Local Uncertainty Sampling for Large-Scale Multi-Class Logistic Regression

Lei Han[†], Ting Yang[†], Tong Zhang[†‡]

[†]Department of Statistics, Rutgers University, NJ, USA

[‡]Baidu Inc. Beijing, China

Abstract

A major challenge for building statistical models in the big data era is that the available data volume may exceed the computational capability. A common approach to solve this problem is to employ a subsampled dataset that can be handled by the available computational resources. In this paper, we propose a general subsampling scheme for large-scale multi-class logistic regression, and examine the variance of the resulting estimator. We show that asymptotically, the proposed method always achieves a smaller variance than that of the uniform random sampling. Moreover, when the classes are conditional imbalanced, significant improvement over uniform sampling can be achieved. Empirical performance of the proposed method is compared to other methods on both simulated and real-world datasets, and these results confirm our theoretical analysis.

1 Introduction

In recent years, data volume grows exponentially in the society, which has created demands for building statistical models with huge datasets. A major challenge is that the size of these datasets may exceed the available computational capability at hand. For example, when the dataset size is large, it may become infeasible to perform standard statistical procedures on a single machine. Although one remedy is to develop sophisticated distributed computing systems that can directly handle big data, the increased system complexity makes this approach not suitable for all scenarios. Another remedy to this problem is to employ a subsampled dataset that can be handled by the existing computational resources. This approach is widely applicable; however, since fewer data are used due to subsampling, statistical accuracy is lost. Therefore a natural question is to tradeoff computational efficiency and statistical accuracy by designing an effective sampling scheme that can minimize the reduction of statistical accuracy given a certain computational capacity.

In this paper, we examine the subsampling approach for solving big data multi-class logistic regression problems that are common in practical applications. The general idea of subsampling is to assign an accept-probability for each data point and select observations according to the assigned probabilities. After subsampling, only a small portion of the data are extracted from the full dataset, which means that the model built on the subsampled data will not be as accurate as that of the full data. The required computational resource can be measured by the number of subsampled data, and our key challenge is to design a good sampling scheme together with the corresponding estimation procedure so that the loss of statistical accuracy is minimized given a fixed number of sampled data.

There has been substantial work on subsampling methods for large-scale statistical estimation problems [6, 5, 7, 15, 21, 22, 23]. The simplest method is to subsample the data uniformly. However, uniform subsampling assigns the same acceptance probability to every data point, which fails to differentiate the importance among the samples. For example, a particular scenario, often encountered in practical applications of logistic regression, is when the class labels are imbalanced. This problem has attracted significant interests in the machine learning literature (see survey papers in[4, 9]). Generally, there are two types of commonly encountered class imbalance situations: *marginal imbalance* and *conditional imbalance*. In the case of marginal imbalance, some classes are much rarer than other classes. This situation often occurs in applications such as fraud and intrusion detection [1, 11], disease diagnoses [20], and protein fold classification [19], etc. On the other hand, conditional imbalance is the case that the label C of most observations $\boldsymbol{X}$ are very easy to predict. This happens in applications with highly accurate classifiers such as handwriting digits recognition [13] and email spam filtering. Note that marginally imbalance implies conditional imbalance, while the reverse is not necessarily true.

For binary marginal imbalance classification problems, case-control subsampling (CC), which uniformly selects an equal number of samples from each class, has been widely used in practice, including epidemiology and social science studies [14]. As a result, equal number of examples from each class are subsampled, and hence the sampled data are marginally balanced. It is known that case-control subsampling is more efficient than uniform subsampling when we deal with marginally imbalanced datasets. However, since the accept-probability relies on the response variable in CC subsampling, the distribution of subsampled data is skewed by the sample selection process [3]. It follows that correction methods are necessary to adjust the selection bias [2, 12]. Another method to remove bias in CC subsampling is to importance weight each sampled data point by the inverse of its acceptance probability. This is known as the weighted case-control method, which has been shown to be consistent and unbiased [10], but may increase the variance of the resulting estimator [18, 16, 17].

One drawback of the standard case-control subsampling is that it does not consider the situation where data are conditionally imbalanced. This issue was addressed in [8], which proposed an improved subsampling scheme called *local case-control* (LCC) sampling for binary logistic regression. The LCC method assigns each data point an acceptance probability determined not only by its label but also by the observation covariates. It puts more importance on data points that are easy to be mis-classified according to a pilot estimator (which is an approximate conditional probability estimator possibly obtained using a small number of uniformly sampled data). The method proposed in [8] tries to fit a logistic model with the LCC sampled data, and then apply a post-estimation correction to the resulting estimator using the pilot estimate. Therefore, the LCC sampling approach belongs to the correction based methods like [2, 12]. It was shown in [8] that the LCC estimator is consistent with an asymptotic variance that may significantly outperform that of the uniform sampling and CC based sampling methods when the data is strongly conditional imbalanced.

In this paper, we propose an effective sampling strategy for large-scale multi-class logistic regression problems that generalizes LCC. The general sampled estimation procedure can be summarized in the following two steps:

(a) Assign an accept-probability for each data point and select observations according to the assigned probabilities.

(b) Fit a logistic model with sampled observations to obtain the unknown model parameter.

In the above framework, the acceptance probability for each data point can be obtained using an

arbitrary probability function. Unlike correction based methods [12, 8] that are specialized for certain models, we propose a maximum likelihood estimate that integrates the correction into the MLE formulation, and this approach allows us to deal with arbitrary sampling probability and produces a consistent estimator within the original model family as long as the underlying logistic model is correctly specified. This new integrated estimation method avoids the post-estimation correction step used in the existing literature.

