Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: www.tandfonline.com/journals/ucgs20

# Optimal Subsampling for Data Streams with Measurement Constrained Categorical Responses

Jun Yu, Zhiqiang Ye, Mingyao Ai & Ping Ma

**To cite this article:** Jun Yu, Zhiqiang Ye, Mingyao Ai & Ping Ma (18 Dec 2024): Optimal Subsampling for Data Streams with Measurement Constrained Categorical Responses, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2024.2421990

**To link to this article:** https://doi.org/10.1080/10618600.2024.2421990

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2024, VOL. 00, NO. 0, 1–11
https://doi.org/10.1080/10618600.2024.2421990

Taylor & Francis
Taylor & Francis Group

# Optimal Subsampling for Data Streams with Measurement Constrained Categorical Responses

Jun Yu[a], Zhiqiang Ye[b], Mingyao Ai[b], and Ping Ma[c]

[a]School of Mathematics and Statistics, Beijing Institute of Technology, Beijing, China; [b]LMAM, School of Mathematical Sciences and Center for Statistical Science, Peking University, Beijing, China; [c]Department of Statistics, University of Georgia, Athens, GA

**ABSTRACT**
High-velocity, large-scale data streams have become pervasive. Frequently, the associated labels for such data prove costly to measure and are not always available upfront. Consequently, the analysis of such data poses a significant challenge. In this article, we develop a method that addresses this challenge by employing an online subsampling procedure and a multinomial logistic model for efficient analysis of high-velocity, large-scale data streams. Our algorithm is designed to sequentially update parameter estimation based on the A-optimality criterion. Moreover, it significantly increases computational efficiency while imposing minimal storage requirements. Theoretical properties are rigorously established to quantify the asymptotic behavior of the estimator. The method's efficacy is further demonstrated through comprehensive numerical studies on both simulated and real-world datasets. Supplementary materials for this article are available online.

**ARTICLE HISTORY**
Received February 2024
Accepted October 2024

**KEYWORDS**
Massive data; Multinomial logistic model; Online updating; Poisson sampling; Semi-supervised learning

## 1. Introduction

Data streams represent a sequence of data points, with each point arriving sequentially—a phenomenon that has become ubiquitous in contemporary settings (Li et al. 2019; Xie, Bai, and Ma 2023). Unlike the analysis of static data, the examination of data streams presents two distinct challenges. First, data volumes often surpass the storage capacity of a single device, especially when dealing with high-velocity data streams. This poses a significant hurdle in storing the data points and providing timely analysis. Second, the process of labeling data can be both expensive and time-consuming, leading to a scenario where labels for many records may be unavailable, despite the presence of covariates. This challenge is commonly referred to as measurement constraints or semi-supervised learning in the statistical literature. For example, in the analysis of forest cover types, environmental information can be easily obtained by remote sensing, but relaying that information from the remote sensor to the forest cover types is usually handled manually and slower than the data collection velocity of the sensors. Addressing these multifaceted challenges is imperative to unlock the full potential of data-driven analytics.

To analyze high-velocity, unlabeled streaming data effectively, a desired statistical approach needs to meet the following three principles. First, online analytical processing should provide a timely and accurate response. Second, to solve the measurement constraint problem, the algorithm can sample a relatively small portion of the data batch at a time when the expensive responses will be measured and the resultant sampling estimator is statistically efficient. Finally, the analysis should reduce data storage, and the analytical processing tries to avoid accessing the historical data (Schifano et al. 2016).

To meet the first two principles, one natural solution is subsampling. Ample evidence shows that the subsampling efficiently balances the statistical accuracy and computational cost. Typical works in this line of research include but are not limited to, Wang, Zhu, and Ma (2018), Wang, Yang, and Stufken (2019), Han et al. (2020), and Wang et al. (2021). Particularly, the leverage score subsampling (Ma, Mahoney, and Yu 2015; Ma et al. 2022), Lowcon (Meng et al. 2020), and OSMUC subsampling (Zhang, Ning, and Ruppert 2021b) were proposed for linear/generalized linear model raise an efficient solution on analysis the massive data under measurement constraint. Recently, Wu et al. (2024) proposed a predictive-inference-based subsampling for semi-supervised learning. Unfortunately, these approaches typically assume access to the full sample before initiating the subsampling procedure. Thus, they may fail to give a timely response to analyzing high-velocity data streams. Moreover, most subsampling strategies do not consider how the selected subsample at the current stage affects the online data analysis in the future stages. Methods that hold promise in addressing the challenges of analyzing data streams are distributed subsampling methods. These methods divide the data into several subsets and then apply the subsampling method or subdata selection to each subset separately. Examples of such works include Yu et al. (2022) and Zhang and Wang (2021), along with mini-batch stochastic gradient descent methods proposed by Robbins and Monro (1951), Dekel et al. (2012), and Zhang et al. (2021a). Nevertheless, these methods largely adhere to the three principles when applied directly to high-velocity data streams, especially in cases where storing or accessing historical data is not feasible. In an online learning framework, the full dataset cannot be used by iterating over it, as unselected data points are discarded and cannot be revisited. As

**CONTACT** Mingyao Ai myai@math.pku.edu.cn School of Mathematical Sciences, Peking University, No. 5 Yiheyuan Rd, Haidian District, Beijing 10087, P.R. China.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

a result, conventional mini-batch stochastic gradient descent may become inefficient, and leveraging informative data points for parameter updates can naturally enhance statistical efficiency. However, the large sampling variance from simple random sampling leads to inexact gradient information, affecting the optimality, convergence, and overall implementation of algorithms. Therefore, an optimal sequential subsampling method and an online updating estimator, based on selected subsamples, are crucial to addressing the challenges posed by data storage and measurement constraints in large-scale, high-velocity data streams.

The primary goal of our work is to develop a sequential subsampling strategy for online estimating procedures for measurement constrained categorical responses when data arrives in a high-velocity stream. Here we restrict ourselves to categorical responses since they are prevalent in big data, particularly under measurement constraints. In pursuing this goal, we opt for the multinomial logistic model, widely recognized for its efficacy in modeling categorical responses across diverse scientific disciplines. Note that multinomial logistic models encompass logistic regression and softmax regression as special cases. In our proposed method, we introduce an online updating estimation approach based on the sequentially selected subsample with theoretical backups, including consistency and asymptotic normality. We employ a first-order gradient-based method to meet the needs of online updating without revisiting the historical information. Moreover, we develop optimal sequential subsampling procedures that minimize the asymptotic MSE of one-step updating and online updating with the number of blocks going to infinity, respectively. We show that consecutively applying the optimal probabilities for a one-step estimator does not yield an optimal solution for online data analysis. Furthermore, the first-order gradient-based method can further accelerate computation compared to Newton-type methods, making the method more responsive to high-velocity data streams. More precisely, the proposed method eliminates the need to update the Hessian matrix during online updating, making it particularly well-suited for scenarios where the selected subsampling in the data chunk is small, and the corresponding Hessian matrix does not exist.

The rest of this article is organized as follows. In Section 2, we introduce the multinomial logistic models and illustrate the general online subsampling and estimation procedure. The consistency of the resulting estimator is also derived. In Section 3, the optimal subsampling probabilities for one-step updating and online updating are derived. Some practical implementation of the optimal subsampling procedures is also considered. Simulation studies and real data analyses are provided in Section 4. Section 5 concludes this article. All proofs are postponed to the supplementary material.

## 2. Multinomial Logistic Model and its Online Updating

### 2.1. Multinomial Logistic Model

Let $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ be a sequence of independent and identically distributed random variables with response $y_i \in \{1, \dots, J\} (J \geq 2)$. The response $y_i$ can be regarded as a random variable coming from a multinomial distribution with $\pi_{ij}(\beta)$ being the conditional probability that the response $y_i$ is $j$ $(j = 1, \dots, J)$ given the covariate $\boldsymbol{x}_i$, $\boldsymbol{\beta}$ is the unknown parameter vector. More precisely, the $\pi_{ij}$ is characterized by the following parametric form:

$$\pi_{ij}(\boldsymbol{\beta}) = \frac{\exp(f(\boldsymbol{x}_i, \boldsymbol{\beta}_0, \boldsymbol{\beta}_j))}{1 + \sum_{k=1}^{J-1} \exp(f(\boldsymbol{x}_i, \boldsymbol{\beta}_0, \boldsymbol{\beta}_k))}, \quad \text{for } j = 1, \dots, J-1,$$

$$\pi_{iJ}(\boldsymbol{\beta}) = \frac{1}{1 + \sum_{k=1}^{J-1} \exp(f(\boldsymbol{x}_i, \boldsymbol{\beta}_0, \boldsymbol{\beta}_k))},$$

where $f(\cdot)$ is a known function, $\boldsymbol{\beta}_0 \in \mathbb{R}^{d_0}$ stands for the common parameters among all categories, $\boldsymbol{\beta}_j \in \mathbb{R}^{d_j} (j = 1, \dots, J-1)$ stands for the individual parameters belonging to the $j$th category only, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_0^{\mathrm{T}}, \dots, \boldsymbol{\beta}_{J-1}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^d$ being parameters of interest for the entire model. Here $d = d_0 + \cdots + d_{J-1}$ denotes the dimension of the parameter space. Consequently, we can rewrite the model as

$$\log\left(\frac{\pi_{ij}(\boldsymbol{\beta})}{\pi_{iJ}(\boldsymbol{\beta})}\right) = f(\boldsymbol{x}_i, \boldsymbol{\beta}_0, \boldsymbol{\beta}_j), \tag{1}$$

which is known as the multinomial logistic model with the baseline category link function. The true parameter vector $\boldsymbol{\beta}_{\text{true}}$ is the maximizer of the expected population likelihood

$$\mathrm{EL}(\boldsymbol{\beta}) = \mathrm{E}\left(\sum_{j=1}^{J-1} \mathbb{I}(y = j) f(\boldsymbol{x}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_j) - \log\left(1 + \sum_{j=1}^{J-1} \exp\left(f(\boldsymbol{x}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_j)\right)\right)\right), \tag{2}$$

where $\mathbb{I}(\cdot)$ is an indicator function. Clearly, when $f(\boldsymbol{x}_i, \beta_0, \boldsymbol{\beta}_j) = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_j$ with $\beta_0 = 0$, Model (1) reduces to be the softmax regression whose subsampling strategies are considered in Yao and Wang (2019). More specifically, for the case $J = 2$, Models (1) become the well-known logistic regression, and the subsampling methods are studied in Wang, Zhu, and Ma (2018). It is worth mentioning that the baseline category function is not the only link function associated with a multinomial logistic model. The cumulative link function, adjacent-categories link function, and continuation-ratio link functions are also widely adopted in practice. For clarity, we state the results for Model (1) in the main text and relegate the results for the other three links to the supplementary material.

