A comparative study on sampling with replacement vs Poisson sampling in optimal subsampling

HaiYing Wang
University of Connecticut

Jiahui Zou
Chinese Academy of Sciences

Abstract

Faced with massive data, subsampling is a commonly used technique to improve computational efficiency, and using nonuniform subsampling probabilities is an effective approach to improve estimation efficiency. For computational efficiency, subsampling is often implemented with replacement or through Poisson subsampling. However, no rigorous investigation has been performed to study the difference between the two subsampling procedures such as their estimation efficiency and computational convenience. In the context of maximizing a general target function, this paper derives optimal subsampling probabilities for both subsampling with replacement and Poisson subsampling. The optimal subsampling probabilities minimize variance functions of the subsampling estimators. Furthermore, they provide deep insights on the theoretical similarities and differences between subsampling with replacement and Poisson subsampling. Practically implementable algorithms are proposed based on the optimal structural results, which are evaluated by both theoretical and empirical analysis.

1 Introduction

With fast development of technology, data collecting is becoming easier and easier, and the volumes of available data sets are increasing exponentially. To extract useful information from these massive data, a major challenge lies with the thirst for computing resources. Subsampling is a commonly used technique to reduce computational burden, and it has been an important topic in computer science and statistics with a long standing of literature, such as Drineas et al. (2006a,b,c); Mahoney and Drineas (2009); Drineas et al. (2011); Mahoney (2011); Clarkson and Woodruff (2013); Kleiner et al. (2014); McWilliams et al. (2014); Yang et al. (2016); Wang and Ma (2020); Yu et al. (2020).

To improve the estimation efficiency, nonuniform subsampling probabilities are often used so that more informative data points are sampled with higher probabilities. A popular choice is the leverage-based subsampling in which the subsampling distribution is the normalized statistical leverage scores of the design matrix (Drineas et al., 2012; Ma et al., 2015). Yang et al. (2015) showed that if statistical leverage scores are very nonuniform, then using their normalized square roots as the subsampling distribution yields better approximation. For logistic regression, Wang et al. (2018) derived an optimal subsampling distribution that minimizes the asymptotic variance of the subsampling estimator, and Wang (2019) further developed a more efficient estimation approach based on the selected subsample. Ting and Brochu (2018) investigated optimal subsampling with influence functions. Wang et al. (2019) proposed a method called information-based optimal subdata selection which selects data points deterministically for linear regression. The subsampling approach has a close connection to the technique of coreset approximation (Campbell and Broderick, 2018, 2019), which also use a subset of the data with associated weights instead of the full data to perform calculations. The coreset approximation is often used in Bayes analysis and the problem is often to better approximate the objective function in a functional space, while this paper focuses on approximating the full data estimator.

For computational efficiency, subsampling is often implemented with replacement or through Poisson subsampling. Nonuniform subsampling without replacement for a fixed sample size can also be implemented with one-pass of the data through reservoir sampling

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

(Efraimidis and Spirakis, 2006; Tillé, 2019). However, these algorithms are not widely implemented in existing software such as in the R programming language (R Core Team, 2020), so we do not consider this sampling procedure in this paper. Subsampling with replacement needs to use all subsampling probabilities simultaneously to generate random numbers from a multinomial distribution. The resultant subsample observations are independent and identically distributed (i.i.d.) conditional on the full data, but their unconditional distributions are not independent. Poisson subsampling looks at each data point and determine if it should be included in the subsample by generating a random number from the uniform distribution. If the subsampling probabilities in Poisson subsample are all equal, then the subsampling procedure is also called the Bernoulli subsampling (Särndal et al., 2003). For Poisson subsampling, the resultant subsample observations do not have identical distributions, but their unconditional distributions can be independent.

This paper has the following major contributions.

- For estimators obtained through maximizing target functions, we derive asymptotic distributions for both subsampling with replacement and Poisson subsampling. These asymptotic distributions accurately characterize the subsampling approximation errors, and we derive general structure results of optimal subsampling probabilities to minimize these errors for the two subsampling procedures.
- We systematically compare subsampling with replacement and Poisson subsampling, both theoretically and empirically, and identify conditions when they are equivalent and when they differ.
- Based on the optimal subsampling probabilities, we propose practical algorithms and evaluate their performance.

The rest of the paper is organized as follows. We present the model setup and asymptotic distributions in Section 2. In Section 3, we derive optimal subsampling probabilities and propose practical algorithms. We will also obtain theoretical properties for the practical algorithms. In Section 4, we perform numerical experiments demonstrating the performance of the proposed methods. Proofs of our theoretical results are provided in the appendix.

Here are some notation conventions to be used in the paper. We use * to indicate subsample quantities; use $\hat{}$ to indicate full data estimator; use $\tilde{}$ to indicate subsample estimator; use $_R$ and $_P$ to indicate subsampling with replacement and Poisson subsampling, respectively; use $\dot{m}$ and $\ddot{m}$ to denote the gradient and Hessian matrix of a function m; and use $\|\boldsymbol{v}\|$ to denote the spectral norm of a vector or matrix $\boldsymbol{v}$.

2 Problem setup and asymptotic distributions

Suppose that a set of training data $\mathcal{D}_n = \{Z_i\}_{i=1}^n$ consists of independent observations from a common distribution. To estimate some parameter $\boldsymbol{\theta} \in \mathbb{R}^d$ about the data distribution, we want to calculate $\hat{\boldsymbol{\theta}}$, the maximizer of

$$M_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n m(Z_i, \boldsymbol{\theta}).$$

Here the dimension of Z_i does not have to be the same as $\boldsymbol{\theta}$, e.g., in softmax regression. Usually, there is no closed-form solution to $\hat{\boldsymbol{\theta}}$, and an iterative algorithm is required to find the solution numerically. For massive data, iterative calculations on the full data of size n are often too expensive to afford, so subsampling is adopted to produce a subsampling estimator $\tilde{\boldsymbol{\theta}}$ to approximate $\hat{\boldsymbol{\theta}}$. Nonuniform subsampling probabilities are often used to improve the estimation efficiency.

Let $\boldsymbol{\pi} = \{\pi_i\}_{i=1}^n$ be a subsampling distribution such that $\pi_i \geq 0$ and $\sum_{i=1}^n \pi_i = 1$. For Poisson subsampling, we further assume that $\pi_i \leq s^{-1}$, where s is the expected subsample size. As stated earlier, we use * to indicate quantities with randomness due to subsampling. For instance, $\{Z_1^*, ..., Z_s^*\}$ denote the data observations in a subsample and $\{\pi_1^*, ..., \pi_s^*\}$ are the associated subsampling probabilities.

We present the general subsampling estimators $\tilde{\boldsymbol{\theta}}_R$ based on subsampling with replacement and $\tilde{\boldsymbol{\theta}}_P$ based on Poisson subsampling in the following.

