^\top x - b_i)^2$. When convenient, we identify the weighted subset with its corresponding row-sampling operator.

Krylov spaces. Fix an initial point $x_0 \in \mathbb{R}^d$ and a horizon $T \geq 1$, and set $H = A^\top A$, $r_0 = Ax_0 - b$, and $g_0 = A^\top r_0$. Define the Krylov subspace $\mathcal{V}_T = \operatorname{span}\{g_0, Hg_0, \dots, H^{T-1}g_0\}$, and its affine translation by $\mathcal{K}_T = x_0 + \mathcal{V}_T$. The corresponding residual space is $\mathcal{U}_T = \operatorname{span}\big(\{r_0\} \cup A\mathcal{V}_T\big)$. For any $x \in \mathcal{K}_T$, the residual $Ax - b$ lies in $\mathcal{U}_T$, which satisfies $\dim(\mathcal{U}_T) \leq T + 1$. We write $F_T^\star = \min_{x\in\mathcal{K}_T} F(x)$ for the optimal objective value over the affine Krylov space.

Leverage scores. Let $\mathcal{U} \subseteq \mathbb{R}^n$ be an r-dimensional subspace, and let $Q \in \mathbb{R}^{n\times r}$ have orthonormal columns spanning $\mathcal{U}$. The row leverage scores are $\tau_i = \|e_i^\top Q\|_2^2$ with $\sum_{i=1}^n \tau_i = r$. For $r \geq 1$, set $p_i = \tau_i/r$ and draw indices $i_1, \dots, i_m$ independently with replacement from this distribution. Let $W \in \mathbb{R}^{m\times n}$ have jth row $e_{i_j}^\top/\sqrt{mp_{i_j}}$. For $\epsilon, \delta \in (0, 1)$, standard leverage-score sampling Cohen & Peng (2015) guarantees that $m = O(\epsilon^{-2} r \log(r/\delta))$ samples suffice to ensure that, with probability at least $1 - \delta$,

$$(1 - \epsilon)\|u\|_2^2 \leq \|Wu\|_2^2 \leq (1 + \epsilon)\|u\|_2^2 \tag{1}$$

simultaneously for all $u \in \mathcal{U}$. Here, repeated samples of the same row are combined by summing their weights, resulting in a weighted subset of size at most m.

Definition 2.1 (Strong Krylov coreset)**.** Let $\epsilon \in (0,1)$. A weighted subset (S, w) is an ϵ-strong Krylov coreset for $\mathcal{K}_T$ if

$$(1-\epsilon)F(x) \le F_S(x) \le (1+\epsilon)F(x)$$

simultaneously for all $x \in \mathcal{K}_T$.

Stable families. For the weak guarantee, we consider affine combinations of the exact GD iterates. Fix a step size $\eta > 0$, and let $x_0, \dots, x_T$ satisfy $x_{t+1} = x_t - \eta A^\top (Ax_t - b)$. For $1 \le t \le T$, $x_t - x_0 = -\eta \sum_{j=0}^{t-1} (I - \eta H)^j g_0$. The coefficient of $H^{t-1} g_0$ is $(-\eta)^t$, so the displacement vectors $x_t - x_0$ span $\mathcal{V}_T$. Hence, $\mathrm{aff}\{x_0, \dots, x_T\} = \mathcal{K}_T$. Define $r_t = Ax_t - b$. Then, it holds that $\mathcal{U}_T = \mathrm{span}\{r_0, \dots, r_T\}$. To avoid affine combinations with large signed coefficients, we restrict the ℓ_1 norm of the coefficients. For $R \ge 1$, define $\Lambda_R = \left\{\lambda \in \mathbb{R}^{T+1} : \mathbf{1}^\top \lambda = 1,\ \|\lambda\|_1 \le R\right\}$, and let

$$\mathcal{X}_{T,R} = \left\{ \sum_{t=0}^{T} \lambda_t x_t : \lambda \in \Lambda_R \right\}.$$

The affine constraint ensures that $\mathcal{X}_{T,R} \subseteq \mathcal{K}_T$, while the ℓ_1 constraint limits signed combinations of the iterates. In particular, $\mathrm{conv}\{x_0, \dots, x_T\} = \mathcal{X}_{T,1} \subseteq \mathcal{X}_{T,R} \subseteq \mathcal{K}_T$. For $x(\lambda) = \sum_{t=0}^{T} \lambda_t x_t$, its residual satisfies $Ax(\lambda) - b = \sum_{t=0}^{T} \lambda_t c_t$. Let $F^\star_{T,R} = \min_{x \in \mathcal{X}_{T,R}} F(x)$. Since every GD iterate belongs to $\mathcal{X}_{T,R}$ and $\mathcal{X}_{T,R} \subseteq \mathcal{K}_T$, we have

$$F^\star_T \le F^\star_{T,R} \le \min_{0 \le t \le T} F(x_t). \tag{2}$$

We next define the weak coreset property for a general trajectory matrix. For $Z = [z_0, \dots, z_T] \in \mathbb{R}^{d \times (T+1)}$, define $\mathcal{X}_R(Z) = \{Z\lambda : \lambda \in \Lambda_R\}$.

Definition 2.2 (Weak Krylov coreset)**.** Let $\epsilon \in (0,1)$ and $R \ge 1$. A weighted subset (S, w) is an ϵ-weak Krylov coreset for Z over Λ_R if every $\widehat{\lambda} \in \arg\min_{\lambda \in \Lambda_R} F_S(Z\lambda)$ satisfies

$$F(Z\widehat{\lambda}) \le (1+\epsilon) \min_{\lambda \in \Lambda_R} F(Z\lambda).$$

For the exact trajectory $X = [x_0, \dots, x_T]$, $\xi_R(X) = \xi_{X,R}$ and the comparison optimum is $F^\star_{T,R}$.

3 Strong Krylov Coreset

We construct a strong Krylov coreset by preserving only the residuals generated by $\mathcal{K}_T$. For every $x \in \mathcal{K}_T$, the residual $Ax - b$ lies in the subspace $\mathcal{U}_T = \mathrm{span}(r_0 \cup A\mathcal{V}_T)$, whose dimension is at most $T+1$. Hence, a subspace embedding for $\mathcal{U}_T$ immediately gives uniform preservation of $F(x)$ over $\mathcal{K}_T$, reducing the relevant sampling dimension from d to T. The main algorithmic task is to construct the leverage scores of $\mathcal{U}_T$ without forming $H = A^\top A$ or separately factorizing the completed residual space.

Residual-space reduction. The Krylov restriction also limits the residuals that must be preserved. For any $x = x_0 + v \in \mathcal{K}_T$ with $v \in \mathcal{V}_T$, $Ax - b = r_0 + Av$, where $r_0 = Ax_0 - b$. Hence, every residual generated by $\mathcal{K}_T$ lies in $\mathcal{U}_T = \mathrm{span}(\{c_0\} \cup A\mathcal{V}_T)$, whose dimension is at most $T+1$. Therefore, preserving the least-squares objective over $\mathcal{K}_T$ reduces to preserving squared Euclidean norms over $\mathcal{U}_T$.

Let Q have orthonormal columns spanning $\mathcal{U}_T$, and define the row leverage scores $\tau_i = \|e_i^\top Q\|_2^2$. Since $\sum_{i=1}^{n} \tau_i = \dim(\mathcal{U}_T) \le T+1$, the leverage-score embedding in Equation (1) yields a weighted subset of $O\left(\frac{T}{\epsilon^2} \log \frac{T}{\delta}\right)$ rows that preserves $F(x)$ simultaneously for all $x \in \mathcal{K}_T$ with probability at least $1 - \delta$. Thus, the sample complexity is controlled by the optimization horizon T. The remaining task is to obtain the leverage scores of $\mathcal{U}_T$ without explicitly forming $H = A^\top A$ or separately factorizing the completed residual space.

Algorithm 1 Strong Krylov coreset

Input: $A \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^n$, initial point $x_0 \in \mathbb{R}^d$, integer horizon $T \geq 1$, and $\epsilon, \delta \in (0,1)$.
Output: A weighted original-row subset $S = (\mathcal{I}, \{\omega_i\}_{i\in\mathcal{I}})$.

Compute $r \leftarrow Ax_0 - b$.
if $\|r\|_2 = 0$ **then**
  return the empty coreset.
end if
Set $\beta \leftarrow C\epsilon^{-2}\log((T+1)/\delta)$.
Compute $U \leftarrow [\,]$, $W \leftarrow [r/\|r\|_2]$, and $v \leftarrow A^\top r$.
Draw independent $\xi_i \sim \mathrm{Unif}(0,1)$ for $i \in [n]$.
Initialize leverage scores $\tau_i \leftarrow r_i^2/\|r\|_2^2$ and inclusion probabilities $p_i \leftarrow \min\{1, \beta\tau_i\}$ for $i \in [n]$.
Set $\mathcal{I} \leftarrow \{i \in [n] : \xi_i \leq p_i\}$.
for $t = 1, \ldots, T$ **do**
  Orthogonalize $v \leftarrow v - U(U^\top v)$.
  if $\|v\|_2 = 0$ **then**
    break.
  end if
  Set $q \leftarrow v/\|v\|_2$ and append $U \leftarrow [U, q]$.
  Compute $y \leftarrow Aq$ and $r \leftarrow y - W(W^\top y)$.
  if $\|r\|_2 > 0$ **then**
    Set $w \leftarrow r/\|r\|_2$.
    Append $W \leftarrow [W, w]$.
    Update $\tau_i \leftarrow \tau_i + w_i^2$ for $i \in [n]$.
    Set $p_i \leftarrow \min\{1, \beta\tau_i\}$ for $i \in [n]$.
    Update $\mathcal{I} \leftarrow \mathcal{I} \cup \{i \in [n] : \xi_i \leq p_i\}$.
  end if
  if $t < T$ **then**
    Set $v \leftarrow A^\top y$.
  end if
end for
Set $\omega_i \leftarrow 1/p_i$ for $i \in \mathcal{I}$, using the final probabilities.
return $S = (\mathcal{I}, \{\omega_i\}_{i\in\mathcal{I}})$.

Incremental Krylov construction. We construct the Krylov and residual bases simultaneously, avoiding both the explicit formation of $H = A^\top A$ and a separate factorization of the completed residual space. Suppose q is a newly added orthonormal Krylov direction and set $y = Aq$. Orthogonalizing y against the current residual basis Q produces the next residual direction, while $A^\top y = Hq$ gives the next Krylov candidate. Thus, one multiplication by A and one by $A^\top$ advance both spaces. If a normalized residual direction w is appended to Q, the row leverage scores admit the exact update $\tau_i \leftarrow \tau_i + w_i^2$ for every $i \in [n]$. Hence, the leverage scores are maintained as the residual space grows without an additional factorization.

Since each leverage score is nondecreasing during the construction, the sampled support can be maintained with a single random threshold each row. We draw independent $\xi_i \sim \mathrm{Unif}(0,1)$ and maintain $p_i = \min\{1, \beta\tau_i\}$, where $\beta = O(\epsilon^{-2}\log(T/\delta))$. The ith row is added once $\xi_i \leq p_i$. The basis construction is independent of these thresholds, so after the final leverage scores are obtained, each row is selected independently with probability p_i. The following lemma summarizes the correctness of the incremental construction.

Lemma 3.1. *At termination of Algorithm 1, the maintained matrices U and Q have orthonormal columns spanning $\mathcal{V}_T$ and $\mathcal{U}_T$, respectively, with the corresponding stabilized spaces used if the Krylov sequence terminates before step T. The maintained scores satisfy $\tau_i = \|e_i^\top Q\|_2^2$ for every $i \in [n]$. Moreover, with $p_i = \min\{1, \beta\tau_i\}$, the final inclusion indicators are independent Bernoulli variables with probabilities p_i, and every selected row receives weight $1/p_i$.*

The incremental construction, together with leverage-score sampling on the final residual space, gives the strong Krylov coreset guarantee, whose proof is provided in Section A in Appendix.

Theorem 3.2. *Given $A \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^n$, $x_0 \in \mathbb{R}^d$, $T \geq 1$, and $\epsilon, \delta \in (0,1)$, Algorithm 1 returns, with probability at least $1-\delta$, a weighted subset of $O\left(\frac{T}{\epsilon^2}\log\frac{T}{\delta}\right)$ rows satisfying*

$$(1-\epsilon)F(x) \leq F_S(x) \leq (1+\epsilon)F(x)$$

Algorithm 2 Weak Krylov Coreset

Input: $A \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^n$, horizon $T \geq 1$, step size $\eta > 0$ with $\eta\|A\|_2^2 \leq 1$, an initial point $x_0 \in \mathbb{R}^d$ with certified factor $\rho \geq 1$, radius $R \geq 1$, an integer block size $B \leq \sqrt{T}$, and $\epsilon, \delta \in (0,1)$.

Output: A weighted row sampler S, and the approximate trajectory $\widehat{X}$ for subsequent minimization.

