# TODO

Kenny Chiu

February 19, 2022

# 1 Summary

## 1.1 Context and background

Lacotte et al. [12] study the performance of iterative Hessian sketch (IHS) [16] 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. [12] 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. [12] include several theoretical results that describe the (asymptotically) optimal value of the parameters for IHS with refreshed 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 refreshed Haar sketches, the optimal convergence rate $\rho_H$ 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 $\rho_G$ 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 $\rho_S$ for IHS with refreshed SRHT sketches is equal to $\rho_H$ under an additional mild assumption on the initialization of the least squares problem (which was not needed for Haar sketches due to existing results from 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

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

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 when there is no momentum and therefore so is the optimal step size. 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. [12] 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) [21] 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. [12] 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. [12] 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 [7, 16]. While certain criteria are intrinsically related [8], they may still have differing properties and lead to differing results [7].

The main limitation of the work by Lacotte et al. [12] 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. [12] is said to be inspired by and therefore most similar to the work by Dobriban and Liu [7], 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, 16, 19]. More recently, Lacotte and Pilanci [11] directly extended their analysis of refreshed 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 [14], distributed IHS [5], first-order IHS with adaptive step sizes [24], and Newton sketch [17] (IHS for general convex optimization problems) and its own variants [e.g., 6, 13]. 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

Quantile regression [10] 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 [20]. Given a data matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, observations $\mathbf{y} \in \mathbb{R}^n$ and a quantile $\tau \in (0,1)$, quantile regression fits a linear model on the quantile with the estimated parameters being 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 objective is non-differentiable as-is but can be optimized as a linear program. For large data problems, the interior point method transforms the dual program into a constrained optimization problem using log barriers [18], i.e.,

$$\begin{aligned} &\arg\max_{\mathbf{a}} \quad \mathbf{y}^\top\mathbf{a} & \Longrightarrow \qquad & \arg\max_{\mathbf{a}} \mathbf{y}^\top\mathbf{a} + \mu \sum_{i=1}^{n} \log a_i \\ &\text{s.t.} \begin{cases} \mathbf{X}^\top\mathbf{a} = (1-\tau)\mathbf{X}^\top\mathbf{1}_n \\ \mathbf{a} \in [0,1]^n \end{cases} & & \text{s.t. } \mathbf{X}^\top\mathbf{a} = (1-\tau)\mathbf{X}^\top\mathbf{1}_n \ . \end{aligned}$$

The interior point method solves the dual program by taking a sequence of Newton steps with $\mu \to 0$. The solution to the Newton step in each iteration satisfies the equation

$$\mathbf{X}^\top\mathbf{W}_t\mathbf{X}\mathbf{b}_t = \mathbf{X}^\top\mathbf{W}_t\left(\mathbf{y} + \mu\mathbf{A}^{-1}\mathbf{1}_n\right) \ .$$

where $\mathbf{A}$ is a $n \times n$ diagonal matrix and $\mathbf{W}_t$ is a $n \times n$ diagonal matrix with positive diagonal entries that changes each iteration. Computing $\mathbf{X}^\top\mathbf{W}_t\mathbf{X}$ is the main computational bottleneck that leads to each iteration having a cost of $O(nd^2)$ [1].

In this proposed project, we consider the case where $d \ll n$ and propose a stochastic interior point algorithm that uses sketching matrices to reduce the cost of the iterative updates. Drawing on methods from the sketching literature [17], the idea is to incorporate partial sketching 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 dimension-reducing random matrix regenerated every iteration. 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 $O(nd\log m)$ [12], and so the matrix product above can be computed at a cost of $O(md^2)$. While the sketched solution will only be an approximation of the original, recent work on the convergence of sketching in other optimization problems show promising theoretical and empirical results [e.g., 6, 13, 17]. We note that Yang et al. [23] had previously proposed a stochastic algorithm for quantile regression. However, their method differs 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 obtained from the proposed sketched algorithm and other existing methods, such as the standard interior point method [18] (implemented in R), the stochastic method by Yang et al. [23], a more modern iteration of the interior point method [25], and a modern quantile regression algorithm based on smoothing [9] (also implemented in R), in large dataset settings.

