On Coresets For Regularized Regression

ICML 2020

Rachit Chhaya, Anirban Dasgupta and Supratim Shit

IIT Gandhinagar

June 15, 2020

Motivation

- Coresets : Small summary of data for some cost function as proxy for original data

Motivation

- Coresets : Small summary of data for some cost function as proxy for original data
- Coresets for ridge regression (smaller) shown by [ACW17].

Motivation

- Coresets : Small summary of data for some cost function as proxy for original data
- Coresets for ridge regression (smaller) shown by [ACW17].
- No study of coresets for regularized regression for general p-norm.

Motivation

- Coresets : Small summary of data for some cost function as proxy for original data
- Coresets for ridge regression (smaller) shown by [ACW17].
- No study of coresets for regularized regression for general p-norm.

Our Contributions

- No coreset for $\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{A}\mathbf{x}-\mathbf{b}\|_p^r + \lambda\|\mathbf{x}\|_q^s$, where $r \neq s$ smaller in size than that for $\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{A}\mathbf{x}-\mathbf{b}\|_p^r$

Our Contributions

- No coreset for $\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{A}\mathbf{x}-\mathbf{b}\|_p^r + \lambda\|\mathbf{x}\|_q^s$, where $r \neq s$ smaller in size than that for $\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{A}\mathbf{x}-\mathbf{b}\|_p^r$
  - Implies no coreset for Lasso smaller in size than that of least squares regression
- Introducing modified lasso and building smaller coreset for it.

Our Contributions

- No coreset for $\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{A}\mathbf{x}-\mathbf{b}\|_p^r + \lambda\|\mathbf{x}\|_q^s$, where $r \neq s$ smaller in size than that for $\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{A}\mathbf{x}-\mathbf{b}\|_p^r$
  - Implies no coreset for Lasso smaller in size than that of least squares regression
- Introducing modified lasso and building smaller coreset for it.
- Coresets for ℓ_p-regression with ℓ_p regularization. Extension to multiple response regression

Our Contributions

- No coreset for $\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{A}\mathbf{x}-\mathbf{b}\|_p^r + \lambda\|\mathbf{x}\|_q^s$, where $r \neq s$ smaller in size than that for $\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{A}\mathbf{x}-\mathbf{b}\|_p^r$
  - Implies no coreset for Lasso smaller in size than that of least squares regression
- Introducing modified lasso and building smaller coreset for it.
- Coresets for ℓ_p-regression with ℓ_p regularization. Extension to multiple response regression
- Empirical Evaluations

Coresets

Definition

For $\epsilon > 0$, a dataset $\mathbf{A}$, a non-negative function f and a query space $\mathbf{Q}$, $\mathbf{C}$ is an ϵ-coreset of $\mathbf{A}$ if $\forall q \in \mathbf{Q}$

$$\left| f_{\mathbf{q}}(\mathbf{A}) - f_{\mathbf{q}}(\mathbf{C}) \right| \leq \epsilon f_{\mathbf{q}}(\mathbf{A})$$

We construct coresets which are subsamples (rescaled) from the original data

Sensitivity [LS10]

Definition

The sensitivity of the i^{th} point of some dataset $\mathbf{X}$ for a function f and query space $\mathbf{Q}$ is defined as
$s_i = \sup_{\mathbf{q} \in \mathbf{Q}} \frac{f_{\mathbf{q}}(\mathbf{x_i})}{\sum_{\mathbf{x'} \in \mathbf{X}} f_{\mathbf{q}}(\mathbf{x'})}$.

Sensitivity [LS10]

Definition

The sensitivity of the i^{th} point of some dataset $\mathbf{X}$ for a function f and query space $\mathbf{Q}$ is defined as
$s_i = \sup_{\mathbf{q} \in \mathbf{Q}} \frac{f_{\mathbf{q}}(\mathbf{x_i})}{\sum_{\mathbf{x'} \in \mathbf{X}} f_{\mathbf{q}}(\mathbf{x'})}$.

- Determines highest fractional contribution of point to the cost function

Sensitivity [LS10]

Definition

The sensitivity of the i^{th} point of some dataset $\mathbf{X}$ for a function f and query space $\mathbf{Q}$ is defined as
$s_i = \sup_{\mathbf{q} \in \mathbf{Q}} \frac{f_{\mathbf{q}}(\mathbf{x_i})}{\sum_{\mathbf{x}' \in \mathbf{X}} f_{\mathbf{q}}(\mathbf{x}')}$.

- Determines highest fractional contribution of point to the cost function
- Can be used to create coresets. Coreset size is function of sum of sensitivities and dimension of query space

Sensitivity [LS10]

Definition

The sensitivity of the i^{th} point of some dataset $\mathbf{X}$ for a function f and query space $\mathbf{Q}$ is defined as
$s_i = \sup_{\mathbf{q} \in \mathbf{Q}} \frac{f_{\mathbf{q}}(\mathbf{x_i})}{\sum_{\mathbf{x}' \in \mathbf{X}} f_{\mathbf{q}}(\mathbf{x}')}$.