Sampling with replacement:

- Calculate $\boldsymbol{\pi} = \{\pi_i\}_{i=1}^n$ based on $\mathcal{D}_n$;
- generate s independent random numbers from multinomial distribution with $\boldsymbol{\pi}$ to determine a subsample $\mathcal{D}_s^* = \{Z_1^*, Z_2^*, ..., Z_s^*\}$;
- record $\{\pi_1^*, \pi_2^*, ..., \pi_s^*\}$ in the subsample;
- obtain the subsample estimator

$$\tilde{\boldsymbol{\theta}}_R = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^s \frac{m(Z_i^*, \boldsymbol{\theta})}{ns\pi_i^*}. \tag{1}$$

Poisson Sampling:

- For each $i = 1, ..., n$, calculate an individual π_i such that $\pi_i \leq s^{-1}$ based on Z_i;
- generate $u_i \sim U(0, 1)$;

- if $u_i \le s\pi_i$, include Z_i in the subsample and record π_i;
- obtain the subsample estimator

$$\tilde{\boldsymbol{\theta}}_P = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{s^*} \frac{m(Z_i^*, \boldsymbol{\theta})}{ns^*\pi_i^*}. \tag{2}$$

Remark 1. Subsampling with replacement requires to access the whole sampling distribution $\boldsymbol{\pi} = \{\pi_i\}_{i=1}^n$, i.e., all π_i's. On the other hand, Poisson subsampling only needs to access one π_i in each sampling iteration. This makes the Poisson subsampling more convenient to implement, especially in distributed computing platforms or when the available memory cannot hold all π_i's. For subsampling with replacement, the subsample size is equal to s and there may be replicates in the subsample. Here π_i is the probability that observation Z_i is selected when only one data point is selected, and the probability to include Z_i in the subsample of size s is $1-(1-\pi_i)^s$, which is smaller than $s\pi_i$. For Poisson subsampling, the subsample size s^* is random with $\mathbb{E}(s^*) = s$; there is no replicates in the subsample; and $s\pi_i$ is the probability of including Z_i in the subsample of expected size s.

We now derive asymptotic properties of $\tilde{\boldsymbol{\theta}}_R$ in (1) and $\tilde{\boldsymbol{\theta}}_P$ in (2), respectively, to compare their estimation efficiency theoretically. We need some regularity assumptions listed below.

Assumption 1. *The parameter $\boldsymbol{\theta}$ belongs to a compact set.*

Assumption 2. *The function $m(Z, \boldsymbol{\theta})$ is a concave function of $\boldsymbol{\theta}$ with a unique and finite maximum, and it satisfies that $\mathbb{E}\{m^2(Z, \boldsymbol{\theta})\} < \infty$ for any $\boldsymbol{\theta}$.*

Assumption 3. *Assume that $-\mathbb{E}\{\ddot{m}(Z, \boldsymbol{\theta})\}$ is positive-definite, $\mathbb{E}\{\ddot{m}_{k,l}^2(Z, \boldsymbol{\theta})\} < \infty$, and $\ddot{m}(Z, \boldsymbol{\theta})$ is Lipschitz continuous in $\boldsymbol{\theta}$ so that there exists a function $\psi(z)$ with $\mathbb{E}\{\psi^2(Z)\} < \infty$ and for every $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, $|\ddot{m}_{k,l}(z, \boldsymbol{\theta}_1) - \ddot{m}_{k,l}(z, \boldsymbol{\theta}_2)| \le \psi(z)\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$, $k, l = 1, 2, ..., d$.*

Assumption 4. *Assume that $\mathbb{E}\{\dot{m}(Z, \boldsymbol{\theta})\dot{m}^{\mathrm{T}}(Z, \boldsymbol{\theta})\}$ is a positive-definite matrix and for $\boldsymbol{\theta}$ in the neighborhood of $\hat{\boldsymbol{\theta}}$, $\frac{1}{n}\sum_{i=1}^n \|\dot{m}(Z_i, \boldsymbol{\theta})\|^4 = O_P(1)$, where $O_P(1)$ means bounded in probability (with high probability).*

Assumption 5. *Assume that $\max_{i=1,...,n}(n\pi_i)^{-1} = O_P(1)$.*

Assumptions 1 and 2 are very mild, and they assure that the target function has a finite and unique maximum. Assumptions 3 and 4 impose some constrains on the Hessian matrix and gradient, and they are required so that the asymptotic distributions of parameter estimators are asymptotically normal. Assumption 5 essentially requires that the minimum subsampling probability is at the same order of $\frac{1}{n}$ in probability. Here, π_i can be random as it is allowed to depend on the data, so the notation $O_P(1)$ is used.

The following Theorems 1 and 2 present asymptotic distributions of $\tilde{\boldsymbol{\theta}}_R$ in (1) and $\tilde{\boldsymbol{\theta}}_P$ in (2), respectively.

Theorem 1. *Under Assumptions 1-5, as $s \to \infty$ and $n \to \infty$, the estimator $\tilde{\boldsymbol{\theta}}_R$ in (1) satisfies that,*

$$\sqrt{s}\{V_R(\hat{\boldsymbol{\theta}})\}^{-1/2}(\tilde{\boldsymbol{\theta}}_R - \hat{\boldsymbol{\theta}}) \xrightarrow{D} \mathbb{N}(\mathbf{0}, \boldsymbol{I}), \tag{3}$$

where $\xrightarrow{D}$ means convergence in distribution, $\mathbb{N}(\mathbf{0}, \boldsymbol{I})$ is a multivariate Gaussian distribution with mean $\mathbf{0}$ and variance $\boldsymbol{I}$ (the identity matrix), and $V_R(\hat{\boldsymbol{\theta}}) = \ddot{M}_n^{-1}(\hat{\boldsymbol{\theta}})\Lambda_R(\hat{\boldsymbol{\theta}})\ddot{M}_n^{-1}(\hat{\boldsymbol{\theta}})$,

$$\ddot{M}_n(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^n \ddot{m}(Z_i, \hat{\boldsymbol{\theta}}), \tag{4}$$

$$\Lambda_R(\hat{\boldsymbol{\theta}}) = \frac{1}{n^2}\sum_{i=1}^n \frac{\dot{m}(Z_i, \hat{\boldsymbol{\theta}})\dot{m}^{\mathrm{T}}(Z_i, \hat{\boldsymbol{\theta}})}{\pi_i}. \tag{5}$$

Theorem 2. *Under Assumptions 1-5, as $s \to \infty$ and $n \to \infty$, the estimator $\tilde{\boldsymbol{\theta}}_P$ in (2) satisfies that,*

$$\sqrt{s}\{V_P(\hat{\boldsymbol{\theta}})\}^{-1/2}(\tilde{\boldsymbol{\theta}}_P - \hat{\boldsymbol{\theta}}) \xrightarrow{D} \mathbb{N}(\mathbf{0}, \boldsymbol{I}), \tag{6}$$

where $V_P(\hat{\boldsymbol{\theta}}) = \ddot{M}_n^{-1}(\hat{\boldsymbol{\theta}})\Lambda_P(\hat{\boldsymbol{\theta}})\ddot{M}_n^{-1}(\hat{\boldsymbol{\theta}})$, $\ddot{M}_n(\hat{\boldsymbol{\theta}})$ is the same as in (4), and

$$\Lambda_P(\hat{\boldsymbol{\theta}}) = \Lambda_R(\hat{\boldsymbol{\theta}}) - \frac{s}{n^2}\sum_{i=1}^n \dot{m}(Z_i, \hat{\boldsymbol{\theta}})\dot{m}^{\mathrm{T}}(Z_i, \hat{\boldsymbol{\theta}}). \tag{7}$$

Remark 2. The asymptotic distributions in (3) and (6) mean that given a full data set for any $\delta > 0$, the probability that $\|\tilde{\boldsymbol{\theta}}_R - \hat{\boldsymbol{\theta}}\| > \delta$ is accurately approximated by $\mathbb{P}(\|U_R\| > \delta)$ where $U_R \sim \mathbb{N}(\mathbf{0}, V_R(\hat{\boldsymbol{\theta}}))$, and the probability that $\|\tilde{\boldsymbol{\theta}}_P - \hat{\boldsymbol{\theta}}\| > \delta$ is accurately approximated by $\mathbb{P}(\|U_P\| > \delta)$ where $U_P \sim \mathbb{N}(\mathbf{0}, V_P(\hat{\boldsymbol{\theta}}))$. Thus, a smaller variance means a smaller probability of excess error at the same error bound, or a smaller error bound for the same excess probability.

