Statistical and Computational Guarantees for Influence Diagnostics

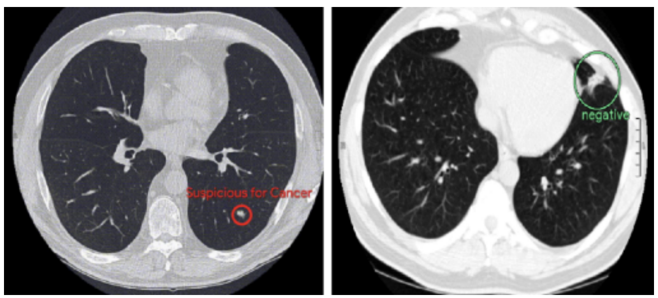
Jillian Fisher[1], Lang Liu[1], Krishna Pillutla[2], Yejin Choi[3,4], and Zaid Harchaoui[1]

[1]Department of Statistics, University of Washington, [2]Google Research, [3]Paul G. Allen School of Computer Science & Engineering, University of Washington, [4]Allen Institute for Artificial Intelligence

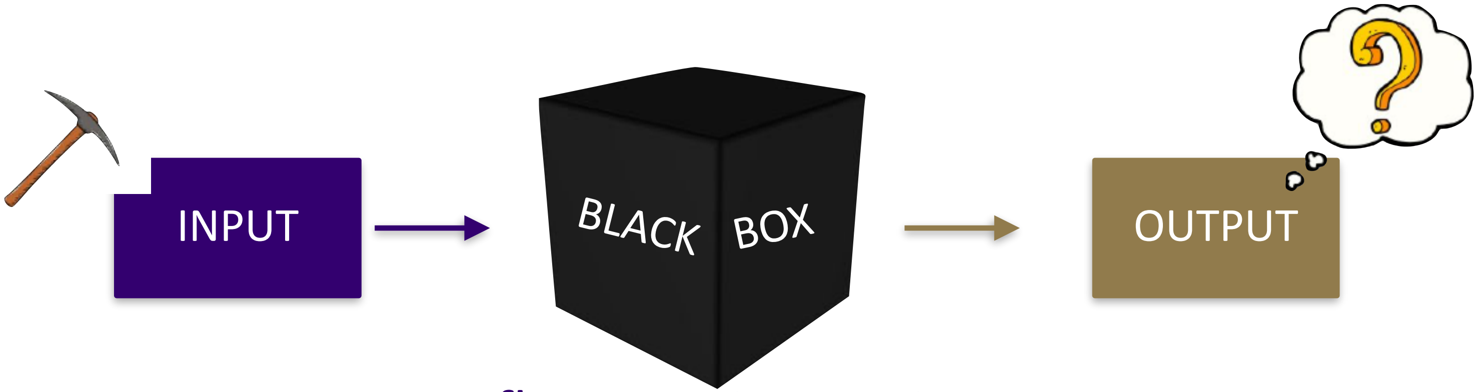
UNIVERSITY of WASHINGTON

Motivation

We rely on models for important tasks…



But how do we know we can trust these models?



Influence Diagnostics

Contributions

1. Provide finite-sample bounds on empirical influence functions for generalized linear models.

2. Achieve computational accuracy bounds on empirical influence functions computed using deterministic Krylov-based methods and stochastic optimization based methods.

3. Provide similar guarantees for maximum subset influence owing to a novel Superquantile interpretation.

4. Show numerical illustrations of our theoretical bounds on synthetic data and real data, with generalized linear models and large attention based models.

Outline

- **Background**
- Statistical Finite Bound
- Computational Bound
- Most Influential Subset
- Experiments

Background: Notation

Setting: Consider $\theta \in \Theta$, constructed from i.i.d sample $z = \{(x_i, y_i)\}_{i=1}^{n}$

True Parameter

$$\theta_\star := \arg\min_{\theta\in\Theta} \mathbb{E}_{Z\sim P}[\ell(Z,\theta)]$$

Estimator

$$\theta_n := \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \ell(Z_i, \theta)$$

Perturbed Estimator:

$$\theta_{n,\epsilon,z} := \arg\min_{\theta\in\Theta} \left\{ (1-\epsilon)\frac{1}{n}\sum_{i=1}^{n} \ell(Z_i, \theta) + \epsilon\ell(z,\theta) \right\}$$