Based on this estimation framework, we propose a new sampling scheme that generalizes LCC as follows. Given a consistent pilot estimator, this scheme preferentially chooses data points with labels that are conditionally uncertain given their local observations based on the prediction of the pilot estimator. The proposed sampling strategy is therefore referred as Local Uncertainty Sampling (LUS). We show that the LUS estimator can achieve an asymptotic variance that is never worse than that of the uniform random sampling. That is, we can achieve variance of no more than $\gamma(\gamma \geq 1)$ times the variance of the full-sampled MLE by using no more than $1/\gamma$ of the sampled data in expectation. Moreover the required sample size can be significantly smaller than $1/\gamma$ of the full data when the classification accuracy of the pilot estimator is relatively high. This generalizes a result for LCC in [8], which reaches a similar conclusion for binary logistic regression when $\gamma \geq 2$. We conduct extensive empirical evaluations on both simulated and real-world datasets, showing that the experimental results match the theoretical conclusions, and the LUS method significantly outperforms the previous approaches in terms of both variance and accuracy.

Our main contributions can be summarized as follows.

- We propose a general estimation framework for large-scale multi-class logistic regression, which can be used with arbitrary sampling probabilities. The procedure always generates a consistent estimator within the original model family when the model is correctly specified. This method can be applied to general logistic models without the need of post-estimation corrections.

- Under this framework, we propose an efficient sampling scheme called local uncertainty sampling. For any $\gamma \geq 1$, the method can achieve asymptotic variance no more than that of the random subsampling with probability $1/\gamma$, using an expected sample size of no more than that of the random subsampling. Moreover the required sample size can be significantly smaller than that of the random subsampling when the the classification accuracy of the underlying problem is relatively high.

2 Preliminaries of Multi-Class Logistic Regression

For a K-class classification problem, we observe random data points $(\boldsymbol{x}, c) \in \mathbb{R}^d \times \{1, 2, \cdots, K\}$ from an unknown underlying distribution $\mathcal{D}$, where $\boldsymbol{x}$ is the feature vector and c is the corresponding label. The label c can be alternatively represented by a K-dimensional vector $[y_1, y_2, \cdots, y_K]$ with only one non-zero element $y_c = 1$ at the corresponding class label c. Given a set of n independently drawn observations $\{(\boldsymbol{X}_i, C_i) : i = 1, \cdots, n\}$ from $\mathcal{D}$, we want to estimate K conditional probabilities $\mathbb{P}_{\mathcal{D}}(C = k|\boldsymbol{X} = \boldsymbol{x})$, $k = 1, 2, \cdots, K$. This paper considers multi-class logistic model with the following parametric form:

$$\mathbb{P}_{\mathcal{D}}(C = k|\boldsymbol{X} = \boldsymbol{x}) = \frac{e^{f(\boldsymbol{x}, \boldsymbol{\theta}_k)}}{1 + \sum_{k'=1}^{K-1} e^{f(\boldsymbol{x}, \boldsymbol{\theta}_{k'})}}, \quad (k = 1, 2, \cdots, K-1)$$

$$\mathbb{P}_{\mathcal{D}}(C = K|\boldsymbol{X} = \boldsymbol{x}) = \frac{1}{1 + \sum_{k'=1}^{K-1} e^{f(\boldsymbol{x}, \boldsymbol{\theta}_{k'})}},$$

where each $\boldsymbol{\theta}_k$ is the model parameter for the k-th class. It implies that

$$f(\boldsymbol{x}, \boldsymbol{\theta}_k) = \log \frac{\mathbb{P}_{\mathcal{D}}(C = k|\boldsymbol{X} = \boldsymbol{x})}{\mathbb{P}_{\mathcal{D}}(C = K|\boldsymbol{X} = \boldsymbol{x})},\ k = 1, \cdots, K-1. \tag{1}$$

Let $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top, \cdots, \boldsymbol{\theta}_{K-1}^\top)^\top$ be the entire model parameter vector. The model in Eq. (1) is specified in terms of $K-1$ log-odds or logit transformations, with the constraint that the probabilities of each class should sum to one. Note that the logistic model uses one reference class as the denominator in the odds-ratios, and the choice of the denominator is arbitrary since the estimates are equivalent under this choice. This paper uses the last class as the reference class in the definition of odds-ratios.

When the underlying model is correctly specified, there exists a true parameter vector $\boldsymbol{\Theta}^* = (\boldsymbol{\theta}_1^{*\top}, \boldsymbol{\theta}_2^{*\top}, \cdots, \boldsymbol{\theta}_{K-1}^{*\top})^\top$ such that

$$f(\boldsymbol{x}, \boldsymbol{\theta}_k^*) = \log \frac{\mathbb{P}_{\mathcal{D}}(C = k|\boldsymbol{X} = \boldsymbol{x})}{\mathbb{P}_{\mathcal{D}}(C = K|\boldsymbol{X} = \boldsymbol{x})}, \tag{2}$$

and $\boldsymbol{\Theta}^*$ is the maximizer of the expected population likelihood:

$$\boldsymbol{\Theta}^* = \arg\max_{\boldsymbol{\Theta}}\ \mathbb{E}_{\mathcal{D}} \left[\sum_{k=1}^{K-1} Y_k \cdot f(\boldsymbol{X}, \boldsymbol{\theta}_k) - \log \left(1 + \sum_{k=1}^{K-1} e^{f(\boldsymbol{X}, \boldsymbol{\theta}_k)} \right) \right],$$

where $[Y_1, \ldots, Y_K] = Y$ is the vector representation of C introduced at the beginning of Section 2. In the maximum likelihood formulation of multi-class logistic regression, the unknown parameter $\boldsymbol{\Theta}^*$ is estimated from the data by maximizing the empirical likelihood:

$$\hat{\boldsymbol{\Theta}}_n = \arg\max_{\boldsymbol{\Theta}} \frac{1}{n} \sum_{i=1}^{n} \left[\sum_{k=1}^{K-1} Y_{i,k} \cdot f(\boldsymbol{X}_i, \boldsymbol{\theta}_k) - \log \left(1 + \sum_{k=1}^{K-1} e^{f(\boldsymbol{X}_i, \boldsymbol{\theta}_k)} \right) \right]. \tag{3}$$

For large-scale multi-class logistic regression problems, n can be extremely large. In such cases, solving the multi-class logistic regression problem (3) may be computationally infeasible due to the limitation of computational resources. To overcome this computational challenge, we will consider a subsampling framework next.