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 in 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 [17] for interior point methods and partial sketches, which would provide a worst-case convergence guarantee for 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. [12] would likely be necessary. However, adapting their least squares approach 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 not be useful for the $n \ll d$ case. 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 [15] did for LASSO, or sketching both the data matrix and the observations as in classical least-squares sketch (although this has been shown to lead to suboptimal performance [16]).

2. Applications of quantile regression or interior point algorithms in general, e.g., composite quantile regression [26] for high-dimensional regression, applications of quadratic progamming, may benefit from the faster algorithm.

3. Sketched-based smoothing algorithms for quantile regression. These algorithms approximate the original optimization problem by a differentiable one and therefore sketching should directly follow from the work of Pilanci and Wainwright [17]. Given the more standard setup, it is likely easier to analyze these algorithms than the interior point algorithms.

# 3 Project report

## 3.1 Introduction

Ridge regression is a special case of regularized least squares where the penalty function is chosen to be the $\ell_2$-norm of the model parameters. Given data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, observations $\mathbf{y} \in \mathbb{R}^n$ and a regularization parameter $\lambda > 0$, ridge regression obtains estimates of the parameters as the solution to the optimization problem

$$\mathbf{b}^* = \arg\min_{\mathbf{b} \in \mathbb{R}^d} \frac{1}{2} \|\mathbf{X}\mathbf{b} - \mathbf{y}\|_2^2 + \frac{\lambda}{2} \|\mathbf{b}\|_2^2 \,.$$

While ridge regression can be motivated as a method for reducing overfitting in ordinary least squares (OLS), it also has its computational and analytical benefits over OLS. When $\mathbf{X}$ does not have full column rank (e.g., when $n < d$), then $\mathbf{X}^\top \mathbf{X}$ is singular and the OLS solution is non-unique. When $\mathbf{X}$ is full rank but ill-conditioned, then small changes in $\mathbf{X}$ lead to large changes in $(\mathbf{X}^\top \mathbf{X})^{-1}$ and consequently in the OLS solution. Ridge regression addresses both of these issues by minimizing the variance and mean squared error at the cost of introducing a small bias [3]. The ridge regression solution is unique and is given by

$$\mathbf{b}^* = \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_d\right)^{-1} \mathbf{X}^\top \mathbf{y} \,.$$

In this report, we analyze the theoretical properties of a partial Newton sketch algorithm [17] as an iterative solver for the ridge regression problem. In particular, we attempt to derive an optimal convergence rate and an optimal step size following the approach of Lacotte et al. [12] for iterative Hessian sketch with OLS using asymptotic results from random matrix theory and free probability. We show that while ridge regression can be considered a simple extension to OLS, extending the analysis approach of Lacotte et al. [12] to partial Newton sketch is not trivial. TODO

This report is organized as follows: Section 3.2 provides background about sketching and describes the OLS results by Lacotte et al. [12] that we aim to extend to ridge regression; Section 3.3 highlights relevant work in the literature; Section 3.5 discusses our attempts to analyze Newton sketch for ridge regression and the key differences from OLS that makes the problem challenging; Section 3.6 describes TODO; and Section 3.7 summarizes our findings and concludes this report.

## 3.2 Background

This section provides additional background about Newton and iterative Hessian sketch. TODO

### 3.2.1 Sketching and Newton sketch

### 3.2.2 Random matrix theory and free probability

### 3.2.3 Notation

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

### 3.3 Related work

Chowdhury et al. [3]

Wang et al. [22]

Cohen et al. [4]

Lacotte et al. [12]

