Sketching Meets Random Projection in the Dual: A Provable Recovery Algorithm for Big and High-dimensional Data

Jialei Wang* **Jason D. Lee$^\sharp$** **Mehrdad Mahdavi†** **Mladen Kolar*** **Nathan Srebro†***

*UChicago $\sharp$USC $\dagger$TTI-Chicago

Abstract

We provide a unified optimization view of iterative Hessian sketch (IHS) and iterative dual random projection (IDRP). We establish a primal-dual connection between the Hessian sketch and dual random projection, and show that their iterative extensions are optimization processes with *preconditioning*. We develop *accelerated* versions of IHS and IDRP based on this insight together with conjugate gradient descent, and propose a *primal-dual sketch* method that simultaneously reduces the sample size and dimensionality.

1 Introduction

Machine learning is nowadays successfully applied to massive data sets collected from various domains. One of the major challenges in applying machine learning methods to massive data sets is how to effectively utilize available computational resources when building predictive and inferential models, while utilizing data in a statistically optimal way. One approach to tackling massive data sets is via building distributed computer systems and distributed learning algorithms. However, distributed systems may not always be available. Furthermore, the cost of running a distributed system can be much higher than one can afford, making distributed learning unsuitable for all scenarios. An alternative approach is to use the state-of-the-art randomized optimization algorithms to accelerate the training process. For example, many optimization algorithms exist for regularized empirical risk minimization problem, with provably fast convergence and low computational cost per iteration (see [9, 25, 4] for examples). It is worth pointing out at this point that the speed of these optimization methods heavily depends on the condition number of the problem at hand, which is undesirable for many real world problems.

Sketching has emerged as a technique for big data analytics [22]. The idea behind sketching is to approximate the solution of the original problem by solving a sketched, smaller scale problem. For example, sketching has been used to *approximately* solve various large-scale problems, ranging from least square regression and robust regression to low-rank approximation and singular value decomposition (see [7, 11, 10, 1, 22, 18, 23, 14, 13, 5] and references therein). However, one major drawback of sketching is that it is typically not suitable in scenarios where a *highly accurate* solution is needed. To obtain a solution with exponentially smaller approximation error, we often also need to increase the sketching dimension *exponentially.*

Recent work on "iterative sketch", iterative Hessian sketch (IHS) [17] and iterative dual random projection (IDRP) [27], has improved the situation. These methods are able to refine the accuracy of their solution by iteratively solving small scale sketched problem. Hessian sketch [17] is designed to reduce the sample size of the original problem, while dual random projection [27] is proposed to reduce the dimensionality of data. As a consequence, when the sample size and feature dimension are both large, IHS and IDRP still need to solve relatively large-scale subproblems as they can only sketch the problem from one perspective.

In this paper, we address the problem of the *recovery* of optimal solution for big and high-dimensional data by solving small sketched problems of original problem. We make the following contributions. First, we propose an accelerated version of IHS that is computationally as effective as IHS at each iteration, but requires provably fewer number of sketching iterations to reach a certain accuracy. Next, we reveal a primal-

Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

dual connection between IHS and IDRP, that were independently proposed. We show that these two methods are equivalent in the sense that the dual random projection is essentially performing the Hessian sketch in the dual space. This connection allows us to provide a unified analysis of IHS and IDRP, and also develop an accelerated sketching schemas. Finally, we alleviate the computational issues raised by big and high-dimensional learning problems. We propose a *primal-dual sketching* method that can simultaneously reduce the sample size and dimension of the problem, and recover the optimal solution to the original large-scale high-dimensional problem with provable convergence guarantees. More details on theoretical and empirical results can be found in a significantly extended version of this paper [21].

2 Iterative Hessian Sketch and Dual Random Projection

In this section, we review the iterative Hessian sketch proposed in [17] and the iterative dual random projection [27] using our notation. This will serve as the basis of our unified view and further acceleration.

2.1 Hessian Sketch

Consider the following ℓ_2 regularized least-squares (a.k.a. ridge regression) problem:

$$\begin{aligned}&\min_{\mathbf{w}\in\mathbb{R}^p} P(\mathbf{X},\mathbf{y};\mathbf{w})\\ &= \min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{2n}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 \qquad (2.1)\\ &= \min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{2n}\|\mathbf{y}\|_2^2 + \frac{1}{2n}\|\mathbf{X}\mathbf{w}\|_2^2 - \frac{1}{n}\langle\mathbf{y},\mathbf{X}\mathbf{w}\rangle + \frac{\lambda}{2}\|\mathbf{w}\|_2^2,\end{aligned}$$

where $\mathbf{X}\in\mathbb{R}^{n\times p}$ is the data matrix, $\mathbf{y}\in\mathbb{R}^n$ is the response vector, and λ is the tuning parameter. Let $\mathbf{w}^*$ denote the optimum of problem (2.1), which can be computed in a closed form as

$$\mathbf{w}^* = \left(\lambda\mathbf{I}_p + \frac{\mathbf{X}^\top\mathbf{X}}{n}\right)^{-1}\frac{\mathbf{X}^\top\mathbf{y}}{n}.$$

Sketching has become a widely used technique for efficiently finding an approximate solution to (2.1) when both n and p are large [6, 11, 22]. To avoid solving a problem of huge sample size, the traditional sketching techniques (for example, [20, 16]) reduce the sample size from n to m, with $m \ll n$, and solve the following sketched ℓ_2 regularized least-squares problem:

$$\begin{aligned}&\min_{\mathbf{w}\in\mathbb{R}^p} P(\mathbf{\Pi}^\top\mathbf{X},\mathbf{\Pi}^\top\mathbf{y};\mathbf{w})\\ &= \min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{2n}\left\|\mathbf{\Pi}^\top\mathbf{y}-\mathbf{\Pi}^\top\mathbf{X}\mathbf{w}\right\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2, \qquad (2.2)\end{aligned}$$

where $\mathbf{\Pi}\in\mathbb{R}^{n\times m}$ is a sketching matrix. The problem (2.2) can be solved faster and with less storage as long as we can choose $m \ll n$. Typical choice of $\mathbf{\Pi}$ includes a random matrix with Gaussian or Rademacher entries, sub-sampled randomized Hadamard transform [2], and sub-sampled randomized Fourier transform [19]. See discussion in Section 2.1 of [17] for more details.

Though the classical sketching has been successful in various problems and has provable guarantees, as shown in [17], there is an approximation limit for classical sketching methods to be practically useful. To obtain an approximate solution with high precision, the sketching dimension m often needs to be of the same order as n. This is impractical as the goal of sketching is to speed up the algorithms via reducing the sample size.

