TODO

Kenny Chiu

February 6, 2022

1 Summary

1.1 Context and background

Lacotte et al. [10] study the performance of iterative Hessian sketch (IHS) [14] for overdetermined least squares problems of the form

$$\mathbf{b}^* = \arg\min_{\mathbf{b}\in\mathbb{R}^d} \left\{ f(\mathbf{b}) = \frac{1}{2}\|\mathbf{X}\mathbf{b} - \mathbf{y}\|^2 \right\}$$

where $\mathbf{X} \in \mathbb{R}^{n\times d}$, $n \geq d$, is a given full rank data matrix and $\mathbf{y} \in \mathbb{R}^n$ is a vector of observations. IHS is an iterative method based on random projections that is effective for large data and ill-conditioned problems. Given step sizes $\{\alpha_t\}$ and momentum parameters $\{\beta_t\}$, the IHS solution is iteratively updated using

$$\mathbf{b}_{t+1} = \mathbf{b}_t - \alpha_t \mathbf{H}_t^{-1} \nabla f(\mathbf{b}_t) + \beta_t(\mathbf{b}_t - \mathbf{b}_{t-1})$$

where $\mathbf{H}_t = \mathbf{X}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{X}$ is an approximation of the Hessian $\mathbf{H} = \mathbf{X}^\top\mathbf{X}$ given refreshed (i.i.d.) $m \times n$ sketching (random) matrices $\{\mathbf{S}_t\}$ with $m \ll n$. The performance of IHS with Gaussian sketches (i.e., where $(\mathbf{S}_t)_{ij}$ are i.i.d. $N(0, m^{-1})$) has been studied, but IHS with other sketches have only been empirically studied. In their work, Lacotte et al. [10] draw on results from random matrix and free probability theory and show that the following sketches (asymptotically) converge faster to the optimal solution compared to Gaussian sketches:

1. Truncated Haar sketch, where the rows of $\mathbf{S}_t$ are orthonormal. Orthogonal sketches are preferred over i.i.d. sketches as they do not distort the projection, but orthogonality in general Haar matrices come at the expense of requiring the Gram-Schmidt procedure, which has cost $O(nm^2)$ larger than the $O(nmd)$ cost when using Gaussian sketches.

2. A version of the subsampled randomized Hadamard transform (SRHT), with $\mathbf{S}_t$ constructed from $\mathbf{R}_t = n^{-\frac{1}{2}}\mathbf{B}_t\mathbf{W}_n\mathbf{D}_t\mathbf{P}_t$ where $\mathbf{B}_t$ is a $n \times n$ diagonal matrix of i.i.d. Bernoulli$\left(\frac{m}{n}\right)$ samples, $\mathbf{D}_t$ is a $n \times n$ diagonal matrix of i.i.d. Rademacher samples, $\mathbf{P}_t$ is a $n \times n$ uniformly sampled row permutation matrix, and $\mathbf{W}_n$ is the $n \times n$ Walsh-Hadamard matrix defined recursively as

$$\mathbf{W}_n = \begin{bmatrix} \mathbf{W}_{\frac{n}{2}} & \mathbf{W}_{\frac{n}{2}} \\ \mathbf{W}_{\frac{n}{2}} & -\mathbf{W}_{\frac{n}{2}} \end{bmatrix}$$

   where $\mathbf{W}_1 = 1$. $\mathbf{S}_t$ is taken as $\mathbf{R}_t$ with the zeros rows removed (as selected by $\mathbf{B}_t$). Note that due to this subsampling, $\mathbf{S}_t$ is a $M \times n$ matrix with $\mathbb{E}[M] = m$. By construction, SRHT sketches are orthogonal. Sketching with SRHT only requires $O(nd \log M)$.

