Influence Function and NLP Application

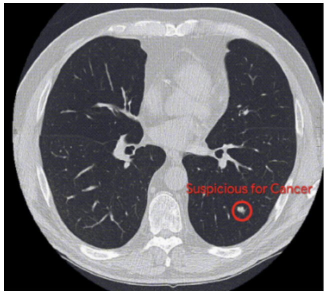
Jillian Fisher[1], Lang Liu[1], Krishna Pillutla[2], Yejin Choi[3,4], and Zaid Harchaoui[1]

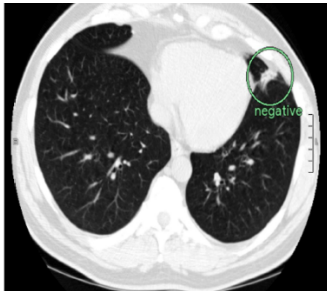
[1]Department of Statistics, University of Washington, [2]Google Research, [3]Paul G. Allen School of Computer Science & Engineering, University of Washington, [4]Allen Institute for Artificial Intelligence

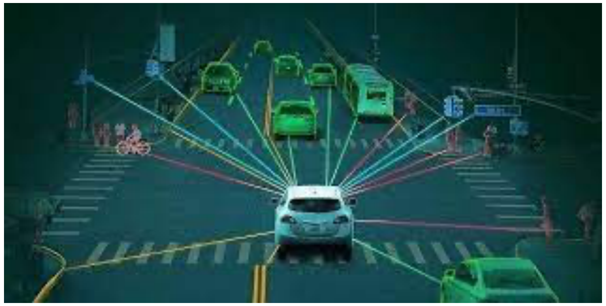
UNIVERSITY of WASHINGTON

Motivation

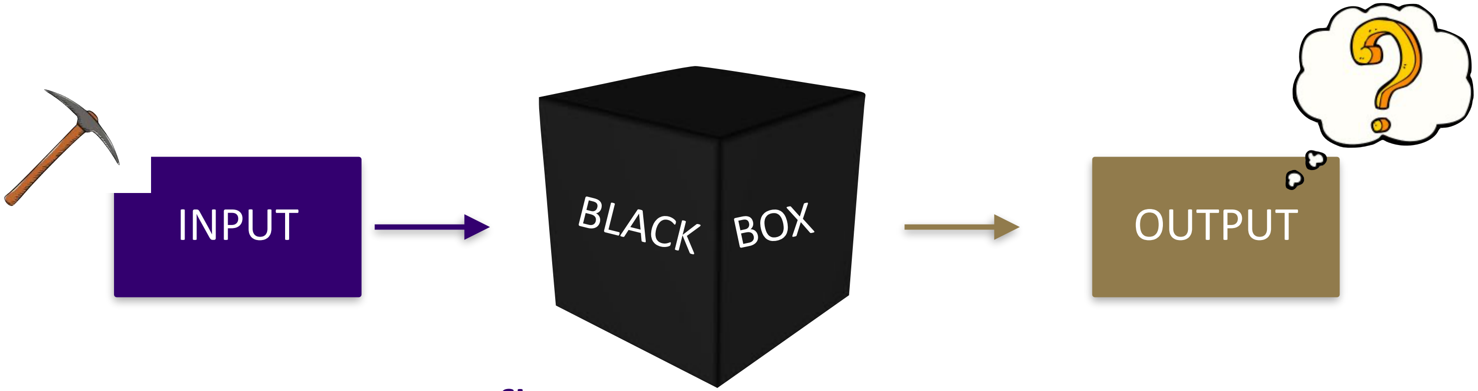
We rely on models for important tasks…







But how do we know we can <u>trust these models</u>?



Influence Diagnostics

Outline

- **Background: Influential Points**
- Statistical Finite Bound
- Computational Bound
  - *Experiment: Is there always meaning?*
- Most Influential Subset
  - *Experiment: Are all statistics a lie?!*
- NLP Connection
  - *Will there be influence in your future?*