Based on the following equivalent formulation of (2.1)

$$\begin{aligned}&\min_{\mathbf{w}\in\mathbb{R}^p} P(\mathbf{X},\mathbf{y};\mathbf{w})\\ &= \min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{2n}\|\mathbf{y}\|_2^2 + \frac{1}{2n}\|\mathbf{X}\mathbf{w}\|_2^2 - \frac{1}{n}\langle\mathbf{y},\mathbf{X}\mathbf{w}\rangle + \frac{\lambda}{2}\|\mathbf{w}\|_2^2, \qquad (2.3)\end{aligned}$$

the Hessian sketch [17] only sketches the quadratic part $\|\mathbf{X}\mathbf{w}\|_2^2$ with respect to $\mathbf{X}$, but not the linear part $\langle\mathbf{y},\mathbf{X}\mathbf{w}\rangle$, leading to the following problem

$$\begin{aligned}&\min_{\mathbf{w}\in\mathbb{R}^p} P_{\text{HS}}(\mathbf{X},\mathbf{y};\mathbf{\Pi},\mathbf{w})\\ &= \min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{2n}\|\mathbf{y}\|_2^2 + \frac{1}{2n}\left\|\mathbf{\Pi}^\top\mathbf{X}\mathbf{w}\right\|_2^2 - \frac{1}{n}\langle\mathbf{y},\mathbf{X}\mathbf{w}\rangle + \frac{\lambda}{2}\|\mathbf{w}\|_2^2. \qquad (2.4)\end{aligned}$$

Iterative Hessian Sketch. The Hessian sketch suffers from the same approximation limit as the classical sketch. However, one notable feature of the Hessian sketch is that one can implement an iterative extension to refine the accuracy of the approximation. [17] showed that the approximation error of IHS is exponentially decreasing with the number of sketching iterations. Thus IHS can find an approximate solution with an ϵ-approximation error within $\mathcal{O}(\log(1/\epsilon))$ iterations, as long as the sketching dimension m is large enough. IHS was originally developed for the least-squares problem in (2.1), the idea can be extended to solve more general problems, such as constrained least-squares [17], optimization with self-concordant loss [15], as well as non-parametric methods [24].

2.2 Dual Random Projection

While Hessian sketch [17] tries to resolve the issue of huge sample size, Dual Random Projection [26, 27] is aimed at resolving the issue of high-dimensionality by using random projections as a tool for reducing the dimension of data. Again, we consider the standard ridge regression problem in (2.1). A random projection is now used to transform the original problem (2.1) to a low-dimensional problem:

$$\min_{\mathbf{z}\in\mathbb{R}^p} P_{\mathrm{RP}}(\mathbf{XR},\mathbf{y};\mathbf{z}) = \min_{\mathbf{z}\in\mathbb{R}^d} \frac{1}{2n}\|\mathbf{y}-\mathbf{XRz}\|_2^2 + \frac{\lambda}{2}\|\mathbf{z}\|_2^2, \tag{2.5}$$

where $\mathbf{R}\in\mathbb{R}^{p\times d}$ is a random projection matrix, and $d \ll p$.

Let $\widehat{\mathbf{z}} = \arg\min_{\mathbf{z}} P_{\mathrm{RP}}(\mathbf{XR},\mathbf{y};\mathbf{z})$. If we want to recover the original high-dimensional solution, [27] observed that the naive solution $\widehat{\mathbf{w}}_{\mathrm{RP}} = \mathbf{R}\widehat{\mathbf{z}}$ results in a bad approximation. Instead, the optimal solution of the original problem, $\mathbf{w}^*$, is recovered from the dual solution, leading to the dual random projection (DRP) approach that we explain below. The dual problem of the optimization problem in (2.1) is

$$\begin{aligned}&\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} D(\mathbf{X},\mathbf{y};\boldsymbol{\alpha})\\ &= \max_{\boldsymbol{\alpha}\in\mathbb{R}^n} -\frac{1}{2n}\boldsymbol{\alpha}^\top\boldsymbol{\alpha} + \frac{\mathbf{y}^\top\boldsymbol{\alpha}}{n} - \frac{1}{2\lambda n^2}\boldsymbol{\alpha}^\top\mathbf{XX}^\top\boldsymbol{\alpha}.\end{aligned} \tag{2.6}$$

Let $\boldsymbol{\alpha}^* = \arg\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} D(\mathbf{X},\mathbf{y};\boldsymbol{\alpha})$ be the dual optimal solution. By the standard primal-dual theory [3], we have the following connection between the optimal primal and dual solutions:

$$\boldsymbol{\alpha}^* = \mathbf{y} - \mathbf{Xw}^* \quad \text{and} \quad \mathbf{w}^* = \frac{1}{\lambda n}\mathbf{X}^\top\boldsymbol{\alpha}^*. \tag{2.7}$$

The dual random projection procedure works as follows. First, we construct and solve the low-dimensional, randomly projected problem (2.5) and obtain the solution $\widehat{\mathbf{z}}$. Next, we calculate the approximated dual variables by

$$\widehat{\boldsymbol{\alpha}}_{\mathrm{DRP}} = \mathbf{y} - \mathbf{XR}\widehat{\mathbf{z}}, \tag{2.8}$$

based on the approximated dual solution $\widehat{\boldsymbol{\alpha}}_{\mathrm{DRP}}$. Finally, we recover the primal solution as:

$$\widehat{\mathbf{w}}_{\mathrm{DRP}} = \frac{1}{\lambda n}\mathbf{X}^\top\widehat{\boldsymbol{\alpha}}_{\mathrm{DRP}}. \tag{2.9}$$

Combining the steps above, it is easy to see that the dual random projection for ridge regression has the following closed form solution:

$$\widehat{\mathbf{w}}_{\mathrm{DRP}} = \frac{\mathbf{X}^\top}{n}\left(\lambda\mathbf{I}_n + \frac{\mathbf{XRR}^\top\mathbf{X}^\top}{n}\right)^{-1}\mathbf{y}. \tag{2.10}$$

The recovered solution from the dual, $\widehat{\mathbf{w}}_{\mathrm{DRP}}$, has much better approximation compared to the solution recovered directly from primal problem $\widehat{\mathbf{w}}_{\mathrm{RP}}$. Specifically, $\widehat{\mathbf{w}}_{\mathrm{RP}}$ is always a poor approximation of $\mathbf{w}^*$, because $\widehat{\mathbf{w}}_{\mathrm{RP}}$ lives in a random subspace spanned by the random projection matrix $\mathbf{R}$, while $\widehat{\mathbf{w}}_{\mathrm{DRP}}$ can be a good approximation of $\mathbf{w}^*$ as long as the projected dimension d is large enough [27]. Finally, an iterative extension of DRP can exponentially reduce the approximation error [27].

