Non-PSD Matrix Sketching with Applications to Regression and Optimization – Supplementary Material

Zhili Feng[1] **Fred Roosta**[2] **David P. Woodruff**[3]

[1]Machine Learning Department, Carnegie Mellon University,
[2]School of Mathematics and Physics, University of Queensland,
[3] Computer Science Department, Carnegie Mellon University,

1 ALGORITHMS

Algorithm 1 Newton-CG With Inexact Hessian

Input: Starting point $\mathbf{x}_0$, line-search parameter $0 < \rho < 1$

for $k = 0, 1, 2, \ldots$ until convergence **do**

(approximately) Solve the following sub-problem using CG

$$\mathbf{H}_k \mathbf{p} = -\mathbf{g}_k$$

Find α_k such that

$$F(\mathbf{x}_{k+1}) \leq F(\mathbf{x}_k) + \rho \alpha_k \langle \mathbf{p}_k, \mathbf{g}_k \rangle$$

Update $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$

Output: $\mathbf{x}_k$

Algorithm 2 Newton-MR With Inexact Hessian

Input: Starting point $\mathbf{x}_0$, line-search parameter $0 < \rho < 1$

for $k = 0, 1, 2, \ldots$ until convergence **do**

(approximately) Solve the following sub-problem

$$\min_{\mathbf{p} \in \mathbb{R}^d} \|\mathbf{p}\| \;\; \text{subject to} \;\; \mathbf{p} \in \arg\min_{\widehat{\mathbf{p}} \in \mathbb{R}^d} \|\mathbf{H}_k \widehat{\mathbf{p}} + \mathbf{g}_k\| .$$

Find α_k such that

$$\|\mathbf{g}_{k+1}\|^2 \leq \|\mathbf{g}_k\|^2 + 2\rho \alpha_k \langle \mathbf{p}_k, \mathbf{H}_k \mathbf{g}_k \rangle$$

Update $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$

Output: $\mathbf{x}_k$

Supplement for the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021).

Algorithm 3 Trust Region with Inexact Hessian

Input: Starting point $\mathbf{x}_0$, initial radius $0 < \Delta_0 < \infty$, hyper-parameters $\eta \in (0,1), \gamma > 1$
for $k = 0, 1, \ldots$ **do**
  Set the approximate Hessian, $\mathbf{H}_k$, as in (4)
  if converged **then**
    Return $\mathbf{x}_k$.
  (approximately) solve the following sub-problem

$$\min_{\|\mathbf{p}\| \le \Delta_k} m_k(\mathbf{p}) \triangleq \langle \nabla F(\mathbf{x}_k), \mathbf{p} \rangle + \frac{1}{2} \langle \mathbf{p}, \mathbf{H}_k \mathbf{p} \rangle ,$$

  Set $\rho_k \triangleq \dfrac{F(\mathbf{x}_k + \mathbf{p}_k) - F(\mathbf{x}_k)}{m_t(\mathbf{p}_k)}$
  if $\rho_k \ge \eta$ **then**
    $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{p}_k$ and $\Delta_{k+1} = \gamma \Delta_k$
  else
    $\mathbf{x}_{k+1} = \mathbf{x}_k$ and $\Delta_{k+1} = \Delta_k / \gamma$
Output: $\mathbf{x}_k$

2 OMITTED PROOF IN Section 2

2.1 PROOF OF Theorem 1

This theorem and proof mimic Theorem 5 in Cohen et al. [2017].

The statistical leverage score of the i^{th} row of $\mathbf{B} \in \mathbb{C}^{n \times d}$ can also be written as the following:

$$\ell_i = \mathbf{B}_i (\mathbf{B}^* \mathbf{B})^\dagger \mathbf{B}_i^* .$$

Proof. Let $\mathbf{B}^* = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ be the SVD of $\mathbf{B}^*$. We have $\ell_i = \mathbf{B}_i^* (\mathbf{U}\mathbf{\Sigma}^{-2}\mathbf{U}^*)\mathbf{B}_i$. Let $Y = \mathbf{\Sigma}^{-1}\mathbf{U}^* \left(\mathbf{C}^\intercal \mathbf{C} - \mathbf{B}^\intercal \mathbf{B}\right) \mathbf{U}\mathbf{\Sigma}^{-1}$. Then we write

$$\mathbf{Y} = \sum_{j=1}^{t} \left(\mathbf{\Sigma}^{-1}\mathbf{U}^* \left(\mathbf{C}_j^\intercal \mathbf{C}_j - \frac{1}{t} \mathbf{B}^\intercal \mathbf{B} \right) \mathbf{U}\mathbf{\Sigma}^{-1} \right) \triangleq \sum_{j=1}^{t} \mathbf{X}_j$$

where $\mathbf{C}_j$ is the j^{th} row of $\mathbf{C}$. Note with probability p_i

$$\mathbf{X}_j = \frac{1}{t} \mathbf{\Sigma}^{-1}\mathbf{U}^* \left(\frac{1}{p_i} \mathbf{B}_i^\intercal \mathbf{B}_i - \mathbf{B}^\intercal \mathbf{B} \right) \mathbf{U}\mathbf{\Sigma}^{-1}.$$

Since $\mathbb{E}[\frac{1}{p_i}\mathbf{B}_i^\intercal \mathbf{B}_i - \mathbf{B}^\intercal \mathbf{B}] = 0$ we have $\mathbb{E}[\mathbf{Y}] = 0$. Also we have $\mathbf{C}^\intercal \mathbf{C} = \mathbf{U}\mathbf{\Sigma}\mathbf{Y}\mathbf{\Sigma}\mathbf{U}^* + \mathbf{B}^\intercal \mathbf{B}$. Because $\mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^* = \mathbf{B}^*\mathbf{B}$ it suffices to show that $\|\mathbf{Y}\| \le \epsilon$, which gives $-\epsilon \mathbf{I} \preceq \mathbf{Y} \preceq \epsilon \mathbf{I}$, and consequently:

$$\mathbf{B}^\intercal \mathbf{B} - \epsilon \mathbf{B}^* \mathbf{B} \preceq \mathbf{C}^\intercal \mathbf{C} \preceq \mathbf{B}^\intercal \mathbf{B} + \epsilon \mathbf{B}^* \mathbf{B}.$$

A useful tool for proving $\|\mathbf{Y}\|$ is small is the matrix Bernstein inequality Tropp et al. [2015]. We remark that the version we use is suitable for complex matrices as well.

