REGULARIZED MOMENTUM ITERATIVE HESSIAN SKETCH FOR LARGE SCALE LINEAR SYSTEM OF EQUATIONS[*]

IBRAHIM KURBAN OZASLAN[†], MERT PILANCI[‡], AND ORHAN ARIKAN[†]

Abstract. In this article, Momentum Iterative Hessian Sketch (M-IHS) techniques, a group of solvers for large scale linear Least Squares (LS) problems, are proposed and analyzed in detail. The proposed techniques are obtained by incorporating the Heavy Ball Acceleration into the Iterative Hessian Sketch algorithm and they provide significant improvements over the randomized preconditioning techniques. Through the error analyses of the M-IHS variants, lower bounds on the sketch size for various randomized distributions to converge at a pre-determined rate with a constant probability are established. The bounds present the best results in the current literature for obtaining a solution approximation and they suggest that the sketch size can be chosen proportional to the statistical dimension of the regularized problem regardless of the size of the coefficient matrix. The statistical dimension is always smaller than the rank and it gets smaller as the regularization parameter increases. By using approximate solvers along with the iterations, the M-IHS variants are capable of avoiding all matrix decompositions and inversions, which is one of the main advantages over the alternative solvers such as the Blendenpik and the LSRN. Similar to the Chebyshev Semi-iterations, the M-IHS variants do not use any inner products and eliminate the corresponding synchronizations steps in hierarchical or distributed memory systems, yet the M-IHS converges faster than the Chebyshev Semi-iteration based solvers.

Key words. randomized preconditioning, iterative solvers, Tikhonov regularization, ridge regression, random projection, oblivious subspace embedding, acceleration, parallel and distributed computing

AMS subject classifications. 15B52, 65F08, 65F10, 65F22, 65F50, 68W20, 90C06

1. Introduction.

We are presenting a group of solvers, named as Momentum Iterative Hessian Sketch, M-IHS, that is designed for solving large scale linear system of equations in the form of

$$Ax_0 + \omega = b, \tag{1.1}$$

where $A \in \mathbb{R}^{n\times d}$ is the given data or coefficient matrix, b is the given measurement vector contaminated by the noise or computation/discretization error ω, and x_0 is the vector desired to be recovered. Due to contaminated measurements, solutions can differ significantly according to the constraints imposed on the problem. In this article, we are particularly interested in the $\ell 2$-norm regularized Least Squares (LS) solution:

$$x^* = \operatorname*{argmin}_x \quad \underbrace{\frac{1}{2}\left\|Ax-b\right\|_2^2 + \frac{\lambda}{2}\left\|x\right\|_2^2}_{f(x)}, \tag{1.2}$$

which is known as the Tikhonov Regularization or the Ridge Regression. The problem in (1.2) frequently arises in various large scale applications of science and engineering. For example, it can appear in the discretization of Fredholm Integral Equations

[*] Submitted to the journal's Methods and Algorithms for Scientific Computing December 5, 2019. A preliminary version of this paper appeared in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019.

Funding: The first authors research is funded by the Scientific and Technological Research Council of Turkey (TÜBİTAK) BİDEB 2210-A Scholarship Programme.

[†] EEE Department, Bilkent University (ikozaslan@gmail.com, oarikan@ee.bilkent.edu.tr).

[‡] EE Department, Stanford University (pilanci@stanford.edu).

of the first kind [28]. In those cases, the data matrix might be ill conditioned and the linear system might be either over-determined or square. When the system is under-determined, although sparse solutions are more popular due to the Compressed Sensing [16], the least norm solutions occupy an important place in statistic applications such as the Support Vector Machines [55, 13]. Solutions to the problem in both regimes, i.e, $n \geq d$ and $n < d$, are often required as intermediate steps of rather complicated algorithms such as the Interior Point and the ADMM that are widely used in machine learning and image processing applications [9, 11, 47].

Throughout the manuscript, we are going to assume that a proper regularization parameter λ estimate is available. In terms of complexity and error performance, estimating a proper regularization parameter is equally important as obtaining the regularized solution. Risk estimators such as the Discrepancy Principle, Unbiased Prediction Risk Estimate, Stein's Unbiased Risk Estimate and Generalized Cross Validation can be directly used to estimate the regularization parameter of the moderate size problems [52]. For large scale problems, these risk estimators can be adapted for the lower dimensional sub-problems that arise during the iteration of the first order iterative solvers [34]. A hybrid scheme that adaptively selects the regularization parameter along with the iterations is also suitable for the proposed `M-IHS` solvers. Indeed, we have developed such a technique, but we will present that study in a separate manuscript to keep the length of this document in a reasonable size.

The regularized solution in (1.2) can be obtained by using *direct* methods such as the Cholesky decomposition for square A, or the QR decomposition for rectangular A. However, $O(nd\min(n,d))$ computational complexity of any full matrix decomposition becomes prohibitively large as the dimensions increase. For large scale problems, linear dependence on both dimensions is acceptable and can be obtained by using the first order iterative solvers based on Krylov Subspaces [8]. These methods require only a few matrix-vector and vector-vector multiplications for each iteration, but the number of iterations that is needed to reach a certain level of tolerance is highly sensitive to the condition number of the coefficient matrix. If the largest and the smallest singular values of A are known, the optimal and un-improvable convergence rate of the first order iterative solvers is $O(1/k^2)$ which can be obtained by using Nesterov's Accelerated Gradient Descent or Polyak's Heavy Ball Method (HBM), unfortunately such information on the largest and the smallest singular values of A is rarely available in practice [40, 48]. In the absence of this information, the Conjugate Gradient (CG) technique achieves the same rate by adaptively tuning the *momentum* parameters through additional calculations at each iteration [12]. For the problem in (1.2), techniques such as the LSQR and LSMR, that are based on Golub-Kahan-Lanczos Bidiagonalization, produce more stable results than the CG technique with the same convergence rate [44, 23]. Other techniques that are based on the Krylov subspace approach with different convergence behaviours can be added to this list [5, 25]. When these techniques are applied to the problem in (1.2), their common convergence rate is characterized by the following inequality:

$$\left\|x^i - x^*\right\|_2 \leq \left(\frac{\sqrt{\kappa(A^TA + \lambda I_d)} - 1}{\sqrt{\kappa(A^TA + \lambda I_d)} + 1}\right)^i \left\|x^1 - x^*\right\|_2, \ 1 < i,$$

where x^* is the optimal solution of (1.2), x^1 is the initial guess, x^i is the i-th iterate of the solver and the condition number $\kappa(\cdot)$ is defined as the ratio of the largest singular value to the smallest singular value of its argument. Since for ill conditioned matrices $\kappa(A^TA + \lambda I_d)$ becomes large, the rate of convergence can be extremely slow.

The computational complexity of Krylov subspace-based iterative solvers is $O(nd)$ for each iteration, which is significantly less than $O(nd\min(n,d))$ if the number of iterations can be significantly fewer than $\min(n,d)$. However, in applications such as big data where A is very large dimensional, the computational complexity is not the only metric for feasibility of the algorithms. For example, if the coefficient matrix is too large to fit in a single working memory and it could be merely stored in a number of distributed computational nodes, then at least two distributed computations of matrix-vector multiplications are required at each iteration of algorithms such as the CGLS or the LSQR (please see [31, 6] for such applications). This suggests that the number of iterations should also be counted as an important metric to measure the overall complexity of an algorithm. One way to reduce the number of iterations in the iterative solvers is to use preconditioning which is a linear mapping of the solution domain for transforming an ill conditioned problem to a well conditioned problem. In the deterministic settings, finding a low-cost and effective preconditioning matrix is still a challenging task.

In addition to the number of iterations, the number of inner products at each iteration also plays an important role in the overall complexity. Each inner product calculation constitutes a synchronization point in parallel computing and therefore is undesirable for distributed or hierarchical memory systems [5]. The Chebyshev Semi-iterative (CS) technique can be preferred in this kind of applications, since it does not use any inner products and therefore eliminates some of the synchronization steps that are required by the techniques such as the CG or the GMRES. However, the CS requires prior information about the ellipsoid that contains all the eigenvalues of A, which is commonly not available in practice [26].

The Random Projection (RP) techniques that are based on the Johnson-Lindenstrauss Lemma have found diverse field of applications [30, 17]. These algorithms are capable of both reducing the dimensions and bounding the number of iterations with statistical guarantees, while they are convenient for parallel and distributed computations as much as the direct methods. The development and the applications of the RP based algorithms can be found in [54, 35] and references therein.

In the application of the RP to the regularized LS problem in (1.2), there are two main approaches. In the first approach, that is referred to as *classical sketching*, the coefficient matrix A and the measurement vector b are projected down onto a lower dimensional $(m \ll n)$[1] subspace using a randomly constructed *sketch matrix* $S \in \mathbb{R}^{m\times n}$. Then, the aim is to obtain an ζ-optimal solution with high probability in terms of the *cost approximation* [18, 45]:

$$\widetilde{x} = \underset{x}{\operatorname{argmin}} \ \frac{1}{2}\left\|SAx - Sb\right\|_2^2 + \frac{\lambda}{2}\left\|x\right\|_2^2, \text{ such that } f(\widetilde{x}) \leq (1+\zeta)f(x^*).$$

The best lower bounds on the sketch size for obtaining an ζ-optimal cost approximation have been derived for both sparse and dense sketch matrices by Avron et al. [1]. They showed that the sketch size can be chosen proportional to the *statistical dimension*, which is defined as $\mathbf{sd}_\lambda(A) = \mathbf{Tr}\left(A(A^TA+\lambda I_d)^{-1}A^T\right)$. Although the cost approximation is sufficient for many machine learning problems, small distance to the optimal solution, which is defined as the *solution approximation*, is a more preferable metric for the problems arisen from, for example, discretization of Fredholm integrals in the applied linear algebra [8, 52]. However, in [46], the classical sketching is shown

[1] Without loss of generality we can assume the linear system is over-determined, if it is not, then we can take a dual of the problem to obtain an over-determined problem as examined later.

to be sub-optimal in terms of sketch size for obtaining a solution approximation.

In the second approach, that is referred to as *randomized preconditioning*, the algorithms with reasonable sketch sizes obtain an η-optimal solution approximation:

$$\|\widehat{x} - x^*\|_W \leq \eta \|x^*\|_W ,$$

where W is a positive definite weight matrix, by iteratively solving a number of low dimensional sub-problems constituted by $(SA, \nabla f(x^i))$ pairs. To the best of our knowledge, Rokhlin [49] is the first who uses random projection techniques proposed in [18] to construct a preconditioning matrix for CG-like algorithms. He used the inverse of R factor in the QR decomposition of the *sketched matrix* SA for this purpose. Later, implementation of similar ideas resulted in Blendenpik and LSRN which have been shown to be faster than some of the deterministic solvers of LAPACK [3, 36]. To solve the preconditioned problems, as opposed to the Blendenpik which uses the LSQR, the LSRN uses the CS technique for parallelization purposes and deduce the prior information about the eigenvalues by using results from the random matrix theory. The main drawback of the LSRN and Blendenpik is that regardless of the desired accuracy η, one has to pay the whole cost, $O(md^2)$, of a full $m \times d$ dimensional matrix decomposition which is the dominant term in the computational complexity of both algorithms. Iterative Hessian Sketch (IHS) proposed in [46] eliminates the dominant term, $O(md^2)$, by using the inverse of the sketched Hessian as preconditioning matrix in the Gradient Descent method [53]. Therefore, instead of computing a full decomposition or an inversion, a linear system can be approximately solved for a pre-determined tolerance[2]. By adapting this idea into the CG technique, Accelerated IHS (A-IHS) has been obtained in [53]. Lastly, in [42], it has been showed that if the linear system is strongly over-determined, then the momentum parameters of the HBM can be robustly estimated by using Marchenko Pastur (MP) Law [19, 20]. This analysis results in a prototype solver `M-IHS` that we study here in detail.

The statistical lower bounds in the current literature suggest that the sketch size in randomized preconditioning algorithms can be chosen proportional to the rank of the problem which can be much larger than the statistical dimension. Although, some lower bounds on the sketch size that are proportional to the statistical dimension have been obtained in Kernel Ridge Regression [2], it has not obtained in regularized LS problem yet.

2. Contributions.

In this article, we describe a group of random projection based iterative solvers for large scale regularized LS problems. The proposed `M-IHS` variants can be used for any dimension regimes if the statistical dimension of the problem is sufficiently smaller than both size of the coefficient matrix A. In section 5, we give the detailed convergence analyses of the proposed techniques. Our guarantees, presented in Theorem 3.1 and Corollary 3.3, are based on the solution approximation metric as opposed to the results obtained for cost approximation metric in [1]. In Corollary 3.2, we derived the best lower bounds, in the current literature, on the sketch size of various randomized distribution for attaining a certain convergence rate with a constant probability. These guarantees can be easily extended to any other sketch types by using the *Approximate Matrix Multiplication* (AMM) property [15]. If tighter bounds are acquired for the AMM property in the future, our bounds can be automatically improved as well. Although our bounds for the dense sketch matrices such as Subgaussian or Randomized Orthonormal Systems (ROS) are the same as

[2] We are going to explore this idea in detail later in the article.

in [1], we gained slightly better results for the sparse sketch matrices. Additionally, we provide some empirical bounds for the sketch size and the rate of convergence in Corollary 3.4 which is very tight as demonstrated through various numerical experiments. Lastly, in Algorithm 6.1, we extend the idea of LSQR into the linear problems in the form of $A^T Ax = b$ which we need to solve during the iterations of all proposed *Inexact* `M-IHS` variants and of the Newton Sketch [47]. The advantages that the proposed sub-solver, referred to as `AAb_Solver`, provides over the symmetric CG technique are the same as the advantages that are provided by the LSQR over the CGLS technique [44]. The *MATLAB* implementations of the proposed solvers together with the codes that generate the figures in the article can be found in the following link: https://github.com/ibrahimkurban/M-IHS.