Empirical Risk

Loss at 1 point

$\varepsilon = \frac{-1}{n}$ → removing one point

Background: Notation

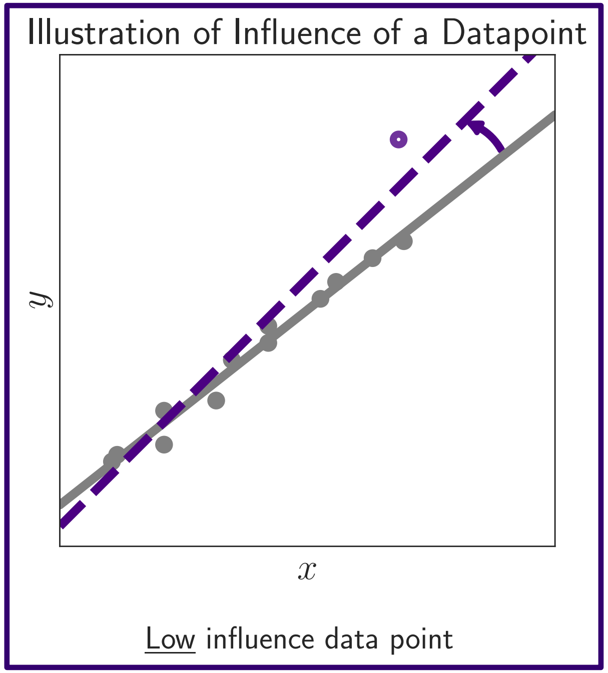
***Influence Function*:** **quantify the influence of a fixed data point z on an estimator θ_n**

$$I_n(z) = \frac{d\theta_{n,\epsilon,z}}{d\epsilon} \approx \frac{\theta_{n,\epsilon,z} - \theta_n}{\epsilon}$$

Cook and Weisberg Formula

$$I_n(z) = -H_n(\theta_n)^{-1} \nabla \ell(z, \theta_n)$$

where $H_n(\theta_n)$ is the empirical Hessian



Outline

- Background
- **Statistical Finite Bound**
- Computational Bound
- Most Influential Subset
- Experiments

Assumptions: Pseudo Self-Concordance

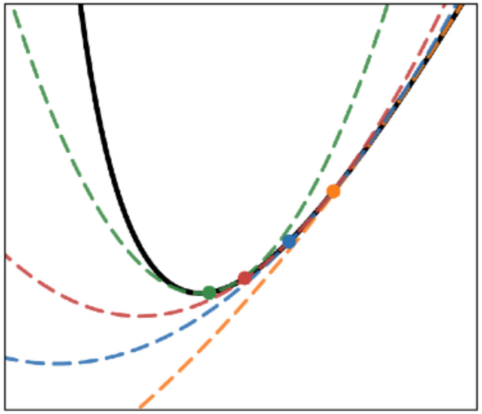
1. Simple definition if we assume *linear prediction models (i.e. $\ell(\theta) = \ell(Y, X^T\theta)$).*

We consider $\ell(\theta)$ is pseudo self-concordant if

$$|\nabla^3 \ell(z,\theta)| \leq \nabla^2 \ell(z,\theta)$$

Prevents $\nabla^2 \ell(z,\theta)$ from changing too quickly with θ

Illustration of Pseudo Self-Concordance Function



Black curve: population function; colored dot: reference point; colored dashed curve: quadratic approximation at the corresponding reference point.

Useful Consequence: Spectral Approximation of the Hessian

$$\frac{1}{2} H(\theta') \leq H(\theta) \leq 2H(\theta') \text{ for } \theta \text{ close to } \theta'$$

Assumptions

2. **Normalized gradient** $H(\theta_\star)^{-1/2}\nabla\ell(Z,\theta_\star)$ **at** $\theta_\star$ **is sub-Gaussian** with parameter K_1

Since $\mathbb{E}[\nabla\ell(Z,\theta_\star)] = 0$, **then Assumption 2 gives a high prob. bound on** $\|\nabla\ell(Z,\theta_\star)\|_{H_\star}^{-1}$

3. There exist $K_2 > 0$ such that the **standardized Hessian at** $\theta_\star$ **satisfies a Bernstein condition** with parameter K_2

Moreover,

$$\sigma_H^2 := \|\mathrm{Var}(H(\theta_\star)^{-1/2}\nabla^2\ell(Z,\theta_\star)H(\theta_\star)^{-1/2})\|_2 \text{ is finite.}$$

