Subsampling for massive data

HaiYing Wang

University of Connecticut

NCKU, December 20, 2023

Outline

1. Introduction

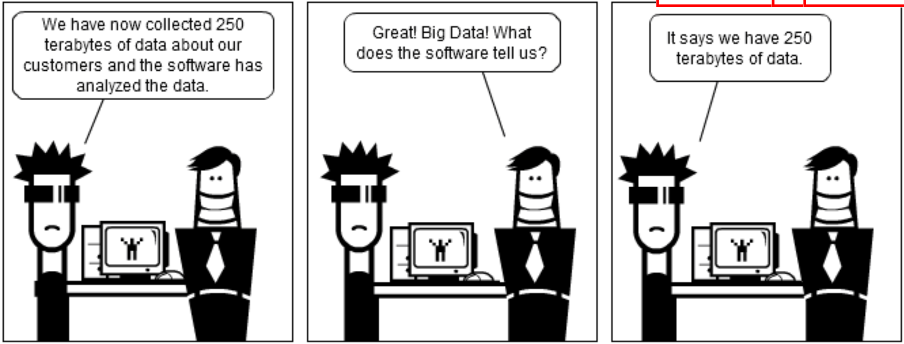
2. Approximate the full data estimator $\widehat{\boldsymbol{\theta}}_{\text{full}}$

3. Estimate the true population parameter $\boldsymbol{\theta}_t$
   - Noninformative random subsampling
   - Response-dependent random subsampling
     - Imbalanced data
   - Deterministic selection, Design based approaches

4. Prediction and other problems

Outline

Big Data challenge and data reduction



- Why not use a subset of the full data, say 250MB?
- Results from using the 250MB is not as precise as using the 250TB, but it gives some information.
- How to choose the 250MB data?

Introduction

- A common challenge from Big Data is how to extract useful information with limited computational costs.
- Subsampling is a commonly used technique to improve computational efficiency.
- It focuses on taking a small proportion of the big data and is a design problem by nature.
- Data-dependent subsampling often provides a better trade-off between computational efficiency and estimation efficiency than uniform subsampling.
- Optimal subsampling applies optimal design of experiments and finds a subsample that "minimize" the asymptotic variance of the resulting subsample estimator.

Notations (full data)

- Independent full data: $\mathcal{D}_N = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N \sim (\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}_t)$, where
  - $\boldsymbol{x}$ is the covariate variable,
  - $\boldsymbol{y}$ is the response variable,
  - $\boldsymbol{\theta} \in \mathbb{R}^d$ is the parameter of interest, and $\boldsymbol{\theta}_t$ is its true value.
- Estimate $\boldsymbol{\theta}$ by

$$\widehat{\boldsymbol{\theta}}_{\text{full}} = \arg\max_{\boldsymbol{\theta}} \left\{ \ell_N(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \ell(\boldsymbol{y}_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}) \right\}. \quad (1)$$

- There is often no closed-form solution to $\widehat{\boldsymbol{\theta}}_{\text{full}}$, and iterative calculations on the full data is not convenient for massive data.
- We want to use a subsample instead of the full data for computational feasibility.

Notations and basic question (subsample

- Let $\boldsymbol{\pi} = \{\pi_i\}_{i=1}^N$ such that $\pi_i \geq 0$ and $\sum_{i=1}^N \pi_i = 1$.
- A subsample taken according to $\boldsymbol{\pi}$: $\mathcal{D}_n^* = \{(\boldsymbol{x}_i^*, \boldsymbol{y}_i^*)\}_{i=1}^n$.
- Define subsample estimator, $\tilde{\boldsymbol{\theta}}$, using $\mathcal{D}_n^*$.
- Basic research question: choice $\boldsymbol{\pi}$ to make $\tilde{\boldsymbol{\theta}}$ "optimal".
- How to define $\tilde{\boldsymbol{\theta}}$?

$$? \qquad \tilde{\boldsymbol{\theta}}_{\text{ipw}} = \arg\max_{\boldsymbol{\theta}} \left\{ \ell^*_{\text{ipw}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \frac{\ell(\boldsymbol{y}_i^* \mid \boldsymbol{x}_i^*; \boldsymbol{\theta})}{N\pi_i^*} \right\} \tag{2}$$

$$? \qquad \tilde{\boldsymbol{\theta}}_{\text{uw}} = \arg\max_{\boldsymbol{\theta}} \left\{ \ell^*_{\text{uw}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{y}_i^* \mid \boldsymbol{x}_i^*; \boldsymbol{\theta}) \right\} \tag{3}$$

$$? \qquad \tilde{\boldsymbol{\theta}} = \text{something else} \tag{4}$$

Here, $\boldsymbol{\pi}$ should be easier to obtain or approximate than $\widehat{\boldsymbol{\theta}}_{\text{full}}$!

Basic sampling approaches

- Sampling with replacement
  - Subsample observations are i.i.d. conditionally on the full data.
  - It is fast to compute.
  - Use all sampling probabilities $\boldsymbol{\pi} = \{\pi_i\}_{i=1}^N$ simultaneously.
  - The subsample observations are not independent unconditionally.
- Poisson subsampling
  - The inclusion probability (often $n\pi_i$) depends only on $(\boldsymbol{x}_i, \boldsymbol{y}_i)$.
  - Subsample observations are independent unconditionally.
  - It is fast to compute.
  - Subsample sizes are random.
- Deterministic selection
  - No additional randomness in subsampling.
  - $N - n$ of π_i's are 0.
- Sampling without replacement for a fixed subsample size
  - Computationally slow.
- Sketching (extensions of subsampling)

Problems of interest

1. Parameters of interest
    1. Full data estimator $\widehat{\boldsymbol{\theta}}_{\text{full}}$
    2. Population parameter $\boldsymbol{\theta}_t$
    3. Prediction (?)
    4. Inference (?)
    5. ......
2. Randomnesses
    1. Randomness of the data
    2. Randomness of the subsampling
    3. Can the extra randomness be beneficial?

Subsampling problems may not be regul

Consider linear regression with N observations,

$$y_i = \beta_0 + \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_1 + \varepsilon_i, \ \varepsilon_i \sim N(0, \sigma^2), \ i = 1, ..., N. \tag{5}$$

If the subsample size n is fixed, can the variance of a noninformative (sampling rules do not involve the responses) subsample estimator go to zero?

[1] Wang, H., Yang, M., and Stufken, J. (2019). Information-based optimal subdata selection for big data linear regression. *JASA* **114**, 525, 393–405

Subsampling problems may not be regula

Consider linear regression with N observations,

$$y_i = \beta_0 + \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_1 + \varepsilon_i, \ \varepsilon_i \sim N(0, \sigma^2), \ i = 1, ..., N. \tag{5}$$

If the subsample size n is fixed, can the variance of a noninformative (sampling rules do not involve the responses) subsample estimator go to zero?

- The IBOSS estimator satisfies [1]

$$\mathbb{V}(\tilde{\beta}_j | \mathbf{X}) = O_P \left\{ \frac{p}{n(x_{(N)j} - x_{(1)j})^2} \right\}, \quad j = 1, ..., p. \tag{6}$$

- The variance goes to zero if the covariate distribution is not bounded.

[1] Wang, H., Yang, M., and Stufken, J. (2019). Information-based optimal subdata selection for big data linear regression. *JASA* **114**, 525, 393–405

An intriguing problem: center or not? [2]

$$y_i = \beta_0 + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma^2),\ i = 1, ..., N, \quad \text{where} \quad \beta_0 \neq 0. \quad (7)$$

- The ordinary least squares (OLS) estimator from a model without an intercept is biased.
- If we centered the data, the OLS estimator for a model without the intercept is unbiased.
- If a subsample is selected from a centered full data, the subsample is typically uncentered.
- Is it still appropriate to fit a model without an intercept?
- Should we recenter the subsample if we use a model without an intercept?

