TODO

Kenny Chiu

January 27, 2022

1 Summary

1.1 Context and background

Lacotte et al. [4] study the theoretical performance of iterative Hessian sketch (IHS) [5] 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 ill-conditioned problems. Given step sizes $\{\alpha_t\}$ and momentum parameters $\{\beta_t\}$, the IHS solution is iteratively updated by

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

where the matrix $\mathbf{H}_t = \mathbf{X}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{X}$ 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 theoretical 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 variants with other sketches have only been empirically studied. In their work, Lacotte et al. [4] draw on results from random matrix and free probability theory and show that the following sketches have faster (asymptotic) convergence rates compared to Gaussian sketches:

1. truncated Haar sketch, where the rows of $\mathbf{S}_t$ are orthonormal. The orthogonality helps to prevent distortions in random projections but 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 $\mathbf{R}_t = n^{-\frac{1}{2}}\mathbf{B}_t(\mathbf{W}_n)_t\mathbf{D}_t\mathbf{P}_t$ 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 $(\mathbf{W}_n)_t$ 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 because of this subsampling, $\mathbf{S}_t$ is a $M \times n$ matrix with $\mathbb{E}[M] = m$. Sketching with SRHT only requires $O(nd \log M)$.

1.2 Main contributions

The main contributions of Lacotte et al. [4] 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 expected squared error is proportional to the optimal rate ρ_G of IHS with Gaussian sketches, and is given by

$$\rho_H = \rho_G \cdot \frac{\xi(1-\xi)}{\gamma^2 + \xi - 2\xi\gamma} \,.$$

The aspect ratio scaling factor is less than 1 and therefore $\rho_H < \rho_G$, implying that IHS with Haar sketches converges at a faster rate than with Gaussian sketches. Theorem 4.1 states a similar conclusion for IHS with SRHT sketches where the optimal rate ρ_S is the same as that with Haar sketches, i.e., $\rho_S = \rho_H$, under an additional mild assumption on the initialization of the least squares problem which is necessary for drawing on existing results 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, i.e.,

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

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

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

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. [4] 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 standard pre-conditioned conjugate gradient method (pCG) [7] by a factor of $\log(d)$. The empirical study verifies that the limiting results can apply to 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, and that the IHS with SRHT sketches refreshed every iteration has faster convergence than pCG on a synthetic ill-conditioned dataset.

1.3 Related literature

Works that focus on analyzing the statistical performance of sketching methods make up a small portion of the sketching literature. The work by Lacotte et al. [4] is said to be inspired

by and therefore most similar to the work by Dobriban and Liu [2], 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 performance of different sketching matrices, which was an analytical challenge in previous works [1, 5, 6].

Other related works in the literature include those that introduce new variants of IHS or similar sketching methods TODO. Analyses of the performance in these works are generally intended as a point of comparison against existing methods, make use of conventional analysis techniques rather than asymptotic theory, and do not examine the impact of specific sketching matrices.

1.4 Limitations

Limitations of the work by Lacotte et al. [4] 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. [4] obtain the convergence rates of IHS with different sketching matrices by drawing on results from asymptotic random matrix theory. The consequence is that these results may not apply in the case of small datasets. While their simulations show that the theory does apply to 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 reconstruction 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 [2, 5]. While certain criteria are intrinsically related [3], they may still have differing properties and lead to differing results [2].

The main limitation of their work 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 particularly be of interest to understand whether there are problems where IHS would be preferred to conventional solvers in the general case.

2 Mini-proposals

2.1 Proposal 1: MY PROPOSAL TITLE

2.2 Proposal 2: MY OTHER PROPOSAL TITLE

3 Project report

References

[1] 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.

[2] 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.

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.