Contents

1 Optimal Iterative Sketching with the Subsampled Randomized Hadamard Transform

Based on [3].

The performance of iterative Hessian sketch (IHS) has only been studied empirically in existing literature. Lacotte et al. [3] 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

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

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. [3] 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. 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 of z outside 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 η-transform is an alternative form of the Stieltjes transform defined for $z \in \mathbb{C}\backslash\mathbb{R}^-$ and given by

$$\eta_\mu(z) = \int \frac{1}{1+zx}\mu(dx) = \frac{1}{z}m_\mu\left(-\frac{1}{z}\right) .$$

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

For a random matrix $\mathbf{X}_n$ in algebra $\mathcal{A}_n$ of $n \times n$ matrices, the normalized trace operator

$$\tau_n(\mathbf{X}_n) = \frac{1}{n}\mathbb{E}\left[\text{trace}(\mathbf{X}_n)\right]$$

is a linear functional (and is analogous to expectations with scalar random variables; see Terry Tao's notes on free probability).

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 $\{\beta_t\}$,

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

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 $t \ge 1$ 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. [3] 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. [3] consider a version of SRHT where the transform $\mathbf{X} \mapsto \mathbf{S}\mathbf{X}$ 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

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

The additional initialization condition can be satisfied by picking $\mathbf{b}_0$ uniformly from the unit d-sphere $\mathbf{S}^{d-1}$ and applying a random signed permutation and scaling to the columns of $\mathbf{X}$ (TODO). The case of general Δ_0 or momentum $\beta_t \neq 0$ has not yet been explored.

Theorem 4.2 (Upper bound of SRHT error): for any $\mathbf{b}_0$ with 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 of error vectors $\{\Delta_t\}$ satisfies

$$\limsup_{n\to\infty} \left(\frac{\mathbb{E}\left[\|\Delta_t\|\right]^2}{d\mathbb{E}\left[\|\Delta_0\|\right]^2}\right)^{\frac{1}{t}} \leq \rho_h \ .$$

(Weaker by a factor of d, but is negligible for large t.)

Lemma 4.3 (First two inverse moments of SRHT sketches): let $\mathbf{S}$ be a $m \times n$ SRHT matrix, $\mathbf{S}_h$ a $m \times n$ Haar matrix and $\mathbf{U}$ a $n \times d$ deterministic matrix with orthonormal columns. Then $\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}\mathbf{U}$ and $\mathbf{U}^\top\mathbf{S}_h^T\mathbf{S}_h\mathbf{U}$ have the same limiting spectral distribution and therefore the same two inverse moments.

1.4 Complexity analysis

Lacotte et al. [3] compare the (asymptotic) complexity of IHS with SRHT embeddings to that of the standard pre-conditioned conjugate gradient method. The latter uses a sketch $\mathbf{S}\mathbf{X}$ to compute a pre-conditioning matrix $\mathbf{P}$ such that $\mathbf{X}\mathbf{P}^{-1}$ has a small condition number and solves the least squares problem $\min_{\mathbf{b}} \|\mathbf{X}\mathbf{P}^{-1}\mathbf{b} - \mathbf{y}\|^2$. The sketch size is prescribed to be $m \approx d\log d$, and the complexity to achieve $\|\Delta_t\|^2 \leq \epsilon$ scales as $C_c \asymp nd\log d + d^3\log d + nd\log\varepsilon^{-1}$ (cost of forming $\mathbf{S}\mathbf{X}$, factoring, and per-iteration cost times number of iterations, respectively). For IHS with SRHT, we can take $m \approx d$ which results in complexity $C_n \asymp (nd\log d + d^3 + nd)\log\varepsilon^{-1}$. Treating $\log\varepsilon^{-1}$ as constant independent of the dimensions, $\frac{C_n}{C_c} \asymp \frac{1}{\log d}$ as $n, m, d \to \infty$.

1.5 Numerical simulations

The main results of the numerical simulations by Lacotte et al. [3] (involving ill-conditioned matrices) include:

- IHS with refreshed Haar/SRHT embeddings (using optimal step size from Theorem 4.1 and finite sample approximations of ξ and γ) converge faster than IHS with refreshed Gaussian embeddings (using parameters α_t and β_t derived from previous work).
- Haar embeddings and SRHT embeddings perform similarly (though SRHT has a computational advantage).
- IHS with refreshed SRHT embeddings every iteration converge faster than the pre-conditioned conjugate gradient method. IHS with sketches refreshed at lower frequencies converge slower.