[2] Wang, H. (2022). A note on centering in subsample selection for linear regression. *Stat* **11**, 1, e525

An intriguing problem: center or not? [2]

$$y_i = \beta_0 + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma^2),\ i = 1, ..., N, \quad \text{where} \quad \beta_0 \neq 0. \quad (7)$$

- The ordinary least squares (OLS) estimator from a model without an intercept is biased.
- If we centered the data, the OLS estimator for a model without the intercept is unbiased.
- If a subsample is selected from a centered full data, the subsample is typically uncentered.
- Is it still appropriate to fit a model without an intercept?
- Should we recenter the subsample if we use a model without an intercept?

The OLS for a model without an intercept using uncentered subsample is unbiased and has a smaller variance.

[2] Wang, H. (2022). A note on centering in subsample selection for linear regression. *Stat* **11**, 1, e525

Outline

Approximate $\widehat{\boldsymbol{\theta}}_{\text{full}}$

Use the inverse provability weighted (IPW) estimator

$$\tilde{\boldsymbol{\theta}}_{\text{ipw}} = \arg\max \left\{ \ell^*_{\text{ipw}}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \frac{\ell(\boldsymbol{y}_i^* \mid \boldsymbol{x}_i^*; \boldsymbol{\theta})}{N\pi_i^*} \right\}. \tag{8}$$

- Under some regularity assumptions, for large n and N,

$$\tilde{\boldsymbol{\theta}}_{\text{ipw}} - \widehat{\boldsymbol{\theta}}_{\text{full}} \overset{a|\mathcal{D}_N}{\sim} \mathbb{N}(\mathbf{0},\ n^{-1}\mathbf{V}_{\tilde{\boldsymbol{\theta}}_{\text{ipw}}|\mathcal{D}_N}), \tag{9}$$

  where $\mathbf{V}_{\tilde{\boldsymbol{\theta}}_{\text{ipw}}|\mathcal{D}_N} = \mathbf{H}_N^{-1}\boldsymbol{\Lambda}_N(\boldsymbol{\pi})\mathbf{H}_N^{-1}$ and $\boldsymbol{\Lambda}_N(\boldsymbol{\pi})$ depends on $\boldsymbol{\pi}$.
- Here $\overset{a|\mathcal{D}_N}{\sim}$ is asymptotic conditional distribution given the full data.
- The randomness of the full data is not considered.

OSMAC [3]

- How to approximate $\widehat{\boldsymbol{\theta}}_{\text{full}}$ better?
- Make the approximation error $\tilde{\boldsymbol{\theta}}_{\text{ipw}} - \widehat{\boldsymbol{\theta}}_{\text{full}}$ small.
- Find $\boldsymbol{\pi} = \{\pi_i\}_{i=1}^N$ that minimizes $\mathbf{V}_{\tilde{\boldsymbol{\theta}}_{\text{ipw}}|\mathcal{D}_N}$.
- Here $\mathbf{V}_{\tilde{\boldsymbol{\theta}}_{\text{ipw}}|\mathcal{D}_N}$ is a matrix. We minimize its trace (A-optimality).
- We call it Optimal Subsampling Method under the A-optimality Criterion.
- If $\mathcal{D}_n^* = \{(\boldsymbol{x}_i^*, \boldsymbol{y}_i^*)\}_{i=1}^n$ are taken by sampling with replacement, the A-optimal probabilities are

$$\pi_i^{\text{osA}} = \frac{\|\mathbf{H}_N^{-1}\dot{\ell}_i(y_i \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}_{\text{full}})\|}{\sum_{j=1}^N \|\mathbf{H}_N^{-1}\dot{\ell}_j(y_j \mid \boldsymbol{x}_j; \widehat{\boldsymbol{\theta}}_{\text{full}})\|}. \tag{10}$$

- Since $\widehat{\boldsymbol{\theta}}_{\text{full}}$ is known, a pilot estimator is needed, say $\tilde{\boldsymbol{\theta}}_{\text{plt}}$.

[3] Wang, H., Zhu, R., and Ma, P. (2018). Optimal subsampling for large sample logistic regression. *JASA* **113**, 522, 829–844

Logistic regression (Wang *et al.*, 2018)

Consider the binary logistic regression model,

$$\mathbb{P}(y = 1 \mid \boldsymbol{x}; \boldsymbol{\theta}) = p(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}) = \frac{1}{1 + e^{-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}}}. \tag{11}$$

Informative subsampling probabilities:

1. A-optimality: minimize $\mathrm{tr}(\mathbf{V}_N)$:

$$\pi_i^{\mathrm{osA}} = \frac{\left|y_i - p(\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\theta}}_{\mathrm{plt}})\right| \|\tilde{\mathbf{H}}^{-1}\boldsymbol{x}_i\|}{\sum_{j=1}^{N}\left|y_j - p(\boldsymbol{x}_j^{\mathrm{T}}\tilde{\boldsymbol{\theta}}_{\mathrm{plt}})\right| \|\tilde{\mathbf{H}}^{-1}\boldsymbol{x}_j\|}. \tag{12}$$

2. L-optimality: minimize $\mathrm{tr}(\boldsymbol{\Lambda}_N)$:
$\sqrt{n}\mathbf{H}_n(\tilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{\mathrm{full}}) \overset{a|\mathcal{D}_N}{\sim} \mathbb{N}\{\mathbf{0},\ \boldsymbol{\Lambda}_N(\boldsymbol{\pi})\}$

$$\pi_i^{\mathrm{osL}} = \frac{\left|y_i - p(\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\theta}}_{\mathrm{plt}})\right| \|\boldsymbol{x}_i\|}{\sum_{j=1}^{N}\left|y_j - p(\boldsymbol{x}_j^{\mathrm{T}}\tilde{\boldsymbol{\theta}}_{\mathrm{plt}})\right| \|\boldsymbol{x}_j\|}. \tag{13}$$

Notes

$$\pi_i^{\text{osA}} \propto |y_i - p(\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\theta}}_{\text{plt}})| \|\tilde{\mathbf{H}}\boldsymbol{x}_i\|, \tag{14}$$

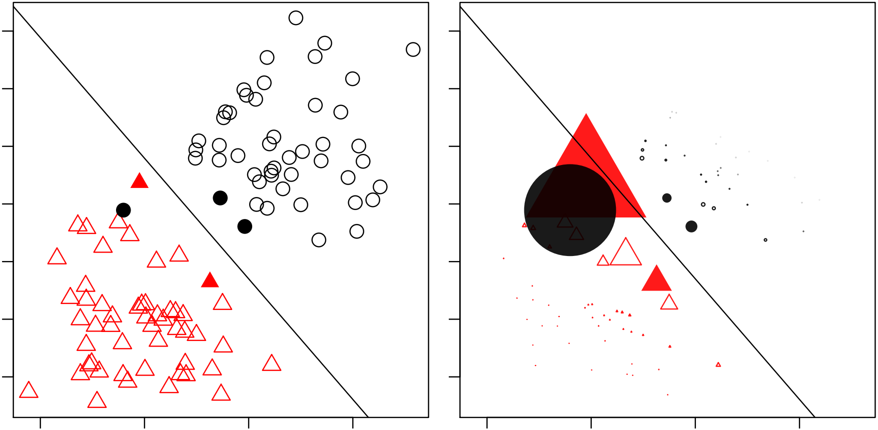
1. Covariate information represented by $\|\tilde{\mathbf{H}}\boldsymbol{x}_i\|$:
   - larger values of $\|\tilde{\mathbf{H}}\boldsymbol{x}_i\|$ indicates larger re-sampling probabilities.