3 Model Parameter Estimation with Subsampling

In this section, we introduce the estimation framework with subsampling for multi-class logistic regression. The proposed approach contains the following steps.

(1) Given an arbitrary sampling probability function $a(\boldsymbol{x}, c) \in [0, 1]$ defined for all data points $(\boldsymbol{x}, c)$. For each $(\boldsymbol{X}_i, C_i)$ $(i = 1, \cdots, n)$, generate a random binary variable $Z_i \in \{0, 1\}$, drawn from the $\{0, 1\}$-valued Bernoulli distribution $\mathcal{B}(\boldsymbol{X}_i, C_i)$ with accept probability

$$\mathbb{P}_{\mathcal{B}(\boldsymbol{X}_i, C_i)}(Z_i = 1) = a(\boldsymbol{X}_i, C_i).$$

(2) Keep the samples with $Z_i = 1$ for $i \in \{1, \cdots, n\}$. Fit a multi-class logistic regression model based on the selected examples by solving the following optimization problem

$$\max_{\boldsymbol{\Theta}} \frac{1}{n} \sum_{i=1}^{n} Z_i \left[\mathbb{I}(C_i \neq K) \cdot f(\boldsymbol{X}_i, \boldsymbol{\theta}_{C_i}) + \log \frac{a(\boldsymbol{X}_i, C_i)}{a(\boldsymbol{X}_i, K)} - \log \left(1 + \sum_{k=1}^{K-1} e^{f(\boldsymbol{X}_i, \boldsymbol{\theta}_k) + \log \frac{a(\boldsymbol{X}_i, k)}{a(\boldsymbol{X}_i, K)}} \right) \right], \tag{4}$$

where $\mathbb{I}(\cdot)$ is the indicator function.

In the following, we shall derive Eq. (4), under the assumption that the logistic model is correctly specified as in Eq. (2). As we will show later, the acceptance probability used in the first step can be an arbitrary function, and the above method always produces a consistent estimator for the original population. The computational complexity in the second step is reduced to $\sum_{i=1}^{n} Z_i$ samples after the subsampling step.

Given $(\boldsymbol{X}, C)$, we may draw Z according to the Bernoulli distribution $\mathcal{B}(\boldsymbol{X}, C)$. This gives the following augmented distribution $\mathcal{A}$ for the joint random variables $(\boldsymbol{X}, C, Z) \in \mathbb{R}^p \times \{1, 2, \ldots, K\} \times \{0, 1\}$ with probability function

$$\mathbb{P}_{\mathcal{A}}(\boldsymbol{X} = \boldsymbol{x}, C = k, Z = z) = \mathbb{P}_{\mathcal{D}}(\boldsymbol{X} = \boldsymbol{x}, C = k)[a(\boldsymbol{x}, k)\mathbb{I}(z = 1) + (1 - a(\boldsymbol{x}, k))\mathbb{I}(z = 0)],$$

where $\mathbb{I}(\cdot)$ is the indicator function. Note that each sampled data $(\boldsymbol{X}_i, C_i) \sim \mathcal{D}$, and Z_i is independently drawn from $\mathcal{B}((\boldsymbol{X}_i, C = C_i)$, it follows that each data point $(\boldsymbol{X}_i, C_i, Z_i)$ is drawn i.i.d. from the distribution $\mathcal{A}$. For the sampled data $(\boldsymbol{X}_i, C_i)$ with $Z_i = 1$, the distribution of random variable $(\boldsymbol{X}, C)$ follows from

$$\mathbb{P}_{\mathcal{A}}(\boldsymbol{X} = \boldsymbol{x}, C = k | Z = 1) \propto \mathbb{P}_{\mathcal{D}}(\boldsymbol{X} = \boldsymbol{x}, C = k) a(\boldsymbol{x}, k).$$

Therefore, we have

$$\log \frac{\mathbb{P}_{\mathcal{A}}(C = k | \boldsymbol{X} = \boldsymbol{x}, Z = 1)}{\mathbb{P}_{\mathcal{A}}(C = K | \boldsymbol{X} = \boldsymbol{x}, Z = 1)} = \log \frac{\mathbb{P}_{\mathcal{D}}(C = k | \boldsymbol{X} = \boldsymbol{x})}{\mathbb{P}_{\mathcal{D}}(C = K | \boldsymbol{X} = \boldsymbol{x})} + \log \frac{a(\boldsymbol{x}, k)}{a(\boldsymbol{x}, K)}.$$

If $f(\cdot)$ is correctly specified for $\mathcal{D}$, then the following function family

$$g(\boldsymbol{x}, \boldsymbol{\theta}_k) = f(\boldsymbol{x}, \boldsymbol{\theta}_k) + \log \frac{a(\boldsymbol{x}, k)}{a(\boldsymbol{x}, K)} \tag{5}$$

is correctly specified for $\mathcal{A}$, i.e., the true parameter $\boldsymbol{\Theta}^*$ in Eq. (2) also satisfies

$$g(\boldsymbol{x}, \boldsymbol{\theta}_k^*) = \log \frac{\mathbb{P}_{\mathcal{A}}(C = k | \boldsymbol{X} = \boldsymbol{x}, Z = 1)}{\mathbb{P}_{\mathcal{A}}(C = K | \boldsymbol{X} = \boldsymbol{x}, Z = 1)}, \quad k = 1, 2, \ldots, K - 1.$$

Therefore, we have the following logistic model under $\mathcal{A}$:

$$\begin{aligned}
\mathbb{P}_{\mathcal{A}}(C = k | \boldsymbol{X} = \boldsymbol{x}, Z = 1) &= \frac{e^{g(\boldsymbol{x}, \boldsymbol{\theta}_k)}}{1 + \sum_{k=1}^{K-1} e^{g(\boldsymbol{x}, \boldsymbol{\theta}_k)}}, \quad k = 1, 2, \ldots, K - 1, \\
\mathbb{P}_{\mathcal{A}}(C = K | \boldsymbol{X} = \boldsymbol{x}, Z = 1) &= \frac{1}{1 + \sum_{k=1}^{K-1} e^{g(\boldsymbol{x}, \boldsymbol{\theta}_k)}}.
\end{aligned}$$