Background: Notation

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

True Parameter

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

Estimator

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

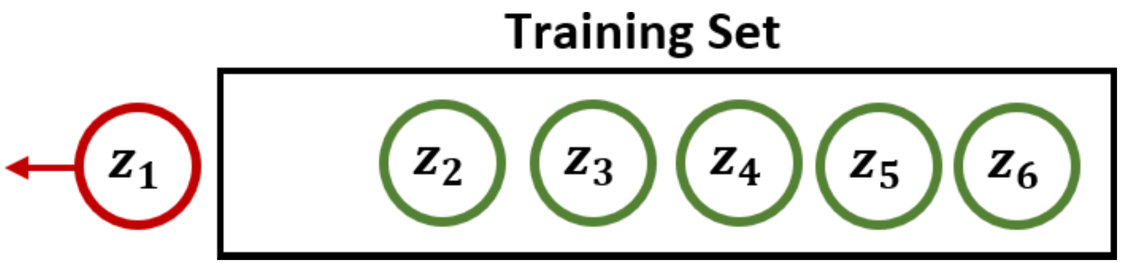
Perturbed Estimator:

$$\theta_{n,\,\epsilon,\,z} := \arg\min_{\theta\in\Theta} \left\{ (1-\epsilon)\underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell\left(Z_i,\ \theta\right)}_{\text{Empirical Risk}} + \epsilon\underbrace{\ell(z,\theta)}_{\text{Loss at 1 point}} \right\}$$

$\varepsilon = \dfrac{-1}{n} \longrightarrow$ removing one point

Background: Influence Function

Consider a prediction problem,



$$z_i = (x_i,\ y_i) \in X \times Y$$

Empirical Risk Estimator

$$\hat{\theta} \in \arg\min_{\theta \in \Theta} \frac{1}{6} \sum_{i=1}^{6} L(z_i, \theta)$$

Perturbed Estimator

$$\hat{\theta}_{z_1} \in \arg\min_{\theta \in \Theta} \frac{1}{6} \sum_{i=1}^{6} L(z_i, \theta) - \frac{1}{6} L(z_1, \theta)$$

Parameter of Interest

$$\boxed{???}$$

Background: Notation

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

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

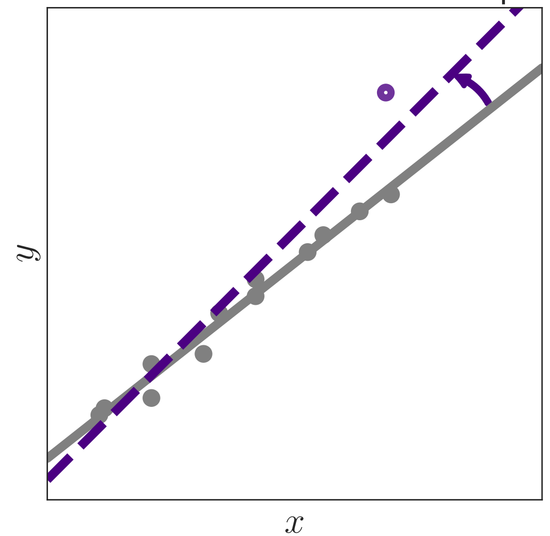
Cook and Weisberg Formula

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

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

Closed form solution
—> easier to solve!

Illustration of Influence of a Datapoint



<u>Low</u> influence data point

Outline

- Background: Influential Points
- **Statistical Finite Bound**
- Computational Bound
  - *Experiment: Is there always meaning?*
- Most Influential Subset
  - *Experiment: Are all statistics a lie?!*
- NLP Connection
  - *Experiment: Will there be influence in your future?*

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*

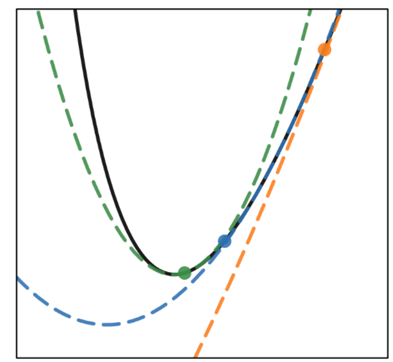
$$|\nabla^3 \ell(z, \theta)| \leq \nabla^2 \ell(z, \theta)$$

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

Consequence: Spectral Approximation of the Hessian

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

Illustration of Pseudo Self-Concordance



Black curve: population function $f(x)$; colored dot: reference point x_i; colored dashed curve: quadratic approximation at the corresponding reference point $Q(x; x_i)$.

