Randomized Algorithms for Least Squares

Krishna Balasubramanian

March 31, 2020

1 Introduction

In this notes, we will study randomized algorithm for computing ordinary least squares and ridge regression. OLS and Ridge regression are popular techniques for studying *linear relationships* between the covariates ($X^{(i)} \in \mathbb{R}^d$) and response ($Y^{(i)} \in \mathbb{R}$) given n samples. We denote the data matrix, with $X^{(i)}$ as its column, by $\mathbf{X} \in \mathbb{R}^{n \times d}$. Similarly, we denote the response vector, with $Y^{(i)}$ as its entries, by the vector $\mathbf{y} \in \mathbb{R}^n$. Then the OLS estimate of the linear relationship between $\mathbf{X}$ and $\mathbf{y}$ is given by the following problem:

$$\boldsymbol{\beta} = \arg\min_{\mathbf{c} \in \mathbb{R}^d} \|\mathbf{X}\mathbf{c} - \mathbf{y}\|_2^2. \tag{1}$$

In statistics, typically the $\mathbf{X}$ is assumed to be random matrix and a linear model assumption is explicitly made between $\mathbf{X}$ and $\mathbf{y}$ to study the statistical properties of the estimator $\boldsymbol{\beta}$. Our discussion below proceeds by conditioning on all the randomness in the data matrix and the response. That is, we assume $\mathbf{X}$ and $\mathbf{y}$ to be deterministic for the purpose of discussion.

The solution of the problem 1 satisfies the following normal equations:

$$\mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^\top \mathbf{y}. \tag{2}$$

When $\mathbf{X}$ has full column-rank, then we can write the solution of 1 in closed form as follows:

$$\boldsymbol{\beta} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}.$$

If we have the following SVD for $\mathbf{X} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^\top$, then we have

$$\boldsymbol{\beta} = \mathbf{V}\boldsymbol{\Lambda}^{-1}\mathbf{U}^\top \mathbf{y}.$$

Furthermore, the residual vector $\mathbf{r}_{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \in \mathbb{R}^n$ is orthogonal to the column space of $\mathbf{X}$. That is, we have

$$\mathbf{r}^\top \mathbf{X} = \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^\top \mathbf{X} = 0$$

Direct methods, that use the closed form solution have a complexity of $\mathcal{O}(nd^2)$. Iterative methods for directly solving the optimization problem above also have similar complexity. When considering the regime of of d being small but $d^2 \leq n \leq e^d$, these methods are computationally prohibitive.

2 Sketching

The idea of sketching is simple: One can potentially construct a matrix $\mathbf{\Phi} \in \mathbb{R}^{r \times n}$ and solve the following **sketched-OLS** problem instead of OLS:

$$\boldsymbol{\beta}_s = \arg\min_{\mathbf{c} \in \mathbb{R}^d} \|\mathbf{\Phi}(\mathbf{X}\mathbf{c} - \mathbf{y})\|_2^2 = \arg\min_{\mathbf{c} \in \mathbb{R}^d} \|\mathbf{\Phi}\mathbf{X}\mathbf{c} - \mathbf{\Phi}\mathbf{y}\|_2^2$$

Similar to the previous case, we have the closed form expression,

$$\boldsymbol{\beta}_s = \left((\mathbf{\Phi}\mathbf{X})^\top \mathbf{\Phi}\mathbf{X}\right)^{-1} (\mathbf{\Phi}\mathbf{X})^\top \mathbf{\Phi}\mathbf{y}. \tag{3}$$

Remark 2.0.1. The matrix $\mathbf{\Phi}$ is called as the **sketching matrix**. It reduces the dimensionality of the problem with respect to the sample size n. The other alternative to sketching is the trivial method: just **throw away** part of the samples. When we instead use the sketching idea and if the matrix $\mathbf{\Phi}$ is constructed **carefully**, we would have a huge computational saving with very little loss in accuracy of the obtained solution.

A popular way to do Sketched-OLS is described in Algorithm 1. Basically, the sketching matrix in this algorithms is defined by $\mathbf{\Phi} = \mathbf{S}^T\mathbf{H}\mathbf{D}$ where $\mathbf{S} \in \mathbb{R}^{n \times r}$ and $\mathbf{H}, \mathbf{D} \in \mathbb{R}^{n \times n}$. Notice that the algorithm fixes $r \approx d\log(n)/\epsilon$. Hence, even when $n \approx e^d$, we have $r \approx d^2/\epsilon$ which provides the computational benefits. Specifically, the sketched-OLS in Equation 3 could be computed in this case in $\mathcal{O}(rd^2) \approx d^4/\epsilon$ running time.