1.6 Appendix

1.6.1 B.2

- Distribution F_γ with density $\gamma\delta_1 + (1-\gamma)\delta_0$
- Distribution F_ξ with density $\xi\delta_1 + (1-\xi)\delta_0$
- System of equations

$$\begin{aligned}
\eta_C(z) &= \int \frac{1}{z\gamma(z)x+1} F_\xi(dx) \\
&= \frac{\xi}{z\gamma(z)+1} + 1 - \xi \\
\gamma(z) &= \int \frac{x}{\eta_C(z) + z\delta(z)x} F_\gamma(dx) \\
&= \frac{\gamma}{\eta_C(z) + z\delta(z)} \\
\delta(z) &= \int \frac{x}{z\gamma(z)x+1} F_\xi(dx) \\
&= \frac{\xi}{z\gamma(z)+1}
\end{aligned}$$

TODO

$$\eta_C(z) = (1-\gamma) + \frac{\gamma}{(1 + z\left(1 + \frac{\xi-1}{\eta_C(z)}\right)}$$

1.7 Code

GitHub

2 Ridge regression

- Sketched Ridge Regression: Optimization Perspective, Statistical Perspective, and Model Averaging
- Input Sparsity Time Low-Rank Approximation via Ridge Leverage Score Sampling
- An Iterative, Sketching-based Framework for Ridge Regression: focuses on underdetermined case $d \gg n$

Ridge regression–does not need assumption that $\mathbf{X}$ is full rank?:

$$\begin{aligned}&\arg\min_{\mathbf{b}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{X}\mathbf{b}-\mathbf{y}\|^2 + \frac{\lambda}{2}\|\mathbf{b}\|^2 \\ &= \arg\min_{\mathbf{b}\in\mathbb{R}^d} \frac{1}{2}\left(\mathbf{b}^\top\mathbf{X}^\top\mathbf{X}\mathbf{b} - 2\mathbf{b}^\top\mathbf{X}^\top\mathbf{y} + \mathbf{y}^T\mathbf{y}\right) + \frac{\lambda}{2}\mathbf{b}^T\mathbf{b}\end{aligned}$$

Gradient:

$$\nabla f(\mathbf{b}_t) = \mathbf{X}^\top\mathbf{X}\mathbf{b}_t - \mathbf{X}^\top\mathbf{y} + \lambda\mathbf{b}_t = \left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d\right)\mathbf{b}_t - \mathbf{X}^\top\mathbf{y}$$

Note: the following are equivalent solutions (useful for overdetermined, underdetermined case?)

$$\begin{aligned}\mathbf{b}^* &= \left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d\right)^{-1}\mathbf{X}^\top\mathbf{y} \\ \mathbf{b}^* &= \mathbf{X}^\top\left(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_n\right)^{-1}\mathbf{y}\end{aligned}$$

Hessian and partial Newton sketch (also considered by Chowdhury et al. [1]):

$$\begin{aligned}\mathbf{H} &= \mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d \\ \mathbf{H}_t &= \mathbf{X}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{X} + \lambda\mathbf{I}_d\end{aligned}$$

IHS with no momentum and partial sketch:

$$\begin{aligned}\mathbf{b}_{t+1} &= \mathbf{b}_t - \alpha_t\mathbf{H}_t^{-1}\nabla f(\mathbf{b}_t) \\ \mathbf{b}_{t+1} &= \mathbf{b}_t - \alpha_t\left(\mathbf{X}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{X} + \lambda\mathbf{I}_d\right)^{-1}\left(\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d\right)\mathbf{b}_t - \mathbf{X}^\top\mathbf{y}\right) \\ &= \mathbf{b}_t - \alpha_t\left(\mathbf{X}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{X} + \lambda\mathbf{I}_d\right)^{-1}\left(\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d\right)\mathbf{b}_t - \left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d\right)\mathbf{b}^*\right) \\ &= \mathbf{b}_t - \alpha_t\left(\mathbf{X}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{X} + \lambda\mathbf{I}_d\right)^{-1}\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d\right)\left(\mathbf{b}_t - \mathbf{b}^*\right) \\ \mathbf{X}\mathbf{b}_{t+1} - \mathbf{X}\mathbf{b}^* &= \mathbf{X}\mathbf{b}_t - \alpha_t\mathbf{X}\left(\mathbf{X}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{X} + \lambda\mathbf{I}_d\right)^{-1}\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d\right)\left(\mathbf{b}_t - \mathbf{b}^*\right) - \mathbf{X}\mathbf{b}^* \\ &= \mathbf{X}\mathbf{b}_t - \alpha_t\left(\mathbf{Q}_t\mathbf{b}_t - \mathbf{Q}_t\mathbf{b}^*\right) - \mathbf{X}\mathbf{b}^*\end{aligned}$$

Let $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^\top$ where $\mathbf{U}$, $\mathbf{V}$ orthogonal and Σ diagonal. TODO: take compact SVD? might break if multiplying by $\mathbf{X}$

$$\begin{aligned}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d &= \mathbf{V}\Sigma^\top\Sigma\mathbf{V}^\top + \lambda\mathbf{V}\mathbf{V}^\top \\ &= \mathbf{V}\left(\Sigma^\top\Sigma + \lambda\mathbf{I}_d\right)\mathbf{V}^\top \\ \left(\mathbf{X}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{X} + \lambda\mathbf{I}_d\right)^{-1} &= \left(\mathbf{V}\Sigma^\top\mathbf{U}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U}\Sigma\mathbf{V}^\top + \lambda\mathbf{V}\mathbf{V}^\top\right)^{-1} \\ &= \mathbf{V}\left(\Sigma^\top\mathbf{U}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U}\Sigma + \lambda\mathbf{I}_d\right)^{-1}\mathbf{V}^\top\end{aligned}$$