Remark 3. Both $\tilde{\boldsymbol{\theta}}_R$ and $\tilde{\boldsymbol{\theta}}_P$ have Gaussian asymptotic distributions, but they have different asymptotic variances $V_R(\hat{\boldsymbol{\theta}})$ and $V_P(\hat{\boldsymbol{\theta}})$, respectively. Under Assumption 4, the second term on the right-hand-side of (7) goes to zero in probability if $s/n \to 0$, and it converges to a positive-definite matrix in probability if $s/n \to c > 0$. Thus, the difference $V_R(\hat{\boldsymbol{\theta}}) - V_P(\hat{\boldsymbol{\theta}}) \to \mathbf{0}$ in probability if $s/n \to 0$, and it converges to a positive-definite matrix in probability if s/n converges to a positive constant. This means that subsampling with replacement and Poisson subsampling have the same asymptotic estimation efficiency only if the subsampling ratio s/n goes to zero; otherwise, Poisson subsampling has a higher estimation efficiency. Thus, to obtain more accurate estimates in practice, Poisson subsampling is recommended unless the subsampling ratio s/n is very small.

3 Optimal subsampling probabilities

From the results in Theorems 1 and 2, the asymptotic variances $V_R(\hat{\boldsymbol{\theta}})$ and $V_P(\hat{\boldsymbol{\theta}})$ depend on $\boldsymbol{\pi} = \{\pi_i\}_{i=1}^n$. To improve the estimation efficiency, we want to choose optimal $\boldsymbol{\pi}$ to minimize $V_R(\hat{\boldsymbol{\theta}})$ or $V_P(\hat{\boldsymbol{\theta}})$. Specifically, we consider the L-optimality criterion (Atkinson et al., 2007). The L-optimality minimizes the trace of the variance matrix for some linear transformation, say L, of the parameter estimator. If we take $L = \boldsymbol{I}$, then the resulting criterion is also called the A-optimality. This is to minimize the average of the variances for all parameter components by minimizing the trace of the variance matrix. For our case, this is to minimize $\mathrm{tr}\{V_R(\hat{\boldsymbol{\theta}})\}$ or $\mathrm{tr}\{V_P(\hat{\boldsymbol{\theta}})\}$. If we take $L = \ddot{M}_n(\hat{\boldsymbol{\theta}})$, then the resultant criterion is to minimize $\mathrm{tr}\{\Lambda_R(\hat{\boldsymbol{\theta}})\}$ or $\mathrm{tr}\{\Lambda_P(\hat{\boldsymbol{\theta}})\}$. This has a computational advantage compared with other choices, so we focus more on this choice in this paper. The following Theorems 3 and 4 present the optimal subsampling probabilities for subsampling with replacement and Poisson subsampling, respectively.

Theorem 3. *For the subsampling with replacement estimator in (1), the L-optimal subsampling probabilities with $L = \ddot{M}_n(\hat{\boldsymbol{\theta}})$ that minimize $\mathrm{tr}\{\Lambda_R(\hat{\boldsymbol{\theta}})\}$ are*

$$\pi_{Ri}^{\mathrm{opt}} = \frac{\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}})\|}{\sum_{j=1}^n \|\dot{m}(Z_j, \hat{\boldsymbol{\theta}})\|}, \quad i = 1, ..., n. \tag{8}$$

Theorem 4. *For the Poisson subsampling estimator in (2), the L-optimal subsampling probabilities with $L = \ddot{M}_n(\hat{\boldsymbol{\theta}})$ that minimize $\mathrm{tr}\{\Lambda_P(\hat{\boldsymbol{\theta}})\}$ are*

$$\pi_{Pi}^{\mathrm{opt}} = \frac{\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}})\| \wedge H}{\sum_{j=1}^n \{\|\dot{m}(Z_j, \hat{\boldsymbol{\theta}})\| \wedge H\}}, \quad i = 1, ..., n, \tag{9}$$

where $a \wedge b = \min(a, b)$,

$$H = \frac{\sum_{i=1}^{n-g} \|\dot{m}(Z, \hat{\boldsymbol{\theta}})\|_{(i)}}{s - g}, \tag{10}$$

$\|\dot{m}(Z, \hat{\boldsymbol{\theta}})\|_{(1)} \leq ... \leq \|\dot{m}(Z, \hat{\boldsymbol{\theta}})\|_{(n)}$ are the order statistics of $\|\dot{m}(Z_1, \hat{\boldsymbol{\theta}})\|, ..., \|\dot{m}(Z_n, \hat{\boldsymbol{\theta}})\|$, and g is an integer such that

$$\frac{\|\dot{m}(Z, \hat{\boldsymbol{\theta}})\|_{(n-g)}}{\sum_{i=1}^{n-g} \|\dot{m}(Z, \hat{\boldsymbol{\theta}})\|_{(i)}} < \frac{1}{s - g}, \tag{11}$$

$$\frac{\|\dot{m}(Z, \hat{\boldsymbol{\theta}})\|_{(n-g+1)}}{\sum_{i=1}^{n-g+1} \|\dot{m}(Z, \hat{\boldsymbol{\theta}})\|_{(i)}} \geq \frac{1}{s - g + 1}, \tag{12}$$

in which we define $\|\dot{m}(Z, \hat{\boldsymbol{\theta}})\|_{(n+1)} = \infty$.

Remark 4. For a general choice of L, we can obtain optimal subsampling probabilities by replacing $\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}})\|$ with $\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}})\|_L = \|L\ddot{M}_n^{-1}(\hat{\boldsymbol{\theta}})\dot{m}(Z_i, \hat{\boldsymbol{\theta}})\|$. However, these quantities require $O(nd^2)$ time to compute when $\ddot{M}_n^{-1}(\hat{\boldsymbol{\theta}})$ and $\dot{m}(Z_i, \hat{\boldsymbol{\theta}})$ are available, where n is the full data sample size and d is dimension of $\hat{\boldsymbol{\theta}}$. On the other hand, it only takes $O(nd)$ time to compute all $\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}})\|$'s. Thus the choice of $L = \ddot{M}_n(\hat{\boldsymbol{\theta}})$ has a significant computational advantage.

Remark 5. In Theorems 3 and 4, π_{Ri}^{opt} in (8) and π_{Pi}^{opt} in (9) have both similarities and differences. Assuming that $\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}})\| > 0$ for all i, then $0 < \pi_{Ri}^{\mathrm{opt}} < 1$ while $0 < \pi_{Pi}^{\mathrm{opt}} \leq \frac{1}{s}$. This means that the inclusion of any data point through optimal subsampling with replacement is random, while the inclusion of data points with $\pi_{Pi}^{\mathrm{opt}} = \frac{1}{s}$ is deterministic through optimal Poisson subsampling. The order statistics constraints in (11) and (12) indicate that if there are data points such that $\frac{s}{n}\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}})\| > \frac{1}{n}\sum_{j=1}^n \|\dot{m}(Z_j, \hat{\boldsymbol{\theta}})\|$, then π_{Ri}^{opt} and π_{Pi}^{opt} are different. This indicates that if the subsampling ratio $\frac{s}{n}$ is larger or if the tail of the distribution of $\|\dot{m}(Z, \hat{\boldsymbol{\theta}})\|$ is heavier, then optimal probabilities for Poisson subsampling and subsampling with replacement are more likely to be different. If $s\|\dot{m}(Z, \hat{\boldsymbol{\theta}})\|_{(n)} < \sum_{i=1}^n \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}})\|$, then π_{Ri}^{opt} and π_{Pi}^{opt} are identical.

Remark 6. In Theorem 4, H is the threshold so that all π_{Pi}^{opt} are no larger than $\frac{1}{s}$, and it satisfies that

$$\|\dot{m}(Z, \hat{\boldsymbol{\theta}})\|_{(n-g)} < H \leq \|\dot{m}(Z, \hat{\boldsymbol{\theta}})\|_{(n-g+1)}. \tag{13}$$

Here g is the number of cases that $\pi_{Pi}^{\mathrm{opt}} = \frac{1}{s}$, i.e., the number of data points that will be included in the subsample for sure.

