Efficient Kernel Learning in the Online and Sliding Window Models: Accuracy Guarantees and Regret Bounds

UMassAmherst
College of Information & Computer Sciences

Archan Ray and Cameron Musco
{ray, cmusco}@cs.umass.edu
College of Information and Computer Sciences, UMASS Amherst

Contributions

- We study sub-linear time online Nyström approximation using ridge leverage scores (RLS) [1] and prove sample size bounds in terms of effective dimension [2].
- We extend these ideas to introduce a online sliding window kernel approximation algorithm and also provide sample size bound in terms of effective dimension.
- We further provide regret bound for exact kernel ridge regression using the approximation algorithm for both unbounded and sliding window scenarios.

Introduction

Online kernels are of particular interest to the community for their short training time when new data arrive leading to very efficient and highly scalable methods. These methods are extremely useful for applications where data arrives sequentially and evolve dynamically. The applications include but not limited to online spam detection and advertising. Algorithms for approximating such kernels in context of ridge regression have been studied in the literature most notably in [3]. General data protection rules necessitates building models based on only a snapshot of the available data has become extremely important. More importantly we want our approximation algorithms to be robust to old data deletion and preferably would not want to recompute approximation using *unimportant* data. Measures like RLS can help us identify important samples, we need a way to use them in the sliding window framework. The effective dimension characterizes the degrees of freedom of the KRR problem, and is equivalent to the sum of RLS. Therefore ideally we would want to express approximation guarantees of the algorithms in terms of effective dimension.

Setup

- **Online ridge leverage score**: $l_i^\lambda(\mathbf{K}_i) \stackrel{\text{def}}{=} (\mathbf{K}_i(\mathbf{K}_i+\lambda\mathbf{I})^{-1})_{ii}$, for any $\lambda>0$, $\mathbf{K}_i=\mathbf{B}_i\mathbf{B}_i^T$ and $\mathbf{B}_i\in\mathbb{R}^{i\times n}$.
- **Approximate online ridge leverage score**: $\tilde{l}_i^\lambda = \frac{3}{2\lambda}\mathbf{K}_{ii} - \frac{3}{2\lambda}(\mathbf{K}_i\mathbf{S}_i(\mathbf{S}_i^T\mathbf{K}_i\mathbf{S}_i+\lambda\mathbf{I})^{-1}\mathbf{S}_i^T\mathbf{K}_i)_{ii}$, which is equivalent to $\frac{3}{2}\mathbf{b}_i^T(\mathbf{B}_i^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{B}_i+\lambda\mathbf{I})^{-1}\mathbf{b}_i$, where $\mathbf{S}_i$ is the sampling matrix.
- **Reverse online ridge leverage scores**: $\zeta_{ji} \stackrel{\text{def}}{=} \min\{\mathbf{b}_j^T(\mathbf{M}_{ji}^T\mathbf{M}_{ji}+\lambda\mathbf{I})^{-1}\mathbf{b}_j, 1\}$, where $\mathbf{K}_i^\tau$ is anti-diagonal transpose of $\mathbf{K}_i$, $\mathbf{K}_i^\tau=\mathbf{M}_i\mathbf{M}_i^T$, and $\mathbf{M}_{ji}$ consists of first j rows of $\mathbf{M}_i$.
- **Approximate reverse online ridge leverage scores**: $\tilde{\zeta}_{ji} = \min(\frac{3}{2\lambda}(\mathbf{K}_i^\tau - (\mathbf{K}_i^\tau\mathbf{S}_i(\mathbf{S}_i^T\mathbf{K}_i^\tau\mathbf{S}_i+\lambda\mathbf{I})^{-1}\mathbf{S}_i^T\mathbf{K}_i^\tau))_{jj}, 1)$, where $\mathbf{S}_i$ is the sampling matrix.
- **Effective dimension**: $d_{\text{eff}}^i \stackrel{\text{def}}{=} \text{tr}(\mathbf{K}_i(\mathbf{K}_i+\lambda\mathbf{I})^{-1})$. It is essentially the optimal rank required to achieve an approximation $\tilde{\mathbf{K}}_i$ with $\|\mathbf{K}_i-\tilde{\mathbf{K}}_i\|_2\le\lambda$.

Results

- **Spectral approximation guarantee (unbounded setting)**. If we sample data points using probability equal to $\min(c\tilde{l}_i^\lambda, 1)$, it holds with probability $1-\frac{1}{n^2}$ that for all $i=1,2,\ldots,n$,

$$\frac{1}{2}(\mathbf{B}_i^T\mathbf{B}_i+\lambda\mathbf{I}) \preceq \mathbf{B}_i^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{B}_i+\lambda\mathbf{I} \preceq \frac{3}{2}(\mathbf{B}_i^T\mathbf{B}_i+\lambda\mathbf{I}).$$

- **Sample size bound (unbounded setting)**. Let s be the number of non-zero entries in $\mathbf{S}_n$, with probability $\ge 1-2/n^2$,

$$s=\mathcal{O}\left(d_{\text{eff}}^\lambda\cdot\log(n)\log\left(\frac{\|\mathbf{K}\|}{\lambda}+1\right)\right),$$

where d_{eff}^λ is the effective dimension of the final kernel matrix $\mathbf{K}\in\mathbb{R}^{n\times n}$.

Our Results contd...

Algorithm 1 ONLINE SLIDING WINDOW

1: If oldest data older than window length W, delete it from the landmark sample set.
2: Add newest data to the list of landmark points.
3: Compute the reverse leverage score of other points in the landmark sample set.
4: Choose to keep samples with probability $\min(c\tilde{\zeta}_{ji}, 1)$.

- **Sample size bound (windowed setting)**. Let s be the number of samples chosen using Algorithm 1, then with probability $1-\frac{1}{W^2}$,

$$s=\Theta(d_{\text{eff}}^\lambda\cdot\log(W)\log\left(\frac{\|\mathbf{K}\|}{\lambda}+1\right)).$$

- **Regret bound for KRR**. Let w^* optimize the oracle algorithm, the regret bound is given by,

$$\mathcal{R}_n\le\lambda\|w^*\|_2^2+d_{\text{eff}}^\lambda,$$

where d_{eff}^λ is the effective dimension of $\mathbf{K}$ in case of the unbounded window and $\mathbf{K}_n$ in case of the sliding window.

Future Work

- Tighter bound for online sliding window regret for ridge regression.
- Application areas for sliding window algorithm.

References

[1] Cohen, M. B., C. Musco, J. Pachocki. Online row sampling. *arXiv preprint arXiv:1604.05448*, 2016.

[2] Musco, C., C. Musco. Recursive sampling for the nyström method. *arXiv preprint arXiv:1605.07583*, 2016.

[3] Calandriello, D., A. Lazaric, M. Valko. Efficient second-order online kernel learning with adaptive embedding. In *Neural Information Processing Systems*. 2017.