Appendix : On Coresets for Regularized Regression

August 10, 2020

Proof of Theorem 3.2

Proof. First we bound the sample size for a fixed query $\mathbf{q} \in Q$. Let s_i be the sensitivity of the i^{th} point $\mathbf{x_i}$ and S be the sum of the sensitivities. Let the sampling probability be $p_i = \frac{s_i}{S}$.
For all $\mathbf{q} \in Q$ and $\mathbf{x_i} \in \mathbf{X}$ define a function $g_{\mathbf{q}}(\mathbf{x_i}) = \frac{f_{\mathbf{q}}(\mathbf{x_i})}{Sp_i \sum_{j=1}^{n} f_{\mathbf{q}}(\mathbf{x_j})}$. So,

$$\mathbb{E}[g_{\mathbf{q}}(\mathbf{x_i})] = \frac{1}{S}$$

and

$$\frac{1}{r} \sum_{\substack{i \in [n] s.t. \\ \tilde{\mathbf{x_i}} \in \mathbf{C}}} g_{\mathbf{q}}(\mathbf{x_i}) = \frac{\sum_{\tilde{\mathbf{x_i}} \in \mathbf{C}} f_{\mathbf{q}}(\tilde{\mathbf{x_i}})}{S \sum_{\mathbf{x_i} \in \mathbf{X}} f_{\mathbf{q}}(\mathbf{x_i})}$$

Let

$$T = \sum_{\substack{i \in [n] s.t. \\ \tilde{\mathbf{x_i}} \in \mathbf{C}}} g_{\mathbf{q}}(\mathbf{x_i})$$

then

$$\mathbb{E}[T] = \sum_{\substack{i \in [n] s.t. \\ \tilde{\mathbf{x_i}} \in \mathbf{C}}} \mathbb{E}[g_{\mathbf{q}}(\mathbf{x_i})] = r/S$$

$$\begin{aligned}
\mathrm{var}(g_{\mathbf{q}}(\mathbf{x_i})) &\leq \mathbb{E}[(g_{\mathbf{q}}(\mathbf{x_i}))^2] \\
&= \sum_{\mathbf{x_i} \in \mathbf{X}} \frac{(f_{\mathbf{q}}(\mathbf{x_i}))^2}{(\sum_{j=1}^{n} f_{\mathbf{q}}(\mathbf{x_j}))^2 S^2 p_i} \\
&\leq \sum_{\mathbf{x_i} \in \mathbf{X}} \frac{(f_{\mathbf{q}}(\mathbf{x_i}))^2 \sum_{j=1}^{n} f_{\mathbf{q}}(\mathbf{x_j})}{(\sum_{j=1}^{n} f_{\mathbf{q}}(\mathbf{x_j}))^2 f_{\mathbf{q}}(\mathbf{x_i}) S} \\
&= 1/S
\end{aligned}$$

We get the third equation by replacing values of p_i and s_i.

Now $\mathrm{var}(g_{\mathbf{q}}(\mathbf{x_i})) \leq \mathbb{E}[(g_{\mathbf{q}}(\mathbf{x_i}))^2] \leq 1/S$. So $\mathrm{var}(T) \leq r/S$.
Now applying Bernstein Inequality as given in [2] we get,

$$\begin{aligned}\Pr(|T - \mathbb{E}[T]| \geq r\epsilon') &\leq \exp(-\frac{r^2\epsilon'^2}{r/S + r\epsilon'/3}) \\ \Pr\left(\Big|\frac{\sum_{\tilde{\mathbf{x_i}}\in\mathbf{C}} f_{\mathbf{q}}(\tilde{\mathbf{x_i}})}{S\sum_{\mathbf{x_i}\in\mathbf{X}} f_{\mathbf{q}}(\mathbf{x_i})} - \frac{1}{S}\Big| \geq \epsilon'\right) &\leq \exp\left(-\frac{r\epsilon'^2}{(1/S)+(\epsilon'/3)}\right)\end{aligned}$$

Replacing ϵ' with ϵ/S we get,

$$\Pr\left(\Big|\sum_{\tilde{\mathbf{x_i}}\in\mathbf{C}} f_{\mathbf{q}}(\tilde{\mathbf{x_i}}) - \sum_{\mathbf{x_i}\in\mathbf{X}} f_{\mathbf{q}}(\mathbf{x_i})\Big| \geq \epsilon\sum_{\mathbf{x_i}\in\mathbf{X}} f_{\mathbf{q}}(\mathbf{x_i})\right) \leq 2\exp\left(\frac{-2r\epsilon^2}{S(1+\frac{\epsilon}{3})}\right)$$