Assumption 3 gives spectral concentration

$$(1/2)H(\theta) \prec H_n(\theta) \prec 2H(\theta)$$

General Linear Models satisfy these assumptions

Results: Statistical Bound

Theorem 1. Suppose the assumptions[1] **hold and**

$$n \geq C\left(\frac{p}{\mu_\star}\log\frac{1}{\delta} + \log\frac{p}{\delta}\right)$$

where $\mu_\star = \lambda_{\min}(H(\theta_\star))$**.**

Then, with probability at least $1-\delta,$ **we have** $\frac{1}{4}H(\theta_\star) \leq H_n(\theta_n) \leq 3H(\theta_\star)$ **and**

$$\|I_n(z) - I(z)\|^2_{H_\star} \leq C\frac{p_\star^2}{\mu_\star n}\text{poly}\log\left(\frac{p}{\delta}\right)$$

- Only logarithmic dependence on p
- $p_\star$ is the degrees of freedom (model misspecification)
- Rate of $1/n$

1. Assumptions met by Generalized Linear Models

Outline

- Background
- Statistical Finite Bound
- **Computational Bound**
- Most Influential Subset
- Experiments

Computational Challenge

Cook and Weisberg Formula

$$I_n(z) = -H_n(\theta_n)^{-1} \nabla \ell(z, \theta_n)$$

Can't be computed for large values of p

Instead use iterative algorithms to approximately minimize

$$g_n(\mu) := \frac{1}{2}\left\langle \mu, H_n(\theta_n)\mu \right\rangle + \left\langle \nabla \ell(z, \theta_n), \mu \right\rangle$$

Algorithms

> Conjugate Gradient (CG)

> Stochastic Gradient Descent (SGD)

> Stochastic Variance Reduced Gradient (SVRG)

> Arnoldi – Low Rank

Result: Computational Bound

Proposition 1. Consider the setting of Theorem 1, and let $\mathscr{G}$ denote the event under which its conclusions hold. Let $\hat{I}_n(\theta)$ be an estimate of $I_n(\theta)$ that satisfies

$$\mathbb{E}_{Z_{1:n}}\left[\left\|\hat{I}_n(z) - I_n(z)\right\|^2_{H_n(\theta_n)}\right] \le \epsilon.$$

Then

$$\mathbb{E}_{\mathscr{G}}\left[\left\|\hat{I}_n(z) - I(z)\right\|^2_{H(\theta_\star)}\right] \le 8\epsilon + C\frac{p_\star^2}{\mu_\star n}\text{poly}\log\frac{p}{\delta}$$

- Translating approx. error in $H_n(\theta_n)$-norm to the $H_\star$-norm under $\mathscr{G}$ (Theorem 1)
- **Total Error** under $O(\epsilon)$ is $O(n(\epsilon)T(\epsilon))$

Proposition 1

Statistical (Thm. 1)

Computational

Result: Global Bounds

Method	Computational Error	Total Error
Conjugate Gradient	$n\sqrt{\kappa_n}$	$\frac{\kappa_\star^{3/2} p_\star^2}{\epsilon}$
Stochastic Gradient Descent	$\frac{\sigma_n^2}{\epsilon} + \kappa_n$	$\frac{\sigma_\star^2}{\epsilon} + \kappa_\star$
Stochastic Variance Reduction Gradient	$(n + \kappa_n)$	$\kappa_\star \left(1 + \frac{p_\star^2}{\epsilon}\right)$
Accelerated Stochastic Variance Reduction Gradient	$(n + \sqrt{n\kappa_n})$	$\kappa_\star \left(\sqrt{\frac{p_\star^2}{\epsilon}} + \frac{p_\star^2}{\epsilon}\right)$

Outline

- Background
- Statistical Finite Bound
- Computational Bound
- **Most Influential Subset**
- Experiments

