Contents

1 Optimal Iterative Sketching with the Subsampled Randomized Hadamard Transform

Based on [1].

The performance of iterative Hessian sketch (IHS) has only been studied empirically in existing literature. Lacotte et al. [1] show that for IHS with random matrices projected via refreshed (i.i.d.) truncated Haar matrices or subsampled randomized Hadamard transform (SRHT), the limiting rate of convergence is expected to be better than that of IHS with Gaussian random projections. Their other theoretical contributions include a closed form optimal (limiting) step size for IHS with Haar sketches, showing that momentum does not improve performance of IHS with refreshed Haar sketches, and an explicit formula for the second inverse moment of Haar sketches.

1.1 Background

1.1.1 Problem and method

Consider 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 $\mathbf{X} \in \mathbb{R}^{n\times d}$ is a given data matrix with $n \geq d$ and $\operatorname{rank}(\mathbf{X}) = d$ and $\mathbf{y} \in \mathbb{R}^n$ is a vector of observations. Iterative Hessian sketch is one iterative method for solving the problem where given step sizes $\{\alpha_t\}$ and momentum $\{\beta_t\}$, the solutions are iteratively updated by

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

The matrix H_t is an approximation of the Hessian $H = \mathbf{X}^\top\mathbf{X}$ and is given by $H_t = \mathbf{X}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{X}$ where $\mathbf{S}_0, \ldots, \mathbf{S}_t, \ldots$ are refreshed (i.i.d.) $m \times n$ sketching (random) matrices with $m \ll n$. The types of sketches discussed by Lacotte et al. [1] include

1. Gaussian sketches where $(\mathbf{S}_t)_{ij} \overset{\text{i.i.d.}}{\sim} N(0, m^{-1})$. Computing the matrix product $\mathbf{SX}$ is $O(mnd)$ in general, which is larger than the cost of $O(nd^2)$ for direct method solvers when $m \geq d$.

2. truncated Haar sketches using Haar matrices $\mathbf{S}_t$ where the rows are orthonormal (TODO: other qualifications? Truncated?). Generating the matrix requires $O(nm^2)$ using a Gram-Schmidt procedure which is larger than $O(nd^2)$.

3. subsampled randomized Hadamard transform where the sketch $\mathbf{SX}$ can be obtained in $O(nd\log m)$ time. Like other orthogonal embeddings, the performance tends to be better than random projections with i.i.d. entries.

1.1.2 Random matrix theory and tools

Let $\{\mathbf{M}_n\}_n$ be a sequence of $n \times n$ Hermitian random matrices. The empirical spectral distribution (e.s.d.) of $\mathbf{M}_n$ is the CDF of its eigenvalues $\lambda_1, \ldots, \lambda_n$ given by $F_{\mathbf{M}_n}(x) = \frac{1}{n}\sum_{j=1}^n \mathbb{1}[\lambda_j \leq x]$ for $x \in \mathbb{R}$. The eigenvalues are random and so $F_{\mathbf{M}_n}$ is also random. The e.s.d.$F_{\mathbf{M}_n}$ converges weakly to the limiting spectral distribution (l.s.d.) of $\mathbf{M}_n$ as $n \to \infty$.

For a probability measure μ with support on $[0, \infty)$, its Stieltjes transform is defined over the complex space complementary to the support of μ as

$$m_\mu(z) = \int \frac{1}{x - z}\mu(dx) \ .$$

The S-transform of μ is unique under certain conditions and is defined as the solution to the equation

$$m_\mu\left(\frac{z+1}{zS_\mu(z)}\right) + zS_\mu(z) = 0\ .$$

The Marchenko-Pastur theorem says that for a $m \times d$ matrix $\mathbf{S}$ where $(\mathbf{S})_{ij} \overset{\text{i.i.d.}}{\sim} N(0, m^{-1})$, then as $m, d \to \infty$ with $\frac{m}{d} \to \rho \in (0,1)$, $\mathbf{S}^\top\mathbf{S}$ has l.s.d.F_ρ with a Stieltjes transform that is the unique solution of a certain fixed point equation and with a density given by

$$\mu_\rho(x) = \frac{\sqrt{((1+\sqrt{\rho})^2 - x)_+(x-(1-\sqrt{\rho})^2)_+}}{2\pi\rho x}$$

where $y_+ = \max\{0, y\}$.

1.1.3 Other notation

Define the aspect ratios $\gamma = \lim_{n,d\to\infty} \frac{d}{n} \in (0,1)$, $\xi = \lim_{n,m\to\infty} \frac{m}{n} \in (\gamma, 1)$ and $\rho_g = \frac{\gamma}{\xi} \in (0,1)$ where subscript g refers to Gaussians and h refers to Haar or Hadamard. For a sequence $\{\mathbf{b}_t\}$, denote the error vector $\Delta_t = \mathbf{U}^\top\mathbf{X}(\mathbf{b}_t - \mathbf{b}^*)$ where $\mathbf{U}$ is the $n \times d$ matrix of left singular vectors of $\mathbf{X}$. Note that $\|\Delta_t\|^2 = \|\mathbf{X}(\mathbf{b}_t - \mathbf{b}^*)\|^2$.

