Adaptive Debiased SGD in High-dimensional GLMs with Streaming Data

Yuanhang Luo[1]
Joint work with Ruijian Han[2], Lan Luo[3], Yuanyuan Lin[4], Jian Huang[2]

[1]Department of Applied Mathematics, The Hong Kong Polytechnic University
[2]Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University
[3]Department of Biostatistics and Epidemiology, Rutgers University
[4]Department of Statistics, The Chinese University of Hong Kong

29th Aug 2024

Table of Contents

Table of Contents

Big Data

Big data is about

> *"datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze" [Manyika et al., 2011]*
>
> *McKinsey Global Institute*

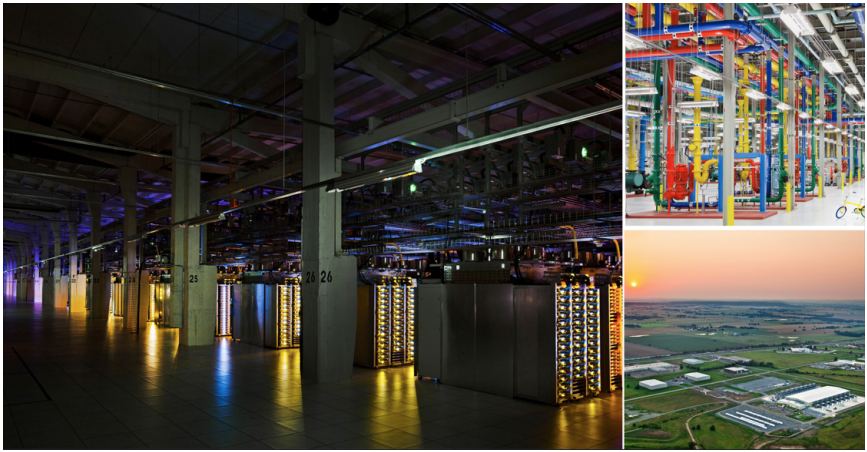


Figure: Google Data Center

Streaming Data

Many applications must process large streams of live data and provide results in near-real-time

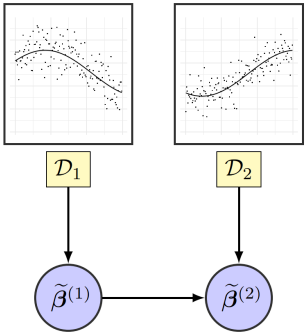
- Algorithmic trading
- Real-time operations management
- Online fraud detection
- Proximity/location tracking
- Intrusion detection systems
- Traffic management
- Real-time recommendations
- Social media/data analytics
- Precision agriculture
- Internet of things
- Gaming data feed
- ...

Some of the major challenges of handling streaming data.

- Historical data is no longer available (needs one pass process).
- Each data has large number of features (high-dimensional).
- Critical memory requirements.
- Limiting computing power.

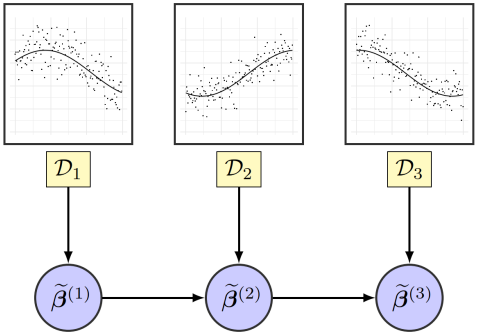
Streaming Data

Proposed solution: Online learning



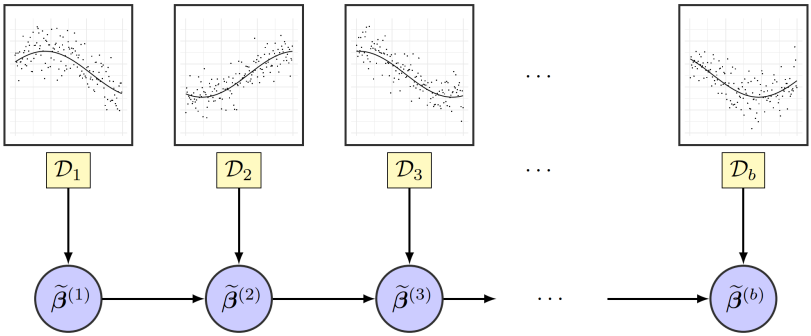
Streaming Data

Proposed solution: Online learning



Streaming Data

Proposed solution: Online learning



Objective: Develop an "online" estimator that

- can be updated sequentially using raw data in current batch only
- requires little time and space complexities
- still achieves desired statistical property