Note that for any i, because $\mathbf{B}^\intercal \mathbf{B}$ has real entries, we have

$$\frac{1}{\ell_i} \mathbf{B}_i^\intercal \mathbf{B}_i \preceq \frac{1}{\ell_i} \mathbf{B}_i^* \mathbf{B}_i \preceq \mathbf{B}^* \mathbf{B}$$

where the first step is by the structure of A, and the second step follows from a known property of leverage scores (see the proof of Lemma 4 in Cohen et al. [2015a]). With this we have:

$$\frac{1}{\ell_i}\mathbf{\Sigma}^{-1}\mathbf{U}^*\mathbf{B}_i^\intercal\mathbf{B}_i\mathbf{U}\mathbf{\Sigma}^{-1} \preceq \mathbf{\Sigma}^{-1}\mathbf{U}^*\left(\mathbf{B}^*\mathbf{B}\right)\mathbf{U}\mathbf{\Sigma}^{-1} = \mathbf{I}$$

Hence

$$\mathbf{X}_j + \frac{1}{t}\mathbf{\Sigma}^{-1}\mathbf{U}^*\mathbf{B}^\intercal\mathbf{B}\mathbf{U}\mathbf{\Sigma}^{-1} \preceq \frac{1}{tp_i}\cdot\ell_i\cdot\mathbf{I} \preceq \frac{\epsilon^2}{c\log(d/\delta)\sum_i\tilde{\ell}_i}\cdot\frac{\sum_i\tilde{\ell}_i}{\tilde{\ell}_i}\ell_i\cdot\mathbf{I} \preceq \frac{\epsilon^2}{c\log(d/\delta)}\mathbf{I}$$

In addition

$$\frac{1}{t}\mathbf{\Sigma}^{-1}\mathbf{U}^*\mathbf{B}^\intercal\mathbf{B}\mathbf{U}\mathbf{\Sigma}^{-1} \preceq \frac{1}{t}\mathbf{\Sigma}^{-1}\mathbf{U}^*\mathbf{B}^*\mathbf{B}\mathbf{U}\mathbf{\Sigma}^{-1} = \frac{\epsilon^2}{c\log(d/\delta)}\mathbf{I}$$

These two give $\|\mathbf{X}_j\| \leq \frac{\epsilon^2}{c\log(d/\delta)}$. We then bound the variance of $\mathbf{Y}$:

$$\begin{aligned}
\mathbb{E}[\mathbf{Y}\mathbf{Y}^*] = \mathbb{E}[\mathbf{Y}^*\mathbf{Y}] &= t\mathbb{E}[\mathbf{X}_j\mathbf{X}_j^*] \\
&\preceq \frac{1}{t}\sum_i p_i\cdot\frac{1}{p_i^2}\mathbf{\Sigma}^{-1}\mathbf{U}^*\mathbf{B}_i^\intercal\mathbf{B}_i\mathbf{U}\mathbf{\Sigma}^{-2}\mathbf{U}^*\mathbf{B}_i^*\bar{\mathbf{B}}_i\mathbf{U}\mathbf{\Sigma}^{-1} \\
&\preceq \frac{1}{t}\sum_i\frac{\sum\tilde{\ell}_i}{\tilde{\ell}_i}\mathbf{\Sigma}^{-1}\mathbf{U}^*\mathbf{B}_i^*\left(\mathbf{B}_i\mathbf{U}\mathbf{\Sigma}^{-2}\mathbf{U}^*\mathbf{B}_i^*\right)\mathbf{B}_i\mathbf{U}\mathbf{\Sigma}^{-1} \\
&\preceq \frac{1}{t}\sum_i\frac{\sum\tilde{\ell}_i}{\tilde{\ell}_i}\cdot\ell_i\mathbf{\Sigma}^{-1}\mathbf{U}^*\mathbf{B}_i^*\mathbf{B}_i\mathbf{U}\mathbf{\Sigma}^{-1} \\
&\preceq \frac{\epsilon^2}{c\log(d/\delta)}\mathbf{\Sigma}^{-1}\mathbf{U}^*\mathbf{B}^*\mathbf{B}\mathbf{U}\mathbf{\Sigma}^{-1} \preceq \frac{\epsilon^2}{c\log(d/\delta)}\mathbf{I}
\end{aligned}$$

By the stable rank matrix Bernstein inequality, we have for large enough c:

$$P(\|\mathbf{Y}\|_2 > \epsilon) \leq \frac{4\text{tr}(\mathbf{I})}{\|\mathbf{I}\|}\exp\left(-\frac{\epsilon^2/2}{\frac{\epsilon^2}{c\log(d/\delta)}(\|\mathbf{I}\|+\epsilon/3)}\right) < \delta$$

where we use the fact that $tr(\mathbf{I}) \leq d$ and $\|\mathbf{I}\| = 1$.

□

2.2 PROOF OF Theorem 2

Proof. Let $\mathbf{B}^* = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ be the SVD of $\mathbf{B}^*$. We have $\ell_i = \mathbf{B}_i^*(\mathbf{U}\mathbf{\Sigma}^{-2}\mathbf{U}^*)\mathbf{B}_i$. Let $\mathbf{Y} = \mathbf{C}^\intercal\mathbf{C} - \mathbf{B}^\intercal\mathbf{B}$. Then we write

$$\mathbf{Y} = \sum_{j=1}^{t}\left(\mathbf{C}_j^\intercal\mathbf{C}_j - \frac{1}{t}\mathbf{B}^\intercal\mathbf{B}\right) \triangleq \sum_{j=1}^{t}\mathbf{X}_j$$

Note with probability p_i

$$\mathbf{X}_j = \frac{1}{t}\left(\frac{1}{p_i}\mathbf{B}_i^\intercal\mathbf{B}_i - \mathbf{B}^\intercal\mathbf{B}\right)$$

Now we bound the variance of $\mathbf{Y}$:

$$\begin{aligned}
\mathbb{E}[\mathbf{Y}^*\mathbf{Y}] &= t\mathbb{E}[\mathbf{X}_j^*\mathbf{X}_j] \\
&\preceq \frac{1}{t}\sum_i p_i\frac{1}{p_i^2}\mathbf{B}_i^*\bar{\mathbf{B}}_i\mathbf{B}_i^\intercal\mathbf{B}_i \\
&= \frac{1}{t}\sum_i\frac{\sum_i\tilde{\ell}_i}{\tilde{\ell}_i}\|\mathbf{B}_i\|^2\mathbf{B}_i^*\mathbf{B}_i \\
&\preceq \frac{\epsilon^2}{c\log(d/\delta)}\mathbf{I}
\end{aligned}$$

By the matrix Chernoff bound Gross and Nesme [2010], we have

$$\mathbf{Pr}(\|\mathbf{Y}\| > \epsilon) \leq 2d \exp\left(-\frac{\epsilon^2}{\frac{4\epsilon^2}{c\log(d/\delta)}}\right) = \mathcal{O}(\delta)$$

We remark that for our particular task, $\mathbf{Y}\mathbf{Y}^* = \mathbf{Y}^*\mathbf{Y}$. In general this is not true. By applying the non-Hermitian matrix Bernstein inequality in Tropp et al. [2015], one can derive the same result off by a multiplicative constant factor. □

2.3 THEORETICAL RESULTS ON THE HYBRID RANDOMIZED-DETERMINISTIC SAMPLING ALGORITHM

We first present a useful inequality from Cohen et al. [2015b] for subspace embeddings in the complex setting.