### 2.2. Online Subsampling and Estimation Procedure

In the online semi-supervised learning framework, the full dataset is no longer provided all at once. Instead, it arrives in a data block $\{\boldsymbol{z}_{ti}\}_{i=1}^n$ for $t = 0, \dots, T$ from a large data stream, where $n$ is the number of observations in each block and $T$ is the number of blocks. We assume the first block contains all the labeled historical data, that is, $\boldsymbol{z}_{ti} = (\boldsymbol{x}_{ti}^{\mathrm{T}}, y_{ti})^{\mathrm{T}}$ for $t = 0$ and the unlabeled data arrives in the rest blocks with $\boldsymbol{z}_{ti} = \boldsymbol{x}_{ti}$ for $t = 1, \dots, T$. An online subsampling is referred to a method that selects $r \leq n$ unlabeled data from $t$th block $\{z_{ti}\}_{i=1}^{n}$ with $t$ ranging from 1 to $T$. Once the selected sample of each block is obtained, we can get access to the labels of them and then update the estimator via the labeled data in the current batch.

To develop an appealing online sampling strategy, we propose to use a gradient-based method that only requires the information on the previous estimator and the selected subsample points, and does not demand the whole data storage. Compared with the divide-and-conquer based approach, our method can handle the scenario where only a small portion of data in each block can be measured, even when the selected data points are less than the number of parameters. To be precise, the estimator of $\boldsymbol{\beta}$ is updated via the following formula when the $t$th block arrives:

$$\tilde{\boldsymbol{\beta}}^{(t)} = \tilde{\boldsymbol{\beta}}^{(t-1)} - \eta_t \tilde{\boldsymbol{g}}_t(\tilde{\boldsymbol{\beta}}^{(t-1)}), \tag{3}$$

where $\tilde{\boldsymbol{\beta}}^{(0)}$ is defined as the initial estimator obtained from the historical data, $\eta_t$ is the pre-specified learning rate, and $\tilde{g}_t(\boldsymbol{\beta})$ is the subsample estimator of the gradient, that is,

$$\boldsymbol{g}_t(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J}\mathbb{I}(y_{ti} = j)\frac{\partial \log \pi_{ti,j}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}. \tag{4}$$

The reasons we adopt (3) as the online updating equation are 2-fold. First, only the gradient and latest estimator are needed to be held in the memory. Thus, data storage is not necessary. Second, the gradient-based method is more computationally tractable compared with Newton-type methods, since it only requires the information on $\tilde{\boldsymbol{\beta}}^{(t-1)}$ and the selected subsample data points. It is worth mentioning that the proposed method can handle the case that only an extremely small portion of the data in the $t$th block can be measured, especially for the scenario that the selected data points are less than the number of parameters.

To further relax memory constraints, it is natural to select a subsample in the $t$th block through Poisson sampling (Särndal, Swensson, and Wretman 2003) since it only needs scanning the data line by line via independent Bernoulli trials. To ease the presentation, denote $\boldsymbol{u}_{ti}(\tilde{\boldsymbol{\beta}}^{(t-1)})$ be the negative score function of $i$th observation in $t$th block, that is, $-\sum_{j=1}^{J}\mathbb{I}(y_{ti} = j)\partial \log \pi_{ti,j}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$. The gradient $\boldsymbol{g}_t(\tilde{\boldsymbol{\beta}}^{(t-1)})$ is naturally estimated by the Horvitz-Thompson method (Horvitz and Thompson 1952), that is,

$$\begin{aligned}\tilde{\boldsymbol{g}}_t(\tilde{\boldsymbol{\beta}}^{(t-1)}) &= \sum_{i=1}^{n}\frac{R_{ti}}{np_{ti}}\boldsymbol{u}_{ti}(\tilde{\boldsymbol{\beta}}^{(t-1)}) \\ &= -\sum_{i=1}^{n}\frac{R_{ti}}{np_{ti}}\sum_{j=1}^{J}\mathbb{I}(y_{ti} = j)\frac{\partial \log \pi_{ti,j}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}},\end{aligned} \tag{5}$$

where $R_{ti}$ is an indicator function with $R_{ti} = 1$ if and only if the $i$th sample in the $t$th block is selected and $p_{ti}$ is the inclusion probability. We denote the expected number of subsampled items as $r$, which implies that the subsampling probabilities are selected such that $p_{t1} + \cdots + p_{tn} = r$. While the subsample sample size in each data block is random, as shown in Ai et al. (2021), the number of subsampled items, say $\hat{r}$, will naturally concentrate around $r$ with $\mathrm{pr}(|\hat{r} - r|/r \leq t) \leq 1 - 2\exp(-2r^2t^2/n)$. Thus, the computational cost is still under control. Without loss of generality, we assume the targeted subsample sizes are the same in all blocks. As suggested in Polyak and Juditsky (1992), the step sizes can be chosen by $\eta_t = \eta_0 t^{-\alpha} (t \geq 1)$ with $\eta_0$ and $\alpha$ being predetermined hyper-parameters. For clarity, the proposed method is summarized in Algorithm 1.

```
Algorithm 1: Online Poisson subsampling for multinomial logistic regression.

Input An initial estimator $\tilde{\boldsymbol{\beta}}^{(0)}$ and sampling probabilities $\{p_{ti}\}_{i=1,\ldots,n,t=1,\ldots,T}$.
for $t = 1, \ldots, T$ do
    Set $\tilde{\boldsymbol{\beta}}^{(t)} \leftarrow \tilde{\boldsymbol{\beta}}^{(t-1)}$.
    for $i = 1, \ldots, n$ do
        Generate a Bernoulli variable $R_{ti} \sim \mathrm{Bernoulli}(p_{ti})$.
        if $R_{ti} = 1$ then
            Measure the response $y_{ti}$ under covariate $\boldsymbol{x}_{ti}$ and update
                $\tilde{\boldsymbol{\beta}}^{(t)} \leftarrow \tilde{\boldsymbol{\beta}}^{(t-1)} - \frac{\eta_t}{np_{ti}}\boldsymbol{u}_{ti}(\tilde{\boldsymbol{\beta}}^{(t-1)})$.
Output: The estimator $\tilde{\boldsymbol{\beta}}^{(T)}$.
```

To establish our asymptotic results, we need the following regularity assumptions.

*Assumption 1.* The regression parameters lie in a compact parameter space $\Lambda$, with the true parameter vector $\boldsymbol{\beta}_{\mathrm{true}}$ being the inner points of this parameter space. Further assume that $\mathrm{E}\boldsymbol{g}_t(\boldsymbol{\beta})$ is well defined for all $\boldsymbol{\beta} \in \mathbb{R}^d$ and $\lim_{\|\boldsymbol{\beta}\|\to\infty}\mathrm{EL}(\boldsymbol{\beta}) = \infty$ with $\mathrm{EL}(\boldsymbol{\beta})$ and $\boldsymbol{g}_t(\boldsymbol{\beta})$ defining in (2) and (5), respectively.

*Assumption 2.* Suppose that $\mathrm{pr}(f(\boldsymbol{x}, \boldsymbol{\beta}) \neq f(\boldsymbol{x}, \boldsymbol{\beta}_{\mathrm{true}})) > 0$ with probability one for all $\boldsymbol{\beta} \neq \boldsymbol{\beta}_{\mathrm{true}}$ and $\boldsymbol{x}$ comes from $F_{\boldsymbol{x}}$, where $F_{\boldsymbol{x}}$ is the distribution of $\boldsymbol{x}$.

*Assumption 3.* Assume that for all $\boldsymbol{\beta} \in \Lambda$, the risk function (2) is twice continuously differentiable with respect to $\boldsymbol{\beta}$. Moreover, the matrix

$$\Psi = -\mathrm{E}\left(\sum_{j=1}^{J}\mathbb{I}(y = j)\frac{\partial^2 \log \pi_j(\boldsymbol{\beta}_{\mathrm{true}})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\mathrm{T}}}\right)$$

is positive definite.

