RandNLA for GLMs with Big Datasets

On the Edge of Inference and Computation

Robert Tjarko Lange

MSc Data Science - Barcelona Graduate School of Economics

Contact: robert.lange@barcelonagse.eu || www.rob-lange.com

Checkout the Code on my GitHub page!

Abstract

This project digs into the potential of randomized algorithms for tall data analysis. In particular, we focus on the implementation of fast approximations to the statistical leverage scores and the properties of the resulting random sampling estimators. These "algorithmic leveraging" estimators (Drineas et al., 2011, 2012) can be used to decrease computational complexity by effectively reducing the dimensionality of the underlying normal equations problem. This analysis extends the established results to the generalized linear model (GLM) setting, where the number of observations (n) is much larger than the number of features (d). Furthermore, we provide a deeper statistical understanding of the resulting estimators. We show that it is possible to obtain similar quality of approximation results as in the simple least squares (LS) case and discuss the convergence behavior of the resulting randomized iterative weighted least squares (IWLS) estimator.

Algorithmic Leveraging: A Generalized Scheme

In general, we are interested in reducing a high-dimensional problem (such as solving a system of normal or likelihood equations where the amount of constraints is much larger than the amounts of variables: $n >> d$) to a much smaller dimensional space. While doing so, we want to maintain as much structural information necessary to obtain a good quality of approximation. Drineas et al. (2011, 2012) proposed a randomized estimator, which samples rows proportional to a fast approximation to the statistical leverage scores. By choosing the amount of row according to a concentration inequality result, one is able to reduce the original complexity of LS or solving likelihood equations from $O(nd^2)$ to $O(ndlog(r))$ where $log(r) < d$. Given $X \in \mathbb{R}^{n \times d}$, $y \in \mathbb{R}^n$ and an ϵ-level of approx. a general static algorithmic leveraging algorithm proceeds in the following way:

1. Construct approx. statistical leverage scores (or another influence measure) in a fast manner.
2. Normalize the approximate influence measure and obtain a discrete sampling distribution.
3. Sample (w. repl.) a specific amount of rows from $(X, y) \in \mathbb{R}^{n \times (d+1)} \Rightarrow (\tilde{X}, \tilde{y}) \in \mathbb{R}^{r \times (d+1)}$.
4. Solve this reduced system of normal equations to obtain an approximate solution vector, $\tilde{\beta}$.

Fast Leverage Score Approximation

Fast approx. algorithms make use of high-dimensional subspace embedding concepts such as the Fast Johnson-Lindenstrauss Transform (FJLT). Unlike the standard JLT, structured random projections are in terms a complete subspace instead of a discrete set of points. The formal definition is the following:

Definition 1 (Fast Subspace Johnson-Lindenstrauss Transform - Drineas (2012, p. 3448))**.** *Let $\epsilon > 0$ and $U \in \mathbb{R}^{n \times d}$ be an orthogonal matrix, viewed as d vectors in $\mathbb{R}^n$. A ϵ-FJLT or subspace Johnson-Lindenstrauss transform maps vectors from $\mathbb{R}^n \rightarrow R^r$ such that the orthogonality of U is preserved. $\Pi \in \mathbb{R}^{r \times n}$ is called an ϵ-FJLT if*

- **Orthogonality preservation:** $||I_d - U^T \Pi^T \Pi U||_2 \leq \epsilon$
- **Fast running time:** $\forall X \in \mathbb{R}^{n \times d}$ we can compute ΠX in $O(ndlog(r))$ time.

By preprocessing with a Hadamard-Walsh transform, one is able to generate a fast sketch of X. Applying a thin SVD to this sketch and Afterwards, we take the euclidean row norm of the matrix of sketched left-singular value matrix, $\tilde{U}^{(X)}$, which yields a fast approx. to the leverage scores.