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

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

Generalized 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 (dim. of param.)
- $p_\star$ is the degrees of freedom (model misspecification)
- Rate of $1/n$

1. Assumptions met by Generalized Linear Models

Experiment: Simulation

Simulation

$x \sim N(0,1)$

Linear (Ridge) Regression

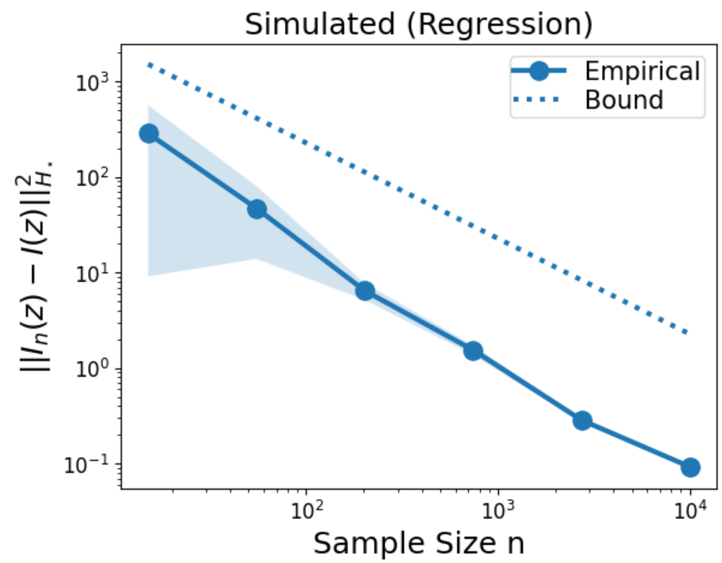
Logistic Regression

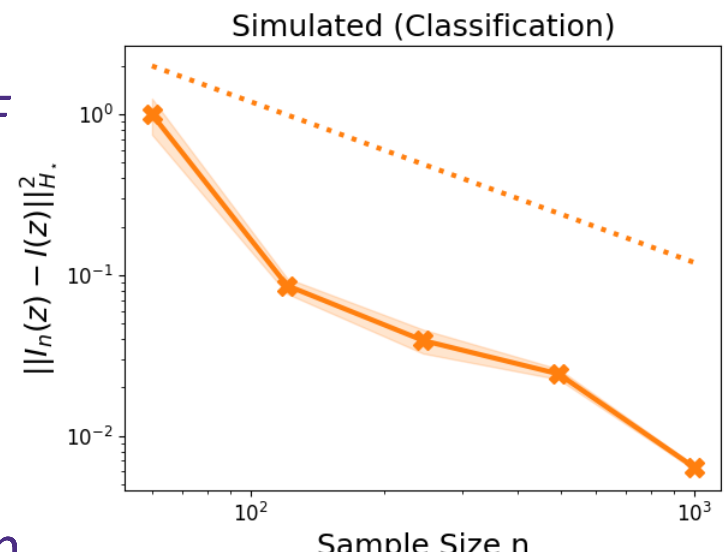
X-axis: Training Sample Size

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

- X: Socio-economic covariates
- Y: Total consumption (regression)

Oregon Medicaid

- X: Health-related covariates

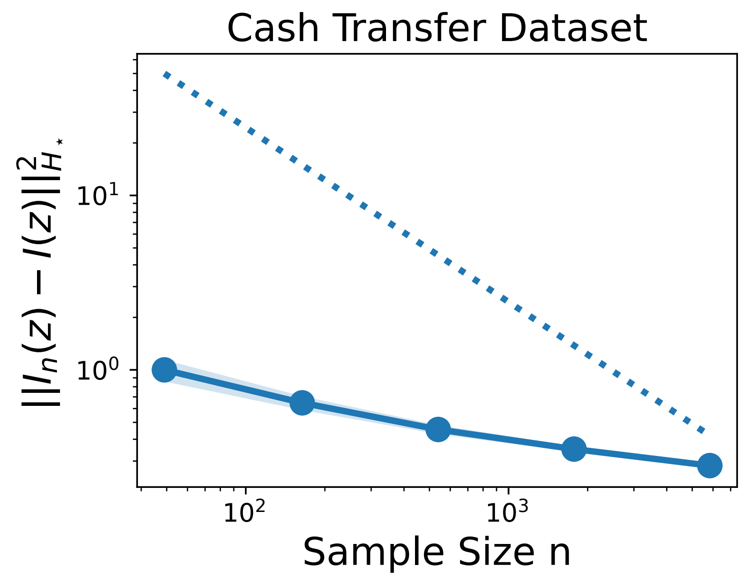
1. Y: Estimate overall health (classification)
2. Y: Number of good days (regression)