2. Classification difficulty represented by $|y_i - p(\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\theta}}_{\text{plt}})|$
   - If $y_i = 0$;
   $$\pi_i^{\text{osA}} \propto p(\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\theta}}_{\text{plt}}) \tag{15}$$
   - If $y_i = 1$
   $$\pi_i^{\text{osA}} \propto 1 - p(\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\theta}}_{\text{plt}}) \tag{16}$$
   - OSMAC protects the separation problems; this echos the result of Silvapulle (JRSSB 1981)[4].

---

[4] Silvapulle, M. (1981). On the existence of maximum likelihood estimators for the binomial response models. *JRSSB* **43**, 3, 310–313

Connection with SVM

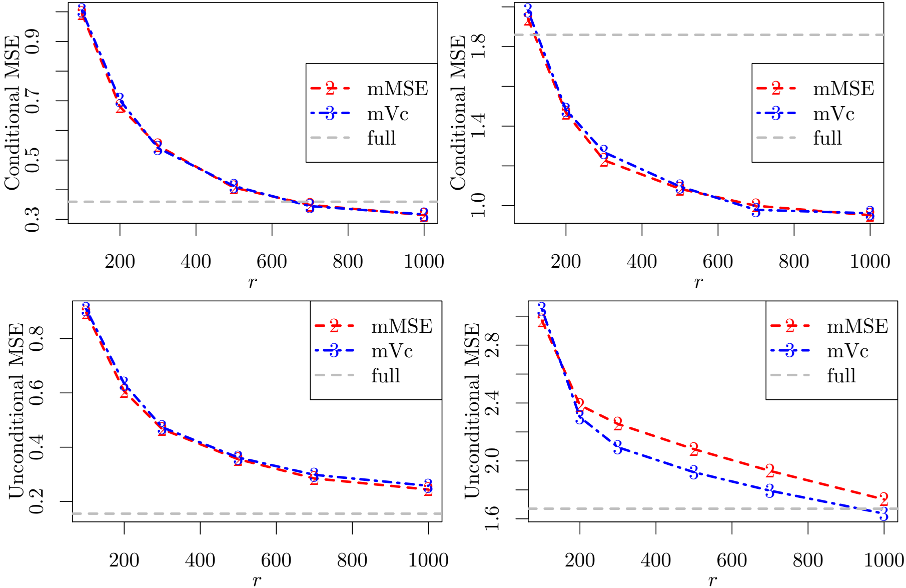


(a) Scatter plot (b) Bubble plot

Figure 1: Support vectors and optimal subsampling probabilities.

Simulation on very imbalanced responses

- Data of size $n = 10,000$ are generated from a logistic model.
- The covariate $\boldsymbol{x}$ follows a multivariate normal distribution.
- The responses are very imbalanced:
  1. 1.01% of the responses are 1's.
  2. 0.14% of the responses are 1's.



(a) 1.01% of y_i's are 1 (b) 0.14% of y_i's are 1

Figure 2: MSEs for rare event data.

Some related work

- Wang, H., Zhu, R., and Ma, P. (2018). Optimal subsampling for large sample logistic regression. *JASA* **113**, 522, 829–844
- Ting, D. and Brochu, E. (2018). Optimal subsampling with influence functions. In *NIPS*, 3650–3659
- Wang, H. and Ma, Y. (2021). Optimal subsampling for quantile regression in big data. *Biometrika* **108**, 1, 99–112
- Ai, M., Yu, J., Zhang, H., and Wang, H. (2021). Optimal subsampling algorithms for big data regressions. *Statistica Sinica* **31**, 2, 749–772
- Yu, J., Wang, H., Ai, M., and Zhang, H. (2022). Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. *JASA* **117**, 537, 265–276
- Keret, N. and Gorfine, M. (2023). Analyzing big EHR data—optimal Cox regression subsampling procedure with rare events. *JASA* 1–14
- ......

Better weights

Use other weights to replace $1/\pi_i^*$:

$$\tilde{\boldsymbol{\theta}}_{\text{w}} = \arg\max \left\{ \ell_{\text{w}}^*(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} w_i^* \ell(\boldsymbol{y}_i^* \mid \boldsymbol{x}_i^*; \boldsymbol{\theta}) \right\} \tag{17}$$

Empirical likelihood weighting:

- Fan, Y., Liu, Y., Liu, Y., and Qin, J. (2022). Nearly optimal capture-recapture sampling and empirical likelihood weighting estimation for m-estimation with big data. *arXiv preprint arXiv:2209.04569*
- Liu, Y. and Fan, Y. (2023). Biased-sample empirical likelihood weighting for missing data problems: an alternative to inverse probability weighting. *JRSSB* **85**, 1, 67–83

Outline

1. Introduction
2. Approximate the full data estimator $\widehat{\boldsymbol{\theta}}_{\text{full}}$
3. Estimate the true population parameter $\boldsymbol{\theta}_t$
   - Noninformative random subsampling
   - Response-dependent random subsampling
   - Deterministic selection, Design based approaches
4. Prediction and other problems

Outline

1. Introduction
2. Approximate the full data estimator $\widehat{\boldsymbol{\theta}}_{\text{full}}$
3. Estimate the true population parameter $\boldsymbol{\theta}_t$
   - Noninformative random subsampling
4. Prediction and other problems

Estimate the true $\boldsymbol{\theta}_t$

It is critical that the model is "correctly" specified!

- Use the unweighted estimator if $\boldsymbol{\pi}$ **does not dependent on** $\boldsymbol{y}_i$**'s**.

$$\tilde{\boldsymbol{\theta}}_{\text{uw}} = \arg\max_{\boldsymbol{\theta}} \left\{ \ell^*_{\text{uw}}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{y}_i^* \mid \boldsymbol{x}_i^*; \boldsymbol{\theta}) \right\} \tag{18}$$

- For large n and N,

$$\tilde{\boldsymbol{\theta}}_{\text{uw}} - \widehat{\boldsymbol{\theta}}_{w\text{full}} \overset{a|\mathcal{D}_N}{\sim} \mathbb{N}(\mathbf{0},\ n^{-1}\mathbf{V}_{\tilde{\boldsymbol{\theta}}_{\text{uw}}|\mathcal{D}_N}), \tag{19}$$

- $\widehat{\boldsymbol{\theta}}_{w\text{full}}$ is a weighted full data estimator, often less efficient than $\widehat{\boldsymbol{\theta}}_{\text{full}}$ in estimating $\boldsymbol{\theta}_t$.
- The unweighted estimator $\tilde{\boldsymbol{\theta}}_{\text{uw}}$ approximates a less efficient full data estimator.

Weighted vs unweighted estimators

$$
\begin{array}{lc|ccc|c}
\text{true parameter} & & & \boldsymbol{\theta}_t & & \\
\text{full data:} & \widehat{\boldsymbol{\theta}}_{\text{full}} & N^{-1}\mathbf{V}_{\widehat{\boldsymbol{\theta}}_{\text{full}}} & \leq & N^{-1}\mathbf{V}_{\widehat{\boldsymbol{\theta}}_{w\text{full}}} & \widehat{\boldsymbol{\theta}}_{w\text{full}} \\
 & \Uparrow & & & & \Uparrow \\
\text{subsample:} & \tilde{\boldsymbol{\theta}}_{\text{ipw}} & n^{-1}\mathbf{V}_{\tilde{\boldsymbol{\theta}}_{\text{ipw}}|\mathcal{D}_N} & \geq & n^{-1}\mathbf{V}_{\tilde{\boldsymbol{\theta}}_{\text{uw}}|\mathcal{D}_N} & \tilde{\boldsymbol{\theta}}_{\text{uw}}
\end{array}
$$

- The unweighted estimator $\tilde{\boldsymbol{\theta}}_{\text{uw}}$ approximates a less efficient full data estimator of $\boldsymbol{\theta}_t$.
- The weighted estimator $\tilde{\boldsymbol{\theta}}_{\text{ipw}}$ approximates a more efficient full data estimator of $\boldsymbol{\theta}_t$.
- Since $n \ll N$ in big data subsampling, $\tilde{\boldsymbol{\theta}}_{\text{uw}}$ is often more efficient than $\tilde{\boldsymbol{\theta}}_{\text{ipw}}$ in estimating $\boldsymbol{\theta}_t$.
- What is the unconditional variance of $\tilde{\boldsymbol{\theta}}_{\text{uw}}$ or $\tilde{\boldsymbol{\theta}}_{\text{ipw}}$?
- What is $\tilde{\boldsymbol{\theta}}_{\text{uw}}$?
- What if $\boldsymbol{\pi}$ depend on $\boldsymbol{y}_i$'s?