3. The proposed solvers for regularized LS problems.

The `M-IHS` is obtained by combining the IHS technique with the Heavy Ball Acceleration. The IHS technique iteratively minimizes the quadratic objective in (1.2) by performing the following updates:

$$\begin{aligned} x^{i+1} &= \underset{x\in\mathbb{R}^d}{\operatorname{argmin}} \ \left\|S_i A(x - x^i)\right\|_2^2 + \lambda \left\|x - x^i\right\|_2^2 + 2\langle\nabla f(x^i),\ x\rangle, \\ &= x^i + (A^T S_i^T S_i A + \lambda I_d)^{-1} \left(A^T(b - Ax^i) - \lambda x^i\right). \end{aligned}$$

In all iterations, the same sketch matrix can be used by adding a proper step size and the convergence can be accelerated by an additional momentum term as:

$$\begin{aligned} \Delta x^i &= \underset{x\in\mathbb{R}^d}{\operatorname{argmin}} \ \left\|SAx\right\|_2^2 + \lambda \left\|x\right\|_2^2 + 2\left\langle\nabla f(x^i),\ x\right\rangle, \\ x^{i+1} &= x^i + \alpha_i\Delta x^i + \beta_i\left(x^i - x^{i-1}\right). \end{aligned} \tag{3.1}$$

If we restrict our attention to the fixed parameters, the optimal momentum parameters α and β, that maximize the convergence rate, can be estimated by using random matrix theory as completely independent of spectral properties of the coefficent matrix A. Here, the linear system is assumed to be strongly over-determined, i.e., $n \gg d$, but by using the dual problem, the theory can be easily extended to the strongly under-determined case of $d \gg n$ as well [10]. A dual of the problem in (1.2) is

$$\nu^* = \underset{\nu\in\mathbb{R}^n}{\operatorname{argmin}} \ \underbrace{\frac{1}{2}\left\|A^T\nu\right\|_2^2 + \frac{\lambda}{2}\left\|\nu\right\|_2^2 - \langle b,\ \nu\rangle}_{g(\nu)}, \tag{3.2}$$

and the relation between the solutions of the primal and dual problem is

$$\nu^* = (b - Ax^*)/\lambda \Longleftrightarrow x^* = A^T\nu^*. \tag{3.3}$$

The dual problem in (3.2) can be minimized by using the same approach as the `M-IHS`:

$$\begin{aligned} \Delta\nu^i &= \underset{\nu\in\mathbb{R}^n}{\operatorname{argmin}} \ \left\|SA^T\nu\right\|_2^2 + \lambda\left\|\nu\right\|_2^2 + 2\left\langle\nabla g(\nu^i),\ \nu\right\rangle, \\ \nu^{i+1} &= \nu^i + \alpha\Delta\nu^i + \beta\left(\nu^i - \nu^{i-1}\right), \end{aligned} \tag{3.4}$$

and the solution of the primal problem can be obtained through the relation in (3.3). We refer to this algorithm as `Dual M-IHS`. The convergence rates of the `M-IHS` and `Dual M-IHS` solvers together with the necessary condition are stated in the Theorem 3.1 below.

Theorem 3.1. *Let A and b are the given data in (1.1); $x^* \in \mathbb{R}^d$ and $\nu^* \in \mathbb{R}^n$ are as in (1.2) and (3.2), respectively. Let $U_1 \in \mathbb{R}^{\max(n,d)\times\min(n,d)}$ consists of the first n rows of an orthogonal basis for $\begin{bmatrix} A \\ \sqrt{\lambda} I_d \end{bmatrix}$ if the problem is over-determined, and consists of the first d rows of an orthogonal basis for $\begin{bmatrix} A^T \\ \sqrt{\lambda} I_n \end{bmatrix}$ if the problem is under-determined. Let the sketching matrix $S \in \mathbb{R}^{m\times\max(n,d)}$ be drawn from a distribution $\mathcal{D}$ such that*

$$\mathbb{P}_{S\sim\mathcal{D}}\left[\left\|U_1^T S^T S U_1 - U_1^T U_1\right\|_2 \geq \epsilon\right] < \delta, \quad \epsilon \in (0,1). \tag{3.5}$$

Then, the `M-IHS` *applied on (1.2) and the* `Dual M-IHS` *applied on (3.2) with fixed momentum parameters*

$$\beta^* = \left(\frac{\sqrt{1+\epsilon} - \sqrt{1-\epsilon}}{\sqrt{1+\epsilon} + \sqrt{1-\epsilon}}\right)^2, \qquad \alpha^* = (1-\beta^*)\sqrt{1-\epsilon^2},$$

converge to the optimal solutions, x^* and ν^*, respectively, at the following rate[3] *with probability at least $1-\delta$:*

$$\left\|x^{i+1} - x^*\right\|_{D^{-1}} \leq \frac{\epsilon}{1+\sqrt{1-\epsilon^2}} \left\|x^i - x^*\right\|_{D^{-1}},$$
$$\left\|\nu^{i+1} - \nu^*\right\|_{D^{-1}} \leq \frac{\epsilon}{1+\sqrt{1-\epsilon^2}} \left\|\nu^i - \nu^*\right\|_{D^{-1}},$$

where D^{-1} is the diagonal matrix whose diagonal entries are $\sqrt{\sigma_i^2 + \lambda}$ and σ_i is the i-th singular value of A.

The proof of Theorem 3.1 can be found in subsection 5.1. The Theorem 3.1 is valid as well for the un-regularized case of $\lambda = 0$ as mentioned in subsection 5.3. Some possibilities for the distributions satisfying the condition in (3.5) of Theorem 3.1 are given in Corollary 3.2.

Corollary 3.2. *The condition in Theorem 3.1 is satisfied if the sketch matrix S is in one of the following types:*

(i) Sparse Subspace Embedding [54] with single nonzero element in each column, sketch size

$$m = \Omega\left(\mathbf{sd}_\lambda(A)^2/(\epsilon^2\delta)\right)$$

and SA is computable in $O(\mathbf{nnz}(A))$;

(ii) Sparse Subspace Embedding with $\alpha > 2$, $\delta < 1/2$, $\epsilon < 1/2$,

$$s = \Omega(\log_\alpha(\mathbf{sd}_\lambda(A)/\delta)/\epsilon)$$

non-zero elements in each column [33, 39], size

$$m = \Omega(\alpha \cdot \mathbf{sd}_\lambda(A)\log(\mathbf{sd}_\lambda(A)/\delta)/\epsilon^2)$$

and SA is computable in $O(s \cdot \mathbf{nnz}(A))$;

(iii) SRHT sketch matrix [15, 46] with size

$$m = \Omega\left(\left(\mathbf{sd}_\lambda(A) + \log(1/\epsilon\delta)\log(\mathbf{sd}_\lambda(A)/\delta)\right)/\epsilon^2\right)$$

and SA is computable in $O(nd\log(m))$;

[3] $\sqrt{\beta} = \frac{\epsilon}{1+\sqrt{1-\epsilon^2}}$

(iv) Sub-Gaussian sketch matrix [45, 15] with size

$$m = \Omega(\mathbf{sd}_\lambda(A)/\epsilon^2)$$

and SA is computable in $O(ndm)$.

The proof of Corollary 3.2 can be found in subsection 5.2. For the solution approximation, we do not need the second condition in Lemma 11 of [1], hence we obtained slightly better results for the sparse subspace embeddings in item (i) and (ii) of Corollary 3.2. The number of iterations to reach a certain level of accuracy is stated in the following corollary.

Corollary 3.3. *For some $\epsilon \in (1, 1/2)$ and arbitrary η, if the sketch size meets the condition of the corresponding distribution in Corollary 3.2 and the fixed momentum parameters are chosen as in Theorem 3.1, then the* `M-IHS` *and the* `Dual M-IHS` *obtain an η-optimal solution approximation in $\ell 2$-norm with total of*

$$N = \left\lceil \frac{\log(\eta)\log(C)}{\log(\epsilon) - \log(1+\sqrt{1-\epsilon^2})} \right\rceil$$

iterations, where the constant C, that is defined as $C = \sqrt{\kappa(A^TA + \lambda I_d)}$ for the `M-IHS` *and $C = \kappa(A)\sqrt{\kappa(AA^T + \lambda I_n)}$ for the* `Dual M-IHS`*, can be removed if the semi-norm in Theorem 3.1 is used as the solution approximation metric instead of $\ell 2$ norm.*

Corollary 3.3 is an immediate result of Theorem 3.1 combined with Corollary 3.2. Choosing a sketch size m and an error constant ϵ, that satisfy the statistical bounds in both Theorem 3.1 and Corollary 3.2, might be challenging for researches who are not specialists in the field. So in Corollary 3.4, we obtained substantially simplified empirical versions of these bounds by using the MP Law [19] and by approximating the Tikhonov regularization filtering coefficients with the binary coefficients.

Corollary 3.4. *Let the entries of the sketch matrix be independent, zero mean, unit variance and that have bounded higher order moments. If the Truncated SVD regularization with truncation parameter $\lceil \mathbf{sd}_\lambda(A) \rceil$ is used, then the* `M-IHS` *and the* `Dual M-IHS` *with the following momentum parameters*

$$\beta = \frac{\mathbf{sd}_\lambda(A)}{m}, \qquad \alpha = (1-\beta)^2$$

will converge to the optimal solutions x^* and ν^*, respectively, with a convergence rate of $\sqrt{\beta}$ as $m \to \infty$ while $\min(n,d)/m$ remains constant. Any sketch size $m > \mathbf{sd}_\lambda(A)$ can be chosen to obtain an η-optimal solution approximation at most $\frac{\log(\eta)}{\log(\mathbf{sd}_\lambda(A)/m)}$ iterations.

Remark 3.5. Although the MP law provides bounds for the singular values of the sketch matrix S in asymptotic regime, i.e., as $m \to \infty$; the bounds become very good estimators of the actual bounds when m takes finite values. As shown in Figure 1, the rate of $\sqrt{\beta}$ creates a remarkable fit for the numerical convergence rate of the `M-IHS` variants when the momentum parameters given in Corollary 3.4 are used for the Tikhonov regularization. This is because the sigmoid-like filtering coefficients in the Tikhonov regularization can be thought of as the smoothed version of the binary coefficients in the TSVD solution and therefore the binary coefficients constitute a good approximation for the filtering coefficients of the Tikhonov regularization as noted in subsection 5.4.

In practice, the M-IHS and Dual M-IHS eliminate all the quadratic terms in the complexity expression by approximately solving the low dimensional linear systems in (3.1) and (3.4) instead of computing a matrix decomposition or an inversion. This *inexact sub-solver* approach constitutes a trade-off between the computational complexity and the accuracy, that is highly desirable in very large dimensional problems where solutions with relatively lower accuracy are acceptable. Unfortunately, forming this trade-off is not possible for the Blendenpik and the LSRN techniques. Inexact sub-solvers have been known to be a good heuristic way to create this trade-off and they are widely used in the algorithms that are based on the Newton Method to solve the large scale normal equations [41]. In these inexact (or truncated) Newton Methods, inner iterations are terminated at the moment that the relative residual error is below some iteration dependent threshold, named as the *forcing terms* [21]. In the literature, there are various techniques to choose these forcing terms that guarantee global convergence, but the number of iterations suggested by these techniques are far above the total number of iterations used in practice. We refer interested reader to [37] and we go forward with the heuristic constant threshold, ϵ_{sub}, that checks the relative residual error of the linear system [7].