It follows that $\boldsymbol{\Theta}$ can be obtained by using MLE with respect to the new population $\mathcal{A}$:

$$\boldsymbol{\Theta}^* = \arg\max_{\boldsymbol{\Theta}} R(\boldsymbol{\Theta}),$$

where

$$R(\boldsymbol{\Theta}) := \mathbb{E}_{(\boldsymbol{X}, C, Z) \sim \mathcal{A}} \, Z \left[\sum_{k=1}^{K-1} Y_k \cdot g(\boldsymbol{X}, \boldsymbol{\theta}_k) - \log \left(1 + \sum_{k=1}^{K-1} e^{g(\boldsymbol{X}, \boldsymbol{\theta}_k)} \right) \right].$$

Practically, the model parameter $\boldsymbol{\Theta}^*$ can be estimated by empirical conditional MLE with respect to the sampled data $\{(\boldsymbol{X}_i, C_i, Z_i) : i = 1, \ldots, n\}$ as

$$\hat{\boldsymbol{\Theta}}_{Sub} = \arg\max_{\boldsymbol{\Theta}} \hat{R}(\boldsymbol{\Theta}),$$

where

$$\hat{R}_n(\boldsymbol{\Theta}) := \frac{1}{n}\sum_{i=1}^{n} Z_i \left[\sum_{k=1}^{K-1} Y_{i,k}\cdot g(\boldsymbol{X}_i,\boldsymbol{\theta}_k) - \log\left(1+\sum_{k=1}^{K-1} e^{g(\boldsymbol{X}_i,\boldsymbol{\theta}_k)}\right)\right]. \tag{6}$$

This is equivalent to Eq. (4). As we will see later, the resulting $\hat{\boldsymbol{\Theta}}_{Sub}$ is a consistent estimator of $\boldsymbol{\Theta}^*$.

4 Asymptotic Analysis

In this section, we examine the asymptotic behavior of the method in Section 3. First, based on the empirical likelihood in Eq. (6), we have the following result for the subsampling based estimator $\hat{\boldsymbol{\Theta}}_{Sub}$.

Theorem 4.1 (Consistency and Asymptotic Normality)**.** *Suppose that the parameter space is compact and $\forall \boldsymbol{\Theta} \neq \boldsymbol{\Theta}^*$ such that we have $\mathbb{P}_{\mathcal{D}}(f(\boldsymbol{X},\boldsymbol{\Theta}) \neq f(\boldsymbol{X},\boldsymbol{\Theta}^*)) > 0$. Moreover, assume the quantities $\|\nabla_{\boldsymbol{\theta}_k} f(\boldsymbol{x},\boldsymbol{\theta}_k)\|$, $\|\nabla^2_{\boldsymbol{\theta}_k} f(\boldsymbol{x},\boldsymbol{\theta}_k)\|$ and $\|\nabla^3_{\boldsymbol{\theta}_k} f(\boldsymbol{x},\boldsymbol{\theta}_k)\|$ for $k = 1,\cdots,K-1$ are bounded. Let $p(\boldsymbol{x},k) = \mathbb{P}_{\mathcal{D}}(C=k|\boldsymbol{X}=\boldsymbol{x})$ and $p_*(\boldsymbol{x},k) = a(\boldsymbol{x},k)p(\boldsymbol{x},k)$. If Eq. (2) is satisfied, then given an arbitrary sampling probability $a(\boldsymbol{x},c)$, as $n\to\infty$, the following claims hold:*

(1) $\hat{\boldsymbol{\Theta}}_{Sub}$ converges to $\boldsymbol{\Theta}^*$;

(2) $\hat{\boldsymbol{\Theta}}_{Sub}$ follows the asymptotic normal distribution:

$$\sqrt{n}\left(\hat{\boldsymbol{\Theta}}_{Sub} - \boldsymbol{\Theta}^*\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \left[\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}} \boldsymbol{\nabla}\mathbf{S}\boldsymbol{\nabla}^\top\right]^{-1}\right), \tag{7}$$

where

$$\boldsymbol{\nabla} = \begin{bmatrix} \nabla_{\boldsymbol{\theta}_1} f(\boldsymbol{x},\boldsymbol{\Theta}^*) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \nabla_{\boldsymbol{\theta}_2} f(\boldsymbol{x},\boldsymbol{\Theta}^*) & \cdots & \mathbf{0} \\ \cdots & \cdots & \ddots & \cdots \\ \mathbf{0} & \mathbf{0} & \cdots & \nabla_{\boldsymbol{\theta}_{K-1}} f(\boldsymbol{x},\boldsymbol{\Theta}^*) \end{bmatrix},$$

$$\mathbf{S} = \begin{bmatrix} p_*(\boldsymbol{x},1) & 0 & \cdots & 0 \\ 0 & p_*(\boldsymbol{x},2) & \cdots & 0 \\ \cdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & p_*(\boldsymbol{x},K-1) \end{bmatrix} - \frac{1}{\sum_{k=1}^{K} p_*(\boldsymbol{x},k)} \begin{bmatrix} p_*(\boldsymbol{x},1) \\ p_*(\boldsymbol{x},2) \\ \cdots \\ p_*(\boldsymbol{x},K-1) \end{bmatrix} \begin{bmatrix} p_*(\boldsymbol{x},1) \\ p_*(\boldsymbol{x},2) \\ \cdots \\ p_*(\boldsymbol{x},K-1) \end{bmatrix}^\top.$$

Theorem 4.1 shows that given an arbitrary sampling probability $a(\boldsymbol{x},k)$, the method in Section 3 can generate a consistent estimator $\hat{\boldsymbol{\Theta}}_{Sub}$ without post-estimation correction (as long as the logistic model is correctly specified). This is different from earlier methods such as the LCC method of [8] which employs post-estimation corrections. One benefit of the method proposed in this paper is that we can still produce a consistent estimator in the original model, and our framework allows different sampling functions for different data points $(\boldsymbol{X}_i, C_i)$. For example in time series analysis,

we may want to sample the older data more aggressively than the more recent data. This can be naturally handled in our framework, but will be impossible to handle using the earlier post-estimation correction approach. Another benefit is that the framework can be naturally applied with regularization, because regularization can be regarded as a restriction on the parameter space for $\boldsymbol{\Theta}$. However, post-estimation correction based methods can not be directly applied to regularized estimators.