Now we discuss examples to illustrate the optimal structural results.

Example 1 (Binary response models)**.** Consider a binary classification model such that

$$\mathbb{P}(y_i = 1) = p(\boldsymbol{x}_i, \boldsymbol{\theta}), \quad i = 1, ..., n,$$

where $y_i \in \{0, 1\}$ is the binary class label, $\boldsymbol{x}_i$ is the covariate, and $\boldsymbol{\theta}$ is the unknown parameter. To estimate $\boldsymbol{\theta}$ using the maximum likelihood estimator (MLE), let $Z_i = (\boldsymbol{x}_i, y_i)$ and $m(Z_i, \boldsymbol{\theta}) = y_i \log\{p(\boldsymbol{x}_i, \boldsymbol{\theta})\} + (1 - y_i)\log\{1 - p(\boldsymbol{x}_i, \boldsymbol{\theta})\}$. Direct calculations yield that

$$\dot{m}(Z_i, \hat{\boldsymbol{\theta}}) = \frac{y_i - \hat{p}_i}{\hat{p}_i(1 - \hat{p}_i)}\dot{\hat{p}}_i, \tag{14}$$

where $\hat{p}_i = p(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}})$, and $\dot{\hat{p}}_i = \dot{p}(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}})$ is the gradient of $p(\boldsymbol{x}_i, \boldsymbol{\theta})$ evaluated at $\hat{\boldsymbol{\theta}}$. We can obtain optimal sampling probabilities by inserting the expression in (14) into Theorems 3 and 4. From (14), the optimal subsampling probabilities are proportional to $|y_i - \hat{p}_i|$. Thus if $y_i = 1$, data points with smaller values of $\hat{p}_i$

are sampled with higher probabilities; if $y_i = 0$, data points with larger values of $\hat{p}_i$ are sampled with higher probabilities. The optimal subsampling probabilities give higher preference to data points that are closer to the class boundary. This increases the classification accuracy because if these data points can be classified correctly, then other data points are easier to classify.

Specifically for Logistic regression in which $p(\boldsymbol{x}_i, \boldsymbol{\theta}) = e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}}/(1 + e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}})$, we have

$$\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}})\| = |y_i - \hat{p}_i|\|\boldsymbol{x}_i\|, \tag{15}$$

With this expression, the structural results for optimal probabilities of subsampling with replacement are identical to those in Wang et al. (2018).

Form Theorem 4 we see that if there are data points such that $\frac{s}{n}|y_i - \hat{p}_i|\|\boldsymbol{x}_i\| > \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{p}_j|\|\boldsymbol{x}_j\|$, then optimal probabilities for Poisson subsampling are different from that for subsampling with replacement.

Example 2 (Least-squares)**.** Consider least-squares estimator

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \{y_i - g(\boldsymbol{x}_i, \boldsymbol{\theta})\}^2,$$

where y_i is the response, $\boldsymbol{x}_i$ is the covariate, and $g(\boldsymbol{x}_i, \boldsymbol{\theta})$ is a smooth function. The least-squares estimator of $\boldsymbol{\theta}$ can be presented in our framework by letting $Z_i = (\boldsymbol{x}_i, y_i)$ and defining $m(Z_i, \boldsymbol{\theta}) = -\frac{1}{2}\{y_i - g(\boldsymbol{x}_i, \boldsymbol{\theta})\}^2$. From direct calculation, we have

$$\dot{m}(Z_i, \hat{\boldsymbol{\theta}}) = \hat{\varepsilon}_i \dot{g}(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}), \tag{16}$$

where $\hat{\varepsilon}_i = y_i - g(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}})$, and $\dot{g}(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}})$is the gradient of $g(\boldsymbol{x}_i, \boldsymbol{\theta})$evaluated at $\hat{\boldsymbol{\theta}}$.

Specifically for ordinary least-squares (OLS) in linear regression, $\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}})\| = |\hat{\varepsilon}_i|\|\boldsymbol{x}_i\|$, and the sampling probabilities reduce to gradient-based sampling probabilities (Zhu, 2016). Furthermore, if we use the L-optimality criterion with $L = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{1/2}$ where $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)^{\mathrm{T}}$, then the optimal probabilities for subsampling with replacement satisfy that

$$\pi_{Ri}^{\mathrm{opt}} \propto |\hat{\varepsilon}_i|\sqrt{h_i}, \quad i = 1, ..., n, \tag{17}$$

where h_i's are statistical leverage scores of $\boldsymbol{X}$. This clearly shows the connection between leverage scores and the L optimality.

Form (17) and Theorem 4, optimal probabilities for Poisson subsampling and subsampling with replacement differ if there are data points such that $\frac{s}{n}|\hat{\varepsilon}_i|\sqrt{h_i} > \frac{1}{n}\sum_{j=1}^{n}|\hat{\varepsilon}_j|\sqrt{h_j}$. This is more likely to happen if $|\hat{\varepsilon}_i|$'s or $\sqrt{h_i}$'s are more nonuniform. Yang et al. (2015) showed that if statistical leverage scores are very nonuniform, then using the square roots of statistical leverage scores to construct subsampling probabilities yields better approximation than using the original leverage scores. An intuitive explanation for their conclusion is that taking square roots on leverage scores has some shrinkage effect on the resulting probabilities toward the uniform subsampling probability. Our results echos their conclusion, and further indicates that for optimal Poisson subsampling it may be necessary to perform truncation for high leverage scores.

Example 3 (Generalized linear models)**.** Let y_i be the response and $\boldsymbol{x}_i$ be the corresponding covariate. A generalized linear model (GLM) assumes that the conditional mean of the response y_i give the covariate $\boldsymbol{x}_i$, $\mathbb{E}(y_i|\boldsymbol{x}_i)$, satisfies

$$g\{\mathbb{E}(y_i|\boldsymbol{x}_i)\} = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta},$$

where g is the link function, $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$ is the linear predictor, and $\boldsymbol{\beta}$ is the regression coefficient. For most of the commonly used GLMs, it is assumed that the distribution of the response y_i given the covariate $\boldsymbol{x}_i$ belongs to the exponential family, namely,

$$f(y_i|\boldsymbol{x}_i; \boldsymbol{\beta}, \phi) = a(y_i, \phi)\exp\Big[\frac{y_i b(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) - c(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}{\phi}\Big],$$

where a, b and c are known scalar functions, and ϕ is the dispersion parameter. Let $Z_i = (\boldsymbol{x}_i, y_i)$. To estimate $\boldsymbol{\theta} = \boldsymbol{\beta}$, the MLE of $\boldsymbol{\theta}$ corresponds to $m(Z_i, \boldsymbol{\theta}) = y_i b(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) - c(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$, and thus

$$\dot{m}(Z_i, \boldsymbol{\theta}) = \{y_i b'(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) - c'(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\}\boldsymbol{x}_i, \tag{18}$$

where b' is the first derivative of b. Optimal sampling probabilities can be obtained by using the expressions in (18) for Theorems 3 and 4.

3.1 Practical algorithms

The optimal subsampling probabilities depend on the full data estimator $\hat{\boldsymbol{\theta}}$, so the structural results in the previous section do not translate into useful algorithms directly. We need a pilot estimator to approximate the optimal subsampling probabilities in order to obtain practically implementable algorithms. This can be done by taking a pilot subsample of size s_0 through a subsampling distribution that does not depend on $\hat{\boldsymbol{\theta}}$. We use the uniform subsampling distribution, and present the approximated optimal subsampling with replacement procedure in Algorithm 1.