1.2 Sketching with Haar matrices

Theorem 3.1 (Optimal IHS with Haar sketches): for refreshed Haar matrices $\{\mathbf{S}_t\}$, step sizes $\alpha_t = \frac{\theta_{1,h}}{\theta_{2,h}}$ (defined in Lemma 3.2) and momentum parameters $\beta_t = 0$, the sequence of error vectors $\{\Delta_t\}$ satisfies

$$\rho_h = \left(\lim_{n\to\infty} \frac{\mathbb{E}\|\Delta_t\|^2}{\|\Delta_0\|^2}\right)^{\frac{1}{t}} = \rho_g \cdot \frac{\xi(1-\xi)}{\gamma^2 + \xi - 2\xi\gamma}\ .$$

For any step sizes $\{a_t\}$ and momentum parameters $\{\beta_t\}$,

$$\rho_h \leq \liminf_{t\to\infty} \left(\lim_{n\to\infty} \frac{\mathbb{E}\|\Delta_t\|^2}{\|\Delta_0\|^2}\right)^{\frac{1}{t}}\ ,$$

i.e., ρ_h is the optimal rate for Haar embeddings.

Theorem 3.1 says that using the optimal parameters (which has closed forms), the rate at any time step $t \geq 1$ is given by

$$\rho_h^t = \lim_{n\to\infty} \frac{\mathbb{E}\|\Delta_t\|^2}{\|\Delta_0\|^2}$$

with $\rho_h < \rho_g$. Momentum also does not provide benefits.

Lemma 3.2 (First two inverse moments of Haar sketches): let $\mathbf{S}$ be a $m \times n$ Haar matrix, $\mathbf{U}$ a $n \times d$ deterministic matrix with orthonormal columns. Then

$$\theta_{1,h} = \lim_{n\to\infty} \frac{1}{d}\text{trace}\left(\mathbb{E}\left[(\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}\mathbf{U})^{-1}\right]\right) = \frac{1-\gamma}{\xi-\gamma}$$
$$\theta_{2,h} = \lim_{n\to\infty} \frac{1}{d}\text{trace}\left(\mathbb{E}\left[(\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}\mathbf{U})^{-2}\right]\right) = \frac{(1-\gamma)(\gamma^2+\xi-2\gamma\xi)}{(\xi-\gamma)^3}$$

(Note that $\theta_{i,h}$ is the average of the eigenvalues of $\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}\mathbf{U}$ to the power of $-i$.)

Lacotte et al. [1] show that as the sketch size m increases relative to n, the convergence ratio of Haar sketches versus Gaussian projections scales as $\frac{\rho_h}{\rho_g} \approx (1-\xi)$.

1.3 Sketching with SRHT matrices

Lacotte et al. [1] consider a version of SRHT where the transform $\mathbf{X} \mapsto \mathbf{SX}$ first randomly permutes the rows of $\mathbf{X}$ before applying the classical transform, i.e., $\mathbf{S} = \frac{1}{\sqrt{n}}\mathbf{B}\mathbf{H}_n\mathbf{D}\mathbf{P}$ where $\mathbf{B}$ is a $n \times n$ diagonal matrix of i.i.d. Bernoulli random variables with success probability $\frac{m}{n}$, $\mathbf{D}$ is a $n \times n$ diagonal matrix of i.i.d. sign random variables with uniform probability, and $\mathbf{P}$ is a $n \times n$ uniformly distributed permutation matrix. $\mathbf{H}_n$ is the $n \times n$ Walsh-Hadamard matrix where for $n = 2^p$ for $p \geq 1$, the matrix is defined recursively as

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

with $\mathbf{H}_1 = 1$. Before applying the transformation to $\mathbf{X}$, the zero rows of $\mathbf{S}$ are discarded and so $\mathbf{S}$ is a $M \times n$ orthogonal matrix with $M \sim \text{Binomial}(\frac{m}{n}, n)$, and $\frac{M}{n} \to \xi$ as $n \to \infty$. Note that $\mathbf{S}$ is still referred to as a $m \times n$ SRHT matrix.

Theorem 4.1 (IHS with SRHT sketches). Suppose that $\mathbf{b}_0$ is random and that the error vector Δ_0 satisfies $\mathbb{E}\left[\Delta_0\Delta_0^\top\right] = d^{-1}\mathbf{I}_d$. Then for refreshed SRHT matrices $\{\mathbf{S}_t\}$, step sizes $\alpha_t = \frac{\theta_{1,h}}{\theta_{2,h}}$ and momentum parameters $\beta_t = 0$, the sequence $\{\Delta_t\}$ satisfies

$$\rho_s = \left(\lim_{n\to\infty} \frac{\mathbb{E}\left[\|\Delta_t\|^2\right]}{\mathbb{E}\left[\|\Delta_0\|^2\right]}\right)^{\frac{1}{t}} = \rho_g \cdot \frac{\xi(1-\xi)}{\gamma^2 + \xi - 2\xi\gamma} = \rho_h \; .$$

TODO

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.