To make the above probability less than δ, we choose $r \geq \frac{S}{2\epsilon^2}(1+\frac{\epsilon}{3})\log\frac{2}{\delta}$ which depends on S for a fixed query $\mathbf{q} \in Q$. Now to bound the number of samples required to give a uniform bound for all queries simultaneously $\forall \mathbf{q} \in Q$, we use the same ϵ-net argument as described in [1]. This part is essentially a repeat of their argument. However we present it here for completeness. Observe that function $g_{\mathbf{q}}(\mathbf{x_i})$ lies in the interval $[0,1]$. Due to the bounded dimension d of Q, the queries in Q span a subspace of $[0,1]^d$. There may be an infinite number of queries in Q. However these may be covered up to L_1 distance $\epsilon/2$ by some set $Q^* \subset Q$ of $O(\epsilon^{-d})$ points [3] as given in [1]. For the ϵ-net argument let $\mathcal{E}$ be the bad event that the coreset property is not satisfied by some $\mathbf{C}$. Therefore

$$\begin{aligned}\Pr(\mathcal{E}) &= \Pr\Big[\exists\mathbf{q}\in Q : \Big|\sum_{\tilde{\mathbf{x_i}}\in\mathbf{C}} f_{\mathbf{q}}(\tilde{\mathbf{x_i}}) - \sum_{\mathbf{x_i}\in\mathbf{X}} f_{\mathbf{q}}(\mathbf{x_i})\Big| > \epsilon\sum_{\mathbf{x_i}\in\mathbf{X}} f_{\mathbf{q}}(\mathbf{x_i})\Big] \\ &\leq \Pr\Big[\exists\mathbf{q}\in Q^* : \Big|\sum_{\tilde{\mathbf{x_i}}\in\mathbf{C}} f_{\mathbf{q}}(\tilde{\mathbf{x_i}}) - \sum_{\mathbf{x_i}\in\mathbf{X}} f_{\mathbf{q}}(\mathbf{x_i})\Big| > \frac{\epsilon}{2}\sum_{\mathbf{x_i}\in\mathbf{X}} f_{\mathbf{q}}(\mathbf{x_i})\Big] \\ &\leq 2|Q^*|\exp\left(\frac{-2r\epsilon^2}{S(1+\frac{\epsilon}{3})}\right)\end{aligned}$$

To make $\mathbf{C}$ an ϵ-coreset with probability at least 1-δ, we choose $r = O(\frac{S}{\epsilon^2}(\log|Q^*| + \log\frac{2}{\delta})$. Now as $|Q^*| \in O(\epsilon^{-d})$ we have $r = O\big(\frac{S}{\epsilon^2}(d\log\frac{1}{\epsilon} + \log\frac{1}{\delta})\big)$.

□

Generalizing the Proof of Corollary 4.1.1

The proof can be generalized to the setting when $\mathbf{b}$ is in the column-space of $\mathbf{A}$ in the following manner. Suppose $\mathbf{b} = \mathbf{Au}$. Also suppose $\mathbf{A_c}$ and $\mathbf{b_c}$ can be obtained as $\mathbf{A_c} = \mathbf{SA}$ and $\mathbf{b_c} = \mathbf{Sb}$ where $\mathbf{S}$ can be either a sampling and reweighing or a sketching matrix. Now we want to prove the following : If $\mathbf{S}$ is a coreset creation matrix for $(\mathbf{A}, \mathbf{b})$ for regression i.e. $\forall\mathbf{x}, \|\mathbf{A_c x} - \mathbf{b_c}\|_p^r \in (1\pm\epsilon)\|\mathbf{Ax} - \mathbf{b}\|_p^r$, then it must be that $\forall\mathbf{x}, \|\mathbf{A_c x}\|_p^r \in (1\pm\epsilon)\|\mathbf{Ax}\|_p^r$. Proving this statement and using