1: Compute $F_0 \leftarrow \frac{1}{2}\|Ax_0 - b\|_2^2$.
2: **if** $F_0 = 0$ **then**
3: Set $\widehat{X} \leftarrow [x_0, \ldots, x_0]$ and
4: **return** $(e_1^\top, \widehat{X})$.
5: **end if**
6: Set $\epsilon_0 \leftarrow C\epsilon$, $\delta_0 \leftarrow C\delta$, and $\alpha^2 \leftarrow \frac{C\epsilon_0^2}{\rho R^2 B \log(T/\delta_0)}$.
7: Construct an approximate blockwise trajectory $\widehat{X} = [\widehat{x}_0, \ldots, \widehat{x}_T]$ with the rowwise anchor energies e^{anc} by GENERATEBLOCKTRAJECTORY$(A, b, x_0, F_0, T, \eta, B, \alpha, \delta_0)$ (See Algorithm 3 in Appendix).
8: Set $M \leftarrow \left\lceil C\rho R^2 \epsilon_0^{-2}\delta_0^{-1/6} T^{1/6}\left(1 + \frac{T}{B} + B\right)\right\rceil$.
9: Construct a reweighted candidate row set S_0 by TRAJECTORYCANDIDATE$(A, b, \widehat{X}, e^{\mathrm{anc}}, M)$ (See Algorithm 4 in Appendix).
10: Set $m \leftarrow \lceil CT\epsilon_0^{-1}\log(T/\delta_0)\rceil$.
11: Set $S \leftarrow$ FINALRESPARSIFY$(S_0, A, b, \widehat{X}, T, m)$ (see Algorithm 5 in Appendix).
12: **return** $(S, \widehat{X})$.

for every $x \in \mathcal{K}_T$. The algorithm runs in $O(Tnd)$ time using $O(T)$ passes over the input. If $Ax_0 = b$, it returns the exact empty coreset.

4 WEAK KRYLOV CORESET

The strong Krylov coreset preserves the objective uniformly over $\mathcal{K}_T$, which is stronger than needed when the goal is to obtain a solution whose loss is near the optimum over the stable family $\mathcal{X}_{T,R}$. Therefore, we study weak Krylov coresets that preserve optimization quality with fewer sampled rows. In particular, the weak guarantee only requires a sufficiently accurate ridge embedding for optimization over Λ_R. A $\Theta(\sqrt{\epsilon})$ distortion, combined with an $O(\epsilon)$ ridge term, is sufficient and reduces the final coreset size from $O(T\epsilon^{-2}\log(T/\delta))$ to $O(T\epsilon^{-1}\log(T/\delta))$.

A second challenge is to construct this smaller coreset without first generating the exact trajectory. We partition the GD horizon into blocks and use each full-data computation at a block anchor to support multiple subsequent updates. At the beginning s of a block, we compute the residual and gradient exactly. For a later iterate $\widehat{x}_t$ in the same block, $\nabla F(\widehat{x}_t) = \nabla F(\widehat{x}_s) + H(\widehat{x}_t - \widehat{x}_s)$. We approximate only the Hessian correction within the block using fresh row samples. The block length B balances the number of exact anchor computations against the error accumulated by these approximate updates.

The same decomposition also provides the information needed for row selection. Each anchor scan gives the row-wise energy of the anchor residual, while the remaining residuals are represented through within-block correction directions. We estimate their row-wise contributions and combine them with the anchor energies to construct a candidate row set without forming all $T+1$ residuals over the original data. The residual basis and its leverage scores are computed only on the candidate rows, followed by a final resparsification step.

Fix a step size $\eta > 0$ satisfying $\eta\|A\|_2^2 \leq 1$, and recall that

$$F_{T,R}^\star = \min_{\lambda\in\Lambda_R} F(X\lambda). \tag{3}$$

We start from a constant-factor approximate solution x_0 satisfying $F_0 := F(x_0) \leq \rho F_\star$, where $F_\star = \min_x F(x)$ and $\rho \geq 1$ is an absolute constant. Such an initialization can be obtained by applying the fast oblivious subspace embedding of Chenakkod et al. (2024) to $[A, b]$ and solving the compressed least-squares problem in $O(nd + d^\omega)$ time with constant success probability, where ω is a fixed admissible matrix-multiplication exponent. For $F_0 > 0$, set $G = F_0/\rho$, which satisfies $G \leq F_\star$ under the initialization guarantee. Let $B \leq \sqrt{T}$ be the block size, and set $\epsilon_0 = C\epsilon$ and $\delta_0 = C\delta$ for a sufficiently small absolute constant $C > 0$.

Blockwise approximation. For each update t, let $s(t)$ denote the anchor of its block and set $v_t = \widehat{x}_t - \widehat{x}_{s(t)}$. GENERATEBLOCKTRAJECTORY (see Algorithm 3 in the Appendix) computes the gradient exactly at each anchor and estimates only the correction Hv_t for the remaining updates. The main difficulty is that the resulting gradient errors propagate through all subsequent steps. A spectral telescoping argument controls the total within-block prediction displacement $\sum_{t<T} \|Av_t\|_2^2$, and a concentration argument converts this control into a uniform prediction-error bound for the approximate trajectory $\widehat{X} = [\widehat{x}_0, \ldots, \widehat{x}_T]$.

Lemma 4.1. *With probability at least $1 - C\delta_0$, the trajectory returned by* GENERATEBLOCKTRAJECTORY *satisfies*

$$\max_{0\le t\le T} \|A(\widehat{x}_t - x_t)\|_2 \le C\frac{\epsilon_0}{R}\sqrt{G}.$$

On the same event, for every $\lambda \in \Lambda_R$, the following inequalities hold

$$(1 - C\epsilon_0)F(X\lambda) \le F(\widehat{X}\lambda) \le (1 + C\epsilon_0)F(X\lambda).$$

The detailed proof appears in Lemma B.4. The pointwise trajectory bound alone is not sufficient for optimization over affine combinations. The constraint $\|\lambda\|_1 \le R$ limits the amplification of these errors and yields the uniform comparison in Theorem 4.1. Consequently, optimization can be carried out over the approximate trajectory $\widehat{X}$ while its solution is compared with $F^\star_{T,R}$.

Candidate reduction. Reducing the cost of trajectory generation alone is not sufficient. Explicitly forming all residuals $A\widehat{x}_t - b$ over the original data would still require $O(Tnd)$ time. We therefore reuse the block decomposition to reduce the number of rows before constructing the residual basis and computing its leverage scores. For a block anchored at s, $A\widehat{x}_t - b = (A\widehat{x}_s - b) + Av_t$, and $v_t = \widehat{x}_t - \widehat{x}_s$. The full-data computation at the anchor gives the row-wise energy of the anchor residual exactly. For the correction vectors Av_t, a constant number of Gaussian probes per block estimates their row-wise energies. TRAJECTORYCANDIDATE (see Algorithm 4 in the Appendix) combines these quantities to construct a reweighted candidate set S_0.

For the analysis, let $\mathcal{C}$ contain the block-anchor residuals, the terminal residual, and the within-block correction vectors. There exists a linear map θ such that $A\widehat{X}\lambda - b = \mathcal{C}\theta(\lambda)$ for every $\lambda \in \Lambda_R$, with $\|\theta(\lambda)\|_2^2 \le 2R^2$. The blockwise displacement bound further gives

$$\|\mathcal{C}\|_F^2 \le C\left(1 + \frac{T}{B} + B\right)F_0.$$

These bounds yield a candidate sample size

$$M = O\left(\rho R^2 \epsilon_0^{-2}\delta_0^{-1/6}T^{1/6}\left(\frac{T}{B} + B\right)\right).$$

The residual basis and its leverage scores are then computed only on S_0, avoiding a factorization over all n rows.

Resparsification for optimization. On the candidate rows, we construct the residual basis and compute its leverage scores. We then apply a final leverage-score sampling step to obtain the weak Krylov coreset. For the weak guarantee, it is sufficient that the sampled Gram matrix approximates the original one in the following sense. Let $\Gamma = \mathcal{C}^\top\mathcal{C}$ and $\Gamma_S = \mathcal{C}^\top S^\top S\mathcal{C}$.

Lemma 4.2. *Suppose $A\widehat{X}\lambda - b = \mathcal{C}\theta(\lambda)$ for $\lambda \in \Lambda_R$, where θ is linear and $\|\theta(\lambda)\|_2^2 \le 2R^2$. Let $\zeta \ge 0$ and let $\gamma > 0$ be smaller than a sufficiently small absolute constant. If $-\gamma(\Gamma + \zeta I) \preceq \Gamma_S - \Gamma \preceq \gamma(\Gamma + \zeta I)$, then every $\widehat{\lambda} \in \arg\min_{\lambda\in\Lambda_R} F_S(\widehat{X}\lambda)$ satisfies*

$$F(\widehat{X}\widehat{\lambda}) \le (1 + C\gamma^2)\min_{\lambda\in\Lambda_R} F(\widehat{X}\lambda) + C\zeta R^2.$$

We set $\zeta = O(\epsilon_0 G/R^2)$ and $\gamma = O(\gamma\sqrt{\epsilon_0})$. Since $\min_{\lambda\in\Lambda_R} F(\widehat{X}\lambda) \ge G$, the lemma gives an $O(\epsilon_0)$ relative optimization error. The candidate sampling and final resparsification together ensure that the above matrix approximation holds for these choices of γ and ζ. Algorithm 2 gives the complete construction, with the proof deferred to Section B in Appendix.

Theorem 4.3. *Under the conditions above and suppose* $n \geq O\left(\frac{\rho R^2 dT}{\epsilon^2}\log\frac{T}{\delta}\right)$*, Algorithm 2 returns* $\widehat{X}$ *and a weighted subset of* $O(\frac{T}{\epsilon}\log\frac{T}{\delta})$ *rows. With probability at least* $1-\delta$*, every* $\widehat{\lambda} \in \arg\min_{\lambda\in\Lambda_R} F_S(\widehat{X}\lambda)$ *satisfies*

$$\max\{F(\widehat{X}\widehat{\lambda}), F(X\widehat{\lambda})\} \leq (1+\epsilon)F^\star_{T,R}.$$

Then, the expected construction time is $O\Big(\big(\frac{T}{B}+\log\frac{1}{\delta}\big)\,nd+\frac{R^2T^{13/6}}{B\epsilon^2\delta^{1/6}}d+d^\omega\log\frac{1}{\delta}\Big)$ *with* $O(\frac{T}{B}+\log\frac{1}{\delta})$ *full-data accesses, where* ω *is a fixed matrix-multiplication exponent, and* $B \leq \sqrt{T}$.

5 Numerical Experiments

We evaluate the proposed StrongKrylov and WeakKrylov coresets on one synthetic and three real-world datasets against standard coreset baselines for least squares. All experiments are implemented in MATLAB R2024b and run on a server with two AMD EPYC 7742 CPUs and 1.5 TB of RAM.

Datasets. We consider the CT Slices dataset ($n = 53{,}500$, $d = 386$), YearPredictionMSD ($n = 515{,}345$, $d = 91$), and MNIST8M ($n = 8{,}100{,}000$, $d = 784$), together with a synthetic instance ($n = 12{,}000$, $d = 64$). For each real-world dataset, we use all columns as the design matrix A. We draw $\theta \sim \mathcal{N}(0, I_d)/\sqrt{d}$, set $s = A\theta$, and generate $b = \widetilde{s} + \xi$, $\widetilde{s} = \frac{s-\mathrm{mean}(s)\mathbf{1}}{\mathrm{std}(s)}$, and $\xi \sim \mathcal{N}(0, I_n)$. For the synthetic instance, we generate a Gaussian design and set $A = G\Sigma Q^\top$ with $\Sigma_{jj} = 10^{-2(j-1)/(d-1)}$ for $j \in [d]$, where G has i.i.d. standard Gaussian entries, Q is a random orthogonal matrix. All three real-world datasets are publicly available.

Algorithms. We compare StrongKrylov and WeakKrylov with four baselines: FullGD, which runs GD on the full data; ImpSamp (Tukan et al., 2020), which performs importance sampling using row scores from the augmented matrix $[A, b]$; SeqCore (Huang et al., 2021), which maintains a local coreset around the current iterate and rebuilds it when the iterate leaves a ball of radius r; and Uniform, which samples rows uniformly at random. These baselines cover full-data optimization, static global sampling, and adaptive local coreset construction. We fix $r = 0.01$ for SeqCore on all datasets and report results for other choices of r in the appendix.

Methodology. All methods use the same initialization, step size, horizon T, and stability radius R. FullGD generates the exact trajectory $X = [x_0, \ldots, x_T]$ on the original data. StrongKrylov recovers the same trajectory from the Krylov basis and minimizes its coreset objective over Λ_R, while WeakKrylov jointly constructs an approximate trajectory $\widehat{X}$ and a weighted subset and performs the same minimization using $\widehat{X}$. For SeqCore, we run its original local reconstruction procedure for T steps and apply the stable-combination postprocessing to the saved iterates using its final active local coreset. This postprocessing is used only for comparison and is not part of the original SeqCore guarantee. Each reported solution is evaluated on the full objective $F(x) = \frac{1}{2}\|Ax-b\|_2^2$. Randomized methods are repeated 10 times. We use $T = 32$, $R = 2$, and target coreset sizes $m \in \{200, 400, \ldots, 2000\}$. WeakKrylov uses $B = 4$, and SeqCore uses $r = 0.01$ in the main experiments.