Any Krylov subspace techniques can be used to solve the sub-problems in (3.1) and (3.4), but LSQR-like solvers that are adapted for the normal equations would require computations of 4 matrix-vector multiplications per iteration. On the other hand, due to the explicit calculation of $(SA)^T(SA)z$, the symmetric CG, that would require only 2 matrix-vector multiplications, might be unstable in the ill-conditioned problems [44]. Therefore, in section 6, we propose a stable sub-solver which is particularly designed for the problems in the form of $A^TAx = b$. The proposed sub-solver, referred to as AAb_Solver, is based on the Golub Kahan Bidiagonalization and it uses a similar approach that the LSQR uses on the LS problem. In addition to the stability advantage over symmetric CG technique, AAb_Solver produces a bidiagonal representation of sketched matrix as a byproduct of the iterations. This bidiagonal form can be used in, for example, Generalized Cross Validation [24, 34] to estimate the problem related parameters including the regularization parameter and the statistical dimension. The inexact versions of the M-IHS and Dual M-IHS that use AAb_Solver are given in Algorithm 3.1 and Algorithm 3.2 where RP_fun represents the function that generates the desired sketched matrix such that $\mathbb{E}\left[S^TS\right] = I_m$ whose implementation details can be found in the relevant references in Corollary 3.2.Setting the

Algorithm 3.1 M-IHS (for $n \geq d$)

```
Input: $A,\ b,\ m,\ \lambda,\ x^1,\ \mathbf{sd}_\lambda(A),\ \epsilon_{sub}$
$SA = $ RP_fun$(A, m)$
  $\beta = \mathbf{sd}_\lambda(A)/m$
  $\alpha = (1-\beta)^2$
while until stopping criteria do
    $g^i = A^T(b - Ax^i) - \lambda x^i$
    $\Delta x^i = $ AAb_Solver$(SA,\ g^i,\ \lambda,\ \epsilon_{sub})$
    $x^{i+1} = x^i + \alpha\Delta x^i + \beta(x^i - x^{i-1})$
end while
```

forcing term ϵ_{sub}, for instance, to 0.1 for all iterations is enough for the inexact M-IHS variants to converge at the same rate $\sqrt{\beta}$ as the exact versions as demonstrated in

Algorithm 3.2 `Dual M-IHS` (for $n \leq d$)

```
Input: $A,\ b,\ m,\ \lambda,\ \mathbf{sd}_\lambda(A),\ \epsilon_{sub}$
$SA^T = \texttt{RP\_fun}(A^T, m)$
    $\beta = \mathbf{sd}_\lambda(A)/m$
    $\alpha = (1-\beta)^2$
    $\nu^0 = 0$
while until stopping criteria do
    $g^i = b - AA^T\nu^i - \lambda\nu^i$
    $\Delta\nu^i = \texttt{AAb\_Solver}(SA^T,\ g^i,\ \lambda,\ \epsilon_{sub})$
    $\nu^{i+1} = \nu^i + \alpha\Delta\nu^i + \beta(\nu^i - \nu^{i-1})$
end while
$x^{N+1} = A^T\nu^{N+1}$
```

Figure 1.

Corollary 3.2 suggests that if the statistical dimension is several times smaller than the dimensions of A, then it is possible to choose a substantially smaller sketch size than the $\min(n,d)$. If this is the case, then the quadratic objective functions in (3.1) and (3.4) become strongly under-determined problems, which makes it possible to approximate the Hessian of the objective functions one more time by taking their convex dual as it has been done in the `Dual M-IHS`. This approach is similar to approximately solving the problems in (3.1) and (3.4) by using the `AAb_Solver`, but this time we are going to apply a second dimension reduction. At the end of two Hessian sketching, the linear sub-problem whose dimensions are reduced from both sides can be efficiently solved by the `AAb_Solver` for a pre-determined tolerance as before. Consider the following dual of the sub-problem in (3.1)

$$z^* = \underset{z \in \mathbb{R}^m}{\operatorname{argmin}} \ \underbrace{\frac{1}{2}\left\|A^TS^Tz + \nabla f(x^i)\right\|_2^2 + \frac{\lambda}{2}\left\|z\right\|_2^2}_{h(x^i,z)}, \tag{3.6}$$

which is a strongly over-determined problem if $m \ll \min(n,d)$. Hence, it can be approximately solved by the `M-IHS` updates:

$$\begin{aligned} \Delta z^j &= \underset{z \in \mathbb{R}^m}{\operatorname{argmin}} \ \left\|WA^TS^Tz\right\|_2^2 + \lambda\left\|z\right\|_2^2 + 2\left\langle \nabla_z h(x^i, z^j),\ z\right\rangle, \\ z^{j+1} &= z^j + \alpha_2\Delta z^j + \beta_2\left(z^j - z^{j-1}\right). \end{aligned} \tag{3.7}$$

After M iterations, the solution of (3.1) can be recovered by using the relation in (3.3) as $\Delta x^i = (\nabla f(x^i) - A^TS^Tz^M)/\lambda$. The same strategy can be applied on the sub-problem in (3.4) by replacing SA with SA^T and $\nabla f(x^i)$ with $\nabla g(\nu^i)$. The resulting algorithms, referred to as `Primal Dual M-IHS`, are given in Algorithm 3.3 and Algorithm 3.4, respectively.

The primal dual idea presented here is first suggested by Zhang et al. in [53]. They used the A-IHS technique to solve the sub-problems that arise during the iterations of the Accelerated Iterative Dual Random Projection (A-IDRP) which is a dual version of the A-IHS. However, since both of the A-IHS and the A-IDRP are based on the CG technique, the convergence rate of the proposed A-IHS, A-IDRP and the primal dual algorithm called as Accelerated Primal Dual Sketch (A-IPDS)

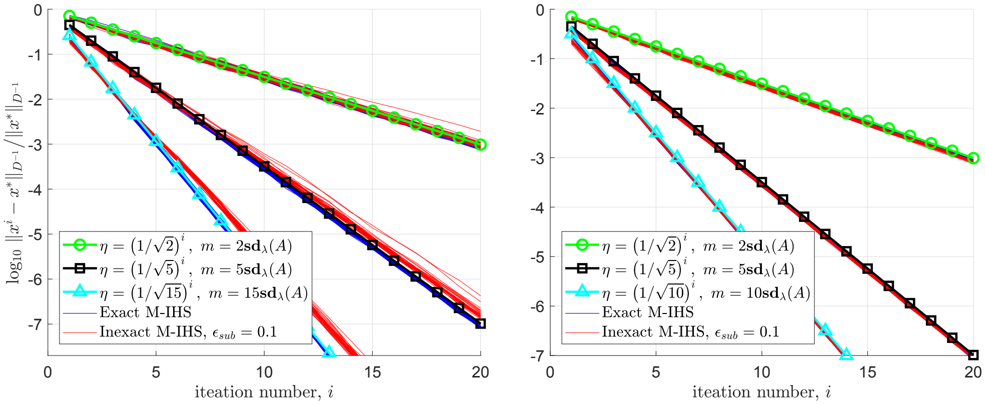


Fig. 1: Convergence rate of the proposed solvers: The coefficient matrix, $A \in \mathbb{R}^{32768 \times 1000}$ with $\kappa(A) = 10^8$, of the left plot is generated by using *philips* singular value profile of RegTool [27]. The details of data generation can be found in section 4. SRHT sketches with Discrete Cosine Transform is used. For the right plot, the coefficient matrix, $A \in \mathbb{R}^{24336 \times 1296}$ is generated by using *sprand* command of MATLAB. We first create a sparse matrix with size $\tilde{A} \in \mathbb{R}^{20 \times 6}$ and sparsity 15%, then the final form is obtain by taking $A = \tilde{A}^{\otimes 4}$ and deleting the all zero rows. The final form, A, has sparsity ratio 0.1% and condition number $\kappa(A) = 10^7$. The CountSketch matrix with single nonzero element in each column is used for the right plot. The noise level $\|w\|_2 / \|Ax_0\|_2 = 1\%$ is used for both problems and the resulting statistical dimensions are 119 and 410, respectively. The lines with different markers show theoretical convergence rate when the corresponding sketch size is used. Both *exact* and *inexact* (seen on Algorithm 3.1) versions of `M-IHS` are run 32 times and the result of each run is plotted as a separate thin line. The empirical momentum parameters given in Corollary 3.4 are used for both schemes. The rate of $\sqrt{\beta}$ provides a remarkable fit to the numerical convergence rate of the exact scheme (thin blue lines), and except for a small degradation in the SRHT case, setting forcing term to a small constant such as $\epsilon_{sub} = 0.1$ is sufficient for the inexact scheme to achieve the same rate as the exact version in these experiments. Both figures suggest that the empirical momentum parameters obtained through the TSVD approximation and the MP law provided in subsection 5.4 successfully models the convergence behaviour of the `M-IHS` variants.

are all degraded in the LS problems with high condition numbers due to the instability issue of the symmetric CG technique [44]. Even if the regularization is used, still the performance of the solvers proposed in [53] are considerably deteriorated compared to the other randomized preconditioning techniques as shown in section 4. Further, applying the preconditioning idea of IHS to the stable techniques such as the LSQR that are adapted for the LS problem is not so efficient as the `M-IHS` variants, because two preconditioning systems are needed to be solved per iteration for such techniques.

The computational saving when we apply a second dimension reduction as in the `Primal Dual M-IHS` may not be significant due to the second gradient computations in Line 10 and 9 of the given algorithms, but the lower dimensional sub-problems that we obtain at the end of the second sketching can be used to estimate several parameters including the regularization parameter itself. Indeed, we are currently working on

Algorithm 3.3 `Primal Dual M-IHS` (for $n \leq d$)

```
Input: $A,\ b,\ m_1,\ m_2,\ \lambda,\ \mathbf{sd}_\lambda(A),\ \epsilon_{sub}$
    $SA^T = \texttt{RP\_fun}(A^T, m_1)$
  $WAS^T = \texttt{RP\_fun}(SA^T, m_2)$
       $\beta_\ell = \mathbf{sd}_\lambda(A)/m_\ell, \quad \ell = 1, 2$
       $\alpha_\ell = (1-\beta_\ell)^2, \quad \ell = 1, 2$
     $\nu^{1,0} = 0,\ z^{1,0} = 0$
for i=1:N do
    $b^i = b - AA^T\nu^i - \lambda\nu^i$
    for j=1:M do
        $g^{i,j} = SA^T(b^i - AS^T z^j) - \lambda z^j$
        $\Delta z^j = \texttt{AAb\_Solver}(WAS^T, g^{i,j},\ \lambda,\ \epsilon_{sub})$
        $z^{j+1} = z^j + \alpha_2\Delta z^j + \beta_2(z^j - z^{j-1})$
    end for
    $\Delta\nu^i = (b^i - AS^T z^{M+1})/\lambda, \quad z^{1,0} = z^{M+1,M}$
    $\nu^{i+1} = \nu^i + \alpha_1\Delta\nu^i + \beta_1(\nu^i - \nu^{i-1})$
end for
$x^{N+1} = A^T\nu^{N+1}$
```

such an algorithm to estimate the regularization parameter of the large scale problems along with the iterations of the `M-IHS` variants without requiring additional accesses to coefficient matrix A.

The `Primal Dual M-IHS` techniques are extension of the inexact schemes. Therefore, their convergence rates depend on their forcing terms that are used to stop the inner iterations [37]. In [53], an upper bound for the error of the primal dual updates is proposed. However as it is detailed in section 8, there are several inaccuracies in the development of the bound. Therefore, finding a provably valid lower bound on the number of inner loop iterations, that guarantee a certain rate of convergence at the main loop, is still an open question for the primal dual algorithms.

The statistical dimension in Algorithm 3.1, 3.2, 3.3 and 3.4 can be estimated by using a Hutchinson-like randomized trace estimator. We refer interested reader to [4] for a detailed comparison of Hutchinson-like estimators. Alternatively, an algorithm is proposed in [1] to estimate $\mathbf{sd}_\lambda(A)$ within a constant factor in $\mathbf{nnz}(A)$ time with a constant probability, if $\mathbf{sd}_\lambda(A) \leq M$ where:

$$M = \min\{n, d, \lfloor (n+d)^{1/3}/\text{poly}(\log(n+d)) \rfloor\}.$$

However, due to the third order root and the division by at least a sixth order polynomial, we can only have such small statistical dimension value when the singular values of A decay severely/exponentially and the signal-to-noise ratio is very low. Therefore, we preferred to use the heuristic trace estimator in Algorithm 3.5. The input matrix SA can be replaced with SA^T or even with WA^TS^T and WAS^T according to the algorithm used. Any estimator in [4] can be substituted for the Hutchinson Estimator and the number of samples, N, can be chosen accordingly. In the experiments that we made with various singular value profiles, small samples sizes such as 2 or 3 and $\epsilon_{tr} = 0.01$ is sufficient to obtain satisfactory results. Noting that the momentum pa-

Algorithm 3.4 `Primal Dual M-IHS` (for $n \geq d$)

```
Input: $A,\ b,\ m_1,\ m_2,\ \lambda,\ x^1,\ \mathbf{sd}_\lambda(A),\ \epsilon_{sub}$
      $SA = \texttt{RP\_fun}(A, m_1)$
$WA^TS^T = \texttt{RP\_fun}(SA, m_2)$
      $\beta_\ell = \mathbf{sd}_\lambda(A)/m_\ell, \quad \ell = 1, 2$
      $\alpha_\ell = (1-\beta_\ell)^2, \quad \ell = 1, 2$
      $x^0 = 0,\ z^{1,0} = 0$
for i=1:N do
    $b^i = A^T(b - Ax^i) - \lambda x^i$
    for j=1:M do
        $g^{i,j} = SA(b^i - A^TS^Tz^j) - \lambda z^j$
        $\Delta z^j = \texttt{AAb\_Solver}(WA^TS^T, g^{i,j},\ \lambda,\ \epsilon_{sub})$
        $z^{j+1} = z^j + \alpha_2 \Delta z^j + \beta_2(z^j - z^{j-1})$
    end for
    $\Delta x^i = (b^i - A^TS^Tz^{M+1})/\lambda, \quad z^{1,0} = z^{M+1,M}$
    $x^{i+1} = x^i + \alpha_1 \Delta x^i + \beta_1(x^i - x^{i-1})$
end for
```

Algorithm 3.5 `Inexact Hutchinson Trace Estimator`

```
Input: $SA \in \mathbb{R}^{m\times d},\ \lambda,\ N,\ \epsilon_{tr}$
$v^\ell = \{-1, +1\}^d, \quad \ell = 1, \dots, N$
$\tau = 0$
for i = 1:N do
    $x = SAv^i$
    $z^i = \texttt{AAb\_Solver}(SA, x, \lambda, \epsilon_{tr})$
    $\tau = \tau + x^Tz^i$
end for
Output: $\tau = \tau/N$
```

rameters suggested in Corollary 3.4 are not sensitive to small errors in $\mathbf{sd}_\lambda(A)$, since $\beta = \mathbf{sd}_\lambda(A)/m$. When the sketch size m exceeds thousands, for example, an error made at the first digit of the statistical dimension estimation, distorts the momentum parameter only at the third decimal.