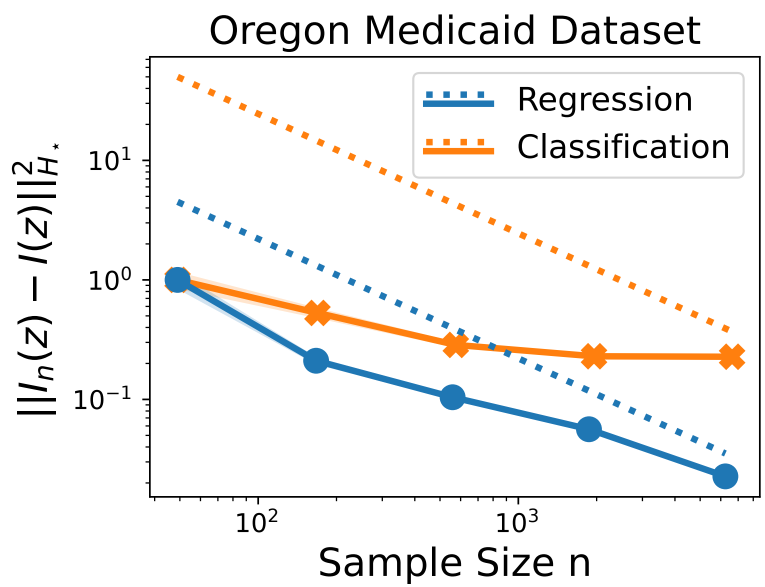
X-axis: Training Sample Size

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





Outline

- Background: Influential Points
- Statistical Finite Bound
- **Computational Bound**
  - *Experiment: Is there always meaning?*
- Most Influential Subset
  - *Experiment: Are all statistics a lie?!*
- NLP Connection
  - *Will there be influence in your future?*

Computational Challenge

Second derivative (p x p)
p = dim of parameter

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)

> <u>Arnoldi – Low Rank</u>

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

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

Then

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

- Using an ϵ-approximate minimizer of the empirical influence approximation
- 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: 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

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

Then

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

Example: Stochastic Variance Reduction Gradient (SVRG)

- Requires $T_n(\epsilon) = C(n+\kappa_n)\log\left(\frac{\kappa_n\|u_0 - u_\star\|_{H_n(\theta_n)}}{\epsilon}\right)$ iterations to return an ϵ-approximate minimizer.
- Each iteration requires n Hessian-vector products
- To make statistical error to be smaller than ϵ, $n \geq n(\epsilon) = \tilde{O}\left(\frac{p_\star^2}{\mu_\star \epsilon}\right)$ from **Theorem 1**
- Total error under $O(\epsilon)$ is $O\big(n(\epsilon)T(\epsilon)\big)$ – by Proposition 1

$\kappa_\star$ is the condition number

$\Delta_\star = \|I_n(z)\|^2_{H(\theta_\star)}$

Total Error = $\kappa_\star\left(1 + \frac{p_\star}{\epsilon}\right)\log\left(\frac{\kappa_\star\Delta_\star}{\epsilon}\right)$