High-dimensional GLM

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

$$\mathbb{P}(y \mid \mathbf{x}; \boldsymbol{\beta}^*) \propto \exp\left\{ \frac{y\left(\mathbf{x}^\top \boldsymbol{\beta}^*\right) - \Phi\left(\mathbf{x}^\top \boldsymbol{\beta}^*\right)}{c(\sigma)} \right\}, \tag{1}$$

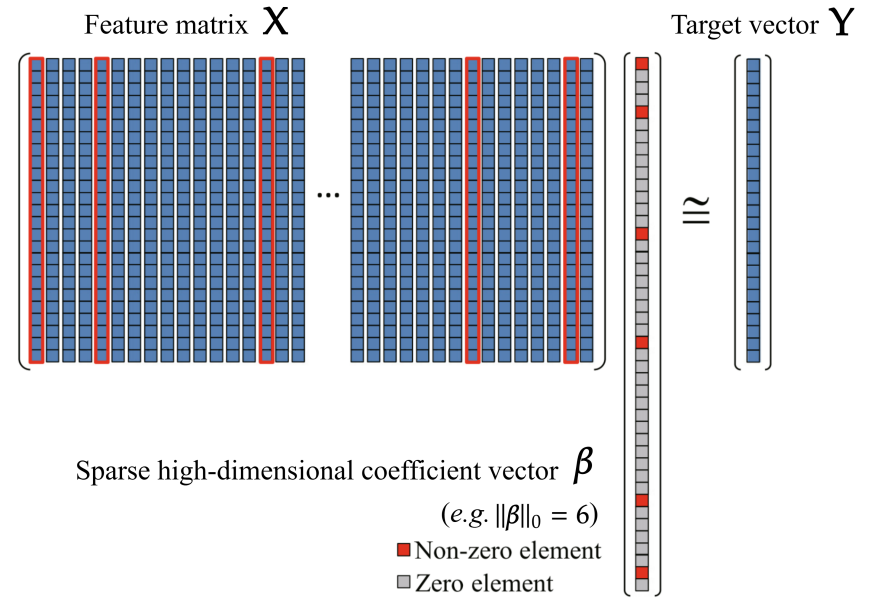
- $\mathbf{x} \in \mathbb{R}^p$. High-dimensional case: $n \ll p$.
- Sparsity: $\|\boldsymbol{\beta}^*\|_0 = s_0$.

The ground truth $\boldsymbol{\beta}^*$ is the minimum of the population loss function

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \mathbb{E}\left[F_n(\boldsymbol{\beta})\right], \tag{2}$$

where $F_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \left\{-y_i(\mathbf{x}_i^\top \boldsymbol{\beta}) + \Phi(\mathbf{x}_i^\top \boldsymbol{\beta})\right\}$, the negative log-likelihood of model $\mathbb{P}$.

Example: High-dimensional Linear Regression



Traditionally, we estimate $\boldsymbol{\beta}^*$ using data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ via two steps:

$$\widetilde{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left\{ F_n(\boldsymbol{\beta}) + \lambda_n \|\boldsymbol{\beta}\|_1 \right\}, \tag{3}$$

and

$$\widehat{\boldsymbol{\beta}}_{j,\mathrm{de}} = \widetilde{\boldsymbol{\beta}}_j - \boldsymbol{\Theta}_j^\top \nabla F_n(\widetilde{\boldsymbol{\beta}}), \tag{4}$$

where $\boldsymbol{\Theta}$ is the inverse of the covariance matrix (information matrix). $\boldsymbol{\Theta}_j$ can be estimated by [Zhang and Zhang, 2014, van de Geer et al., 2014]...