Results for Krylov Coresets. As shown in Figure 1, StrongKrylov and WeakKrylov achieve regression losses very close to FullGD over the tested coreset sizes. For MNIST8M (Table 1) dataset, at $m = 1000$, the loss of WeakKrylov is only 0.005% higher than FullGD, while StrongKrylov matches FullGD to the reported precision. In comparison, the regression loss of Uniform and ImpSamp are 43.2% and 3.6% higher than FullGD, respectively. WeakKrylov further reduces the computational cost of StrongKrylov while preserving nearly the same loss. On MNIST8M, at $m = 1000$, WeakKrylov is $1.85\times$ faster than StrongKrylov and $2.40\times$ faster than FullGD. At $m = 2000$, the corresponding speedups are $2.16\times$ and $2.43\times$. Moreover, SeqCore can also achieve low regression loss with lower runtime, but its performance depends on the local radius r and on repeated coreset reconstruction. The additional results in Section C show that smaller radii reduce the loss but increase runtime because the local coreset is rebuilt more often, while larger radii reduce runtime but increase the loss. Overall, both Krylov coresets achieve regression losses close to

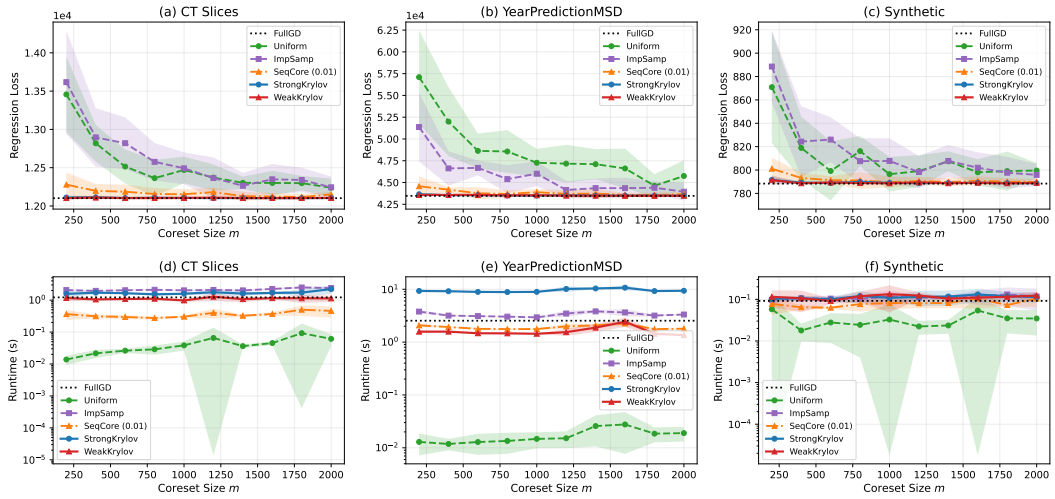

Figure 1: Comparison of coreset methods for gradient descent in terms of regression loss and runtime with varying coreset size m on CT Slices ($n = 53{,}500, d = 386$), YearPredictionMSD ($n = 515{,}345, d = 91$), and Synthetic ($n = 12{,}000, d = 64$). SeqCore uses the fixed radius $r = 0.01$ on all datasets.

Table 1: Comparison of coreset methods for gradient descent on MNIST8M ($n = 8{,}100{,}000, d = 784$) in terms of regression loss and runtime at coreset sizes $m \in \{1000, 2000\}$.

Method	Regression Loss		Time (s)	
	$m = 1{,}000$	$m = 2{,}000$	$m = 1{,}000$	$m = 2{,}000$
FullGD	**1.8394e+07**		**389.1**	
Uniform	2.6337e+07 ± 1.44e+06	2.6037e+07 ± 1.68e+06	0.1042	0.1458
ImpSamp	1.9059e+07 ± 2.58e+06	1.8124e+07 ± 1.16e+06	670.8	639.9
SeqCore (0.01)	1.8464e+07 ± 6.58e+04	1.8435e+07 ± 3.49e+04	71.72	75.14
StrongKrylov	**1.8394e+07 ± 1.55e-01**	**1.8394e+07 ± 6.52e-02**	300.8	345.1
WeakKrylov	1.8395e+07 ± 2.52e+03	1.8395e+07 ± 2.05e+03	162.2	160.0

FullGD using only a small subset of the original rows, while WeakKrylov substantially reduces the computational cost of the strong construction.

6 Conclusion

In this paper, we introduced Krylov coresets for linear regression, motivated by the observation that finite-horizon optimization explores a much smaller space than the full parameter space. In particular, any T-step GD trajectory lies in an affine Krylov space of dimension at most T. This shifts the dependence of the coreset size from the full parameter dimension d to the optimization horizon T. Based on this structure, we developed a strong Krylov coreset of size $O(\frac{T}{\epsilon^2} \log \frac{T}{\delta})$ that preserves the regression objective throughout the entire affine Krylov space. We further developed a weak Krylov coreset that reduces both the coreset size and the construction cost. By combining a blockwise approximation of the GD trajectory with bounded affine combinations and candidate-row reduction, the weak coreset has size $O(\frac{T}{\epsilon} \log \frac{T}{\delta})$ and achieves a $(1+\epsilon)$ approximation relative to the best bounded affine combination of the exact GD iterates. With an appropriate block size, the construction requires only $O(\sqrt{T} + \log(\frac{1}{\delta}))$ full-data accesses in expectation. Experiments on synthetic and real datasets show that both Krylov coresets closely match the regression loss of full-data GD, while the weak construction provides additional computational savings. Overall, our results show that data summarization can exploit the structure induced by the optimization horizon T, rather than by the full dimension d. A natural direction for future work is to develop horizon-dependent coresets for ℓ_p linear regression.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language polishing of the manuscript and assistance in drafting validation and diagnostic code. All AI-assisted arguments were checked by the authors, experimental results were obtained by running the authors' implementations, and cited references were manually verified. The authors reviewed all AI-assisted content and take responsibility for the final text, claims, proofs, experiments, and results presented in this work.

ETHICS STATEMENT

This work studies randomized coreset constructions for least-squares optimization and does not involve human-subject studies or the collection of new personal data. Our experiments use publicly available benchmark datasets and synthetically generated responses, together with a synthetic dataset. We do not foresee ethical concerns arising from this work.

REPRODUCIBILITY STATEMENT

We provide the assumptions, algorithms, and complete proofs for the Strong and Weak Krylov coreset results in the main paper and appendix. The experimental section specifies the datasets, synthetic-data and response-generation procedures, initialization, and other optimization parameters. Additional experimental settings and parameter studies are provided in the appendix to facilitate reproduction of the reported results.

REFERENCES

Haim Avron, Petar Maymounkov, and Sivan Toledo. Blendenpik: Supercharging lapack's least-squares solver. *SIAM Journal on Scientific Computing*, 32(3):1217–1236, 2010.

Oleg Balabanov and Laura Grigori. Randomized gram–schmidt process with application to GMRES. *SIAM Journal on Scientific Computing*, 44(3):A1450–A1474, 2022.

Oleg Balabanov and Laura Grigori. Randomized block gram–schmidt process for the solution of linear systems and eigenvalue problems. *SIAM Journal on Scientific Computing*, 47(1):A553–A585, 2025.

Nikhil Bansal, Vincent Cohen-Addad, Milind Prabhu, David Saulpic, and Chris Schwiegelshohn. Sensitivity sampling for k-means: Worst case and stability optimal coreset bounds. In *Proceedings of the 65th IEEE Annual Symposium on Foundations of Computer Science*, pp. 1707–1723, 2024.

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

Christos Boutsidis, Petros Drineas, and Malik Magdon-Ismail. Near-optimal coresets for least-squares regression. *IEEE Transactions on Information Theory*, 59(10):6880–6892, 2013.

Liam Burke, Stefan Güttel, and Kirk M Soodhalter. Gmres with randomized sketching and deflated restarting. *SIAM Journal on Matrix Analysis and Applications*, 46(1):702–725, 2025.

Richard H Byrd, Gillian M Chin, Jorge Nocedal, and Yuchen Wu. Sample size selection in optimization methods for machine learning. *Mathematical Programming*, 134(1):127–155, 2012.

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*, pp. 1106–1117, 2024.

Julianne Chung and Silvia Gazzola. Randomized krylov methods for inverse problems. *arXiv preprint arXiv:2508.20269*, 2025.

Michael B. Cohen and Richard Peng. ℓ_p row sampling by Lewis weights. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing*, pp. 183–192, 2015.

James Demmel, Ioana Dumitriu, and Olga Holtz. Fast linear algebra is stable. *Numerische Mathematik*, 108(1):59–91, 2007.

Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Sampling algorithms for ℓ_2 regression and applications. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1127–1136, 2006.

Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing*, pp. 569–578, 2011.

Michael P Friedlander and Mark Schmidt. Hybrid deterministic-stochastic methods for data fitting. *SIAM Journal on Scientific Computing*, 34(3):A1380–A1405, 2012.

Stefan Güttel and Igor Simunec. A sketch-and-select arnoldi process. *SIAM Journal on Scientific Computing*, 46(4):A2774–A2797, 2024.

Jiawei Huang, Ruomin Huang, Wenjie Liu, Nikolaos Freris, and Hu Ding. A novel sequential coreset method for gradient descent algorithms. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 4412–4422, 2021.

Claude-Pierre Jeannerod, Clément Pernet, and Arne Storjohann. Rank-profile revealing gaussian elimination and the CUP matrix decomposition. *Journal of Symbolic Computation*, 56:46–68, 2013.

Marius Junge and Qiang Zeng. Noncommutative bennett and rosenthal inequalities. *The Annals of Probability*, 41(6):4287–4316, 2013.

Praneeth Kacham and David Woodruff. Optimal deterministic coresets for ridge regression. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pp. 4141–4150, 2020.

Xiangrui Meng, Michael A Saunders, and Michael W Mahoney. LSRN: A parallel iterative solver for strongly over-or underdetermined systems. *SIAM Journal on Scientific Computing*, 36(2): C95–C118, 2014.

Alexander Munteanu and Simon Omlor. Optimal bounds for ℓ_p sensitivity sampling via ℓ_2 augmentation. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 36769–36796, 2024.

Yuji Nakatsukasa and Joel A Tropp. Fast and accurate randomized algorithms for linear systems and eigenvalue problems. *SIAM Journal on Matrix Analysis and Applications*, 45(2):1183–1214, 2024.

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

Yang Peng, Liangyu Zhang, and Zhihua Zhang. Statistical efficiency of distributional temporal difference learning. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pp. 24724–24761, 2024.

Mert Pilanci and Martin J Wainwright. Iterative hessian sketch: Fast and accurate solution approximation for constrained least-squares. *Journal of Machine Learning Research*, 17(53):1–38, 2016.

Anant Raj, Cameron Musco, and Lester Mackey. Importance sampling via local sensitivity. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pp. 3099–3109, 2020.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

Murad Tukan, Alaa Maalouf, and Dan Feldman. Coresets for near-convex functions. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 997–1009, 2020.

David P. Woodruff and Taisuke Yasuda. Sharper bounds for ℓ_p sensitivity sampling. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 37238–37272, 2023.

David P. Woodruff and Taisuke Yasuda. Coresets for multiple ℓ_p regression. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 53202–53233, 2024.

A STRONG KRYLOV CORESETS

This section proves the strong Krylov coreset guarantee for Algorithm 1. We first identify a low-dimensional space containing the residuals generated over the affine Krylov space. Then, we show that the incremental construction samples rows according to the leverage scores of this residual space.

Let W have orthonormal columns spanning $\mathcal{U}_T$ and define $\tau_i = \|e_i^\top W\|_2^2$. We also define the Krylov sensitivity of row i by

$$\sigma_i = \sup_{\substack{x \in \mathcal{K}_T \\ F(x)>0}} \frac{(a_i^\top x - b_i)^2}{\|Ax-b\|_2^2}.$$

Lemma A.1. *For any choice of scalar GD step sizes, $x_t \in \mathcal{K}_t$ for every $0 \le t \le T$. The residuals over $\mathcal{K}_T$ lie in $\mathcal{U}_T$, where $\dim(\mathcal{U}_T) \le T+1$. Moreover, $\sigma_i \le \tau_i$ for every row i, and $\sum_{i=1}^n \tau_i = \dim(\mathcal{U}_T) \le T+1$. Hence, the total trajectory sensitivity is at most $T+1$.*

Proof. We first show that the GD trajectory remains in the affine Krylov spaces. Let $h_t = x_t - x_0$. Since $A^\top(Ax_t - b) = g_0 + Hh_t$, the GD update gives

$$h_{t+1} = (I - \eta_t H)h_t - \eta_t g_0.$$

Starting from $h_0 = 0$, induction shows that $h_t \in \mathcal{V}_t$. Indeed, if $h_t \in \mathcal{V}_t$, then $Hh_t \in \mathcal{V}_{t+1}$, while both h_t and g_0 also belong to $\mathcal{V}_{t+1}$. Thus, $h_{t+1} \in \mathcal{V}_{t+1}$ and hence $x_t \in \mathcal{K}_t$ for every $t \le T$.