Unconditional distributions

- Approximate $\widehat{\boldsymbol{\theta}}_{\text{full}}$:
  - Considering the conditional distribution is sufficient.
  - Randomness of the full data is not important.
  - Model does not have to be correct.
- Estimate $\boldsymbol{\theta}_t$:
  - Need to consider the unconditional distribution, "unless $n/N = o(1)$".
  - Randomness of the full data is relevant.
  - Model correctness is critical.

Unconditional asymptotic distribution (sampling with replacement)[5]:

$$\tilde{\boldsymbol{\theta}}_{\text{ipw}} - \boldsymbol{\theta}_t \overset{a}{\sim} \mathbb{N}(\mathbf{0},\ n^{-1}\mathbf{V}_{\tilde{\boldsymbol{\theta}}_{\text{ipw}}|\mathcal{D}_N} + N^{-1}\mathbf{V}_{\widehat{\boldsymbol{\theta}}_{\text{full}}}) \tag{20}$$

$$\tilde{\boldsymbol{\theta}}_{\text{uw}} - \boldsymbol{\theta}_t \overset{a}{\sim} \mathbb{N}(\mathbf{0},\ n^{-1}\mathbf{V}_{\tilde{\boldsymbol{\theta}}_{\text{uw}}|\mathcal{D}_N} + N^{-1}\mathbf{V}_{\widehat{\boldsymbol{\theta}}_{w\text{full}}}) \tag{21}$$

[5] Wang, J., Zou, J., and Wang, H. (2022b). Sampling with replacement vs poisson sampling: A comparative study in optimal subsampling. *IEEE Transactions on Information Theory* **68**, 10, 6605–6630

Intuition on $\tilde{\boldsymbol{\theta}}_{\mathrm{uw}}$ and what is it?

$$\tilde{\boldsymbol{\theta}}_{\mathrm{ipw}} = \arg\max_{\boldsymbol{\theta}} \left\{ \ell^*_{\mathrm{ipw}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\ell(\boldsymbol{y}_i^* \mid \boldsymbol{x}_i^*; \boldsymbol{\theta})}{N\pi_i^*} \right\} \tag{22}$$

$$\tilde{\boldsymbol{\theta}}_{\mathrm{uw}} = \arg\max_{\boldsymbol{\theta}} \left\{ \ell^*_{\mathrm{uw}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{y}_i^* \mid \boldsymbol{x}_i^*; \boldsymbol{\theta}) \right\} \tag{23}$$

- $\tilde{\boldsymbol{\theta}}_{\mathrm{ipw}}$ down-weights more informative data points.
- $\tilde{\boldsymbol{\theta}}_{\mathrm{uw}}$ does not penalize more informative data points.
- With sampling with replacement, no exact interpretation for $\tilde{\boldsymbol{\theta}}_{\mathrm{uw}}$.
- With Poisson sampling or deterministic selection,
  - $\tilde{\boldsymbol{\theta}}_{\mathrm{uw}}$ is the maximum (**conditional**) likelihood estimator (MLE) based on the subsample;
  - The distribution of $\boldsymbol{y}^* \mid \boldsymbol{x}^*$ is the same as that of $\boldsymbol{y} \mid \boldsymbol{x}$, if sampling may only depend on $\mathbf{X}$.

Sampling with replacement v.s. Poisson

Under some regularity assumptions, for large n and N,

$$\tilde{\boldsymbol{\theta}}_{\text{ipw}}^{\text{swr}} - \boldsymbol{\theta}_t \overset{a}{\sim} \mathbb{N}(\mathbf{0},\ n^{-1}\mathbf{V}_{\tilde{\boldsymbol{\theta}}_{\text{ipw}}|\mathcal{D}_N}^{\text{swr}} + N^{-1}\mathbf{V}_{\widehat{\boldsymbol{\theta}}_{\text{full}}}) \tag{24}$$

$$\tilde{\boldsymbol{\theta}}_{\text{ipw}}^{\text{poi}} - \boldsymbol{\theta}_t \overset{a}{\sim} \mathbb{N}(\mathbf{0},\ n^{-1}\mathbf{V}_{\tilde{\boldsymbol{\theta}}_{\text{ipw}}|\mathcal{D}_N}^{\text{poi}}) \tag{25}$$

- $\mathbf{V}_{\tilde{\boldsymbol{\theta}}_{\text{ipw}}|\mathcal{D}_N}^{\text{swr}} - \mathbf{V}_{\tilde{\boldsymbol{\theta}}_{\text{ipw}}|\mathcal{D}_N}^{\text{poi}} = o_P(1)$, if $n = o(N)$.
- $\mathbf{V}_{\tilde{\boldsymbol{\theta}}_{\text{ipw}}|\mathcal{D}_N}^{\text{swr}} - \mathbf{V}_{\widehat{\boldsymbol{\theta}}_{\text{full}}} \neq o_P(1)$, regardless the relative rates of n and N.
- $\mathbf{V}_{\tilde{\boldsymbol{\theta}}_{\text{ipw}}|\mathcal{D}_N}^{\text{poi}} - \mathbf{V}_{\widehat{\boldsymbol{\theta}}_{\text{full}}} = o_P(1)$, if $N - n \to 0$. $n/N \to 1$?
- $\tilde{\boldsymbol{\theta}}_{\text{ipw}}^{\text{poi}}$ is better than $\tilde{\boldsymbol{\theta}}_{\text{ipw}}^{\text{swr}}$ when $n \asymp N$.

Relevant work

- Zhang, T., Ning, Y., and Ruppert, D. (2021). Optimal sampling for generalized linear models under measurement constraints. *JCGS* **30**, 1, 106–114

- Wang, J., Wang, H., and Xiong, S. (2022a). Unweighted estimation based on optimal sample under measurement constraints. *Canadian Journal of Statistics* **n/a**, n/a, `https://doi.org/10.1002/cjs.11753`

- Wang, J., Zou, J., and Wang, H. (2022b). Sampling with replacement vs poisson sampling: A comparative study in optimal subsampling. *IEEE Transactions on Information Theory* **68**, 10, 6605–6630

Outline

1. Introduction
2. Approximate the full data estimator $\widehat{\boldsymbol{\theta}}_{\text{full}}$
3. Estimate the true population parameter $\boldsymbol{\theta}_t$
   - Response-dependent random subsampling
     - Imbalanced data
4. Prediction and other problems

When $\boldsymbol{\pi}$ depend on $\boldsymbol{y}_i$

- The subsample is biased; $\boldsymbol{y}^* \mid \boldsymbol{x}^*$ and $\boldsymbol{y} \mid \boldsymbol{x}$ have different distributions.
- A naive unweighted estimator is biased and inconsistent.
- Can we avoid IPW and still have an asymptotically unbiased estimator?