Then

$$\begin{aligned}\mathbf{b}_{t+1} &= \mathbf{b}_t - \alpha_t\left(\mathbf{X}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{X} + \lambda\mathbf{I}_d\right)^{-1}\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d\right)\left(\mathbf{b}_t - \mathbf{b}^*\right) \\ &= \mathbf{b}_t - \alpha_t\mathbf{V}\left(\Sigma^\top\mathbf{U}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U}\Sigma + \lambda\mathbf{I}_d\right)^{-1}\left(\Sigma^\top\Sigma + \lambda\mathbf{I}_d\right)\mathbf{V}^\top\left(\mathbf{b}_t - \mathbf{b}^*\right) \\ \mathbf{V}^\top\left(\mathbf{b}_{t+1} - \mathbf{b}^*\right) &= \mathbf{V}^\top\mathbf{b}_t - \alpha_t\left(\Sigma^\top\mathbf{U}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U}\Sigma + \lambda\mathbf{I}_d\right)^{-1}\left(\Sigma^\top\Sigma + \lambda\mathbf{I}_d\right)\mathbf{V}^\top\left(\mathbf{b}_t - \mathbf{b}^*\right) - \mathbf{V}^\top\mathbf{b}^* \\ &= \left(\mathbf{I}_d - \alpha_t\left(\Sigma^\top\mathbf{U}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U}\Sigma + \lambda\mathbf{I}_d\right)^{-1}\left(\Sigma^\top\Sigma + \lambda\mathbf{I}_d\right)\right)\mathbf{V}^\top\left(\mathbf{b}_t - \mathbf{b}^*\right) \\ \|\mathbf{b}_{t+1} - \mathbf{b}^*\|^2 &= \left(\mathbf{b}_t - \mathbf{b}^*\right)^\top\mathbf{V}\left(\mathbf{I}_d - \alpha_t\left(\Sigma^\top\mathbf{U}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U}\Sigma + \lambda\mathbf{I}_d\right)^{-1}\left(\Sigma^\top\Sigma + \lambda\mathbf{I}_d\right)\right)^2\mathbf{V}^\top\left(\mathbf{b}_t - \mathbf{b}^*\right)\end{aligned}$$

TODO: take $\Delta_t = \mathbf{V}^\top(\mathbf{b}_t - \mathbf{b}^*)$? Interpretation in terms of solution error?

$$\begin{aligned}\|\Delta_{t+1}\|^2 &= \|\mathbf{V}^\top(\mathbf{b}_{t+1} - \mathbf{b}^*)\|^2 \\ &= \|\mathbf{b}_{t+1} - \mathbf{b}^*\|^2\end{aligned}$$

$$\begin{aligned}\|\Delta_{t+1}\|^2 &= \Delta_t^\top \left(\mathbf{I}_d - \alpha_t \left(\Sigma^\top \mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U}\Sigma + \lambda \mathbf{I}_d\right)^{-1} \left(\Sigma^\top\Sigma + \lambda \mathbf{I}_d\right)\right)^2 \Delta_t \\ \mathbb{E}\left[\|\Delta_{t+1}\|^2\right] &= \Delta_t^\top \mathbb{E}\left[\left(\mathbf{I}_d - \alpha_t \left(\Sigma^\top \mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U}\Sigma + \lambda \mathbf{I}_d\right)^{-1} \left(\Sigma^\top\Sigma + \lambda \mathbf{I}_d\right)\right)^2\right] \Delta_t \\ &= \Delta_t^\top \left(\mathbf{I}_d - \alpha_t \mathbb{E}\left[\left(\Sigma^\top \mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U}\Sigma + \lambda \mathbf{I}_d\right)^{-1}\right] \left(\Sigma^\top\Sigma + \lambda \mathbf{I}_d\right)\right. \\ &\quad - \alpha_t \left(\Sigma^\top\Sigma + \lambda \mathbf{I}_d\right) \mathbb{E}\left[\left(\Sigma^\top \mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U}\Sigma + \lambda \mathbf{I}_d\right)^{-1}\right] \\ &\quad \left. + \alpha_t^2 \left(\Sigma^\top\Sigma + \lambda \mathbf{I}_d\right) \mathbb{E}\left[\left(\Sigma^\top \mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U}\Sigma + \lambda \mathbf{I}_d\right)^{-2}\right] \left(\Sigma^\top\Sigma + \lambda \mathbf{I}_d\right)\right) \Delta_t\end{aligned}$$

$\Sigma^\top \mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U}\Sigma$ is positive semidefinite and $\lambda \mathbf{I}_d$ is positive definite. Then $\Sigma^\top \mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U}\Sigma + \lambda \mathbf{I}_d$ is positive definite. Consider the eigendecomposition $\Sigma^\top \mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U}\Sigma + \lambda \mathbf{I}_d = \mathbf{Q}\Lambda\mathbf{Q}^\top$ where Λ is diagonal with positive entries $\lambda_1, \ldots, \lambda_d$ and $\mathbf{Q} \in \mathbb{R}^{d \times d}$ is orthogonal. TODO: can't make rotational invariance argument? Can method for Theorem 4.1 be used here?