2.3 Limitations of IHS and IDRP

Though IHS and IDRP improved the classical sketch and random projection by enabling us to find a high quality approximation more efficiently, they are imperfect due to the following reasons:

- The guarantee that the approximation error decreases exponentially for both IHS and IDRP relies on the sketching dimension being large enough. The necessary sketching dimension depends on the intrinsic complexity of the problem, and, if the sketching dimension is too small, IHS can diverge, obtaining arbitrary worse approximation.
- As we will show later, even when the sketching dimension is large enough, the speed at which the approximation error decreases in IHS and IDRP can be significantly improved.

3 A Unified View of IHS and IDRP

In this section, we present a unified view of the iterative Hessian sketch and iterative dual random projection. We first show that the dual random projection is equivalent to applying the Hessian sketch in the dual space. Next, we demonstrate that both IHS and IDRP can be viewed as an optimization process with preconditioning. This view allows us to develop better iterative algorithms by searching the conjugate directions.

3.1 Dual Random Projection is Hessian Sketch in Dual Space

We present the equivalence between Hessian sketch and dual random projection. Note that the Hessian sketch is used for *sample reduction*, while the dual random projection is utilized for *dimension reduction*. Recall that the dual maximization objective (2.6) is quadratic with respect to $\boldsymbol{\alpha}$. We can write it in the equivalent form as

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \boldsymbol{\alpha}^\top \left(\frac{\mathbf{X}\mathbf{X}^\top}{2\lambda n} + \frac{1}{2}\mathbf{I}_n\right)\boldsymbol{\alpha} - \langle \mathbf{y}, \boldsymbol{\alpha}\rangle. \tag{3.1}$$

By applying the Hessian sketch with sketching matrix $\mathbf{R} \in \mathbb{R}^{p\times d}$, we find an approximate solution for $\boldsymbol{\alpha}^*$ as

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}_{\text{HS}} &= \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \boldsymbol{\alpha}^\top \left(\frac{\mathbf{X}\mathbf{R}\mathbf{R}^\top\mathbf{X}^\top}{2\lambda n} + \frac{1}{2}\mathbf{I}_n\right)\boldsymbol{\alpha} - \langle \mathbf{y}, \boldsymbol{\alpha}\rangle \\ &= \lambda\left(\lambda\mathbf{I}_n + \frac{\mathbf{X}\mathbf{R}\mathbf{R}^\top\mathbf{X}^\top}{n}\right)^{-1}\mathbf{y}.\end{aligned} \tag{3.2}$$

The primal-dual connection (2.7) gives us the following approximation to the original problem

$$\widehat{\mathbf{w}} = \frac{\mathbf{X}^\top}{n}\left(\lambda\mathbf{I}_n + \frac{\mathbf{X}\mathbf{R}\mathbf{R}^\top\mathbf{X}^\top}{n}\right)^{-1}\mathbf{y},$$

which is the same as the DRP approximation in (2.10). From this discussion, we see that *the Dual Random Projection is the Hessian sketch applied in the dual space.* To summarize, for ridge regression problem (2.1) one has closed form solutions for various sketching techniques as:

$$\begin{aligned}\texttt{Original}:\quad \mathbf{w}^* &= \left(\lambda\mathbf{I}_p + \frac{\mathbf{X}^\top\mathbf{X}}{n}\right)^{-1}\frac{\mathbf{X}^\top\mathbf{y}}{n} \\ &= \frac{\mathbf{X}^\top}{n}\left(\lambda\mathbf{I}_n + \frac{\mathbf{X}\mathbf{X}^\top}{n}\right)^{-1}\mathbf{y}\end{aligned}$$

$$\texttt{Classical Sketch}:$$

$$\widehat{\mathbf{w}}_{\text{CS}} = \left(\lambda\mathbf{I}_p + \frac{\mathbf{X}^\top\boldsymbol{\Pi}\boldsymbol{\Pi}^\top\mathbf{X}}{n}\right)^{-1}\frac{\mathbf{X}^\top\boldsymbol{\Pi}\boldsymbol{\Pi}^\top\mathbf{y}}{n}$$

$$\texttt{Random Projection}:$$

$$\widehat{\mathbf{w}}_{\text{RP}} = \mathbf{R}\left(\lambda\mathbf{I}_d + \frac{\mathbf{R}^\top\mathbf{X}^\top\mathbf{X}\mathbf{R}}{n}\right)^{-1}\mathbf{R}^\top\frac{\mathbf{X}^\top\mathbf{y}}{n}$$

$$\texttt{Hessian Sketch}:$$

$$\widehat{\mathbf{w}}_{\text{HS}} = \left(\lambda\mathbf{I}_p + \frac{\mathbf{X}^\top\boldsymbol{\Pi}\boldsymbol{\Pi}^\top\mathbf{X}}{n}\right)^{-1}\frac{\mathbf{X}^\top\mathbf{y}}{n}$$

$$\texttt{Dual Random Projection}:$$

$$\widehat{\mathbf{w}}_{\text{DRP}} = \frac{\mathbf{X}^\top}{n}\left(\lambda\mathbf{I}_n + \frac{\mathbf{X}\mathbf{R}\mathbf{R}^\top\mathbf{X}^\top}{n}\right)^{-1}\mathbf{y}$$

As we can see above, the Hessian sketch is sketching the *covariance matrix*:

$$\mathbf{X}^\top\mathbf{X} \to \mathbf{X}^\top\boldsymbol{\Pi}\boldsymbol{\Pi}^\top\mathbf{X},$$

while DRP is sketching the *Gram matrix*:

$$\mathbf{X}\mathbf{X}^\top \to \mathbf{X}\mathbf{R}\mathbf{R}^\top\mathbf{X}^\top.$$

3.2 IHS and IDRP as Optimization with Preconditionning

Define the initial Hessian sketch approximation as

$$\widehat{\mathbf{w}}_{\text{HS}}^{(1)} = \arg\min_{\mathbf{w}} \mathbf{w}^\top\left(\frac{\mathbf{X}^\top\boldsymbol{\Pi}\boldsymbol{\Pi}^\top\mathbf{X}}{2n} + \frac{\lambda}{2}\mathbf{I}_p\right)\mathbf{w} - \frac{1}{n}\langle\mathbf{y}, \mathbf{X}\mathbf{w}\rangle.$$