This Krylov containment also gives the low-dimensional residual space. Let U and W have orthonormal columns spanning $\mathcal{V}_T$ and $\mathcal{U}_T$, respectively. Define $\tau_i = \|e_i^\top W\|_2^2$ and

$$\sigma_i = \sup_{x \in \mathcal{K}_T, F(x)>0} \frac{(a_i^\top x - b_i)^2}{\|Ax-b\|_2^2}.$$

Let U be an orthonormal basis for $\mathcal{V}_T$. Every $x \in \mathcal{K}_T$ can be written as $x = x_0 + Uz$, and therefore

$$Ax - b = r_0 + AUz \in \mathrm{col}([r_0, AU]) = \mathcal{U}_T.$$

Moreover, U has $\dim(\mathcal{V}_T) \le T$ columns, so $\dim(\mathcal{U}_T) \le T+1$. It remains to bound the trajectory sensitivities. Consider any $x \in \mathcal{K}_T$ with nonzero residual $v = Ax - b$. Since $v \in \mathcal{U}_T$, write $v = Wu$. Since the columns of W are orthonormal, $\|v\|_2 = \|u\|_2$. The fraction of the squared residual contributed by row i is therefore bounded by its leverage score as

$$\frac{v_i^2}{\|v\|_2^2} = \frac{\left((e_i^\top W)u\right)^2}{\|u\|_2^2} \le \|e_i^\top W\|_2^2 = \tau_i,$$

where the inequality follows from Cauchy–Schwarz. Taking the supremum over $x \in \mathcal{K}_T$ gives $\sigma_i \le \tau_i$. Finally, orthonormality of W yields $\sum_i \tau_i = \|W\|_F^2 = \dim(\mathcal{U}_T) \le T+1$, and therefore, we conclude $\sum_i \sigma_i \le T+1$. □

By Lemma A.1, every residual $Ax - b$ with $x \in \mathcal{K}_T$ lies in $\mathcal{U}_T$. Hence, it suffices to preserve squared Euclidean norms over $\mathcal{U}_T$. For a weighted row subset $S = (\mathcal{I}, \{\omega_i\}_{i\in\mathcal{I}})$, let R_S contain the row $\sqrt{\omega_i} e_i^\top$ for each $i \in \mathcal{I}$. Then $F_S(x) = \frac{1}{2}\|R_S(Ax-b)\|_2^2$. Therefore, if

$$(1-\epsilon)\|v\|_2^2 \le \|R_S v\|_2^2 \le (1+\epsilon)\|v\|_2^2$$

holds for every $v \in \mathcal{U}_T$, then $(1-\epsilon)F(x) \le F_S(x) \le (1+\epsilon)F(x)$ for every $x \in \mathcal{K}_T$.

We next prove Lemma 3.1, which establishes that the incremental construction recovers both spaces and produces independent leverage-score sampling from the final residual space.

Proof of Lemma 3.1. If $r_0 = 0$, the algorithm returns immediately and the claim is trivial. We therefore assume $r_0 \neq 0$.

We first show that the accepted parameter directions form an orthonormal basis for the full Krylov space. Initially, U is empty. If $g_0 = 0$, then $\mathcal{V}_T = \{0\}$ and no parameter direction is added. Suppose that $g_0 \neq 0$. The first direction is $q_1 = g_0/\|g_0\|_2$, which spans $\mathcal{V}_1$.

Suppose that after j directions have been accepted, $q_1, \ldots, q_j$ form an orthonormal basis for $\mathcal{V}_j$. If $j < T$, the algorithm orthogonalizes $Hq_j = A^\top(Aq_j)$ against the current basis. Since $q_1, \ldots, q_{j-1}$ span $\mathcal{V}_{j-1}$ and $H\mathcal{V}_{j-1} \subseteq \mathcal{V}_j$, we have $\mathcal{V}_{j+1} = \mathcal{V}_j + \operatorname{span}\{Hq_j\}$. Thus, if Hq_j has a nonzero component orthogonal to $\mathcal{V}_j$, normalizing this component produces a new direction q_{j+1} and extends the basis to $\mathcal{V}_{j+1}$.

It remains to justify early stopping. If the orthogonal component of Hq_j is zero, then $Hq_j \in \mathcal{V}_j$. For every $\ell < j$, we also have $Hq_\ell \in \mathcal{V}_{\ell+1} \subseteq \mathcal{V}_j$. Since $q_1, \ldots, q_j$ span $\mathcal{V}_j$, it follows that $H\mathcal{V}_j \subseteq \mathcal{V}_j$. Together with $g_0 \in \mathcal{V}_j$, this implies that every subsequent Krylov vector $H^k g_0$ remains in $\mathcal{V}_j$. Hence, $\mathcal{V}_j = \mathcal{V}_T$, so stopping at this point does not omit any Krylov direction.

We next show that the residual basis and leverage scores are maintained exactly. Initially, $W = r_0/\|r_0\|_2$ spans $\operatorname{col}(r_0)$, and $\tau_i = (r_0)_i^2/\|r_0\|_2^2 = \|e_i^\top W\|_2^2$. Suppose that the current W is an orthonormal basis for $\operatorname{col}([r_0, AU])$, and a new parameter direction q is appended to U. The residual space then enlarges by Aq. Let $z = Aq - WW^\top Aq$. If $z = 0$, the residual space is unchanged. Otherwise, $w = z/\|z\|_2$ is orthogonal to the current basis, and appending w gives an orthonormal basis for the enlarged residual space. The squared row norms satisfy

$$\|e_i^\top [W, w]\|_2^2 = \|e_i^\top W\|_2^2 + w_i^2,$$

which is exactly the update $\tau_i \leftarrow \tau_i + w_i^2$. Therefore, the maintained scores are the leverage scores of the current residual space throughout the construction. Since the final U spans $\mathcal{V}_T$, the final W spans $\operatorname{col}([r_0, AU]) = \mathcal{U}_T$, and hence $\tau_i = \|e_i^\top W\|_2^2$.

The basis construction and the final leverage scores are independent of the thresholds $\{\xi_i\}_{i=1}^n$. Hence the final probabilities $p_i^\star$ are fixed independently of these thresholds. Since the thresholds are mutually independent, the final selection indicators are independent Bernoulli random variables with probabilities $p_i^\star$. The algorithm assigns weight $1/p_i^\star$ to every selected row, which proves the claim. □

The preceding lemmas identify the residual space and characterize the final sample. We now establish the strong coreset guarantee as follows.

Proof of Theorem 3.2. If $r_0 = Ax_0 - b = 0$, then $\mathcal{K}_T = \{x_0\}$ and F vanishes on this space, so the empty subset is exact. Computing the initial residual takes $O(nd)$ time. Therefore, we assume $r \neq 0$.

We first prove uniform preservation over the residual space $\mathcal{U}_T$. By Lemma 3.1, the final matrix W has orthonormal columns spanning $\mathcal{U}_T$, and row i is selected independently with probability $p_i = \min\{1, \beta\tau_i\}$, where $\beta = C\epsilon^{-2}\log((T+1)/\delta)$ and $\tau_i = \|e_i^\top W\|_2^2$. Let $z_i = W^\top e_i$ and let χ_i be the indicator that row i is selected. Then, $\|z_i\|_2^2 = \tau_i$ and $\sum_i z_i z_i^\top = I$. The error in squared norms over $\mathcal{U}_T$ is represented by

$$W^\top R_S^\top R_S W - I = \sum_i X_i,$$

where $X_i = (\chi_i/p_i - 1)z_i z_i^\top$ for $p_i > 0$ and $X_i = 0$ for $p_i = 0$. The latter case is valid since $p_i = 0$ implies $\tau_i = \|z_i\|_2^2 = 0$.

The matrices X_i are independent, self-adjoint, and mean zero. The cases $p_i \in \{0, 1\}$ contribute zero. If $0 < p_i < 1$, then $p_i = \beta\|z_i\|_2^2$, which gives

$$\|X_i\|_2 \leq \max\left\{\frac{1-p_i}{p_i}\|z_i\|_2^2, \|z_i\|_2^2\right\} \leq \frac{1}{\beta}.$$

The corresponding second moment satisfies

$$\mathbb{E}[X_i^2] = \frac{1-p_i}{p_i}\|z_i\|_2^2 z_i z_i^\top \preceq \frac{1}{\beta} z_i z_i^\top.$$

Hence, $\left\|\sum_i \mathbb{E}[X_i^2]\right\|_2 \leq 1/\beta$. Therefore, the self-adjoint matrix Bernstein inequality (Tropp, 2012, Theorem 6.1) gives

$$\Pr\left[\left\|W^\top R_S^\top R_S W - I\right\|_2 > \epsilon\right] \leq 2\dim(\mathcal{U}_T)\exp\left(-\frac{3\beta\epsilon^2}{8}\right) \leq \frac{\delta}{2},$$

where the last inequality follows from $\dim(\mathcal{U}_T) \le T+1$ and a sufficiently large choice of C.

On the complementary event, every $v = Wu \in \mathcal{U}_T$ satisfies

$$\left|\|R_S v\|_2^2 - \|v\|_2^2\right| \le \epsilon \|v\|_2^2.$$

By Lemma A.1, $Ax - b \in \mathcal{U}_T$ for every $x \in \mathcal{K}_T$. Substituting $v = Ax - b$ and using $F(x) = \frac{1}{2}\|Ax - b\|_2^2$ proves the stated objective guarantee. The same lemma gives $x_t \in \mathcal{K}_T$ for every GD iterate, and since $\mathcal{K}_T$ is affine, $\mathrm{aff}\{x_0, \dots, x_T\} \subseteq \mathcal{K}_T$.

We next bound the number of sampled rows. Let $N = |\mathcal{I}|$. Using the leverage-score identity from Lemma A.1,

$$\mathbb{E}[N] = \sum_i p_i \le \beta \sum_i \tau_i \le \beta(T+1).$$

Since the selection indicators are independent, applying Chernoff bound gives

$$\Pr[N > 2\beta(T+1)] \le \exp\left(-\frac{\beta(T+1)}{3}\right) \le \frac{\delta}{2}$$

for a sufficiently large C. A union bound with the norm-preservation event shows that both guarantees hold with probability at least $1 - \delta$. Combining $N = O(\beta T)$ with the deterministic bound $N \le n$ gives the claimed coreset size.

It remains to bound the construction time. Computing r_0 and $g_0 = A^\top r_0$ takes $O(nd)$ time. Let $k = \dim(\mathcal{V}_T) \le \min\{T, n, d\}$. The products with A and $A^\top$ over the accepted Krylov directions take $O(Tnd)$ time. Orthogonalizing the parameter and residual directions costs $O(dk^2 + nk^2) = O(Tnd)$. Updating the leverage scores, sampling probabilities, and thresholds costs $O(nT)$ and is also dominated by $O(Tnd)$. Hence, the construction takes $O(Tnd)$ time. The products with A and $A^\top$ require $O(T)$ passes over the input. □

B Proofs of the Weak Trajectory Coreset Guarantee

This section proves Theorem 4.3 and provides the complete descriptions and analyses of the three subroutines used in Algorithm 2.

Constant-factor initialization. The initial point required by Algorithm 2 can be obtained by a fast oblivious subspace embedding. The following lemma gives the initialization guarantee used in our construction.

Lemma B.1. *There is a randomized procedure that computes a constant-factor initial point in $O(nd + d^\omega)$ time. With constant probability, it returns x_0 satisfying $F(x_0) \le \rho F_\star$ for an absolute constant $\rho \ge 1$.*

Proof. We apply the fast oblivious subspace embedding from (Chenakkod et al., 2024, Theorem 1.4) to $Y = [A, b]$. With constant probability, the resulting matrix ΠY has $O(d)$ rows and preserves the Euclidean norm of every vector in $\mathrm{col}(Y)$ within constant factors. Define the compressed objective $\widetilde{F}(x) = \frac{1}{2}\|\Pi(Ax - b)\|_2^2$ and let $\widetilde{x}$ be one of its minimizers. The embedding guarantee and optimality of $\widetilde{x}$ imply

$$F(\widetilde{x}) \le C\widetilde{F}(\widetilde{x}) \le C\widetilde{F}(x^\star) \le CF(x^\star) = CF_\star$$

for an absolute constant C.

For rank-deficient ΠA, a compressed minimizer can be obtained by selecting a maximal independent set of its columns using deterministic rank-profile elimination and solving the resulting full-rank least-squares problem. This takes $O(d^\omega)$ time (Jeannerod et al., 2013; Demmel et al., 2007). The embedding can be formed in $O(nd + d^{2+\tau}\,\mathrm{polylog}\, d)$ time for any fixed $\tau > 0$. Choosing $\tau < \omega - 2$ gives the stated $O(nd + d^\omega)$ bound. □

B.1 Approximating the gradient-descent trajectory

GenerateBlockTrajectory approximates the exact GD trajectory using sampled gradient corrections within each block. The subroutine validates the anchor residual energy and within-block

displacements and returns the first accepted trial, while Gaussian probes estimate the total prediction-displacement energy without forming all correction vectors on the original data. We show that each trial is accepted with constant probability and that the returned trajectory has uniformly small prediction error.

Algorithm 3 GENERATEBLOCKTRAJECTORY

Input: $A, b, x_0, F_0, T, \eta, B, \alpha, \delta_0$ with $F_0 = F(x_0) > 0$.
Output: An approximate trajectory $\widehat{X}$ and the row-wise anchor energies e^{anc}.