- Determines highest fractional contribution of point to the cost function
- Can be used to create coresets. Coreset size is function of sum of sensitivities and dimension of query space
- Upper bounds to sensitivities are enough [FL11, BFL16]

Coresets for Regularized Regression

- Regularization is important to prevent overfitting, numerical stability, induce sparsity etc.

Coresets for Regularized Regression

- Regularization is important to prevent overfitting, numerical stability, induce sparsity etc.

We are interested in the following problem : For $\lambda > 0$

$$\min_{\mathbf{x} \in \mathbb{R}^d} \|\mathbf{Ax} - \mathbf{b}\|_p^r + \lambda \|\mathbf{x}\|_q^s$$

for $p, q \geq 1$ and $r, s > 0$.

Coresets for Regularized Regression

- Regularization is important to prevent overfitting, numerical stability, induce sparsity etc.

We are interested in the following problem : For $\lambda > 0$

$$\min_{\mathbf{x} \in \mathbb{R}^d} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^r + \lambda \|\mathbf{x}\|_q^s$$

for $p, q \geq 1$ and $r, s > 0$.
A coreset for this problem is $(\tilde{\mathbf{A}}, \tilde{\mathbf{b}})$ such that $\forall \mathbf{x} \in \mathbb{R}^d$ and $\forall \lambda > 0$,

$$\|\tilde{\mathbf{A}}\mathbf{x} - \tilde{\mathbf{b}}\|_p^r + \lambda \|\mathbf{x}\|_q^s \in (1 \pm \epsilon)(\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^r + \lambda \|\mathbf{x}\|_q^s)$$

Main Question

- Coresets for unregularized regression work for regularized counterpart

Main Question

- Coresets for unregularized regression work for regularized counterpart
- [ACW17] showed coreset for ridge regression using ridge leverage scores. Coreset smaller than coresets for least squares regression

Main Question

- Coresets for unregularized regression work for regularized counterpart
- [ACW17] showed coreset for ridge regression using ridge leverage scores. Coreset smaller than coresets for least squares regression
- **Intuition :** Regularization imposes a constraint on the solution space.

Main Question

- Coresets for unregularized regression work for regularized counterpart
- [ACW17] showed coreset for ridge regression using ridge leverage scores. Coreset smaller than coresets for least squares regression
- **Intuition :** Regularization imposes a constraint on the solution space.
- Can we expect all regularized problems to have a smaller size coresets, than the unregularized version? For e.g. for Lasso

Our Main Result

Theorem

Given a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\lambda > 0$, any coreset for the problem $\|\mathbf{A}\mathbf{x}\|_p^r + \lambda \|\mathbf{x}\|_q^s$, where $r \neq s$, $p, q \geq 1$ and $r, s > 0$, is also a coreset for $\|\mathbf{A}\mathbf{x}\|_p^r$.

Our Main Result

Theorem

Given a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\lambda > 0$, any coreset for the problem $\|\mathbf{A}\mathbf{x}\|_p^r + \lambda\|\mathbf{x}\|_q^s$, where $r \neq s$, $p, q \geq 1$ and $r, s > 0$, is also a coreset for $\|\mathbf{A}\mathbf{x}\|_p^r$.

Implication: Smaller coresets for regularized problem are not obtained when $r \neq s$

Our Main Result

Theorem

Given a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\lambda > 0$, any coreset for the problem $\|\mathbf{A}\mathbf{x}\|_p^r + \lambda\|\mathbf{x}\|_q^s$, where $r \neq s$, $p, q \geq 1$ and $r, s > 0$, is also a coreset for $\|\mathbf{A}\mathbf{x}\|_p^r$.

Implication: Smaller coresets for regularized problem are not obtained when $r \neq s$

The popular Lasso problem falls under this category and hence does not have a coreset smaller than one for least square regression.

Our Main Result

Theorem

Given a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\lambda > 0$, any coreset for the problem $\|\mathbf{Ax}\|_p^r + \lambda\|\mathbf{x}\|_q^s$, where $r \neq s$, $p, q \geq 1$ and $r, s > 0$, is also a coreset for $\|\mathbf{Ax}\|_p^r$.

Implication: Smaller coresets for regularized problem are not obtained when $r \neq s$

The popular Lasso problem falls under this category and hence does not have a coreset smaller than one for least square regression.