*Assumption 4.* For $j = 1, \ldots, J-1$, the quantity $\|\partial f(\boldsymbol{x}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_j)/\partial \boldsymbol{\beta}\| \leq h(\boldsymbol{x})$, such that $\mathrm{E}\,(h^3(\boldsymbol{x})) < \infty$. Furthermore, $\mathrm{E}\,(\|\partial^2 f(\boldsymbol{x}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_j)/\partial \boldsymbol{\beta}^{\mathrm{T}}\partial \boldsymbol{\beta}\|_s) < \infty$, where $\|\cdot\|_s$ denotes the matrix spectrum norm.

*Assumption 5.* The following conditions hold with probability one:

$$\sup_t \mathrm{E}\left(\sum_{i=1}^{n}\frac{rh^2(\boldsymbol{x}_{ti})}{n^2 p_{ti}}\middle|\mathcal{F}_{t-1}\right) < \infty,$$

$$\sup_t \mathrm{E}\left(\frac{r^{3/2}}{n^3}\sum_{i=1}^{n}\frac{h^3(\boldsymbol{x}_{ti})}{p_{ti}^2}\middle|\mathcal{F}_{t-1}\right) < \infty,$$

where $\mathcal{F}_{t-1}$ be the $\sigma$-field generated by $\{(\boldsymbol{x}_{ki}^{\mathrm{T}}, y_{ki}, R_{ki})^{\mathrm{T}} : 1 \le k \le t-1, 1 \le i \le n\}$.

Assumptions 1 are the regularity conditions of multinomial logistic regressions. This assumption is to ensure the boundedness of the sequence $\tilde{\boldsymbol{\beta}}^{(t)} (t = 0, 1, \ldots)$. We also require the boundedness of $\boldsymbol{\beta}_{\text{true}}$ as in McCullagh and Nelder (1989). Assumption 2 is the identification condition to ensure that the $\boldsymbol{\beta}_{\text{true}}$ is unique, which is also required in Han et al. (2020). Assumption 3 can be satisfied when the risk function is convex. Assumptions 4 and 5 are the moment assumptions imposed on the full sample and subsample. Specifically, the envelop function $h(\boldsymbol{x})$ in Assumption 4 can be simply specified as $\|\boldsymbol{x}\|$ for the softmax regression problem. Similar assumptions are used in Wang, Zhu, and Ma (2018), and Yu et al. (2022). The reason for using the conditional expectation is that the subsampling probabilities $\{p_{ti}\}_{i=1}^{n}$ may depend on $\mathcal{F}_{t-1}$. A typical example is that the subsampling probabilities rely on the pilot estimator derived in the previous step.

At the end of this section, we establish the consistency of the estimator achieved by Algorithm 1.

*Theorem 2.1.* If Assumptions 1–5 hold, and $\eta_t = \eta_0 t^{-\alpha}$ for some constants $\eta_0 > 0, \alpha \in (1/2, 1)$, and $r/(\sum_{t=1} t^{-\alpha})^2 \to 0$, then $\tilde{\boldsymbol{\beta}}^{(T)}$ converges to $\boldsymbol{\beta}_{\text{true}}$ almost surely as $T \to \infty$.

Compared with the divide-and-conquer-based online updating methods, such as Schifano et al. (2016), in which the number of blocks is usually less than the root of the total sample size, our method focuses on the case that the number of blocks ($T$) goes to infinity and is probably faster than $n$. Moreover, Theorem 2.1 implies $r$ can be either fixed or go to infinity, which differs from most existing subsampling works. We believe this is more suitable for the online system which needs responsive service for data streams. It is worth mentioning that when $T$ is finite or much slower than $r^{1/3}$, the distributed subsampling method proposed in Yu et al. (2022) can be applied directly. Thus, we focus our attention on the scenario that $T \to \infty$. As for the case that $T$ goes to infinity with a fixed and small $n$, the subsampling is not essential since we can handle each block timely.

## 3. Subsampling Strategies for Online Updating

### 3.1. Optimal Poisson Subsampling for One-Step Updating

For the one-step updating in (4), simple calculation yields that $\mathrm{E}(\|\tilde{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}_{\text{true}}\|^2 \mid \tilde{\boldsymbol{\beta}}^{(t-1)}, X_t)$ can be decomposed into the following three terms: $\|\mathrm{E}(\tilde{\boldsymbol{\beta}}^{(t)} \mid \tilde{\boldsymbol{\beta}}^{(t-1)}, X_t) - \boldsymbol{\beta}_{\text{true}}\|^2$, $\mathrm{tr}[\mathrm{var}\{\mathrm{E}(\tilde{\boldsymbol{\beta}}^{(t)} | \tilde{\boldsymbol{\beta}}^{(t-1)}, X_t, Y_t) \mid \tilde{\boldsymbol{\beta}}^{(t-1)}, X_t\}]$, and $\mathrm{tr}[\mathrm{E}\{\mathrm{var}(\tilde{\boldsymbol{\beta}}^{(t)} \mid \tilde{\boldsymbol{\beta}}^{(t-1)}, X_t, Y_t) \mid \tilde{\boldsymbol{\beta}}^{(t-1)}, X_t\}]$. From the detailed calculation given in Section S.2.2, one can observe that the first two terms are not related to the subsampling probabilities and the third term corresponds to the sampling variance of $\tilde{\boldsymbol{g}}_t(\tilde{\boldsymbol{\beta}}^{(t-1)})$. Thus, it is desirable to design subsampling probabilities to well approximate the gradient information. Due to measurement constraints, the responses are unknown. Thus, we propose to minimize its expectation instead, that is, we resort to minimizing the matrix $\mathrm{E}\{\mathrm{var}(\tilde{\boldsymbol{g}}_t(\tilde{\boldsymbol{\beta}}^{(t-1)}) \mid \tilde{\boldsymbol{\beta}}^{(t-1)}, X_t, Y_t) \mid \tilde{\boldsymbol{\beta}}^{(t-1)}, X_t\}$, where $X_t = \{\boldsymbol{x}_{ti}\}_{i=1}^{n}$, $Y_t = \{y_{ti}\}_{i=1}^{n}$. In general, one cannot expect to optimize the matrix $\mathrm{E}\{\mathrm{var}(\tilde{\boldsymbol{g}}_t(\tilde{\boldsymbol{\beta}}^{(t-1)}) \mid \tilde{\boldsymbol{\beta}}^{(t-1)}, X_t, Y_t) \mid \tilde{\boldsymbol{\beta}}^{(t-1)}, X_t\}$ in the sense of the Loewner ordering (Horn and Johnson 2012). Therefore, we opt to minimize the trace of $\mathrm{E}\{\mathrm{var}(\tilde{\boldsymbol{g}}_t(\tilde{\boldsymbol{\beta}}^{(t-1)}) \mid \tilde{\boldsymbol{\beta}}^{(t-1)}, X_t, Y_t) \mid \tilde{\boldsymbol{\beta}}^{(t-1)}, X_t\}$, which is known as the $A$-optimality criterion in the design of experiments.

The explicit form of $\mathrm{E}\{\mathrm{var}(\tilde{\boldsymbol{g}}_t(\tilde{\boldsymbol{\beta}}^{(t-1)}) \mid \tilde{\boldsymbol{\beta}}^{(t-1)}, X_t, Y_t) \mid \tilde{\boldsymbol{\beta}}^{(t-1)}, X_t\}$ can be derived as follows.

*Lemma 3.1.* For each data block $(X_t, Y_t)$, we have

$$\mathrm{E}\{\mathrm{var}(\tilde{\boldsymbol{g}}_t(\tilde{\boldsymbol{\beta}}^{(t-1)}) \mid \tilde{\boldsymbol{\beta}}^{(t-1)}, X_t, Y_t) \mid \tilde{\boldsymbol{\beta}}^{(t-1)}, X_t\} = \frac{1}{n^2}\sum_{i=1}^{n}\frac{1-p_{ti}}{p_{ti}}\sum_{j=1}^{J}\pi_{ti,j}(\boldsymbol{\beta}_{\text{true}}) \frac{\partial \log \pi_{ti,j}(\tilde{\boldsymbol{\beta}}^{(t-1)})}{\partial \boldsymbol{\beta}}\left(\frac{\partial \log \pi_{ti,j}(\tilde{\boldsymbol{\beta}}^{(t-1)})}{\partial \boldsymbol{\beta}}\right)^{\mathrm{T}}. \tag{6}$$

From Lemma 3.1, the optimal subsampling probabilities that **m**inimize the expected **v**ariance of the **g**radient $\tilde{\boldsymbol{g}}_t(\tilde{\boldsymbol{\beta}}^{(t-1)})$, that is the trace of (6), are derived as follows. The corresponding subsampling probabilities are so-called "MVg".

*Theorem 3.2.* Let $\hbar_{ti}^{\mathrm{MVg}} = (\sum_{j=1}^{J}\pi_{ti,j}(\boldsymbol{\beta}_{true})\|\partial \log \pi_{ti,j}(\tilde{\boldsymbol{\beta}}^{(t-1)})/\partial\boldsymbol{\beta}\|^2)^{1/2}$. Then the asymptotic mean square error of the gradient, that is, $\mathrm{tr}(\mathrm{E}[\mathrm{var}\{\tilde{\boldsymbol{g}}_t(\tilde{\boldsymbol{\beta}}^{(t-1)}) \mid \tilde{\boldsymbol{\beta}}^{(t-1)}, X_t, Y_t\} \mid \tilde{\boldsymbol{\beta}}^{(t-1)}, X_t])$ attains its minimum if $p_{ti}$'s are chosen to be

$$p_{ti}^{\mathrm{MVg}} = r\frac{\hbar_{ti}^{\mathrm{MVg}} \wedge M_t^{\mathrm{MVg}}}{\sum_{j=1}^{n}\hbar_{tj}^{\mathrm{MVg}} \wedge M_t^{\mathrm{MVg}}}, \quad i = 1, \ldots, n,$$

where $M_t^{\mathrm{MVg}}$ is the maximum number such that $r(\hbar_{ti}^{\mathrm{MVg}} \wedge M_t^{\mathrm{MVg}}) \le \sum_{j=1}^{n}\hbar_{tj}^{\mathrm{MVg}} \wedge M_t^{\mathrm{MVg}}$ and $a \wedge b = \min(a, b)$.