Traditionally, we estimate $\boldsymbol{\beta}^*$ using data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ via two steps:

$$\widetilde{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left\{ F_n(\boldsymbol{\beta}) + \lambda_n \|\boldsymbol{\beta}\|_1 \right\}, \tag{3}$$

and

$$\widehat{\boldsymbol{\beta}}_{j,\mathrm{de}} = \widetilde{\boldsymbol{\beta}}_j - \boldsymbol{\Theta}_j^\top \nabla F_n(\widetilde{\boldsymbol{\beta}}), \tag{4}$$

where $\boldsymbol{\Theta}$ is the inverse of the covariance matrix (information matrix). $\boldsymbol{\Theta}_j$ can be estimated by [Zhang and Zhang, 2014, van de Geer et al., 2014]...

By doing so, we have asymptotic normality

$$(\widehat{\boldsymbol{\beta}}_{j,\mathrm{de}} - \boldsymbol{\beta}_j^*)/\widehat{\sigma}_j \to \mathcal{N}(0,1) \text{ in distribution as } n \to \infty. \tag{5}$$

And we are able to make all kinds of statistical inference based on

$$\mathbb{P}\{\boldsymbol{\beta}_j^* \in (\widehat{\boldsymbol{\beta}}_{j,\mathrm{de}} - z_{\alpha/2}\widehat{\sigma}_j, \widehat{\boldsymbol{\beta}}_{j,\mathrm{de}} + z_{\alpha/2}\widehat{\sigma}_j)\} \to 1-\alpha \text{ as } n \to \infty. \tag{6}$$

To do

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

and

$$\widehat{\boldsymbol{\beta}}_{j,\mathrm{de}} = \widetilde{\boldsymbol{\beta}}_j - \boldsymbol{\Theta}_j^\top \nabla F_n(\widetilde{\boldsymbol{\beta}}),$$

Remark

we need to store data of size $\mathcal{O}(n \times p)$, and do calculations involving the covariance matrix of size $\mathcal{O}(p \times p)$. In large datasets (*e.g.* text data, web corpus, gene expression data...) p can be extensive. Let's say $p = 3,000,000$, then,

$$p \times p = 9,000,000,000,000. \tag{7}$$

To operate this $p \times p$ matrix of double-precision floating-point (float64) numbers, you would need at least approximately 65.54 terabytes (TB) of storage.

Our Solution

Question

Can we bring it "online"?

Our Solution

Question

Can we bring it "online"?

Table: A comparison in time and space complexities for statistical inference of one single coefficient. LM: linear model; GLM: generalized linear model.

Works	Model	One-pass?	Space complexity
[Chen et al., 2020]	LM	Yes	$\mathcal{O}(np)$
[Deshpande et al., 2023]	LM	No	$\mathcal{O}\left(p^2\right)$
[Han et al., 2023]	LM	Yes	$\mathcal{O}(p)$
[Shi et al., 2021]	GLM	No	$\mathcal{O}(np)$
[Luo et al., 2023]	GLM	No	$\mathcal{O}\left(p^2\right)$
Proposed Method	GLM	Yes	$\mathcal{O}(p)$

Our Solution

Question

Can we bring it "online"?

Table: A comparison in time and space complexities for statistical inference of one single coefficient. LM: linear model; GLM: generalized linear model.

Works	Model	One-pass?	Space complexity
[Chen et al., 2020]	LM	Yes	$\mathcal{O}(np)$
[Deshpande et al., 2023]	LM	No	$\mathcal{O}\left(p^2\right)$
[Han et al., 2023]	LM	Yes	$\mathcal{O}(p)$
[Shi et al., 2021]	GLM	No	$\mathcal{O}(np)$
[Luo et al., 2023]	GLM	No	$\mathcal{O}\left(p^2\right)$
Proposed Method	GLM	Yes	$\mathcal{O}(p)$

Main Idea

Optimization + Statistics

Table of Contents

Overview of the Algorithm

Upon the arrival of the k-th data, update $\widehat{\boldsymbol{\beta}}^{(i)}$ and $\widehat{\boldsymbol{\gamma}}_{\cdot,j}^{(i)}$ with a variant of SGD, then construct the following summary statistics:

$$\boldsymbol{A}_1^{(k)} = \sum_{i=1}^{k} \boldsymbol{x}_i\{-y_i + \dot{\Phi}(\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}^{(i)})\}, \quad \boldsymbol{A}_2^{(k)} = \sum_{i=1}^{k} \boldsymbol{x}_i^\top \widehat{\boldsymbol{\gamma}}_{\cdot,j}^{(i)} \ddot{\Phi}(\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}^{(i)}) \boldsymbol{x}_i,$$