Lemma 1. *Let* $\mathbf{S}$ *be an* ϵ*-subspace embedding for span*$(\mathbf{A}, \mathbf{B})$*, where* $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{n\times d}$*. Then we have:*

$$\|\mathbf{A}^*\mathbf{S}^*\mathbf{S}\mathbf{B} - \mathbf{A}^*\mathbf{B}\| \leq \epsilon\|\mathbf{A}\|\|\mathbf{B}\|.$$

Proof of Lemma 1. W.l.o.g., we assume that $\|\mathbf{A}\| = \|\mathbf{B}\| = 1$, since we can divide both sides by $\|\mathbf{A}\|\|\mathbf{B}\|$. Let $\mathbf{U}$ be an orthonormal matrix of which the columns form a basis for span$(\mathbf{A}, \mathbf{B})$. Note since $\|\mathbf{A}\| = \|\mathbf{B}\| = 1$, for any $\mathbf{x}, \mathbf{y}$, we have $\mathbf{A}\mathbf{x} = \mathbf{U}\mathbf{s}$ and $\mathbf{B}\mathbf{y} = \mathbf{U}\mathbf{t}$ such that $\|\mathbf{s}\| \leq \|\mathbf{x}\|$ and $\|\mathbf{t}\| \leq \|\mathbf{y}\|$. Now:

$$\begin{aligned}
&\|\mathbf{A}^*\mathbf{S}^*\mathbf{S}\mathbf{B} - \mathbf{A}^*\mathbf{B}\| \\
=&\sup_{\|\mathbf{x}\|=\|\mathbf{y}\|=1} |\langle \mathbf{S}\mathbf{A}\mathbf{x}, \mathbf{S}\mathbf{B}\mathbf{y}\rangle - \langle \mathbf{A}\mathbf{x}, \mathbf{B}\mathbf{y}\rangle| \\
=&\sup_{\|\mathbf{s}\|,\|\mathbf{t}\|\leq 1} |\langle \mathbf{S}\mathbf{U}\mathbf{s}, \mathbf{S}\mathbf{U}\mathbf{t}\rangle - \langle \mathbf{U}\mathbf{s}, \mathbf{U}\mathbf{t}\rangle| \\
=&\|\mathbf{U}^*\mathbf{S}^*\mathbf{S}\mathbf{U} - \mathbf{I}\| \leq \epsilon
\end{aligned}$$

□

We are now ready to prove Theorem 3.

Proof of Theorem 3. If $\mathbf{x} \in \ker(\mathbf{D}_N^{1/2}\mathbf{A}_N)$, then the statement holds trivially. Assume without loss of generality that $\mathbf{x} \notin \ker(\mathbf{D}_N^{1/2}\mathbf{A}_N)$.

We first show that

$$\|\mathbf{A}_N^\intercal\mathbf{D}_N^{1/2}\mathbf{S}^\intercal\mathbf{S}\mathbf{D}_N^{1/2}\mathbf{A}_N - \mathbf{A}_N^\intercal\mathbf{D}_N\mathbf{A}_N\|_2 \leq \epsilon\|\mathbf{A}_N^\intercal|\mathbf{D}_N|\mathbf{A}_N\| \tag{1}$$

By Lemma 1, it suffices to show that $\mathbf{S}$ is a subspace embedding for span $\left(\mathbf{D}_N^{1/2}\mathbf{A}_N, (\mathbf{D}_N^{1/2})^*\mathbf{A}_N\right)$. Since $\mathbf{D}_N^{1/2}\mathbf{A}_N$ has the relaxed RIP, for $\mathbf{T}$ being a sampling matrix that randomly samples $t \triangleq O(d^2/\epsilon)$ rows of $\mathbf{D}_N^{1/2}\mathbf{A}_N$, we have:

$$\mathbf{Pr}\left(\forall \mathbf{x} : \|\mathbf{T}\mathbf{D}_N^{1/2}\mathbf{A}_N\mathbf{x}\|^2 = \rho^2(1\pm\epsilon)\|\mathbf{x}\|^2\right) \geq 1 - \frac{1}{n}$$

Since $\mathbf{S} = \sqrt{\frac{n}{t}}\mathbf{T}$, this leads to

$$\|\mathbf{S}\mathbf{D}_N^{1/2}\mathbf{A}_N\mathbf{x}\|^2 = \rho^2(1\pm\epsilon)\cdot\frac{n}{t}\|\mathbf{x}\|^2 = (1\pm\epsilon)\|\mathbf{D}_N^{1/2}\mathbf{A}_N\mathbf{x}\|^2$$

The reason for the last step is the following: we randomly partition $\mathbf{D}_N^{1/2}\mathbf{A}_N$ into $\frac{n}{t}$ chunks of rows, where each chunk has t rows. Denote the i^{th} chunk as $\mathbf{D}_{N_i}^{1/2}\mathbf{A}_{N_i}$ and correspondingly $\mathbf{A}_{N_i}$. By the relaxed RIP and union bound, we have with probability $1 - \frac{1}{t}$ that all $\frac{n}{t}$ chunks have $\|\mathbf{D}_{N_i}^{1/2}\mathbf{A}_{N_i}\mathbf{x}\|^2 = (1\pm\epsilon)\rho^2\|\mathbf{x}\|^2$. So in total:

$$\|\mathbf{D}_N^{1/2}\mathbf{A}_N\mathbf{x}\|^2 = \sum_{i=1}^{n/t}\|\mathbf{D}_{N_i}^{1/2}\mathbf{A}_{N_i}\mathbf{x}\|^2 = (1\pm\epsilon)\frac{n\rho^2}{t}\|\mathbf{x}\|^2$$

The same proof holds for showing $\mathbf{S}$ is an ϵ-subspace embedding for span $\left((\mathbf{D}_N^{1/2})^*\mathbf{A}_N\right)$.