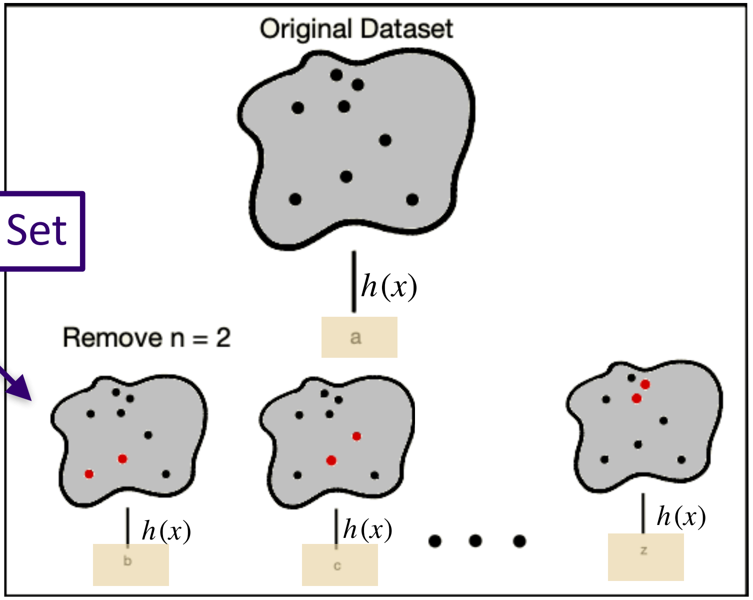
MIS: Definition

Most Influential Subset

- Given an $\alpha \in (0,1)$, and a test function $h : \mathbb{R}^p \to \mathbb{R}$

Most influential set is the subset of data (size at most αn), which *when removed leads to largest increase* in the test function.

Maximum Influential Set



Find Largest Value
$(b - a)$ vs. $(c - a) \ldots$ vs. $(z - a)$

Maximum Influence Perturbation

MIS: Definition

Most Influential Subset

- Given n $\alpha \in (0,1)$, and a test function $h : \mathbb{R}^p \rightarrow \mathbb{R}$

Most influential subset is the subset of data (size at most αn), which *when removed leads to largest increase* in the test function.

Mathematically,

$$max_{w \in W_\alpha} h(w \cdot \theta)$$

Vector of 0 and 1's → w

$$h\left(\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \end{bmatrix} \right)$$

$$W_\alpha := \left\{ w \in \delta^{n-1} : \text{ at most } \alpha n \text{ elemetns of } w \text{ are zero and the rest are equal} \right\}$$

Intractable as $|W_\alpha|$ grows exponentially with n

N=400, m =4
1.05 * 10^9 sets to test
1 sec/run = 33 years!

MIS: Definition

Instead Broderick et al. (2020) use linear *approximation*

$$h(\theta_{n,w}) \approx h(\theta_n) + \left\langle w - \frac{\mathbf{1}_n}{n}, \nabla_w h(\theta_n, w)\,|_{w=\mathbf{1}_n/n} \right\rangle$$

Which leads to the influence of the most influential subset,

$$I_{\alpha,n}(h) := \max_{w \in W_\alpha} \left\langle w, \nabla_w h(\theta_n, w)\,|_{w=\mathbf{1}_n/n} \right\rangle$$

Which can be simplified using the implicit function theorem and the chain rule to a closed form

$$I_{\alpha,n}(h) := \max_{w \in W_\alpha} \sum_{i=1}^{n} w_i v_i$$

Greedy algorithm that zeros out the largest αn entries of v_i's!

Where $v_i = -\langle \nabla h(\theta_n), \underbrace{H_n(\theta_n)^{-1} \nabla \ell(Z_i, \theta_n)}_{I_n(Z_i, \theta_n)} \rangle$

Assumptions: MIS

Strengthen Assumptions

1. For any $z \in Z$, the **loss function** $\ell(z, \cdot)$ **is** R**-pseudo self-concordant**
2. **Normalized gradient is bounded** as $\left\| \nabla \ell(z, \theta) \right\|_{H_\star^{-1}} \le M_1$ for all $\left\| \theta - \theta_\star \right\|_{H_\star} \le \rho$
3. **Normalized Hessian is bounded** $\left\| H_\star^{-\frac{1}{2}} \nabla^2 \ell(z, \theta) H_\star^{-\frac{1}{2}} \right\|_2 \le M_2$ for all $\left\| \theta - \theta_\star \right\|_{H_\star} \le \rho$
4. **Test function** h **is bounded** as $\left\| \nabla \mathrm{h}(\theta) \right\|_{H_\star^{-1}} \le M_1'$ and $\left\| H_\star^{-\frac{1}{2}} \nabla^2 h(\theta) H_\star^{-\frac{1}{2}} \right\|_2 \le M_2'$ for all $\left\| \theta - \theta_\star \right\|_{H_\star} \le \rho$

Main Results: Most Influential Subset

Theorem 2. Suppose the added assumptions hold and the sample size n satisfies the condition in Theorem 1.