Specific results on logistic regression

1. Consider the binary logistic regression model,

$$\mathbb{P}(y=1 \mid \boldsymbol{x};\boldsymbol{\theta}) = p(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}) = \frac{1}{1+e^{-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}}}. \tag{26}$$

2. The informative subsampling probabilities satisfy

$$\pi(\boldsymbol{x}_i, y_i) \propto \left| y_i - p(\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\theta}}_{\mathrm{plt}}) \right| h(\boldsymbol{x}_i). \tag{27}$$

Use the subsample to calculate the unweighted estimator $\tilde{\boldsymbol{\theta}}_{\mathrm{uw}}$,

$$\tilde{\boldsymbol{\theta}}_{\mathrm{uw}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \delta_i \left\{ y_i \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta} - \log\left(1 + e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}}\right) \right\}. \tag{28}$$

Correct the bias with

$$\check{\boldsymbol{\theta}}_{\mathrm{uw}} = \tilde{\boldsymbol{\theta}}_{\mathrm{uw}} + \tilde{\boldsymbol{\theta}}_{\mathrm{plt}}. \tag{29}$$

The $\check{\boldsymbol{\theta}}_{\mathrm{uw}}$ is asymptotically unbiased and more efficient (Wang, 2019):

$$\mathbb{V}_a(\check{\boldsymbol{\theta}}_{\mathrm{uw}}) \le \mathbb{V}_a(\check{\boldsymbol{\theta}}_{\mathrm{ipw}}). \tag{30}$$

Unweighted with bias correction

- What is $\check{\boldsymbol{\theta}}_{\mathrm{uw}}$? It is the maximum sampled conditional likelihood estimator shown in (34) later in the slides.
- The bias correction in (29) only works for binary logistic regression with the specific form of $\pi(\boldsymbol{x}, y)$.

Big binary imbalanced data [6] [7]

- Let $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ be training data that satisfies

$$\mathbb{P}(y = 1 \mid \boldsymbol{x}) = p(\boldsymbol{x}; \boldsymbol{\theta}). \tag{31}$$

- Let N_1 be the number of ones, and N_0 be the number of zeros.
- For very imbalanced data, $N_1 \ll N_0$, it is more appropriate to assume that N_1 increases in a slower rate compared with N_0,

$$\frac{N_1}{N_0} \xrightarrow{P} 0 \quad \text{and} \quad N_1 \xrightarrow{P} \infty \qquad \text{as} \quad N \to \infty.$$

- This requires $\mathbb{P}(y = 1) \to 0$ as $N \to \infty$ on the model side.

[6] Wang, H. (2020). Logistic regression for massive data with rare events. In *ICML*

[7] Wang, H., Zhang, A., and Wang, C. (2021a). Nonuniform negative sampling and log odds correction with rare events data. In *NeurIPS*

Model that allows $\mathbb{P}(y = 1) \to 0$

- Let $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$ and write the log odds as

$$g(\boldsymbol{x}; \boldsymbol{\theta}) := \log\left\{\frac{p(\boldsymbol{x}; \boldsymbol{\theta})}{1 - p(\boldsymbol{x}; \boldsymbol{\theta})}\right\} = \alpha + f(\boldsymbol{x}; \boldsymbol{\beta})$$

- Here $f(\boldsymbol{x}; \boldsymbol{\beta})$ is a smooth function of $\boldsymbol{\beta}$, such as a neural net.
- Assume that $\alpha_t \to -\infty$ as $N \to \infty$ and $\boldsymbol{\beta}_t$ is fixed.
- A diverging α_t and a fixed $\boldsymbol{\beta}_t$ indicates that the both the marginal and conditional probabilities for a positive instance are small.
- This means a covariate change does not convert a small-probability-event to a large-probability-event.

How much information do we really have

Under some moment assumptions, as $N \to \infty$,

$$\sqrt{N_1}(\widehat{\boldsymbol{\theta}}_{\mathrm{f}} - \boldsymbol{\theta}_t) \xrightarrow{D} \mathbb{N}(\mathbf{0},\ \mathbf{V}_{\mathrm{f}}).$$

Table 1: Numerical illustration

	Correct model			Mis-sprcified model		
$(N,\ N_1^a)$	$\mathrm{tr}(\hat{\mathbf{V}}_{\mathrm{e}})$	$N_1^a\mathrm{tr}(\hat{\mathbf{V}}_{\mathrm{e}})$	$N\mathrm{tr}(\hat{\mathbf{V}}_{\mathrm{e}})$	$\mathrm{tr}(\hat{\mathbf{V}}_{\mathrm{e}})$	$N_1^a\mathrm{tr}(\hat{\mathbf{V}}_{\mathrm{e}})$	$N\mathrm{tr}(\hat{\mathbf{V}}_{\mathrm{e}})$
$(10^3,\ 32)$	0.169	5.41	169.17	0.969	30.99	968.70
$(10^4,\ 64)$	0.097	6.20	969.29	0.322	20.59	3217.12
$(10^5,\ 128)$	0.045	5.76	4497.24	0.135	17.32	13527.60
$(10^6,\ 256)$	0.018	4.62	18048.40	0.046	11.74	45847.40

Here, $\hat{\mathbf{V}}_{\mathrm{e}}$ is the empirical variance of $\widehat{\boldsymbol{\theta}}_{\mathrm{f}}$ and $N_1^a = \mathbb{E}(N_1)$.

General negative sampling algorithm

Algorithm 1 Negative sampling

For $i = 1, ..., N$:

1. if $y_i = 1$, record $\{\boldsymbol{x}_i, y_i, \pi(\boldsymbol{x}_i, y_i) = 1\}$ in the sample;
2. if $y_i = 0$, with probability $\pi(\boldsymbol{x}_i, y_i)$,
   include $\{\boldsymbol{x}_i, y_i, \pi(\boldsymbol{x}_i, y_i) = \rho\varphi(\boldsymbol{x}_i)\}$ in the sample.

- ρ: sampling rate on the negative class.
- $\varphi(\boldsymbol{x}) > 0$: a function with $\mathbb{E}\{\varphi(\boldsymbol{x})\} = 1$.

Note: selected subsamples are biased!

Inverse probability weighting (IPW)

Under some moment assumptions,

$$\sqrt{N_1}(\widehat{\boldsymbol{\theta}}_{\mathrm{ipw}} - \boldsymbol{\theta}_t) \xrightarrow{D} \mathbb{N}(\mathbf{0},\ \mathbf{V}_{\mathrm{ipw}}),$$

where $\mathbf{V}_{\mathrm{ipw}} = \mathbf{V}_{\mathrm{f}} + \mathbf{V}_{\mathrm{sub}}$.

- $\mathbf{V}_{\mathrm{sub}} = \mathbf{0}$ if $\dfrac{N_1}{N_0\rho} \to 0$ $(c = 0)$:
  - No asymptotic efficiency loss.
  - No need to design better sampling function.
- $\mathbf{V}_{\mathrm{sub}} > \mathbf{0}$ if $\lim \dfrac{N_1}{N_0\rho} > 0$ $(c > 0)$:
  - Variance inflation due to subsampling.
  - **A well designed sampling function $\varphi(\boldsymbol{x})$ is useful.**

A real application case

- An online recommendation system has over 10 billion impressions each day, but only about 1.25% are clicked (zeros/ones ≈ 80:1.).
- Due to limited storage and computational resources, the goal is to reduce zeros/ones to 4:1, i.e., keep about 5% of the data.

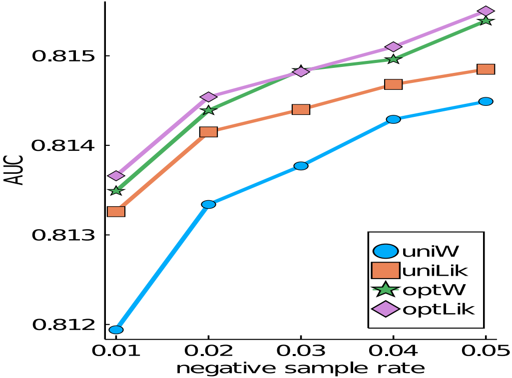


Figure 3: Testing AUC of subsample estimators (the larger the better).