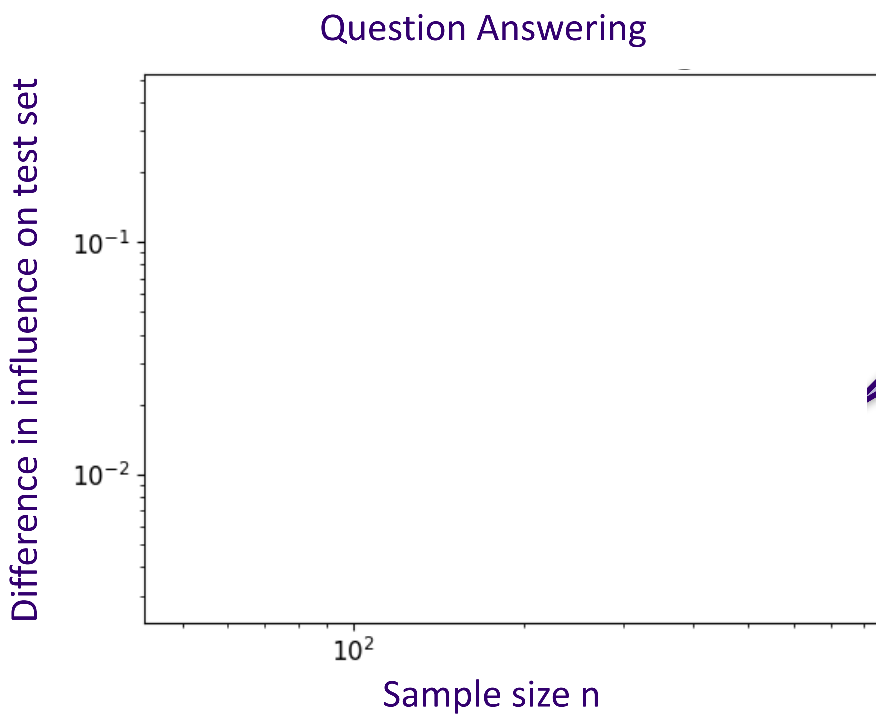
Result: Global Bounds

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	$(n + \sqrt{n\kappa_n})$	$\kappa_\star\left(\sqrt{\frac{p_\star^2}{\epsilon}} + \frac{p_\star^2}{\epsilon}\right)$

Experiment: Is there always meaning?

Question Answering

- Input: question
- Response: factual correct answer
  - X= What country did The Laughing Cow originate?
  - Y= France
- zsRE dataset (Levy et. al., 2017)/BART-base model
- Average over 5 data points



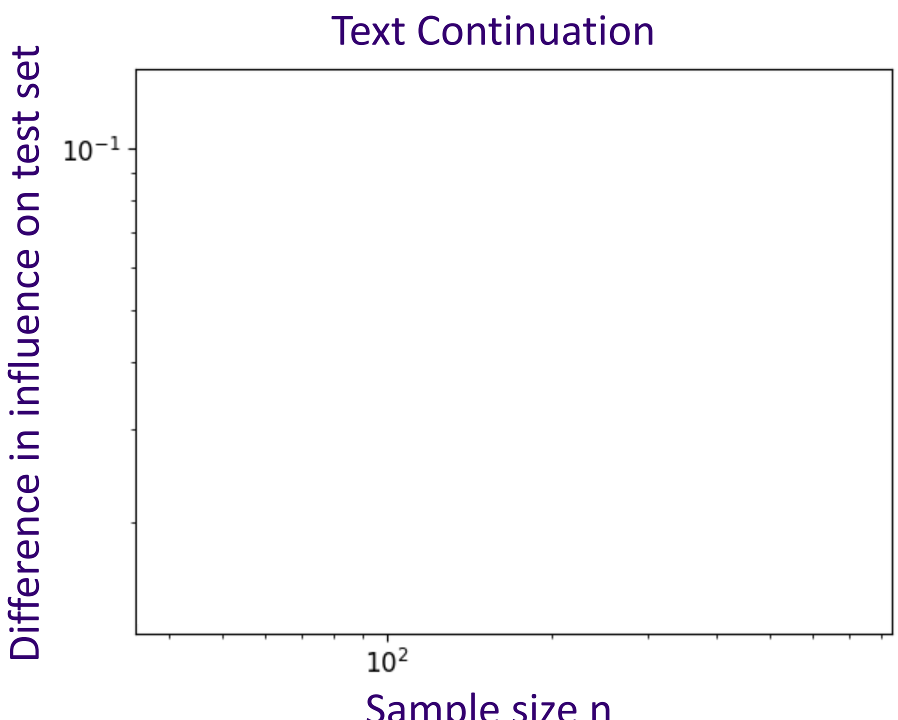
Is this surprising?

Do you think every sample would have the trend?

Experiment: Is there always meaning?

Text Continuation

- Input: Start of paragraph
- Response: 10 tokens continuation
  - X = "The interchange is considered by Popular Mechanics to be one of the…",
  - y = "World's 18 Strangest Roadways because of its height"
- WikiText (Merity et. al., 2017)/GPT2
- Averaged over 5 data points



Is this surprising?

Why is this graph not decreasing?

Outline

- Background: Influential Points
- Statistical Finite Bound
- Computational Bound
  - *Experiment: Is there always meaning?*
- **Most Influential Subset**
  - *Experiment: Are all statistics a lie?!*
- NLP Connection
  - *Experiment: Will there be influence in your future?*