Let $E = \cup_{i=1}^T E_i$. By (1) and the fact that $c\|\mathbf{A}_N^\intercal|\mathbf{D}_N|\mathbf{A}_N\| \leq \|\sum_i \mathbf{E}^i\|$:

$$\begin{aligned}
&\|\mathbf{A}_N^\intercal\mathbf{D}_N^{1/2}\mathbf{S}^\intercal\mathbf{S}\mathbf{D}_N^{1/2}\mathbf{A}_N - \mathbf{A}_N^\intercal\mathbf{D}_N\mathbf{A}_N\| \\
=&\|\mathbf{A}_N^\intercal\mathbf{D}_N^{1/2}\mathbf{S}^\intercal\mathbf{S}\mathbf{D}_N^{1/2}\mathbf{A}_N - \mathbf{A}_N^\intercal\mathbf{D}_N\mathbf{A}_N + \mathbf{A}_E^\intercal\mathbf{D}_E\mathbf{A}_E - \mathbf{A}_E^\intercal\mathbf{D}_E\mathbf{A}_E\| \\
=&\|\sum_{i=1}^T \mathbf{E}^i + \mathbf{A}_N^\intercal\mathbf{D}_N^{1/2}\mathbf{S}^\intercal\mathbf{S}\mathbf{D}_N^{1/2}\mathbf{A}_N - \mathbf{A}^\intercal\mathbf{D}\mathbf{A}\|_2 \\
\leq&\epsilon\|\mathbf{A}_N^\intercal|\mathbf{D}_N|\mathbf{A}_N\| \leq \frac{\epsilon}{c}\|\sum_i \mathbf{E}^i\| \\
\leq&\frac{\epsilon}{c-1}\left(\|\sum_i \mathbf{E}^i\| - \|\mathbf{A}_N^\intercal|\mathbf{D}_N|\mathbf{A}_N\|\right) \\
\leq&\frac{\epsilon}{c-1}\left(\|\sum_i \mathbf{E}^i\| - \|\mathbf{A}_N^\intercal\mathbf{D}_N\mathbf{A}_N\|\right) \\
\leq&\frac{\epsilon}{c-1}\|\mathbf{A}^\intercal\mathbf{D}\mathbf{A}\|
\end{aligned}$$

□

2.4 FAST COMPUTATION OF LEVERAGE SCORES

Despite the nice properties of leverage scores, they are data-dependent features and quite expensive to compute. In this section, we show how one can efficiently approximate all the leverage scores simultaneously.

Theorem 1. *Let $\mathbf{B} \in \mathbb{C}^{n\times d}$ and let $\mathbf{S} \in \mathbb{C}^{s\times n}$ be an ϵ-subspace embedding of span$(\mathbf{B})$. Let $\mathbf{SB} = \mathbf{QR}^{-1}$ be a QR-factorization of $\mathbf{SB}$, where $\mathbf{Q} \in \mathbb{C}^{s\times d}$ has orthonormal columns and $\mathbf{R}^{-1} \in \mathbb{C}^{d\times d}$. Let $\mathbf{G} \in \mathbb{R}^{d\times \log n}$ be a random Gaussian matrix. We define the i^{th} approximate leverage score to be: $\tilde{\ell}_i = \|\mathbf{e}_i^\intercal\mathbf{BRG}\|^2$ Then $\tilde{\ell}_i = (1\pm\epsilon)\ell_i$ for all i with high probability, and all $\tilde{\ell}_i$ can be calculated simultaneously in $\mathcal{O}\left((\mathbf{nnz}(\mathbf{A}) + d^2)\log n\right)$ time.*

Proof. Define

$$\ell_i' = \|\mathbf{e}_i^\intercal\mathbf{BR}\|^2$$

- We first show that $\ell_i' = O(1\pm\epsilon)\ell_i$ for all $i \in [n]$. Let $\mathbf{B} = \mathbf{U\Sigma V}^*$. Since $\mathbf{BR}$ has the same column space as $\mathbf{B}$, we have $\mathbf{BR} = \mathbf{UT}^{-1}$, for some matrix $\mathbf{T}$. We have:

  $$\|\mathbf{x}\| = \|\mathbf{Qx}\| = \|\mathbf{SBRx}\| = (1\pm\epsilon)\|\mathbf{BRx}\|$$

  Hence

  $$\|\mathbf{T}^{-1}\mathbf{x}\| = \|\mathbf{UT}^{-1}\mathbf{x}\| = \|\mathbf{BRx}\| = (1\pm O(\epsilon))\|\mathbf{x}\|$$

  This implies that $\mathbf{T}^{-1}$ is well-conditioned: all singular values of $\mathbf{T}^{-1}$ are of order $1\pm O(\epsilon)$. With this property:

  $$\begin{aligned}
  \ell_i' &= \|\mathbf{e}_i^\intercal\mathbf{BR}\|^2 = (1\pm O(\epsilon))\|\mathbf{e}_i^\intercal\mathbf{BRT}\|^2 \\
  &= (1\pm O(\epsilon))\|\mathbf{e}_i^\intercal\mathbf{U}\|^2 = (1\pm O(\epsilon))\ell_i
  \end{aligned}$$

- The second step is to show that $\tilde{\ell}_i = (1\pm\epsilon)\ell_i'$. Recall the Johnson-Lindenstrauss lemma: let $\mathbf{G}$ be as defined above. Then for all vectors $\mathbf{z} \in \mathbb{C}^d$:

  $$\mathbf{Pr}\left(\|\mathbf{z}^\intercal\mathbf{G}\|^2 = (1\pm\epsilon)\|\mathbf{z}\|^2\right) \geq 1-\delta$$

  We remark that the JL lemma holds for complex vectors $\mathbf{z}$ as in Krahmer and Ward [2011]. Now set $\mathbf{z} = \mathbf{e}_i^\intercal\mathbf{BR}$:

  $$\mathbf{Pr}\left(\|\mathbf{e}_i^\intercal\mathbf{BRG}\|^2 = (1\pm\epsilon)\|\mathbf{e}_i^\intercal\mathbf{BR}\|^2\right) \geq 1-\delta$$

  and we get the desired result.

- The time complexity for such a construction is the same as the construction for real matrices, which takes $\mathcal{O}((\mathbf{nnz}(\mathbf{A}) + d^2)\log n)$ time.