The matrix $\mathbf{S}$ and $\mathbf{D}$ are described in Algorithm 1. The matrix $\mathbf{H}$ is called as a *normalized* Hadamard transform matrix. First, the Hadamard transformation matrix is defined recursively as follows.

$$\tilde{\mathbf{H}}_2 = \begin{pmatrix} +1 & +1 \\ +1 & -1 \end{pmatrix} \qquad \text{and} \qquad \tilde{\mathbf{H}}_n = \begin{pmatrix} \tilde{\mathbf{H}}_{n/2} & \tilde{\mathbf{H}}_{n/2} \\ \tilde{\mathbf{H}}_{n/2} & -\tilde{\mathbf{H}}_{n/2} \end{pmatrix}$$

Then, the normalized Hadamard transformation matrix is defined as $\mathbf{H} = (1/\sqrt{n})\tilde{\mathbf{H}}_n$. The matrix $\mathbf{HD}$ is called as **randomized Hadamard transform** as $\mathbf{D}$ is random and has nice properties. Specifically, the matrix-vector multiplication $\mathbf{HDe}$ could be done in $n\log(n)$ time for any vector $\mathbf{e}$. **Note: A naive implementation might take $\mathcal{O}(n^2)$ time but the fast Hadamard transform approach outlined here has the above mentioned complexity**. The matrix $\mathbf{S}$ is a *sub-sampling* matrix which also contributes to the randomness in the algorithm. Recall that, we assume that $\mathbf{X}$ and $\mathbf{y}$ are deterministic in our discussion. That is, you can think as follows: all of the discussion above is done condition on $\mathbf{X}$ and $\mathbf{y}$ if they are random. So in the algorithm as such, the the only source of randomness in through $\mathbf{D}$ and $\mathbf{S}$. So, when you get to Fact 2.1.1 and encounter the phrase, **with high probability** we refer to this source of randomness only.

Algorithm 1 The **Sketched-OLS** algorithm

Input: $\mathbf{X} \in \mathbb{R}^{n \times d}$, $\mathbf{y} \in \mathbb{R}^n$, and an error parameter $\epsilon \in (0, 1)$.

Output: $\boldsymbol{\beta}_s \in \mathbb{R}^d$.

1. Let $r \approx \frac{d \log(n)}{\epsilon}$.
2. Let $\mathbf{S}$ be an empty matrix.
3. **For** $t = 1, \ldots, r$ (i.i.d. trials with replacement) **select uniformly at random** an integer from $\{1, 2, \ldots, n\}$.
   - **If** i is selected, **then** append the column vector $\left(\sqrt{n/r}\right) \mathbf{e}_i$ to $\mathbf{S}$, where $\mathbf{e}_i \in \mathbb{R}^n$ is the i-th canonical vector.
4. Let $\mathbf{H} \in \mathbb{R}^{n \times n}$ be the normalized Hadamard transform matrix.
5. Let $\mathbf{D} \in \mathbb{R}^{n \times n}$ be a diagonal matrix with
$$\mathbf{D}_{ii} = \begin{cases} +1 & \text{, with probability } 1/2 \\ -1 & \text{, with probability } 1/2 \end{cases}$$
6. Compute and return $\boldsymbol{\beta}_s = \left(\mathbf{S}^T \mathbf{H} \mathbf{D} \mathbf{X}\right)^\dagger \mathbf{S}^T \mathbf{H} \mathbf{D} \mathbf{y}$, where for a matrix $\mathbf{A}$, $\mathbf{A}^\dagger = (\mathbf{A}^\top \mathbf{A})^{-1}$.

2.1 What is a *good* sketching matrix ?

We provide some general statements that need to be satisfied for a matrix $\boldsymbol{\Phi}$ to be a "good" sketching matrix. Basically, any matrix $\boldsymbol{\Phi}$ that satisfies the two assumptions stated in Lemma 1 is called as a good sketching matrix. Instead of stating the assumption upfront, we outline the proofs from which we can see where the stated assumptions arise from. To fix notation, we denote the residual vectors with respect to OLS solution as $\mathbf{r}_{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ and the residual with the estimator $\boldsymbol{\beta}_s$ as $\mathbf{r}_{\boldsymbol{\beta}_s} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}_s$.