*Remark 1.* In Theorem 3.2, the thresholds $M_t^{\mathrm{MVg}}$s' are used to ensure the inclusion probabilities lie between zero and one. When $r/n \to 0$, which is common in big data subsampling settings, one may expect that $r\hbar_{ti}^{\mathrm{MVg}}/(\sum_{j=1}^{n}\hbar_{tj}^{\mathrm{MVg}}) < 1$. As a result, the optimal subsampling probabilities in each data block reduce to $p_{ti}^{\mathrm{MVg}} = r\hbar_{ti}^{\mathrm{MVg}}/(\sum_{j=1}^{n}\hbar_{tj}^{\mathrm{MVg}})$. As for the cases that $r/n > 0$, it may exist some $i$ such that $r\hbar_{ti}^{\mathrm{MVg}}/(\sum_{j=1}^{n}\hbar_{tj}^{\mathrm{MVg}}) > 1$ for the $t$th block. The explicit form solution is presented given in the supplementary material. As pointed out in Yu et al. (2022), finding the exact solution of $M_t^{\mathrm{MVg}}$ needs $O(n + r\log r)$ times. To reduce the computational cost, the $M_t^{\mathrm{MVg}}$ can be estimated by some quantile of $\hbar_{ti}^{\mathrm{MVg}}$ under this scenario.

### 3.2. Optimal Poisson Subsampling for Online Updating

The subsampling method proposed in Section 3.1 leads to an estimator that achieves the smallest mean squared error of $\tilde{\boldsymbol{\beta}}^{(t)}$ in the $t$th step. One of the naive ideas is to adopt the proposed subsampling probabilities in each step recursively which is known as the greedy method in numerical analysis. It is well known that the greedy method will not always lead to an optimal solution for the final estimator (Cormen et al. 2009). Thus, some calibration is needed on the subsampling probabilities to achieve the minimal asymptotic mean squared error (AMSE) of $\boldsymbol{\beta}$ as $T \to \infty$.

In our online learning setting, we not only allow $p_{it}$ to depend on $\{\boldsymbol{x}_{ti}\}_{i=1}^{n}$ at the $t$th batch but also on $\mathcal{F}_{t-1}$ which is defined in Assumption 5. To calibrate the greedy subsampling probabilities obtained in Theorem 3.2, we restrict our attention to the following scenario in the rest of this article.

*Assumption 6.* The subsampling probabilities are designed such that $\mathrm{E}(\Gamma_t(\boldsymbol{\beta}_{\text{true}})|\mathcal{F}_{t-1})$ is a stable quantity against $t$, where $\Gamma_t(\boldsymbol{\beta}) = rn^{-2}\sum_{i=1}^{n} p_{ti}{}^{-1}\boldsymbol{u}_{ti}(\boldsymbol{\beta})\boldsymbol{u}_{ti}^{\mathrm{T}}(\boldsymbol{\beta})$. That is to say, as $T \to \infty$, $\mathrm{E}(\Gamma_t(\boldsymbol{\beta}_{\text{true}})|\mathcal{F}_{t-1}) \to \Gamma$ almost surely, where $\Gamma$ is a positive definite matrix.

This assumption is to ensure that the sampling effect is stable against $t$ and can be satisfied when the proposed subsampling probability in Theorems 3.2 or 3.4 (will be shown later) is adopted. Clearly, the uniform subsampling also satisfied this assumption under Assumption 2. The asymptotic normality of $\tilde{\boldsymbol{\beta}}^{(T)}$ as $T \to \infty$ is presented as follows.

*Theorem 3.3.* If Assumptions 1–6 hold, and $\eta_t = \eta_0 t^{-\alpha}$ for some constants $\eta_0 > 0, \alpha \in (1/2, 1)$, then the following results hold.

(i) As $T, n, r \to \infty$ with $r/(\sum_{t=1} t^{-\alpha})^2 \to 0$, and further assume that $r$ is functional independent of $t, T$, it follows that $(r^{1/\alpha}T)^{\alpha/2}(\tilde{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}_{\text{true}}) \to N(0, \eta_0 V)$, where $V$ is the unique solution of the following equation:

$$\Psi V + V\Psi = \Gamma, \tag{7}$$

with $\Psi$ defined in Assumption 3.

(ii) As $T, n \to \infty$ with fixed $r$, $T^{\alpha/2}(\tilde{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}_{\text{true}}) \to N(0, \eta_0 V_{\text{fix}})$, where $V_{\text{fix}}$ is the unique solution of the equation $\Psi V_{\text{fix}} + V_{\text{fix}}\Psi = r^{-1}\Gamma$.

Theorem 3.3 characterizes the asymptotic behavior when both the number of data blocks $T$ and the size of each block $n$ goes to infinity. These results enable us to calibrate the subsampling probabilities in Theorem 3.2 such that $\tilde{\boldsymbol{\beta}}^{(T)}$ achieves the **m**inimum asymptotic **v**ariance (and the minimum AMSE due to the asymptotic unbiasedness). Thus, we denote the sampling probabilities by "MV".

*Theorem 3.4.* Let $\hbar_{ti}^{\mathrm{MV}} = \{\sum_{j=1}^{J} \pi_{ti,j}(\boldsymbol{\beta}_{true})\|\Psi^{-1/2}(\partial \log \pi_{ti,j}(\boldsymbol{\beta}_{true})/\partial\boldsymbol{\beta})\|^2\}^{1/2}$. Then the AMSE of $\tilde{\boldsymbol{\beta}}^{(T)}$ attains its minimum if $p_{ti}$'s are chosen to be

$$p_{ti}^{\mathrm{MV}} = r\frac{\hbar_{ti}^{\mathrm{MV}} \wedge M_t^{\mathrm{MV}}}{\sum_{j=1}^{n} \hbar_{tj}^{\mathrm{MV}} \wedge M_t^{\mathrm{MV}}}, \quad t = 1, \dots, T, i = 1, \dots, n,$$

where $M_t^{\mathrm{MV}}$ is the maximum number such that $r(\hbar_{ti}^{\mathrm{MV}} \wedge M_t^{\mathrm{MV}}) \le \sum_{j=1}^{n} \hbar_{tj}^{\mathrm{MV}} \wedge M_t^{\mathrm{MV}}$.

Clearly, the proposed optimal subsampling probabilities depend only on the covariates, but not on the responses. Thus, it is applicable to scenarios with measurement constraints. Moreover, one can clearly see that the optimal subsampling probabilities for minimizing the AMSE of $\tilde{\boldsymbol{\beta}}^{(T)}$ are different from the optimal subsampling probabilities for minimizing the expected variance of the gradient in each step. In contrast to the subsampling probabilities which focus on well estimating the gradient information in each step, the Fisher information matrix also plays an important role in the subsampling probabilities to achieve the AMSE of $\tilde{\boldsymbol{\beta}}^{(T)}$.

To take a close look at the proposed subsampling probabilities, we take the logistic regression as an example. Clearly, it is a special case of Model (1) with $J = 2$ and $f(\boldsymbol{x}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_1$. As discussed in Remark 1, we focus on the cases that $r\hbar_{ti}^{\mathrm{MV}}/(\sum_{j=1}^{n} \hbar_{tj}^{\mathrm{MV}}) < 1$ for all $i$. Simple calculation yields that $p_{ti}^{\mathrm{MV}} \propto \{\pi_{ti,1}(\boldsymbol{\beta}_{\text{true}})(1 - \pi_{ti,1}(\boldsymbol{\beta}_{\text{true}}))\boldsymbol{x}_i^{\mathrm{T}}\Psi^{-1}\boldsymbol{x}_i\}^{1/2}$, which corresponds to the square root leverage score of logistic models (Lee 1987). Similar to the results of leverage score sampling for the linear model discussed in Ma et al. (2022), the data point that has a large impact on the prediction naturally contains more information. Thus, it is reasonable to select the large leverage score points to improve the estimation performance. For the one step updating considered in Section 3.1, the optimal subsampling probabilities turn into $p_{ti}^{MV} \propto \{\pi_{ti,1}(\boldsymbol{\beta}_{\text{true}})(1 - \pi_{ti,1}(\boldsymbol{\beta}_{\text{true}}))\}^{1/2}\|\boldsymbol{x}_i\|_2$ which can be regarded as $L$-optimal sampling probabilities under measurement constraints with $T = 1, r, n \to \infty$. It is also interesting to see that when the learning rate in (3) is selected as the second-order derivative of the loss function $L_{nT}(\boldsymbol{\beta})$ at $\boldsymbol{\beta}^{(t-1)}$, the optimal subsampling probabilities reduce to $p_{ti}^{\mathrm{MV}} \propto \{\pi_{ti,1}(\boldsymbol{\beta}_{\text{true}})(1 - \pi_{ti,1}(\boldsymbol{\beta}_{\text{true}}))\}^{1/2}\|\Psi^{-1}\boldsymbol{x}_i\|_2$, which is the $A$-optimal sampling probabilities under measurement constraints with $T = 1$, $r, n \to \infty$ (Zhang, Ning, and Ruppert 2021b). Compared with the optimal subsampling strategies for one-step updating described in Theorem 3.2, one can find that the greedy methods will not lead to the optimal solution as $T \to \infty$. Instead, the squared root leverage score subsampling is much more suitable for online learning.