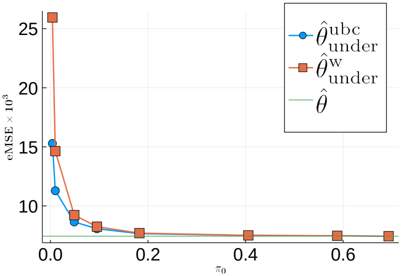
Subsampling and Oversampling together

Oversample the positives for $v_{\boldsymbol{x}}$ times, where $v_{\boldsymbol{x}} \mid \boldsymbol{x} \sim \mathbb{POI}\{\lambda_N(\boldsymbol{x})\}$.
Under some moment assumptions,

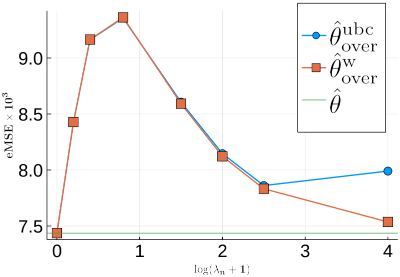
$$\sqrt{N_1}(\widehat{\boldsymbol{\theta}}_{\mathrm{ipw}} - \boldsymbol{\theta}_t) \xrightarrow{D} \mathbb{N}(\mathbf{0},\ \mathbf{V}_{\mathrm{ipw}}),$$

where $\mathbf{V}_{\mathrm{ipw}} = \mathbf{V}_{\mathrm{f}} + \mathbf{V}_{\mathrm{sub}} + \mathbf{V}_{\mathrm{over}}$.

- Oversampling reduce the estimation efficiency!



(a) MSE for Under-Sampling



(b) MSE for Over-Sampling

Figure 4: MSEs of under-sampled and over-sampled estimators.

Nonuniform log odds correction

- Let $\delta \in \{0,1\}$ be the indicator that $(\boldsymbol{x}, y)$ is included in the subsample.
- Conditional on $(\boldsymbol{x}, y)$, δ is Bernoulli with

$$\mathbb{P}(\delta = 1 \mid \boldsymbol{x}, y) = \pi(\boldsymbol{x}, y).$$

- By Bayes' theorem, conditional on $\{\delta = 1\}$, the probability

$$\mathbb{P}(y = 1 \mid \boldsymbol{x}, \delta = 1) = \frac{1}{1 + e^{-\{g(\boldsymbol{x};\boldsymbol{\theta}) + l\}}}, \tag{32}$$

where

$$l = \log\left\{\frac{\pi(\boldsymbol{x}, 1)}{\pi(\boldsymbol{x}, 0)}\right\} \tag{33}$$

- This gives the distribution of $y \mid \boldsymbol{x}$ for the subsample, which allows the conditional likelihood estimator.

Linear logistic regression

For the special case of logistic regression:

- $g(\boldsymbol{x}_i; \boldsymbol{\theta}) = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\theta}$;
- if $\pi_{\mathrm{lcc}}(\boldsymbol{x}_i, y_i) \propto |y_i - p(\boldsymbol{x}_i; \tilde{\boldsymbol{\theta}}_{\mathrm{plt}})|$, then $l_i = \boldsymbol{x}_i^{\mathrm{T}} \tilde{\boldsymbol{\theta}}_{\mathrm{plt}}$;
- The conditional log-likelihood $\ell_{\mathrm{lik}}(\boldsymbol{\theta})$ is

$$\ell_{\mathrm{lik}}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \delta_i \left[y_i \boldsymbol{x}_i^{\mathrm{T}} (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}_{\mathrm{plt}}) - \log \left\{ 1 + e^{\boldsymbol{x}_i^{\mathrm{T}} (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}_{\mathrm{plt}})} \right\} \right]. \qquad (34)$$

- The unweighted estimator with bias correction in (29) is a special case of the likelihood based estimator.

Sampled data conditional likelihood [8]

By Bayes' theorem, the density function for sampled data is

$$f(y_i \mid \boldsymbol{x}_i, \delta_i = 1; \boldsymbol{\theta}) = \frac{f(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta})\pi(\boldsymbol{x}_i, y_i)}{\int f(y \mid \boldsymbol{x}_i; \boldsymbol{\theta})\pi(\boldsymbol{x}_i, y)\mathrm{d}y}. \tag{35}$$

Thus, for the sampled data, the conditional log-likelihood function is

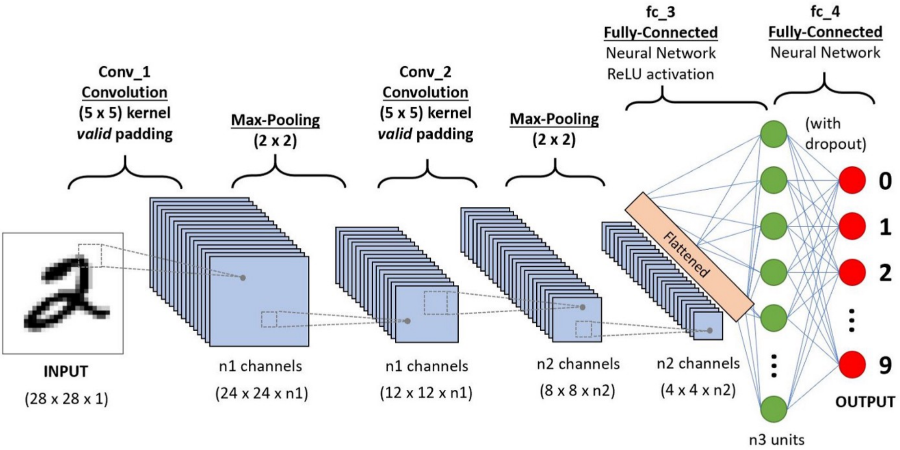
$$\ell_{\text{lik}}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \delta_i \left\{ \log f(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}) - \log \int f(y \mid \boldsymbol{x}_i; \boldsymbol{\theta})\pi(\boldsymbol{x}_i, y)\mathrm{d}y \right\} + C, \tag{36}$$

where C does not contain $\boldsymbol{\theta}$.

[8] Wang, H. and Kim, J. K. (2022). Maximum sampled conditional likelihood for informative subsampling. *JMLR* **23**, 332, 1–50

Application to the MNIST data

The convolutional neural network LeNet-



- The model has 44,426 parameters.

Results

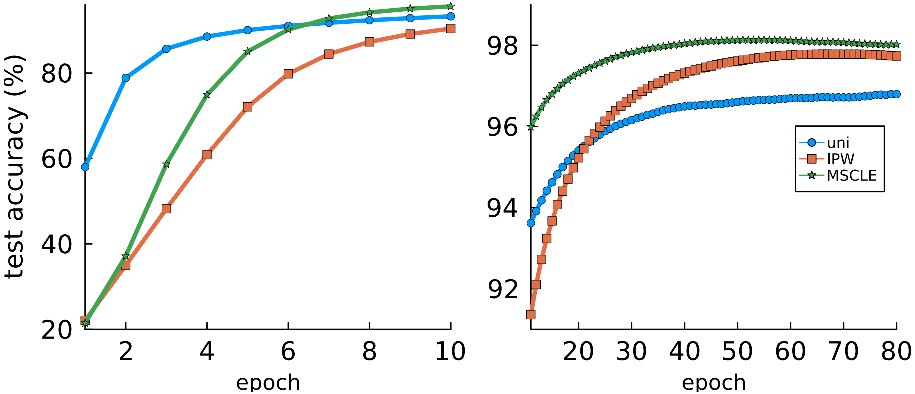


Figure 5: Classification accuracy (in percentage) on the test data against epoch in the training using subsamples of size $n = 5,000$ from the MNIST data.