1.2 Main contributions

The main contributions of Lacotte et al. [10] include several theoretical results that describe the (asymptotically) optimal value of the parameters for IHS with Haar or SRHT sketches, the corresponding convergence rates of IHS with these parameters, and closed form expressions for the inverse moments of SRHT sketches. These results are obtained based on asymptotic results from random matrix theory, in which it is assumed that the matrix dimensions satisfy the aspect ratios $\frac{d}{n} \to \gamma \in (0,1)$ and $\frac{m}{n} \to \xi \in (\gamma, 1)$ as $n, d, m \to \infty$.

The main results are Theorems 3.1 and 4.1. Theorem 3.1 says that for IHS with Haar sketches, the optimal convergence rate ρ_H of the relative prediction error is

$$\rho_H = \left(\lim_{n\to\infty} \frac{\mathbb{E}\left[\|\mathbf{X}(\mathbf{b}_t - \mathbf{b}^*)\|^2\right]}{\|\mathbf{X}(\mathbf{b}_0 - \mathbf{b}^*)\|^2} \right)^{\frac{1}{t}} = \rho_G \cdot \frac{\xi(1-\xi)}{\gamma^2 + \xi - 2\xi\gamma}$$

where ρ_G is the optimal rate of IHS with Gaussian sketches. The aspect ratio scaling factor is less than 1, implying that $\rho_H < \rho_G$ and that IHS with Haar sketches converges faster than with Gaussian sketches. Theorem 4.1 states that the rate ρ_S for IHS with SRHT sketches is equal to ρ_H under an additional mild assumption on the initialization of the least squares problem (which was not needed for Haar sketches due to their properties known in random matrix theory). Theorem 3.1 also states that the optimal convergence rate for IHS with Haar sketches is obtained using momentum values $\beta_t = 0$ (i.e., momentum does not help) and step sizes $\alpha_t = \frac{\theta_{1,H}}{\theta_{2,H}}$ where $\theta_{k,H}$ is the k-th inverse moment of the Haar sketch defined as

$$\theta_{k,H} = \lim_{n\to\infty} \frac{1}{d} \mathbb{E}\left[\operatorname{trace}\left((\mathbf{U}^\top \mathbf{S}^\top \mathbf{S} \mathbf{U})^{-k}\right)\right]$$

for $m \times n$ Haar matrix $\mathbf{S}$ and $n \times d$ deterministic matrix $\mathbf{U}$ with orthonormal columns. Closed-form expressions for the first two inverse moments are provided in Lemma 3.2 and are given by

$$\theta_{1,H} = \frac{1-\gamma}{\xi-\gamma}\,, \qquad \theta_{2,H} = \frac{(1-\gamma)(\gamma^2+\xi-2\gamma\xi)}{(\xi-\gamma)^3}\,.$$

Theorem 4.1 and Lemma 4.3 together state that the limiting distribution of Haar and SRHT sketches are the same and therefore so is the optimal step size when there is no momentum. However, the optimality of $\beta_t = 0$ for IHS with SRHT sketches is only a conjecture based on numerical simulations.

Other contributions of Lacotte et al. [10] include a complexity analysis of IHS with SRHT sketches and an empirical study of the theoretical results. The complexity analysis concludes that the asymptotic performance of IHS with SRHT sketches is faster than that of the preconditioned conjugate gradient method (pCG) [19] by a factor of $\log(d)$. The empirical study verifies that the limiting results can apply in the finite case where the convergence of IHS with Haar and with SRHT sketches are similar and faster than that of Gaussian sketches on ill-conditioned synthetic and real datasets of moderate size ($n \geq 4000$, $d \geq 200$), and that the IHS with SRHT sketches refreshed every iteration has faster convergence than pCG on a similar synthetic dataset.

1.3 Limitations

Limitations of the work by Lacotte et al. [10] include the reliance on asymptotic theory, the empirical evaluation of results on mostly synthetic datasets, the comparison of sketches based on a single criterion, and the unclear generalizability of results to more complicated problems.