Set $q_i \leftarrow \min\{1, \eta\|a_i\|_2^2/\alpha^2\}$ for each $i \in [n]$.
Partition $\{0, \dots, T-1\}$ into blocks $I_\ell = [s_\ell, e_\ell)$ of length at most B.
Set $q_{\mathrm{val}} \leftarrow \lceil C \log(1/\delta_0) \rceil$.
repeat
  Set $\widehat{x}_0 \leftarrow x_0$ and $e^{\mathrm{anc}} \leftarrow 0 \in \mathbb{R}^n$.
  for each block $I_\ell = [s, e)$ **do**
    Compute $r_\ell \leftarrow A\widehat{x}_s - b$ and $g_\ell^{\mathrm{anc}} \leftarrow A^\top r_\ell$.
    Set $e_i^{\mathrm{anc}} \leftarrow e_i^{\mathrm{anc}} + (r_\ell)_i^2$ for each $i \in [n]$.
    Generate independent Bernoulli sets $\mathcal{J}_t$ with row probabilities $\{q_i\}_{i=1}^n$ for each $t \in I_\ell$.
    for $t = s, \dots, e-1$ **do**
      Set $v_t \leftarrow \widehat{x}_t - \widehat{x}_s$ and $h_t \leftarrow \sum_{i \in \mathcal{J}_t} q_i^{-1} a_i (a_i^\top v_t)$.
      Set $\widehat{x}_{t+1} \leftarrow \widehat{x}_t - \eta(g_\ell^{\mathrm{anc}} + h_t)$.
    end for
  end for
  Compute $r_T \leftarrow A\widehat{x}_T - b$ and set $e_i^{\mathrm{anc}} \leftarrow e_i^{\mathrm{anc}} + (r_T)_i^2$ for each $i \in [n]$.
  Set $E_A \leftarrow \sum_i e_i^{\mathrm{anc}}$ and $U \leftarrow \eta^{-1} \max_{t<T} \|v_t\|_2^2$.
  Set $V \leftarrow [v_0, \dots, v_{T-1}]$ and draw independent $g_h \sim N(0, I_T)$ for $h \in [q_{\mathrm{val}}]$.
  Set $\widehat{D} \leftarrow q_{\mathrm{val}}^{-1} \sum_{h=1}^{q_{\mathrm{val}}} \|AVg_h\|_2^2$.
until $E_A \le C(1 + T/B)F_0$, $U \le CBF_0$, and $\widehat{D} \le CBF_0$
return $(\widehat{X} = [\widehat{x}_0, \dots, \widehat{x}_T], e^{\mathrm{anc}})$.

In the following, $C > 0$ denotes a sufficiently large absolute constant. For $t \in I_\ell = [s, e)$, we use $v_t = \widehat{x}_t - \widehat{x}_s$ and $\chi_{t,i} = \mathbf{1}\{i \in \mathcal{J}_t\}$. Since $\nabla F(\widehat{x}_t) = g_\ell^{\mathrm{anc}} + Hv_t$, the difference between the sampled and exact corrections is

$$\xi_t = h_t - Hv_t = \sum_{i: q_i > 0} \left(\frac{\chi_{t,i}}{q_i} - 1 \right) a_i (a_i^\top v_t).$$

Thus, the approximate trajectory follows GD with an additive gradient error

$$\widehat{x}_{t+1} = \widehat{x}_t - \eta\big(\nabla F(\widehat{x}_t) + \xi_t\big). \tag{4}$$

For the analysis, let $\mathcal{F}_t$ contain all randomness revealed before the samples for step t. We reveal the Bernoulli samples only when their corresponding step is reached. Thus, v_t is $\mathcal{F}_t$-measurable, while the current indicators remain independent of $\mathcal{F}_t$.

The accumulated gradient error depends on the total within-block prediction displacement. For a single trajectory trial before conditioning on acceptance, we write $D = \sum_{t<T} \|v_t\|_H^2$ for the total prediction displacement and $\nu = \eta \sum_{t<T} \|\xi_t\|_2^2$ for the accumulated squared gradient error, where $\|v\|_H^2 = v^\top H v = \|Av\|_2^2$. We use the potential $\Phi_t = 2(F(\widehat{x}_t) - F_\star)$ to relate these quantities to the loss decrease across blocks. In particular, $\Phi_0 \le 2F_0$.

Lemma B.2. *The sampled corrections are conditionally unbiased, and their variance is bounded by the within-block prediction displacement. For every $t < T$, $\mathbb{E}[\xi_t \mid \mathcal{F}_t] = 0$, and*

$$\eta\, \mathbb{E}[\|\xi_t\|_2^2 \mid \mathcal{F}_t] \le \alpha^2 \|v_t\|_H^2. \tag{5}$$

If $4\alpha^2 B \le 1/2$, the total displacement satisfies $\mathbb{E}D \le 4B\Phi_0 \le 8BF_0$. The accumulated gradient error satisfies $\mathbb{E}\nu \le 8\alpha^2 BF_0$.

Proof. Conditional on $\mathcal{F}_t$, the row contributions to ξ_t are independent and have mean zero. Therefore, $\mathbb{E}[\xi_t \mid \mathcal{F}_t] = 0$, and their variances sum to $\mathbb{E}[\|\xi_t\|_2^2 \mid \mathcal{F}_t] = \sum_{i:q_i>0}(q_i^{-1} - 1)\|a_i\|_2^2 (a_i^\top v_t)^2$. Terms with $q_i = 1$ vanish. For $0 < q_i < 1$, the sampling rule gives $\|a_i\|_2^2/q_i = \alpha^2/\eta$, so this sum is at most

$(\alpha^2/\eta)\|v_t\|_H^2$. This proves Equation (5). Taking expectations and summing over the updates gives $\mathbb{E}\nu \le \alpha^2 \mathbb{E}D$.

To bound D, we compare the displacement within each block with the decrease in Φ_t. Since $AH^\dagger A^\top$ projects onto $\mathrm{col}(A)$ and the orthogonal residual has squared norm $2F_\star$, the potential also satisfies $\Phi_t = \|\nabla F(\widehat{x}_t)\|_{H^\dagger}^2$. The step-size condition gives $0 \preceq P \preceq I$ for $P = I - \eta H$. For a block $I_\ell = [s, e)$, we denote by $\bar{x}_t$ the exact GD iterates starting from $\bar{x}_s = \widehat{x}_s$. Their displacements satisfy $v_t^\circ = \bar{x}_t - \widehat{x}_s = -H^\dagger(I - P^{t-s}) g_\ell^{\mathrm{anc}}$, and their endpoint potential is $\Phi_e^\circ = \|P^{e-s} g_\ell^{\mathrm{anc}}\|_{H^\dagger}^2$. For an eigenvalue $h > 0$ of H, the corresponding eigenvalue $p = 1 - \eta h$ of P lies in $[0, 1]$. The inequality $(1 - p^{t-s})^2 \le 1 - p^{2(e-s)}$ bounds the squared displacement in this direction by its contribution to the potential decrease. Zero eigenvalues contribute nothing because $g_\ell^{\mathrm{anc}} \in \mathrm{range}(H)$. Summing over the eigendirections and the at most B updates gives

$$\sum_{t=s}^{e-1} \|v_t^\circ\|_H^2 \le B(\Phi_s - \Phi_e^\circ).$$

The sampled updates introduce an additional displacement $\Delta_t = \widehat{x}_t - \bar{x}_t$, which satisfies $\Delta_{t+1} = P\Delta_t - \eta\xi_t$ and $\Delta_s = 0$. We write $\nu_\ell = \eta \sum_{j=s}^{e-1} \|\xi_j\|_2^2$ for the accumulated gradient error in this block. The conditional mean-zero property removes cross terms between errors at different steps after taking expectations. Using $\eta H \preceq I$ and $0 \preceq P \preceq I$, the additional prediction displacement is bounded by

$$\mathbb{E}\|\Delta_t\|_H^2 = \eta^2 \sum_{j=s}^{t-1} \mathbb{E}\left[\xi_j^\top P^{t-1-j} H P^{t-1-j} \xi_j\right] \le \eta \sum_{j=s}^{t-1} \mathbb{E}\|\xi_j\|_2^2 \le \mathbb{E}\nu_\ell,$$

for every $s \le t \le e$.

Combining $v_t = v_t^\circ + \Delta_t$ with the bounds above gives the expected displacement in this block by $2B(\mathbb{E}\Phi_s - \mathbb{E}\Phi_e^\circ) + 2B\mathbb{E}\nu_\ell$. To sum these bounds across consecutive blocks, we replace Φ_e° by the potential at the computed endpoint. The endpoint gradient is $\nabla F(\widehat{x}_e) = P^{e-s} g_\ell^{\mathrm{anc}} + H\Delta_e$. Since $\mathbb{E}[\Delta_e \mid \mathcal{F}_s] = 0$, the cross term in its potential has mean zero, giving $\mathbb{E}\Phi_e = \mathbb{E}\Phi_e^\circ + \mathbb{E}\|\Delta_e\|_H^2 \le \mathbb{E}\Phi_e^\circ + \mathbb{E}\nu_\ell$. Substituting this inequality gives the block displacement bound in terms of the computed endpoints

$$\sum_{t=s}^{e-1} \mathbb{E}\|v_t\|_H^2 \le 2B(\mathbb{E}\Phi_s - \mathbb{E}\Phi_e) + 4B\mathbb{E}\nu_\ell.$$

The endpoint of each block is the next anchor, so the potential differences telescope. Summing over blocks and using $\Phi_T \ge 0$ gives $\mathbb{E}D \le 2B\Phi_0 + 4B\mathbb{E}\nu$. Combining this with $\mathbb{E}\nu \le \alpha^2 \mathbb{E}D$ yields $(1 - 4\alpha^2 B)\mathbb{E}D \le 2B\Phi_0$. Therefore, the condition $4\alpha^2 B \le 1/2$ gives $\mathbb{E}D \le 4B\Phi_0 \le 8BF_0$. Substituting this bound into $\mathbb{E}\nu \le \alpha^2 \mathbb{E}D$ proves the accumulated gradient-error bound. □

The preceding lemma gives $\mathbb{E}D = O(BF_0)$. To obtain a high-probability prediction-error bound, we analyze the tests used to accept a trajectory. The anchor-energy test also limits the contribution of anchor residuals to candidate sampling. Its tested quantity is the sum of squared residual norms at the block anchors and the final iterate

$$E_A = \sum_{\ell=0}^{J-1} \|A\widehat{x}_{s_\ell} - b\|_2^2 + \|A\widehat{x}_T - b\|_2^2.$$

The final term accounts for $\widehat{x}_T$, which is not included among the block anchors. The other two tests evaluate $U = \eta^{-1} \max_{t<T} \|v_t\|_2^2$ directly and estimate $D = \|AV\|_F^2$ using Gaussian probes independent of the trajectory. The following lemma bounds the estimation error and shows that each trial is accepted with constant probability.

Lemma B.3. *The Gaussian test estimates the total prediction-displacement energy within a constant factor, and each trajectory trial is accepted with constant probability. Conditional on a completed trajectory, $D/2 \le \widehat{D} \le 2D$ holds with probability at least $1 - C\delta_0$. Moreover, each trial is accepted with probability at least an absolute constant $p_0 > 0$, so the expected number of trials is $O(1)$.*

Proof. We first bound the error of the Gaussian estimate. Conditional on V, denote the eigenvalues of $V^\top HV$ by $a_j \geq 0$. Each probe satisfies $\|AVg_h\|_2^2 = \sum_j a_j Z_{h,j}^2$, where the $Z_{h,j}$ are independent standard Gaussian variables and $\sum_j a_j = D$. If $D = 0$, then $\widehat{D} = 0$. For $D > 0$, the moment-generating function of a centered probe satisfies

$$\log \mathbb{E} \exp\left(u \sum_j a_j (Z_{h,j}^2 - 1)\right) = \sum_j \left(-ua_j - \frac{1}{2}\log(1 - 2ua_j)\right) \leq Cu^2 \sum_j a_j^2$$

for $|u| \leq (4 \max_j a_j)^{-1}$. Since $\sum_j a_j^2 \leq D^2$ and $\max_j a_j \leq D$, averaging the q_{val} independent probes and applying the exponential Markov inequality gives $\Pr[\widehat{D} \notin [D/2, 2D] \mid V] \leq 2\exp(-cq_{\mathrm{val}})$. The choice $q_{\mathrm{val}} = \lceil C \log(1/\delta_0) \rceil$ makes this probability at most $C\delta_0$.

We next show that a trajectory satisfying moderate displacement and gradient error passes all three tests. Expanding the quadratic loss in Equation (4) and using $\eta H \preceq I$ gives

$$F(\widehat{x}_{t+1}) \leq F(\widehat{x}_t) - \frac{\eta}{2}\|\nabla F(\widehat{x}_t)\|_2^2 + \frac{\eta}{2}\|\xi_t\|_2^2.$$

On the event $\nu \leq F_0$, every iterate therefore satisfies $F(\widehat{x}_t) \leq 3F_0/2$. Hence, $E_A \leq 3(J+1)F_0 \leq 6(1 + T/B)F_0$. Summing the same descent inequality also gives $\eta \sum_{t<T} \|\nabla F(\widehat{x}_t)\|_2^2 \leq 2F_0 + \nu$. The total squared update norm is consequently bounded by

$$\eta \sum_{t<T} \|\nabla F(\widehat{x}_t) + \xi_t\|_2^2 \leq 4F_0 + 4\nu.$$

For t in a block anchored at s, Cauchy–Schwarz applied to $v_t = -\eta \sum_{j=s}^{t-1} (\nabla F(\widehat{x}_j) + \xi_j)$ gives $U \leq B(4F_0 + 4\nu) \leq 8BF_0$.