From Theorem 4.1, the resulted estimator $\hat{\boldsymbol{\Theta}}_{Sub}$ follows the asymptotic normal distribution in Eq. (7) with zero mean and a variance of $\left[\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\boldsymbol{\nabla}\mathbf{S}\boldsymbol{\nabla}^\top\right]^{-1}$. Although given a data point $\boldsymbol{x}$, the sampling probability $a(\boldsymbol{x}, c)$ can be arbitrary probability, it is natural to select a sampling probability such that the variance is as small as possible. In the following, we study a specific choice of $a(\boldsymbol{x}, c)$ that achieves the purpose.

Denote by $\mathbf{S}_{full}$ the corresponding S matrix when we set $a(\boldsymbol{x}, k) \equiv 1$ (i.e., we accept all data points in the dataset), then

$$\mathbf{S}_{full} = \begin{bmatrix} p(\boldsymbol{x},1) & 0 & \cdots & 0 \\ 0 & p(\boldsymbol{x},2) & \cdots & 0 \\ \cdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & p(\boldsymbol{x},K-1) \end{bmatrix} - \begin{bmatrix} p(\boldsymbol{x},1) \\ p(\boldsymbol{x},2) \\ \cdots \\ p(\boldsymbol{x},K-1) \end{bmatrix} \begin{bmatrix} p(\boldsymbol{x},1) \\ p(\boldsymbol{x},2) \\ \cdots \\ p(\boldsymbol{x},K-1) \end{bmatrix}^\top. \quad (8)$$

Moreover, if we set $a(\boldsymbol{x}, k) \equiv \frac{1}{\gamma}$ for some $\gamma \geq 1$ (i.e., we sample uniformly at random a fraction $\frac{1}{\gamma}$ of the full dataset), denote the corresponding S matrix as $\mathbf{S}_{US:\frac{1}{\gamma}}$, then

$$\mathbf{S}_{US:\frac{1}{\gamma}} = \frac{1}{\gamma}\mathbf{S}_{full}.$$

In the following, we denote the asymptotic variance of our subsampling based estimator in Eq. (4), the full-sample based estimator and the estimator obtained from $\frac{1}{\gamma}$ uniformly sampled data by

$$\mathcal{V}_{Sub} = \left[\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\boldsymbol{\nabla}\mathbf{S}\boldsymbol{\nabla}^\top\right]^{-1}, \quad \mathcal{V}_{full} = \left[\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\boldsymbol{\nabla}\mathbf{S}_{full}\boldsymbol{\nabla}^\top\right]^{-1}, \quad \mathcal{V}_{US:\frac{1}{\gamma}} = \gamma\mathcal{V}_{full},$$

respectively.

Our purpose is to find a better sampling strategy with lower variance than that of uniform sampling. That is, we want to choose an acceptance probability function $a(\boldsymbol{x}, k)$ such that there exists some scalar $\gamma \geq 1$ making

$$\mathcal{V}_{Sub} \preceq \gamma\mathcal{V}_{full} = \mathcal{V}_{US:\frac{1}{\gamma}},$$

under the constraint that

$$\mathbb{E}_{C|\boldsymbol{X}}\left[a(\boldsymbol{X}, C)\right] \leq 1/\gamma$$

for all $\boldsymbol{X}$. The constraint means that the expected sample size $\sum_{i=1}^n Z_i$ is no more than n/γ: that is, we sample no more than $1/\gamma$ fraction of the full data.

Theorem 4.2 (Sampling Strategy)**.** *For any data point $\boldsymbol{x}$, let*

$$q(\boldsymbol{x}) = \max\left(0.5, p(\boldsymbol{x},1), \cdots, p(\boldsymbol{x},K)\right).$$

Given any $\gamma \geq 1$, consider the following choice of acceptance probability function:

(1) for $\gamma \geq 2q(\boldsymbol{x})$, set $a(\boldsymbol{x}, k)$ as

$$a(\boldsymbol{x}, k) = \begin{cases} \frac{2(1-q(\boldsymbol{x}))}{\gamma}, & \text{if } p(\boldsymbol{x}, k) = q(\boldsymbol{x}) \geq 0.5 \\ \frac{2q(\boldsymbol{x})}{\gamma}, & \text{otherwise} \end{cases}, \ k = 1 \cdots, K; \tag{9}$$

(2) for $1 \leq \gamma < 2q(\boldsymbol{x})$, set $a(\boldsymbol{x}, k)$ as

$$a(\boldsymbol{x}, k) = \begin{cases} \frac{1-q(\boldsymbol{x})}{\gamma - q(\boldsymbol{x})}, & \text{if } p(\boldsymbol{x}, k) = q(\boldsymbol{x}) \geq 0.5 \\ 1, & \text{otherwise} \end{cases}, \ k = 1 \cdots, K, \tag{10}$$

then, we always have

$$\mathcal{V}_{Sub} \preceq \gamma \mathcal{V}_{full} = \mathcal{V}_{US:\frac{1}{\gamma}}, \tag{11}$$

and the expected number of subsampled examples is

$$n_{Sub} = n\mathbb{E}_{(\boldsymbol{X},C)\sim\mathcal{D}}\ a(\boldsymbol{X}, C) \leq \frac{n}{\gamma}. \tag{12}$$

It is easy to check that the assigned acceptance probability in Theorem 4.2 is always valid (that is, it is a value in $[0, 1]$). With the sampling strategy in Theorem 4.2, we always use less than a fraction $\frac{1}{\gamma}$ of the full data to achieve less than γ times the variance of the full-sampled MLE. It implies that the method is never worse than the uniform sampling method. Moreover, the required sample size n_{Sub} can be significantly smaller than n/γ under favorable conditions. For example, when $\gamma \geq 2$, then it is easy to verify that $n_{Sub}/(n/\gamma) = \mathbb{E}_{\boldsymbol{x}} 4q(\boldsymbol{x})(1 - q(\boldsymbol{x}))$, which is close to zero when $q(\boldsymbol{x}) \approx 1$ for most $\boldsymbol{x}$, which happens when the classification accuracy is high.