Assuming $\mathbf{X}$ full rank, if taking $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^\top$ as thin SVD:

$$\begin{aligned}\mathbf{b}_{t+1} &= \mathbf{b}_t - \alpha_t \mathbf{V} \left(\Sigma \mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U}\Sigma + \lambda \mathbf{I}_d\right)^{-1} \left(\Sigma^2 + \lambda \mathbf{I}_d\right) \mathbf{V}^\top (\mathbf{b}_t - \mathbf{b}^*) \\ &= \mathbf{b}_t - \alpha_t \mathbf{V}\Sigma^{-1} \left(\mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U} + \lambda \Sigma^{-2}\right)^{-1} \left(\Sigma + \lambda \Sigma^{-1}\right) \mathbf{V}^\top (\mathbf{b}_t - \mathbf{b}^*) \\ &= \mathbf{b}_t - \alpha_t \mathbf{V}\Sigma^{-1} \left(\mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U} + \lambda \Sigma^{-2}\right)^{-1} \left(\mathbf{I}_d + \lambda \Sigma^{-2}\right) \Sigma \mathbf{V}^\top (\mathbf{b}_t - \mathbf{b}^*)\end{aligned}$$

If $\Delta_t = \Sigma\mathbf{V}^\top(\mathbf{b}_t - \mathbf{b}^*)$:

$$\begin{aligned}\Sigma\mathbf{V}^\top(\mathbf{b}_{t+1} - \mathbf{b}^*) &= \Sigma\mathbf{V}^\top \mathbf{b}_t - \alpha_t \left(\mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U} + \lambda \Sigma^{-2}\right)^{-1} \left(\mathbf{I}_d + \lambda \Sigma^{-2}\right) \Sigma \mathbf{V}^\top (\mathbf{b}_t - \mathbf{b}^*) - \Sigma\mathbf{V}^\top \mathbf{b}^* \\ &= \left(\mathbf{I}_d - \alpha_t \left(\mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U} + \lambda \Sigma^{-2}\right)^{-1} \left(\mathbf{I}_d + \lambda \Sigma^{-2}\right)\right) \Sigma \mathbf{V}^\top (\mathbf{b}_t - \mathbf{b}^*) \\ \|\Delta_{t+1}\|^2 &= \Delta_t^\top \left(\mathbf{I}_d - \alpha_t \left(\mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U} + \lambda \Sigma^{-2}\right)^{-1} \left(\mathbf{I}_d + \lambda \Sigma^{-2}\right)\right)^2 \Delta_t \\ \mathbb{E}\left[\|\Delta_{t+1}\|^2\right] &= \Delta_t^\top \mathbb{E}\left[\left(\mathbf{I}_d - \alpha_t \left(\mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U} + \lambda \Sigma^{-2}\right)^{-1} \left(\mathbf{I}_d + \lambda \Sigma^{-2}\right)\right)^2\right] \Delta_t\end{aligned}$$

If $\Delta_t = \mathbf{U}^\top \mathbf{X}\left(\mathbf{b}_t - \mathbf{b}^*\right)$:

$$\begin{aligned}
\mathbf{X}(\mathbf{b}_{t+1} - \mathbf{b}^*) &= \mathbf{X}\mathbf{b}_t - \alpha_t \mathbf{U}\left(\mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U} + \lambda\Sigma^{-2}\right)^{-1}\left(\mathbf{I}_d + \lambda\Sigma^{-2}\right)\Sigma\mathbf{V}^\top\left(\mathbf{b}_t - \mathbf{b}^*\right) - \mathbf{X}\mathbf{b}^* \\
&= \left(\mathbf{I}_d - \alpha_t \mathbf{U}\left(\mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U} + \lambda\Sigma^{-2}\right)^{-1}\left(\mathbf{I}_d + \lambda\Sigma^{-2}\right)\mathbf{U}^T\right)\mathbf{X}\left(\mathbf{b}_t - \mathbf{b}^*\right) \\
\mathbf{U}^\top\mathbf{X}(\mathbf{b}_{t+1} - \mathbf{b}^*) &= \left(\mathbf{U}^\top - \alpha_t \left(\mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U} + \lambda\Sigma^{-2}\right)^{-1}\left(\mathbf{I}_d + \lambda\Sigma^{-2}\right)\mathbf{U}^T\right)\mathbf{X}\left(\mathbf{b}_t - \mathbf{b}^*\right) \\
&= \left(\mathbf{I}_d - \alpha_t \left(\mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U} + \lambda\Sigma^{-2}\right)^{-1}\left(\mathbf{I}_d + \lambda\Sigma^{-2}\right)\right)\mathbf{U}^\top\mathbf{X}\left(\mathbf{b}_t - \mathbf{b}^*\right) \\
\Delta_{t+1} &= \left(\mathbf{I}_d - \alpha_t \left(\mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U} + \lambda\Sigma^{-2}\right)^{-1}\left(\mathbf{I}_d + \lambda\Sigma^{-2}\right)\right)\Delta_t \\
\|\Delta_{t+1}\|^2 &= \Delta_t^\top\left(\mathbf{I}_d - \alpha_t \left(\mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U} + \lambda\Sigma^{-2}\right)^{-1}\left(\mathbf{I}_d + \lambda\Sigma^{-2}\right)\right)^2\Delta_t \\
\mathbb{E}\left[\|\Delta_{t+1}\|^2\right] &= \Delta_t^\top\mathbb{E}\left[\left(\mathbf{I}_d - \alpha_t \left(\mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U} + \lambda\Sigma^{-2}\right)^{-1}\left(\mathbf{I}_d + \lambda\Sigma^{-2}\right)\right)^2\right]\Delta_t \\
&= \Delta_t^\top\Big(\mathbf{I}_d - \alpha_t\mathbb{E}\left[\left(\mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U} + \lambda\Sigma^{-2}\right)^{-1}\right]\left(\mathbf{I}_d + \lambda\Sigma^{-2}\right) \\
&\quad - \alpha_t\left(\mathbf{I}_d + \lambda\Sigma^{-2}\right)\mathbb{E}\left[\left(\mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U} + \lambda\Sigma^{-2}\right)^{-1}\right] \\
&\quad + \alpha_t^2\left(\mathbf{I}_d + \lambda\Sigma^{-2}\right)\mathbb{E}\left[\left(\mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U} + \lambda\Sigma^{-2}\right)^{-2}\right]\left(\mathbf{I}_d + \lambda\Sigma^{-2}\right)\Big)\Delta_t
\end{aligned}$$