Remarks on optimality

- Optimal probabilities are only defined for the IPW estimator $\tilde{\boldsymbol{\theta}}_{\text{ipw}}$.
- Achieve optimality for $\tilde{\boldsymbol{\theta}}_{\text{ipw}}$ in practice?
  - Optimal probabilities depend on unknowns, $\pi(\boldsymbol{x}, y) = \pi(\boldsymbol{x}_i, y_i; \boldsymbol{\vartheta})$.
  - Even $\boldsymbol{\vartheta}$ is consistently estimated, $\tilde{\boldsymbol{\theta}}_{\text{ipw}}$ may not have optimal variance (Wang *et al.*, 2021a, 2022b),
  - because $\pi(\boldsymbol{x}_i, y_i; \boldsymbol{\vartheta})$ is in the denominator of the target function.
- For $\tilde{\boldsymbol{\theta}}_{\text{uw}}$, $\check{\boldsymbol{\theta}}_{\text{uw}}$, and $\tilde{\boldsymbol{\theta}}_{\text{lik}}$,
  - subsamples are not optimal for these estimators;
  - they are less sensitive to variations in $\tilde{\boldsymbol{\vartheta}}_{\text{plt}}$.
- Can we derive optimal probabilities for $\tilde{\boldsymbol{\theta}}_{\text{lik}}$?
  - In general, no!
  - For noninformative sampling ($\tilde{\boldsymbol{\theta}}_{\text{uw}}$),
    - optimal probabilities are either 0 or 1, i.e., $\pi(\boldsymbol{x}_i, y) \in \{0, 1\}$;
    - the problem becomes a deterministic design problem!
    - the correctness of model assumptions are crucial;
    - optimal design to appropriate the full data estimator?

Fithian, W. and Hastie, T. (2014). Local case-control sampling: Efficient subsampling in imbalanced data sets. *Annals of statistics* **42**, 5, 1693

Wang, H. (2019). More efficient estimation for logistic regression with optimal subsamples. *JMLR* **20**, 132, 1–59

Han, L., Tan, K. M., Yang, T., Zhang, T., *et al.* (2020). Local uncertainty sampling for large-scale multiclass logistic regression. *Annals of Statistics* **48**, 3, 1770–1788

Wang, H., Zhang, A., and Wang, C. (2021a). Nonuniform negative sampling and log odds correction with rare events data. In *NeurIPS*

Wang, H. and Kim, J. K. (2022). Maximum sampled conditional likelihood for informative subsampling. *JMLR* **23**, 332, 1–50

Outline

1. Introduction
2. Approximate the full data estimator $\widehat{\boldsymbol{\theta}}_{\text{full}}$
3. Estimate the true population parameter $\boldsymbol{\theta}_t$
   - Deterministic selection, Design based approaches
4. Prediction and other problems

D-optimality motivated IBOSS algorithm

Assume linear regression:

$$y_i = \beta_0 + \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_1 + \varepsilon_i, \quad i = 1, ..., N. \tag{37}$$

1. Use a partition-based selection algorithm.
2. For each covariate, include $r = [N/(2p)]$ data points with the smallest covariate values and r data points with the largest covariate values to the subdata.
3. Exclude data points that were previously selected.

For the selected subdata $(\mathbf{X}_D^*, \boldsymbol{y}_D^*)$, use the OLS

$$\hat{\boldsymbol{\beta}}^D = \{(\mathbf{X}_D^*)^{\mathrm{T}} \mathbf{X}_D^*\}^{-1} (\mathbf{X}_D^*)^{\mathrm{T}} \boldsymbol{y}_D^*.$$

[9] Wang, H., Yang, M., and Stufken, J. (2019). Information-based optimal subdata selection for big data linear regression. *JASA* **114**, 525, 393–405

Orders of variances of $\hat{\boldsymbol{\beta}}^D$

Methods	Covariate $\boldsymbol{x} \sim t_\nu$		
	β_0	β_1	
		$\nu \geq 3$	$\nu < 3$
IBOSS	$\frac{1}{n}$	$\frac{1}{n\mathbf{N}^{\mathbf{2}/\nu}}$	$\frac{1}{n\mathbf{N}^{\mathbf{2}/\nu}}$
UNIF	$\frac{1}{n}$	$\frac{1}{n}$	slower than $\frac{1}{n\mathbf{N}^{(\mathbf{2}/\nu-\mathbf{1}+\alpha)}}$ for any $\alpha > 0$
FULL	$\frac{1}{N}$	$\frac{1}{N}$	slower than $\frac{1}{\mathbf{N}^{(\mathbf{2}/\nu+\alpha)}}$ for any $\alpha > 0$

If $\boldsymbol{x}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \boldsymbol{\Phi}\boldsymbol{\rho}\boldsymbol{\Phi}$,

$$\mathbb{V}\left(\hat{\boldsymbol{\beta}}^D \middle| \mathbf{X}\right) = \begin{bmatrix} \frac{\sigma^2}{n} & \mathbf{0} \\ \mathbf{0} & \frac{1}{\log \mathbf{N}} \frac{p\sigma^2}{2n} (\boldsymbol{\Phi}\boldsymbol{\rho}^2\boldsymbol{\Phi})^{-1} \end{bmatrix} + O_P \begin{bmatrix} \frac{1}{\sqrt{\log N}} & \frac{1}{\log N} \\ \frac{1}{\log N} & \frac{1}{(\log N)^{3/2}} \end{bmatrix}.$$

Some related literature

Wang, L., Elmstedt, J., Wong, W. K., and Xu, H. (2021b). Orthogonal subsampling for big data linear regression. *The Annals of Applied Statistics* **15**, 3, 1273–1290

Joseph, V. R. and Mak, S. (2021). Supervised compression of big data. *Statistical Analysis and Data Mining* **14**, 3, 217–229

Pronzato, L. and Wang, H. (2021). Sequential online subsampling for thinning experimental designs. *JSPI* **212**, 169 – 193

Yu, J., Ai, M., and Ye, Z. (2023a). A review on design inspired subsampling for big data. *Statistical Papers*

Yu, J., Liu, J., and Wang, H. (2023b). Information-based optimal subdata selection for non-linear models. *Statistical Papers* **64**, 4, 1069–1093

Reuter, T. and Schwabe, R. (2023). D-optimal subsampling design for massive data linear regression. *arXiv preprint arXiv:2307.02236*

Outline

1. Introduction
2. Approximate the full data estimator $\widehat{\boldsymbol{\theta}}_{\text{full}}$
3. Estimate the true population parameter $\boldsymbol{\theta}_t$
4. Prediction and other problems

Estimation vs Prediction

- Most existing work focus on estimation.
- The prediction error with a linear regression

$$\mathbb{E}\{(y_{new} - \hat{y}_{new})^2\} \tag{38}$$
$$= \mathbb{E}[\{y_{new} - \mathbb{E}(y_{new})\}^2] \quad + \quad \mathbb{E}[\{\mathbb{E}(y_{new}) - \hat{y}_{new}\}^2] \tag{39}$$
$$= \sigma^2 \quad + \quad \mathbb{E}[\{\boldsymbol{x}_{new}^{\mathrm{T}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\}^2] \tag{40}$$

model error estimation error

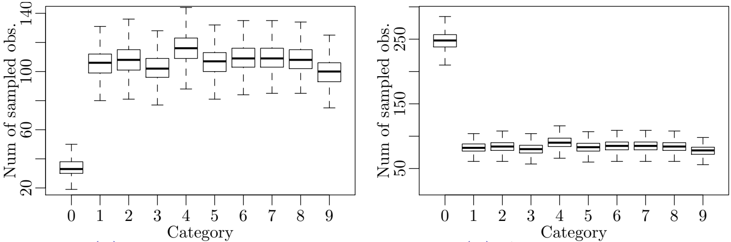
MSPE

- The variance $\mathbb{V}(y_{new}) = \sigma^2$ is the dominating term, and it cannot be reduced by a better subsample.
- Focusing on the MSPE, $\mathbb{E}[\{\boldsymbol{x}_{new}^{\mathrm{T}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\}^2]$ is essentially estimating of the mean responses.
- Focusing on the MSPE does have benefits.