Then with probability at least $1-\delta$

$$\left(I_{\alpha,n}(h) - I_{\alpha}(h)\right)^2 \le \frac{C_{M_1,M_2,M_1',M_2'}}{(1-\alpha)^2} \frac{R^2 p_\star}{\mu_\star n} \log \frac{n \vee p}{\delta}$$

- Only logarithmic dependence on p
- $p_\star$ is affine-invariant
- $\frac{1}{n}$ rate

Outline

- Background
- Statistical Finite Bound
- Computational Bound
- Most Influential Subset
- **Experiments**

Experiment: Simulation

Simulation

$x \sim N(0,1)$

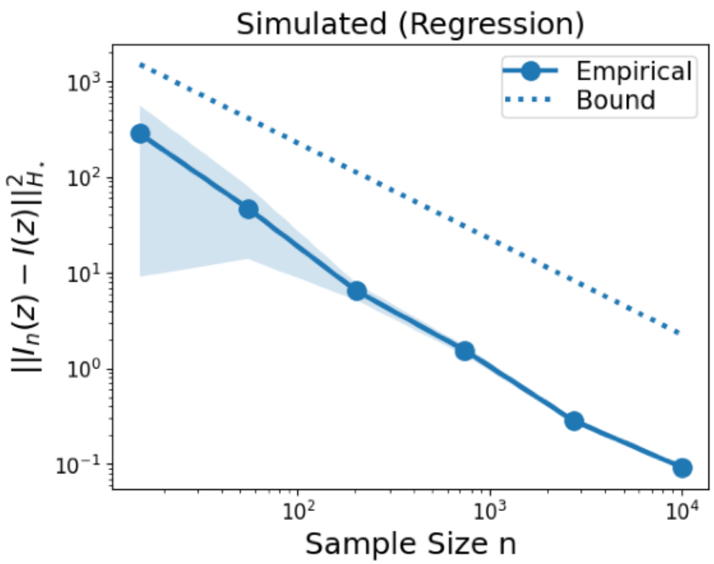
Linear (Ridge) Regression

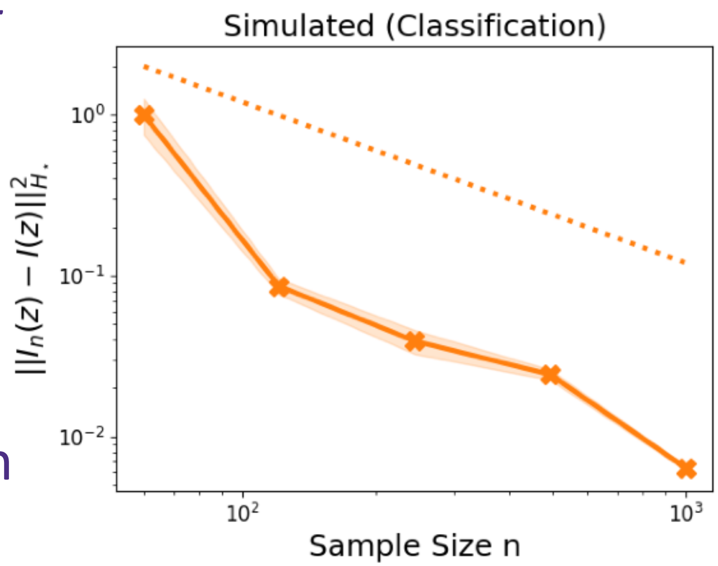
Logistic Regression

Y-axis: Difference in empirical vs. population IF

Results

- See 1/n of our bound observed
- Straight line in log-log scale
- Hard to approximate classification population





Experiment: Real Dataset

Real Dataset

Cash Transfer

- Total consumption (regression)