### 3.4 Newton sketch for ridge regression

Consider the ridge regression loss function

$$f(\mathbf{b}) = \frac{1}{2}\|\mathbf{X}\mathbf{b} - \mathbf{y}\|_2^2 + \frac{\lambda}{2}\|\mathbf{b}\|_2^2$$

for $\mathbf{b} \in \mathbb{R}^d$ given data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ with $d \ll n$, responses $\mathbf{y} \in \mathbb{R}^n$ and a regularization parameter $\lambda > 0$. The gradient and Hessian of the function are respectively given by

$$\begin{aligned} \nabla f(\mathbf{b}) &= \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_d\right) \mathbf{b} - \mathbf{X}^\top \mathbf{y} \ , \\ \mathbf{H} = \nabla^2 f(\mathbf{b}) &= \mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_d \ . \end{aligned}$$

An iterative solver based on Newton's method for minimizing the loss function computes the updates

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

We consider a partial Newton sketch variant of this update that approximates the Hessian by the sketched version

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

where $\mathbf{S}_t$ is a $m \times n$ sketching matrix, $m \ll n$, that is resampled every iteration. TODO:Chowdhury et al. [3] appendix results?. The partial Newton sketch updates for ridge regression are then given by

$$\mathbf{b}_{t+1} = \mathbf{b}_t - \alpha_t \left(\mathbf{X}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{X} + \lambda \mathbf{I}_d\right)^{-1} \left(\left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_d\right) \mathbf{b}_t - \mathbf{X}^\top \mathbf{y}\right) \ .$$

Note that Chowdhury et al. [3] and Wang et al. [22]) also considered this sketched update for ridge regression. However, our analysis of this method differs from theirs in that we adopt the asymptotic random matrix theoretic approach from Lacotte et al. [12]. Also note that we do not consider updates with momentum as Lacotte et al. [12] did as we show that extending their analysis approach from OLS to ridge regression is already non-trivial.

### 3.5 Analysis attempt based on random matrix theory

In this section, we show that the proof technique used to obtain Theorems 3.1 and 4.1 of [12] do not easily generalize to the partial Newton sketch updates for ridge regression. We follow the general procedure of the proofs and show how far we can get with the ridge regression setup. We also highlight the key differences between OLS and ridge regression that leads to problems in the proof and discuss possible solutions for rectifying these problems in future work.

The following conjecture formalizes the result analogous to Theorems 3.1 and 4.1 that we would like to prove. Note that additional assumptions will be added to the conjecture as we progress through the proof.

**Conjecture 1.** *Consider the partial Newton update for ridge regression described in Section 3.4. For some optimal step size $\alpha_t$, the sequence of error vectors $\{\Delta_t\}$ satisfies*

$$\rho = \left( \lim_{n\to\infty} \frac{\mathbb{E}\left[\|\Delta_t\|_2^2\right]}{\|\Delta_0\|_2^2} \right)^{\frac{1}{t}}$$

*where $\rho$ is the rate of convergence with some closed-form expression.*

We begin our attempt to prove Conjecture 1 following the proofs by Lacotte et al. [12]. Using the fact that the ridge regression solution satisfies the equation

$$(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d)\mathbf{b}^* = \mathbf{X}^\top\mathbf{Y} \ ,$$

the update can be rewritten as

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

Let $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^\top$ be the thin singular value decomposition of $\mathbf{X}$ where $\mathbf{U}$ is a $n \times d$ semi-orthogonal matrix, $\mathbf{V}$ is a $d \times d$ orthogonal matrix, and $\Sigma$ is a $d \times d$ diagonal matrix with the singular values of $\mathbf{X}$ on the diagonal. Then we can write

$$\begin{aligned}
\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d &= \mathbf{V}\Sigma^2\mathbf{V}^\top + \lambda\mathbf{V}\mathbf{V}^\top \\
&= \mathbf{V}\left(\Sigma^2 + \lambda\mathbf{I}_d\right)\mathbf{V}^\top \ , \\
\left(\mathbf{X}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{X} + \lambda\mathbf{I}_d\right)^{-1} &= \left(\mathbf{V}\Sigma\mathbf{U}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U}\Sigma\mathbf{V}^\top + \lambda\mathbf{V}\mathbf{V}^\top\right)^{-1} \\
&= \mathbf{V}\left(\Sigma\mathbf{U}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U}\Sigma + \lambda\mathbf{I}_d\right)^{-1}\mathbf{V}^\top \ .
\end{aligned}$$