Compared with the exact π_{Ri}^{opt}, the approximated $\tilde{\pi}_{Ri}^{\mathrm{opt}}$ in (21) are subject to additional disturbance due to the randomness of $\tilde{\boldsymbol{\theta}}_R^{0*}$, the maximizer of (19). From Theorem 1, the subsampling probabilities are in the

Algorithm 1 Practical algorithm based on optimal subsampling with replacement

• *Pilot subsampling*: use sampling with replacement with $\boldsymbol{\pi}^{\text{uni}} = \{\pi_i = \frac{1}{n}\}_{i=1}^n$ to obtain $\{Z_1^{0*}, ..., Z_{s_0}^{0*}\}$; obtain $\tilde{\boldsymbol{\theta}}_R^{0*}$ through maximizing

$$M_R^{0*}(\boldsymbol{\theta}) = \sum_{i=1}^{s_0} \frac{m(Z_i^{0*}, \boldsymbol{\theta})}{s_0}. \tag{19}$$

• *Approximated optimal subsampling:* calculate the whole subsampling distribution $\tilde{\boldsymbol{\pi}}_{R\alpha i} = \{\tilde{\pi}_{R\alpha i}^{\text{opt}}\}_{i=1}^n$, where $\alpha \in (0, 1)$,

$$\tilde{\pi}_{Ri}^{\text{opt}} = \frac{\|\dot{m}(Z_i, \tilde{\boldsymbol{\theta}}_R^{0*})\|}{\sum_{j=1}^n \|\dot{m}(Z_j, \tilde{\boldsymbol{\theta}}_R^{0*})\|}, \tag{20}$$

$$\text{and} \quad \tilde{\pi}_{R\alpha i}^{\text{opt}} = (1-\alpha)\tilde{\pi}_{Ri}^{\text{opt}} + \alpha\frac{1}{n}; \tag{21}$$

use $\tilde{\boldsymbol{\pi}}_{R\alpha i}$ to take a subsample $\{Z_1^*, ..., Z_s^*\}$, and record the corresponding probabilities $\{\tilde{\pi}_{R\alpha 1}^{\text{opt}*}, ..., \tilde{\pi}_{R\alpha s}^{\text{opt}*}\}$.

• *Estimation:* obtain $\tilde{\boldsymbol{\theta}}_R^{\alpha}$ through maximizing

$$M_{R\alpha}^*(\boldsymbol{\theta}) = \sum_{i=1}^{s} \frac{m(Z_i^*, \boldsymbol{\theta})}{ns\tilde{\pi}_{R\alpha i}^{\text{opt}*}}. \tag{22}$$

Algorithm 2 Practical algorithm based on optimal Poisson subsampling

• *Pilot subsampling*: use Poisson sampling with $\boldsymbol{\pi}^{\text{uni}}$ to obtain $\{Z_1^{0*}, ..., Z_{s_0^*}^{0*}\}$;

obtain $\tilde{\boldsymbol{\theta}}_P^{0*}$ through maximizing

$$M_P^{0*}(\boldsymbol{\theta}) = \sum_{i=1}^{s_0^*} \frac{m(Z_i^{0*}, \boldsymbol{\theta})}{s_0^*}; \tag{23}$$

calculate

$$H^{0*} = \|\dot{m}(Z_i^{0*}, \tilde{\boldsymbol{\theta}}_P^{0*})\|_{\frac{s}{bn}}, \tag{24}$$

$$\Psi^{0*} = \sum_{i=1}^{s_0^*} \frac{\{\|\dot{m}(Z_i^{0*}, \tilde{\boldsymbol{\theta}}_P^{0*})\| \wedge H^{0*}\}}{s_0^*}. \tag{25}$$

• *Approximated optimal subsampling:* For each i of $i = 1, ..., n$, calculate

$$\tilde{\pi}_{Pi}^{\text{opt}} = \frac{\|\dot{m}(Z_i, \tilde{\boldsymbol{\theta}}_P^{0*})\| \wedge H^{0*}}{n\Psi^{0*}}, \tag{26}$$

$$\tilde{\pi}_{P\alpha i}^{\text{opt}} = (1-\alpha)\tilde{\pi}_{Pi}^{\text{opt}} + \alpha\frac{1}{n}; \tag{27}$$

generate $u_i \sim U(0, 1)$;
if $u_i \leq s\tilde{\pi}_{P\alpha i}^{\text{opt}}$, include Z_i in the subsample and record $\tilde{\pi}_{P\alpha i}^{\text{opt}}$.

• *Estimation:* obtain $\tilde{\boldsymbol{\theta}}_P^{\alpha}$ through maximizing

$$M_{P\alpha}^*(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{s^*} \frac{m(Z_i^*, \boldsymbol{\theta})}{(s\tilde{\pi}_{P\alpha i}^{\text{opt}*}) \wedge 1}. \tag{28}$$

denominators of $\Lambda_R(\hat{\boldsymbol{\theta}})$. Thus the additional disturbance may be amplified for data points with π_{Ri}^{opt} being close to zero, and this may inflate the asymptotic variance of the subsample estimator. To protect the estimator from these data points, we adopt the idea of defensive importance sampling (Hesterberg, 1995; Owen and Zhou, 2000) and mix the approximated optimal subsampling distribution with the uniform subsampling distribution. Specifically, we use $\tilde{\pi}_{R\alpha i}^{\text{opt}}$ in (21) instead of $\tilde{\pi}_{Ri}^{\text{opt}}$ in (20) to perform the subsampling. The same idea was also adopted in Ma et al. (2015).

In $\tilde{\boldsymbol{\pi}}_{R\alpha i} = \{\tilde{\pi}_{R\alpha i}^{\text{opt}}\}_{i=1}^n$, α controls the proportion of mixture, and $\tilde{\boldsymbol{\pi}}_{R\alpha i}$ is close to the optimal subsampling distribution if α is close to 0 while it is close to the uniform subsampling distribution if α is close to 1. If $\alpha > 0$, then $n\pi_{R\alpha i}^{\text{opt}}$ are bounded away from zero, which add to robustness of the subsampling estimator.

For the optimal Poisson subsampling probability π_{Pi}^{opt}, we also need to use the pilot subsample to approximate H and $\Psi = \frac{1}{n}\sum_{i=1}^n \{\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}})\| \wedge H\}$ in order to determine the inclusion probability based on each data point itself, as described in Algorithm 2. From (13), H is between the $(n-g)$-th and the $(n-g+1)$-th order statistics of $\{\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}})\|\}_{i=1}^n$, and g is between 0 and s, so we can roughly approximate H with $\|\dot{m}(Z_i^{0*}, \tilde{\boldsymbol{\theta}}_P^{0*})\|_{\frac{s}{bn}}$, the upper $\frac{s}{bn}$-th sample quantile of $\{\|\dot{m}(Z_i^{0*}, \tilde{\boldsymbol{\theta}}_P^{0*})\|\}_{i=1}^{s_0}$, where $b \geq 1$ is a tuning parameter. Since g is typically closer to 0 and farther from s, taking $b = 1$ underestimates H and the resulting subsampling probabilities lean towards the uniform subsampling probability (if $H \leq \|\dot{m}(Z, \hat{\boldsymbol{\theta}})\|_{(1)}$, then π_{Pi}^{opt} would be all equal to $\frac{1}{n}$). When subsampling from massive data, s is often much smaller than n and the number of cases for $\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}})\|$ to be larger than H is small. For this scenario, one may simply ignore H and use ∞ to replace H. This simple option in general overestimates H, but it may perform reasonably well for small subsampling ratios. For Ψ, it can be approximated by Ψ^{0*} defined in (25).

When we use Ψ^{0*} and H^{0*} to replace Ψ and H in (27), it is possible that some $\tilde{\pi}_{Pi}^{\text{opt}}$ in (27) are larger than $\frac{1}{s}$ and thus $s\tilde{\pi}_{Pi}^{\text{opt}}$ are larger than one. Thus, we use one as a threshold in the denominator of (28).