It remains to show that these conditions occur with constant probability. By Lemma B.2 and Markov's inequality,

$$\Pr[\nu > F_0] \leq 8\alpha^2 B, \qquad \Pr[D > (C/2)BF_0] \leq 16/C.$$

By choosing the constant in α sufficiently small and C sufficiently large, the events $\nu \leq F_0$ and $D \leq (C/2)BF_0$ hold simultaneously with constant probability. On this event, the preceding bounds ensure that the first two acceptance tests pass. Moreover, whenever $D/2 \leq \widehat{D} \leq 2D$, the third test also passes. Since the Gaussian validation fails with probability at most $C\delta_0$, choosing the internal constant in $\delta_0 = c\delta$ sufficiently small gives an absolute constant $p_0 > 0$ such that every trial is accepted with probability at least p_0. Since different trials use fresh randomness, the expected number of trials is at most $1/p_0 = O(1)$. □

The bounds $D = O(BF_0)$ and $U = O(BF_0)$ provide the two ingredients needed for concentration. The bound on D controls the predictable variance of the accumulated prediction error, while the bound on U controls each row increment. The next lemma uses these bounds to establish the uniform prediction-error guarantee stated in the main text and records the energy bounds needed for the subsequent row-sampling analysis.

Lemma B.4. *With probability at least $1 - C\delta_0$, the trajectory returned by* GENERATEBLOCKTRAJECTORY *satisfies the uniform prediction-error bound $\max_{0 \leq t \leq T} \|A(\widehat{x}_t - x_t)\|_2 \leq C\frac{\epsilon_0\sqrt{G}}{R}$. On the same event, every $\lambda \in \Lambda_R$ satisfies the relative loss bound*

$$(1 - C\epsilon_0)F(X\lambda) \leq F(\widehat{X}\lambda) \leq (1 + C\epsilon_0)F(X\lambda).$$

Moreover, the returned trajectory also satisfies $E_A \leq C(1+T/B)F_0$, $U \leq CBF_0$, and $D \leq 2CBF_0$ on this event.

Proof. We first bound the prediction error in a single trial. With $e_t = \widehat{x}_t - x_t$ and $P = I - \eta H$, subtracting the exact GD update from the approximate update gives $e_{t+1} = Pe_t - \eta\xi_t$, with $e_0 = 0$. For a fixed time t, expanding this recurrence expresses the prediction error as a sum of propagated gradient errors

$$H^{1/2} e_t = -\eta \sum_{s<t} H^{1/2} P^{t-1-s} \xi_s.$$

To bound the accumulated variance and each row contribution, we truncate this sum. Before revealing the samples at step s, we check whether $\sum_{j\le s}\|v_j\|_H^2 > 2CBF_0$ or $\|v_s\|_2^2/\eta > CBF_0$. If either bound is exceeded, the contributions from step s onward are set to zero. This change is used only in the proof and does not alter the algorithm. Since v_s is determined by earlier updates, the decision does not depend on the samples at step s. Therefore, the retained row contributions still have conditional mean zero when the row indicators are revealed one at a time. We denote the truncated sum by M_t.

Since $\|\eta H^{1/2}P^{t-1-s}\|_2^2 \le \eta$, Equation (5) bounds the sum of conditional row variances at each included step by $\alpha^2\|v_s\|_H^2$. Therefore, the total conditional variance is at most $2C\alpha^2BF_0$. For $0 < q_i < 1$, we have

$$\left\|\eta H^{1/2}P^{t-1-s}a_i\left(\frac{\chi_{s,i}}{q_i}-1\right)a_i^\top v_s\right\|_2 \le \frac{\eta\sqrt{\|H\|_2}\|a_i\|_2^2\|v_s\|_2}{q_i} \le \frac{\alpha^2\|v_s\|_2}{\sqrt{\eta}} \le C\alpha^2\sqrt{BF_0}.$$

If $q_i = 1$, row i is always sampled and its correction is exact. If $q_i = 0$, then $a_i = 0$, so the row has no contribution to Hv_s. Applying the Hilbert-space Freedman inequality (Peng et al., 2024, Theorem 5.1) with these variance and increment bounds gives, with probability at least $1 - C\delta_0/T$,

$$\|M_t\|_2 \le C\left(\alpha\sqrt{BF_0\log(T/\delta_0)} + \alpha^2\sqrt{BF_0}\log(T/\delta_0)\right).$$

Substituting the choice of α and $F_0 = \rho G$ bounds the right-hand side by $C\epsilon_0\sqrt{G}/R$. Thus, a union bound over $t \in [T]$ gives $\max_{t\in[T]}\|M_t\|_2 \le C\epsilon_0\sqrt{G}/R$ with probability at least $1 - C\delta_0$.

We now apply the concentration bound to the returned trajectory. If a trial is accepted and $D/2 \le \widehat{D} \le 2D$, then the acceptance conditions imply $D \le 2CBF_0$ and $U \le CBF_0$. Hence, the truncation used in the concentration argument is never triggered, and the bound for M_t applies to the actual prediction error. Since $\|Ae_t\|_2 = \|H^{1/2}e_t\|_2$, every such trial satisfies

$$\max_{0\le t\le T}\|A(\widehat{x}_t - x_t)\|_2 \le C\frac{\epsilon_0\sqrt{G}}{R}$$

except with probability at most $C\delta_0$.

For a single trial, the probability that it is accepted while either the Gaussian validation or the prediction-error bound fails is at most $C\delta_0$. By Lemma B.3, each trial is accepted with probability at least $p_0 > 0$. Since different trials use fresh randomness, the first accepted trial violates these bounds with probability at most $C\delta_0/p_0 = O(\delta_0)$. The acceptance conditions also give $E_A \le C(1+T/B)F_0$ and $U \le CBF_0$, while correct Gaussian validation gives $D \le 2CBF_0$. □

B.2 Candidate Row Reduction

Although GenerateBlockTrajectory reduces the number of exact full-gradient evaluations, constructing the residual basis directly from the $T+1$ approximate iterates would still require evaluating their residuals on all n rows. To reduce the rows before constructing this basis, we use the blockwise residual decomposition to estimate the contribution of each row without forming all $T+1$ residual vectors. For each block, the anchor scan provides the row-wise anchor energies e_i^{anc} exactly, and the remaining contribution comes from the corrections Av_t. TrajectoryCandidate estimates the row-wise correction energy $\sum_{t<T}(a_i^\top v_t)^2$ using a constant number of Gaussian probes per block and combines it with e_i^{anc} to form the sampling scores. Then, it draws M original rows according to these scores and uses the sampled rows for the subsequent residual-basis construction. Repeated rows are merged by adding their objective weights, which preserves $S_0^\top S_0$. In the following, we condition on a trajectory satisfying Lemma B.4 and use fresh randomness independent of the trajectory construction.

Residual representation. To obtain row scores that are valid uniformly over all $\lambda \in \Lambda_R$, we use the blockwise residual representation. For every $t \in I_\ell = [s_\ell, e_\ell)$, the approximate residual satisfies $\widehat{r}_t = \widehat{r}_{s_\ell} + Av_t$, where $\widehat{r}_t = A\widehat{x}_t - b$. Hence, every feasible residual is generated by the block-anchor residuals and the within-block corrections.

We collect these vectors in $R_A = [\widehat{r}_{s_0}, \dots, \widehat{r}_{s_{J-1}}, \widehat{r}_T]$ and $Z = [Av_0, \dots, Av_{T-1}]$, and set $\mathcal{C} = [R_A, Z]$. For $\lambda \in \Lambda_R$, define $\beta_\ell(\lambda) = \sum_{t\in I_\ell}\lambda_t$ and

$$\theta(\lambda) = \big(\beta_0(\lambda), \dots, \beta_{J-1}(\lambda), \lambda_T, \lambda_0, \dots, \lambda_{T-1}\big)^\top.$$

Algorithm 4 TRAJECTORYCANDIDATE

Input: $A, b, \widehat{X}, e^{\mathrm{anc}}, B$, and M.
Output: A weighted candidate row sampler S_0.

Initialize $S_0 \leftarrow \emptyset$.
if $M \geq n$ **then**
  return I_n
end if
Set $T \leftarrow \mathrm{cols}(\widehat{X}) - 1$ and partition $\{0, \ldots, T-1\}$ into blocks $I_\ell = [s_\ell, e_\ell)$ of length at most B.
Set $\widehat{e}_i \leftarrow e_i^{\mathrm{anc}}$ and $q_0 \leftarrow 48$.
for each block $I_\ell = [s_\ell, e_\ell)$ **do**
  Draw independent $\omega_{\ell,h} \sim N(0, I_{e_\ell - s_\ell})$ for $h \in [q_0]$ and form $w_{\ell,h} = \sum_{t=s_\ell}^{e_\ell - 1} (\omega_{\ell,h})_{t-s_\ell+1} (\widehat{x}_t - \widehat{x}_{s_\ell})$.
  Compute $z_{\ell,h} = A w_{\ell,h}$ and update $\widehat{e}_i \leftarrow \widehat{e}_i + q_0^{-1} \sum_{h=1}^{q_0} (z_{\ell,h})_i^2$.
end for
Set $p_i \leftarrow \widehat{e}_i / \sum_j \widehat{e}_j$ and draw M independent indices from p.
Let N_i be the number of times row i is sampled.
For each i with $N_i > 0$, include row i in S_0 with scaling $\sqrt{N_i/(Mp_i)}$.
return S_0.

The residual decomposition gives

$$A\widehat{X}\lambda - b = \mathcal{C}\theta(\lambda). \tag{6}$$

Moreover, $\sum_\ell |\beta_\ell(\lambda)| + |\lambda_T| \leq R$ and $\sum_{t<T} |\lambda_t| \leq R$, so $\|\theta(\lambda)\|_2^2 \leq 2R^2$. This yields a common importance score for every feasible residual. For row i, define $e_i = \|\mathcal{C}_{i,:}\|_2^2$. Then, for every $\lambda \in \Lambda_R$,

$$\left(a_i^\top \widehat{X}\lambda - b_i\right)^2 = \left(\mathcal{C}_{i,:}\theta(\lambda)\right)^2 \leq 2R^2 e_i.$$

Thus, e_i controls the contribution of row i uniformly over the entire stable family. The total mass of these row scores is also bounded by the trajectory estimates. Since $Av_t = \widehat{r}_t - \widehat{r}_{s_\ell}$, all columns of $\mathcal{C}$ lie in $\mathrm{span}\{\widehat{r}_0, \ldots, \widehat{r}_T\}$, and hence $\mathrm{rank}(\mathcal{C}) \leq T + 1$. Furthermore, we have

$$E := \sum_i e_i = \|\mathcal{C}\|_F^2 = E_A + D \leq C\left(1 + \frac{T}{B} + B\right) F_0. \tag{7}$$

Computing the exact scores e_i would require forming all corrections Av_t. Here, the anchor part is available in e_i^{anc}, while the Gaussian probes estimate the correction part. Therefore, the algorithm uses

$$\widehat{e}_i = e_i^{\mathrm{anc}} + \sum_{\ell=0}^{J-1} \frac{1}{q_0} \sum_{h=1}^{q_0} (a_i^\top w_{\ell,h})^2$$

and samples with probabilities $p_i = \widehat{e}_i / \sum_j \widehat{e}_j$. Since the first anchor contributes $\|Ax_0 - b\|_2^2 = 2F_0 > 0$, these probabilities are well defined. Rows with $e_i = 0$ contribute nothing to any feasible residual and can be omitted from the analysis.

Lemma B.5. *The Gaussian score estimates have bounded inverse moments relative to the exact row energies. With $q_0 = 48$, every row with $e_i > 0$ and every integer $1 \leq k \leq 11$ satisfy*

$$\mathbb{E} p_i^{-k} \leq C\left(\frac{E}{e_i}\right)^k,$$

where the expectation is over the Gaussian probes.

Proof. For a fixed row i and block I_ℓ, the exact correction energy is $e_{i,\ell}^Z = \sum_{t \in I_\ell} (a_i^\top v_t)^2$. The Gaussian estimate of this quantity has the form $e_{i,\ell}^Z Y_\ell$, where $Y_\ell \sim \chi_{q_0}^2 / q_0$. Hence, with $R_i = \widehat{e}_i / e_i$, the variable R_i is a convex combination of 1 and the variables $\{Y_\ell\}$. Since $q_0 = 48$, both positive and negative moments of Y_ℓ up to order 22 are bounded by absolute constants. By Jensen's inequality, we have $\mathbb{E} R_i^s \leq C$ and $\mathbb{E} R_i^{-s} \leq C$ for every $1 \leq s \leq 22$.