TODOAssuming $\mathbf{Q}_t = \mathbf{I}_d - \alpha_t\left(\mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U} + \lambda\Sigma^{-2}\right)^{-1}\left(\mathbf{I}_d + \lambda\Sigma^{-2}\right)$ satisfies conditions (limiting spectral distribution as $n \to \infty$ and others?), under the additional assumption that Δ_0 is random and $\mathbb{E}\left[\Delta_0\Delta_0^\top\right] = \frac{\mathbf{I}_d}{d}$, we have $\Delta_{t+1} = Q_t\Delta_t$ and so

$$\begin{aligned}
\mathbb{E}\left[\|\Delta_{t+1}\|^2\right] &= \operatorname{trace}\left(\mathbb{E}\left[\Delta_0^\top\mathbf{Q}_0\ldots\mathbf{Q}_{t-1}\mathbf{Q}_{t-1}\ldots\mathbf{Q}_0\Delta_0\right]\right) \\
&= \operatorname{trace}\left(\mathbb{E}\left[\mathbf{Q}_0\ldots\mathbf{Q}_{t-1}\mathbf{Q}_{t-1}\ldots\mathbf{Q}_0\Delta_0\Delta_0^\top\right]\right) \\
&= \operatorname{trace}\left(\mathbb{E}\left[\mathbf{Q}_0\ldots\mathbf{Q}_{t-1}\mathbf{Q}_{t-1}\ldots\mathbf{Q}_0\right]\mathbb{E}\left[\Delta_0\Delta_0^\top\right]\right) \\
&= \frac{1}{d}\operatorname{trace}\left(\mathbb{E}\left[\mathbf{Q}_0\ldots\mathbf{Q}_{t-1}\mathbf{Q}_{t-1}\ldots\mathbf{Q}_0\right]\right)
\end{aligned}$$

using the independence of Δ_0 and $\mathbf{Q}_i$. Then

$$\begin{aligned}
\mathbb{E}\left[\|\Delta_{t+1}\|^2\right] &= \frac{1}{d}\operatorname{trace}\left(\mathbb{E}\left[\mathbf{Q}_1\ldots\mathbf{Q}_{t-1}\mathbf{Q}_{t-1}\ldots\mathbf{Q}_0^2\right]\right) \\
\Rightarrow \qquad \lim_{n\to\infty}\mathbb{E}\left[\|\Delta_{t+1}\|^2\right] &= \lim_{n\to\infty}\frac{1}{d}\operatorname{trace}\left(\mathbb{E}\left[\mathbf{Q}_1\ldots\mathbf{Q}_{t-1}\mathbf{Q}_{t-1}\ldots\mathbf{Q}_0^2\right]\right) \\
&= \lim_{n\to\infty}\frac{1}{d}\operatorname{trace}\left(\mathbb{E}\left[\mathbf{Q}_0^2\right]\right)\lim_{n\to\infty}\frac{1}{d}\operatorname{trace}\left(\mathbb{E}\left[\mathbf{Q}_2\ldots\mathbf{Q}_{t-1}\mathbf{Q}_{t-1}\ldots\mathbf{Q}_1^2\right]\right) \\
&= \prod_{i=0}^{t-1}\lim_{n\to\infty}\frac{1}{d}\operatorname{trace}\left(\mathbb{E}\left[\mathbf{Q}_i^2\right]\right)
\end{aligned}$$

using the fact that $\mathbf{Q}_0^2$ is asymptotically free from $\mathbf{Q}_{t-1}\ldots\mathbf{Q}_1$ and recursing. The expectation of $\mathbf{Q}_i$ is given by