□

2.5 PROOF OF Theorem 4

Proof. By Lemma 1, we have that:

$$\|\mathbf{A}^\intercal\mathbf{T}^\intercal\mathbf{T}\mathbf{D}\mathbf{A} - \mathbf{A}^\intercal\mathbf{D}\mathbf{A}\| \leq \epsilon\|\mathbf{A}\|\|\mathbf{D}\mathbf{A}\|$$

and

$$\|\mathbf{A}^\intercal\mathbf{D}^{1/2}\mathbf{S}^\intercal\mathbf{S}\mathbf{D}^{1/2}\mathbf{A} - \mathbf{A}^\intercal\mathbf{D}\mathbf{A}\| \leq \epsilon\|\mathbf{D}^{1/2}\mathbf{A}\|\|(\mathbf{D}^{1/2})^*\mathbf{A}\|\| = \epsilon\|\mathbf{D}^{1/2}\mathbf{A}\|_2^2$$

Note that

$$\begin{aligned}&\|\mathbf{D}^{1/2}\mathbf{A}\|_2^2 = \lambda_{\max}(\mathbf{A}^\intercal(\mathbf{D}^{1/2})^*\mathbf{D}^{1/2}\mathbf{A})\\ =&\lambda_{\max}(\mathbf{A}^\intercal|\mathbf{D}|\mathbf{A}) = \|\mathbf{A}^\intercal|\mathbf{D}|\mathbf{A}\|_2\\ \leq&\|\mathbf{A}\|\||\mathbf{D}|\mathbf{A}\| = \|\mathbf{A}\|\|\mathbf{D}\mathbf{A}\|\end{aligned}$$

So sampling in the latter way is always as good as the former.

Now we give a simple example that the first sampling scheme can give an arbitrarily worse bound. Let $\mathbf{A} = \begin{bmatrix} a_1 & 0 \\ 0 & a_2 \end{bmatrix}$ and $\mathbf{D} = \begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix}$, where $1 = a_1 > a_2$ and $1 = |d_1| < |d_2|$.

Hence $\mathbf{D}\mathbf{A} = \begin{bmatrix} a_1 d_1 & 0 \\ 0 & a_2 d_2 \end{bmatrix}$ and $\mathbf{A}^\intercal|\mathbf{D}|\mathbf{A} = \begin{bmatrix} |d_1|a_1^2 & 0 \\ 0 & |d_2|a_2^2 \end{bmatrix}$

By the above calculation, $\|\mathbf{D}^{1/2}\mathbf{A}\|_2^2 = \max\{1, |d_2|a_2^2\}$, and $\|\mathbf{A}\|\|\mathbf{D}\mathbf{A}\| = \max\{1, |d_2|a_2\}$. Let $a_2 = \Theta(\sqrt{1/|d_2|})$ and making $|d_2|$ arbitrarily large, we then have $\|\mathbf{A}\|\|\mathbf{D}\mathbf{A}\| \gg \|\mathbf{D}^{1/2}\mathbf{A}\|_2^2$. □

2.6 FAST LOCAL CONVERGENCE OF NEWTON-CG

Theorem 2 (Fast Local Convergence)**.** *Let $\mathbf{S}$ be the leverage score sampling matrix as in Theorem 1 with precision ϵ. Let $r(\mathbf{x}) = \lambda\|\mathbf{x}\|^2/2$ and $\lambda \geq 4\|\mathbf{A}\|^2 h$ where h is the Lipschitz continuity constant of the derivative, i.e., $|f_i''(t)| \leq h$ for some $h < \infty$. Then for sub-sampled Newton-CG with initial point satisfying $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq \mu/(4L)$, step-size $\alpha_k = 1$, and the approximate Hessian $\mathbf{H} = \mathbf{A}^\intercal\mathbf{D}^{1/2}\mathbf{S}^\intercal\mathbf{S}\mathbf{D}^{1/2}\mathbf{A} + \lambda\mathbf{I}$, we have the following error recursion $\|\mathbf{x}_k - \mathbf{x}^\star\| \leq C_q \cdot \|\mathbf{x}_k - \mathbf{x}^\star\|^2 + C_l \cdot \|\mathbf{x}_k - \mathbf{x}^\star\|$, where $\mathbf{x}^\star$ is the optimal solution, $C_q = \frac{2L}{(1-O(\epsilon))\mu}$, $C_l = \frac{3\epsilon}{1-O(\epsilon)}\sqrt{\kappa}$, L is the Lipschitz continuity constant of the Hessian, $\mu = \lambda_{\min}\left(\nabla^2 F(\mathbf{x}^\star)\right) > 0$, $\nu = \lambda_{\max}\left(\nabla^2 F(\mathbf{x}^\star)\right) < \infty$, and $\kappa = \nu/\mu$ is the condition number.*

Proof. Let $\mathbf{B} = \mathbf{D}^{1/2}\mathbf{A}$, $\mathbf{S}$ be the sketching matrix, and $\mathbf{C} = \mathbf{S}\mathbf{D}^{1/2}\mathbf{A}$. By Theorem 1, we have:

$$-\epsilon\mathbf{A}^\intercal|\mathbf{D}|\mathbf{A} \preceq \mathbf{A}^\intercal\mathbf{D}^{1/2}\mathbf{S}^\intercal\mathbf{S}\mathbf{D}^{1/2}\mathbf{A} - \mathbf{A}^\intercal\mathbf{D}\mathbf{A} \preceq \epsilon\mathbf{A}^\intercal|\mathbf{D}|\mathbf{A}. \tag{2}$$

Rewrite

$$\mathbf{A}^\intercal|\mathbf{D}|\mathbf{A} = \sum_{i=1}^n |\mathbf{D}_{i,i}|\mathbf{A}_i^\intercal\mathbf{A} = \sum_{i=1}^n \mathbf{D}_{i,i}\mathbf{A}_i^\intercal\mathbf{A} - 2\sum_{i:\mathbf{D}_{i,i}<0} \mathbf{D}_{i,i}\mathbf{A}_i^\intercal\mathbf{A} \preceq \mathbf{A}^\intercal\mathbf{D}\mathbf{A} + \mathbf{Q}$$

where $\mathbf{Q} \triangleq \lambda\mathbf{I}$ as defined in the theorem. The above inequality then holds by the definition of λ. Therefore, by (2) we have

$$-\epsilon(\mathbf{A}^\intercal\mathbf{D}\mathbf{A} + \mathbf{Q}) \preceq (\mathbf{A}^\intercal\mathbf{D}^{1/2}\mathbf{S}^\intercal\mathbf{S}\mathbf{D}^{1/2}\mathbf{A} + \mathbf{Q}) - (\mathbf{A}^\intercal\mathbf{D}\mathbf{A} + \mathbf{Q}) \preceq \epsilon(\mathbf{A}^\intercal\mathbf{D}\mathbf{A} + \mathbf{Q}).$$