A refinement of $\widehat{\mathbf{w}}_{\text{HS}}^{(1)}$ can be obtained by considering the following optimization problem

$$\begin{aligned}&\min_{\mathbf{u}} \frac{1}{2n}\left\|\mathbf{y} - \mathbf{X}(\mathbf{u} + \widehat{\mathbf{w}}_{\text{HS}}^{(1)})\right\|_2^2 + \frac{\lambda}{2}\left\|(\mathbf{u} + \widehat{\mathbf{w}}_{\text{HS}}^{(1)})\right\|_2^2 \\ &= \min_{\mathbf{u}} \mathbf{u}^\top\left(\frac{\mathbf{X}^\top\mathbf{X}}{2n} + \frac{\lambda}{2}\mathbf{I}_p\right)\mathbf{u} \\ &\qquad - \left\langle \frac{\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\widehat{\mathbf{w}}_{\text{HS}}^{(t)})}{n} - \lambda\widehat{\mathbf{w}}_{\text{HS}}^{(t)}, \mathbf{u}\right\rangle,\end{aligned}$$

whose optimum is $\mathbf{w}^* - \widehat{\mathbf{w}}_{\text{HS}}^{(1)}$. The main idea of the iterative Hessian sketch is to approximate the residual solution $\mathbf{w}^* - \widehat{\mathbf{w}}_{\text{HS}}^{(1)}$ by the Hessian sketch. At iteration t, $\mathbf{w}^* - \widehat{\mathbf{w}}_{\text{HS}}^{(t)}$ is approximated by $\widehat{\mathbf{u}}^{(t)}$ that minimizes the following problem

$$\begin{aligned}\min_{\mathbf{u}}\ &\mathbf{u}^\top\left(\frac{\mathbf{X}^\top\boldsymbol{\Pi}\boldsymbol{\Pi}^\top\mathbf{X}}{2n} + \frac{\lambda}{2}\mathbf{I}_p\right)\mathbf{u} \\ &- \left\langle \frac{\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\widehat{\mathbf{w}}_{\text{HS}}^{(t)})}{n} - \lambda\widehat{\mathbf{w}}_{\text{HS}}^{(t)}, \mathbf{u}\right\rangle\end{aligned} \tag{3.3}$$

and the new approximation $\widehat{\mathbf{w}}_{\text{HS}}^{(t+1)}$ is updated as

$$\widehat{\mathbf{w}}_{\text{HS}}^{(t+1)} = \widehat{\mathbf{w}}_{\text{HS}}^{(t)} + \widehat{\mathbf{u}}^{(t)}.$$

Also notice that (3.3) is a sketched problem with sample size m, and therefore it can be solved more efficiently than the original problem (2.1). Furthermore, we can reuse the previous sketched data matrix $\boldsymbol{\Pi}^\top\mathbf{X}$ without constructing any new random sketching matrices.

Next, we show that the iterative Hessian sketch is in fact an optimization process with *preconditioning*. Let

$$\mathbf{H} = \frac{\mathbf{X}^\top\mathbf{X}}{n} + \lambda\mathbf{I}_p \quad\text{and}\quad \widetilde{\mathbf{H}} = \frac{\mathbf{X}^\top\boldsymbol{\Pi}\boldsymbol{\Pi}^\top\mathbf{X}}{n} + \lambda\mathbf{I}_p.$$

Let

$$\nabla P(\mathbf{w}) = -\frac{\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\mathbf{w})}{n} + \lambda\mathbf{w}$$

denote the gradient of $P(\mathbf{X}, y; \mathbf{w})$ with respect to $\mathbf{w}$. Then the IHS algorithm can be seen as performing the following iterative update

$$\widehat{\mathbf{w}}_{\text{HS}}^{(t+1)} = \widehat{\mathbf{w}}_{\text{HS}}^{(t)} - \widetilde{\mathbf{H}}^{-1}\nabla P(\widehat{\mathbf{w}}_{\text{HS}}^{(t)}),$$

which is like a Newton update with the true Hessian $\mathbf{H}$ being replaced by the sketched Hessian $\widetilde{\mathbf{H}}$. Another way to think about this update is via the change of variable $\mathbf{z} = \widetilde{\mathbf{H}}^{1/2}\mathbf{w}$ and then applying the gradient descent in the $\mathbf{z}$ space

$$\begin{aligned}\widehat{\mathbf{z}}^{(t+1)} =& \widehat{\mathbf{z}}^{(t)} - \nabla_{\mathbf{z}} P(\widetilde{\mathbf{H}}^{-1/2}\mathbf{z}) \\ =& \widehat{\mathbf{z}}^{(t)} - \widetilde{\mathbf{H}}^{-1/2}\nabla P(\widetilde{\mathbf{H}}^{-1/2}\widehat{\mathbf{z}}^{(t)}).\end{aligned}$$

Multiplying by $\widetilde{\mathbf{H}}^{-1/2}$, changes the update back to the original space, leading back to the IHS update

$$\widehat{\mathbf{w}}_{\text{HS}}^{(t+1)} = \widehat{\mathbf{w}}_{\text{HS}}^{(t)} - \widetilde{\mathbf{H}}^{-1}\nabla P(\widehat{\mathbf{w}}_{\text{HS}}^{(t)}).$$

From the discussion above, we see that the iterative Hessian sketch is an optimization process with the sketched Hessian as preconditioning.

With this unified optimization view of IHS and IDRP, we can provide a unified theoretical analysis for Hessian sketch and dual random projection, see [21] for more details. Moreover, based on the preconditioned gradient descent view we propose faster convergent iterative algorithms via searching the conjugate direction, as well as primal-dual sketch method to simultaneously reduce the sample and feature dimension. We introduce these methods in the following sections, while their corresponding theoretical analysis can be found in [21].

4 Accelerated IHS via Preconditioned Conjugate Gradient

In this section, we present the accelerated iterative Hessian sketch (Acc-IHS) algorithm by utilizing the idea of preconditioned conjugate gradient. Conjugate gradient is known to have better convergence properties than gradient descent in solving linear systems [8, 12]. Since the iterative Hessian sketch is performing the gradient descent in the transformed space $\mathbf{z} = \widetilde{\mathbf{H}}^{1/2}\mathbf{w}$, it can be accelerated by performing the conjugate gradient descent instead. Equivalently, we can implicitly transform the space by defining inner product as $\langle \mathbf{x}, \mathbf{y}\rangle = \mathbf{x}^\top \widetilde{\mathbf{H}}\mathbf{y}$.