Remark 7. In Algorithm 1, $\tilde{\boldsymbol{\theta}}_R^{0*}$ and $\tilde{\boldsymbol{\theta}}_R^{\alpha}$ can be combined to obtain an aggregated estimator, $\check{\boldsymbol{\theta}}_R = \{s_0\ddot{M}_R^{0*} + s\ddot{M}_R^*\}^{-1} \times \{s_0\ddot{M}_R^{0*} \times \tilde{\boldsymbol{\theta}}_R^{0*} + s\ddot{M}_R^* \times \tilde{\boldsymbol{\theta}}_R^{\alpha}\}$. where $\ddot{M}_R^{0*}$ is the Hessian matrix of $M_R^{0*}(\boldsymbol{\theta})$ in (19) evaluated at $\tilde{\boldsymbol{\theta}}_R^{0*}$ and $\ddot{M}_R^*$ is the Hessian matrix of $M_R^*(\boldsymbol{\theta})$ in (22) evaluated at $\tilde{\boldsymbol{\theta}}_R^{\alpha}$. Here, $\check{\boldsymbol{\theta}}_R$ is obtained as a linear combination of $\tilde{\boldsymbol{\theta}}_R^{0*}$ and $\tilde{\boldsymbol{\theta}}_R^{\alpha}$ in a way similar to the aggregation step in the divide-and-conquer method (Lin and Xie, 2011; Schifano et al., 2016). This fur-

ther improves the estimation efficiency. Similarly, in Algorithm 2, $\tilde{\boldsymbol{\theta}}_P^{0*}$ and $\tilde{\boldsymbol{\theta}}_P^{\alpha}$ can be combined to obtain an aggregated estimator.

3.2 Theoretical analysis of practical algorithms

We obtain the following distributional results in Theorems 5 and 6 for Algorithms 1 and 2, respectively.

Theorem 5. *For $\tilde{\boldsymbol{\theta}}_R^{\alpha}$ obtained from Algorithm 1, under Assumptions 1-4, as s_0, s, and n get large, the following result hold.*

$$\sqrt{s}\{V_R^{\alpha}(\hat{\boldsymbol{\theta}})\}^{-1/2}(\tilde{\boldsymbol{\theta}}_R^{\alpha}-\hat{\boldsymbol{\theta}})\xrightarrow{D}\mathbb{N}(\mathbf{0},\boldsymbol{I}),$$

where $V_R^{\alpha}(\hat{\boldsymbol{\theta}}) = \ddot{M}_n^{-1}(\hat{\boldsymbol{\theta}})\Lambda_R^{\alpha}(\hat{\boldsymbol{\theta}})\ddot{M}_n^{-1}(\hat{\boldsymbol{\theta}})$, $\pi_{R\alpha i}^{\mathrm{opt}}(\hat{\boldsymbol{\theta}}) = (1-\alpha)\pi_{Ri}^{\mathrm{opt}}(\hat{\boldsymbol{\theta}})+\alpha\frac{1}{n}$, and

$$\Lambda_R^{\alpha}(\hat{\boldsymbol{\theta}}) = \frac{1}{n^2}\sum_{i=1}^{n}\frac{\dot{m}(Z_i,\hat{\boldsymbol{\theta}})\dot{m}^{\mathrm{T}}(Z_i,\hat{\boldsymbol{\theta}})}{\pi_{R\alpha i}^{\mathrm{opt}}(\hat{\boldsymbol{\theta}})}.$$

Theorem 6. *For $\tilde{\boldsymbol{\theta}}_P^{\alpha}$ obtained from Algorithm 2, under Assumptions 1-4, as s_0, s, and n get large, if $s_0 = o(n)$, $\varrho_n = s/(bn)\to\varrho\in[0,1)$, and the distribution of Z is continuous, the following result hold. If $\varrho = 0$, then*

$$\sqrt{s}\{V_P^{\alpha}(\hat{\boldsymbol{\theta}})\}^{-1/2}(\tilde{\boldsymbol{\theta}}_P^{\alpha}-\hat{\boldsymbol{\theta}})\xrightarrow{D}\mathbb{N}(\mathbf{0},\boldsymbol{I}),$$

where $V_P^{\alpha}(\hat{\boldsymbol{\theta}}) = \ddot{M}_n^{-1}(\hat{\boldsymbol{\theta}})\Lambda_P^{\alpha}(\hat{\boldsymbol{\theta}})\ddot{M}_n^{-1}(\hat{\boldsymbol{\theta}})$, $\pi_{P\alpha i}^{\mathrm{opt}}(\hat{\boldsymbol{\theta}}) = (1-\alpha)\pi_{Pi}^{\mathrm{opt}}(\hat{\boldsymbol{\theta}})+\alpha\frac{1}{n}$, and

$$\Lambda_P^{\alpha}(\hat{\boldsymbol{\theta}}) = \frac{1}{n^2}\sum_{i=1}^{n}\frac{\{1-s\pi_{P\alpha i}^{\mathrm{opt}}(\hat{\boldsymbol{\theta}})\}\dot{m}(Z_i,\hat{\boldsymbol{\theta}})\dot{m}^{\mathrm{T}}(Z_i,\hat{\boldsymbol{\theta}})}{\pi_{P\alpha i}^{\mathrm{opt}}(\hat{\boldsymbol{\theta}})};$$

if $\varrho>0$, then $\pi_{Pi}^{\mathrm{opt}}(\hat{\boldsymbol{\theta}})$ in $\Lambda_P^{\alpha}(\hat{\boldsymbol{\theta}})$ is replaced by $\pi_{Pi}^{\mathrm{opt}} = \frac{\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}})\|\wedge H_{\varrho n}}{\sum_{j=1}^{n}\{\|\dot{m}(Z_j,\hat{\boldsymbol{\theta}})\|\wedge H_{\varrho n}\}}$, where H_{ϱ_n} is the ϱ_n-th upper sample quantile of $\|\dot{m}(Z_1,\hat{\boldsymbol{\theta}})\|,...,\|\dot{m}(Z_n,\hat{\boldsymbol{\theta}})\|$.

Remark 8. Denote $\Lambda_R^{\mathrm{opt}}(\hat{\boldsymbol{\theta}})$ and $\Lambda_P^{\mathrm{opt}}(\hat{\boldsymbol{\theta}})$ as $\Lambda_R(\hat{\boldsymbol{\theta}})$ and $\Lambda_P(\hat{\boldsymbol{\theta}})$ with optimal subsampling probabilities that produce the minimum trace values, respectively. In Theorems 5 and 6, $\Lambda_R^{\alpha}(\hat{\boldsymbol{\theta}})$ and $\Lambda_P^{\alpha}(\hat{\boldsymbol{\theta}})$ are different from $\Lambda_R^{\mathrm{opt}}(\hat{\boldsymbol{\theta}})$ and $\Lambda_P^{\mathrm{opt}}(\hat{\boldsymbol{\theta}})$, respectively. However, it can be shown that

$$\mathrm{tr}\{\Lambda_R^{\mathrm{opt}}(\hat{\boldsymbol{\theta}})\}<\mathrm{tr}\{\Lambda_R^{\alpha}(\hat{\boldsymbol{\theta}})\}<\frac{\mathrm{tr}\{\Lambda_R^{\mathrm{opt}}(\hat{\boldsymbol{\theta}})\}}{1-\alpha},\ \text{ and}$$
$$\mathrm{tr}\{\Lambda_P^{\mathrm{opt}}(\hat{\boldsymbol{\theta}})\}<\mathrm{tr}\{\Lambda_P^{\alpha}(\hat{\boldsymbol{\theta}})\}<\frac{\mathrm{tr}\{\Lambda_P^{\mathrm{opt}}(\hat{\boldsymbol{\theta}})\}}{1-\alpha}.$$

Thus, if α is small enough, $\mathrm{tr}\{\Lambda_R^{\alpha}(\hat{\boldsymbol{\theta}})\}$ and $\mathrm{tr}\{\Lambda_R^{\mathrm{opt}}(\hat{\boldsymbol{\theta}})\}$ can be arbitrarily close, and $\mathrm{tr}\{\Lambda_P^{\alpha}(\hat{\boldsymbol{\theta}})\}$ and $\mathrm{tr}\{\Lambda_P^{\mathrm{opt}}(\hat{\boldsymbol{\theta}})\}$ can be arbitrarily close.