Algorithm 1 Drineas et al. (2012, p. 3451) - FJLT approximation for leverage scores

Input: $X \in \mathbb{R}^{n \times d}$ with SVD $X = U^{(X)} \Sigma V^T$ and an ϵ-level of approximation
Output: Approximate leverage scores, $\tilde{l}_i, i = 1, ..., n$

1: Let $\Pi_1 \in \mathbb{R}^{r_1 \times n}$ be an ϵ-FJLT for $U^{(X)}$ with $r_1 = \Omega\left(\frac{dlog(n)}{\epsilon^2} log\left(\frac{dlog(n)}{\epsilon^2}\right)\right)$.
2: Compute $\tilde{X} = \Pi_1 X$ and its SVD/QR where $\tilde{R} = \Sigma^{\tilde{X}} V^{\tilde{X}|T}$.
3: View rows of $X\tilde{R}^{-1} \in \mathbb{R}^{n \times d}$ as n vectors in $\mathbb{R}^d$. Let $\Pi_2 \in \mathbb{R}^{d \times r_2}$ be an ϵ-JLT for n^2 vectors, with $r_2 = O(\frac{log(n)}{\epsilon^2})$.
4: **return** $\tilde{l}_i = ||(XR^{-1}\Pi_2)_i||_2^2$, an ϵ-approximation of l_i.

- Computational Complexity: $O(ndlog(d/\epsilon) + nd\epsilon^{-2}log(n) + d^3\epsilon^{-2}log(n)log(d\epsilon^{-1}))$
- Same concentration result can be obtained without $\Pi_{JLT} \rightarrow$ But: Running time improvements
- Orthog. preservation condition $\Leftrightarrow$ Preprocessing by random version of identity: $\mathbb{E}(\Pi^T \Pi) = I_n$
- For Hadamard-based construction of Π_{FJLT} please consult Drineas et al. (2012).

Approximate Iterative Weighted Least Squares

IWLS Problem Formulation:

$W, X \in \mathbb{R}^{n \times n}$, $z \in \mathbb{R}^n$. Iteratively solve $\beta_{(k+1)} = (X^T W_{(k)} X)^{-1} X^T W_{(k)} z_{(k)}$ until convergence.

$\rightarrow$ Linear predictor: $\eta = X\beta$ and $\mu = (\mu_1, ..., \mu_n)$ where $\mu_i = \mathbb{E}(y_i)$

$\rightarrow$ Working variates: $z_{(k)} = X\beta_{(k-1)} + (y - \mu) diag\left(\frac{\partial \eta_i}{\partial \mu_i}\right)$

$\rightarrow$ Weighting matrix: $W_{(k)} = diag(w_{1|(k)}, ..., w_{n|(k)})$ where $w_{i|(k)} = Var(\mu_i)^{-1}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$

Randomized Problem Formulation:

$\Pi_{(k)} \in \mathbb{R}^{r \times n}$: Random sampling matrix constructed at each iteration $\rightarrow$ each row contains one non-zero and rescaled value. $\tilde{W}$: Diag. matrix of weights that correspond to the subsampled rows.

$$\left((\Pi_{(k)}X)^T \tilde{W}_{(k)} \Pi_{(k)} X\right) \beta_{(k+1)} = \left((\Pi_{(k)}X)^T \tilde{W}_{(k)} \Pi_{(k)} z_{(k)}\right)$$
$$\tilde{X}^T \tilde{W}_{(k)} \tilde{X} \tilde{\beta}_{(k+1)} = \tilde{X}^T \tilde{W}_{(k)} \tilde{z}_{(k)}$$

Theorem 1 (Asymptotic Converge of RandIWLS Algorithm)**.** *For a random sampling IWLS estimator, which iteratively constructs unbiased random approximations, it can be shown that*

$$\lim_{k\to\infty} \hat{z}_{(k)} = \lim_{k\to\infty} \prod_{j=1}^{k-1} (1 \pm \epsilon_j) z_{(k)} = z_{(k)} \text{ and } \lim_{k\to\infty} \hat{W}_{(k)} = \lim_{k\to\infty} \prod_{j=1}^{k-1} (1 \pm \epsilon_j) W_{(k)} = W_{(k)}$$
$$\lim_{k\to\infty} \tilde{\eta} = \lim_{k\to\infty} \prod_{j=1}^{k-1} (1 \pm \epsilon_j) X\beta_{(k)} = X\beta_{(k)} \text{ and } \tilde{\mu} = \mathbb{E}(X \prod_{j=1}^{k} (1 \pm \epsilon_j)\beta_{(k)}) = \mathbb{E}(X\beta_{(k)})$$

The estimator asymptotically converges to the converged IWLS estimator, $\lim_{k\to\infty} \tilde{\beta}_{(k)} = \beta_{GLM}$.