This leads to the algorithm Acc-IHS detailed in Algorithm 1. At each iteration, the solver is called for the following sketched linear system

$$\min_{\mathbf{u}} \mathbf{u}^\top \left(\frac{\mathbf{X}^\top \boldsymbol{\Pi}\boldsymbol{\Pi}^\top \mathbf{X}}{2n} + \frac{\lambda}{2}\mathbf{I}_p\right)\mathbf{u} - \left\langle \mathbf{r}^{(t)}, \mathbf{u}\right\rangle. \tag{4.1}$$

Unlike IHS, which uses $\widetilde{\mathbf{H}}^{-1}\nabla P(\widehat{\mathbf{w}}_{\text{HS}}^{(t)})$ as the update direction at iteration t, Acc-IHS uses $\mathbf{p}^{(t)}$ as the update direction where $\mathbf{p}^{(t)}$ is chosen to satisfy the conjugate condition: $\forall t_1, t_2 \geqslant 0, t_1 \neq t_2$

Algorithm 1: Accelerated Iterative Hessian Sketch

```
Input: Data $\mathbf{X}, \mathbf{y}$, sketching matrix $\boldsymbol{\Pi}$.
Initialization: $\widehat{\mathbf{w}}_{\text{HS}}^{(0)} = \mathbf{0}, \mathbf{r}^{(0)} = -\frac{\mathbf{X}^\top\mathbf{y}}{n}$.
Compute $\widehat{\mathbf{u}}^{(0)}$ by solving (4.1), and update
    $\mathbf{p}^{(0)} = -\widehat{\mathbf{u}}^{(0)}$, calculate $\mathbf{v}^{(0)} = \left(\frac{\mathbf{X}^\top\mathbf{X}}{n} + \lambda\mathbf{I}_p\right)\mathbf{p}^{(0)}$.
for $t = 0, 1, 2, \ldots$ do
    Calculate $\alpha^{(t)} = \frac{\langle \mathbf{r}^{(t)}, \mathbf{u}^{(t)}\rangle}{\langle \mathbf{p}^{(t)}, \mathbf{v}^{(t)}\rangle}$
    Update the approximation by
        $\widehat{\mathbf{w}}_{\text{HS}}^{(t+1)} = \widehat{\mathbf{w}}_{\text{HS}}^{(t)} + \alpha^{(t)}\mathbf{p}^{(t)}$.
    Update $\mathbf{r}^{(t+1)} = \mathbf{r}^{(t)} + \alpha^{(t)}\mathbf{v}^{(t)}$.
    Update $\mathbf{u}^{(t+1)}$ by solving (4.1).
    Update $\beta^{(t+1)} = \frac{\langle \mathbf{r}^{(t+1)}, \mathbf{u}^{(t)}\rangle}{\langle \mathbf{r}^{(t)}, \mathbf{r}^{(t)}\rangle}$.
    Update $\mathbf{p}^{(t+1)} = -\mathbf{u}^{(t+1)} + \beta^{(t+1)}\mathbf{p}^{(t)}$.
    Update $\mathbf{v}^{(t+1)} = \left(\frac{\mathbf{X}^\top\mathbf{X}}{n} + \lambda\mathbf{I}_p\right)\mathbf{p}^{(t+1)}$.
end
```

$$\left(\mathbf{p}^{(t_1)}\right)^\top \widetilde{\mathbf{H}}^{-1/2}\mathbf{H}\widetilde{\mathbf{H}}^{-1/2}\mathbf{p}^{(t_2)} = 0.$$

Since the updating direction is conjugate to the previous directions, it is guaranteed that after p iterations we reach the exact minimizer, that is,

$$\widehat{\mathbf{w}}_{\text{HS}}^{(t)} = \mathbf{w}^*, \quad \forall t \geqslant p.$$

Moreover, Acc-IHS has the same computational cost as the standard IHS in solving each sketched subproblem. However, the convergence rate of Algorithm 1 is much faster than IHS, that is, it requires solving much smaller number of sketched sub-problems compared to IHS to reach the same approximation accuracy.

4.1 Accelerated Iterative Dual Random Projection

We recall the standard IDRP [27]. At iteration t, let $\widehat{\mathbf{w}}_{\text{DRP}}^{(t)}$ denote the approximate solution. The following optimization problem

$$\min_{\mathbf{u}\in\mathbb{R}^p} \frac{1}{2n}\left\|\mathbf{y} - \mathbf{X}(\mathbf{u} + \widehat{\mathbf{w}}_{\text{DRP}}^{(t)})\right\|_2 + \frac{\lambda}{2}\left\|\mathbf{u} + \widehat{\mathbf{w}}_{\text{DRP}}^{(t)}\right\|_2^2, \tag{4.2}$$

has the solution $\mathbf{w}^* - \widehat{\mathbf{w}}_{\text{DRP}}^{(t)}$. The idea of the iterative dual random projection is to approximate the residual $\mathbf{w}^* - \widehat{\mathbf{w}}_{\text{DRP}}^{(t)}$ by applying dual random projection to

(4.2). Given $\widehat{\mathbf{w}}^{(t)}_{\text{DRP}}$, let $\widehat{\mathbf{z}}^{(t)}$ to be the solution of

$$\min_{\mathbf{z}\in\mathbb{R}^d} \frac{1}{2n}\left\|\mathbf{y}-\mathbf{X}\mathbf{w}^{(t)}_{\text{DRP}}-\mathbf{X}\mathbf{R}\mathbf{z}\right\|_2^2+\frac{\lambda}{2}\left\|\mathbf{z}+\mathbf{R}^\top\mathbf{w}^{(t)}_{\text{DRP}}\right\|_2^2. \tag{4.3}$$

Using $\widehat{\mathbf{z}}^{(t)}$, the dual variables are updated as

$$\widehat{\boldsymbol{\alpha}}^{(t+1)}_{\text{DRP}}=\mathbf{y}-\mathbf{X}\mathbf{w}^{(t)}_{\text{DRP}}-\mathbf{X}\mathbf{R}\widehat{\mathbf{z}}$$

and the primal variables as

$$\widehat{\mathbf{w}}^{(t+1)}_{\text{DRP}}=\frac{1}{\lambda n}\mathbf{X}^\top\widehat{\boldsymbol{\alpha}}^{(t+1)}_{\text{DRP}}.$$

Based on the equivalence between the dual random projection and the Hessian sketch established in Section 3.2, we propose an accelerated iterative dual random projection algorithm, which improves the convergence speed of the standard iterative DRP procedure [27]. At each iteration t, we call the solver for the following randomly projected problem based on the residual $\mathbf{r}^{(t)}$ (see Algorithm 4 in [21]):

$$\min_{\mathbf{z}\in\mathbb{R}^d}\mathbf{z}^\top\left(\frac{\mathbf{R}^\top\mathbf{X}^\top\mathbf{X}\mathbf{R}}{2n}+\frac{\lambda}{2}\mathbf{I}_d\right)\mathbf{z}-\langle\mathbf{R}^\top\mathbf{X}^\top\mathbf{r}^{(t)},\mathbf{z}\rangle. \tag{4.4}$$

The accelerated IDRP algorithm runs the Acc-IHS Algorithm 1 in the dual space. However, Acc-IDRP is still a primal algorithm, since it updates the corresponding dual variables after solving the randomly projected primal problem (4.4). The dual version of Acc-IDRP algorithm would at each iteration solve the following dual optimization problem

$$\min_{\mathbf{u}\in\mathbb{R}^n}\mathbf{u}^\top\left(\frac{\mathbf{X}\mathbf{R}\mathbf{R}^\top\mathbf{X}^\top}{2n}+\frac{\lambda}{2}\mathbf{I}_n\right)\mathbf{u}-\langle\mathbf{r}^{(t)},\mathbf{u}\rangle, \tag{4.5}$$

where $\mathbf{r}^{(t)}$ is the dual residual. This, however, is not a practical algorithm as it requires solving relatively more expensive dual problem.