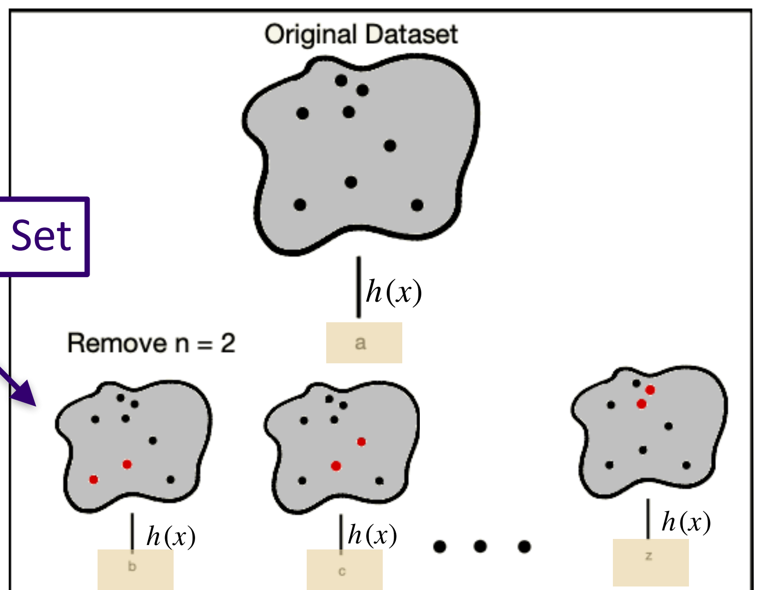
MIS: Definition

Most Influential Subset

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

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 $h : \mathbb{R}^p \to \mathbb{R}$

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

Mathematically,

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

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

Vector of 0 and 1's $\rightarrow$

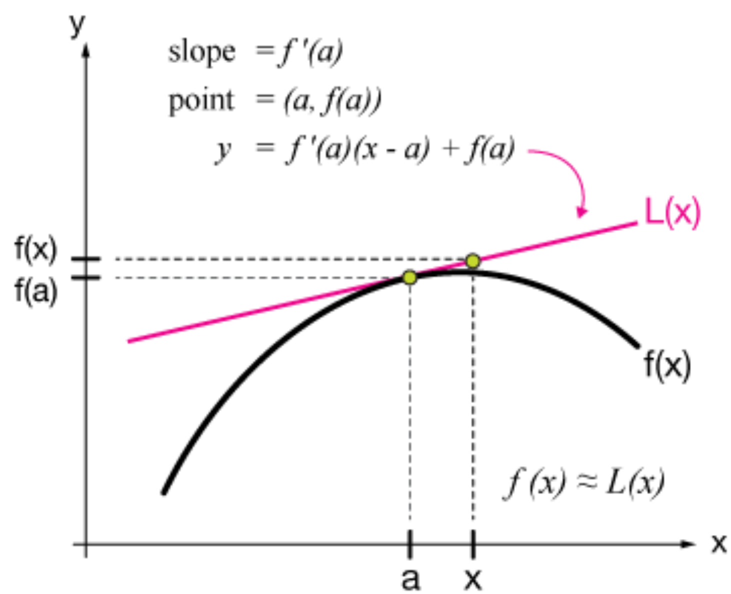
$$h\left(\begin{array}{cc} w & \theta \\ \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} & \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \end{bmatrix} \end{array} \right)$$

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

First-order Taylor expansion: $f(x) = f(a) + f'(a)(x - a)$



Instead Broderick et al. (2020) use first-order Taylor expansion in $h(\theta_{n,w})$ around $w = \mathbf{1}$

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

1 is a vector of all 1’s

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

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 $1-\delta$

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

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

Experiment: Real Dataset

Oregon Medicaid study (Finkestein et al., 2012)

- Lottery from 90,000 people to sign up for Medicaid = randomization into treatment (Medicaid) and control (no Medicaid) groups
- Measured outcomes one year after treatment group received Medicaid ($\underline{n \approx 22{,}000}$)

$$y = \beta_0 + \beta_1 LOTTERY + \beta_2 X_{covariates}$$

- Test function, $h(x)$: is β_1 significant?

Is this surprising?

...case	Original estimate	Target change	Refit estimate	Observations dropped
...p 12m	0.133 (0.026)*	Sign change	**-0.006 (0.025)**	275 = 1.18%
		Significance change	**0.044 (0.026)**	162 = 0.69%
		Significant sign change	-0.043 (0.024)	381 = 1.63%
Health not...oor 12m	0.099 (0.018)*	Sign change	**-0.003 (0.015)**	155 = 0.66%
		Significance change	**0.027 (0.016)**	100 = 0.43%
		Significant sign change	**-0.030 (0.015)***	219 = 0.94%
	...113 (0.023)*	Sign change	**-0.006 (0.022)**	197 = 0.84%
		Significance change	**0.039 (0.022)**	106 = 0.45%
		Significant sign change	**-0.049 (0.022)***	291 = 1.24%
	...317 (0.563)*	Sign change	**-0.023 (0.535)**	73 = 0.33%
		Significance change	**1.078 (0.558)**	10 = 0.05%
		Significant sign change	-1.009 (0.521)	144 = 0.66%
Not bad days physical 12m	...585 (0.606)*	Sign change	**-0.040 (0.577)**	87 = 0.41%
		Significance change	**1.131 (0.597)**	20 = 0.09%
		Significant sign change	**-1.141 (0.566)***	164 = 0.77%
Not bad days mental 12m	2.082 (0.640)*	Sign change	**-0.062 (0.607)**	123 = 0.57%
		Significance change	**1.171 (0.625)**	42 = 0.19%
		Significant sign change	**-1.201 (0.594)***	212 = 0.98%