$$a_3^{(k)} = \sum_{i=1}^{k} \boldsymbol{x}_i^\top \widehat{\boldsymbol{\gamma}}_{\cdot,j}^{(i)} \ddot{\Phi}(\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}^{(i)}) \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}^{(i)}, \quad a_4^{(k)} = -\sum_{i=1}^{k} \boldsymbol{x}_i^\top \widehat{\boldsymbol{\gamma}}_{\cdot,j}^{(i)} (\boldsymbol{x}_i)_j \{\ddot{\Phi}(\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}^{(i)})\}.$$

All are p-dimensional vectors and scalars. The approximated debiased lasso (ADL) estimator

$$\widetilde{\boldsymbol{\beta}}_{j,\mathrm{de}}^{(k)} := \widehat{\boldsymbol{\beta}}_j^{(k)} - \frac{\{\boldsymbol{A}_1^{(k)}\}^\top \widehat{\boldsymbol{\gamma}}_{\cdot,j}^{(k)} + \{\boldsymbol{A}_2^{(k)}\}^\top \widehat{\boldsymbol{\beta}}^{(k)} - a_3^{(k)}}{a_4^{(k)}}.$$

We update summary statistics online:

$$\boldsymbol{A}_1^{(k)} \longleftarrow \boldsymbol{A}_1^{(k-1)} + \boldsymbol{x}_k\{-y_k + \dot{\Phi}(\boldsymbol{x}_k^\top \widehat{\boldsymbol{\beta}}^{(k)})\}.$$

Overview of the Algorithm

Upon the arrival of the k-th data, update $\widehat{\boldsymbol{\beta}}^{(i)}$ and $\widehat{\boldsymbol{\gamma}}_{\cdot,j}^{(i)}$ with a variant of SGD, then construct the following summary statistics:

$$\boldsymbol{A}_1^{(k)} = \sum_{i=1}^{k} \boldsymbol{x}_i\{-y_i + \dot{\Phi}(\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}^{(i)})\}, \quad \boldsymbol{A}_2^{(k)} = \sum_{i=1}^{k} \boldsymbol{x}_i^\top \widehat{\boldsymbol{\gamma}}_{\cdot,j}^{(i)} \ddot{\Phi}(\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}^{(i)}) \boldsymbol{x}_i,$$

$$a_3^{(k)} = \sum_{i=1}^{k} \boldsymbol{x}_i^\top \widehat{\boldsymbol{\gamma}}_{\cdot,j}^{(i)} \ddot{\Phi}(\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}^{(i)}) \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}^{(i)}, \quad a_4^{(k)} = -\sum_{i=1}^{k} \boldsymbol{x}_i^\top \widehat{\boldsymbol{\gamma}}_{\cdot,j}^{(i)} (\boldsymbol{x}_i)_j \{\ddot{\Phi}(\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}^{(i)})\}.$$

All are p-dimensional vectors and scalars. The approximated debiased lasso (ADL) estimator

$$\widetilde{\boldsymbol{\beta}}_{j,\mathrm{de}}^{(k)} := \widehat{\boldsymbol{\beta}}_j^{(k)} - \frac{\{\boldsymbol{A}_1^{(k)}\}^\top \widehat{\boldsymbol{\gamma}}_{\cdot,j}^{(k)} + \{\boldsymbol{A}_2^{(k)}\}^\top \widehat{\boldsymbol{\beta}}^{(k)} - a_3^{(k)}}{a_4^{(k)}}.$$

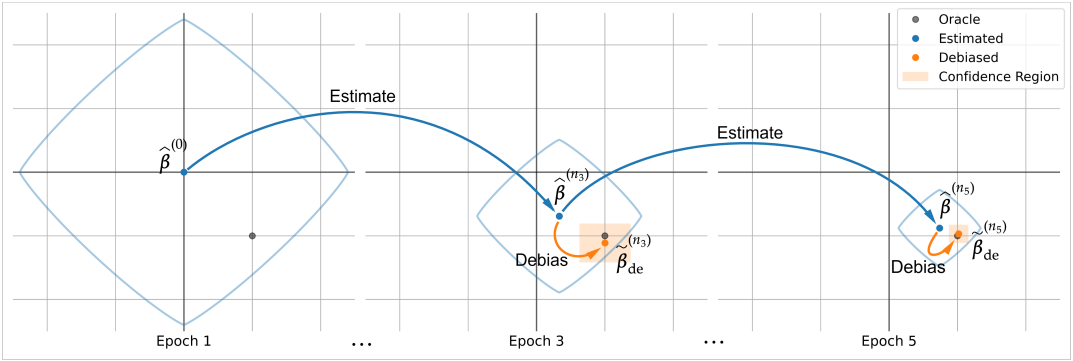
We update summary statistics online:

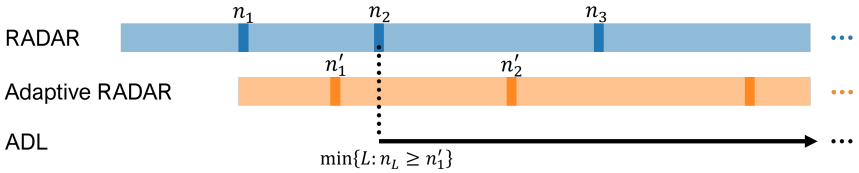
$$\boldsymbol{A}_1^{(k)} \longleftarrow \boldsymbol{A}_1^{(k-1)} + \boldsymbol{x}_k\{-y_k + \dot{\Phi}(\boldsymbol{x}_k^\top \widehat{\boldsymbol{\beta}}^{(k)})\}.$$

Remark

Now we can recurse $\{\widehat{\boldsymbol{\beta}}^{(k)}, \widehat{\boldsymbol{\gamma}}_{\cdot,j}^{(k)}, \boldsymbol{A}_1^{(k)}, \boldsymbol{A}_2^{(k)} a_3^{(k)}, a_4^{(k)}\}$ and clear all other variables in memory.

Overview of the Algorithm





Complexities

Table: A comparison in time and space complexities for statistical inference of one single coefficient. LM: linear model; GLM: generalized linear model.

Works	Model	One-pass?	Space complexity
[Chen et al., 2020]	LM	Yes	$\mathcal{O}(np)$
[Deshpande et al., 2023]	LM	No	$\mathcal{O}\left(p^2\right)$
[Han et al., 2023]	LM	Yes	$\mathcal{O}(p)$
[Shi et al., 2021]	GLM	No	$\mathcal{O}(np)$
[Luo et al., 2023]	GLM	No	$\mathcal{O}\left(p^2\right)$
Proposed Method	GLM	Yes	$\mathcal{O}(p)$

Table of Contents

Theorem (Convergence Rate)

With some regularity conditions and proper choice of hyperparameters, and initial error $d_0 = \|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|_1$ *the following events*

$$\|\widehat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*\|_1 \leq C_1 s_0 d_0 \sqrt{\frac{(\log p)^3}{i}}, \quad \|\widehat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*\|_2 \leq C_1 d_0 \sqrt{\frac{s_0 (\log p)^3}{i}} \tag{8}$$

for $i \geq n_1$ *hold uniformly for a universal constant* C_1 *with high probability.*

Theoretical Results

Theorem (Convergence Rate)

With some regularity conditions and proper choice of hyperparameters, and initial error $d_0 = \|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|_1$ *the following events*

$$\|\widehat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*\|_1 \leq C_1 s_0 d_0 \sqrt{\frac{(\log p)^3}{i}}, \quad \|\widehat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*\|_2 \leq C_1 d_0 \sqrt{\frac{s_0 (\log p)^3}{i}} \tag{8}$$

for $i \geq n_1$ *hold uniformly for a universal constant* C_1 *with high probability.*

Theorem (Asymptotic Normality)

Further with

$$\left(s_0 s_1 d_0 d_1 + s_0^2 d_0^2 + s_0 s_1 d_0^2\right) (\log p)^4 \log_2(n) = o_{\mathbb{P}}(\sqrt{n}),$$

then, for any fixed $j \in [p]$,

$$\left(\widetilde{\boldsymbol{\beta}}_{j,de}^{(n)} - \boldsymbol{\beta}_j^*\right) / \widetilde{\nu}_j^{(n)} \to \mathcal{N}(0,1) \text{ as } n \to \infty \text{ in distribution.}$$

Remark

Most existing methods rely on the bounded individual probability condition that

$$P(y_i = 1 \mid X_i) \in (c, 1-c)$$

for all $1 \leq i \leq n$ and some $c \in (0, 1/2)$. To the best of our knowledge, only [Guo et al., 2021, Cai et al., 2023] and ours remove this.

