Online Inference in High-Dimensional Models

Luo Yuanhang Supervisors: Han Ruijian, Huang Jian

Departament of Applied Mathematics, The Hong Kong Polytechnic University.

DEPARTMENT OF APPLIED MATHEMATICS
應 用 數 學 系

Objective

In this research, we aim to develop online statistical inference approach in high-dimensional generalized linear models that sequentially updates both coefficient and variance estimates upon the arrival of a new data point, without retrieving historical raw data of space complexity. We propose an online debiasing procedure with space complexity $\mathcal{O}(p)$. This algorithm is built on top of the regularization annealed epoch dual averaging (RADAR) proposed by Agarwal et al. (2012).

Introduction

A sequence of n data points $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ are independently and identically sampled from a generalized linear model:

$$\mathbb{P}(y \mid \mathbf{x}; \beta^*) \propto \exp\left\{\frac{y(\mathbf{x}^\top\beta^*) - \Phi(\mathbf{x}^\top\beta^*)}{c(\sigma)}\right\}$$

- $\mathcal{D}_i = (\mathbf{x}_i, y_i), i \in [n]$.
- $\mathbf{x}_i \in \mathbb{R}^p$. High-dimensional case: $n \ll p$.
- Sparsity: $\|\beta^*\|_0 = s_0$.

Offline approach:

$$\beta^{(n)} = \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \left\{-y_i(\mathbf{x}_i^\top\beta) + \Phi(\mathbf{x}_i^\top\beta)\right\} + \lambda_n\|\beta\|_1,$$

However, retrieving historical raw data of space complexity $\mathcal{O}(np)$ or saving summary statistics of space complexity $\mathcal{O}(p^2)$. We can solve the problem in an online manner.

Method

We propose the Approximated Debiased Lasso (ADL) for online estimation and inference via a variant of online stochastic gradient descent.
Let $l_n(\beta) = \frac{1}{n}\sum_{i=1}^n \left\{-y_i\left(\mathbf{x}_i^\top\beta\right) + \Phi\left(\mathbf{x}_i^\top\beta\right)\right\}$.

- Receive incoming data.
- Update estimates $\widetilde{\beta}_j^{(k)} \Leftarrow \mathbf{RADAR}(\widetilde{\beta}_j^{(k-1)}, \mathcal{D}_k)$
- De-bias via one-step Newton correction

$$\widetilde{\beta}_{j,\mathrm{de}}^{(k)} = \widetilde{\beta}_j^{(k)} - \left[\left\{\nabla^2 l_k\left(\widetilde{\beta}^{(k)}\right)\right\}^{-1}\right]_j \nabla l_k\left(\widetilde{\beta}^{(k)}\right),$$

  where
  - $[\{\nabla^2 l_n(\widetilde{\beta}^{(n)})\}^{-1}]_j \Leftarrow \mathbf{RADAR}(\widetilde{\gamma}_{\cdot,j}^{(n)}, \mathcal{D}_n)$, the node-wise Lasso.
  - $\nabla l_n(\widetilde{\beta}^{(n)}) \Leftarrow$ Taylor's expansion.
  - Obtain some statistics $s^{(n)}$ and $\widetilde{\tau}_j^{(n)}$
- The $(1-\alpha)$-confidence interval for β_j^* with $0 < \alpha < 1$ at time k is given as:

$$\left(\widetilde{\beta}_{j,\mathrm{de}}^{(k)} - z_{\alpha/2}\widetilde{\tau}_j^{(k)}, \widetilde{\beta}_{j,\mathrm{de}}^{(k)} + z_{\alpha/2}\widetilde{\tau}_j^{(k)}\right), k \in [n].$$

- Recurse the variables $(\widetilde{\beta}_j^{(k)}, \widetilde{\gamma}_{\cdot,j}^{(k)}, s^{(k)}, \widetilde{\tau}_j^{(k)})$ and clear other variables in memory.