This form satisfies the fast convergence condition in [Xu et al., 2016, Lemma 7]. Applying their lemma leads to our conclusion. □

3 SKETCHING FOR OPTIMIZATION–MORE DETAILS AND EXPERIMENTS

More Background on Some Optimization Methods

– **Convex Optimization: Sub-sampled Newton-CG.** In strongly convex settings where $\nabla^2 F(\mathbf{x}) \succeq \mu \mathbf{I}$ for some $\mu > 0$, the Hessian matrix is positive definite, and the k^{th} iteration of the sub-sampled Newton-CG method is often written as $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$, where $\mathbf{p}_k$ is an approximate solution to the linear system $\mathbf{H}_k \mathbf{p} = -\nabla F(\mathbf{x}_k)$, obtained using the conjugate gradient (CG) algorithm Saad [2003], and $0 < \alpha_k \leq 1$ is an appropriate step-size, which satisfies the Armijo-type line search Nocedal and Wright [2006] condition stating that $F(\mathbf{x}_{k+1}) \leq F(\mathbf{x}_k) + \rho \alpha_k \langle \mathbf{p}_k, \mathbf{g}_k \rangle$, where $0 < \rho < 1$ is a given line-search parameter (see Algorithm 1 in Section 1).

– **Non-convex Optimization: Sub-sampled Newton-MR.** In non-convex settings, the Hessian matrix could be indefinite and possibly rank-deficient. In light of this, in the k^{th} iteration, Newton-MR Roosta et al. [2018] with an approximate Hessian involves iterations of the form $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$ where $\mathbf{p}_k \approx -[\mathbf{H}_k]^{\dagger} \nabla F(\mathbf{x}_k)$ is obtained by a variety of least-squares iterative solvers such as MINRES-QLP Choi et al. [2011], and $0 < \alpha_k \leq 1$ is such that $\|\mathbf{g}_{k+1}\|^2 \leq \|\mathbf{g}_k\|^2 + 2\rho\alpha_k \langle \mathbf{p}_k, \mathbf{H}_k \mathbf{g}_k \rangle$ (see Algorithm 2 in Section 1). It has been shown that Newton-MR achieves fast local and global convergence rates when applied to a class of non-convex problems known as invex Roosta et al. [2018], whose stationary points are global minima. From Liu and Roosta [2019, Corollary 1] with ϵ small enough in (4), Algorithm 2 converges to an ϵ_g-approximate first-order stationary point $\|\nabla F(\mathbf{x}_k)\| \leq \epsilon_g$ in at most $k \in \mathcal{O}\left(\log\left(1/\epsilon_g\right)\right)$ iterations. Every iteration of MINRES-QLP requires one Hessian-vector product, which using the full Hessian, amounts to a complexity of $\mathcal{O}\left(\mathbf{nnz}(\mathbf{A})\right)$. In the worst case, MINRES-QLP requires $\mathcal{O}(d)$ iterations to obtain a solution. Putting this all together, the overall running time of Newton-MR with exact Hessian to achieve an ϵ_g-approximate first-order stationary point is $k \in \mathcal{O}\left(\mathbf{nnz}(\mathbf{A}) d \log\left(1/\epsilon_g\right)\right)$. However, with the complex leverage score sampling of Algorithm 1 (cf. Theorem 1), the running time then becomes $k \in \mathcal{O}\left(\left(\mathbf{nnz}(\mathbf{A}) \log n + d^3\right) \log\left(1/\epsilon_g\right)\right)$.

– **Non-convex Optimization: Sub-sampled Trust Region.** As a more versatile alternative to line-search, trust-region Sorensen [1982], Conn et al. [2000] is an elegant globalization strategy that has attracted much attention. Recently, Xu et al. [2019] theoretically studied the variants of trust-region in which the Hessian is approximated as in (4). The crux of each iteration of the resulting algorithm is the (approximate) solution to a constrained quadratic sub-problem of the form $\min_{\|\mathbf{p}\| \leq \Delta_k} m_k(\mathbf{p}) \triangleq \langle \nabla F(\mathbf{x}_k), \mathbf{p} \rangle + \frac{1}{2} \langle \mathbf{p}, \mathbf{H}_k \mathbf{p} \rangle$, for which a variety of methods exists, e.g, CG-Steihaug Steihaug [1983], Toint [1981], and the generalized Lanczos based methods Gould et al. [1999], Lenders et al. [2016] (see Algorithm 3 in Section 1). Suppose for $i \in [n]$, $\|\mathbf{a}_i\|^2 \sup_{\mathbf{x} \in \mathbb{R}^d} |f_i''(\mathbf{x})| \leq K_i$ and define $K_{\max} \triangleq \max_{i=1,\ldots,n} K_i$, $\widehat{K} \triangleq \sum_{i=1}^n K_i / n$. By considering uniform and row-norm sampling of $\mathbf{D}^{1/2}\mathbf{A}$ with respective sampling complexities of $|\mathcal{S}| \in \mathcal{O}(K_{\max}^2 \epsilon^{-2} \log d)$ and $|\mathcal{S}| \in \mathcal{O}(\widehat{K}^2 \epsilon^{-2} \log d)$, Xu et al. [2019] showed that one can guarantee (4) with high-probability, and as a result Algorithm 3 achieves an optimal iteration complexity, i.e., it converges to an $(\epsilon_g, \epsilon_{\mathbf{H}})$-approximate second-order stationary point $\|\nabla F(\mathbf{x}_k)\| \leq \epsilon_g$ and $\lambda_{\min}(\nabla^2 F(\mathbf{x}_k)) \geq -\epsilon_{\mathbf{H}}$ in at most $k \in \mathcal{O}(\max\{\epsilon_g^{-2} \epsilon_{\mathbf{H}}^{-1}, \epsilon_{\mathbf{H}}^{-3}\})$ iterations.

Sub-sampling Schemes. Recall the following terms:

– **Uniform**: For this sampling, we have $p_i = 1/n,\ i = 1, \ldots, n$.

– **Leverage Score (LS)**: Complex leverage score sampling by considering the leverage scores of $\mathbf{D}^{1/2}\mathbf{A}$ as in Algorithm 1.