Lacotte et al. [10] obtain the convergence rates of IHS with different sketching matrices by drawing on results from asymptotic random matrix theory. While their simulations show that the theory does apply in moderately-sized datasets, the datasets that they examine are primarily synthetic and designed to satisfy assumptions even if ill-conditioned. However, there is also the counterargument that IHS would only be considered over standard solvers for large data problems, and so these limitations are relatively minor.

Another limitation of their work is that only a single criterion—namely the prediction error between the sketched solution and the optimal solution—is used to compare the performance of the sketching matrices. Other criteria have also been considered in the literature, such as those based on other losses or those based on out-of-sample prediction [5, 14]. While certain criteria are intrinsically related [6], they may still have differing properties and lead to differing results [5].

The main limitation of the work by Lacotte et al. [10] is the simple problem context that the results are derived for. While the theory shows that IHS is promising for large data, overdetermined least squares problems, standard solvers would still be preferred over IHS in large data problems if the appropriate computational resources were available. It is unclear whether the theory could generalize to more complicated problems, such as to undetermined least squares problems or optimization problems with other losses. It would be particularly useful to understand whether there are problems for which IHS would be preferred over conventional solvers in the general case.

1.4 Related literature

Works that analyze the impact of sketch type in sketching methods make up a small portion of the sketching literature. The work by Lacotte et al. [10] is said to be inspired by and therefore most similar to the work by Dobriban and Liu [5], which appears to be the first in the literature to leverage results from asymptotic random matrix theory. Analysis in the asymptotic regime appears to be key in being able to differentiate between the analytical performance of different sketching matrices, which was a challenge in previous works [2, 14, 17]. More recently, Lacotte and Pilanci [9] directly extended their analysis of sketches in IHS to fixed sketches in a related first-order method that has better guarantees.

Recent related works in the literature also include those that propose extensions of IHS, e.g., IHS with momentum and fixed sketches [12], distributed IHS [3], first-order IHS with adaptive step sizes [21], and Newton sketch [15] (IHS for general convex optimization problems) and its own variants [e.g., 4, 11]. Analyses of the performance in these works generally are intended as a point of comparison against existing methods, are done empirically or make use of conventional analysis techniques rather than asymptotic theory, and do not particularly examine the impact of specific sketching matrices.

2 Mini-proposals

2.1 Proposal 1: A sketched interior point algorithm for quantile regression

Whereas linear regression fits a linear model on the conditional mean, quantile regression [8] fits a linear model on a conditional quantile. Quantile regression offers several advantages over linear regression, such as being able to model different quantiles (as opposed to only a mean), being free from assumptions regarding the parametric form of the response and homoscedasticity, and being transformation equivariant in its response [18]. Given a data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, observations $\mathbf{y} \in \mathbb{R}^n$ and a quantile $\tau \in (0, 1)$ of interest, the estimated parameters of the linear model are the solution to the optimization problem

$$\min_{\mathbf{b} \in \mathbb{R}^d} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \mathbf{b}) \left(\tau - \mathbb{1}[y_i - \mathbf{x}_i^\top \mathbf{b} < 0]\right) \ .$$

The problem is non-differentiable as-is but can be optimized as a linear program. For large datasets, the conventional approach to solving the linear program is to use a constrained primal-dual interior point method [16]. The interior point method involves iterative updates that are obtained as the solution to a linear system derived from a Newton step. The computational bottleneck in each update comes from computing $\mathbf{X}^\top \mathbf{W}_t \mathbf{X}$ where $\mathbf{W}_t$ is a diagonal matrix that changes every iteration [1]. This computation results in each iteration having a cost of $O(nd^2)$.