More precisely, we have the following explicit formula for the expected conditional sampling probability:

$$\mathbb{E}_{C|\boldsymbol{X}\sim\mathcal{D}}\ a(\boldsymbol{X}, C) = \begin{cases} \frac{4}{\gamma}\ q(\boldsymbol{X})(1 - q(\boldsymbol{X})), & \text{under case (1) in Theorem 4.2} \\ \frac{\gamma(1-q(\boldsymbol{X}))}{\gamma - q(\boldsymbol{X})}, & \text{under case (2) in Theorem 4.2} \end{cases}$$

Therefore in the favorable case where most $q(\boldsymbol{x}) \approx 1$ for $\boldsymbol{x} \sim \mathcal{D}$ (i.e., the data are conditionally imbalanced), Theorem 4.2 implies that the method will subsample very few examples to achieve the desired variance comparable to that of $1/\gamma$ random sampling.

The choice of Theorem 4.2 reduces to the sampling strategy of local case control in [8] when $\gamma \geq 2$ and $K = 2$. Although a method was proposed in [8] for $\gamma \in [1, 2)$, it is different from our sampling strategy, and there is no theoretical guarantee for that strategy. In fact, the choice in [8] for $\gamma < 2$ may lead to a variance larger than that of the random sampling. The empirical performance can also be inferior to our method.

In the multi-class case, our method is not a natural extension of local case control (which would imply a method to set all class probabilities to $1/K$ after sampling). Instead, we will only assign a smaller sampling probability for $(\boldsymbol{x}, c)$ when $p(\boldsymbol{x}, c) \geq 0.5$. The method is less likely to select a sample when c coincides with the the prediction of the underlying true model, while it will likely be selected if c contradicts the underlying true model. Since the sampling strategy prefers data points with uncertain labels, we call it *local uncertainty sampling* (LUS).

5 Local Uncertainty Sampling

In order to apply Theorem 4.2 empirically, the main idea is to employ a pilot estimator $\tilde{\boldsymbol{\Theta}}$, which is a rough but consistent estimate of the true parameter, and then assign the sampling probability according to this estimator. In many real-world applications, such a pilot is easy to obtain. For example, when data arrive in time sequence, the estimator trained on previous observations can be used as a pilot for fitting a new model when new observations are coming in. Moreover, a rough estimator obtained on a small subset of the full population can be used as a pilot for training on the entire dataset. In our experiments, a small uniformly subsampled subset of the original population is used to obtain the pilot. As we will see later, it is sufficient for our method to obtain good practical performance.

Given a pilot estimator $\tilde{\boldsymbol{\Theta}}$ and $\gamma \geq 1$, the LUS algorithm can be described in Algorithm 1.

Algorithm 1 The LUS Algorithm for Multi-Class Logistic Regression.

1: Given a pilot estimator $\tilde{\boldsymbol{\Theta}}$ and $\gamma \geq 1$.

2: Scan the data once and generate the random variables $Z_i \sim \text{Bernoulli}(a(\boldsymbol{X}_i, C_i) : Z_i = 1)$ based on the acceptance probability $a(\boldsymbol{X}_i, C_i)$ defined as

$$a(\boldsymbol{X}_i, C_i) = \begin{cases} \frac{1-\tilde{q}}{\gamma - \max(\tilde{q}, 0.5\gamma)}, & \text{if } \tilde{p}_{C_i} = \tilde{q} \geq 0.5 \\ \min(1, 2\tilde{q}/\gamma), & \text{otherwise} \end{cases},$$

where $\tilde{q} = \max(0.5, \tilde{p}_1, \cdots, \tilde{p}_K)$ and

$$\tilde{\boldsymbol{p}} = (\tilde{p}_1, \cdots, \tilde{p}_K)^\top = \left(\frac{e^{f(\boldsymbol{X}_i, \tilde{\theta}_1)}}{1 + \sum_{k=1}^{K-1} e^{f(\boldsymbol{X}_i, \tilde{\theta}_k)}}, \cdots, \frac{e^{f(\boldsymbol{X}_i, \tilde{\theta}_{K-1})}}{1 + \sum_{k=1}^{K-1} e^{f(\boldsymbol{X}_i, \tilde{\theta}_k)}}, \frac{1}{1 + \sum_{k=1}^{K-1} e^{f(\boldsymbol{X}_i, \tilde{\theta}_k)}} \right)^\top.$$

3: Fit a multi-class logistic regression model to the subsample set $\{(\boldsymbol{X}_i, Y_i) : Z_i = 1\}$ with $g(\cdot)$ defined in Eq. (5):

$$\hat{\boldsymbol{\Theta}}_{LUS} = \arg\max_{\boldsymbol{\Theta}} \sum_{i=1}^{n} Z_i \left(\sum_{k=1}^{K-1} Y_{i,k} \cdot g(\boldsymbol{X}_i, \boldsymbol{\theta}_k) - \log \left(1 + \sum_{k=1}^{K-1} e^{g(\boldsymbol{X}_i, \boldsymbol{\theta}_k)} \right) \right). \tag{13}$$

4: Output $\hat{\boldsymbol{\Theta}}_{LUS}$.