- IWLS solves a weighted normal/likelihood equation problem at each iteration.
- We adapt the LS approach of Drineas et al. (2011, 2012) to this dynamic setting.
- Challenge: Convergence behavior due to additional introduction of sampling variance.

$\Rightarrow$ Possible Solution: Introduce dynamic influence measures according to which we sample.

Dynamic Influence Measures for Iterative Schemes

Simulations in the static LS case showed, that one could obtain decreases in the quality of approximation (QoA) variance by sampling according to approx. influence scores. We compare three different types of non-adaptive and adaptive influence measures:

Definition 2 (Statistical Leverage Score (Drineas et al., 2012, p. 3450))**.** *Let $X \in \mathbb{R}^{n \times d}$ and its SVD/QR decomposition be denoted by $X = U^{(X)} \Sigma V^T = Q^{(X)} R$. Furthermore, let e_i denote a standard basis vector acting as an indicator for row i. Also let $X^\dagger = V\Sigma^{-1} U^{(X)T}$ denote the generalized Moore-Penrose inverse. The leverage scores of X are then defined to be*

$$l_i = \left(X(X^TX)^{-1}X^T\right)_{ii} = ||U_i^{(X)}||_2^2 = ||Q_i^{(X)}||_2^2 .$$

Definition 3 (Weighted Influence Scores (Jinzhu Jia (2014)))**.** *At iteration k of the Randomized IWLS scheme the weighted leverage scores of X are defined as $WL_i = ||w_{i|(k)} U_i^{(X)}||_2^2$.*

Definition 4 (Working Variate Influence Score)**.** *At iteration k of the Randomized IWLS scheme the working variate influence scores are defined as the leverage scores of the concatenated matrix $(X, z_{(k)}) \in \mathbb{R}^{n \times (d+1)}$, which we denote as $wvI_i = ||U_i^{(X, z_{(k)})}||_2^2$.*

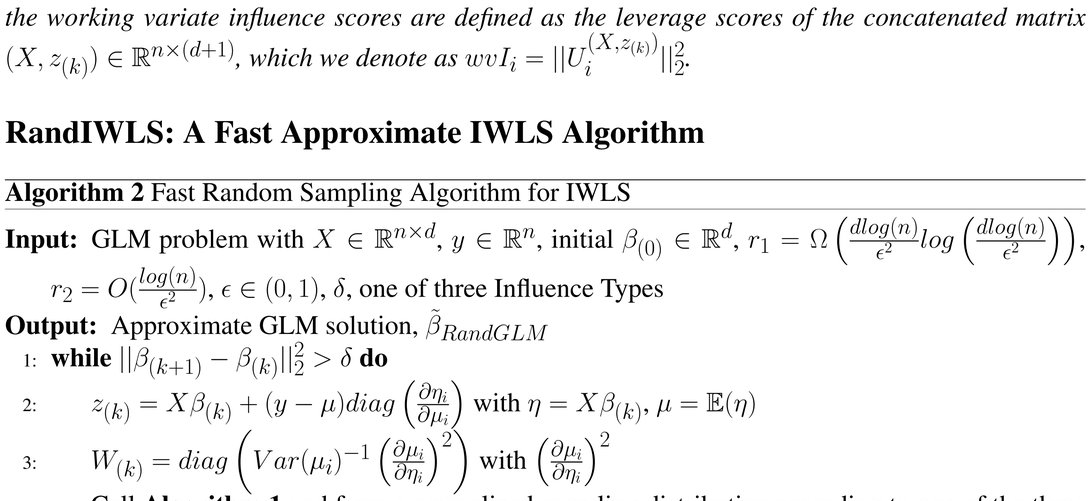
RandIWLS: A Fast Approximate IWLS Algorithm