4. Numerical Experiments and Comparisons.

For a fair comparison, we have implemented all the proposed algorithms in this manuscript as well as those that are used for comparisons in MATLAB. All codes can be found in the link[4]. The coefficient matrix $A \in \mathbb{R}^{n\times d}$ is generated for various sizes as follows: we first sample the entries of A from distribution $\mathcal{N}(1_d, \Gamma)$ where $\Gamma_{ij} = 5 \cdot 0.9^{|i-j|}$ so that the columns are highly correlated with each other. Then, by using the SVD, we replace the singular values with *philips* profile provided in RegTool. We scale the singular values to set the condition number $\kappa(A)$ to 10^8 and we use the same input signal provided by RegTool [27]. In this way, we have obtained a challenging setup for any first order iterative

[4] https://github.com/ibrahimkurban/M-IHS

solvers to compare their performances. In all experiments, the same setup has been used unless indicated. We counted the number of operations according to Hunger's report [29]. All results have been obtained by averaging over 32 MC simulations.

We compared the proposed algorithms with the state of the art randomized preconditioning techniques which can reach any level of desired accuracy within a bounded number of iterations. The compared methods can be briefly described as follows. The Blendenpik uses the R factor in QR decomposition of the sketched matrix SA as the preconditioning matrix for the LSQR algorithm just like the method proposed by Rokhlin et al [3, 49] and it uses Randomized Orthonormal System (ROS) to generate the sketched matrix [45]. The LSRN uses the V factor in the SVD similar to the Blendenpik. In spite of its high running complexity, the Gaussian sketch matrices are preferred in the LSRN. In addition to the LSQR, the CS can be used in the LSRN as the core solver for parallelization purposes without calculating the singular values explicitly [36]. The IHS uses the sketched Hessian as the preconditioning matrix for the Gradient Descent. The Accelerated IHS (A-IHS) uses this idea for the CG algorithm in over-determined problems. The dual counter-part of the A-IHS algorithm, A-IDRP, is shown to be faster than the Dual Random Projection algorithm proposed in [55], so we did not include the DRP in the simulations. Additionally, we include a CS variant of the IHS (IHS-CS) to the comparisons. We combined the randomized preconditioning idea of IHS with the preconditioned CS method [5]. We found the bounds for the eigenvalues in the same way as the LSRN. We have solved the low dimensional sub-problems required by all IHS variants by taking the QR decomposition, but for *inexact* schemes, we have used the proposed `AAb_Solver` with constant forcing term $\epsilon_{sub} = 0.1$. We did not include inexact versions of the accelerated algorithms proposed in [53], since their exact versions are outperformed by the *Exact* `M-IHS` variants in all settings. Except for the LSRN variants which use Gaussian sketch matrices, we used Discrete Cosine Transform in the ROS for all techniques.

We compare the operation counts required by the algorithms to obtain a certain level of accuracy for the solution approximation metric. In the first experiment, we did not include noise in the linear system to emphasize the convergence rate that the algorithms can provide in such severely ill posed problems. To make the problem more challenging, we sampled the input vector, x_0, from uniform distribution $\mathrm{Uni}(-1, 1)$ for this experiment only. In such scenarios, convergence rates of Krylov subspace-based iterative solvers without preconditioning fall its minimum value. The results are shown in Figure 2. Due to high running time of the Gaussian sketches, $O(mnd)$, the LSRN variants require more operations (for the size of the problems considered here approximately 10 times larger) than the others. Due to the lack of inner product calculations, the `M-IHS` requires slightly fewer operations than the Blendenpik, nonetheless it reaches the same accuracy with the LSRN-LSQR. The A-IHS algorithm has the lowest performance which is expected in the un-regularized problems, since it is adapted on the CG technique that can be unstable for the un-regularized LS problems due to the high condition number [44]. The convergence of the CS-based techniques, both of the IHS and the LSRN variants, are substantially slower than the `M-IHS`, which suggests that the `M-IHS` algorithm can take the CSs place in those applications where parallel computation is viable. A similar comparison of the `M-IHS` with Accelerated Randomized Kaczmarz (ARK) and CGLS without preconditioning can be seen on Figure 2 of [42].

By using additive i.i.d Gaussian noise at various levels of variance, we tested the aforementioned algorithms on regularized LS problems. Results for over-determined and under-determined cases can be seen on Figure 3 and Figure 4, respectively. We

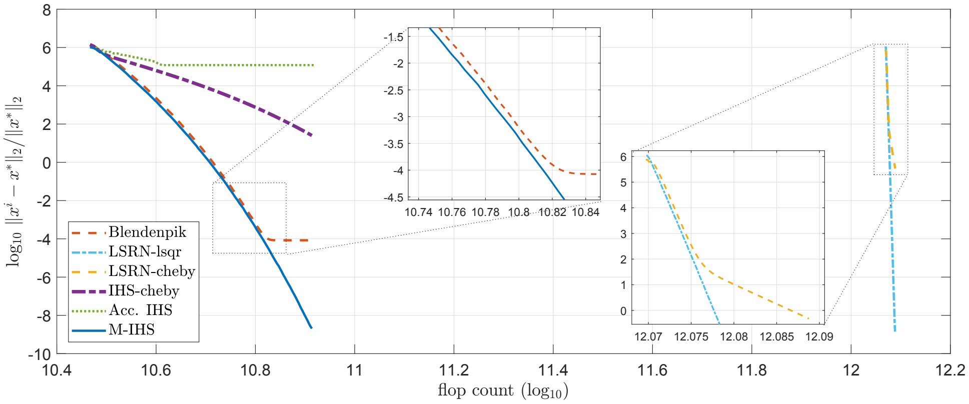


Fig. 2: LS problem without regularization: The problem size is $2^{16} \times 2000$. In order to compare the convergence rates, all solvers are allowed to do $N = 100$ iterations with same sketch size, $m = 4000$. According to the Corollary 3.4, we expect the M-IHS to reach an accuracy: $\left\|x^N - x_0\right\|_2 \leq \kappa(A) \left\|x_0\right\|_2 \left(1/\sqrt{2}\right)^N = 9 \cdot 10^{-8}$, which is exactly the case.

used a sketch size $m = \min(n, d)$ to emphasis the promise of the RP techniques although such sizes are not applicable for the LSRN variants. Even if the sketch size is increased further, the convergence rates of the LSRN variants were considerably lower than the others; so we leave out the LSRN variants from the comparison set in the regularized settings. The A-IHS and A-IDRP methods are slower than the Blendenpik, IHS-CS and M-IHS variants in the regularized setup as well. Besides, the inexact schemes proposed for the M-IHS and Dual M-IHS requires significantly less operations to reach the same level of accuracy as the exact versions. Although the inexact schemes require approximately 10 times less operations then the exact versions in these setups; the saving gets larger as the sketch size increases, because while any full decomposition requires $O(m \min(n, d)^2)$ operations, approximately solving the sub-problem requires only $O(m \min(n, d))$ operations.

Corollary 3.2 implies eligibility of sketch sizes that are smaller than the rank, $m \leq \min(n, d)$ which suggest that the RP techniques can be used for all dimension regimes whether it is strong or not, as long as the statistical dimension of the problem is small compared to the dimensions of coefficient matrix A. This implication can be verified in Figure 5 on which we showed the performance of the Primal Dual M-IHS techniques. As discussed in section 3, the Primal Dual M-IHS techniques seem to increase the complexity in the contrary of saving, but they obtained lower dimensional sub-problems than the M-IHS and Dual M-IHS, which suggests that some problem related parameters can be deduced more cheaply during iterations of the Primal Dual M-IHS technique. Lastly, the Primal Dual M-IHS variants have a noticeably higher rate of convergence than the A-IPDS algorithm which is based on the CG technique.

5. Theoretical analysis of proposed approaches. In this section, we are going to provide a unified analysis of the regularized LS problems through the proofs by including many comments and explanations. Throughout the analysis we denote $A = U\Sigma V^T$ as the compact SVD with $U \in \mathbb{R}^{n\times r}$, $\Sigma = \text{diag}(\sigma_1, \ldots, r) \in \mathbb{R}^{r\times r}$ and

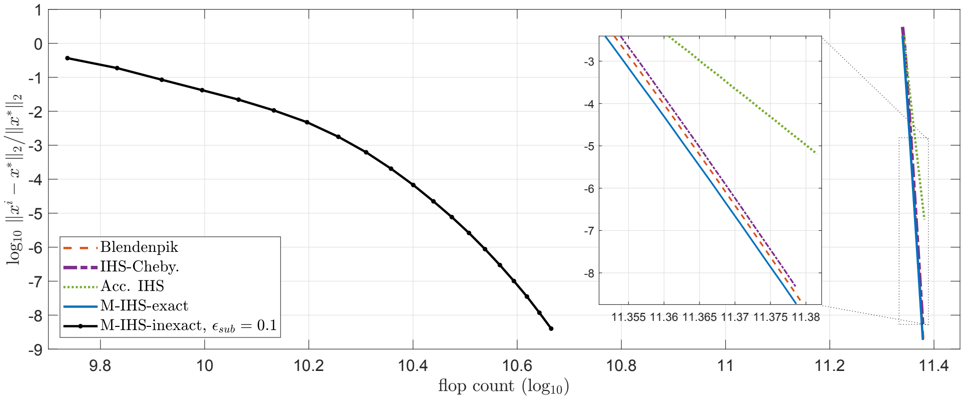


Fig. 3: Regularized LS problem ($n \gg d$): The problem size is $2^{16} \times 4000$. The noise level is $\|w\|_2 \big/ \|Ax_0\|_2 = 1\%$. For this noise level, the statistical dimension of the problem is $\mathbf{sd}_\lambda(A) = 443$. The optimal regularization parameter that minimizes the $\|x^*(\lambda) - x_0\|_2$ is provided to all techniques. All methods are allowed to do $N = 20$ iterations with the same sketch size of $m = 4000$. According to the Corollary 3.4, `M-IHS` is expected to satisfy: $\left\|x^N - x^*\right\|_2 \leq \|x^*\|_2 \sqrt{\kappa(A^TA + \lambda I_d)} \left(\sqrt{443/4000}\right)^N = 6 \cdot 10^{-9}$ which is almost exactly the case. The *Inexact* `M-IHS` requires significantly fewer operations to reach the same accuracy as others. For example to obtain an ($\eta = 10^{-4}$)-optimal solution approximation, the *Inexact* `M-IHS` requires approximately 10 times less operations than any techniques that need factorization of or inversion of the sketched matrix.

$V \in \mathbb{R}^{d \times r}$ where $r = \min(n, d)$.

5.1. Proof of Theorem 3.1. To prove the theorem for the `M-IHS` and the `Dual M-IHS`, we mainly combine the idea of *partly exact* sketching, that is proposed in [1], with the Lyapunov analysis, that we use in [42]. In parallel to [1], we define diagonal matrix $D := (\Sigma^2 + \lambda I_r)^{-1/2}$ and the *partly exact* sketching matrix as:

$$\widehat{S} = \begin{bmatrix} S & \mathbf{0} \\ \mathbf{0} & I_r \end{bmatrix}, \quad S \in \mathbb{R}^{m \times \max(n,d)}.$$

The Proof for M-IHS:. Let

$$\widehat{A} = \begin{bmatrix} U\Sigma D \\ \sqrt{\lambda} V D \end{bmatrix} = \begin{bmatrix} U_1 \\ U_2 \end{bmatrix}, \quad \widehat{A}^T \widehat{A} = I_d, \quad \widehat{b} = \begin{bmatrix} b \\ \mathbf{0} \end{bmatrix},$$

so that U_1 is the first n rows of an orthogonal basis for $\begin{bmatrix} A \\ \sqrt{\lambda} I_d \end{bmatrix}$ as required by the condition in (3.5) of the theorem. We have the following equality

$$\|Ax - b\|_2^2 + \lambda \|x\|_2^2 = \left\|\widehat{A}y - \widehat{b}\right\|_2^2, \text{ for } \left\{\forall x \in \mathbb{R}^d \mid y = D^{-1}V^Tx\right\},$$