6 Experiments

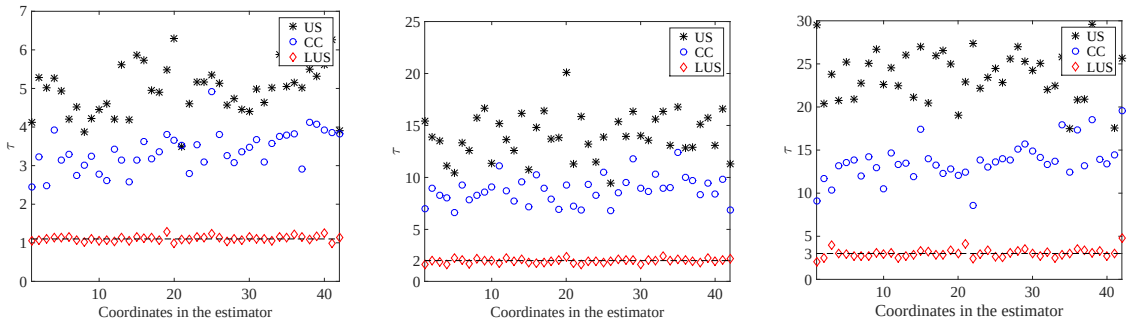
In this section, we evaluate the performance of LUS and compare it with the uniform sampling (US) and case-control (CC) sampling methods on both simulated and real-world datasets. For the CC sampling method, we extend the standard CC considered in the binary classification problem to multi-class case by sampling equal number of data for each class. Under marginal imbalance, if some minority classes do not have enough samples, we keep all data for those classes and subsample equal number of the remaining data points from other classes. In addition, we also compare the LUS and LCC sampling methods on the Web Spam dataset, which is a binary classification problem.

6.1 Simulation: Marginal Imbalance

We first simulate the case where the data is marginally imbalanced. We generate a 3-class Gaussian model according to $(\boldsymbol{X}|C=k) \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, which is the true data distribution $\mathcal{D}$. We set the number of features as $d = 20$, and $\boldsymbol{\mu}_1 = [\underbrace{1,1,\cdots,1}_{10}, \underbrace{0,0,\cdots,0}_{10}]$, $\boldsymbol{\mu}_2 = [\underbrace{0,0,\cdots,0}_{10}, \underbrace{1,1,\cdots,1}_{10}]$ and $\boldsymbol{\mu}_3 = [\underbrace{0,0,\cdots,0}_{20}]$. The covariance matrices for classes $k = 1,2,3$ are assigned to be the same $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = \mathbf{I}_d$, where $\mathbf{I}_d$ is a $d \times d$ identity matrix. So the true log-odds function f is linear and we use a linear model to fit the data. Moreover, we set $\mathbb{P}(C=1) = 0.1$, $\mathbb{P}(C=2) = 0.8$, $\mathbb{P}(C=3) = 0.1$, i.e., the data is marginally imbalanced and the second class dominates the population.

Since the true data distribution $\mathcal{D}$ is known in this case, we directly generate the full dataset from the distribution $\mathcal{D}$. For the full dataset, we generate $n = 50,000$ data points. The entire procedure is repeated for 200 times to obtain the variance of different estimators. For the LUS method, we randomly generate 5000 data points (10% of the full data) from $\mathcal{D}$ to obtain a pilot estimator and fix it before the 200 repetitions. Moreover, we generate another $n_{test} = 100,000$ data points to test the prediction accuracy.

Recall that γ controls the desired variance of the LUS estimator according to Theorem 4.2. In the following experiments, we will test different values of $\gamma = \{1.1, 1.2, \cdots, 1.9, 2, 3, \cdots, 9, 10\}$, respectively. Given the value of γ, suppose the LUS method will subsample a number of n_{Sub} data points. Then, we let the US and CC sampling methods to select the same amount n_{Sub} of examples for fair comparison.

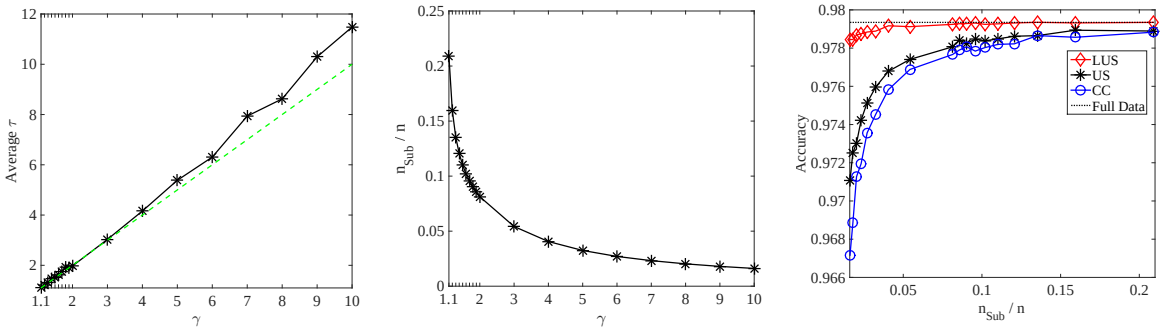


(a) $\gamma = 1.1$ (b) $\gamma = 2$ (c) $\gamma = 3$

Figure 1: The τ value for each coordinate under different values of γ. τ denotes the ratio between the variance of each coordinate in the subsampling based estimator and the variance of the coordinate in the full-sample MLE, i.e., $\tau = \mathrm{Var}(\hat{\theta}_{sub})/\mathrm{Var}(\hat{\theta}_{full})$.

Since $\boldsymbol{\theta}_k \in \mathbb{R}^d$ $(k = 1, \cdots, K-1)$ and there is an additional intercept parameter, the estimator contains a total number of $(d+1)(K-1)$ coordinates. Now, denote by τ the empirical coordinate-wise ratio between the variance of the coordinate in the candidate estimator and the variance of the coordinate in the full-sample MLE, we show the τ value for each coordinate under different values of γ. The results under $\gamma = 1.1, 2$ and 3 are shown in Fig. 1. In this case, there are 42 coordinates. From the figures, we observe that the τ value for each coordinate of the LUS method is approximately γ, which matches our theoretical analysis in Theorem 4.2. On the other side, the variances of the US and CC sampling methods are much higher than that of the LUS method.

