# TODO

Kenny Chiu

January 23, 2022

# 1 Summary

## 1.1 Context and background

Lacotte et al. [1] study the theoretical performance of iterative Hessian sketch (IHS) 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. [1] 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 with i.i.d. Bernoulli$\left(\frac{m}{n}\right)$ samples on the diagonal, $\mathbf{D}_t$ is a $n \times n$ diagonal matrix with uniformly sampled ±1 on the diagonal, $\mathbf{P}_t$ 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. [1] include several theoretical results that prescribe 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 expression for the second inverse moment 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 rate $\rho_H$ at which the relative expected squared error decreases in each iteration is constant and proportional to the optimal rate $\rho_G$ 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 constant 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 $\rho_S$ 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 known random matrix results. TODOoptimal parameters

## 1.3 Related literature

## 1.4 Limitations

# 2 Mini-proposals

## 2.1 Proposal 1: MY PROPOSAL TITLE

## 2.2 Proposal 2: MY OTHER PROPOSAL TITLE

# 3 Project report

# References

[1] 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.