– **Row Norm (RN)**: Row-norm sampling of $\mathbf{D}^{1/2}\mathbf{A}$ using (3) where $s((\mathbf{D}^{1/2}\mathbf{A})_i) = |f_i''(\langle \mathbf{a}_i, \mathbf{x} \rangle)| \, \|\mathbf{a}_i\|_2^2$

– **Mixed Leverage Score (LS-MX)**: A mixed leverage score sampling strategy arising from a non-symmetric viewpoint of the product $\mathbf{A}^{\intercal}(\mathbf{DA})$ using (2) with $s(\mathbf{A}_i) = \ell_i(\mathbf{A})$ and $S((\mathbf{DA})_i) = \ell_i(\mathbf{DA})$.

– **Mixed Norm Mixture (RN-MX)**: A mixed row-norm sampling strategy with the same non-symmetric viewpoint as in (2) with $s(\mathbf{A}_i) = \|(\mathbf{A})_i\|$ and $S((\mathbf{DA})_i) = \|(\mathbf{DA})_i\|$.

– **Hybrid Randomized-Deterministic (LS-Det)**: Hybrid deterministic-leverage score sampling of Algorithm 2.

– **Full**: In this case, the exact Hessian is used.

Datasets. The datasets used in our experiments for this section are listed in Table 1. All datasets are publicly available from the UC Irvine Machine Learning Repository Dua and Graff [2017].

Name	n	d
`Drive Diagnostics`	50,000	48
`covertype,`	581,012	54
`UJIIndoorLoc`	19,937	520

Table 1: Data sets used for our experiments.

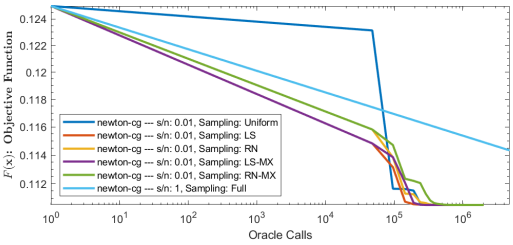
Hyper-parameters. Algorithms 1 to 3 are always initialized at $\mathbf{x}_0 = \mathbf{0}$. In all of our experiments, we run each method until either a maximum number of iterations or a maximum number of function evaluations is reached. The maximum number of CG iterations within Newton-CG, MINRES-QLP iterations within Newton-MR and CG-Steihaug within trust-region methods are all set to 100. The parameter of line-search ρ in Newton-MR is set to 10^{-4}. For trust-region, we set $\Delta_0 = 1$, $\eta = 0.8$ and $\gamma = 1.2$.

Performance Evaluation. In all of our experiments, we plot the objective value or the gradient norm vs. the total number of oracle calls of function, gradient, and Hessian-vector products. This is because comparing algorithms in terms of "wall-clock" time can be highly affected by their particular implementation details as well as system specifications. In contrast, counting the number of oracle calls, as an implementation and system independent unit of complexity, is most appropriate and fair. More specifically, after computing each function value, computing the corresponding gradient is equivalent to one additional function evaluation. Our implementations are Hessian-free, i.e., we merely require Hessian-vector products instead of using the explicit Hessian. For this, each Hessian-vector product involving $\mathbf{ADA}$ amounts to two additional function evaluations, as compared with gradient evaluation. In this light, each matrix-vector product involving $\mathbf{D}^{1/2}\mathbf{A}$ for approximating the underlying complex leverage scores is equivalent to one gradient evaluation.

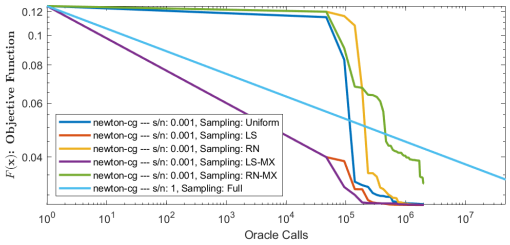
Following the theory of Newton-MR, whose convergence is measured by the norm of the gradient, we evaluate Algorithm 2 with various sampling schemes by plotting $\|\nabla F(\mathbf{x}_k)\|$ vs. the total number of oracle calls, whereas for Algorithms 1 and 3, which guarantees descent in objective function, we plot $F(\mathbf{x}_k)$ vs. the total number of oracle calls.

3.1 COMPARISON AMONG VARIOUS SKETCHING TECHNIQUES

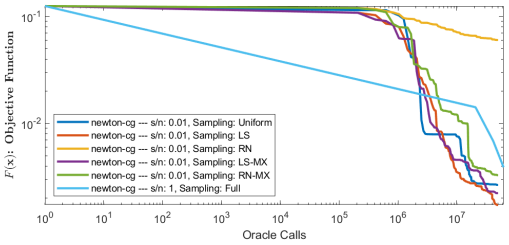
To verify the result of Theorem 4, in this section we present empirical evaluations of Uniform, LS, RN, LS-MX, RN-MX and Full in the context of Algorithms 1 to 3. The results are depicted in Figures 1 to 2. It can be clearly seen that for both algorithms, LS and LS-MX sampling amounts to a more efficient algorithm than that with RN and RN-MX variants, and at times this difference is more pronounced than other times.



(a) $F(x_k)$ vs. Oracle calls (`Drive Diagnostics`)

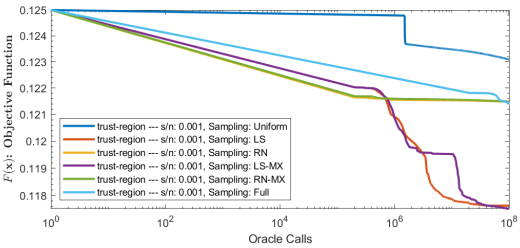


(b) $F(x_k)$ vs. Oracle calls (`Cover Type`)

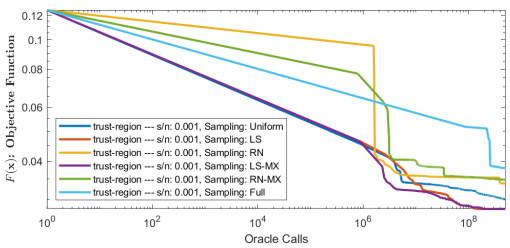


(c) $F(x_k)$ vs. Oracle calls (`UJIIndoorLoc`)

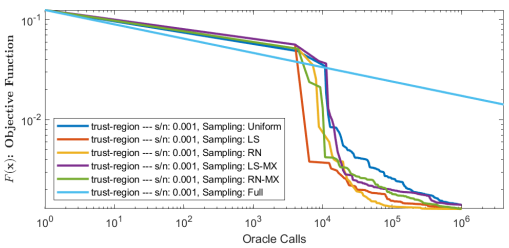
Figure 1: Comparison of Newton-CG (Algorithm 1) using various sampling schemes.