$$\begin{aligned}
\mathbb{E}\left[\mathbf{Q}_i^2\right] &= \mathbf{I}_d - \alpha_i\mathbb{E}\left[\left(\mathbf{U}^\top \mathbf{S}_i^\top \mathbf{S}_i \mathbf{U} + \lambda\Sigma^{-2}\right)^{-1}\right]\left(\mathbf{I}_d + \lambda\Sigma^{-2}\right) \\
&\quad - \alpha_i\left(\mathbf{I}_d + \lambda\Sigma^{-2}\right)\mathbb{E}\left[\left(\mathbf{U}^\top \mathbf{S}_i^\top \mathbf{S}_i \mathbf{U} + \lambda\Sigma^{-2}\right)^{-1}\right] \\
&\quad + \alpha_i^2\left(\mathbf{I}_d + \lambda\Sigma^{-2}\right)\mathbb{E}\left[\left(\mathbf{U}^\top \mathbf{S}_i^\top \mathbf{S}_i \mathbf{U} + \lambda\Sigma^{-2}\right)^{-2}\right]\left(\mathbf{I}_d + \lambda\Sigma^{-2}\right)
\end{aligned}$$

and the normalized limiting trace is given by

$$\begin{aligned}
\lim_{n\to\infty}\frac{1}{d}\operatorname{trace}\left(\mathbb{E}\left[\mathbf{Q}_i^2\right]\right) &= 1 - \frac{2\alpha_i}{d}\lim_{n\to\infty}\operatorname{trace}\left(\mathbb{E}\left[\left(\mathbf{U}^\top \mathbf{S}_i^\top \mathbf{S}_i \mathbf{U} + \lambda\Sigma^{-2}\right)^{-1}\right]\left(\mathbf{I}_d + \lambda\Sigma^{-2}\right)\right) \\
&\quad + \frac{\alpha_i^2}{d}\lim_{n\to\infty}\operatorname{trace}\left(\mathbb{E}\left[\left(\mathbf{U}^\top \mathbf{S}_i^\top \mathbf{S}_i \mathbf{U} + \lambda\Sigma^{-2}\right)^{-2}\right]\left(\mathbf{I}_d + \lambda\Sigma^{-2}\right)^2\right)
\end{aligned}$$

TODOproblem: cannot derive limiting spectral distribution of $\mathbf{U}^\top \mathbf{S}_i^\top \mathbf{S}_i \mathbf{U} + \lambda \Sigma^{-2}$ following approach of Lemma A.1 as matrix is not orthogonal.

2.1 Full Newton sketch

TODO: instead of sketching $\mathbf{X}^\top \mathbf{X}$, can we sketch $\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_d$? For example, take Cholesky decomposition $\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_d = \mathbf{A}\mathbf{A}^\top$ and sketch $\mathbf{A}\mathbf{S}^\top \mathbf{S}\mathbf{A}^\top$. We only have to compute Cholesky once, but $\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_d$ still needs to be computed beforehand. Are there still computational benefits of sketching in this case?

2.2 Scratch notes

Newton's method with least squares:

$$\begin{aligned}
\mathbf{b}_{t+1} &= \mathbf{b}_t - \alpha_t \mathbf{H}^{-1} \nabla f(\mathbf{b}_t) \\
\mathbf{b}_{t+1} &= \mathbf{b}_t - \alpha_t \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \left(\mathbf{X}^\top \mathbf{X}\mathbf{b}_t - \mathbf{X}^\top \mathbf{y}\right) \\
&= \mathbf{b}_t - \alpha_t \left(\mathbf{b}_t - \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}\right) \\
\mathbf{X}\mathbf{b}_{t+1} &= \mathbf{X}\mathbf{b}_t - \alpha_t \left(\mathbf{X}\mathbf{b}_t - \mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}\right) \\
\mathbf{X}(\mathbf{b}_{t+1} - \mathbf{b}^*) &= \mathbf{X}\mathbf{b}_t - \alpha_t \left(\mathbf{X}\mathbf{b}_t - \mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}\right) - \mathbf{X}\mathbf{b}^* \\
&= \mathbf{X}\mathbf{b}_t - \alpha_t \left(\mathbf{X}\mathbf{b}_t - \mathbf{X}\mathbf{b}^*\right) - \mathbf{X}\mathbf{b}^* \\
&= \left(\mathbf{I}_n - \alpha_t \mathbf{I}_n\right) \mathbf{X}(\mathbf{b}_t - \mathbf{b}^*)
\end{aligned}$$