Algorithm 2 Fast Random Sampling Algorithm for IWLS

Input: GLM problem with $X \in \mathbb{R}^{n \times d}$, $y \in \mathbb{R}^n$, initial $\beta_{(0)} \in \mathbb{R}^d$, $r_1 = \Omega\left(\frac{dlog(n)}{\epsilon^2} log\left(\frac{dlog(n)}{\epsilon^2}\right)\right)$, $r_2 = O(\frac{log(n)}{\epsilon^2})$, $\epsilon \in (0, 1)$, δ, one of three Influence Types
Output: Approximate GLM solution, $\tilde{\beta}_{RandGLM}$

1: **while** $||\beta_{(k+1)} - \beta_{(k)}||_2^2 > \delta$ **do**
2: $z_{(k)} = X\beta_{(k)} + (y - \mu) diag\left(\frac{\partial \eta_i}{\partial \mu_i}\right)$ with $\eta = X\beta_{(k)}$, $\mu = \mathbb{E}(\eta)$
3: $W_{(k)} = diag\left(Var(\mu_i)^{-1}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\right)$ with $\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$
4: Call **Algorithm 1** and form a normalized sampling distribution according to one of the three proposed static/dynamic influence measures.
5: Randomly sample $r = O(\frac{dlog(d)}{\epsilon})$ rows of X, W and z. Rescale them by $\frac{1}{\sqrt{rp_i}}$, form $\tilde{X} \in R^{r \times d}, \tilde{W}_{(k)} \in \mathbb{R}^{r \times r}, \tilde{z}_{(k)} \in R^r$.
6: Solve $\beta_{(k)} = (\tilde{X}^T \tilde{W}_{(k)} \tilde{X})^{-1} \tilde{X}^T \tilde{W}_{(k)} \tilde{z}_{(k)}$.
7: **end while**
8: **return** $\tilde{\beta}_{RandGLM}$, an approximation of β_{GLM}.

Quality of Approximation Results and Convergence Behavior



(a) Standardized IWLS Quality of Approximation

(b) Estimator Trace RandIWLS

Figure 1: Simulation Results

- Comparison of three different data-generating processes (uniformity of leverage scores)
- QoA statistic follows a bimodal distribution

$\Rightarrow$ "Bad" mode results from failure to converge

- Working Variate Influence Score performs the best in terms of variance
- Random fluctuations around the true converged IWLS Logit estimator
- Strongest variance increase for very nonuniform leverage datasets
- Again, Working Variate Influence Score seems to be the most robust measure.

Conclusions, Forthcoming Research and Acknowledgements

This project has theoretically and empirically analyzed potential extensions of the randomized numerical linear algebra framework to the application of GLMs. Furthermore, it has attempted to extend and unify this exciting field of computer science with a statistical learning perspective.

Forthcoming research questions include the following:

1. How do different convergence criteria (e.g. information criterion) influence the conv. behavior?
2. Is it possible to adapt a form of row amount scheduling, that allows us to increase the amount of rows sampled whenever the estimator varies to much from one iteration to the next?
3. How can we define and exploit pull from gravity/influence in a GLM setting?

I thankfully acknowledge the support of both of my Master thesis supervisors, Prof. Omiros Papaspiliopoulos (UPF) and Prof. Ioannis Kosmidis (UCL).

References

DRINEAS, P., M. MAGDON-ISMAIL, M. W. MAHONEY, AND D. P. WOODRUFF (2012): "Fast approximation of matrix coherence and statistical leverage," *Journal of Machine Learning Research*, 13, 3475–3506.

DRINEAS, P., M. W. MAHONEY, S. MUTHUKRISHNAN, AND T. SARLÓS (2011): "Faster least squares approximation," *Numerische Mathematik*, 117, 219–249.