Theorem 2.1. *Let $\boldsymbol{\beta}$ be the OLS solution and $\boldsymbol{\beta}_s$ be the Sketched-OLS solution. Let the sketching matrix $\boldsymbol{\Phi}$ satisfy the two conditions stated in Lemma 1. Then we have*

$$\|\mathbf{r}_{\boldsymbol{\beta}_s}\|_2^2 \leq (1 + \epsilon)\|\mathbf{r}_{\boldsymbol{\beta}}\|_2^2$$
$$\|\boldsymbol{\beta} - \boldsymbol{\beta}_s\|_2^2 \leq \frac{\|\mathbf{r}_{\boldsymbol{\beta}}\|_2^2 \epsilon}{\sigma_{\min}^2(\mathbf{X})}$$

$$
\begin{aligned}
\|\mathbf{r}_{\boldsymbol{\beta}_s}\|_2^2 &= \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} + \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta}_s\|_2^2 \\
&= \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta}_s\|_2^2 \\
&= \|\mathbf{r}_{\boldsymbol{\beta}}\|_2^2 + \| - \mathbf{U}\boldsymbol{\theta}\|_2^2 \\
&\leq \|\mathbf{r}_{\boldsymbol{\beta}}\|_2^2 + \epsilon\|\mathbf{r}_{\boldsymbol{\beta}}\|_2^2 \qquad [\textit{by Equation } 4]. \\
&\leq (1+\epsilon)\|\mathbf{r}_{\boldsymbol{\beta}}\|_2^2.
\end{aligned}
$$

Lemma 1. *Consider the optimal value of the following two optimization problems:*

$$
\min_{\mathbf{c}\in\mathbb{R}^d} \|\boldsymbol{\Phi}\mathbf{X}\mathbf{c} - \boldsymbol{\Phi}\mathbf{y}\|_2^2 \qquad \min_{\mathbf{e}\in\mathbb{R}^d} \|\boldsymbol{\Phi}\mathbf{U}\mathbf{e} - \boldsymbol{\Phi}\mathbf{r}_{\boldsymbol{\beta}}\|_2^2.
$$

Then both values are the same, i.e., $\min_{\mathbf{c}\in\mathbb{R}^d} \|\boldsymbol{\Phi}\mathbf{X}\mathbf{c} - \boldsymbol{\Phi}\mathbf{y}\|_2^2 = \min_{\mathbf{e}\in\mathbb{R}^d} \|\boldsymbol{\Phi}\mathbf{U}\mathbf{e} - \boldsymbol{\Phi}\mathbf{r}_{\boldsymbol{\beta}}\|_2^2$. *Furthermore, if a vector* $\boldsymbol{\theta}$ *satisfies:*

$$
\mathbf{U}\boldsymbol{\theta} = \mathbf{X}(\boldsymbol{\beta}_s - \boldsymbol{\beta})
$$

then, it achieves the minimum of the optimization problem in the right hand size.

Proof. To see the first statement, note that

$$
\begin{aligned}
&\min_{\mathbf{c}\in\mathbb{R}^d} \|\boldsymbol{\Phi}\mathbf{X}\mathbf{c} - \boldsymbol{\Phi}\mathbf{y}\|_2^2 \\
&= \min_{\mathbf{c}\in\mathbb{R}^d} \|\boldsymbol{\Phi}\mathbf{X}\mathbf{c} - \boldsymbol{\Phi}(\mathbf{X}\boldsymbol{\beta} + \mathbf{r}_{\boldsymbol{\beta}})\|_2^2 \\
&= \min_{\mathbf{d}\in\mathbb{R}^d} \|\boldsymbol{\Phi}\mathbf{X}(\boldsymbol{\beta} + \mathbf{d}) - \boldsymbol{\Phi}(\mathbf{X}\boldsymbol{\beta} + \mathbf{r}_{\boldsymbol{\beta}})\|_2^2 \\
&= \min_{\mathbf{d}\in\mathbb{R}^d} \|\boldsymbol{\Phi}\mathbf{X}\mathbf{d} - \boldsymbol{\Phi}\mathbf{r}_{\boldsymbol{\beta}}\|_2^2 \\
&= \min_{\mathbf{e}\in\mathbb{R}^d} \|\boldsymbol{\Phi}\mathbf{U}\mathbf{e} - \boldsymbol{\Phi}\mathbf{r}_{\boldsymbol{\beta}}\|_2^2
\end{aligned}
$$

In the above derivation, in the last step, we are allowed to replace $\mathbf{X}$ with $\mathbf{U}$ as singular vectors in $\mathbf{U}$ form a basis for the column space of $\mathbf{X}$. Hence, we just have different combination of the vectors in $\mathbf{U}$ instead of vectors in $\mathbf{X}$. This proves the first statement.