Do you think large datasets (like the ones we use in LLM pretraining) are this sensitive to change?

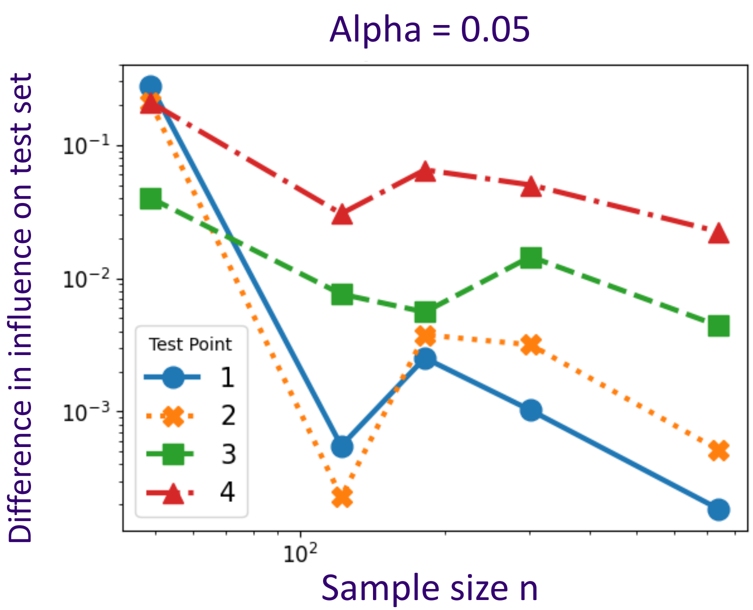
On average the removal of <.05% of the data changes sign of significance!!

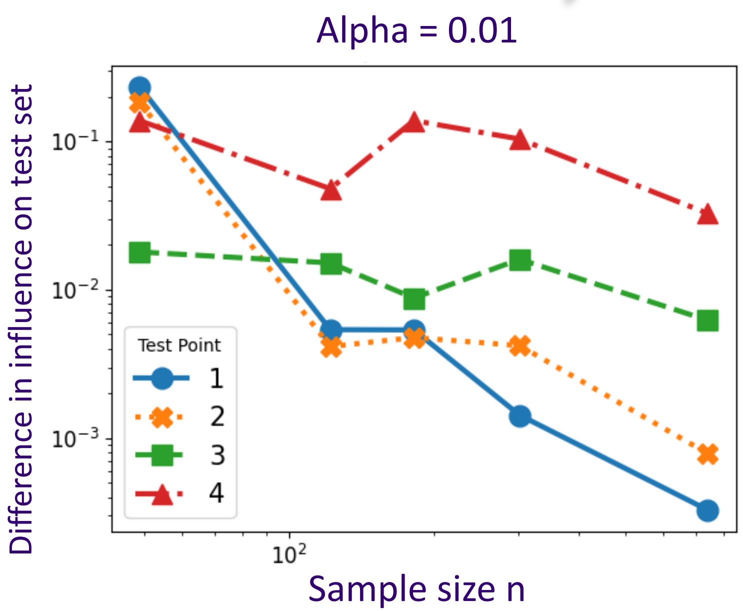
Experiment: Most Influential Subset

MIS (Question Answering)

- 4 different test points (questions/answer)
- $\alpha = 0.05,.0.1$ (size of subset)
- Arnoldi method was used to approximate influence

Why is there such variety in slopes?





- Downward trend —> similar to influence of 1 datapoint

Experiment: Most Influential Subset

z_{test}
Q: Where is Venera 9 found?
A: Venus

If the MIS was perfect, how could this be used to help in NLP tasks?

Is there alternative methods to MIS that identify influential sets?