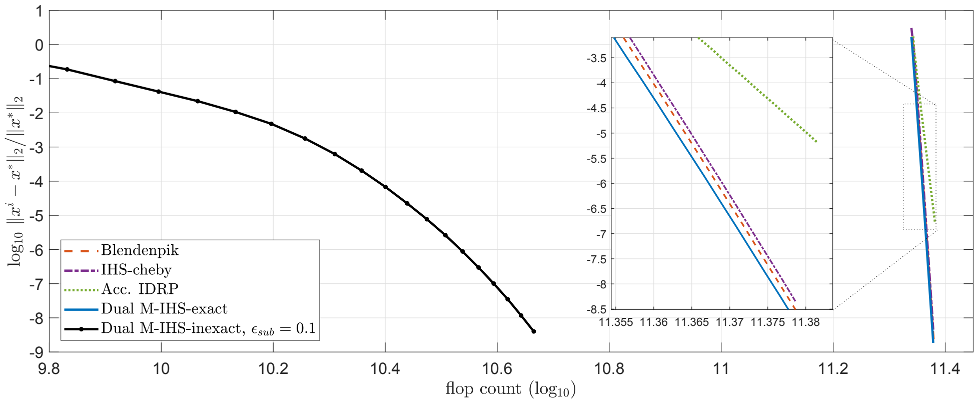


Fig. 4: Regularized LS problem ($n \ll d$): The problem size is 4000×2^{16}. The noise level is $\|w\|_2 / \|Ax_0\|_2 = 1\%$. For this noise level, the statistical dimension of the problem is $\mathbf{sd}_\lambda(A) = 462$. The optimal regularization parameter, in the same sense as Figure 3, is provided to the all techniques. All methods are allowed to do $N = 20$ iterations with same sketch size $m = 4000$. The comments as Figure 3 are valid in here as well. The *Inexact* scheme for `Dual M-IHS` is capable of reducing the complexity in considerable amounts.

and the sketched matrix

$$\widehat{S}\widehat{A} = \begin{bmatrix} SU\Sigma D \\ \sqrt{\lambda} V D \end{bmatrix} = \begin{bmatrix} SU_1 \\ U_2 \end{bmatrix}.$$

Since the following Hessian Sketch (HS) sub-problem

$$\Delta y^i = \operatorname*{argmin}_y \ \left\|\widehat{S}\widehat{A}y\right\|_2^2 - 2\langle \widehat{A}^T(\widehat{b} - \widehat{A}y^i),\ y\rangle$$

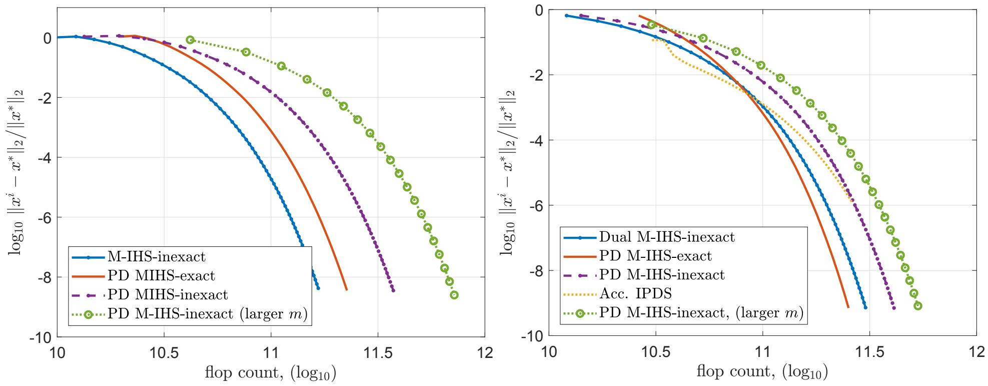
is equivalent to (3.1), we can examine the following bipartite transformation to find out the convergence properties of the `M-IHS`:

$$\begin{bmatrix} y^{i+1} - y^* \\ y^i - y^* \end{bmatrix} = \underbrace{\begin{bmatrix} (1+\beta)I_d - \alpha(\widehat{A}^T\widehat{S}^T\widehat{S}\widehat{A})^{-1} & -\beta I_d \\ I_d & \mathbf{0} \end{bmatrix}}_{T} \begin{bmatrix} y^i - y^* \\ y^{i-1} - y^* \end{bmatrix},$$

where $y^* = D^{-1}V^Tx^*$ and $\widehat{A}^T\widehat{A} = I_d$. The eigenvalues and therefore the contraction ratio of the transformation can be found analytically by converting the matrix T into a block diagonal form through the same similarity transformation given in [42]:

$$T = P^{-1}\operatorname{diag}(T_1, \dots, T_d)P, \quad T_i := \begin{bmatrix} 1 + \beta - \alpha\lambda_i & \beta \\ 1 & 0 \end{bmatrix},$$

$$P = \begin{bmatrix} \Psi & 0 \\ 0 & \Psi \end{bmatrix} \Pi, \qquad \Pi_{i,j} = \begin{cases} 1 & \text{i is odd } j = i, \\ 1 & \text{i is even } j = r + i, \\ 0 & \text{otherwise,} \end{cases}$$



(a) $n \geq d$

(b) $n \leq d$

Fig. 5: Regularized LS problem for all dimension regimes: The problem dimensions are set to $\max(n, d) = 5 \cdot 10^4$ and $\min(n, d) = 10^4$. The noise level is set to $\|w\|_2 \big/ \|Ax_0\|_2 = 10\%$. All techniques are provided with the optimal regularization parameter which is chosen in the same way as Figure 3. The *Inexact* schemes of the `M-IHS` and `Dual M-IHS` use a sketch size $m = 2 \cdot \mathbf{sd}_\lambda(A)$. The primal dual schemes use $m_1 = m_2 = 2 \cdot \mathbf{sd}_\lambda(A)$ except for the green `Primal Dual M-IHS` which uses $m_1 = m_2 = 8 \cdot \mathbf{sd}_\lambda(A)$. The statistical dimension for (a) is $\mathbf{sd}_\lambda(A) = 690$ and for (b) is $\mathbf{sd}_\lambda(A) = 825$. All methods are allowed to do $N = 60$ iterations except for that the `Primal Dual M-IHS` with larger sketch size is allowed to do only 20 iterations. At the end of iterations, the `Primal Dual M-IHS` with larger sketch size reaches the same accuracy level as the one with smaller sketch size, but it require slightly more operations, which suggests that complexity cannot be reduced by choosing smaller sketch sizes below some certain level that depends on the dimensions $(n, d, \mathbf{sd}_\lambda(A))$. The number of inner iterations are restricted by $M = 25$ for all primal dual schemes. Instead of limiting the number of iterations, forcing terms that checks the relative residual error can also be used for the stopping criterion of the inner loop iterations. Lastly, fixed forcing term $\epsilon_{sub} = 0.1$ is used in the `AAb_Solver` for all inexact schemes.

where $\Psi\Lambda\Psi^T$ is the Spectral decomposition of $(\widehat{A}^T\widehat{S}^T\widehat{S}\widehat{A})^{-1}$ and λ_i is the i^{th} eigenvalue. The characteristic polynomials of each block is

$$u^2 - (1 + \beta - \alpha\lambda_i)u + \beta = 0, \quad \forall i \in [r].$$

If the following condition holds

$$\beta \geq (1 - \sqrt{\alpha\lambda_i})^2, \quad \forall i \in [r], \tag{5.1}$$

then both of the roots are imaginary and both have a magnitude $\sqrt{\beta}$ for all λ's. In this case, all linear dynamical systems driven by above characteristic polynomial will be in the under-damped regime and the contraction rate of the transformation T, through all directions, not just one of them, will be exactly $\sqrt{\beta}$. If the condition in (5.1) is not satisfied for a λ_j with $j \in [r]$, then the linear dynamical system corresponding to λ_j will be in the over-damped regime and the contraction rate in the direction through the eigenvector corresponding to this over-damped system will be smaller compared to the others. As a result, the overall algorithm will be slowed down substantially (see [43] for details). If the condition in (3.5) of Theorem 3.1 holds,

$$\left\|\widehat{A}^T\widehat{S}^T\widehat{S}\widehat{A}-I_r\right\|_2=\left\|U_1^TS^TSU_1+U_2^TU_2-I_r\right\|_2=\left\|U_1^TS^TSU_1-U_1^TU_1\right\|_2\leq\epsilon,$$

then

$$\sup_{\|v\|_2=1} v^T\widehat{A}^T\widehat{S}^T\widehat{S}\widehat{A}v\leq 1+\epsilon \quad \text{and} \quad \inf_{\|v\|_2=1} v^T\widehat{A}^T\widehat{S}^T\widehat{S}\widehat{A}v\geq 1-\epsilon,$$

and

$$\underset{i\in[r]}{\text{maximize}}\ \lambda_i\leq\frac{1}{1-\epsilon} \quad \text{and} \quad \underset{i\in[r]}{\text{minimize}}\ \lambda_i\geq\frac{1}{1+\epsilon}.$$

Consequently, the condition in (5.1) can be satisfied for all λ_i's by the following choice of β that maximizes the convergence rate over step size α

$$\sqrt{\beta^*}=\underset{\alpha}{\text{minimize}}\left(\max\left\{1-\frac{\sqrt{\alpha}}{\sqrt{1+\epsilon}},\ \frac{\sqrt{\alpha}}{\sqrt{1-\epsilon}}-1\right\}\right)=\frac{\epsilon}{1+\sqrt{1-\epsilon^2}},$$

where the minimum is achieved at $\alpha^*=\frac{4(1-\epsilon^2)}{(\sqrt{1+\epsilon}+\sqrt{1-\epsilon})^2}=(1-\beta^*)\sqrt{1-\epsilon^2}$ as claimed.

The Proof for Dual M-IHS:. The proof for the dual counter-part is almost same as the proof for `M-IHS` except for the following modifications. Let

$$\widehat{A}^T=\begin{bmatrix}V\Sigma D\\ \sqrt{\lambda}UD\end{bmatrix}=\begin{bmatrix}U_1\\ U_2\end{bmatrix},\ \widehat{A}\widehat{A}^T=I_n \ \text{ and } \ \widehat{S}\widehat{A}^T=\begin{bmatrix}SV\Sigma D\\ \sqrt{\lambda}UD\end{bmatrix}=\begin{bmatrix}SU_1\\ U_2\end{bmatrix},$$

so that U_1 is the first d rows of an orthogonal basis for $\begin{bmatrix}A^T\\ \sqrt{\lambda}I_n\end{bmatrix}$ as required by the theorem. We have the following equality:

$$\frac{1}{2}\left\|A^T\nu\right\|_2^2+\frac{1}{2}\left\|\nu\right\|_2^2-\langle b,\ \nu\rangle=\frac{1}{2}\left\|\widehat{A}^Tw\right\|_2^2-\langle DU^Tb,\ w\rangle,$$

for $\left\{\forall\nu\in\mathbb{R}^n \mid w=D^{-1}U^T\nu\right\}$. Thus, following HS sub-problem is equivalent to (3.4)

$$\Delta w^{i+1}=\underset{y}{\text{argmin}}\ \left\|\widehat{S}\widehat{A}^Tw\right\|_2^2-2\langle DU^Tb-\widehat{A}\widehat{A}^Tw^i,\ w\rangle.$$

We can analyze the following bipartite transformation to figure out the convergence properties of the `Dual M-IHS`

$$\begin{bmatrix}w^{i+1}-w^*\\ w^i-w^*\end{bmatrix}=\underbrace{\begin{bmatrix}(1+\beta)I_n-\alpha(\widehat{A}\widehat{S}^T\widehat{S}\widehat{A}^T)^{-1} & -\beta I_n\\ I_n & \mathbf{0}\end{bmatrix}}_{T}\begin{bmatrix}w^i-w^*\\ w^{i-1}-w^*\end{bmatrix},$$

where $w^*=D^{-1}U^Tv^*$. The rest of the proof can be completed straightforwardly by following the same analysis steps as in the proof for the `M-IHS` case.

5.2. Analysis of the Condition in Theorem 3.1. Theorem 3.1 is valid if the sketch matrix S satisfies the following condition:

$$\left\|U_1^TS^TSU_1-U_1^TU_1\right\|_2\leq\left\|U_1^TS^TSU_1-U_1^TU_1\right\|_F\leq\epsilon, \tag{5.2}$$

where

$$\|U_1\|_F^2 = \|U\Sigma D\|_F^2 = \left\|\Sigma(\Sigma^2+\lambda I_r)^{-1/2}\right\|_F^2 = \sum_{i=1}^{r}\frac{\sigma_i^2}{\sigma_i^2+\lambda} = \mathbf{sd}_\lambda(A),$$

and $\|U_1\|_2^2 = \frac{\sigma_1^2}{\sigma_1^2+\lambda} \approx 1$ for a properly chosen regularization parameter λ. In this subsection before giving the proof of Corollary 3.2, we summarize some of the current results and tools from the random matrix theory literature that is used to prove the condition in (5.2) for different applications such as low rank approximation, matrix ridge regression, kernel ridge regression, ℓ_p regression, k-means clustering.

If the sketch matrix S is drawn from a randomized distribution $\mathcal{D}$ over matrices $\mathbb{R}^{m\times n}$, then the condition in (5.2) can be met with a desired level of probability.

Definition 5.1. *(JL Lemma [30]) For any integer $n > 0$, $\epsilon, \delta < 1/2$, the distribution $\mathcal{D}$ over $S \in \mathbb{R}^{m\times n}$ is called as $(\epsilon,\delta)-$JL distribution if for any $x \in \mathbb{R}^n$,*

$$\mathbb{P}_{S\sim\mathcal{D}}\left[(1-\epsilon)\|x\|_2 \le \|Sx\|_2 \le (1+\epsilon)\|x\|_2\right] > 1-\delta,$$

holds for $m = \Theta(\epsilon^{-2}\log(1/\delta))$.