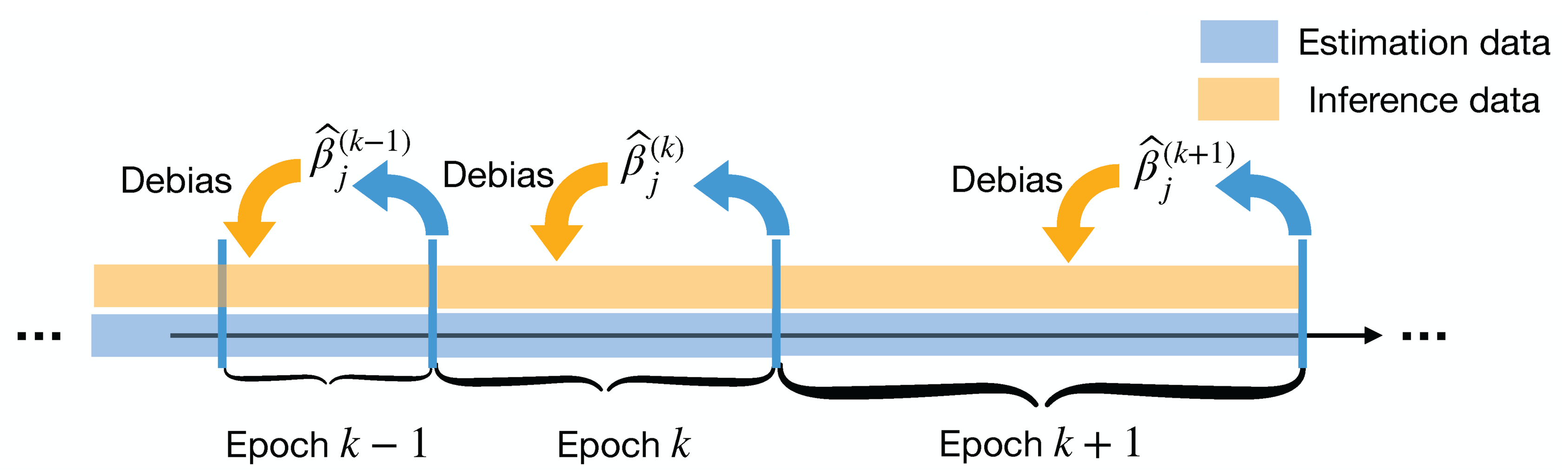

Figure 1: Approximated Debiased Lasso

For statistical inference of the j-th element in β^*, we only need to compute and store up to some p-dimensional vectors instead of some $p \times p$ matrix.

Theorem (Convergence rate)

Under some conditions, the following events

$$\|\widetilde{\beta}^{(k)} - \beta^*\|_1 \lesssim \|\widetilde{\beta}^{(0)} - \beta^*\|_1 c_1 s_0\sqrt{\frac{\log p}{k}},$$
$$\|\widetilde{\beta}^{(k)} - \beta^*\|_2^2 \lesssim \|\widetilde{\beta}^{(0)} - \beta^*\|_2^2 c_2 s_0 \frac{\log p}{k}.$$

for $k \geq n_1$ hold uniformly for universal constants n_1, c_1 and c_2 with high probability.

Theorem (Asymptotic normality)

Under some Assumptions,

$$\left(\widetilde{\beta}_{j,de}^{(n)} - \beta_j^*\right)/\widetilde{\tau}_j^{(n)} \to \mathcal{N}(0,1)$$

in distribution as $n \to \infty$.

Numerical Results

Logistic regression with synthetic data:

Table 1: $n = 200$, $p = 500$, $s_0 = 6$, $\Sigma = 0.1 \times \{0.5^{|i-j|}\}_{i,j=1,\ldots,p}$. Simulation results are summarized over 200 replications.

	β_k^*	deLasso	LSW	ODL			ADL		
Sample size n		200	200	40	120	200	40	120	200
Coverage probability	0	0.95	0.98	0.97	0.96	0.95	0.97	0.96	0.95
	1	0.93	0.92	0.96	0.95	0.94	0.97	0.95	0.95
	-1	0.95	0.94	0.97	0.95	0.95	0.97	0.97	0.95
Absolute bias	0	0.41	0.36	0.06	0.04	0.03	0.73	0.46	0.36
	1	0.43	0.53	0.25	0.21	0.18	0.80	0.47	0.38
	-1	0.41	0.52	0.23	0.20	0.17	0.78	0.45	0.37
Coverage length	0	2.03	2.26	4.06	2.29	1.77	4.24	2.47	1.88
	1	2.03	2.26	4.05	2.29	1.76	4.23	2.47	1.87
	-1	2.04	2.28	4.09	2.30	1.77	4.26	2.48	1.89
Time (s)		7.21	20.22		0.54			0.32	