4 Numerical experiments

We compare the estimation efficiency for the two subsampling procedures using both synthetic and real data sets.

Example 4 (Logistic regression). Form model $\mathbb{P}(y_i=1|\boldsymbol{x}_i) = e^{\theta_0+\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}_1}/(1+e^{\theta_0+\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}_1})$, $i=1,...,n$, we generate synthetic data sets by setting $n=10^5$, $\theta_0=0.5$, and $\boldsymbol{\theta}_1$ to be a 9 dimensional vector of 0.5. We consider the following three cases to generate $\boldsymbol{x}_i$. In Cases 1 and 3, the responses y_i are balanced, while in Case 2 about 98% of the data points are with $y_i=1$.

Case 1: **Normal**. Generate $\boldsymbol{x}_i$ from a multivariate normal distribution, $\mathbb{N}(\mathbf{0},\boldsymbol{\Sigma})$, where the (i,j)-th element of $\boldsymbol{\Sigma}$ is $\Sigma_{ij}=0.5^{I(i\neq j)}$ and $I()$ is the indicator function. This distribution is symmetric with light tails.

Case 2: **LogNormal**. Generate $\boldsymbol{v}_i$ from $\mathbb{N}(\mathbf{0},\boldsymbol{\Sigma})$ as defined in Case 1 and then set $\boldsymbol{x}_i=e^{\boldsymbol{v}_i}$, where the exponentiation is element-wise. This distribution is asymmetric and positively skewed.

Case 3: $\boldsymbol{T}_3$. We generate $\boldsymbol{x}_i$ from a multivariate t distribution with three degrees of freedom $t_3(\mathbf{0},\boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ defined in Case 1. This distribution is symmetric with heavy tails.

We also consider two real data sets: the covtype data from the LIBSVM data website (`https://www.csie.ntu.edu.tw/~cjlin/libsvm/`) and the SUSY data (Baldi et al., 2014), as Cases 4 and 5 below.

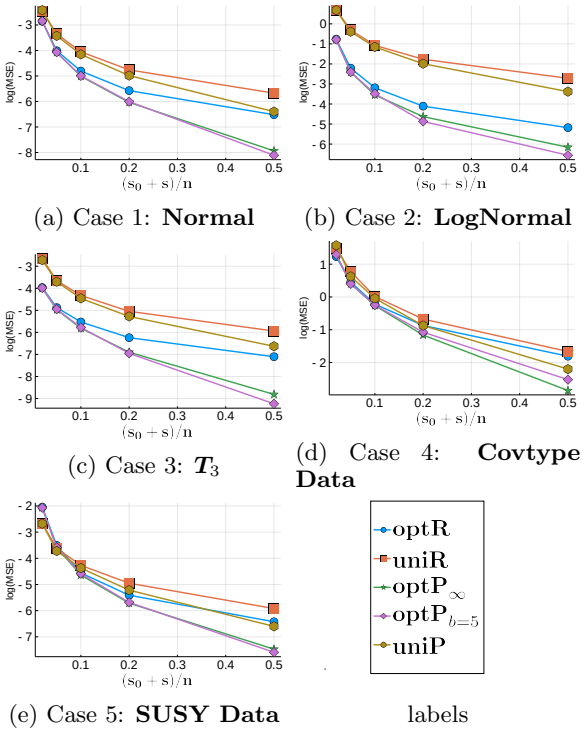
Case 4: **Covtype Data**. It has $n=581,012$ observations with about 48.76% of the responses are $y_i=1$. We use the ten quantitative covariate variables as $\boldsymbol{x}_i$'s.

Case 5: **SUSY Data**. It has $n=5,000,000$ observations with about 54.24% of the responses are $y_i=1$. We use the 18 kinematic features to classify whether new SUSY particles are produced.

We set $\alpha=0.1$, and choose $s_0=0.01n$ and different values for s so that the sampling ratio $(s_0+s)/n=$ 0.02, 0.05, 0.1, 0.2, and 0.5. Two different options of H^{0*} are considered: $H^{0*}=\|\dot{m}(Z_i^{0*},\tilde{\boldsymbol{\theta}}_P^{0*})\|_{\frac{s}{5n}}$ and $H^{0*}=\infty$. We aggregate the pilot estimator using the procedure described in Remark 7. For comparison, we also implement the uniform subsampling method with expected subsample sizes s_0+s. Newton's method is used for optimization on all subsamples. We repeat the simulation for $T=1000$ times to calculate the empirical mean squared error (MSE).

Figure 1 plots log(MSE) against $(s_0+s)/n$. When the subsampling ratio $(s_0+s)/n$ is close to zero, subsampling with replacement and Poisson subsampling

have similar performance for both approximated optimal subsampling and uniform subsampling. However, when $(s_0+s)/n$ gets larger, Poisson subsampling outperforms subsampling with replacement, and the improvement from subsampling with replacement to Poisson subsampling is more significant for approximated optimal subsampling than for uniform subsampling. For Poisson subsampling, the results for the two choices of H^{0*}, $H^{0*}=\infty$ and $H^{0*}=\|\dot{m}(Z_i^{0*},\tilde{\boldsymbol{\theta}}_P^{0*})\|_{\frac{s}{5n}}$, are similar when $(s_0+s)/n$ is small, but they start to differ for larger $(s_0+s)/n$.



(a) Case 1: **Normal** (b) Case 2: **LogNormal**

(c) Case 3: $\boldsymbol{T}_3$ (d) Case 4: **Covtype Data**

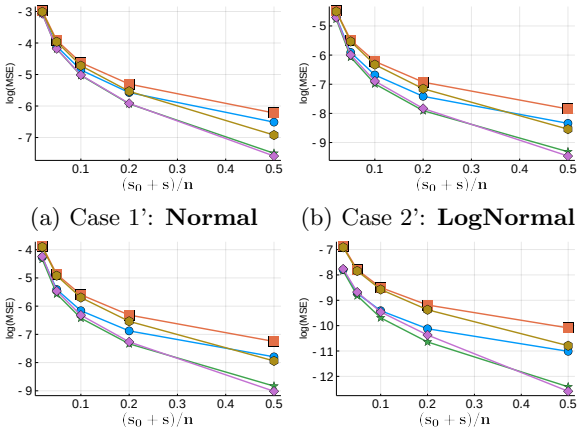
(e) Case 5: **SUSY Data**

Figure 1: Log(MSE) against $(s_0+s)/n$ for logistic regression. Here, "optR" means optimal subsampling with replacement; "uniR" means uniform subsampling with replacement; "optP$_\infty$" means approximated optimal Poisson subsampling with $H^{0*}=\infty$; "optP$_{b=5}$" means approximated optimal Poisson subsampling with $H^{0*}=\|\dot{m}(Z_i^{0*},\tilde{\boldsymbol{\theta}}_P^{0*})\|_{\frac{s}{5n}}$; and "uniP" means uniform Poisson subsampling.