Theorem 4.1 essentially proves the corollary for the more general setting of $\mathbf{b}$ in column space of $\mathbf{A}$. To prove the statement we use contradiction. Let us suppose that the statement is false. Then $\exists \mathbf{v} \in \mathbb{R}^d$ s.t. $\|\mathbf{A_c v}\|_p^r > (1+\epsilon)\|\mathbf{Av}\|_p^r$. We will create a $\mathbf{y}$ s.t that $\|\mathbf{A_c y} - \mathbf{b_c}\|_p^r > (1+\epsilon)\|\mathbf{Ay} - \mathbf{b}\|_p^r$. Consider the ratio

$$\frac{\|\mathbf{S}(\mathbf{Ay}-\mathbf{b})\|_p^r}{\|\mathbf{Ay}-\mathbf{b}\|_p^r} = \frac{\|\mathbf{SA}(\mathbf{y}-\mathbf{u})\|_p^r}{\|\mathbf{A}(\mathbf{y}-\mathbf{u})\|_p^r}$$

Now if we choose $\mathbf{y} = \mathbf{u} + \mathbf{v}$ then we have $\frac{\|\mathbf{SAv}\|_p^r}{\|\mathbf{Av}\|_p^r} > (1+\epsilon)$. This a contradiction to the fact that $(\mathbf{SA}, \mathbf{Sb})$ is coreset for $\|\mathbf{Ay} - \mathbf{b}\|_p^r$. Hence our assumption is false. So $\forall \mathbf{x}, \|\mathbf{A_c x}\|_p^r \leq (1+\epsilon)\|\mathbf{Ax}\|_p^r$. The other direction for coreset definition is proved in similar manner. This combined with the Theorem 4.1 gives our corollary

Proof of Corollary 6.1.1

Proof. For $\hat{\mathbf{A}} = [\mathbf{A} \quad -\mathbf{B}]$ and $\hat{\mathbf{X}} = \begin{bmatrix} \mathbf{X} \\ \mathbf{I_k} \end{bmatrix}$ where $\mathbf{I_k}$ is k-dimensional identity matrix, the sensitivity of Multiresponse RLAD problem is given as

$$s_i = \sup_{\hat{\mathbf{X}}} \frac{\|\hat{\mathbf{a}}_i^T \hat{\mathbf{X}}\|_1 + \frac{\lambda \|\hat{\mathbf{X}}\|_1}{n}}{\sum_j \|\hat{\mathbf{a}}_j^T \hat{\mathbf{X}}\|_1 + \lambda \|\hat{\mathbf{X}}\|_1}$$

Let $\hat{\mathbf{A}} = \mathbf{UY}$ where $\mathbf{U}$ is an $(\alpha, \beta, 1)$ well conditioned basis for $\hat{\mathbf{A}}$. So $\hat{\mathbf{a}}_j^T \hat{\mathbf{X}} = \mathbf{u_j}^T \mathbf{Y}\hat{\mathbf{X}}$. Let $\mathbf{Y}\hat{\mathbf{X}} = \mathbf{Z}$. So the sensitivity equation becomes

$$\begin{aligned} s_i \quad &= \sup_{\mathbf{Z}} \frac{\|\mathbf{u}_i^T \mathbf{Z}\|_1 + \frac{\lambda \|\mathbf{Y}^{-1}\mathbf{Z}\|_1}{n}}{\sum_j \|\mathbf{u}_j^T \mathbf{Z}\|_1 + \lambda \|\mathbf{Y}^{-1}\mathbf{Z}\|_1} \\ &\leq \sup_{\mathbf{Z}} \frac{\|\mathbf{u}_i^T \mathbf{Z}\|_1}{\sum_j \|\mathbf{u}_j^T \mathbf{Z}\|_1 + \lambda \|\mathbf{Y}^{-1}\mathbf{Z}\|_1} + \frac{1}{n} \end{aligned}$$

Instead of supremum of the first quantity on the right hand side, we take the infimum of its reciprocal. Lets call it m.

$$\begin{aligned} m \quad &= \inf_{\mathbf{Z}} \frac{\sum_j \|\mathbf{u}_j^T \mathbf{Z}\|_1 + \lambda \|\mathbf{Y}^{-1}\mathbf{Z}\|_1}{\|\mathbf{u}_i^T \mathbf{Z}\|_1} \\ &\geq \inf_{\mathbf{Z}} \frac{\sum_j \|\mathbf{u}_j^T \mathbf{Z}\|_1}{\|\mathbf{u}_i^T \mathbf{Z}\|_1} + \inf_{\mathbf{Z}} \frac{\lambda \|\mathbf{Y}^{-1}\mathbf{Z}\|_1}{\|\mathbf{u}_i^T \mathbf{Z}\|_1} \end{aligned}$$