## 3.3. *Practical Considerations*

Note that the optimal subsampling probabilities presented in both Theorems 3.2 and 3.4 depend on the unknown true parameter vector $\boldsymbol{\beta}_{\text{true}}$. Thus, the optimal subsampling strategies cannot be applied directly. A practical way is to replace the unknown true parameter in the $t$th iteration with the estimator obtained in the $(t-1)$th iteration, that is, $\tilde{\boldsymbol{\beta}}^{(t-1)}$, since it is a consistent estimator of $\boldsymbol{\beta}_{\text{true}}$ as $t$ goes to infinity. To better approximate the subsampling probabilities, we use a consistent estimator estimated by the historical data or the data in the first block as the initial values in Algorithm 1. In the streaming setting, if the historical data is limited, we can initialize $\Psi$ based on the historical data at first, and periodically update it using new stream subsamples to ensure that $\Psi$ is consistent. As suggested in Xie et al. (2019), an infrequent update of $\Psi$ with a small probability is sufficient.

In the big data era, it is common to see that the subsample budget $r$ in each data block is relatively small compared with

**Algorithm 2:** Optimal online Poisson subsampling for multinomial logistic regression.

```
**Input** Historical labeled data $\mathcal{F}_0 = \{(\boldsymbol{x}_{0i}^{\mathrm{T}}, y_{0i})^{\mathrm{T}}\}_{i=1}^{n}$ and unlabeled data stream $\{\boldsymbol{x}_{ti}\}_{i=1}^{n}$ with $t = 1, \ldots, T$. The (shrinkage) parameter $\rho$.
**Initial estimator** Find the maximum likelihood estimator $\tilde{\boldsymbol{\beta}}^{(0)}$ for multinomial logistic regression based on $\mathcal{F}_0$. Calculate $\hat{\Psi} = -\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{J}\mathbb{I}(y_{0i} = j)\frac{\partial^2 \log \pi_{0i,j}(\tilde{\boldsymbol{\beta}}^{(0)})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}\right)$.
for $t = 1, \ldots, T$ do
    Set $\tilde{\boldsymbol{\beta}}^{(t)} \leftarrow \tilde{\boldsymbol{\beta}}^{(t-1)}$.
    Calculate $\lambda^{\mathrm{MV}} \leftarrow \sum_{i=1}^{n}\left\{\sum_{j=1}^{J}\pi_{ti,j}(\tilde{\boldsymbol{\beta}}^{(t-1)})\|\hat{\Psi}^{-1/2}(\partial \log \pi_{ti,j}(\tilde{\boldsymbol{\beta}}^{(t-1)})/\partial\boldsymbol{\beta})\|^2\right\}^{1/2}$.
    for $i = 1, \ldots, n$ do
        Calculate $\grave{p}_{ti}^{\mathrm{MV}} \leftarrow \left[(1-\rho)\frac{r\left\{\sum_{j=1}^{J}\pi_{ti,j}(\tilde{\boldsymbol{\beta}}^{(t-1)})\|\hat{\Psi}^{-1/2}(\partial \log \pi_{ti,j}(\tilde{\boldsymbol{\beta}}^{(t-1)})/\partial\boldsymbol{\beta})\|^2\right\}^{1/2}}{\lambda^{\mathrm{MV}}} + \rho\frac{r}{n}\right] \wedge 1$.
        Generate a Bernoulli variable $R_{ti} \sim \mathrm{Bernoulli}(\grave{p}_{ti}^{\mathrm{MV}})$.
        if $R_{ti} = 1$ then
            Measure the response $y_{ti}$ under covariate $\boldsymbol{x}_{ti}$ and update
            $$\tilde{\boldsymbol{\beta}}^{(t)} \leftarrow \tilde{\boldsymbol{\beta}}^{(t-1)} - \frac{\eta_t}{n\grave{p}_{ti}^{\mathrm{MV}}}\boldsymbol{u}_{ti}(\tilde{\boldsymbol{\beta}}^{(t-1)}).$$
**Output:** The estimator $\tilde{\boldsymbol{\beta}}^{(T)}$.
```

the number of observations $n$ in each block. As a result, it is seldom to meet the event that $\hbar_{ti}^{\mathrm{MVg}} > M_t^{\mathrm{MVg}}$ or $\hbar_{ti}^{\mathrm{MV}} > M_t^{\mathrm{MV}}$ happens. In practice, when $r/n$ is relatively small, the $M_t^{\mathrm{MVg}}$ or $M_t^{\mathrm{MV}}$ can be taken as infinity with no significant effect on the optimal subsampling probabilities in Theorems 3.2 and 3.4.

Let $\lambda^{\mathrm{MVg}}$ or $\lambda^{\mathrm{MV}}$ denote the denominator for optimal one-step subsampling and optimal online subsampling, respectively. To make the estimators more stable and robust, Ma, Mahoney, and Yu (2015) first proposed a shrinkage-based subsampling method, which is a linear combination of the optimal probabilities and the uniform subsampling probabilities. Thereby it is natural to obtain the benefits of both two methods.

The computational cost is $O(nT)$ of Algorithm 2. The storage demand is $O(1)$ since the analysis only need to keep $\tilde{\boldsymbol{\beta}}^{(t-1)}$ and $\boldsymbol{u}_{ti}(\tilde{\boldsymbol{\beta}}^{(t-1)})$. When the pilot estimator or the initial guess is consistent, the denominators in the subsampling probabilities are very similar across different data blocks. Thus, we can use the first block to estimate the denominator. In this scenario, $\grave{p}_{ti}^{\mathrm{MV}}$ only depends on the $i$th data points itself together with the estimated denominator. Thus, the cost of the RAM is also $O(1)$ since the Poisson subsampling can be implemented by loading the data line-by-line. Note that the responses of the data stream are expensive. It should be emphasis that the main cost in our problem is the data annotation.

The results of consistency and asymptotic normality for the practical implementation are presented in the following theorems.

*Assumption 7.* Suppose that

$$\mathrm{E}\left(\frac{r}{n^2}\sum_{i=1}^{n}\frac{\boldsymbol{u}_{1i}(\boldsymbol{\beta}_{\mathrm{true}})\boldsymbol{u}_{1i}^{\mathrm{T}}(\boldsymbol{\beta}_{\mathrm{true}})}{\left\{(1-\rho)\frac{r\hbar_{1i}^{\mathrm{MV}}}{\sum_{i=1}^{n}\hbar_{1i}^{\mathrm{MV}}} + \rho\frac{r}{n}\right\} \wedge 1}\right) = \Gamma^{\mathrm{MV}} \qquad (8)$$

holds almost surely with $\Gamma^{\mathrm{MV}}$ being a positive definite matrix, where $\hbar_{1i}^{\mathrm{MV}}$ is defined in Theorem 3.4.

*Theorem 3.5.* If Assumptions 1–4 and 7 hold, and $\eta_t = \eta_0 t^{-\alpha}$ for some constants $\eta_0 > 0$, $\alpha \in (1/2, 1)$, then $\tilde{\boldsymbol{\beta}}^{(t)}$ achieved by Algorithm 1 with $\{\grave{p}_{ti}^{\mathrm{MV}}\}$ converges to $\boldsymbol{\beta}_{\mathrm{true}}$ almost surely as $t \to \infty$. Moreover, the following results hold.

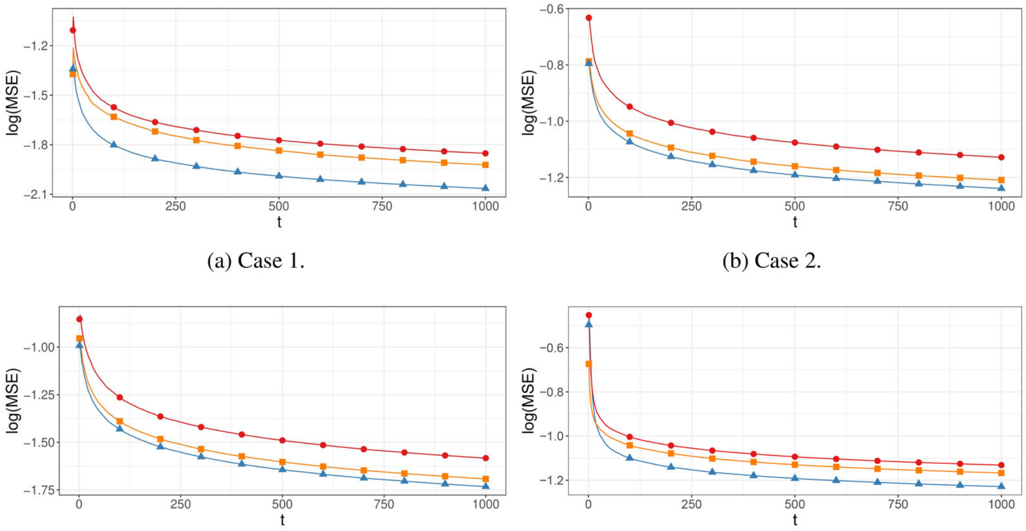
(i) As $T, n, r \to \infty$ with $r/(\sum_{t=1} t^{-\alpha})^2 \to 0$ and $r$ is functionally independent of $t, T$, $(r^{1/\alpha}T)^{\alpha/2}(\tilde{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}_{\mathrm{true}}) \to N(0, \eta_0 V_{\mathrm{prac}})$, where $V$ is the unique solution of the equation $\Psi V_{\mathrm{prac}} + V_{\mathrm{prac}}\Psi = \Gamma^{\mathrm{MV}}$.