The sketch size given in the definition is optimal [33]. By assuming x is unit norm, the probability in Definition 5.1 can be bounded by the Markov Inequality [54, 39]:

$$\begin{aligned}\mathbb{P}_{S\sim\mathcal{D}}\left[\big|\,\|Sx\|_2-1\big| > \epsilon\right] &= \mathbb{P}_{S\sim\mathcal{D}}\left[\big|\,\|Sx\|_2-1\big|^\ell > \epsilon\right]\\ &\le \epsilon^{-\ell}\cdot\mathbb{E}_{S\sim\mathcal{D}}\left[\big|\,\|Sx\|_2-1\big|^\ell\right],\end{aligned}$$

which leads to a fundamental property:

Definition 5.2. *(JL-Moment Property Definition 6.1 in [33]) Distribution $\mathcal{D}$ over $\mathbb{R}^{m\times n}$ has (ϵ,δ,ℓ)-JL moment property if for all the unit norm vectors $x \in \mathbb{R}^n$*

$$\mathbb{E}_{S\sim\mathcal{D}}\left[\big|\,\|Sx\|_2^2-1\big|^\ell\right] \le \epsilon^\ell\cdot\delta.$$

Having JL-moment property is a stronger condition than being a JL-distribution, but if a distribution $\mathcal{D}$ is a $(\epsilon, e^{-\Omega(\epsilon^2 m)})$-JL distribution, then it is called as *Strong JL Distribution* [32] and it automatically verifies the $(\epsilon, e^{-\Omega(\epsilon^2 m)}, \min(\epsilon^2 m,\ \epsilon m))$-JL moment property [33]. The JL-moment property is important to reach our goal due to the following property called the Approximate Matrix Multiplication (AMM):

Theorem 5.3. *(AMM Property, Theorem 6.2 in [33]) Given $\epsilon,\delta \in (0,1/2)$, let $\mathcal{D}$ be any distribution over matrices in $\mathbb{R}^{m\times n}$ with the (ϵ,δ,ℓ)-JL moment property for some $\ell \ge 2$. Then for any $A, B \in \mathbb{R}^{n\times d}$ real matrices,*

$$\mathbb{P}_{S\sim\mathcal{D}}\left[\left\|A^TS^TSB - A^TB\right\|_F < 3\epsilon\|A\|_F\|B\|_F\right] > 1-\delta.$$

If the sketch matrix satisfies the JL-moment property, setting $A = B = U_1$ in the AMM property gives the condition in 5.2, but except for the sparse sketch matrices with single non-zero element in each column, the AMM property that is based on the JL-Distribution is too restrictive for the other sketch types mentioned in Corollary 3.2 to satisfy the condition. For them, the $\ell 2$ subspace embeddings, which is introduced in [50], can be used.

DEFINITION 5.4. *(Oblivious $\ell 2$ Subspace Embedding (OSE)) If a distribution $\mathcal{D}$ satisfies*

$$\mathbb{P}_{S\sim\mathcal{D}}\left[\left\|US^TSU - I\right\|_2 > \epsilon\right] < \delta,$$

with $U \in \mathbb{R}^{n\times k},\ U^TU = I_k,\ S \in \mathbb{R}^{m\times n}$, then it is called (ϵ, δ, k)-OSE.

The JL-distributions are special case of (ϵ, δ, k)-OSE with $k = n$. While the JL distributions projects the whole $\mathbb{R}^n$, the OSE distributions embed only a subspace $\mathbf{span}(U) \subseteq \mathbb{R}^n$. Consequently, the subspace embeddings provide a more general moment property.

DEFINITION 5.5. *(OSE Moment Property [15]) A distribution $\mathcal{D}$ over $S \in \mathbb{R}^{m\times n}$ has $(\epsilon, \delta, k, \ell)$-OSE Moment Property, if for all matrices $U \in \mathbb{R}^{n\times k}$ with orthonormal columns,*

$$\mathbb{E}_{S\sim\mathcal{D}}\left[\left\|U^TS^TSU - I_k\right\|_2^\ell\right] < \epsilon^{-\ell}\cdot\delta.$$

The JL moment property is the special case of the OSE moment property with $k = n$ and the relation between them is given in the following lemma.

LEMMA 5.6. *(Lemma 4 of [15]) If a distribution $\mathcal{D}$ satisfies the (ϵ, δ, ℓ)-JL moment property, then $\mathcal{D}$ satisfies the $(2\epsilon, 9^k\delta, k, \ell)$-OSE moment property.*

By using the OSE Moment property to obtain a generalized AMM property is also possible:

THEOREM 5.7. *((ϵ, δ, k)-AMM Property [15]) If a distribution $\mathcal{D}$ over $S \in \mathbb{R}^{m\times n}$ has the $(\epsilon, \delta, 2k, \ell)$-OSE moment property for some $\delta < 1/2$ and $\ell \geq 2$, then for any A, B:*

$$\mathbb{P}_{S\sim\mathcal{D}}\left[\left\|A^TS^TSB - A^TB\right\|_2 > \epsilon\sqrt{\left(\|A\|_2^2 + \frac{\|A\|_F^2}{k}\right)\left(\|B\|_2^2 + \frac{\|B\|_F^2}{k}\right)}\right] < \delta.$$

Theorem 5.7 breaks the dependence between the error constant ϵ and the sketch size m in the AMM property given in Theorem 5.3. By Theorem 5.7, the sketch sizes can be chosen in accordance with the embedding size k to satisfy the condition in (5.2) as we will show next.

***Proof of Corollary 3.2**, Item (i).* Count Sketch with a single nonzero element in each column and size $m \geq 2/(\epsilon'^2\delta)$ has $(\epsilon', \delta, 2)$-JL moment property [51]. By Theorem 5.3:

$$\left\|U_1S^TSU_1 - U_1^TU_1\right\|_F < 3\epsilon'\left\|U_1\right\|_F^2 = 3\epsilon'\mathbf{sd}_\lambda(A) \leq \epsilon$$

for $\epsilon' = \epsilon/(3\mathbf{sd}_\lambda(A))$. So, condition in (5.2) holds with probability at least $1-\delta$, if $m = O(\mathbf{sd}_\lambda(A)^2/(\epsilon^2\delta))$.

***Proof of Corollary 3.2**, Item (ii).* Combining Theorem 4.2 of [14] and Remark 2 of [15] implies that any sketch matrix drawn from an OSNAP [39] with the conditions given in the item (ii) of Corollary 3.2 satisfies the $(\epsilon', \delta, k, \log(k/\delta))$-OSE moment property, thus the $(\epsilon', \delta, k/2)$-AMM Property. Setting $A = B = U_1$ and $k = \mathbf{sd}_\lambda(A)/2$ in Theorem 5.7 gives:

$$\left\|U_1^TS^TSU_1 - U_1^TU_1\right\|_2 \leq \epsilon'(\|U_1\|_2^2 + 2) \leq 3\epsilon' \leq \epsilon$$

with probability at least $1-\delta$.

Remark 5.8. Based on the lower bounds established for any OSE in [38], the Conjecture 14 in [39] states that any OSNAP with $m = \Omega((k + \log(1/\delta))/\epsilon^2)$ and $s = \Omega(\log(k/\delta)/\epsilon)$ have the $(\epsilon, \delta, k, \ell)$-OSE moment property for $\ell = \Theta(\log(k/\delta))$, an even integer. As a result of the conjecture and Theorem 5.7, the condition in (5.2) can be satisfied with probability at least $1 - \delta$ by using an OSNAP matrix with size $m = \Omega((\mathbf{sd}_\lambda(A) + \log(1/\delta))/\epsilon^2)$ and sparsity $s = \Omega(\log(\mathbf{sd}_\lambda(A)/\delta)/\epsilon)$.

***Proof of Corollary 3.2**, Item (iii).* By Theorem 9 of [15], SRHT with the sketch size given in item *(iii)* has the $(\epsilon', \delta, 2\mathbf{sd}_\lambda(A), \log(\mathbf{sd}_\lambda(A)/\delta))$-OSE moment property and thus by Theorem 5.7, it provides $(\epsilon', \delta, \mathbf{sd}_\lambda(A))$-AMM property. Again, setting $A = B = U_1$ and $k = \mathbf{sd}_\lambda(A)$ produces the desired result.

***Proof of Corollary 3.2**, Item (iv).* The Subgaussian matrices having entries with mean zero and variance $1/m$ satisfy the Definition 5.1 with optimal sketch size. Also, they have the $(\epsilon/2, \delta, \Theta(\log(1/\delta)))$-JL moment property [32]. Thus by Lemma 5.6 such matrices have $(\epsilon, \delta, k, \Theta(k + \log(1/\delta)))$-OSE moment property for $\delta < 9^{-k}$, which means $m = \Omega(k/\epsilon^2)$. Again, by setting $A = B = U_1$ and $k = \mathbf{sd}_\lambda(A)$ in Theorem 5.7 gives the desired result.

5.3. Convergence proof without regularization. If the LS problem without any regularization is tried to be solved, then we need the following inequality to hold:

$$\left\|U_1^T S^T S U_1 - U_1^T U_1\right\|_2 \leq \epsilon \xrightarrow[n \geq d]{\lambda \to 0} \left\|U^T S^T S U - I_r\right\|_2 \leq \epsilon,$$

which can be satisfied with a constant probability, if the sketch matrix is drawn from a distribution $\mathcal{D}$ that has the OSE Moment property defined in Definition 5.5. Then, the bound for the sketch size can be derived by using the other results presented above. The details of how to obtain the above inequality can be found in [42].

5.4. Proof of Corollary 3.4. Consider the regularized LS solution with parameter λ and the Truncated SVD solution with parameter $\lceil \mathbf{sd}_\lambda(A) \rceil$:

$$x^* = \sum_{i=1}^{r} \frac{\sigma_i^2}{\sigma_i^2 + \lambda} \frac{u_i^T b}{\sigma_i} v_i \quad \text{and} \quad x^\dagger = \sum_{i=1}^{\lceil \mathbf{sd}_\lambda(A) \rceil} \frac{u_i^T b}{\sigma_i} v_i$$

where u_i's and v_i's are columns of U and V matrices in the SVD. The Tikhonov regularization with the closed form solution is preferred in practise to avoid the high computational cost of the SVD. The filtering coefficients of the Tikhonov regularization, $\frac{\sigma_i^2}{\sigma_i^2 + \lambda}$, become very close to the binary filtering coefficients of the TSVD, as the decay rate of the singular values of A increases. In these cases, both the solution x^* and $x^\dagger$ become almost indistinguishable. Thereby, the diagonal matrix, ΣD, which is used in the proof of Corollary 3.2 can be approximated by the diagonal matrix Π where

$$\Pi_{ii} = \begin{cases} 1 & \text{if } i \leq \mathbf{sd}_\lambda(A) \leq r \\ 0 & \textit{otherwise} \end{cases},$$

which is equivalent to replace the Tikhonov coefficients by the binary TSVD coefficients. Then, we have

$$\begin{aligned}\left(\widehat{A}^T\widehat{S}^T\widehat{S}\widehat{A}\right)^{-1} &= \left(D\Sigma U^T S^T S U\Sigma D + \lambda D^2\right)^{-1} \\ &\approx \left(\Pi(SU)^T(SU)\Pi + I_r - \Pi\right)^{-1} = \left[\begin{array}{c|c} \overline{S}^T\overline{S} & \mathbf{0} \\ \hline \mathbf{0} & I_{(r-\mathbf{sd}_\lambda(A))}\end{array}\right],\end{aligned}$$

where $\overline{S} = SU\Pi \in \mathbb{R}^{m\times \mathbf{sd}_\lambda(A)}$ has the same distribution as S, since $U\Pi$ is an orthonormal transformation. By the Marchenko Pastur Law, the minimum and the maximum eigenvalues of this approximation converge to $\left(1 \pm \sqrt{\frac{\mathbf{sd}_\lambda(A)}{m}}\right)^{-2}$ as $m\to\infty$ and while $\mathbf{sd}_\lambda(A)/m$ remains constant [19, 20]. The rest of the proof follows from the analysis given in subsection 5.1.