In our proposed project, we consider the case where $d \ll n$ and propose a stochastic interior point algorithm that uses sketching matrices to reduce the computational cost of the iterative updates. Drawing on proven methods in the sketching literature [15], the idea is to incorporate a partial sketching step into the original algorithm where instead of computing $\mathbf{X}^\top \mathbf{W}_t \mathbf{X}$, we compute

$$\mathbf{X}^\top \mathbf{W}_t^{\frac{1}{2}} \mathbf{S}_t^\top \mathbf{S}_t \mathbf{W}_t^{\frac{1}{2}} \mathbf{X} \ .$$

The matrix $\mathbf{S}_t \in \mathbb{R}^{m \times n}$, $m \ll n$, is a random matrix regenerated every iteration that is introduced for reducing the dimension. For example, the subsampled randomized Hadamard transform allows the sketch $\mathbf{S}_t \mathbf{W}_t^{\frac{1}{2}} \mathbf{X}$ to be formed at a cost of $O(nd \log m)$ [10], and the matrix product above can then be computed at a cost of $O(md^2)$. While the sketched solution will only be an approximation to the original solution, recent work on the convergence of sketched solutions in other optimization problems show promising theoretical and empirical results [e.g., 4, 11, 15]. We also note that Yang et al. [20] had previously proposed a stochastic algorithm for quantile regression. However, their method differs greatly from ours in that they construct a random preconditioning matrix before using standard methods to solve the optimization problem on the conditioned data matrix.

The main contributions of this project would be as follows:

1. A sketched interior point algorithm for optimizing quantile regression problems that is expected to be faster than standard methods currently used in practice.

2. A theoretical analysis of the proposed sketched interior point algorithm that provides convergence guarantees.

3. An empirical comparison of quantile regression models fitted on large datasets obtained from the proposed sketched interior point algorithm and other existing methods, such as the standard interior point method [16] (implemented in R), the stochastic method by Yang et al. [20], a more modern iteration of the interior point method [22], and a modern quantile regression algorithm based on smoothing [7] (also implemented in R).

4. An implemention of the sketch interior point algorithm, e.g., in R, if found to have practical advantages over the existing algorithms.

The main challenge of this project would be the theoretical analysis of the sketched interior point algorithm. The most feasible analysis approach would likely be following that of Pilanci and Wainwright [15] for interior point methods and partial sketches, which would provide a worst-case result about the number of iterations needed to obtain a solution within a desired tolerance. The effect of the sketching matrix may also be of interest, but an analysis approach similar to the asymptotic approach of Lacotte et al. [10] would likely be necessary. However, adapting their approach for least squares to that of quantile regression is not straightforward and would likely be more suited for a follow-up project.

Following the completion of this project, there are multiple directions of future work that may be of interest:

1. The proposed sketched interior point algorithm would be useful for the $d \ll n$ case but not for the $n \ll d$ case. For the latter, a sketch-based method would likely still be possible but would need to use sketches differently, e.g., directly sketching the data matrix as Pham and El Ghaoui [13] did for LASSO or sketching both the data matrix and the observations as in classical least-squares sketch (although sketching both has been shown to be suboptimal [14]).

2. Exploration of applications that may benefit from a faster quantile regression or interior point algorithm, e.g., composite quantile regression [23] for high-dimensional regression or applications of quadratic progamming.

3. Investigation of sketched quantile regression algorithms based on smoothing. These algorithms approximate the original optimization problem by a differentiable one and therefore sketching should directly follow from the work of Pilanci and Wainwright [15]. Given the more standard setup, it is likely easier to adapt the approach of Lacotte et al. [10] to these problems than it is to the interior point algorithms for studying the effect of specific sketching matrices in sketched quantile regression.

2.2 Proposal 2: MY OTHER PROPOSAL TITLE

3 Project report

References

[1] Colin Chen and Ying Wei. Computational issues for quantile regression. *Sankhyā: The Indian Journal of Statistics*, pages 399–417, 2005.

[2] Krzysztof Choromanski, Mark Rowland, and Adrian Weller. The unreasonable effectiveness of structured random orthogonal embeddings. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 218–227, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[3] Michał Dereziński, Burak Bartan, Mert Pilanci, and Michael W Mahoney. Debiasing distributed second order optimization with surrogate sketching and scaled regularization. *arXiv preprint arXiv:2007.01327*, 2020.