However, in order to later on obtain an expression in terms of $\Delta_t$ as in the original proof, we require that the data matrix be full column rank. This is a less than ideal assumption to make as one of the advantages of ridge regression is being able to obtain an unique solution with non-full rank data matrices. We return to this point in a later discussion how we may avoid this assumption. TODO

**Assumption 1.** The data matrix $\mathbf{X}$ has full column rank.

Under Assumption 1, the singular values of $\mathbf{X}$ are non-zero and so the above matrices can be rewritten as

$$\begin{aligned}
\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d &= \mathbf{V}\Sigma\left(\mathbf{I}_d + \lambda\Sigma^{-2}\right)\Sigma\mathbf{V}^\top \ , \\
\left(\mathbf{X}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{X} + \lambda\mathbf{I}_d\right)^{-1} &= \mathbf{V}\Sigma^{-1}\left(\mathbf{U}^\top\mathbf{S}_t^\top\mathbf{S}_t\mathbf{U} + \lambda\Sigma^{-2}\right)^{-1}\Sigma^{-1}\mathbf{V}^\top \ .
\end{aligned}$$

Replacing the corresponding matrices in the update with these identities gives

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

Multiplying both sides by $\mathbf{U}^\top\mathbf{X}$ gives

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

and then subtracting both sides by $\mathbf{U}^\top\mathbf{X}\mathbf{b}^*$ gives

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

Let $\mathbf{Q}_t = \mathbf{I}_d - \alpha_t\left(\mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U} + \lambda\Sigma^{-2}\right)^{-1}\left(\mathbf{I}_d + \lambda\Sigma^{-2}\right)$. Therefore by definition, we have $\Delta_{t+1} = \mathbf{Q}_t\Delta_t$ and

$$\|\Delta_{t+1}\|^2 = \Delta_t^\top\mathbf{Q}_t^\top\mathbf{Q}_t\Delta_t \ .$$

Taking the expectation with respect to $\mathbf{S}_t$, we get

$$\mathbb{E}\left[\|\Delta_{t+1}\|^2\right] = \Delta_t^\top\mathbb{E}\left[\mathbf{Q}_t^\top\mathbf{Q}_t\right]\Delta_t$$

where

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

At this point, we run into our first major obstacle that prevents us from applying the key step of the proof of Theorem 3.1. In Theorem 3.1 for OLS, the expression that is obtained from taking the expectation is

$$\mathbb{E}\left[\|\Delta_{t+1}\|^2\right] = \Delta_t^\top\mathbb{E}\left[\mathbf{R}_t^2\right]\Delta_t$$

where

$$\mathbb{E}\left[\mathbf{R}_t^2\right] = \mathbf{I}_d - 2\alpha_t\mathbb{E}\left[\left(\mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U}\right)^{-1}\right] + \alpha_t^2\mathbb{E}\left[\left(\mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U}\right)^{-2}\right] \ .$$

The proof of Theorem 3.1 proceeds to recognize that the matrix $\mathbf{S}_t\mathbf{U}$ can be embedded into a Haar matrix and is therefore rotationally invariant. Using exchangeability arguments, the expectations $\mathbb{E}\left[\left(\mathbf{U}^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{U}\right)^{-p}\right]$ have a simple closed-form expression in terms of the inverse moments from which the rest of the proof follows. We do not have rotational invariance in our ridge regression case, and so we follow the proof of Theorem 4.1 from this point onwards. We require an additional assumption on the initialization of the problem.