(ii) As $T, n \to \infty$ with fixed $r$, $T^{\alpha/2}(\tilde{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}_{\mathrm{true}}) \to N(0, \eta_0 V_{\mathrm{prac,f}})$, where $V$ is the unique solution of the equation $\Psi V_{\mathrm{prac,f}} + V_{\mathrm{prac,f}}\Psi = r^{-1}\Gamma^{\mathrm{MV}}$.

The Assumption 7 is to ensure that when the $\boldsymbol{\beta}_{\mathrm{true}}$ is known, the proposed subsampling method will not lead to a trivial solution, that is, the $V_{\mathrm{prac}}$ goes to zero. Furthermore, if (8) also holds when $\hbar_{1i}^{\mathrm{MV}}$ is replaced by $\hbar_{1i}^{\mathrm{MVg}}$ and denote the expectation as $\Gamma^{\mathrm{MVg}}$ with $\Gamma^{\mathrm{MVg}}$ being a positive definite matrix. The results are also hold when $\{\grave{p}_{ti}^{\mathrm{MVg}}\}$ are adopted by replacing $\Gamma^{\mathrm{MV}}$ with $\Gamma^{\mathrm{MVg}}$.

## 4. Numerical Studies

This section is to evaluate the performances of the methods proposed in Section 3 via some simulated and real datasets. To account for the randomness in the subsampling phases, the implementation is repeated 1000 times for each case in this section. The empirical mean squared error (MSE) of the resultant estimator is used to evaluate the accuracy of the algorithms in estimating the model parameter. To be precise, for the $t$th iteration ($t = 1, \ldots, T$), we calculate $500^{-1}\sum_{k=1}^{500}\|\tilde{\boldsymbol{\beta}}_{\boldsymbol{p},k}^{(t)} - \boldsymbol{\beta}_{\mathrm{true}}\|^2$,



(a) Case 1.

(b) Case 2.

(c) Case 3.

(d) Case 4.

**Figure 1.** The log of MSEs for Model (9) with the number of updates based on the MV (triangle), MVg (square), and uniform subsampling (circle) methods.

where $\tilde{\boldsymbol{\beta}}^{(t)}_{\boldsymbol{p},k}$ is the estimator in the $t$th iteration with probability scheme $\boldsymbol{p} = \{p_{ti} : t = 1, \ldots, T; i = 1, \ldots, n\}$ from the $k$th subsample. All the computations are performed via R (R Core Team 2021).

### 4.1. Simulations

In the following, we take Model (1) with $J = 3$ as an example to illustrate our methods. More precisely, we consider the following model:

$$\log\left(\frac{\pi_{i1}(\boldsymbol{\beta})}{\pi_{i3}(\boldsymbol{\beta})}\right) = \boldsymbol{x}_{ti(0)}^{\mathrm{T}}\boldsymbol{\beta}_0 + \boldsymbol{x}_{ti(1)}^{\mathrm{T}}\boldsymbol{\beta}_1,$$
$$\log\left(\frac{\pi_{i2}(\boldsymbol{\beta})}{\pi_{i3}(\boldsymbol{\beta})}\right) = \boldsymbol{x}_{ti(0)}^{\mathrm{T}}\boldsymbol{\beta}_0 + \boldsymbol{x}_{ti(2)}^{\mathrm{T}}\boldsymbol{\beta}_2. \qquad (9)$$

Here we set the true parameter vector $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = (1, 1, 1, 1)^{\mathrm{T}}$. The number of data blocks is set to $T = 1000$ with $n = 1000$ observations in each block. The corresponding covariates $\boldsymbol{x}_{ti} = (\boldsymbol{x}_{ti(0)}^{\mathrm{T}}, \boldsymbol{x}_{ti(1)}^{\mathrm{T}}, \boldsymbol{x}_{ti(2)}^{\mathrm{T}})^{\mathrm{T}}$, are independent and identically distributed and generated as follows.

Case 1 The covariate $x$ follows a multivariate t distribution with degree four, that is, $t_4(\mathbf{0}, \Sigma_1)$, where $\Sigma_1$ is a matrix with all diagonal elements equal to one and off-diagonal elements equal to 0.5.

Case 2 The covariate $x$ follows a multivariate normal distribution $N(\mathbf{0}, 2\Sigma_1)$, where $\Sigma_1$ is the same as in Case 1.

Case 3 The covariate $x$ is the same as Case 1 expect the correlation structure is changed to $\Sigma_2$, that is, $t_4(\mathbf{0}, \Sigma_2)$. Here $\Sigma_2$ is a matrix with the $(i, j)$th entry being $0.5^{|i-j|}$.

Case 4 The covariate $x$ follows a mixture distribution $0.5N(\mathbf{0.5}, 2\Sigma_1) + 0.5t_4(\mathbf{0}, \Sigma_1)$.

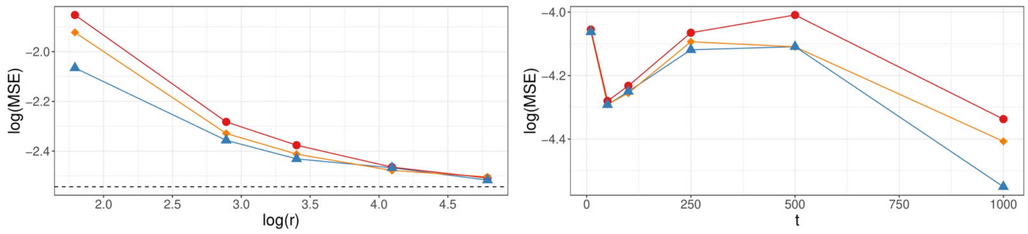
We first test our method in a multivariate t distribution case since it is common to see the heavy-tailed distribution in the big data era. Here we select four as the degree of multivariate t distribution since we need to ensure that the third order moment exists as we imposed in Assumption 4. Case 2 considers a lighter-tailed distribution compared with Case 1. The constant in front of the correlation matrix is to ensure the variances of the four cases are the same. Case 3 considers a different correlation structure. Case 4 further considers the case that covariates come from mixture distribution and are asymmetric.

Now we evaluate the performances of our methods based on different choices of subsampling probabilities. The uniform subsampling method is also considered for comparison. We regard the first block as historical data and set the hyperparameters $\rho = 0.2, \eta_0 = 1$, and $\alpha = 0.99$ throughout this section. To verify the convergence of the subsample-based estimator, the empirical log of MSEs with different numbers of updates are presented in Figure 1, where the number of updates $t$ ranges from 1 to 1000 and the target subsample size is fixed at $r = 6$.

From Figure 1, one can see that the MSEs for all subsampling methods decrease as $t$ increases, which confirms the theoretical result in Theorem 2.1. In addition, subsampling methods based on the MV and MVg always result in smaller empirical MSEs compared with uniform subsampling. Clearly, when the number of updates is small, the MVg method may perform better than the MV method. However, the MV method results in smaller MSEs than those in the MVg method as $t$ grows up. This is because the MVg method is a greedy approach and may not lead to the smallest MSE of the final estimator.

To evaluate the influence of the target subsample size $r$, we implement Algorithm 1 by fixing $T$ at 100, changing $r$ among 6, 18, 30, 60, and 120. Case 1 is selected as an example and the results are reported in Figures 2(a). The results for all the data in each block are also reported by the dashed line. We also illustrate the case that the total target subsample size, that is, $rT$, is fixed at 6000, which can explore the effect of the allocations between $r$ and $T$. The results are reported in Figure 2(b).

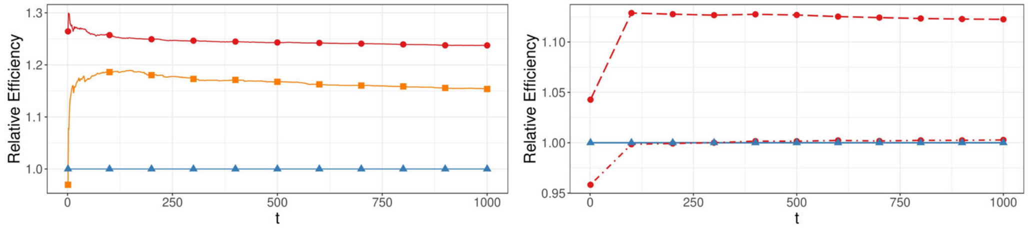
Clearly, the MSEs for all subsampling methods decrease as $r$ increases since more data are used. One can find that there is a linear relationship between the log(MSE) and log($r$) with



(a) Fixed $T = 1,000$.

(b) Fixed $rT = 6,000$.

**Figure 2.** The log of MSEs for Model (9) with different combinations of *r* and *T* based on the MV (triangle), MVg (square), and uniform subsampling (circle) methods, where $n_0$ in all cases are fixed at 1000 for a fair comparison. The dashed line in (a) is the log MSE when all data in each block have been used. The data is generated as mentioned in Case 1.



(a) Relative efficiency vs different method.

(b) Relative efficiency vs different $r$.