Table of Contents

Simulation with Synthetic Data (1)

True coefficients: $\beta^*_{D_1} = 1, \beta^*_{D_2} = -1$ and others are set to 0.

Table: Case 2: $n = 200$, $p = 500$, $s_0 = 6$, $\mathbf{\Sigma} = \{0.5^{|i-j|}\}_{i,j=1,\ldots,p}$. Simulation results are averaged over 200 replications. Other methods: deLasso [van de Geer et al., 2014], LSW [Cai et al., 2023], ODL [Luo et al., 2023].

	β^*_k	deLasso	LSW	ODL			ADL		
Sample size n		200	200	40	120	200	40	120	200
Coverage probability	0	0.983	0.996	1.000	0.999	0.998	0.951	0.951	0.961
	1	0.598	0.936	1.000	0.880	0.705	0.891	0.955	0.943
	-1	0.610	0.946	1.000	0.861	0.670	0.881	0.961	0.945
Absolute bias	0	0.114	0.113	0.022	0.019	0.017	0.901	0.425	0.309
	1	0.325	0.260	0.657	0.590	0.539	1.008	0.423	0.328
	-1	0.326	0.259	0.660	0.598	0.548	1.034	0.426	0.324
Coverage length	0	0.753	1.207	3.423	1.691	1.287	4.419	2.053	1.532
	1	0.769	1.261	3.434	1.689	1.286	4.555	2.086	1.536
	-1	0.767	1.271	3.416	1.648	1.282	4.431	2.078	1.541
Time (s)		6.550	20.04		0.628			0.365	

Simulation with Synthetic Data (1)

True coefficients: $\beta^*_{D_1} = 1, \beta^*_{D_2} = -1$ and others are set to 0.

Table: Case 2: $n = 200$, $p = 500$, $s_0 = 6$, $\mathbf{\Sigma} = \{0.5^{|i-j|}\}_{i,j=1,\ldots,p}$. Simulation results are averaged over 200 replications. Other methods: deLasso [van de Geer et al., 2014], LSW [Cai et al., 2023], ODL [Luo et al., 2023].

	β^*_k	deLasso	LSW	ODL			ADL		
Sample size n		200	200	40	120	200	40	120	200
Coverage probability	0	0.983	0.996	1.000	0.999	0.998	0.951	0.951	0.961
	1	0.598	0.936	1.000	0.880	0.705	0.891	0.955	0.943
	-1	0.610	0.946	1.000	0.861	0.670	0.881	0.961	0.945
Absolute bias	0	0.114	0.113	0.022	0.019	0.017	0.901	0.425	0.309
	1	0.325	0.260	0.657	0.590	0.539	1.008	0.423	0.328
	-1	0.326	0.259	0.660	0.598	0.548	1.034	0.426	0.324
Coverage length	0	0.753	1.207	3.423	1.691	1.287	4.419	2.053	1.532
	1	0.769	1.261	3.434	1.689	1.286	4.555	2.086	1.536
	-1	0.767	1.271	3.416	1.648	1.282	4.431	2.078	1.541
Time (s)		6.550	20.04		0.628			0.365	

Remark

deLasso and ODL fail because they rely on the bounded individual probability condition.

Simulation with Synthetic Data (2)