Proof by Contradiction

Modified Lasso

$$\min_{\mathbf{x}\in\mathbf{R}^d} ||\mathbf{A}\mathbf{x} - \mathbf{b}||_2^2 + \lambda||\mathbf{x}||_1^2$$

- Constrained version same as lasso

Modified Lasso

$$\min_{\mathbf{x} \in \mathbf{R}^d} ||\mathbf{A}\mathbf{x} - \mathbf{b}||_2^2 + \lambda ||\mathbf{x}||_1^2$$

- Constrained version same as lasso
- Empirically shown to induce sparsity like lasso

Modified Lasso

$$\min_{\mathbf{x}\in\mathbf{R}^d} ||\mathbf{A}\mathbf{x} - \mathbf{b}||_2^2 + \lambda||\mathbf{x}||_1^2$$

- Constrained version same as lasso
- Empirically shown to induce sparsity like lasso
- Allows smaller coreset than least squares regression

Coreset for Modified Lasso

Theorem

Given a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$, corresponding vector $\mathbf{b} \in \mathbb{R}^n$, any coreset for the function $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p + \lambda\|\mathbf{x}\|_p^p$ is also a coreset of the function $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p + \lambda\|\mathbf{x}\|_q^p$ where $q \leq p$, $p, q \geq 1$.

Coreset for Modified Lasso

Theorem

Given a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$, corresponding vector $\mathbf{b} \in \mathbb{R}^{n}$, any coreset for the function $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p + \lambda\|\mathbf{x}\|_p^p$ is also a coreset of the function $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p + \lambda\|\mathbf{x}\|_q^p$ where $q \leq p$, $p, q \geq 1$.

- Implication: Coresets for ridge regression also work for modified lasso

Coreset for Modified Lasso

Theorem

Given a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$, corresponding vector $\mathbf{b} \in \mathbb{R}^n$, any coreset for the function $\|\mathbf{Ax} - \mathbf{b}\|_p^p + \lambda\|\mathbf{x}\|_p^p$ is also a coreset of the function $\|\mathbf{Ax} - \mathbf{b}\|_p^p + \lambda\|\mathbf{x}\|_q^p$ where $q \leq p$, $p, q \geq 1$.

- Implication: Coresets for ridge regression also work for modified lasso
- Coreset of size $O(\frac{sd_\lambda(\mathbf{A}) \log sd_\lambda(\mathbf{A})}{\epsilon^2})$ with a high probability for modified lasso
- $sd_\lambda(\mathbf{A}) = \sum_{j \in [d]} \frac{1}{1 + \frac{\lambda}{\sigma_j^2}} \leq d$

Coresets for ℓ_p Regression with ℓ_p Regularization

The ℓ_p Regression with ℓ_p Regularization is given as

$$\min_{\mathbf{x} \in \mathbb{R}^d} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p + \lambda \|\mathbf{x}\|_p^p$$

Coresets for ℓ_p regression constructed using the well conditioned basis

Well conditioned Basis [DDH+09]

A matrix $\mathbf{U}$ is called an (α, β, p) well-conditioned basis for $\mathbf{A}$ if $\|\mathbf{U}\|_p \leq \alpha$ and $\forall \mathbf{x} \in \mathbb{R}^d, \|\mathbf{x}\|_q \leq \beta \|\mathbf{U}\mathbf{x}\|_p$ where $\frac{1}{p} + \frac{1}{q} = 1$.

- Sampling using the p^{th} power of the p norm of rows of the (α, β, p) well-conditioned basis of $[\mathbf{A}, \mathbf{b}]$, we can obtain a coreset of size $\tilde{O}(\alpha\beta)^p$ with high probability for ℓ_p regression
- For ℓ_p Regression with ℓ_p Regularization we bound the sensitivities by $s_i \leq \frac{\beta^p \|\mathbf{u}_i\|_p^p}{1+\frac{\lambda}{\|\mathbf{A}'\|_{(p)}^p}} + \frac{1}{n}$

- Sampling using the p^{th} power of the p norm of rows of the (α, β, p) well-conditioned basis of $[\mathbf{A}, \mathbf{b}]$, we can obtain a coreset of size $\tilde{O}(\alpha\beta)^p$ with high probability for ℓ_p regression
- For ℓ_p Regression with ℓ_p Regularization we bound the sensitivities by $s_i \leq \frac{\beta^p \|\mathbf{u}_i\|_p^p}{1+\frac{\lambda}{\|\mathbf{A}'\|_{(p)}^p}} + \frac{1}{n}$
- Sum of sensitivities is bound by $S \leq \frac{(\alpha\beta)^p}{1+\frac{\lambda}{\|\mathbf{A}'\|_{(p)}^p}} + 1$