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? …

Outline

- Background: Influential Points
- Statistical Finite Bound
- Computational Bound
  - *Discussion: Is there always meaning?*
- Most Influential Subset
  - *Discussion: Are all statistics a lie?!*
- **NLP Connection**
  - *Will there be influence in your future?*

Related Work in NLP

Influential points

- Leave one out training (data point importance)
- Saliency maps (token importance)
- Self-influence (Bejan et al. , 2023)
- Influence function for NLP…. Still in development

Machine Unlearning

- Quark - reinforcement learning (Lu et al., 2022)
- SISA Training (Kumar et al., 2022)

Explaining Black Box Predictions and Unveiling Data Artifacts through Influence Functions

Xiaochuang Han, Byron C. Wallace, Yulia Tsvetkov

INFLUENCE FUNCTIONS IN DEEP LEARNING ARE FRAGILE

Samyadeep Basu,* Phillip Pope *& Soheil Feizi
Department of Computer Science
University of Maryland, College Park
{sbasu12,pepope,sfeizi}@cs.umd.edu

AUG 27 2020

Influence Functions Do Not Seem to Predict Usefulness in NLP Transfer Learning

Author(s): Vid Kocijan and Samuel R. Bowman
Publication date: August 27 2020
Reviewer: Alex Wang
Editor: Kyunghyun Cho

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

Future Extension

- Non-convex/Non-smooth penalized M-estimation
- Application for toxicity/bias removal in NLP

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

1: $u_0 = 0, r_0 = -v - \mathrm{HVP}_n(u_0), d_0 = r_0$
2: **for** $t = 0, ..., T-1$ **do**
3: $\quad \alpha_t = \frac{d_t^\top r_t}{d_t^\top \mathrm{HVP}_n(d_t)}$
4: $\quad u_{t+1} = u_t + \alpha_t d_t$
5: $\quad r_{t+1} = -v - \mathrm{HVP}_n(u_{t+1})$
6: $\quad \beta_t = \frac{r_{t+1}^\top r_{t+1}}{r_t^\top r_t}$
7: $\quad d_{t+1} = r_{t+1} + \beta_t d_t$
8: **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 $\mathrm{HVP}(i, u) = \nabla^2 \ell(z_i, \theta_n) u$, number of epochs S, number of iterations per epoch T, learning rate γ

1: $u_T^{(0)} = 0$
2: **for** $s = 1, 2, ..., S$ **do**
3: $\quad u_0^{(s)} = u_T^{(s-1)}$
4: $\quad \tilde{u}_0^{(s)} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{HVP}(u_0^{(s)}) - v$
5: $\quad$ **for** $t = 0, ..., T-1$ **do**
6: $\quad\quad$ Sample $i_t \sim \mathrm{Unif}([n])$
7: $\quad\quad u_{t+1}^{(s)} = u_t^{(s)} - \gamma(\mathrm{HVP}(i_t, u_t^{(s)}) - \mathrm{HVP}(i_t, u_0^{(s)}) + \tilde{u}_0^{(s)})$
8: **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 $\text{HVP}_n(u) = H_n(\theta_n)u$, 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 Λ, G = ARNOLDI(u_0, T, k)                                ▷ Cache the results for future calls
return ⟨G∇h(θ), Λ^{-1}Gv⟩

procedure ARNOLDI(u_0, T, k)
    w_0 = 1 = u_0/‖u_0‖_2
    A = 0_{T+1×T}
    for t = 1, ..., T do
        Set u_t = HVP_n(w_t) − Σ_{j=1}^{t} ⟨u_t, w_j⟩ w_j
        Set A_{j,t} = ⟨u_t, w_j⟩ for j = 1, ..., t and A_{t+1,t} = ‖u_t‖_2
        Update w_{t+1} = u_t/‖u_t‖
    Set Ã = A[1 : T, :] ∈ ℝ^{T×T} (discard the last row)
    Compute an eigenvalue decomposition Ã = Σ_{j=1}^{T} λ_j e_j e_j^⊤ with λ_j's in descending order
    Define G : ℝ^p → ℝ^k as the operator Gu = (⟨u, W^⊤e_1⟩, ··· , ⟨u, W^⊤e_k⟩), where W = (w_1^⊤; ··· ; w_T^⊤) ∈ ℝ^{T×p}
    return diagonal matrix Λ = Diag(λ_1, ··· , λ_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***

Experiment: Most Influential Subset

1. What position did Víctor Vázquez Solsona play? Midfielder
2. Was Goldmoon male or female? Female
3. Where is Venera 9 found? Venus
4. In which language Nintendo La Rivista Ufficiale monthly football magazine reporting? Italian