IHS: let $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^\top$ where $\mathbf{U}$, $\mathbf{V}$ orthogonal and Σ diagonal. Then

$$\begin{aligned}
\mathbf{X}\left(\mathbf{X}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{X}\right)^{-1} \mathbf{X}^T &= \mathbf{U}\Sigma\mathbf{V}^\top \left(\mathbf{V}\Sigma\mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U}\Sigma\mathbf{V}^\top\right)^{-1} \mathbf{V}\Sigma\mathbf{U}^\top \\
&= \mathbf{U}\left(\mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U}\right)^{-1} \mathbf{U}^\top
\end{aligned}$$

$$\begin{aligned}
\mathbf{b}_{t+1} &= \mathbf{b}_t - \alpha_t \left(\mathbf{X}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{X}\right)^{-1} \left(\mathbf{X}^\top \mathbf{X}\mathbf{b}_t - \mathbf{X}^\top \mathbf{y}\right) \\
&= \mathbf{b}_t - \alpha_t \left(\mathbf{X}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{X}(\mathbf{b}_t - \mathbf{b}^*) \\
\mathbf{X}\mathbf{b}_{t+1} - \mathbf{X}\mathbf{b}^* &= \mathbf{X}\mathbf{b}_t - \alpha_t \mathbf{X}\left(\mathbf{X}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{X}(\mathbf{b}_t - \mathbf{b}^*) - \mathbf{X}\mathbf{b}^* \\
&= \mathbf{X}\mathbf{b}_t - \alpha_t \mathbf{U}\left(\mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U}\right)^{-1} \mathbf{U}^\top \mathbf{X}(\mathbf{b}_t - \mathbf{b}^*) - \mathbf{X}\mathbf{b}^* \\
&= \left(\mathbf{I}_n - \alpha_t \mathbf{U}\left(\mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U}\right)^{-1} \mathbf{U}^\top\right) \mathbf{X}(\mathbf{b}_t - \mathbf{b}^*)
\end{aligned}$$

Let $\Delta_t = \mathbf{U}^\top \mathbf{X}(\mathbf{b}_t - \mathbf{b}^*)$. Then

$$\begin{aligned}
\Delta_{t+1} &= \left(\mathbf{U}^\top - \alpha_t \mathbf{U}^\top \mathbf{U}\left(\mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U}\right)^{-1} \mathbf{U}^\top\right) \mathbf{X}(\mathbf{b}_t - \mathbf{b}^*) \\
&= \left(\mathbf{I}_d - \alpha_t \left(\mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U}\right)^{-1}\right) \Delta_t \\
\|\Delta_{t+1}\|^2 &= \Delta_t^\top \left(\mathbf{I}_d - \alpha_t \left(\mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U}\right)^{-1}\right) \Delta_t
\end{aligned}$$

Note: Hessian approximated before simplification; otherwise work does not follow

$$\begin{aligned}
\mathbf{X}(\mathbf{b}_{t+1} - \mathbf{b}^*) &= \mathbf{X}\mathbf{b}_t - \alpha_t \left(\mathbf{X}\mathbf{b}_t - \mathbf{X}\left(\mathbf{X}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}\right) - \mathbf{X}\mathbf{b}^* \\
&= \mathbf{X}\mathbf{b}_t - \alpha_t \left(\mathbf{X}\mathbf{b}_t - \mathbf{U}\left(\mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U}\right)^{-1} \mathbf{U}^\top \mathbf{y}\right) - \mathbf{X}\mathbf{b}^*
\end{aligned}$$

3 Scratch notes

3.1 Random matrix theory

- Stieltjes transform:
  https://mathoverflow.net/questions/79109/intuitive-understanding-of-the-stieltjes-transform
  https://terrytao.wordpress.com/2010/02/02/254a-notes-4-the-semi-circular-law/
  Berkeley lecture notes: https://www.stat.berkeley.edu/~songmei/Teaching/STAT260_Spring2021/Lecture_notes/scribe_lecture17.pdf
  Distribution of sums of independent commutative r.v.'s can be quickly computed using characteristic function; analogous for distribution of sums of freely independent non-commutative r.v.'s is using the Stieltjes transform
  Stieltjes transform of e.s.d.of a random matrix is easy to compute [2]
- S-transform: under suitable conditions, the S-transform of the l.s.d.of a matrix product $\mathbf{AB}$ is the product of the S-transforms of the l.s.d.of $\mathbf{A}$ and the l.s.d.of $\mathbf{B}$ [2]
- η-transform: purpose mainly to simplify long derivations involving Stieltjes transform [2]
- Free probability (non-commutative random variables, e.g., random matrices):
  https://terrytao.wordpress.com/2010/02/10/245a-notes-5-free-probability/
  Light intro: https://arxiv.org/pdf/1902.10763.pdf [4]
  - Freeness between non-commutative random variables analogous to independence in commutative random variables
  - Semicircular distribution of free CLT analogous to Gaussian distribution of classical CLT; the asymptotic distributions of eigenvalues of Hermition Gaussian random matrices (Wigner's semicircle law); sub-Gaussian distribution
    Example: https://mathworld.wolfram.com/WignersSemicircleLaw.html
  - Cauchy transform = Stieltjes transform (or closely related)? Theorem 3 says that Cauchy transforms and probability measures on $\mathbb{R}$ are in one-to-one correspondence
  - One method to construct Haar unitary matrices: take $n \times n$ random matrix with i.i.d. Gaussian random variables and apply the Gram-Schmidt procedure
- Marchenko-Pastur proof given in [2]

3.2 Newton's method

Problem: for f twice-differentiable, minimize

$$\min_{x \in \mathbb{R}} f(x) \ .$$

Second-order Taylor expansion around iterate x_t:

$$f(x_t + a) \approx f(x_t) + f'(x_t)a + \frac{1}{2} f''(x_t)a^2 \ .$$

If second derivative positive, minimum at derivative zero and so

$$f'(x_t) + f''(x_t)a = 0 \qquad \Rightarrow \qquad a = -\frac{f'(x_t)}{f''(x_t)} \ .$$

Newton update:

$$x_{t+1} = x_t + a = x_t - \frac{f'(x_t)}{f''(x_t)} \ .$$

3.3 Underdetermined least squares

Is there anything interesting about IHS in the underdetermined least squares case?