(a) $F(x_k)$ vs. Oracle calls (`Drive Diagnostics`)



(b) $F(x_k)$ vs. Oracle calls (`Cover Type`)



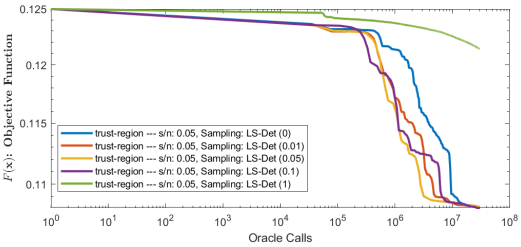
(c) $F(x_k)$ vs. Oracle calls (`UJIIndoorLoc`)

Figure 2: Comparison of Trust-region (Algorithm 3) using various sampling schemes.

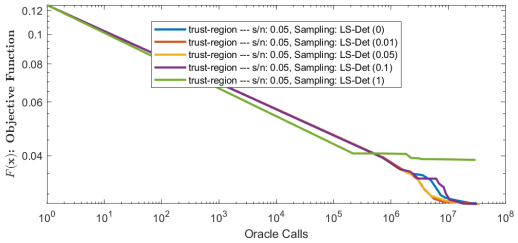
3.2 EVALUATION OF HYBRID SKETCHING TECHNIQUES

Here, to verify the result of Theorem 3, we evaluate the performance of Algorithm 3 by varying the terms involved in $\mathbf{E}$. We do this for a simple splitting of $\mathbf{H} = \mathbf{E} + \mathbf{N}$, i.e., $T = 1$ in Theorem 3. We fix the overall sample size and change the fraction of samples that are deterministically picked in $\mathbf{E}$. The results are depicted in Figure 3. The value in brackets in front of LS-Det is the fraction of samples that are included in $\mathbf{E}$, i.e., deterministic samples. "LS-Det (0)" and "LS-Det (1)" correspond to $\mathbf{E} = \mathbf{0}$ and $\mathbf{N} = \mathbf{0}$, respectively. The latter strategy has been used in low rank matrix approximations McCurdy [2018]. As it can be seen, the hybrid sampling approach is always competitive with, and at times significantly

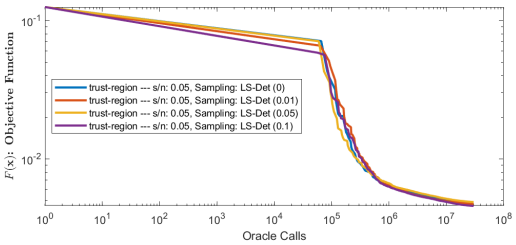
better than, LS-Det (0). As expected, LS-Det (1), which amounts to entirely deterministic samples, consistently performs worse. This can be easily attributed to the high bias of such a deterministic estimator.



(a) $F(x_k)$ vs. Oracle calls (`Drive Diagnostics`)



(b) $F(x_k)$ vs. Oracle calls (`Cover Type`)



(c) $F(x_k)$ vs. Oracle calls (`UJIIndoorLoc`)

Figure 3: Comparison of Algorithm 3 using hybrid randomized-deterministic sampling schemes. For all runs, the overall sample/mini-batch size for estimating the Hessian matrix is $s = 0.05n$. The values in parentheses in front of LS-Det is the fraction of samples that are taken deterministically and included in $\mathbf{E}$.

References

Sou-Cheng T Choi, Christopher C Paige, and Michael A Saunders. MINRES-QLP: A Krylov subspace method for indefinite or singular symmetric systems. *SIAM Journal on Scientific Computing*, 33(4):1810–1836, 2011.

Michael B Cohen, Yin Tat Lee, Cameron Musco, Christopher Musco, Richard Peng, and Aaron Sidford. Uniform sampling for matrix approximation. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pages 181–190, 2015a.

Michael B Cohen, Jelani Nelson, and David P Woodruff. Optimal approximate matrix product in terms of stable rank. *arXiv preprint arXiv:1507.02268*, 2015b.

Michael B Cohen, Cameron Musco, and Christopher Musco. Input sparsity time low-rank approximation via ridge leverage score sampling. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1758–1777. SIAM, 2017.

Andrew R Conn, Nicholas IM Gould, and Philippe L Toint. *Trust region methods*. SIAM, 2000.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Nicholas IM Gould, Stefano Lucidi, Massimo Roma, and Philippe L Toint. Solving the trust-region subproblem using the Lanczos method. *SIAM Journal on Optimization*, 9(2):504–525, 1999.

David Gross and Vincent Nesme. Note on sampling without replacing from a finite collection of matrices. *arXiv preprint arXiv:1001.2738*, 2010.

Felix Krahmer and Rachel Ward. New and improved johnson–lindenstrauss embeddings via the restricted isometry property. *SIAM Journal on Mathematical Analysis*, 43(3):1269–1281, 2011.

Felix Lenders, Christian Kirches, and Andreas Potschka. trlib: A vector-free implementation of the gltr method for iterative solution of the trust region problem. *arXiv preprint arXiv:1611.04718*, 2016.

Yang Liu and Fred Roosta. Stability Analysis of Newton-MR Under Hessian Perturbations. *arXiv preprint arXiv:1909.06224*, 2019.

Shannon McCurdy. Ridge regression and provable deterministic ridge leverage score sampling. In *Advances in Neural Information Processing Systems*, pages 2463–2472, 2018.

Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

Farbod Roosta, Yang Liu, Peng Xu, and Michael W Mahoney. Newton-MR: Newton's Method Without Smoothness or Convexity. *arXiv preprint arXiv:1810.00303*, 2018.

Yousef Saad. *Iterative methods for sparse linear systems*, volume 82. SIAM, 2003.

Danny C Sorensen. Newton's method with a model trust region modification. *SIAM Journal on Numerical Analysis*, 19(2): 409–426, 1982.

Trond Steihaug. The conjugate gradient method and trust regions in large scale optimization. *SIAM Journal on Numerical Analysis*, 20(3):626–637, 1983.

Philippe L Toint. Towards an efficient sparsity exploiting Newton method for minimization. *Sparse matrices and their uses*, page 1981, 1981.

Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8 (1-2):1–230, 2015.

Peng Xu, Jiyan Yang, Farbod Roosta, Christopher Ré, and Michael W Mahoney. Sub-sampled newton methods with non-uniform sampling. In *Advances in Neural Information Processing Systems*, pages 3000–3008, 2016.

Peng Xu, Farbod Roosta, and Michael W Mahoney. Newton-type methods for non-convex optimization under inexact Hessian information. *Mathematical Programming*, 2019. doi:10.1007/s10107-019-01405-z.