To see the next statement, take the objective function of the optimization problems in the right hand side and note that

$$
\begin{aligned}
\|\boldsymbol{\Phi}\mathbf{U}\boldsymbol{\theta} - \boldsymbol{\Phi}\mathbf{r}_{\boldsymbol{\beta}}\|_2^2 &= \|\boldsymbol{\Phi}\mathbf{X}(\boldsymbol{\beta}_s - \boldsymbol{\beta}) - \boldsymbol{\Phi}\mathbf{r}_{\boldsymbol{\beta}}\|_2^2 \\
&= \|\boldsymbol{\Phi}\mathbf{X}\boldsymbol{\beta}_s - \boldsymbol{\Phi}\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\Phi}\mathbf{r}_{\boldsymbol{\beta}}\|_2^2 \\
&= \|\boldsymbol{\Phi}\mathbf{X}\boldsymbol{\beta}_s - \boldsymbol{\Phi}(\mathbf{X}\boldsymbol{\beta} + \mathbf{r}_{\boldsymbol{\beta}})\|_2^2 \\
&= \|\boldsymbol{\Phi}\mathbf{X}\boldsymbol{\beta}_s - \boldsymbol{\Phi}\mathbf{y}\|_2^2 \\
&= \min_{\mathbf{c}\in\mathbb{R}^d} \|\boldsymbol{\Phi}\mathbf{X}\mathbf{c} - \boldsymbol{\Phi}\mathbf{y}\|_2^2 \\
&= \min_{\mathbf{e}\in\mathbb{R}^d} \|\boldsymbol{\Phi}\mathbf{U}\mathbf{e} - \boldsymbol{\Phi}\mathbf{r}_{\boldsymbol{\beta}}\|_2^2
\end{aligned}
$$

Hence, we have shown that the vector $\boldsymbol{\theta}$ indeed is indeed the vector that achieves the minimum value of the optimization problem on the right hand side. □

Lemma 2. *Assume the following*

- *Isometry assumption 1:*

$$\sigma^2_{\min}(\boldsymbol{\Phi}\mathbf{U}) \geq \frac{1}{\sqrt{2}}$$

- *Isometry assumption 2:*

$$\|(\boldsymbol{\Phi}\mathbf{U})^\top \boldsymbol{\Phi}\mathbf{r}_{\boldsymbol{\beta}}\|_2^2 \leq \frac{\epsilon\|\mathbf{r}_{\boldsymbol{\beta}}\|_2^2}{2}$$

Then, we have the following bound for θ, from Lemma 1.

$$\|\boldsymbol{\theta}\|_2^2 \leq \epsilon\|\mathbf{r}_{\boldsymbol{\beta}}\|_2^2 \tag{4}$$

Proof. Because $\boldsymbol{\theta}$ is the solution of the optimization problem $\min_{\mathbf{e}\in\mathbb{R}^d}\|\boldsymbol{\Phi}\mathbf{U}\mathbf{e} - \boldsymbol{\Phi}\mathbf{r}_{\boldsymbol{\beta}}\|_2^2$, it should satisfy the normal equation:

$$(\boldsymbol{\Phi}\mathbf{U})^\top \boldsymbol{\Phi}\mathbf{U}\boldsymbol{\theta} = (\boldsymbol{\Phi}\mathbf{U})^\top \boldsymbol{\Phi}\mathbf{r}_{\boldsymbol{\beta}}$$

Hence, under our assmption, we have

$$\frac{\|\boldsymbol{\theta}\|_2^2}{2} \leq \|(\boldsymbol{\Phi}\mathbf{U})^\top \boldsymbol{\Phi}\mathbf{U}\boldsymbol{\theta}\|_2^2 = \|(\boldsymbol{\Phi}\mathbf{U})^\top \boldsymbol{\Phi}\mathbf{r}_{\boldsymbol{\beta}}\|_2^2 \leq \frac{\epsilon\|\mathbf{r}_{\boldsymbol{\beta}}\|_2^2}{2}$$

□

Remark 2.1.1. Any sketching matrix $\boldsymbol{\Phi}$ that satisfies the two assumptions stated in Lemma 2 is called as a "good" sketching matrix. The numbers in the assumptions are fixed just arbitrarily, i.e., instead of $\sigma^2_{\min}(\boldsymbol{\Phi}\mathbf{U}) \geq \frac{1}{\sqrt{2}}$, we can have $\sigma^2_{\min}(\boldsymbol{\Phi}\mathbf{U}) > 0$ and things will still work.

Remark 2.1.2. The assumptions are called as **isometry** assumptions because of a reason. A matrix or a transformation is called as isometric, if it preserves norms. That is, $\|\mathbf{A}\mathbf{e}\|_2^2 \approx \|\mathbf{e}\|_2^2$ for all $\mathbf{e}$ or **most** vectors $\mathbf{e}$. Sketching is exactly doing something like this, while reducing the dimension as well. The two assumptions are basically a way to characterize when the sketching matrix is isometric.

Fact 2.1.1. The Sketching matrix $\boldsymbol{\Phi} = \mathbf{S}^T\mathbf{H}\mathbf{D}$ constructed in the previous section, satisfies the properties of a good sketching matrix **with high probability** (as the matrix is a random matrix). If you are curious how to show this fact, you may refer to [Mah16].

References

[Mah16] Michael W Mahoney, *Lecture notes on randomized linear algebra*, arXiv preprint arXiv:1608.04481 (2016).