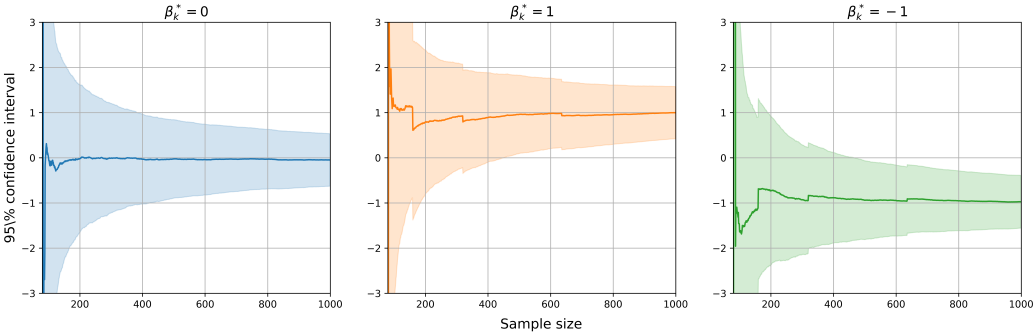


Figure: $n = 1000$, $p = 20000$, $s_0 = 20$, $\mathbf{\Sigma} = \{0.5^{|i-j|}\}_{i,j=1,\ldots,p}$. Simulation results are averaged over 200 replications (around 110.4s for each replication).

Large-scale URLs Data

Real data application: Detection of malicious websites [Ma et al., 2010].

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

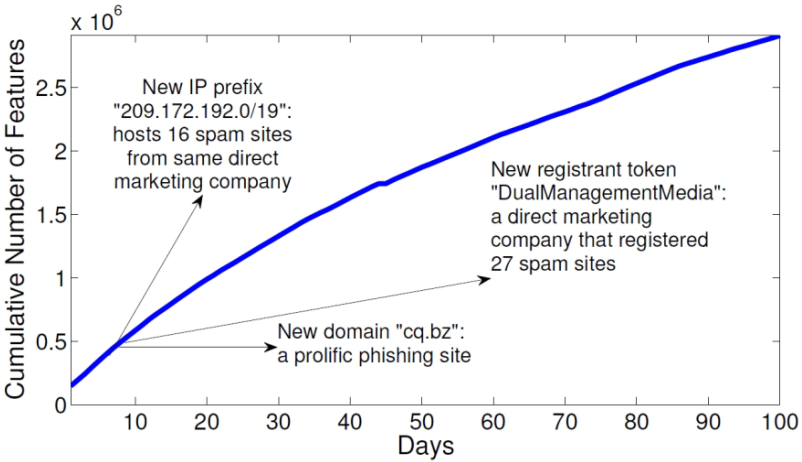
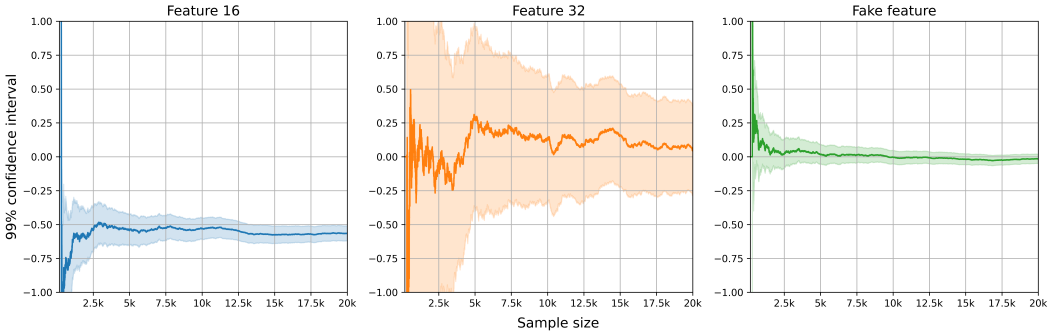


Figure: The number of features also grows with incoming data.

Large-scale URLs Data

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?



- An incoming data needs around 2s to arrive on average.
- ADL consumes around 0.3s on each data.

10-K Financial Report Data

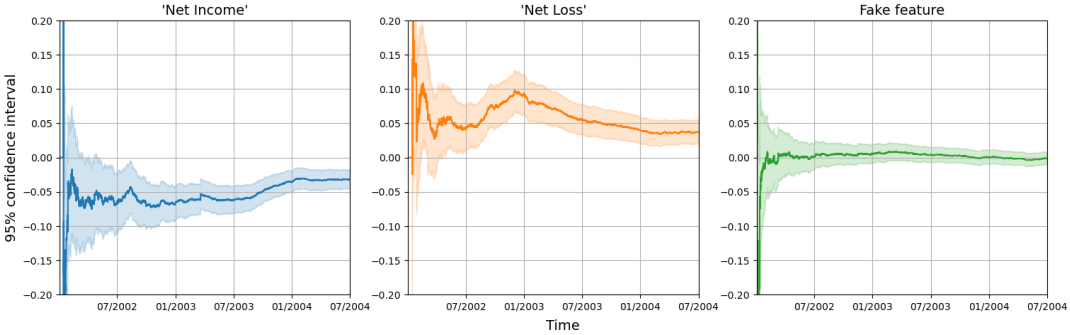
Data source: [Kogan et al., 2009]. We used more than 10,000 10-K forms from 2002 to 2004 to predict future market volatility.