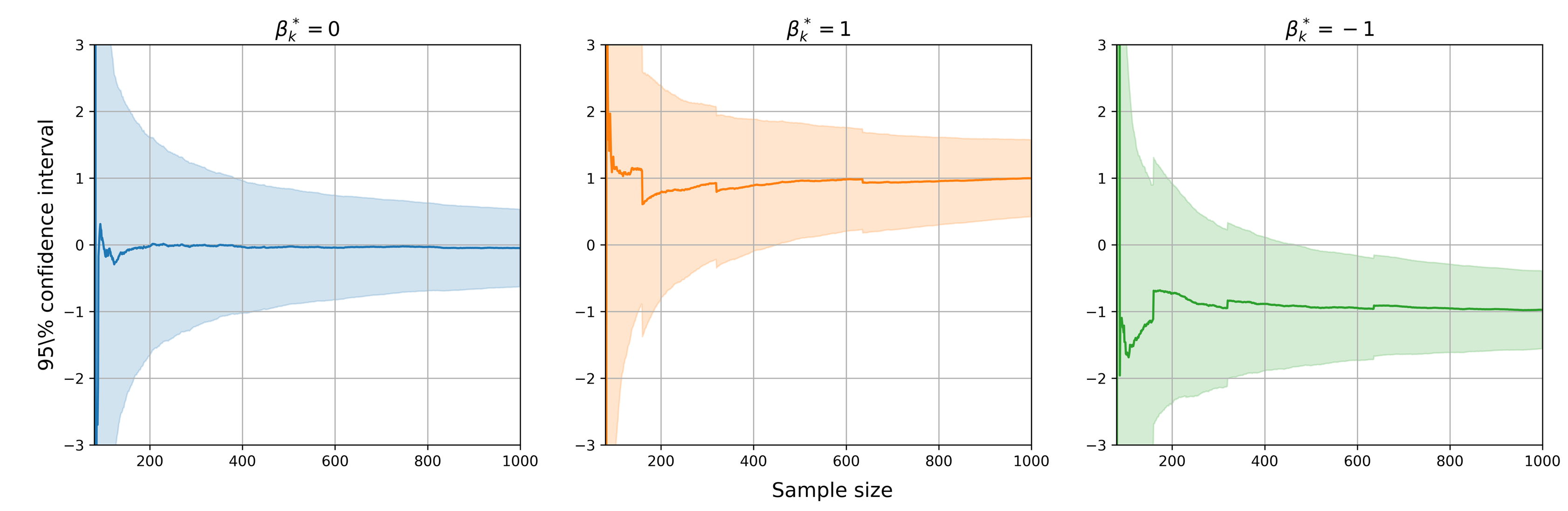

Figure 2: $n = 1000$, $p = 20000$, $s_0 = 20$, $\Sigma = \{0.5^{|i-j|}\}_{i,j=1,\ldots,p}$. Simulation results are averaged over 200 replications (around 110.4s for each)

Real data application on logistic regression:

- x: 3,231,961 features containing lexical, host-based information...
- y: Binary, is this URL a phishing site or not?

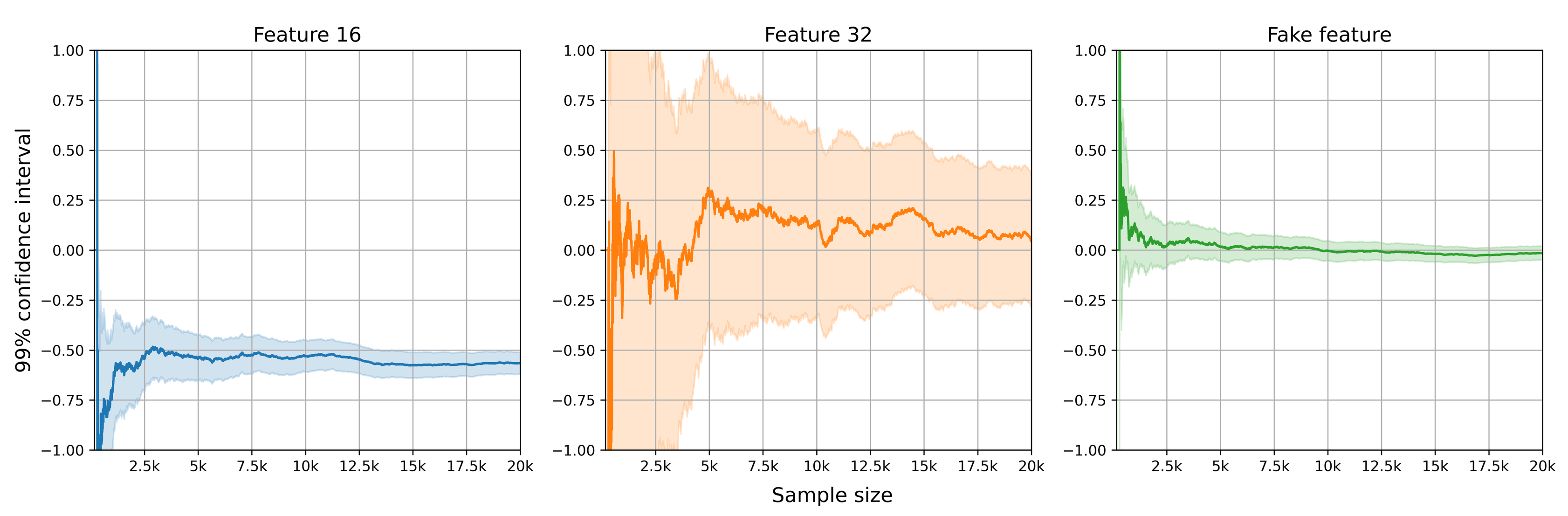

Figure 3: Trace plots of impacts of three features on the detection of malicious websites.

Reference

Agarwal, A., Negahban, S., & Wainwright, M. J. (2012). Stochastic optimization and sparse statistical recovery: Optimal algorithms for high dimensions. *Advances in Neural Information Processing Systems, 25.*

Han, R., Luo, L., Luo, Y., Lin, Y., & Huang, J. (2023). Online Inference in High-dimensional Models. *To submit.*

https://chattelionluo.github.io/ chattelion.luo@connect.polyu.hk