6. The proposed technique to approximately solve the sub-problems.

The linear sub-problems in the form of $(A^TA+\lambda I)x = b$, whose solutions are required by all four `M-IHS` variants, can be approximately solved by using the *bidiag2* procedure described in [44]. The *bidiag2* procedure produces an upper bidiagonal matrix as following:

$$P_k^T A V_k = R_k = \begin{bmatrix}\rho_1 & \theta_2 & & \\ & \ddots & \ddots & \\ & & \rho_{k-1} & \theta_k \\ & & & \rho_k\end{bmatrix} \in \mathbb{R}^{k\times k}, \qquad \begin{array}{l} P_k \in \mathbb{R}^{n\times k},\ V_k\in\mathbb{R}^{d\times k} \\ P_k^TP_k = V_k^TV_k = I_k\end{array} .$$

The upper bidiagonal decomposition R_k is computed by using the Lanczos-like three term recurrence:

$$\begin{array}{r} AV_k = P_kR_k \\ A^TP_k = V_kR_k^T + \theta_{k+1}v^{k+1}e_k^T \end{array} \Longrightarrow \begin{array}{rll} Av^1 &= \rho_1 p^1, & \\ A^Tp^j &= \rho_j v^j + \theta_{j+1}v^{j+1} & j\le k, \\ Av^j &= \theta_j p^{j-1} + \rho_j p^j, & j \le k,\end{array}$$

where θ_j's and ρ_j's are chosen so that $\|v^j\|_2 = \|p^j\|_2 = 1$, respectively. Noting that P_k and V_k are not needed to be orthogonal in `AAb_Solver`, therefore we do not need any reorthogonalization steps. As different from the LSQR, we choose $\theta_1 v^1 = b$ with $\theta_1 = \|b\|_2$ so that the columns of matrix V_k constitute an orthonormal basis for the k-th order Krylov Subspace:

$$\text{span}\{v^1,\ldots,v^k\} = \mathcal{K}_k(A^TA,\ b) = \mathcal{K}_k(A^TA+\mu I_d,\ b),\ \forall\mu\in\mathbb{R}_+.$$

Regularization does not affect this property since the Krylov Subspace is invariant under a constant shift. In the k-th iteration of the proposed sub-solver, `AAb_Solver`, let the solution estimate of the linear system be $x^k = V_ky^k$ for some vector $y^k\in\mathbb{R}^k$, i.e., $x^k\in\mathcal{K}_k(A^TA,\ b)$, then we have:

$$(A^TA+\lambda I_d)V_ky^k = b \Longrightarrow \overline{R}_ky^k = \overline{R}_k^{-T}V_k^Tb = \theta_1\overline{R}_k^{-T}e_1 = f^k = [\phi_1,\ldots,\phi_k]^T,$$

where $\overline{R}_k$ is obtained by applying a sequence of Givens rotation on $\begin{bmatrix} R_k \\ \sqrt{\lambda}I_k\end{bmatrix}$ in order to eliminate the sub-diagonal elements coming from the regularization [22]. One instance of this procedure is

$$\begin{bmatrix} \bar{\rho}_k & \theta_{k+1} \\ 0 & \rho_{k+1} \\ 0 & 0 \\ 0 & \sqrt{\lambda} \end{bmatrix} \rightarrow \begin{bmatrix} \bar{\rho}_k & c_k\theta_{k+1} \\ 0 & \rho_{k+1} \\ 0 & 0 \\ 0 & \overline{\lambda}_{k+1} \end{bmatrix} \rightarrow \begin{bmatrix} \bar{\rho}_k & \bar{\theta}_{k+1} \\ 0 & \bar{\rho}_{k+1} \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \xrightarrow[\text{iteration}]{\text{next}} \begin{bmatrix} \bar{\rho}_{k+1} & \theta_{k+2} \\ 0 & \rho_{k+2} \\ 0 & 0 \\ 0 & \sqrt{\lambda} \end{bmatrix},$$

where

$$\begin{aligned} c_k &= \rho_k/\bar{\rho}_k, & s_k &= \overline{\lambda}_k/\bar{\rho}_k, \\ \bar{\theta}_{k+1} &= c_k\theta_{k+1}, & \overline{\lambda}_{k+1}{}^2 &= \lambda + (s_k\theta_{k+1})^2, \\ & & \bar{\rho}_{k+1} \quad &= \quad \sqrt{\rho_{k+1}{}^2 + \overline{\lambda}_{k+1}{}^2}. \end{aligned}$$

Since $\overline{R}_k$ is an upper bidiagonal matrix, the inverse always exists and ϕ_j can be found analytically as:

$$\phi_1 = \frac{\theta_1}{\bar{\rho}_1} \quad \text{and} \quad \phi_k = -\phi_{k-1}\frac{\bar{\theta}_k}{\bar{\rho}_k}.$$

Furthermore, the solution at the k-th iteration, $x^k = V_k\overline{R}_k^{-1}f^k$, can be obtain without computing any inversions by using the forward substitution. Define $D_k = V_k\overline{R}_k^{-1}$:

$$\left.\begin{aligned} [D_{k-1},\ d^k]\begin{bmatrix} \overline{R}_{k-1} & e_{k-1}\bar{\theta}_k \\ 0 & \bar{\rho}_k \end{bmatrix} &= [V_{k-1},\ v^k] \\ D_{k-1}\overline{R}_{k-1} &= V_{k-1} \\ \bar{\theta}_k d^{k-1} + \bar{\rho}_k d^k &= v^k \end{aligned}\right\} \begin{aligned} d^k &= (v^k - \bar{\theta}_k d^{k-1})/\bar{\rho}_k \\ x^k &= x^{k-1} + \phi_k d^k. \end{aligned}$$

and the relative residual error, to use as stopping criterion, can be found as:

$$\begin{aligned} \|A^TAx^k + \lambda x^k - b\|_2^2 &= \|A^TAV_ky^k + \lambda V_ky^k - b\|_2^2 = \|A^TP_k\overline{R}_ky^k - b\|_2^2 \\ &= \|\left(V_k\overline{R}_k^T + \theta_{k+1}v^{k+1}e_k^T\right)\overline{R}_ky^k - b\|_2^2 \\ &\overset{(i)}{=} \|\overline{R}_k^T\overline{R}_ky^k - V_k^Tb\|_2^2 + \|\theta_{k+1}v^{k+1}e_k^T\overline{R}_ky^k - \left(I - V_kV_k^T\right)b\|_2^2 \\ &= \left|\phi_k\bar{\theta}_{k+1}\right| = |\phi_{k+1}\bar{\rho}_{k+1}|. \end{aligned}$$

The first norm in (i) is zero since the linear system is always consistent. The second term in the second norm is also zero, since $b \in \text{span}(V_k)$ by the initial choice of $\theta_1v^1 = b$. By definition, $f^k = \overline{R}_ky^k$ gives the final results. The overall algorithm is given in Algorithm 6.1. The `AAb_Solver` also is a Krylov Subspace method, therefore, it finds the solution at most in $\min(n, d, m)$ iterations in the exact arithmetic, but far fewer number of iterations is sufficient for our purpose.

7. Conclusions.

We studied the `M-IHS` techniques, a group of solvers for large scale LS problems, which are obtained by incorporating the Heavy Ball Acceleration into the iterations of the IHS algorithm. We obtained the optimal fixed momentum parameters that maximize the convergence rate by the help of the results in random matrix theory. We examined the effect of $\ell 2$ norm regularization on the optimal momentum parameters. We showed that the `M-IHS` variants can be used for any dimension regimes if the statistical dimension is sufficiently smaller than the dimensions of the coefficient matrix and we obtained lower bounds on the sketch size for several randomized distributions in order to get a pre-determined convergence rate with a

Algorithm 6.1 `AAb_Solver` (for problems in the form of $(A^TA + \lambda I)x = b$)

```
Input: $A, b, \lambda, \epsilon$                    ▷ choose $\rho$ and $\theta$ to make $\|p\| = \|v\| = 1$
$\theta_1 v = b$
$\rho p = Av$

$\bar{\rho} = \sqrt{\rho^2 + \lambda}, \quad c = \rho/\bar{\rho}, \quad s = \sqrt{\lambda/\bar{\rho}}, \quad \phi = \theta_1/\bar{\rho}$
$d = v/\bar{\rho}$
$x = \phi d$
while $t \geq \epsilon$ do
  $\theta v := A^T p - \rho v$
  $\rho p := Av - \theta p$

  $\bar{\lambda}^2 := \lambda + (s\theta)^2, \quad \bar{\theta} = c\theta$
    $\bar{\rho} := \sqrt{\rho^2 + \bar{\lambda}^2}, \quad c = \rho/\bar{\rho}, \quad s = \bar{\lambda}/\bar{\rho}$

    $d := (v - \bar{\theta} d)/\bar{\rho}$
    $\phi := -\phi\bar{\theta}/\bar{\rho}$
    $x := x + \phi d$
      $t = |\phi\bar{\rho}|/\theta_1$
end while
```

constant probability. The `M-IHS` variants outperform the best existing randomized preconditioning techniques such as the Blendenpik, A-IHS and LSRN in all numerical experiments. Moreover, the `M-IHS` variants do not use any inner products during the iterations and they are shown to be faster than CS-based randomized preconditioning algorithms, hence `M-IHS` variants can be substituted for the CS-based techniques in parallel or distributed architectures. Furthermore, by the proposed *Inexact* schemes, the `M-IHS` variants eliminate all, sketched or not, matrix decompositions and inversions; thus they automatically speed up in the applications with sparse data matrices or in the applications with linear operators that allow fast matrix-vector multiplications.

REFERENCES

[1] H. Avron, K. L. Clarkson, and D. P. Woodruff, *Sharper bounds for regularized data fitting*, arXiv preprint arXiv:1611.03225, (2016).

[2] H. Avron, K. L. Clarkson, and D. P. Woodruff, *Faster kernel ridge regression using sketching and preconditioning*, SIAM J. Matrix Anal. Appl., 38 (2017), pp. 1116–1138.

[3] H. Avron, P. Maymounkov, and S. Toledo, *Blendenpik: Supercharging lapack's least-squares solver*, SIAM J. Sci. Comput., 32 (2010), pp. 1217–1236.

[4] H. Avron and S. Toledo, *Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix*, J. ACM, 58 (2011), p. 8.

[5] R. Barrett, M. W. Berry, T. F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H. Van der Vorst, *Templates for the solution of linear systems: building blocks for iterative methods*, vol. 43, SIAM, 1994.

[6] B. Bartan and M. Pilanci, *Polar coded distributed matrix multiplication*, arXiv preprint arXiv:1901.06811, (2019).

[7] A. S. Berahas, R. Bollapragada, and J. Nocedal, *An investigation of newton-sketch and subsampled newton methods*, arXiv preprint arXiv:1705.06211, (2017).

[8] Å. Björck, *Numerical methods for least squares problems*, SIAM, 1996.

[9] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al., *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Found. Trends

Mach. Learn., 3 (2011), pp. 1–122.

[10] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge university press, 2004.

[11] S. Bubeck et al., *Convex optimization: Algorithms and complexity*, Found. Trends Mach. Learn., 8 (2015), pp. 231–357.

[12] A. Chambolle, *Continuous optimization, an introduction*, Lecture Notes, (2016).

[13] A. Chowdhury, J. Yang, and P. Drineas, *An iterative, sketching-based framework for ridge regression*, in International Conference on Machine Learning, 2018, pp. 988–997.

[14] M. B. Cohen, *Nearly tight oblivious subspace embeddings by trace inequalities*, in Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms, SIAM, 2016, pp. 278–287.

[15] M. B. Cohen, J. Nelson, and D. P. Woodruff, *Optimal approximate matrix product in terms of stable rank*, arXiv preprint arXiv:1507.02268, (2015).

[16] D. L. Donoho et al., *Compressed sensing*, IEEE Trans. Inform. Theory, 52 (2006), pp. 1289–1306.

[17] P. Drineas and M. W. Mahoney, *Randnla: randomized numerical linear algebra*, Comm. ACM, 59 (2016), pp. 80–90.

[18] P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós, *Faster least squares approximation*, Numer. Math., 117 (2011), pp. 219–249.

[19] A. Edelman, *Eigenvalues and condition numbers of random matrices*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 543–560.

[20] A. Edelman and Y. Wang, *Random matrix theory and its innovative applications*, in Advances in Applied Mathematics, Modeling, and Computational Science, Springer, 2013, pp. 91–116.

[21] S. C. Eisenstat and H. F. Walker, *Choosing the forcing terms in an inexact newton method*, SIAM J. Sci. Comput., 17 (1996), pp. 16–32.

[22] L. Elden, *Algorithms for the regularization of ill-conditioned least squares problems*, BIT, 17 (1977), pp. 134–145.

[23] D. C.-L. Fong and M. Saunders, *Lsmr: An iterative algorithm for sparse least-squares problems*, SIAM J. Sci. Comput., 33 (2011), pp. 2950–2971.

[24] G. H. Golub, M. Heath, and G. Wahba, *Generalized cross-validation as a method for choosing a good ridge parameter*, Technometrics, 21 (1979), pp. 215–223.

[25] A. Greenbaum, *Iterative methods for solving linear systems*, vol. 17, SIAM, 1997.

[26] M. H. Gutknecht and S. Röllin, *The chebyshev iteration revisited*, Parallel Comput., 28 (2002), pp. 263–283.

[27] P. C. Hansen, *Regularization tools: a matlab package for analysis and solution of discrete ill-posed problems*, Numer. Algorithms, 6 (1994), pp. 1–35.

[28] P. C. Hansen, *Discrete inverse problems: insight and algorithms*, vol. 7, SIAM, 2010.

[29] R. Hunger, *Floating point operations in matrix-vector calculus*, Technical University of Munich, 2007.

[30] W. B. Johnson and J. Lindenstrauss, *Extensions of lipschitz mappings into a hilbert space*, Contemp. Math., 26 (1984), p. 1.

[31] E. Jonas, Q. Pu, S. Venkataraman, I. Stoica, and B. Recht, *Occupy the cloud: Distributed computing for the 99%*, in Proceedings of the 2017 Symposium on Cloud Computing, ACM, 2017, pp. 445–451.

[32] D. Kane, R. Meka, and J. Nelson, *Almost optimal explicit johnson-lindenstrauss families*, in Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, Springer, 2011, pp. 628–639.