The normalized total score also satisfies $\widehat{E}/E = \sum_i (e_i/E) R_i$, which is a convex combination of the R_i. Thus, $\mathbb{E}(\widehat{E}/E)^s \leq C$ for the same range of s. Since $p_i = \widehat{e}_i / \widehat{E}$, applying Cauchy–Schwarz

inequality gives

$$\begin{aligned}\mathbb{E}p_i^{-k} &= \left(\frac{E}{e_i}\right)^k \mathbb{E}\left[\left(\frac{\widehat{E}}{E}\right)^k R_i^{-k}\right] \\ &\le \left(\frac{E}{e_i}\right)^k \left[\mathbb{E}\left(\frac{\widehat{E}}{E}\right)^{2k}\right]^{1/2} \left[\mathbb{E}R_i^{-2k}\right]^{1/2} \\ &\le C\left(\frac{E}{e_i}\right)^k,\end{aligned}$$

where the last step follows from $2k \le 22$. □

The candidate rows should preserve the squared norm of every residual represented by $\mathcal{C}$. Let $\Gamma = \mathcal{C}^\top\mathcal{C}$ and $\Gamma_0 = \mathcal{C}^\top S_0^\top S_0\mathcal{C}$ denote the original and sampled Gram matrices. We measure their difference relative to the ridge-regularized matrix $\Gamma + \zeta I$, where $\zeta = C\epsilon_0 G/R^2$. For every feasible coefficient vector, the ridge contributes only $\zeta\|\theta(\lambda)\|_2^2 \le 2\zeta R^2 = O(\epsilon_0 G)$.

To analyze the sampled Gram matrix, define $u_i = (\Gamma + \zeta I)^{-1/2}\mathcal{C}_{i,:}^\top$ and $\ell_i = \|u_i\|_2^2$. Since $\Gamma + \zeta I \succeq \zeta I$, we have $\ell_i \le e_i/\zeta$, while $\sum_i \ell_i \le \operatorname{rank}(\mathcal{C}) \le T+1$. Moreover, $\sum_i u_i u_i^\top \preceq I$.

The sampling probabilities are random because they are obtained from Gaussian score estimates. The inverse-moment bound above allows us to control the sampled Gram error without requiring each estimated score to approximate e_i individually. The next lemma gives the resulting ridge approximation.

Lemma B.6. *The candidate sample has a bounded ridge-normalized Gram error. For the sampling branch $M < n$, define $\Delta_0 = (\Gamma + \zeta I)^{-1/2}(\Gamma_0 - \Gamma)(\Gamma + \zeta I)^{-1/2}$. With $q_0 = 48$, the expectation over both the probes and the sampled rows satisfies*

$$\left(\mathbb{E}\|\Delta_0\|_2^{12}\right)^{1/12} \le C(T+1)^{1/12}\left[\sqrt{\frac{E}{\zeta M}} + \left(\frac{E}{\zeta M}\right)^{11/12}\right].$$

If $M \ge n$, then $\Delta_0 = 0$.

Proof. We condition first on the Gaussian probes, so the probabilities p_i are fixed. We use $\mathbb{E}_s$ for expectation over the subsequent independent row draws. If i_j denotes the index selected in draw j, then we have $\Delta_0 = \frac{1}{M}\sum_{j=1}^M Z_j$ and $Z_j = \frac{u_{i_j}u_{i_j}^\top}{p_{i_j}} - K_0$, where $K_0 = \sum_i u_i u_i^\top \preceq I$. Conditional on the probes, the matrices Z_j are independent, self-adjoint, and have mean zero.

We first control their conditional variance. For one sampled index, write $B = u_i u_i^\top/p_i$. Since $\ell_i = \|u_i\|_2^2$, we get

$$\mathbb{E}_s Z_1^2 \preceq W = \sum_i \frac{\ell_i}{p_i} u_i u_i^\top.$$

To bound the sixth trace moment of W, denote by $\mathcal{U}$ the matrix with rows $u_i^\top$. Then $\mathcal{U}^\top\mathcal{U} = K_0 \preceq I$ and $W = \mathcal{U}^\top D\mathcal{U}$ for $D = \operatorname{diag}(\ell_i/p_i)$. Convexity of x^6 applied in an eigenbasis of W gives $\operatorname{tr}(W^6) \le \sum_i \ell_i^7 p_i^{-6}$. Taking expectation over the probes and using Theorem B.5 with $k = 6$ yields

$$\mathbb{E}\operatorname{tr}(W^6) \le CE^6\sum_i \frac{\ell_i^7}{e_i^6} \le C(E/\zeta)^6\sum_i \ell_i \le C(E/\zeta)^6(T+1).$$

We also need to control the size of one sampled matrix. For a matrix X, we use the Schatten-12 norm $\|X\|_{S_{12}} = (\operatorname{tr}|X|^{12})^{1/12}$. Since B is rank one, $\|B\|_{S_{12}} = \ell_i/p_i$. Centering changes the twelfth moment only by an absolute constant, and Theorem B.5 with $k = 11$ gives

$$\begin{aligned}\mathbb{E}\mathbb{E}_s\|Z_1\|_{S_{12}}^{12} &\le C\sum_i \ell_i^{12}\mathbb{E}p_i^{-11} \\ &\le CE^{11}\sum_i \frac{\ell_i^{12}}{e_i^{11}} \le C(E/\zeta)^{11}(T+1).\end{aligned}$$

The fixed-order noncommutative Rosenthal inequality (Junge & Zeng, 2013, Theorem 0.4), applied after conditioning on the probes, gives

$$\left(\mathbb{E}_{\mathrm{s}}\|\Delta_0\|_{S_{12}}^{12}\right)^{1/12} \le CM^{-1/2}\,\mathrm{tr}(W^6)^{1/12} + CM^{-11/12}\left(\mathbb{E}_{\mathrm{s}}\|Z_1\|_{S_{12}}^{12}\right)^{1/12}.$$

Taking the twelfth moment over the Gaussian probes and substituting the two bounds above yields

$$\left(\mathbb{E}\|\Delta_0\|_{S_{12}}^{12}\right)^{1/12} \le C(T+1)^{1/12}\left[\sqrt{\frac{E}{\zeta M}} + \left(\frac{E}{\zeta M}\right)^{11/12}\right].$$

Finally, $\|\Delta_0\|_2 \le \|\Delta_0\|_{S_{12}}$, which proves the claim. □

Corollary B.7. *The candidate rows preserve the residual Gram matrix up to a small relative error after adding the ridge. For sufficiently large C, with conditional probability at least $1 - C\delta_0$,*

$$-\gamma_0(\Gamma + \zeta I) \preceq \Gamma_0 - \Gamma \preceq \gamma_0(\Gamma + \zeta I),$$

where $\gamma_0 = O(\sqrt{\epsilon_0})$.

Proof. If $M \ge n$, then all rows are retained and $\Gamma_0 = \Gamma$. Suppose that $M < n$. The energy bound in Equation (7) and the definition of ζ give

$$\frac{E}{\zeta} \le C\rho R^2 \epsilon_0^{-1}\left(1 + \frac{T}{B} + B\right).$$

Using the chosen value of M, we obtain $\frac{E}{\zeta M} \le C\epsilon_0\delta_0^{1/6}T^{-1/6}$. In particular, $E/(\zeta M) \le 1$. Applying Theorem B.6 and using $T + 1 \le 2T$ gives

$$\left(\mathbb{E}\|\Delta_0\|_2^{12}\right)^{1/12} \le C\sqrt{\epsilon_0}\,\delta_0^{1/12},$$

where $\Delta_0 = (\Gamma + \zeta I)^{-1/2}(\Gamma_0 - \Gamma)(\Gamma + \zeta I)^{-1/2}$. Using Markov's inequality gives $\|\Delta_0\|_2 \le \gamma_0$ with probability at least $1 - C\delta_0$, for $\gamma_0 = O(\sqrt{\epsilon_0})$. This operator bound is equivalent to $-\gamma_0 I \preceq \Delta_0 \preceq \gamma_0 I$. Multiplying on both sides by $(\Gamma + \zeta I)^{1/2}$ proves the claim. □

B.3 Final resparsification

The candidate set from the previous subsection already provides the required ridge approximation, but its row count can be larger than the desired weak coreset size. In the following, we construct the residual basis only on the candidate rows and resample them using leverage scores. Since the candidate residual space has rank at most $T + 1$, an $O(\sqrt{\epsilon})$ relative embedding requires only $O(\frac{T}{\epsilon}\log\frac{T}{\delta})$ sampled rows. The key point is that this weaker distortion is sufficient for optimization over Λ_R. The analysis below first shows that the final sampling preserves the candidate Gram matrix, and then converts the resulting $O(\sqrt{\epsilon})$ distortion into an $O(\epsilon)$ objective error.

Algorithm 5 FinalResparsify

Input: $S_0, A, b, \widehat{X}, T, m$.
Output: A weighted original-row sampler S.

Set $n_0 \leftarrow \mathrm{rows}(S_0)$.
if $n_0 \le m$ **then**
  return S_0
end if
Form $A_0 = S_0 A$ and $b_0 = S_0 b$. If $T < d$, set $Y_0 = A_0\widehat{X} - b_0\mathbf{1}^\top$; otherwise set $Y_0 = [A_0, b_0]$.
Compute $r = \mathrm{rank}(Y_0)$ and an orthonormal basis Q_0 for $\mathrm{col}(Y_0)$.
if $r = 0$ **then**
  return any one weighted row of S_0
end if
Set $\mu_i = \|e_i^\top Q_0\|_2^2/r$, draw m independent indices from μ, and form the sampler S_1 by rescaling each selected row by $(m\mu_i)^{-1/2}$.
return the composed sampler $S = S_1 S_0$.

The final sampling only needs to preserve the residual space represented by $S_0\mathcal{C}$. To compute leverage scores without explicitly forming this matrix, FINALRESPARSIFY constructs a matrix Y_0 whose column space contains all columns of $S_0\mathcal{C}$. When $T < d$, the sampled trajectory residuals already span the sampled anchor residuals and corrections, so we use $Y_0 = A_0\widehat{X} - b_0\mathbf{1}^\top$. When $d \le T$, the same residuals lie in $\mathrm{col}([S_0A, S_0b])$, and we use $Y_0 = [S_0A, S_0b]$. In either case, $\mathrm{col}(S_0\mathcal{C}) \subseteq \mathrm{col}(Y_0)$ and $\mathrm{rank}(Y_0) \le T + 1$. Thus, leverage-score sampling on Y_0 preserves every candidate residual while working in a space of rank at most $T + 1$. Let $\Gamma_1 = \mathcal{C}^\top S^\top S\mathcal{C}$ denote the resulting residual Gram matrix.

Lemma B.8. *The final sampling preserves the candidate residual Gram matrix. With probability at least* $1 - C\delta_0$*,*

$$-\gamma_1\Gamma_0 \preceq \Gamma_1 - \Gamma_0 \preceq \gamma_1\Gamma_0, \tag{8}$$

where $\gamma_1 = O(\sqrt{\epsilon_0})$*. The output contains at most* m *distinct original rows.*

Proof. If $n_0 \le m$, the algorithm returns S_0, so $\Gamma_1 = \Gamma_0$. If $r = 0$, then $S_0\mathcal{C} = 0$ and both Gram matrices vanish. Therefore, we consider $n_0 > m$ and $r > 0$.

The probabilities $\mu_i = \|e_i^\top Q_0\|_2^2/r$ are the normalized leverage scores of the residual space. Since Q_0 has orthonormal columns, the sampled embedding satisfies $\mathbb{E}[Q_0^\top S_1^\top S_1 Q_0] = I_r$. The sampled matrices are positive semidefinite and have operator norm r/m. By the matrix Chernoff bound (Tropp, 2012), we have

$$\Pr\left[\|Q_0^\top S_1^\top S_1 Q_0 - I_r\|_2 > \gamma_1\right] \le 2r \exp\left(-\frac{cm\gamma_1^2}{r}\right).$$

Choosing $m = O(\frac{T}{\epsilon}\log\frac{T}{\delta})$, this probability is at most $C\delta_0$ for $\gamma_1 = O(\sqrt{\epsilon_0})$.

Since $\mathrm{col}(S_0\mathcal{C}) \subseteq \mathrm{col}(Q_0)$, the same embedding holds for every column of $S_0\mathcal{C}$, which gives Equation (8). Finally, at most m indices are sampled, so merging repeated indices leaves at most m distinct original rows. □

Combining the candidate approximation with the final embedding gives

$$\Gamma_1 - \Gamma = (\Gamma_1 - \Gamma_0) + (\Gamma_0 - \Gamma).$$

By Theorem B.7, we also have $\Gamma_0 \preceq (1 + \gamma_0)(\Gamma + \zeta I)$. Therefore, we obtain

$$-\gamma(\Gamma + \zeta I) \preceq \Gamma_1 - \Gamma \preceq \gamma(\Gamma + \zeta I),$$

where $\gamma = \gamma_0 + (1 + \gamma_0)\gamma_1 = O(\sqrt{\epsilon_0})$. This distortion is larger than the target $O(\epsilon_0)$ loss error. The next lemma shows that, after minimizing over the convex coefficient set, the excess loss depends quadratically on γ, yielding an $O(\epsilon_0)$ optimization error.