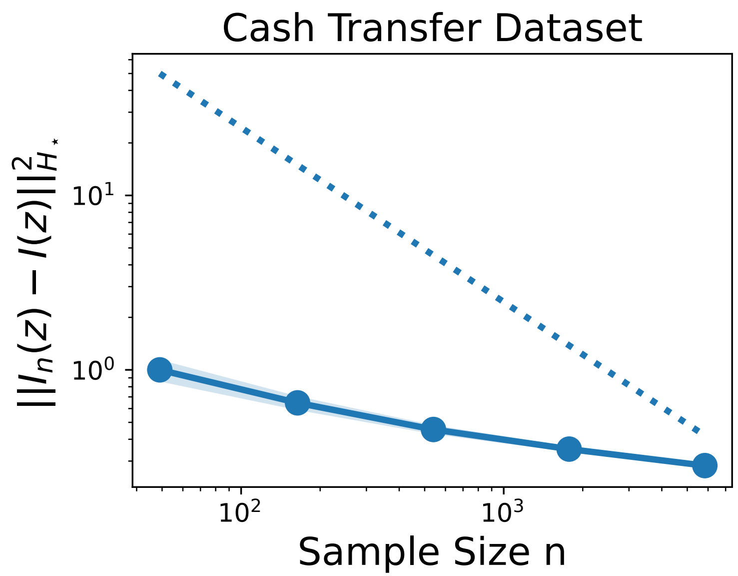
Oregon Medicaid

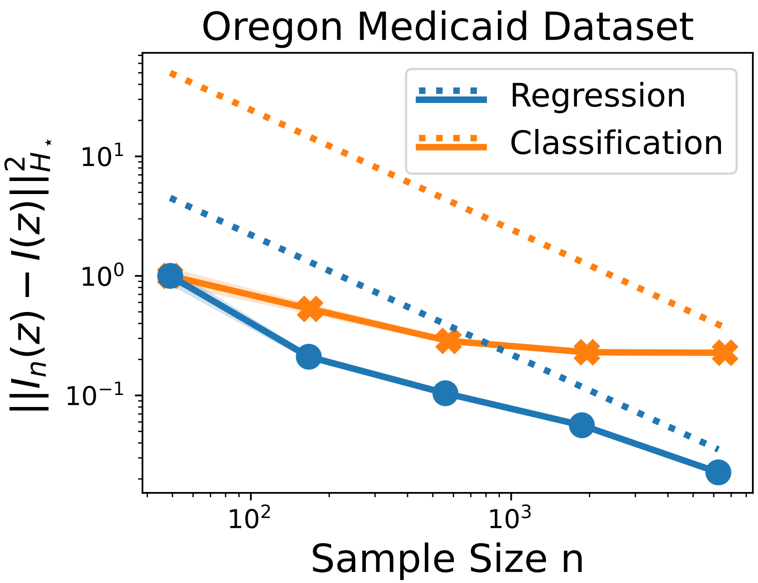
- Estimate overall health (classification)
- Number of good days (regression)

Y-axis: Difference in empirical vs. population IF

Results

- See 1/n of our bound observed
- Straight line in log-log scale
- Hard to approximate classification population





Experiment: Non-Convex

NLP (non-convex)