[33] D. M. Kane and J. Nelson, *Sparser johnson-lindenstrauss transforms*, J. ACM., 61 (2014), p. 4.

[34] M. E. Kilmer and D. P. O'Leary, *Choosing regularization parameters in iterative methods for ill-posed problems*, SIAM J. Matrix Anal. Appl., 22 (2001), pp. 1204–1221.

[35] M. W. Mahoney et al., *Randomized algorithms for matrices and data*, Found. Trends Theor. Mach. Learn, 3 (2011), pp. 123–224.

[36] X. Meng, M. A. Saunders, and M. W. Mahoney, *Lsrn: A parallel iterative solver for strongly over-or underdetermined systems*, SIAM J. Sci. Comput., 36 (2014), pp. C95–C118.

[37] S. G. Nash, *A survey of truncated-newton methods*, J. Comput. Appl. Math., 124 (2000), pp. 45–59.

[38] J. Nelson and H. L. NguyÅn, *Lower bounds for oblivious subspace embeddings*, in International Colloquium on Automata, Languages, and Programming, Springer, 2014, pp. 883–894.

[39] J. Nelson and H. L. Nguyên, *Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings*, in 2013 IEEE 54th Annual Symposium on Foundations of Computer

Science, IEEE, 2013, pp. 117–126.

[40] Y. Nesterov, *Introductory lectures on convex programming volume i: Basic course*, Lecture notes, 3 (1998), p. 5.

[41] J. Nocedal and S. Wright, *Numerical optimization*, Springer Science & Business Media, 2006.

[42] I. K. Ozaslan, M. Pilanci, and O. Arikan, *Iterative hessian sketch with momentum*, 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (2019).

[43] B. Odonoghue and E. Candes, *Adaptive restart for accelerated gradient schemes*, Found. Comput. Math., 15 (2015), pp. 715–732.

[44] C. C. Paige and M. A. Saunders, *Lsqr: An algorithm for sparse linear equations and sparse least squares*, ACM Trans. Math. Software, 8 (1982), pp. 43–71.

[45] M. Pilanci and M. J. Wainwright, *Randomized sketches of convex programs with sharp guarantees*, IEEE Trans. Inform. Theory, 61 (2015), pp. 5096–5115.

[46] M. Pilanci and M. J. Wainwright, *Iterative hessian sketch: Fast and accurate solution approximation for constrained least-squares*, J. Mach. Learn. Res., 17 (2016), pp. 1842–1879.

[47] M. Pilanci and M. J. Wainwright, *Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence*, SIAM J. Optim., 27 (2017), pp. 205–245.

[48] B. T. Polyak, *Some methods of speeding up the convergence of iteration methods*, Comput. Math. Math. Phys, 4 (1964), pp. 1–17.

[49] V. Rokhlin and M. Tygert, *A fast randomized algorithm for overdetermined linear least-squares regression*, Proc. Nat. Acad. Sci. India Sect. A, 105 (2008), pp. 13212–13217.

[50] T. Sarlos, *Improved approximation algorithms for large matrices via random projections*, in 2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06), IEEE, 2006, pp. 143–152.

[51] M. Thorup and Y. Zhang, *Tabulation-based 5-independent hashing with applications to linear probing and second moment estimation*, SIAM J. Comput., 41 (2012), pp. 293–331.

[52] C. R. Vogel, *Computational methods for inverse problems*, vol. 23, SIAM, 2002.

[53] J. Wang, J. D. Lee, M. Mahdavi, M. Kolar, N. Srebro, et al., *Sketching meets random projection in the dual: A provable recovery algorithm for big and high-dimensional data*, Electron. J. Stat., 11 (2017), pp. 4896–4944.

[54] D. P. Woodruff et al., *Sketching as a tool for numerical linear algebra*, Found. Trends Theor. Comput. Sci., 10 (2014), pp. 1–157.

[55] L. Zhang, M. Mahdavi, R. Jin, T. Yang, and S. Zhu, *Recovering the optimal solution by dual random projection*, in Conference on Learning Theory, 2013, pp. 135–157.

8. Appendices: Discussion on the Proposed Error Upper Bound for Iterations of Primal Dual Algorithms in [53]. In this appendix, we provide details of a critical discussion on the steps of the derivation that leads to an error upper bound for the iterations of the primal dual algorithms given in [53]. First, we provide a short list of minor issues that can easily be corrected.

1. During the initialization stage in *Line 2* of both *Algorithm 4* in page 4097 and *Algorithm 5* in page 4098, the residual error vector $\mathbf{r}^{(0)}$ must be set to $-\lambda\mathbf{y}$ instead of $-\mathbf{y}$, otherwise iterates of the both of the algorithms diverge from the optimal solution.
2. During the initialization stage in *Line 2* of *Algorithm 7* in 4912, the dual residual error vector $\mathbf{r}_{\text{Dual}}^{(0)}$ must be set to $-\lambda\mathbf{y}$ instead of $-\mathbf{y}$ and during the initialization stage of the inner loop iterations in *Line 15*, the primal residual error vector $\mathbf{r}_{\text{P}}^{(0)}$ must be set to $-\mathbf{R}^T\mathbf{X}^T\mathbf{r}_{\text{D}}^{(t+1)}$; otherwise iterates of the algorithm diverges from the optimal solution. The MATLAB codes provided in the link includes these corrections.

In addition to the above mentioned minor issues, there are some major issues as well. Unfortunately, we could not obtain corrective actions on these major issues as we could have done on the minor issues mentioned above. Therefore, a lower bound on the number of inner loop iterations, that guarantee a certain rate of convergence at the main loop, is still an open question for the primal dual algorithms. In the remaining of this appendix, we will provide steps of the derivation presented in [53], along with our critical remarks on their validity.

Consider the following A-IHS updates

$$\widehat{\mathbf{w}}^{t+1} = \widehat{\mathbf{w}}^{t} + \widehat{\mathbf{u}}^{t}.$$

We are going to use exactly the same notation as [53] except for that *HS* subscript for the A-IHS iterates are omitted. In the primal dual algorithms, instead of exact sequence $\{\widehat{\mathbf{w}}^t\}$, a sequence $\{\widetilde{\mathbf{w}}^t\}$ is obtained due to the approximate minimizers that are used in place of $\widehat{\mathbf{u}}^t$. Consequently while the sequence $\{\widehat{\mathbf{w}}^t\}$ is obtained after t exact iterations of the A-IHS algorithm, sequence $\{\widetilde{\mathbf{w}}^t\}$ is obtained after t primal dual iterations in each of which k inner loop updates are used to approximate $\widehat{\mathbf{u}}^t$'s. The details of the inner and outer loops can be found in Algorithm 7 of [53]. The aim of the *Theorem 9* is to establish an upper bound for

$$\left\|\widetilde{\mathbf{w}}^{t+1} - \mathbf{w}^*\right\|_{\mathbf{X}}$$

where $\mathbf{w}^*$ is the true minimizer of the primal objective function. The triangle inequality and the convergence rate of the A-IHS that is established in *Theorem 2* of [53] is used to find an upper bound for this error norm:

$$\begin{aligned} \left\|\widetilde{\mathbf{w}}^{t+1} - \mathbf{w}^*\right\|_{\mathbf{X}} &\leq \left\|\widehat{\mathbf{w}}^{t+1} - \mathbf{w}^*\right\|_{\mathbf{X}} + \left\|\widetilde{\mathbf{w}}^{t+1} - \widehat{\mathbf{w}}^{t+1}\right\|_{\mathbf{X}} \\ &\leq \alpha^t \left\|\mathbf{w}\right\|_{\mathbf{X}} + \left\|\widetilde{\mathbf{w}}^{t+1} - \widehat{\mathbf{w}}^{t+1}\right\|_{\mathbf{X}}, \end{aligned} \tag{8.1}$$

where $\alpha = \frac{C_0\sqrt{\mathbb{W}^2(\mathbf{X}\mathbb{R}^p \cap \mathcal{S}^{n-1})\log(1/\delta)}}{1-C_0\sqrt{\mathbb{W}^2(\mathbf{X}\mathbb{R}^p \cap \mathcal{S}^{n-1})\log(1/\delta)}}$. At this point a new iterate, $\overline{\mathbf{w}}^{t+1}$, is introduced, which is the result of one exact step of the IHS initialized at $\widetilde{\mathbf{w}}^t$. The inner loop iterations at the t-th outer (main) loop iteration of the primal dual iterations are expected to converge $\overline{\mathbf{w}}^{t+1}$. Therefore,

$$\begin{aligned} \left\|\widetilde{\mathbf{w}}^{t+1} - \widehat{\mathbf{w}}^{t+1}\right\|_{\mathbf{X}} &\leq \left\|\widetilde{\mathbf{w}}^{t+1} - \overline{\mathbf{w}}^{t+1}\right\|_{\mathbf{X}} + \left\|\overline{\mathbf{w}}^{t+1} - \widehat{\mathbf{w}}^{t+1}\right\|_{\mathbf{X}}, \\ \left\|\widetilde{\mathbf{w}}^{t+1} - \overline{\mathbf{w}}^{t+1}\right\|_{\mathbf{X}} &\leq \lambda_{max}\left(\frac{\mathbf{X}^T\mathbf{X}}{n}\right)\beta^k \left\|\overline{\mathbf{w}}^{t+1}\right\|_2 \\ &\leq \lambda_{max}\left(\frac{\mathbf{X}^T\mathbf{X}}{n}\right)\beta^k \left(\left\|\overline{\mathbf{w}}^{t+1} - \mathbf{w}^*\right\|_2 + \left\|\mathbf{w}^*\right\|_2\right) \\ &\leq 2\lambda_{max}\left(\frac{\mathbf{X}^T\mathbf{X}}{n}\right)\beta^k \left\|\mathbf{w}^*\right\|_2, \end{aligned} \tag{8.2}$$

where $\beta = \frac{C_0\sqrt{\mathbb{W}^2(\mathbf{X}^T\mathbb{R}^p \cap \mathcal{S}^{p-1})\log(1/\delta)}}{1-C_0\sqrt{\mathbb{W}^2(\mathbf{X}\mathbb{R}^p \cap \mathcal{S}^{p-1})\log(1/\delta)}}$. The last inequality is not valid unless

$$\left\|\overline{\mathbf{w}}^{t+1} - \mathbf{w}^*\right\|_2 \leq \left\|\mathbf{w}^*\right\|_2.$$

However, particularly during the initial phases of the main iterations this condition can be violated. Therefore, this step of the proof requires a major revision. Assuming that such revision is possible, up to this point, the following is obtained:

$$\left\|\widetilde{\mathbf{w}}^{t+1} - \mathbf{w}^*\right\|_{\mathbf{X}} \leq \alpha^t \left\|\mathbf{w}\right\|_{\mathbf{X}} + 2\lambda_{max}\left(\frac{\mathbf{X}^T\mathbf{X}}{n}\right)\beta^k \left\|\mathbf{w}^*\right\|_2 + \left\|\overline{\mathbf{w}}^{t+1} - \widehat{\mathbf{w}}^{t+1}\right\|_{\mathbf{X}}. \tag{8.3}$$

To proceed for the final form of the upper bound, the following upper bound on the last term of (8.3) is given in [53]:

$$\left\|\overline{\mathbf{w}}^{t+1}-\widehat{\mathbf{w}}^{t+1}\right\|_{\mathbf{X}} \leq \left\|\widetilde{\mathbf{H}}^{-1}\right\|_2 \left\|\widetilde{\mathbf{H}}-\mathbf{H}\right\|_2 \left\|\widetilde{\mathbf{w}}^{t}-\widehat{\mathbf{w}}^{t}\right\|_{\mathbf{X}} \leq \frac{4\lambda_{max}\left(\frac{\mathbf{X}^T\mathbf{X}}{n}\right)}{\lambda} \left\|\widetilde{\mathbf{w}}^{t}-\widehat{\mathbf{w}}^{t}\right\|_{\mathbf{X}},$$

which is a valid bound. Then in [53] the following upper bound is given without necessary justification:

$$\left\|\widetilde{\mathbf{w}}^{t}-\widehat{\mathbf{w}}^{t}\right\|_{\mathbf{X}} \leq 2\lambda_{max}\left(\frac{\mathbf{X}^T\mathbf{X}}{n}\right)\beta^k \left\|\mathbf{w}^*\right\|_2.$$

to reach the final form of the error upper bound:

$$\left\|\widetilde{\mathbf{w}}^{t+1}-\mathbf{w}^*\right\|_{\mathbf{X}} \leq \alpha^t \left\|\mathbf{w}\right\|_{\mathbf{X}} + \frac{10\lambda_{max}^2\left(\frac{\mathbf{X}^T\mathbf{X}}{n}\right)}{\lambda}\beta^k \left\|\mathbf{w}^*\right\|_2.$$

However, this final form of the upper bound is not supported in detail as part of the presented proof. Because of the following major issue, we conclude that the proposed bound remains an unproven conjecture. The bound established for $\left\|\widetilde{\mathbf{w}}^{t+1}-\overline{\mathbf{w}}^{t+1}\right\|_{\mathbf{X}}$ in (8.2) is used to upper bound $\left\|\widetilde{\mathbf{w}}^{t}-\widehat{\mathbf{w}}^{t}\right\|_{\mathbf{X}}$. This is not justified as part of the proof in [53].