**Figure 3.** The relative efficiencies against the MV methods (fixed $r = 6$) for Model (9) with the number of updates. (a) The relative efficiencies of MVg (square), and uniform subsampling (circle) with *r* fixed at 6. (b) The relative efficiencies for different subsample sizes are based on the uniform subsampling method. The log dash line and dot-dash line stand for the cases $r = 8$ and 10, respectively. The data is generated as mentioned in Case 1.

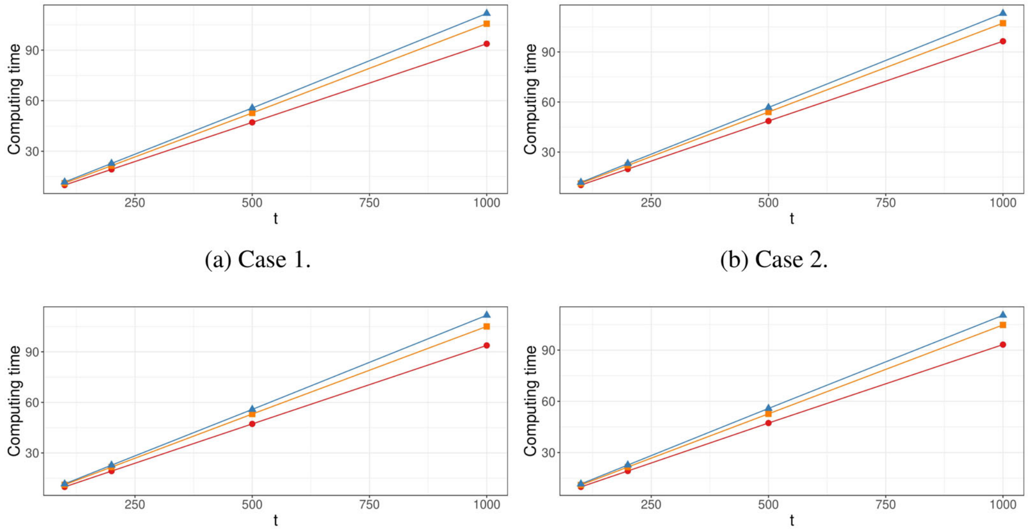
the slope around −0.5, which implies as *r* increasing the convergence rate is around $r^{1/2}$. This echoes the convergence results in the first part of Theorem 3.3. As *r* increased, the difference between our method and uniform sampling became smaller, since the uniform subsampling has more chance to select the informative points. One can see that when $r = 120$, the MSE is very close to the MSE that all the data in the block is used, although only 12% data in each block is used. This reveals that our methods are particularly useful for the scenario where the responses are expensive to be measured when the data is updated rapidly. In Figure 2(b), the $T$ ranges among 10, 50, 100, 250, 500, and 1000. One can see that the MVg and MV have similar performance when the number of blocks is small. This may be because the gradient-based method may not lead to a convergence result when the iteration is small. The phenomenon also reflects that the *r* can not be too big compared with $T$, which echos the assumption that $r/(\sum_{t=1} t^{-\alpha})^2 \to 0$ in Theorem 2.1.

Note that for the measurement constraints problem, the main cost is the data labeling. In the language of experimental design, relative design A-efficiency is a measure that compares the cost of any two design strategies. For example, the efficiency of design $\xi_1$ is $\kappa$ compared with design $\xi_2$, the design $\xi_1$ would need $100(\kappa - 1)\%$ more subjects than the design $\xi_2$ to achieve the same accuracy of estimators in the sense of asymptotic mean squared error. Thus, we also report the relative design A-efficiency calculated by the ratio of MSE for the two sampling strategies, namely uniform and MVg methods, over MSE for the MV method. The results for Case 1 are reported in Figure 3(a). For reference, we also drew a horizontal line at one. To numerically verify the effect of *r*, we also compared MV methods with the uniform subsampling method with $r = 8$ and 10. The corresponding results are reported in Figure 3(b).

From Figure 3, we can see that the MVg method is sometimes superior to the MV method in the first few rounds. This echos the results in Theorems 3.2 and 3.4. As time passed, the MV method is better than MVg, which implies the MV method is more suitable for continuous online updating under measurement constraints. Simple calculation yields that the relative efficiencies are around 115% and 105% for the uniform sampling and MVg methods, respectively. Moreover, from Figure 3(b), one may expect to label additional 60% samples in each iteration for the uniform subsampling method to achieve a similar performance.

To further evaluate the tradeoff between statistical efficiency and computational cost, we report the computing time plots for the four cases. All the computations are carried out on an iMac with a 3.6GHz Intel Core i9 processor. To mimic the cost of data annotating, we use `Sys.sleep()` function in `R` and assume the time spent on labeling each instance is only 0.01 sec. To mimic the situation that data arrives block by block, we use `fread()` function in `data.table` package (Dowle and Srinivasan 2021) to load the data in each block. The running time is recorded by the `Sys.time()` function and repeat each case 50 times. The averaged computing times are reported in Figure 4.

From Figure 4, we can see that the uniform subsampling method uses less time than the MV and MVg methods. This is because the uniform method does not need to compute the optimal subsampling probabilities. As expected, the computation time is very sensitive to the number of updates for all methods since the labeling cost takes most of the time even though we only assume that each subject only requires 0.01 sec to label. These facts echo the conclusion in Section 4.1 that the main cost is the data labeling for the measurement constraints problem. Since annotating large datasets is a tedious and time-intensive



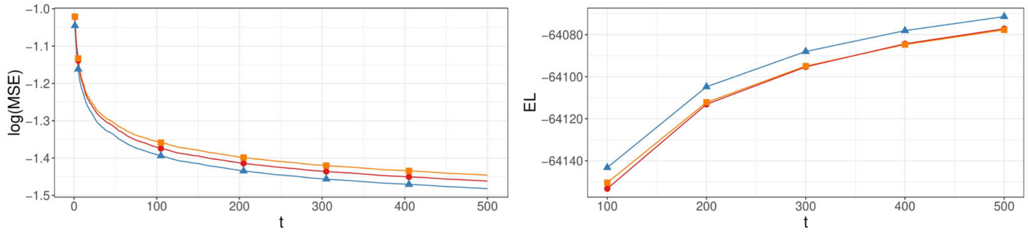
(a) Case 1.

(b) Case 2.

(c) Case 3.

(d) Case 4.

**Figure 4.** The Computational time for different numbers of updates based on the MV (triangle), MVg (square), and uniform subsampling (circle) methods.



(a) Log of MSEs.

(b) Expected log-likelihood gain.

**Figure 5.** The log of MSEs and expected log-likelihood gain on the testing set with different numbers of updates based on the MV (triangle), MVg (square), and uniform subsampling (circle) methods.

task, the MV subsampling method has its advantages in dealing with massive data under measurement constraints.

## 4.2. Forest Cover Type Dataset

The Forest cover type dataset from the UCI Machine Learning repository at *https://archive.ics.uci.edu/dataset/31/covertype* is being analyzed to predict the forest cover type using geometric information. The cover type is divided into seven categories in the research areas. Among the seven categories, spruce or fir, lodgepole pine are the primary major tree species with proportions of 36.5%, 48.8%, respectively. The primary goal is to distinguish forest cover type from spruce or fir (type 1), lodgepole pine (type 2), and the rest. The model uses elevation, aspect in degrees azimuth, slope in degrees, and horizontal distance to the nearest surface water feature as covariates. The data contains 581,012 instances, and the proportions of the seven cover types are given. The data is divided into a training set of 85% (501,000 instances) and a testing set of 15% (80,012 instances).

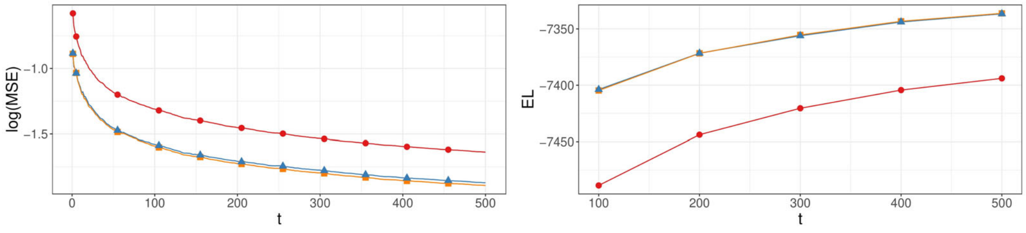
We again set the number of blocks to be 500, and use the first $n_0 = 1000$ instances in the training set as historical data. The target subsample size of each block is set to 30. The maximum likelihood estimator calculated by the entire training set is treated as "true" parameters. The empirical MSE is reported in Figure 5(a).

From Figure 5(a), it is clear that our methods outperform the uniform subsampling method. To measure the prediction performance, we report the expected log-likelihood gain (EL) as adopted in Ando and Chau Li (2017). To be precise, the EL is calculated by $\text{EL} = \sum_{i=1}^{n_{\text{test}}} \sum_{j=1}^{J} \mathbb{I}(y_i = j) \log \pi_{ij}(\tilde{\boldsymbol{\beta}}_{\mathbf{p},k}^{(t)})$ with $n_{\text{test}}$ being the size of the test data. Since it is the likelihood of the test data, a larger EL corresponds to a better method. As expected, the performance of the MV method in both estimation and prediction is better than the uniform subsampling method.

## 4.3. Census Income Dataset

We also apply our method on a census income dataset, which is provided in the UCI Machine Learning repository at *https://archive.ics.uci.edu/dataset/20/census+income*.