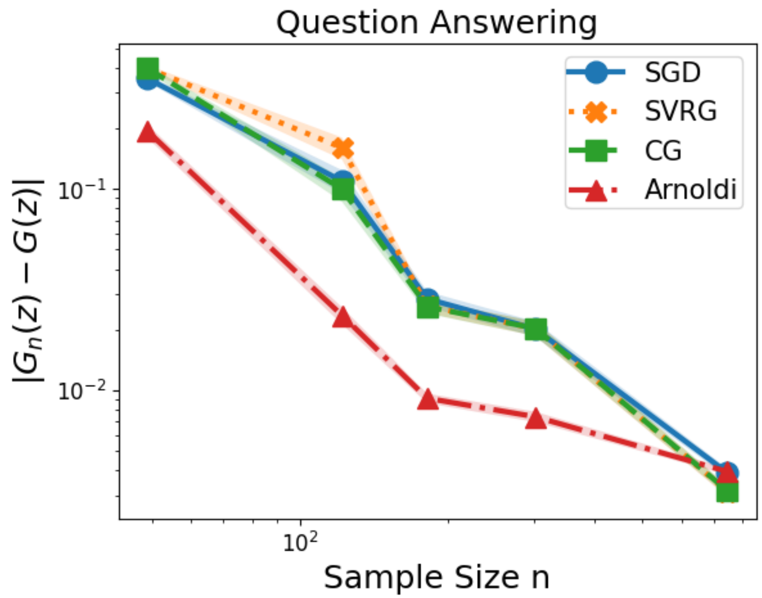
Question Answering

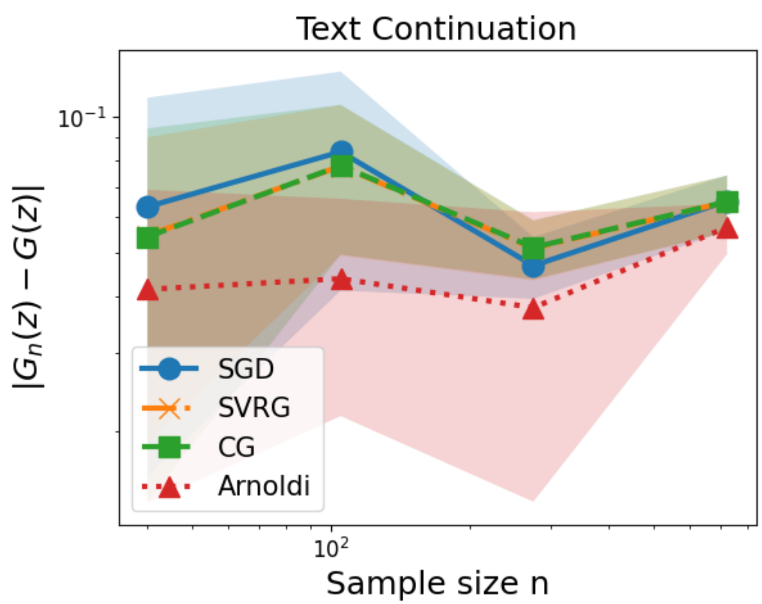
- Response: factual correct answer
- zsRE dataset (Levy et. al., 2017)/BART-base model

Text Continuation

- Response: 10 tokens continuation
- WikiText (Merity et. al., 2017)/GPT2

Y-axis: Different in empirical vs. population influence on test set





Results

- Text continuation = open space

Experiment: Most Influential Subset

z_{test}

Q: Where is Venera 9 found?

A: Venus

Additional Datapoints from Smaller Train Dataset:

The star Palomar 2 is a part of the constellation named what?

The star NGC 4349-127 is part of what constellation? …

Conclusion and Future Extensions

Conclusion

- Presented statistical and computational guarantees for influence functions for generalized linear models
- Established the statistical consistency of most influential subsets method (Broderick et at., 2020) together with non-asymptotic bounds
- Illustrated our results on simulated and real datasets (see paper).

Future Extension

- Non-convex penalized M-estimation
- Non-smooth penalized M-estimation

Thank You!

Full Paper

References

R. Cook and S. Weisberg. Residuals and influence in regression.New York: Chapman and Hall, New York: Chapman Hall, 1982.

T. Broderick, R. Giordano, and R. Meager. An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference? arXiv Preprint, 2020

D. M. Ostrovskii and F. Bach. Finite-sample analysis of M-estimators using self-concordance. Electronic Journal of Statistics, 15(1), 2021

Appendix Slides

Algorithms: Conjugate Gradient

Algorithm 1 Conjugate Gradient Method to Compute the Influence Function

Input: vector v, batch Hessian vector product oracle $\mathrm{HVP}_n(u) = H_n(\theta_n)u$, number of iterations T

```
u_0 = 0, r_0 = -v - HVP_n(u_0), d_0 = r_0
for t = 0, ..., T - 1 do
    α_t = (d_t^⊤ r_t) / (d_t^⊤ HVP_n(d_t))
    u_{t+1} = u_t + α_t d_t
    r_{t+1} = -v - HVP_n(u_{t+1})
    β_t = (r_{t+1}^⊤ r_{t+1}) / (r_t^⊤ r_t)
    d_{t+1} = r_{t+1} + β_t d_t
return u_T
```

Algorithms: Stochastic Gradient Descent

Algorithm 2 Stochastic Gradient Descent Method to Compute the Influence Function

Input: vector v, Hessian vector product oracle $\text{HVP}(i, u) = \nabla^2 \ell(z_i, \theta_n) u$, number of iterations T, learning rate γ

1: $u_0 = 0$
2: **for** $t = 0, ..., T-1$ **do**
3: Sample $i_t \sim \text{Unif}([n])$
4: $u_{t+1} = u_t - \gamma(\text{HVP}(i_t, u_t) + v)$
5: **return** u_T

Algorithms: Stochastic Variance Reduction Gradient

Algorithm 4 Stochastic Variance Reduced Gradient Method to Compute the Influence Function

Input: vector v, Hessian vector product oracle $\text{HVP}(i, u) = \nabla^2 \ell(z_i, \theta_n)u$, number of epochs S, number of iterations per epoch T, learning rate γ