Let us consider the first part. $\mathbf{U}$ is an $(\alpha, \beta, 1)$- well conditioned basis for $\hat{\mathbf{A}}$. Hence by definition $\|\mathbf{U}\|_1 \leq \alpha$ and $\forall \mathbf{z} \in \mathbb{R}^{d+k}, \|\mathbf{z}\|_\infty \leq \beta \|\mathbf{Uz}\|_1$. So the first term in the infimum

$$\begin{aligned}
&\inf_{\mathbf{Z}} \frac{\sum_j \|\mathbf{u}_j^T \mathbf{Z}\|_1}{\|\mathbf{u}_i^T \mathbf{Z}\|_1} \\
&= \inf_{\mathbf{Z}} \frac{\sum_{l=1}^k \|\mathbf{U}\mathbf{z}^l\|_1}{\sum_{l=1}^k |\mathbf{u_i}^T \mathbf{z}^l|} \\
&\geq \frac{\frac{1}{\beta}\sum_{l=1}^k \|\mathbf{z}^l\|_\infty}{\|\mathbf{u_i}\|_1 \sum_{l=1}^k \|\mathbf{z}^l\|_\infty} \\
&= \frac{1}{\beta\|\mathbf{u_i}\|_1}
\end{aligned}$$

Now for the second term in the infimum let us consider instead

$$\begin{aligned}
&\inf_{\mathbf{Z}} \frac{\|\mathbf{A}\mathbf{Y^{-1}}\mathbf{Z}\|_1}{\|\mathbf{u}_i^T \mathbf{Z}\|_1} \\
&= \inf_{\mathbf{Z}} \frac{\|\mathbf{U}\mathbf{Z}\|_1}{\|\mathbf{u}_i^T \mathbf{Z}\|_1} \\
&\geq \frac{1}{\beta\|\mathbf{u_i}\|_1}
\end{aligned}$$

Now $\|\mathbf{A}\mathbf{Y^{-1}}\mathbf{Z}\|_1 \leq \|\mathbf{A}\|_{(1)}\|\mathbf{Y^{-1}}\mathbf{Z}\|_1$. Therefore

$$\begin{aligned}
&\inf_{\mathbf{Z}} \frac{\|\mathbf{Y^{-1}}\mathbf{Z}\|_1}{\|\mathbf{u}_i^T \mathbf{Z}\|_1} \\
&\geq \inf_{\mathbf{Z}} \frac{\|\mathbf{A}\mathbf{Y^{-1}}\mathbf{Z}\|_1}{\|\mathbf{A}\|_{(1)}\|\mathbf{u}_i^T \mathbf{Z}\|_1} \\
&\geq \frac{1}{\beta\|\mathbf{A}\|_{(1)}\|\mathbf{u_i}\|_1}
\end{aligned}$$

Combining both these

$$m \geq \frac{1}{\beta\|\mathbf{u_i}\|_1}\left(1 + \frac{\lambda}{\|\mathbf{A}\|_{(1)}}\right)$$

Now sensitivity of i^{th} point is bounded as $s_i \leq \frac{1}{m} + \frac{1}{n}$. Therefore $s_i \leq \frac{\beta\|\mathbf{u}_i\|_1}{1+\frac{\lambda}{\|\mathbf{A}\|_{(1)}}} + \frac{1}{n}$. So the sum of sensitivities is bounded by $S \leq \frac{\alpha\beta}{1+\frac{\lambda}{\|\mathbf{A}\|_{(1)}}} + 1$. This fact combined with fact that dimension of $\mathbf{X}$ is dk and applying theorem 3.2 proves the corollary □

References

[1] Olivier Bachem, Mario Lucic, and Andreas Krause. Practical coreset constructions for machine learning. *arXiv preprint arXiv:1703.06476*, 2017.

[2] Devdatt P Dubhashi and Alessandro Panconesi. *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, 2009.

[3] David Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *J. Comb. Theory, Ser. A*, 69(2):217–232, 1995.