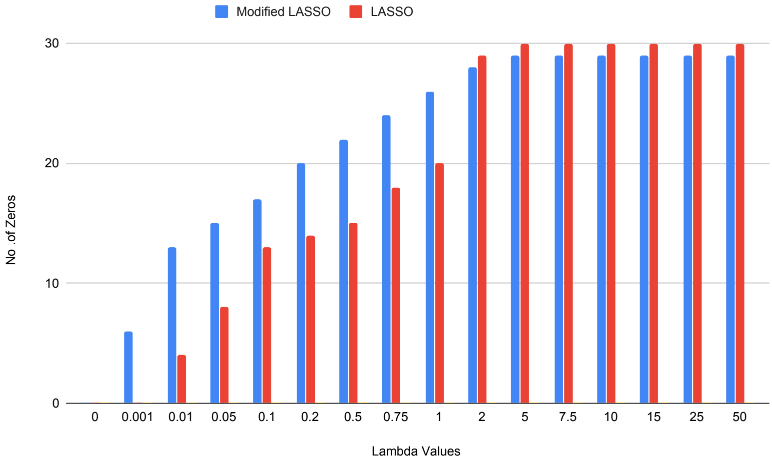
- The coreset size is $O\left(\frac{(\alpha\beta)^p d \log \frac{1}{\epsilon}}{\left(1+\frac{\lambda}{\|\mathbf{A}'\|_{(p)}^p}\right)\epsilon^2}\right)$ whp

- The coreset size is $O\left(\frac{(\alpha\beta)^p d \log \frac{1}{\epsilon}}{\left(1+\frac{\lambda}{\|\mathbf{A}'\|_{(p)}^p}\right)\epsilon^2}\right)$ whp
- Coreset size is decreasing in λ

- The coreset size is $O\left(\frac{(\alpha\beta)^p d \log\frac{1}{\epsilon}}{\left(1+\frac{\lambda}{\|\mathbf{A}'\|_{(p)}^p}\right)\epsilon^2}\right)$ whp
- Coreset size is decreasing in λ
- Specifically for Regularized Least Deviation problem we get coreset of size $O\left(\frac{d^{5/2}\log\frac{1}{\epsilon}}{\left(1+\frac{\lambda}{\|\mathbf{A}'\|_{(1)}}\right)\epsilon^2}\right)$
- Results extend to Multiresponse Regularized Regression also

Empirical Results

Sparsity Induced by Modified Lasso



Comparison with Uniform Sampling

Matrix size : 100000×30
Matrix with non uniform leverage scores [YMM15]

Table 1: Relative error of different coreset sizes for Modified Lasso, $\lambda = 0.5$

Sample Size	Ridge Leverage Scores Sampling	Uniform Sampling
30	**0.059**	0.8289
50	**0.044**	0.8289
100	**0.031**	0.8286
150	**0.028**	0.8286
200	**0.013**	0.8287

Table 2: Relative error of different coreset sizes for RLAD, $\lambda = 0.5$

Sample Size	Sensitivity based Sampling	Uniform Sampling
30	**0.69**	385.99
50	**0.65**	112.70
100	**0.34**	98.53
150	**0.19**	96.09
200	**0.17**	27.49

Conclusion and Future Work

- We present first work on coresets for regularized regression for general p norm.

Open Questions

- Tighter bounds on sensitivity scores
- Coresets for other models with regularization and/or constraints.

References I

- Haim Avron, Kenneth L Clarkson, and David P Woodruff, *Sharper bounds for regularized data fitting*, Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2017), Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.
- Vladimir Braverman, Dan Feldman, and Harry Lang, *New frameworks for offline and streaming coreset constructions*, arXiv preprint arXiv:1612.00889 (2016).
- Anirban Dasgupta, Petros Drineas, Boulos Harb, Ravi Kumar, and Michael W Mahoney, *Sampling algorithms and coresets for ℓ_p regression*, SIAM Journal on Computing **38** (2009), no. 5, 2060–2078.

References II

Petros Drineas, Michael W Mahoney, and Shan Muthukrishnan, *Sampling algorithms for l 2 regression and applications*, Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm, Society for Industrial and Applied Mathematics, 2006, pp. 1127–1136.

Dan Feldman and Michael Langberg, *A unified framework for approximating and clustering data*, Proceedings of the forty-third annual ACM symposium on Theory of computing, ACM, 2011, pp. 569–578.

David Haussler, *Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension*, Journal of Combinatorial Theory, Series A **69** (1995), no. 2, 217–232.

References III

Michael Langberg and Leonard J Schulman, *Universal ε-approximators for integrals*, Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms, SIAM, 2010, pp. 598–607.

Mert Pilanci and Martin J Wainwright, *Randomized sketches of convex programs with sharp guarantees*, IEEE Transactions on Information Theory **61** (2015), no. 9, 5096–5115.

Jiyan Yang, Xiangrui Meng, and Michael W Mahoney, *Implementing randomized matrix algorithms in parallel and distributed environments*, Proceedings of the IEEE **104** (2015), no. 1, 58–92.

More references in the paper....

Thank You

Hope to get your feedback and answer your questions at the live chat session

Take Care