- For example, gradient descent in underdetermined least squares converges to the minimum norm solution (assuming solution initialized in orthogonal complement of null space of data matrix, i.e., row space).
  Implicit Regularization in Matrix Factorization; SE

- If $n \ll d$, does it make more sense to reduce the feature dimension, i.e., $\tilde{A} = AS$ for A $ntimesd$, S $d \times m$, $m \ll d$? For fixed sketches? Otherwise are there any computational savings from sketching? Note A^TA not invertible in this case but AA^T is.
  High-Dimensional Optimization in Adaptive Random Subspaces: uses right sketching matrix, works with AA^T.
  Effective Dimension Adaptive Sketching Methods for Faster Regularized Least-Squares Optimization: dual of underdetermined is overdetermined problem?

- Ridge regression allows for an unique solution.
  An Iterative, Sketching-based Framework for Ridge Regression

3.4 Quantile regression

Are there problems for which IHS could perform better than conventional methods?

- Models conditional quantiles. Makes no assumptions about distribution of response nor homoscedasticity. Is equivariant to transformation. Need more data than linear regression and is computationally intensive.
  What is quantile regression
  Five Things You Should Know about Quantile Regression
  Quantile Regression
  An Interior Point Algorithm for Nonlinear Quantile Regression
  An Improved Interior Point Algorithm for Quantile Regression
  Computational Issues for Quantile Regression
  Smoothed Quantile Regression with Large-Scale Inference
  Computational Methods for Quantile Regression
  Quantile Regression for Large-scale Applications: pre-IHS stochastic algorithm?

  The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators: implemented in R; interior point method works by maximizing (differentiable) dual of primal function (minimizing check loss) and translating constraints into a barrier function. Under conditions (?), overall complexity is $O_p(n^{1+a}p^3 \log n)$ given a $n \times p$ constraint matrix and $a < \frac{1}{2}$.
  Primal-Dual Interior-Point Methods
  On the Implementation of a Primal-Dual Interior Point Method

  Newton Sketch: A Linear-time Optimization Algorithm with Linear-Quadratic Convergence

3.5 Gradient flow

Proximal algorithms?

- Proximal algorithms

3.6 Related literature

- Extensions of IHS:
  Distributed Averaging Methods for Randomized Second Order Optimization
  Iterative Hessian Sketch with Momentum TODO: how does this work differ or not contradict the

current work?
Optimal Randomized First-Order Methods for Least-Squares Problems: extension of current work with non-refreshed sketches
Adaptive Newton Sketch: Linear-time Optimization with Quadratic Convergence and Effective Hessian Dimensionality
Newton-LESS: Sparsification without Trade-offs for the Sketched Newton Update: sparse IHS?
Asymptotics for Sketching in Least Squares: referenced paper in current work; compares one-step sketch matrices

- Applications:
  Randomized sketches for kernels: Fast and optimal nonparametric regression
- Possibly helpful:
  On randomized sketching algorithms and the Tracy-Widom law
  Practical sketching algorithms for low-rank matrix approximation
  A Statistical Perspective on Randomized Sketching for Ordinary Least-Squares: complementary to 2016 IHS paper
  Efficient Matrix Sketching over Distributed Data: PCA?
  Randomized Iterative Methods for Linear Systems: classical sketch
  Optimal approximate matrix product in terms of stable rank: error bounds on matrix approximations
- Related to expectation of solution?
  High-Dimensional Optimization in Adaptive Random Subspaces
  Randomized sketches for kernels: Fast and optimal nonparametric regression
- Underdetermined?
  Effective Dimension Adaptive Sketching Methods for Faster Regularized Least-Squares Optimization: refreshed embeddings does not improve on fixed embedding?
  - Faster Least Squares Optimization
  - Optimal Randomized First-Order Methods for Least-Squares Problems: first-order method with fixed sketches that provide better guarantees.

References

[1] Agniva Chowdhury, Jiasen Yang, and Petros Drineas. An iterative, sketching-based framework for ridge regression. In *International Conference on Machine Learning*, pages 989–998. PMLR, 2018.

[2] Romain Couillet and Mérouane Debbah. *The Stieltjes transform method*, page 35–70. Cambridge University Press, 2011. doi: 10.1017/CBO9780511994746.005.

[3] 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.

[4] Xiang-Gen Xia. A simple introduction to free probability theory and its application to random matrices. *arXiv preprint arXiv:1902.10763*, 2019.