Though the computational cost per iteration of Acc-IDRP and standard IDRP is the same, Acc-IDRP converges faster and is more robust than IDRP.

5 Iterative Primal-Dual Sketch

In this section, we combine the idea of the iterative Hessian sketch and iterative dual random projection from the primal-dual point of view. We propose a more efficient sketching technique named *Iterative Primal-Dual Sketch* (IPDS), which simultaneously reduces the sample size and dimensionality of the problem, while recovering the original solution to a high precision.

Algorithm 2: Iterative Primal-Dual Sketch (IPDS).

```
Input: Data $\mathbf{X}\in\mathbb{R}^{n\times p},\mathbf{y}\in\mathbb{R}^n$, sketching matrix
    $\mathbf{R}\in\mathbb{R}^{p\times d},\mathbf{\Pi}\in\mathbb{R}^{n\times m}$.
Initialization: $\widehat{\mathbf{w}}^{(0)}_{\text{DS}}=\mathbf{0}$.
for $t=0,1,2,\ldots$ do
    Initialization: $\widetilde{\mathbf{z}}^{(0)}=\mathbf{0},k=0$
    if not converged then
        Solve the sketched problem in (5.2) and
          obtain solution $\Delta\mathbf{z}^{(k)}$.
        Update $\widetilde{\mathbf{z}}^{(k+1)}=\widetilde{\mathbf{z}}^{(k)}+\Delta\mathbf{z}^{(k)}$.
        Update $k=k+1$.
    end
    Update dual approximation:
      $\widehat{\boldsymbol{\alpha}}^{(t+1)}_{\text{DS}}=\mathbf{y}-\mathbf{X}\widehat{\mathbf{w}}^{(t)}_{\text{DS}}-\mathbf{X}\mathbf{R}\widetilde{\mathbf{z}}^{(k+1)}$.
    Update primal approximation:
      $\widehat{\mathbf{w}}^{(t+1)}_{\text{DS}}=\frac{1}{\lambda n}\mathbf{X}^\top\widehat{\boldsymbol{\alpha}}^{(t+1)}_{\text{DS}}$.
end
```

The Hessian sketch is particularly suitable for the case where the sample size is much larger than the problem dimension and the computational bottleneck is "big n". Here the Hessian sketch reduces the sample size significantly, and as a consequence, speeds up the computation. By utilizing the iterative extension approximation error can be further reduced to recover the original solution to a high precision. In contrast, the dual random projection is aimed at dimensionality reduction and is suitable for the case of high-dimensional data, with relatively small sample size. Here the computational bottleneck is "large p" and the random projection is used to reduce dimensionality and speedup computations.

The iterative Primal-Dual Sketch only involves solving small scale problems. For the original problem (2.1) with data $\{\mathbf{X},\mathbf{y}\}$, we first construct the randomly projected data, as well as the *doubly sketched* data, as follows:

$$\mathbf{X}\to\mathbf{X}\mathbf{R},\qquad \mathbf{X}\mathbf{R}\to\mathbf{\Pi}^\top\mathbf{X}\mathbf{R},$$

where $\mathbf{X}\mathbf{R}$ is the randomly projected data, and $\mathbf{\Pi}^\top\mathbf{X}\mathbf{R}$ is doubly sketched data. Let $\widehat{\mathbf{w}}^{(0)}_{\text{DS}}=\mathbf{0}$. At every iteration of IPDS, we first apply random projection on the primal problem (which is equivalent to the Hessian sketch on the dual problem), and obtain the following problem:

$$\min_{\mathbf{z}\in\mathbb{R}^d}\frac{1}{2n}\left\|\mathbf{y}-\mathbf{X}\widehat{\mathbf{w}}^{(t)}_{\text{DS}}-\mathbf{X}\mathbf{R}\mathbf{z}\right\|_2^2+\frac{\lambda}{2}\left\|\mathbf{z}+\mathbf{R}^\top\widehat{\mathbf{w}}^{(t)}_{\text{DS}}\right\|_2^2, \tag{5.1}$$

which is the same as the iterative dual random projection subproblem (4.3). However, different from IDRP, we do not directly solve (5.1), but apply the iterative Hessian sketch in the inner loop to find an approximate solution to

$$\min_{\mathbf{z}\in\mathbb{R}^d} \ \mathbf{z}^\top \left(\frac{\mathbf{R}^\top\mathbf{X}^\top\mathbf{X}\mathbf{R}}{2n} + \frac{\lambda}{2}\mathbf{I}_d\right)\mathbf{z} - \left\langle \frac{(\mathbf{y}-\mathbf{X}\widehat{\mathbf{w}}_{\mathrm{DS}}^{(t)})^\top\mathbf{X}\mathbf{R}}{n} - \lambda\mathbf{R}^\top\widehat{\mathbf{w}}_{\mathrm{DS}}^{(t)}, \mathbf{z} \right\rangle .$$

We initialize $\widetilde{\mathbf{z}}^{(0)} = \mathbf{0}$. At iteration k in the inner loop, we solve the following sketched problem:

$$\min_{\Delta\mathbf{z}} \ \Delta\mathbf{z}^\top \left(\frac{\mathbf{R}^\top\mathbf{X}^\top\mathbf{\Pi}\mathbf{\Pi}^\top\mathbf{X}\mathbf{R}}{2n} + \frac{\lambda}{2}\mathbf{I}_d\right)\mathbf{z} - \left\langle \frac{\mathbf{R}^\top\mathbf{X}^\top(\mathbf{y}-\mathbf{X}\widehat{\mathbf{w}}_{\mathrm{DS}}^{(t)} - \mathbf{X}\mathbf{R}\widetilde{\mathbf{z}}^{(k)})}{n} - \lambda\mathbf{R}^\top\widehat{\mathbf{w}}_{\mathrm{DS}}^{(t)} - \lambda\widetilde{\mathbf{z}}^{(k)}, \Delta\mathbf{z} \right\rangle . \tag{5.2}$$

and update $\widetilde{\mathbf{z}}^{(k+1)}$ as

$$\widetilde{\mathbf{z}}^{(k+1)} = \widetilde{\mathbf{z}}^{(k)} + \Delta\mathbf{z}^{(k)}.$$

The key point is that for the subproblem (5.2), the sketched data matrix is only of size $m \times d$, compared to the original problem size $n \times p$, where $n \gg m, p \gg d$. In contrast, the IHS still needs to solve sub-problems of size $m \times p$, while IDRP needs to solve sub-problems of size $n \times d$. We only need to call solvers of $m \times d$ problem (5.2) logarithmic times to obtain a solution of high approximation quality.