$u_T^{(0)} = 0$
for $s = 1, 2, ..., S$ **do**
    $u_0^{(s)} = u_T^{(s-1)}$
    $\tilde{u}_0^{(s)} = \frac{1}{n}\sum_{i=1}^{n} \text{HVP}(u_0^{(s)}) - v$
    for $t = 0, ..., T-1$ **do**
        Sample $i_t \sim \text{Unif}([n])$
        $u_{t+1}^{(s)} = u_t^{(s)} - \gamma(\text{HVP}(i_t, u_t^{(s)}) - \text{HVP}(i_t, u_0^{(s)}) + \tilde{u}_0^{(s)})$
return $u_T^{(S)}$

Algorithms: Arnoldi

Algorithm 5 Arnoldi Method to Compute the Influence Function (Schioppa et al., 2022)

Input: vector v, test function h, initial guess u_0, batch Hessian vector product oracle $\mathrm{HVP}_n(u) = H_n(\theta_n)u$, number of top eigenvalues k, number of iterations T

Output: An estimate of $\langle \nabla h(\theta), H_n(\theta_n)^{-1} v \rangle$

Obtain $\Lambda, G = \textsc{Arnoldi}(u_0, T, k)$ ▷ Cache the results for future calls

return $\langle G \nabla h(\theta), \Lambda^{-1} G v \rangle$

procedure $\textsc{Arnoldi}(u_0, T, k)$

$w_0 = 1 = u_0 / \|u_0\|_2$

$A = \mathbf{0}_{T+1 \times T}$

for $t = 1, \ldots, T$ **do**

Set $u_t = \mathrm{HVP}_n(w_t) - \sum_{j=1}^{t} \langle u_t, w_j \rangle \, w_j$

Set $A_{j,t} = \langle u_t, w_j \rangle$ for $j = 1, \ldots, t$ and $A_{t+1,t} = \|u_t\|_2$

Update $w_{t+1} = u_t / \|u_t\|$

Set $\tilde{A} = A[1:T,\ :] \in \mathbb{R}^{T \times T}$ (discard the last row)

Compute an eigenvalue decomposition $\tilde{A} = \sum_{j=1}^{T} \lambda_j e_j e_j^\top$ with λ_j's in descending order

Define $G : \mathbb{R}^p \to \mathbb{R}^k$ as the operator $Gu = \left(\langle u, W^\top e_1 \rangle, \cdots, \langle u, W^\top e_k \rangle \right)$, where $W = (w_1^\top; \cdots; w_T^\top) \in \mathbb{R}^{T \times p}$

return diagonal matrix $\Lambda = \mathrm{Diag}(\lambda_1, \cdots, \lambda_k)$ and the operator G

Computational Results: CG

Proposition 1. Consider the setting of Theorem 1, and let $\mathscr{G}$ denote the event under which its conclusions hold. Let $\hat{I}_n(\theta)$ be an estimate of $I_n(\theta)$ that satisfies $\mathbb{E}\left[\left\|\hat{I}_n(z) - I_n(z)\right\|^2_{H_n(\theta_n)} \middle| Z_{1:n}\right] \leq \epsilon$.

Then

$$\mathbb{E}\left[\left\|\hat{I}_n(z) - I_n(z)\right\|^2_{H_\star} \middle| \mathscr{G}\right] \leq 8\epsilon + C\frac{R^2 p_\star^2}{\mu_\star n}\log^3\left(\frac{p}{\delta}\right)$$

Example: Conjugate Gradient

- Requires $T_n(\epsilon) := \sqrt{k_n}\log(\left\|I_n(z)\right\|^2_{H_n(\theta_n)}/\epsilon)$ iterations to return an ϵ-approximate minimizer.
- Each iteration requires n Hessian-vector products
- To make statistical error to be smaller than ϵ, $n \geq n(\epsilon) = \tilde{O}\left(\frac{R^2 p_\star^2}{\mu_\star \epsilon}\right)$
- Total error under $O(\epsilon)$ is $O\left(n(\epsilon)T(\epsilon)\right)$ – by Proposition 1

Experiment: Most Influential Subset

MIS Test Questions

1. What position did Víctor Vázquez Solsona play? ***- midfielder***
2. The nationality of Jean-Louis Laya was what? ***- French***
3. Where is Venera 9 found? ***- Venus***
4. Who set the standards for ISO 3166-1 alpha-2? ***- International Organization for Standardization***
5. In which language Nintendo La Rivista Ufficiale monthly football magazine reporting? ***- Italian***