**Assumption 2.** The initial error vector $\Delta_0$ is random and satisfies $\mathbb{E}\left[\Delta_0\Delta_0^\top\right] = \frac{\mathbf{I}_d}{d}$.[1]

Under Assumption 2, taking the expectation with respect to $\mathbf{S}_t$ instead gives

$$\begin{aligned}\mathbb{E}\left[\|\Delta_{t+1}\|^2\right] &= \mathbb{E}\left[\Delta_t^\top \mathbf{Q}_t^\top \mathbf{Q}_t \Delta_t\right]\\ &= \mathbb{E}\left[\Delta_0^\top \mathbf{Q}_0^\top \dots \mathbf{Q}_t^\top \mathbf{Q}_t \dots \mathbf{Q}_0 \Delta_0\right]\\ &= \mathbb{E}\left[\operatorname{trace}\left(\Delta_0^\top \mathbf{Q}_0^\top \dots \mathbf{Q}_t^\top \mathbf{Q}_t \dots \mathbf{Q}_0 \Delta_0\right)\right]\\ &= \operatorname{trace}\left(\mathbb{E}\left[\mathbf{Q}_0^\top \dots \mathbf{Q}_t^\top \mathbf{Q}_t \dots \mathbf{Q}_0 \Delta_0 \Delta_0^\top\right]\right)\end{aligned}$$

TODO: are $\mathbf{Q}$ and rank one matrix $\Delta_0\Delta_0^\top$ free?

TODOsatisfies conditions (limiting spectral distribution as $n \to \infty$ and others?). Then we have $\Delta_{t+1} = \mathbf{Q}_t\Delta_t$ and so

$$\begin{aligned}\mathbb{E}\left[\|\Delta_{t+1}\|^2\right] &= \operatorname{trace}\left(\mathbb{E}\left[\Delta_0^\top \mathbf{Q}_0 \dots \mathbf{Q}_{t-1}\mathbf{Q}_{t-1} \dots \mathbf{Q}_0\Delta_0\right]\right)\\ &= \operatorname{trace}\left(\mathbb{E}\left[\mathbf{Q}_0 \dots \mathbf{Q}_{t-1}\mathbf{Q}_{t-1} \dots \mathbf{Q}_0\Delta_0\Delta_0^\top\right]\right)\\ &= \operatorname{trace}\left(\mathbb{E}\left[\mathbf{Q}_1 \dots \mathbf{Q}_{t-1}\mathbf{Q}_{t-1} \dots \mathbf{Q}_1\mathbf{Q}_0^2\right]\mathbb{E}\left[\Delta_0\Delta_0^\top\right]\right)\\ &= \frac{1}{d}\operatorname{trace}\left(\mathbb{E}\left[\mathbf{Q}_1 \dots \mathbf{Q}_{t-1}\mathbf{Q}_{t-1} \dots \mathbf{Q}_1\mathbf{Q}_0^2\right]\right)\end{aligned}$$

using the independence of $\Delta_0$ and $\mathbf{Q}_i$ and using Assumption 2. Then taking the limit in $n$ and recursively applying the fact that $\mathbf{Q}_0^2$ is asymptotically free from $\mathbf{Q}_{t-1}\dots\mathbf{Q}_1$ TODO, we get

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

TODOThe expectation of $\mathbf{Q}_i$ is given by

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

and the normalized limiting trace is given by

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

[1] The convergence rate $\rho$ in the statement of Conjecture 1 is also redefined as $\left(\lim_{n\to\infty}\frac{\mathbb{E}\left[\|\Delta_t\|_2^2\right]}{\mathbb{E}\left[\|\Delta_0\|_2^2\right]}\right)^{\frac{1}{t}}$.

#### 3.5.1 Full column rank assumption

### 3.6 Empirical experiments

### 3.7 Conclusion

# 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] 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.

[4] Michael B Cohen, Cameron Musco, and Christopher Musco. Input sparsity time low-rank approximation via ridge leverage score sampling. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1758–1777. SIAM, 2017.

[5] 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.

[6] 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.

[7] 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.

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

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

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] Shusen Wang, Alex Gittens, and Michael W Mahoney. Sketched ridge regression: Optimization perspective, statistical perspective, and model averaging. In *International Conference on Machine Learning*, pages 3608–3616. PMLR, 2017.

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

[24] 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.

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

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