Example 5 (Linear regression). We consider a linear model $y_i=\theta_0+\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}_1+\varepsilon_i$, $i=1,...,n$, with $n=10^5$, $\theta_0=1$, $\boldsymbol{\theta}_1$ being a 50 dimensional vector of ones, and ε_i being i.i.d. $\mathbb{N}(0,1)$. We use the same distributions in Cases 1-3 to generate $\boldsymbol{x}_i$ and refer them as Cases 1'-3'. We also consider a gas sensor data (Fonollosa et al., 2015) from the UCI data repository (Dheeru and Karra Taniskidou, 2017), as Case 6.

Case 6: **Gas Sensor Data**. After cleaning, the data contain $n=4,188,261$ readings on 15 sensors. We use log of readings from the last sensor as responses and log of other readings as covariates.

We use the same setup for α, s_0, s, and H^{0*}, as used in logistic regression. Figure 2 presents the results. The overall pattern in Figure 2 is similar to that in Figure 1. We observe that the advantage of Poisson subsmapling over subsampling with replacement is more significant for approximated optimal subsampling, and the advantage of Poisson subsampling compared with subsampling with replacement is more significant. For example, in Case 4', the synthetic data sets with $\boldsymbol{x}_i$'s from the t_3 distribution, the uniform Poisson subsampling can even outperform the approximated optimal subsampling with replacement when $(s_0+s)/n=0.5$. We also observe that approximated optimal subsampling methods outperform the uniform subsampling methods, and the gap between their performance in terms of estimation efficiency is larger for larger $(s_0+s)/n$. Another pattern is that when the approximated optimal subsampling probabilities are more nonuniform, their advantage over uniform subsampling is more significant.



(a) Case 1': **Normal** (b) Case 2': **LogNormal**

(c) Case 3': $\boldsymbol{T}_3$ (d) Case 6: **Gas Sensor**

Figure 2: Log Empirical MSEs (y-axis) against subsampling ratio $(s_0+s)/n$ (x-axis) for linear regression. Here, "optR" means optimal subsampling with replacement; "uniR" means uniform subsampling with replacement; "optP$_\infty$" means approximated optimal Poisson subsampling with $H^{0*}=\infty$; "optP$_{b=5}$" means approximated optimal Poisson subsampling with $H^{0*}=\|\dot{m}(Z_i^{0*},\tilde{\boldsymbol{\theta}}_P^{0*})\|_{\frac{s}{5n}}$; and "uniP" means uniform Poisson subsampling.

5 Conclusion and Discussion

In this paper, we derived optimal subsampling probabilities in the context of maximizing an additive target function for both subsampling with replacement

and Poisson subsampling. Theoretical and empirical results show that the two different subsampling procedure have similar performance when the subsampling ratio is small. However, when subsampling ratio does not converge to zero, Poisson subsampling has a higher estimation efficiency. One problem warrants for further investigation is how to chose the tuning parameter b in Algorithm 2 so that the approximated optimal subsampling probabilities produce an estimator with an asymptotic variance-covariance matrix that is near optimal even when the subsampling ratio does not converge to zero.

Acknowledgments

The authors are deeply grateful to Michael Mahoney for the insightful comments and suggestions that significantly improved the manuscript. The first author acknowledges NSF for providing partial support of this work.

References

Atkinson, A., Donev, A., and Tobias, R. (2007). *Optimum experimental designs, with SAS*, volume 34. Oxford University Press.

Baldi, P., Sadowski, P., and Whiteson, D. (2014). Searching for exotic particles in high-energy physics with deep learning. *Nature Communications*, 5(4308):http://dx.doi.org/10.1038/ncomms5308.

Campbell, T. and Broderick, T. (2018). Bayesian coreset construction via greedy iterative geodesic ascent. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 698–706, Stockholmsmässan, Stockholm Sweden. PMLR.

Campbell, T. and Broderick, T. (2019). Automated scalable bayesian inference via hilbert coresets. *The Journal of Machine Learning Research*, 20(1):551–588.

Clarkson, K. L. and Woodruff, D. P. (2013). Low rank approximation and regression in input sparsity time. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 81–90. ACM.

Dheeru, D. and Karra Taniskidou, E. (2017). UCI machine learning repository.

Drineas, P., Kannan, R., and Mahoney, M. W. (2006a). Fast monte carlo algorithms for matrices i: Approximating matrix multiplication. *SIAM Journal on Computing*, 36(1):132–157.

Drineas, P., Kannan, R., and Mahoney, M. W. (2006b). Fast monte carlo algorithms for matrices ii: Computing a low-rank approximation to a matrix. *SIAM Journal on Computing*, 36(1):158–183.

Drineas, P., Kannan, R., and Mahoney, M. W. (2006c). Fast monte carlo algorithms for matrices iii: Computing a compressed approximate matrix decomposition. *SIAM Journal on Computing*, 36(1):184–206.

Drineas, P., Magdon-Ismail, M., Mahoney, M., and Woodruff, D. (2012). Faster approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13:3475–3506.

Drineas, P., Mahoney, M., Muthukrishnan, S., and Sarlos, T. (2011). Faster least squares approximation. *Numerische Mathematik*, 117:219–249.

Efraimidis, P. S. and Spirakis, P. G. (2006). Weighted random sampling with a reservoir. *Information Processing Letters*, 97(5):181–185.

Fonollosa, J., Sheik, S., Huerta, R., and Marco, S. (2015). Reservoir computing compensates slow response of chemosensor arrays exposed to fast varying gas concentrations in continuous monitoring. *Sensors and Actuators B: Chemical*, 215:618–629.

Hesterberg, T. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–194.

Kleiner, A., Talwalkar, A., Sarkar, P., and Jordan, M. I. (2014). A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):795–816.

Lin, N. and Xie, R. (2011). Aggregated estimating equation estimation. *Statistics and Its Interface*, 4:73–83.

Ma, P., Mahoney, M., and Yu, B. (2015). A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research*, 16:861–911.

Mahoney, M. W. (2011). Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224.

Mahoney, M. W. and Drineas, P. (2009). CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697–702.

McWilliams, B., Krummenacher, G., Lucic, M., and Buhmann, J. M. (2014). Fast and robust least squares estimation in corrupted linear models. In *Advances in Neural Information Processing Systems*, pages 415–423.

Owen, A. and Zhou, Y. (2000). Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model assisted survey sampling*. Springer Science & Business Media.

Schifano, E. D., Wu, J., Wang, C., Yan, J., and Chen, M.-H. (2016). Online updating of statistical inference in the big data setting. *Technometrics*, 58(3):393–403.

Tillé, Y. (2019). A general result for selecting balanced unequal probability samples from a stream. *Information Processing Letters*, 152:105840.

Ting, D. and Brochu, E. (2018). Optimal subsampling with influence functions. In *Advances in Neural Information Processing Systems 31*, pages 3654–3663. Curran Associates, Inc.

Wang, H. (2019). More efficient estimation for logistic regression with optimal subsamples. *Journal of Machine Learning Research*, 20(132):1–59.

Wang, H. and Ma, Y. (2020). Optimal subsampling for quantile regression in big data. *Biometrika*, .:DOI:10.1093/biomet/asaa043.

Wang, H., Yang, M., and Stufken, J. (2019). Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association*, 114(525):393–405.

Wang, H., Zhu, R., and Ma, P. (2018). Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, 113(522):829–844.

Yang, T., Zhang, L., Jin, R., and Zhu, S. (2015). An explicit sampling dependent spectral error bound for column subset selection. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 135–143.

Yang, Y., Pilanci, M., and Wainwright, M. J. (2016). Randomized sketches for kernels: Fast and optimal non-parametric regression. *The Annals of Statistics,*, page forthcoming.

Yu, J., Wang, H., Ai, M., and Zhang, H. (2020). Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. *Journal of the American Statistical Association*, 0(0):1–12.

Zhu, R. (2016). Gradient-based sampling: An adaptive importance sampling for least-squares. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 406–414. Curran Associates, Inc.