Softmax regression [10]

$$\mathbb{P}(y = k|\boldsymbol{x}_i) = p_k(\boldsymbol{x}, \boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_k)}{\sum_{l=0}^{K}\exp(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_l)}, \quad k = 0, 1, ..., K.$$

- Not all $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_k$ are estimable, because $\sum p_k(\boldsymbol{x}, \boldsymbol{\beta}) = 1$.
- The baseline constraint for identifiability assumes $\boldsymbol{\beta}_0 = \mathbf{0}$.



(a) L-optimality (b) A-optimality

Figure 6: "Optimal" subsamples from balanced full data.

[10] Yao, Y., Zou, J., and Wang, H. (2023). Model constraints independent optimal subsampling probabilities for softmax regression. *JSPI* **225**, 188–201

Focus on the MSPE

- Define $\boldsymbol{\pi}$ to minimize the MSPE.

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\boldsymbol{p}_i(\tilde{\boldsymbol{\beta}})-\boldsymbol{p}_i(\hat{\boldsymbol{\beta}})\right\|^2.$$

- Different identifiability constraints produce identical probabilities.
- Subsamples are balanced.

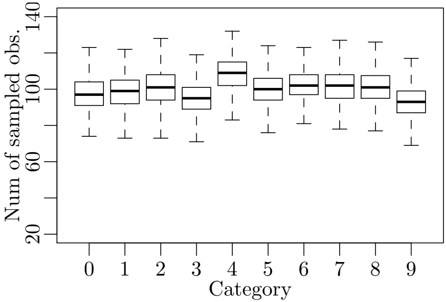


Figure 7: "Optimal" prediction subsamples from balanced full data.

Other problems

- More complicated data structure
  - Censored data:
    - Effect of censoring?
    - Naively using existing approaches results in zero inclusion probabilities for censored observations.
    - Some relevant work: Zuo *et al.* (2021), Yang *et al.* (2022), Keret and Gorfine (2023), Zhang *et al.* (2023)
  - Image data
  - Missing data
  - Network data
  - ......
- Hypothesis test
- Model selection and variable selection
- Model checking
- Model misspecification
- ......

Thank you!

Ai, M., Yu, J., Zhang, H., and Wang, H. (2021). Optimal subsampling algorithms for big data regressions. *Statistica Sinica* **31**, 2, 749–772.

Fan, Y., Liu, Y., Liu, Y., and Qin, J. (2022). Nearly optimal capture-recapture sampling and empirical likelihood weighting estimation for m-estimation with big data. *arXiv preprint arXiv:2209.04569* .

Fithian, W. and Hastie, T. (2014). Local case-control sampling: Efficient subsampling in imbalanced data sets. *Annals of statistics* **42**, 5, 1693.

Han, L., Tan, K. M., Yang, T., Zhang, T., *et al.* (2020). Local uncertainty sampling for large-scale multiclass logistic regression. *Annals of Statistics* **48**, 3, 1770–1788.

Joseph, V. R. and Mak, S. (2021). Supervised compression of big data. *Statistical Analysis and Data Mining* **14**, 3, 217–229.

Keret, N. and Gorfine, M. (2023). Analyzing big EHR data—optimal Cox regression subsampling procedure with rare events. *JASA* 1–14.

Liu, Y. and Fan, Y. (2023). Biased-sample empirical likelihood weighting for missing data problems: an alternative to inverse probability weighting. *JRSSB* **85**, 1, 67–83.

Ma, P., Mahoney, M., and Yu, B. (2015). A statistical perspective on algorithmic leveraging. *JMLR* **16**, 861–911.

Meng, C., Xie, R., Mandal, A., Zhang, X., Zhong, W., and Ma, P. (2021). Lowcon: A design-based subsampling approach in a misspecified linear model. *JCGS* **30**, 3, 694–708.

Pronzato, L. and Wang, H. (2021). Sequential online subsampling for thinning experimental designs. *JSPI* **212**, 169 – 193.

Reuter, T. and Schwabe, R. (2023). D-optimal subsampling design for massive data linear regression. *arXiv preprint arXiv:2307.02236* .

Silvapulle, M. (1981). On the existence of maximum likelihood estimators for the binomial response models. *JRSSB* **43**, 3, 310–313.

Ting, D. and Brochu, E. (2018). Optimal subsampling with influence functions. In *NIPS*, 3650–3659.

Wang, H. (2019). More efficient estimation for logistic regression with optimal subsamples. *JMLR* **20**, 132, 1–59.

Wang, H. (2020). Logistic regression for massive data with rare events. In *ICML*.

Wang, H. (2022). A note on centering in subsample selection for linear regression. *Stat* **11**, 1, e525.

Wang, H. and Kim, J. K. (2022). Maximum sampled conditional likelihood for informative subsampling. *JMLR* **23**, 332, 1–50.

Wang, H. and Ma, Y. (2021). Optimal subsampling for quantile regression in big data. *Biometrika* **108**, 1, 99–112.

Wang, H., Yang, M., and Stufken, J. (2019). Information-based optimal subdata selection for big data linear regression. *JASA* **114**, 525, 393–405.

Wang, H., Zhang, A., and Wang, C. (2021a). Nonuniform negative sampling and log odds correction with rare events data. In *NeurIPS*.

Wang, H., Zhu, R., and Ma, P. (2018). Optimal subsampling for large sample logistic regression. *JASA* **113**, 522, 829–844.

Wang, J., Wang, H., and Xiong, S. (2022a). Unweighted estimation based on optimal sample under measurement constraints. *Canadian*

Journal of Statistics **n/a**, n/a,
https://doi.org/10.1002/cjs.11753.

Wang, J., Zou, J., and Wang, H. (2022b). Sampling with replacement vs poisson sampling: A comparative study in optimal subsampling. *IEEE Transactions on Information Theory* **68**, 10, 6605–6630.

Wang, L., Elmstedt, J., Wong, W. K., and Xu, H. (2021b). Orthogonal subsampling for big data linear regression. *The Annals of Applied Statistics* **15**, 3, 1273–1290.

Yang, Z., Wang, H., and Yan, J. (2022). Optimal subsampling for parametric accelerated failure time models with massive survival data. *Statistics in Medicine* **41**, 27, 5421–5431.

Yao, Y., Zou, J., and Wang, H. (2023). Model constraints independent optimal subsampling probabilities for softmax regression. *JSPI* **225**, 188–201.

Yu, J., Ai, M., and Ye, Z. (2023a). A review on design inspired subsampling for big data. *Statistical Papers* .

Yu, J., Liu, J., and Wang, H. (2023b). Information-based optimal subdata selection for non-linear models. *Statistical Papers* **64**, 4, 1069–1093.

Yu, J., Wang, H., Ai, M., and Zhang, H. (2022). Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. *JASA* **117**, 537, 265–276.

Zhang, H., Zuo, L., Wang, H., and Sun, L. (2023). Approximating partial likelihood estimators via optimal subsampling. *JCGS* **0**, 0, 1–13.

Zhang, T., Ning, Y., and Ruppert, D. (2021). Optimal sampling for generalized linear models under measurement constraints. *JCGS* **30**, 1, 106–114.

Zuo, L., Zhang, H., Wang, H., and Liu, L. (2021). Sampling-based estimation for massive survival data with additive hazards model. *Statistics in medicine* **40**, 2, 441–450.