Proof of Lemma 4.2. We compare an arbitrary sampled minimizer $\widehat{\lambda}$ with a minimizer $\lambda^\star$ of $F(\widehat{X}\lambda)$ over Λ_R. Their residual coefficient vectors are $\widehat{\theta} = \theta(\widehat{\lambda})$ and $\theta^\star = \theta(\lambda^\star)$. With $h = \widehat{\theta} - \theta^\star$ and $\Delta = \Gamma_S - \Gamma$, first-order optimality at $\widehat{\lambda}$ gives $\widehat{\theta}^\top\Gamma_S h \le 0$, since θ is linear and Λ_R is convex. Expanding the difference in original losses and using this inequality gives

$$F(\widehat{X}\widehat{\lambda}) - F(\widehat{X}\lambda^\star) = \widehat{\theta}^\top\Gamma h - \frac{1}{2}h^\top\Gamma h \le -(\theta^\star)^\top\Delta h - h^\top\Delta h - \frac{1}{2}h^\top\Gamma h.$$

We next bound the two terms involving Δ. For $\zeta > 0$, the ridge inequality gives $\|(\Gamma + \zeta I)^{-1/2}\Delta(\Gamma + \zeta I)^{-1/2}\|_2 \le \gamma$. Since $\|\theta^\star\|_2^2 \le 2R^2$ and $\|h\|_2^2 \le 8R^2$, the mixed term is bounded by

$$|(\theta^\star)^\top\Delta h| \le \gamma\sqrt{(\theta^\star)^\top\Gamma\theta^\star + 2\zeta R^2}\sqrt{h^\top\Gamma h + 8\zeta R^2} \le \gamma^2(\theta^\star)^\top\Gamma\theta^\star + \frac{1}{4}h^\top\Gamma h + C\zeta R^2,$$

where the second inequality uses the Young inequality. The first inequality also holds when $\zeta = 0$, by using $\Gamma + \tau I$ for $\tau > 0$ and taking $\tau \downarrow 0$. The quadratic error term satisfies $|h^\top\Delta h| \le \gamma h^\top\Gamma h + 8\gamma\zeta R^2$ directly from the assumed matrix bound.

Substituting these estimates into the preceding optimality bound leaves a nonpositive coefficient on $h^\top \Gamma h$. For a sufficiently small internal accuracy ϵ_0, we have $\gamma \le 1/4$. Therefore, the excess loss satisfies

$$F(\widehat{X}\widehat{\lambda}) - F(\widehat{X}\lambda^\star) \le \gamma^2 (\theta^\star)^\top \Gamma \theta^\star + \left(\gamma - \frac{1}{4}\right) h^\top \Gamma h + C\zeta R^2$$
$$\le 2\gamma^2 F(\widehat{X}\lambda^\star) + C\zeta R^2.$$

Optimality of $\lambda^\star$ proves the claim. □

Applying the lemma to the final sampler, with $\Gamma_S = \Gamma_1$, gives the optimization guarantee on $\widehat{X}$. Since $F(\widehat{X}\lambda) \ge F_\star \ge G$ for every $\lambda \in \Lambda_R$, the choice $\zeta = C\epsilon_0 G/R^2$ bounds the additive term by $C\epsilon_0 \min_{\lambda\in\Lambda_R} F(\widehat{X}\lambda)$. Combining this with $\gamma^2 = O(\epsilon_0)$ shows that every sampled minimizer satisfies

$$F(\widehat{X}\widehat{\lambda}) \le (1 + C\epsilon_0) \min_{\lambda\in\Lambda_R} F(\widehat{X}\lambda). \tag{9}$$

B.4 Omitted Proof of Weak Krylov Coreset

The main approximation result for weak coreset is presented as follows.

Proof of Theorem 4.3. We first prove the loss guarantee for an initial point satisfying $F(x_0) \le \rho F_\star$. If $F_0 = 0$, both trajectories are constant and the returned subset is exact. Hence, assume $F_0 > 0$. By Lemma B.4 and the sampling guarantee in Equation (9), with probability at least $1 - C\delta_0$, every sampled minimizer $\widehat{\lambda}$ satisfies

$$F(\widehat{X}\widehat{\lambda}) \le (1 + C\epsilon_0) \min_{\lambda\in\Lambda_R} F(\widehat{X}\lambda).$$

Let $\lambda^\star \in \arg\min_{\lambda\in\Lambda_R} F(X\lambda)$. The uniform trajectory comparison then gives

$$\begin{aligned} F(\widehat{X}\widehat{\lambda}) &\le (1 + C\epsilon_0) F(\widehat{X}\lambda^\star) \\ &\le (1 + C\epsilon_0)^2 F(X\lambda^\star) \le (1 + C'\epsilon_0) F^\star_{T,R}. \end{aligned} \tag{10}$$

Applying the reverse trajectory comparison to the same $\widehat{\lambda}$ gives $F(X\widehat{\lambda}) \le F(\widehat{X}\widehat{\lambda})/(1 - C\epsilon_0)$. Thus, choosing $\epsilon_0 = C\epsilon$ and $\delta_0 = C\delta$ with sufficiently small C proves the two loss bounds with probability at least $1 - \delta$. By Theorem B.8, the output contains at most $m = O(T\epsilon^{-1}\log(T/\delta))$ distinct original rows.

For the construction time, the assumed lower bound on n ensures that the sampled corrections within each block are covered by the $O(nd)$ anchor scan. Hence, generating one trajectory trial costs $O((T/B + \log(1/\delta))nd)$ time, including the Gaussian tests. Since each trial is accepted with constant probability by Lemma B.3, repeating until acceptance preserves this bound in expectation.

Moreover, the candidate probes require another $O((T/B)nd)$ time, and constructing the residual basis and leverage scores on the candidate rows costs $O(MdT)$. Substituting the candidate size gives

$$O(MdT) = O\left(\frac{\rho R^2 T^{13/6} d}{B\epsilon^2 \delta^{1/6}}\right).$$

The constant-factor initialization, amplified to failure probability $O(\delta)$, adds $O((nd + d^\omega)\log(1/\delta))$ time. Summing these terms and absorbing the constant ρ gives the stated expected construction time. □

C Additional Experimental Results

This section provides additional experiments complementing the main results. We first report comparisons with additional local coreset variants, and then present further results for the Strong coreset and the effect of the stable radius R.

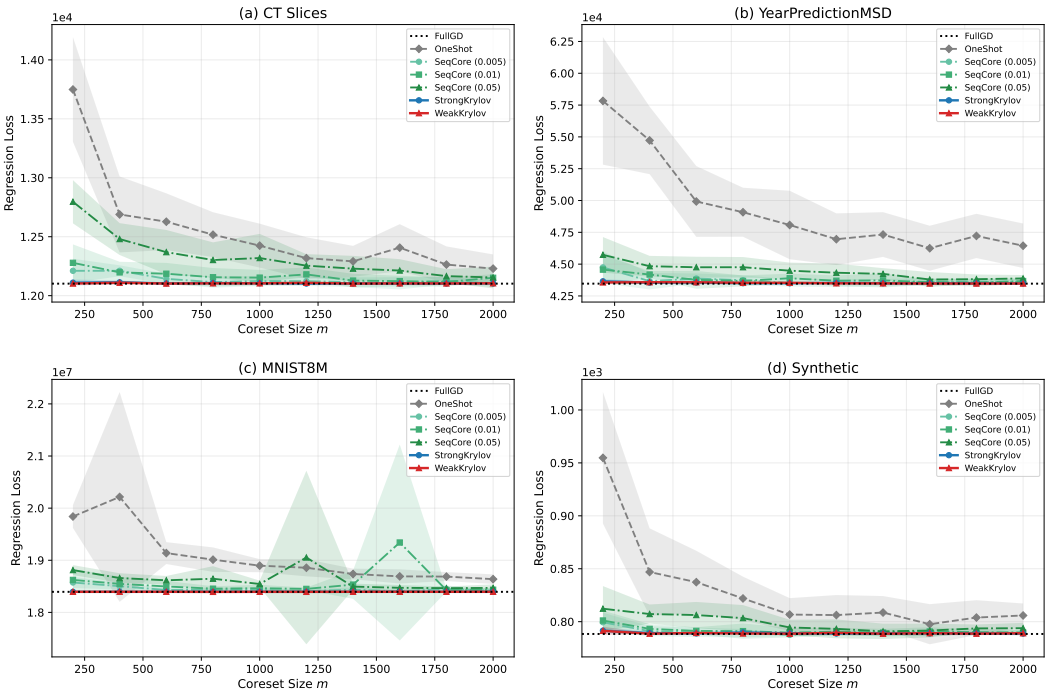

Figure 2: Comparison of regression loss for OneShot, SeqCore with $r \in \{0.005, 0.01, 0.05\}$, StrongKrylov, and WeakKrylov algorithms on CT Slices, YearPredictionMSD, MNIST8M, and Synthetic with varying coreset size m.

OneShot and SeqCore sensitivity. Following the settings in the main experiments, we use the same $T = 32$ and coreset sizes $m \in \{200, 400, \ldots, 2000\}$ and compare StrongKrylov and WeakKrylov with OneShot and SeqCore on all four datasets. OneShot is the no-rebuild variant of SeqCore: it constructs a local coreset once at the initial point and keeps it fixed throughout the iterations. For SeqCore, we vary the radius over $r \in \{0.005, 0.01, 0.05\}$.

As shown in Figures 2 and 3, OneShot has the lowest runtime in most cases, but yields higher regression loss. For SeqCore, a smaller r reduces the loss but increases runtime due to more frequent rebuilding, while a larger r reduces runtime but increases the loss. StrongKrylov and WeakKrylov give losses close to FullGD, and WeakKrylov is faster than StrongKrylov, especially on YearPredictionMSD and MNIST8M.

Additional MNIST8M results. Figure 4 reports the full loss and runtime curves on MNIST8M over all tested coreset sizes, complementing the selected results reported in the main text.

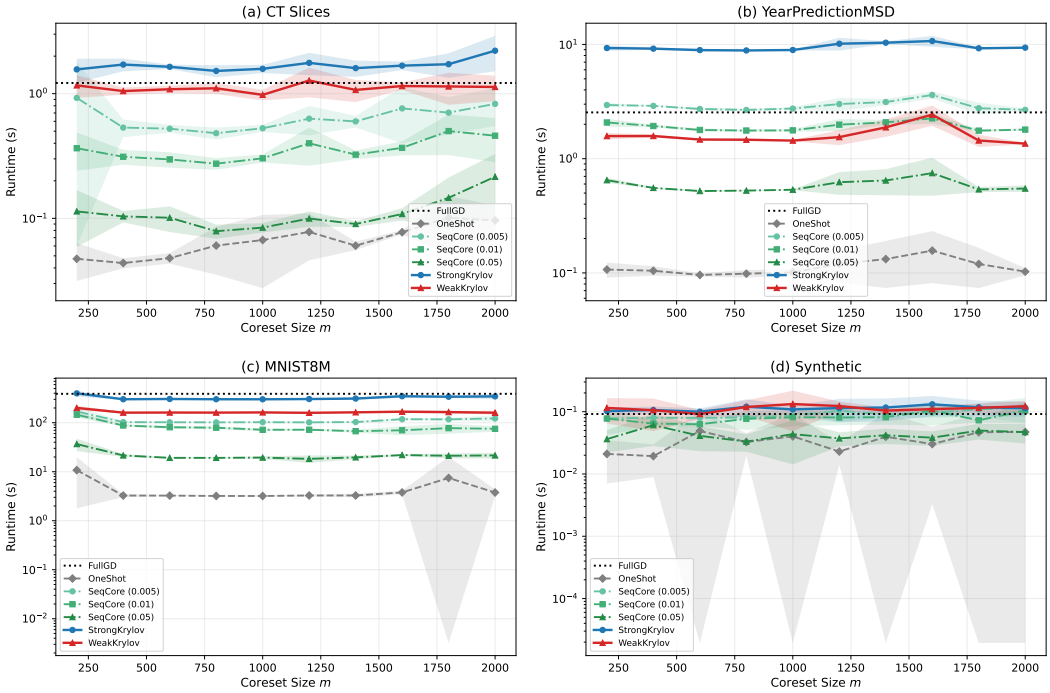

Figure 3: Comparison of runtime for OneShot, SeqCore with $r \in \{0.001, 0.005, 0.01, 0.05\}$, StrongKrylov, and WeakKrylov on CT Slices, YearPredictionMSD, MNIST8M, and Synthetic with varying coreset size m.

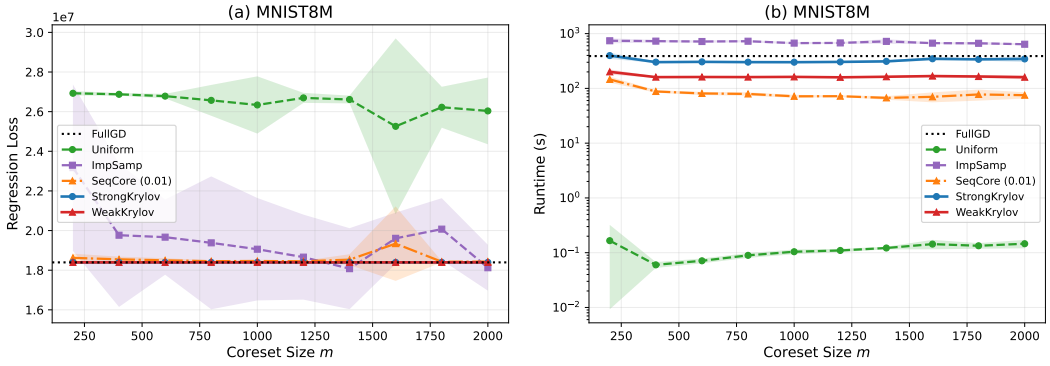

Figure 4: Comparison of coreset methods for gradient descent on MNIST8M ($n = 8{,}100{,}000, d = 784$) in terms of regression loss and runtime with varying coreset sizes.