- x: 4,500,000 features containing unigram and bigram frequency.
- y: Increment of market volatility (log scale).
- Linear regression

Figure: Form 10-K is an annual report required by the U.S. Securities and Exchange Commission (SEC), that gives a comprehensive summary of a company's financial performance including, organizational structure, executive compensation, equity, subsidiaries, and audited financial statements....

10K Financial Report Data

- x: 4,500,000 features containing unigram and bigram frequency.
- y: Increment of market volatility (log scale).



- ADL consumes around 0.45s on each data.

Remark (Leverage/Announcement Effect)

"Bad news" has a larger impact on market volatility than "good news".

Conclusion & Remarks

In this research, we develop online statistical inference approach in high-dimensional GLM that

- sequentially updates without retrieving historical data.
- has a space complexity $\mathcal{O}(p)$, instead of $\mathcal{O}(p^2)$ nor $\mathcal{O}(np)$.
- provides theoretical guarantees for convergence and asymptotic normality.
- achieves similar performance as offline counterparts.

Future works:

- non i.i.d. data, covariate shift
- non-convex objectives
- ...

Suggested reference:

- Han, R., Luo, L., Luo, Y., Lin, Y. and Huang, J. (2024). "Adaptive debiased SGD in high-dimensional GLMs with streaming data". Manuscript submitted for publication.

Thanks for watching!

$\mathcal{Q}\&\mathcal{A}$

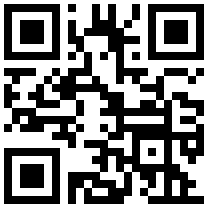


Figure: Contact info: chattelion.luo@connect.polyu.hk

Reference I


Cai, T. T., Guo, Z., and Ma, R. (2023).
Statistical inference for high-dimensional generalized linear models with binary outcomes.
Journal of the American Statistical Association, 118(542):1319–1332.


Chen, X., Lee, J. D., Tong, X. T., and Zhang, Y. (2020).
Statistical inference for model parameters in stochastic gradient descent.
The Annals of Statistics, 48(1):251 – 273.


Deshpande, Y., Javanmard, A., and Mehrabi, M. (2023).
Online debiasing for adaptively collected high-dimensional data with applications to time series analysis.
Journal of the American Statistical Association, 118(542):1126–1139.


Guo, Z., Rakshit, P., Herman, D. S., and Chen, J. (2021).
Inference for the case probability in high-dimensional logistic regression.
Journal of Machine Learning Research, 22(254):1–54.


Han, R., Luo, L., Lin, Y., and Huang, J. (2023).
Online inference with debiased stochastic gradient descent.
Biometrika, 111(1):93–108.


Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., and Smith, N. A. (2009).
Predicting risk from financial reports with regression.
USA. Association for Computational Linguistics.


Luo, L., Han, R., Lin, Y., and Huang, J. (2023).
Online inference in high-dimensional generalized linear models with streaming data.
Electronic Journal of Statistics, 17(2):3443 – 3471.


Ma, J., Kulesza, A., Dredze, M., Crammer, K., Saul, L., and Pereira, F. (2010).
Exploiting feature covariance in high-dimensional online learning.
In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 493–500.


Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Hung Byers, A. (2011).
Big data: The next frontier for innovation, competition, and productivity.


Shi, C., Song, R., Lu, W., and Li, R. (2021).
Statistical inference for high-dimensional models via recursive online-score estimation.
Journal of the American Statistical Association, 116(535):1307–1318.

Reference II

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014).
On asymptotically optimal confidence regions and tests for high-dimensional models.
The Annals of Statistics, 42(3):1166 – 1202.

Zhang, C.-H. and Zhang, S. S. (2014).
Confidence intervals for low dimensional parameters in high dimensional linear models.
Journal of the Royal Statistical Society Series B: Statistical Methodology, 76(1):217–242.