In this dataset, the responses are labeled as 1 and 2, where 1 stands for the person who makes less than or equal 50K a year and 2 for the rest. The following five covariates are considered: age ($x_1$), final weight ($x_2$), highest level of education ($x_3$), capital loss ($x_4$), and hours worked per week ($x_5$). Here, the final weight ($x_2$) represents the weights related to people who have similar socio-economic characteristics, which is measured by the Population Division at the Census Bureau. The variable $x_4$ is the loss of income due to some bad investments. Readers may refer to Kohavi (1996) for more details. A total of 48,842 observations



(a) Log of MSEs.

(b) Expected log-likelihood gain.

**Figure 6.** The log of MSEs and Expected log-likelihood gain on the testing set with the number of updates based on the MV (triangle), MVg (square), and uniform subsampling (circle) methods.

are in this dataset. The proportion of individuals whose income is less than or equal to 50K a year and exceeds 50K is 76.07% and 23.93%, respectively. In addition, this dataset is partitioned into a training set with 32,561 instances and a testing set with 16,281 instances by the UCI Machine Learning repository.

We adopt the baseline-category logit model under the proportional odds assumption here, which is also known as logistic regression. The regression coefficients calculated using the full training set are regarded as “true” parameters in this example. The total number of blocks is set to 500, equivalently the size of each block is 64. In addition, the first $n_0 = 561$ instances are used as historical data. The target subsample size is chosen as $r = 3$. The empirical MSE and expected log-likelihood gain for the MV, MVg, and uniform subsampling are shown in Figure 6.

As expected, the MV and MVg methods perform similarly and both of them dominate the uniform subsampling method. This pattern is similar to that in the simulation studies. We also consider the classification accuracy on the testing set with different numbers of updates when the classification probability threshold is selected as 0.5. The results are presented in Figure 6(b). It is clear to see that our methods also have better performance than the uniform subsampling method in prediction since our methods provide more accurate estimators. The superiority of the MV method can also be found in the sense of prediction, which agrees with Theorem 3.3.

## 5. Conclusion

In this article, we present an optimal online subsampling algorithm for multinomial logistic models and simultaneously introduce sequential updating estimators for the regression coefficients. Unlike most existing works which focus on the case that the full data are given before subsampling, our work gives a sequential subsampling solution to accommodate the online big data analysis. The proposed method breaks the storage barrier and the computational barrier under high-speed data streams and is particularly useful for data under measurement constraints. Note that the proposed methods need $\tilde{\boldsymbol{\beta}}^{(t)}$ together with a crude estimate of the Hessian matrix based on a pilot sample. It is natural to see the subsampling procedure can be performed at the sensor side with a small data transaction cost on synchronizing $\tilde{\boldsymbol{\beta}}^{(t)}$. Then perform the analysis on the data center side. Since solar-powered Internet of Things (IoT) devices are getting cheaper, our method can be a choice in reducing the data transaction. The consistency and asymptotic normality of the subsample-based estimator are derived. Based on the asymptotic variance of the resultant estimator, the optimal subsampling probabilities are obtained. Both theoretical and numerical results demonstrate the great potential of the proposed methods in extracting useful information from online data with categorical responses under measurement constraints.

In this work, we focus on the scenario that the number of data blocks grows to infinity at a rate faster than the size of each block, which is common in high-velocity data streams. As for the scenarios that the data arrival rate is slower, in the sense $T$ is finite or much slower than $r^{1/3}$, the distributed subsampling method proposed in Yu et al. (2022) is recommended to replace the proposed sampling probability by the OSMUC subsampling probability. Due to data storage constraints, the proposed method tries to avoid access to the historical data. Another interesting direction is how to further use the labeled subdata in the analysis processing when accessing the historically labeled subdata is allowed.

## Supplementary Materials

All technical proofs and additional simulation results are included in the supplementary material.

## Acknowledgments

The authors sincerely thank the editor, associate editor, and two referees for their valuable comments and insightful suggestions, which lead to further improvement of this work.

## Disclosure Statement

The authors report there are no competing interests to declare.

## Funding

Ai’s work is supported by NSFC grants 12071014 and 12131001, National Key R&D Program of China 2020YFE0204200, and LMEQF. Yu’s work is supported by NSFC grant 12471244 and Beijing Municipal Natural Science Foundation grant 1232019.

## ORCID

Mingyao Ai http://orcid.org/0000-0002-0421-0051

## References

Ai, M., Yu, J., Zhang, H., and Wang, H. (2021), "Optimal Subsampling Algorithms for Big Data Regressions," *Statistica Sinica*, 31, 749–772. [3]

Ando, T., and Chau Li, K. (2017), "A Weight-Relaxed Model Averaging Approach for High-Dimensional Generalized Linear Models," *The Annals of Statistics*, 45, 2654–2679. [9]

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2009), *Introduction to Algorithms*, Cambridge, MA: MIT Press. [5]

Dekel, O., Gilad-Bachrach, R., Shamir, O., and Xiao, L. (2012), "Optimal Distributed Online Prediction Using Mini-Batches," *Journal of Machine Learning Research*, 13, 165–202. [1]

Dowle, M., and Srinivasan, A. (2021), *data.table: Extension of 'data.frame'*, R package version 1.14.2. [8]

Han, L., Tan, K. M., Yang, T., and Zhang, T. (2020), "Local Uncertainty Sampling for Large-Scale Multiclass Logistic Regression," *The Annals of Statistics*, 48, 1770–1788. [1,4]

Horn, R. A., and Johnson, C. R. (2012), *Matrix Analysis* (2nd ed.), Cambridge: Cambridge University Press. [4]

Horvitz, D. G., and Thompson, D. J. (1952), "A Generalization of Sampling Without Replacement from a Finite Universe," *Journal of the American statistical Association*, 47, 663–685. [3]

Kohavi, R. (1996), "Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 202–207. [9]

Lee, A. H. (1987), "Diagnostic Displays for Assessing Leverage and Influence in Generalized Linear Models," *Australian Journal of Statistics*, 29, 233–243. [5]

Li, F., Xie, R., Wang, Z., Guo, L., Ye, J., Ma, P., and Song, W. (2019), "Online Distributed Iot Security Monitoring with Multidimensional Streaming Big Data," *IEEE Internet of Things Journal*, 7, 4387–4394. [1]

Ma, P., Chen, Y., Zhang, X., Xing, X., Ma, J., and Mahoney, M. W. (2022), "Asymptotic Analysis of Sampling Estimators for Randomized Numerical Linear Algebra Algorithms," *Journal of Machine Learning Research*, 23, 1–45. [1,5]

Ma, P., Mahoney, M. W, and Yu, B. (2015), "A Statistical Perspective on Algorithmic Leveraging," *Journal of Machine Learning Research*, 16, 861–911. [1,6]

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), Boca Raton, FL: Chapman & Hall/CRC. [4]

Meng, C., Xie, R., Mandal, A., Zhang, X., Zhong, W., and Ma, P. (in press), "Lowcon: A Design-based Subsampling Approach in a Misspecified Linear Model," *Journal of Computational and Graphical Statistics*. [1]

Polyak, B. T., and Juditsky, A. B. (1992), "Acceleration of Stochastic Approximation by Averaging," *SIAM Journal on Control and Optimization*, 30, 838–855. [3]

R Core Team. (2021), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. [7]

Robbins, H., and Monro, S. (1951), "A Stochastic Approximation Method," *The Annals of Mathematical Statistics*, 22, 400–407. [1]

Särndal, C.-E., Swensson, B., and Wretman, J. (2003), "*Model Assisted Survey Sampling* (2nd ed.), New York: Springer. [3]

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, 393–403. [1,4]

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, 393–405. [1]

Wang, H., Zhu, R., and Ma, P. (2018), "Optimal Subsampling for Large Sample Logistic Regression," *Journal of the American Statistical Association*, 113, 829–844. [1,2,4]

Wang, L., Elmstedt, J., Wong, W. K., and Xu, H. (2021), "Orthogonal Subsampling for Big Data Linear Regression," *The Annals of Applied Statistics*, 15, 1273–1290. [1]

Wu, X., Huo, Y., Ren, H., and Zou, C. (2024), "Optimal Subsampling via Predictive Inference," *Journal of the American Statistical Association*, 1–13. [1]

Xie, R., Bai, S., and Ma, P. (2023), "Optimal Sampling Designs for Multidimensional Streaming Time Series with Application to Power Grid Sensor Data," *The Annals of Applied Statistics*, 17, 3195–3215. [1]

Xie, R., Wang, Z., Bai, S., Ma, P., and Zhong, W. (2019), "Online Decentralized Leverage Score Sampling for Streaming Multidimensional Time Series," in *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2301–2311, PMLR. [5]

Yao, Y., and Wang, H. (2019), "Optimal Subsampling for Softmax Regression," *Statistical Papers*, 60, 585–599. [2]

Yu, J., Wang, H., Ai, M., and Zhang, H. (2022), "Optimal Distributed Subsampling for Maximum Quasi-Likelihood Estimators with Massive Data," *Journal of the American Statistical Association*, 117, 265–276. [1,4,10]

Zhang, A., Lipton, Z. C., Li, M., and Smola, A. J. (2021a), "Dive into Deep Learning," arXiv preprint arXiv:2106.11342. [1]

Zhang, H., and Wang, H. (2021), "Distributed Subdata Selection for Big Data via Sampling-based Approach," *Computational Statistics and Data Analysis*, 153, 107072. [1]

Zhang, T., Ning, Y., and Ruppert, D. (2021b), "Optimal Sampling for Generalized Linear Models Under Measurement Constraints," *Journal of Computational and Graphical Statistics*, 30, 106–114. [1,5]