The pseudo code of Iterative Primal-Dual Sketch (IPDS) is summarized in Algorithm 2. It is also possible to perform iterative Primal-Dual Sketch via another direction, that is, first perform primal Hessian sketch, and then apply dual Hessian sketch to solve the sketched primal problem:

$$\mathbf{X} \to \mathbf{\Pi}^\top\mathbf{X}, \qquad \mathbf{\Pi}^\top\mathbf{X} \to \mathbf{\Pi}^\top\mathbf{X}\mathbf{R}.$$

The idea presented in Section 4 can also be adopted to further reduce the number of calls to $m \times d$ scale sub-problems, which leads to the accelerated iterative primal-dual sketch (Acc-IPDS) algorithm. In Acc-IPDS, we maintain both the vectors in the primal space $\mathbf{u}_{\mathrm{P}}, \mathbf{v}_{\mathrm{P}}, \mathbf{r}_{\mathrm{P}}$ and the vectors in the dual space $\mathbf{u}_{\mathrm{D}}, \mathbf{v}_{\mathrm{D}}, \mathbf{r}_{\mathrm{D}}$, to make sure that the updating directions for both primal variables and dual variables are conjugate with previous updating directions. Moreover, based on the residual vector $\mathbf{r}_{\mathrm{P}}$, Acc-IPDS iteratively calls the solver to find a solution of the following sketched linear system of scale $m \times d$:

$$\min_{\mathbf{u}} \mathbf{u}^\top \left(\frac{\mathbf{R}^\top\mathbf{X}^\top\mathbf{\Pi}\mathbf{\Pi}^\top\mathbf{X}\mathbf{R}}{2n} + \frac{\lambda}{2}\mathbf{I}_d\right)\mathbf{u} - \left\langle \mathbf{r}_{\mathrm{P}}^{(k)}, \mathbf{u} \right\rangle .$$

Table 1: List of real-world data sets used in the experiments.

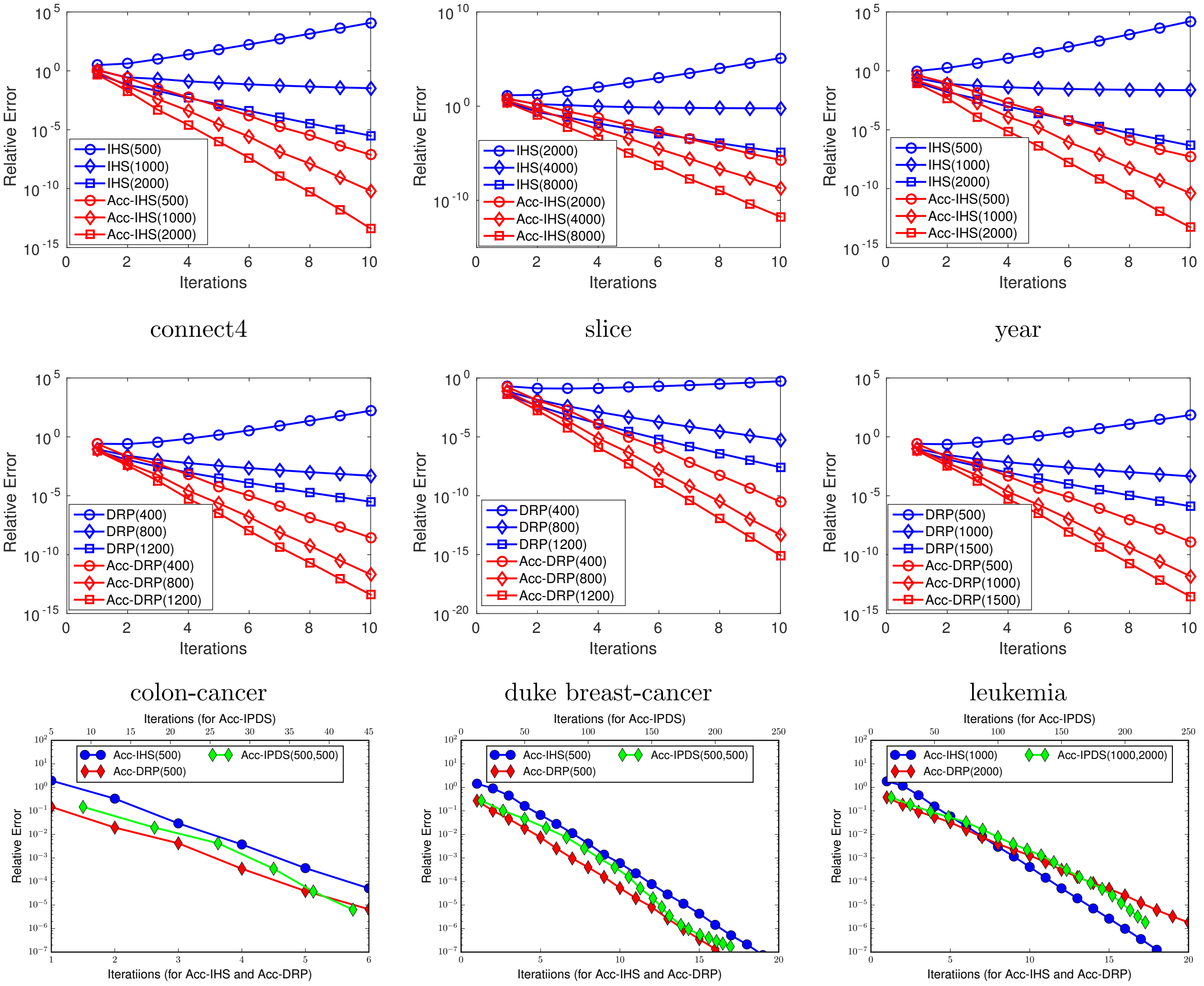
Name	#Instances	#Features
connect4	67,557	126
slice	53,500	385
year	51,630	90
colon-cancer	62	2,000
duke breast-cancer	44	7,129
leukemia	72	7,129
cifar	4,047	3,072
gisette	6,000	5,000
sector	6,412	15,000