[4] Michał Dereziński, Jonathan Lacotte, Mert Pilanci, and Michael W Mahoney. Newton-LESS: Sparsification without trade-offs for the sketched Newton update. *Advances in Neural Information Processing Systems*, 34, 2021.

[5] Edgar Dobriban and Sifan Liu. Asymptotics for sketching in least squares. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.

[6] Petros Drineas, Michael W Mahoney, Shan Muthukrishnan, and Tamás Sarlós. Faster least squares approximation. *Numerische mathematik*, 117(2):219–249, 2011.

[7] Xuming He, Xiaoou Pan, Kean Ming Tan, and Wen-Xin Zhou. Smoothed quantile regression with large-scale inference. *Journal of Econometrics*, 2021.

[8] Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

[9] Jonathan Lacotte and Mert Pilanci. Optimal randomized first-order methods for least-squares problems. In *International Conference on Machine Learning*, pages 5587–5597. PMLR, 2020.

[10] Jonathan Lacotte, Sifan Liu, Edgar Dobriban, and Mert Pilanci. Optimal iterative sketching methods with the subsampled randomized Hadamard transform. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9725–9735. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/6e69ebbfad976d4637bb4b39de261bf7-Paper.pdf.

[11] Jonathan Lacotte, Yifei Wang, and Mert Pilanci. Adaptive Newton sketch: Linear-time optimization with quadratic convergence and effective Hessian dimensionality. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR, 2021. URL http://proceedings.mlr.press/v139/lacotte21a.html.

[12] Ibrahim Kurban Ozaslan, Mert Pilanci, and Orhan Arikan. Iterative Hessian sketch with momentum. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7470–7474, 2019. doi: 10.1109/ICASSP.2019.8682720.

[13] Vu Pham and Laurent El Ghaoui. Robust sketching for multiple square-root LASSO problems. In *Artificial Intelligence and Statistics*, pages 753–761. PMLR, 2015.

[14] Mert Pilanci and Martin J Wainwright. Iterative Hessian sketch: Fast and accurate solution approximation for constrained least-squares. *The Journal of Machine Learning Research*, 17(1):1842–1879, 2016.

[15] Mert Pilanci and Martin J Wainwright. Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence. *SIAM Journal on Optimization*, 27(1): 205–245, 2017.

[16] Stephen Portnoy and Roger Koenker. The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statistical Science*, 12 (4):279–300, 1997.

[17] Garvesh Raskutti and Michael W. Mahoney. A statistical perspective on randomized sketching for ordinary least-squares. *Journal of Machine Learning Research*, 17(213): 1–31, 2016. URL http://jmlr.org/papers/v17/15-440.html.

[18] Robert N Rodriguez and Yonggang Yao. Five things you should know about quantile regression. In *Proceedings of the SAS global forum 2017 conference, Orlando*, pages 2–5, 2017.

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

[20] Jiyan Yang, Xiangrui Meng, and Michael Mahoney. Quantile regression for large-scale applications. In *International Conference on Machine Learning*, pages 881–887. PMLR, 2013.

[21] Aijun Zhang, Hengtao Zhang, and Guosheng Yin. Adaptive iterative Hessian sketch via A-optimal subsampling. *Statistics and Computing*, 30(4):1075–1090, jul 2020. ISSN 0960-3174. doi: 10.1007/s11222-020-09936-8. URL https://doi.org/10.1007/s11222-020-09936-8.

[22] Pan Zhao and Shenghua Yu. An improved interior point algorithm for quantile regression. *IEEE Access*, 8:139647–139657, 2020.

[23] Hui Zou and Ming Yuan. Composite quantile regression and the oracle model selection theory. *The Annals of Statistics*, 36(3):1108–1126, 2008.