6 Experiments

We conduct experiments on real-world data sets where their statistics are summarized in Table 1. Among all the data sets, the first 3 are cases where the sample size is significantly larger than the dimension of data and we use them to compare the IHS and Acc-IHS algorithms; the middle 3 data sets are high-dimensional data sets with small sample size and we use them to compare DRP and Acc-DRP procedures; the last 3 data sets are cases where both the sample size and dimension are relatively large and therefore are suitable for iterative primal-dual sketching methods. For the last 3 data sets we found that standard IHS and DRP often fail (unless a very large sketching dimension is used), thus we compared with Acc-IHS and Acc-DRP. The convergence plots are summarized in Figure 1. We have the following observations:

- For both IHS and Acc-IHS, the larger the sketching dimension m, the faster the iterative methods converge to the optimum, which is consistent with the theory, as also observed in [17] and [27] for IHS and IDRP algorithms. Acc-IHS and Acc-DRP converge significantly faster than IHS and DRP, respectively. In particular, accelerated algorithms converges faster than their non-accelerated counterpart even when their sketching dimensions are only 1/3 of the sketching dimensions in IHS and DRP. Moreover, when the sketching dimension is small, IHS can diverge and go far away from the optimum, while Acc-IHS still converges.

- For the last 3 data sets where n and p are both large, and the data are not exactly low-rank: IHS, DRP and IPDS often diverge because the requirements on the sketching dimension that ensure convergence are not satisfied. On the other hand, the accelerated versions still converge to the optimum,



connect4 slice year

colon-cancer duke breast-cancer leukemia

cifar gisette sector

Figure 1: Comparion of various iterative sketching approaches on real-world datasets, Top row: Acc-IHS versus IHS, middle row: Acc-DRP versus DRP, bottom row: Acc-IPDS versus Acc-IHS and Acc-DRP.

so we only show their accelerated version in the plot. It is notable that the Acc-IPDS only require solving several least squares problems where both sample size and dimension are relatively small.

7 Conclusion and Discussion

We focused on sketching techniques for solving large-scale ℓ_2 regularized least squares problem. We established the equivalence between the recently proposed two emerging techniques (Hessian sketch and dual random projection) from a primal-dual point of view. We also proposed accelerated methods for IHS and IDRP, from the preconditioned optimization perspective. Finally, by combining the primal and dual sketching ideas, we proposed a novel iterative primal-dual sketching approach, which substantially reduced the computational cost in solving sketched sub problems.

The proposed approach can be extended to solving more general problems. For example, by sketching the Newton step in the second-order optimization methods, as was done in the “Newton Sketch” paper [15], we will be able to solve regularized risk minimization problem with self-concordant losses. It will be interesting to examine its empirical performance compared to existing approaches. More generally, Hessian sketch and dual random projection are designed for solving convex problems and it will be interesting to extend them for some structured non-convex problems such as principle component analysis.

References

[1] A. E. Alaoui and M. W. Mahoney. Fast randomized kernel methods with statistical guarantees. *arXiv preprint arXiv:1411.0306*, 2014.

[2] C. Boutsidis and A. Gittens. Improved matrix algorithms via the subsampled randomized hadamard transform. *SIAM Journal on Matrix Analysis and Applications*, 34(3):1301–1340, 2013.

[3] S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

[4] A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.

[5] P. Drineas and M. W. Mahoney. Randnla: randomized numerical linear algebra. *Communications of the ACM*, 59(6):80–90, 2016.

[6] P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós. Faster least squares approximation. *Numerische Mathematik*, 117(2):219–249, 2011.

[7] N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[8] M. R. Hestenes and E. Stiefel. *Methods of conjugate gradients for solving linear systems*, volume 49. 1952.

[9] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

[10] Y. Lu, P. Dhillon, D. P. Foster, and L. Ungar. Faster ridge regression via the subsampled randomized hadamard transform. In *Advances in neural information processing systems*, pages 369–377, 2013.

[11] M. W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011.

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

[13] S. Oymak and J. A. Tropp. Universality laws for randomized dimension reduction, with applications. *arXiv preprint arXiv:1511.09433*, 2015.

[14] S. Oymak, B. Recht, and M. Soltanolkotabi. Isometric sketching of any set via the restricted isometry property. *arXiv preprint arXiv:1506.03521*, 2015.

[15] M. Pilanci and M. J. Wainwright. Newton sketch: A linear-time optimization algorithm with linear-quadratic convergence. *arXiv preprint arXiv:1505.02250*, 2015.

[16] M. Pilanci and M. J. Wainwright. Randomized sketches of convex programs with sharp guarantees. *Information Theory, IEEE Transactions on*, 61(9):5096–5115, 2015.

[17] M. Pilanci and M. J. Wainwright. Iterative hessian sketch: Fast and accurate solution approximation for constrained least-squares. *Journal of Machine Learning Research*, 2016.

[18] G. Raskutti and M. Mahoney. A statistical perspective on randomized sketching for ordinary least-squares. *stat*, 1050:25, 2015.

[19] V. Rokhlin and M. Tygert. A fast randomized algorithm for overdetermined linear least-squares regression. *Proceedings of the National Academy of Sciences*, 105(36):13212–13217, 2008.

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

[21] J. Wang, J. D. Lee, M. Mahdavi, M. Kolar, and N. Srebro. Sketching meets random projection in the dual: A provable recovery algorithm for big and high-dimensional data. *arXiv preprint arXiv:1610.03045*, 2016.

[22] D. P. Woodruff. Sketching as a tool for numerical linear algebra. *arXiv preprint arXiv:1411.4357*, 2014.

[23] T. Yang, L. Zhang, R. Jin, and S. Zhu. Theory of dual-sparse regularized randomized reduction. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 305–314, 2015.

[24] Y. Yang, M. Pilanci, and M. J. Wainwright. Randomized sketches for kernels: Fast and optimal non-parametric regression. *arXiv preprint arXiv:1501.06195*, 2015.

[25] L. Zhang, M. Mahdavi, and R. Jin. Linear convergence with condition number independent access of full gradients. In *Advances in Neural Information Processing Systems*, pages 980–988, 2013.

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

[27] L. Zhang, M. Mahdavi, R. Jin, T. Yang, and S. Zhu. Random projections for classification: A recovery approach. *Information Theory, IEEE Transactions on*, 60(11):7300–7316, 2014.