In Fig. 2(a), we plot the relationship between the average τ for all coordinates against γ.



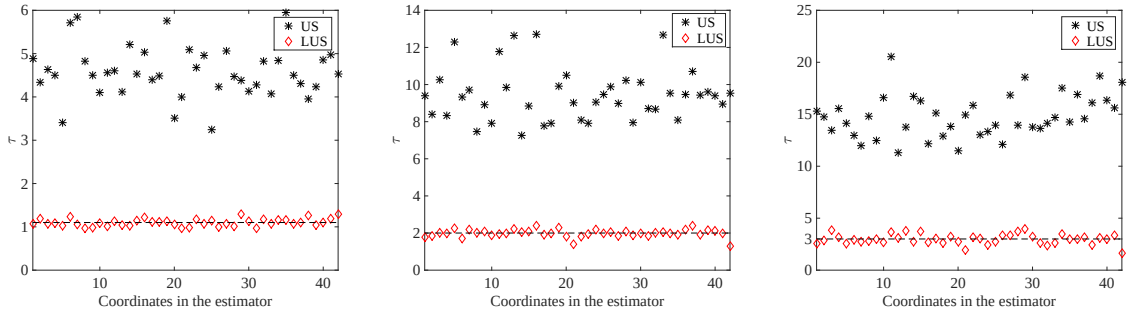
(a) Average τ v.s. γ (b) n_{Sub}/n v.s. γ (c) Accuracy v.s. n_{Sub}/n

Figure 2: Plots in the first simulation.

Observe that the relationship is close to $y = x$ (the dashed green line), which shows that τ approximately equals γ. These experimental results support our theoretical analysis. Fig. 2(b) reports the relationship between n_{Sub}/n and γ. Fig. 2(c) shows the relationship between the prediction accuracy on the testing data and the subsampling proportion n_{Sub}/n. From the figure, when n_{Sub}/n decreases, the prediction accuracy of all the methods decreases, while the LUS method shows much slower degradation compared to the US and CC methods. Moreover, according to the results of LUS in Fig. 2(c), we only need about 10% of the full data to achieve the same prediction accuracy as the full MLE, implying that the LUS method is very effective for reducing the computational cost while preserving high accuracy.

6.2 Simulation: Marginal Balance

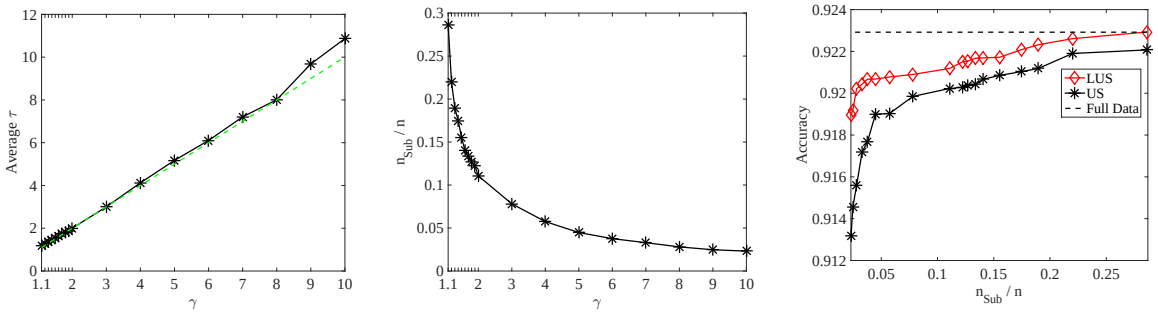
In the second simulation, we generate marginally balanced data with conditional imbalance. Under this situation, the CC sampling method is identical to US, and hence we omit it in our comparison. The settings are exactly the same as those in the previous simulation, except that we let $\mathbb{P}(C = 1) = \mathbb{P}(C = 2) = \mathbb{P}(C = 3) = \frac{1}{3}$ here, i.e., the data is marginally balanced. However, this simulated data is conditional imbalanced as we will see later.



(a) $\gamma = 1.1$ (b) $\gamma = 2$ (c) $\gamma = 3$

Figure 3: The τ value for each coordinate under different values of γ. $\tau = \mathrm{Var}(\hat{\theta}_{sub})/\mathrm{Var}(\hat{\theta}_{full})$.

The τ value for each coordinate when $\gamma = 1.1, 2$ and 3 is shown in Fig. 3. The relationship between the average τ for all the coordinates and γ is plotted in Fig. 4(a). Fig. 4(b) reports the



(a) Average τ v.s. γ (b) n_{Sub}/n v.s. γ (c) Accuracy v.s. n_{Sub}/n

Figure 4: Plots in the second simulation.

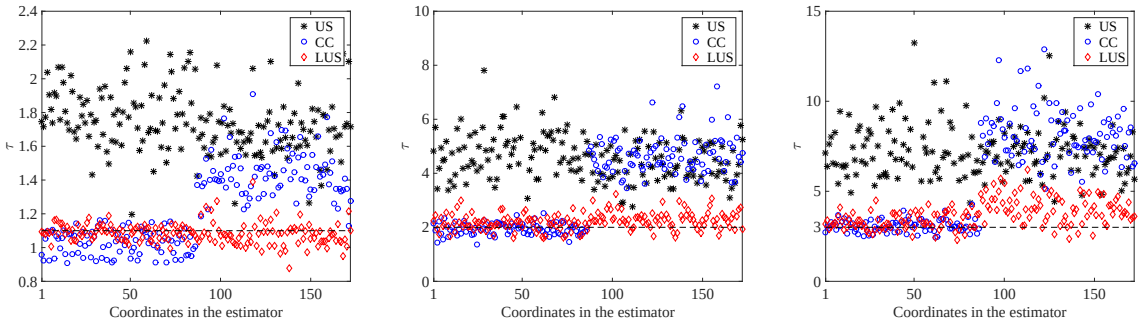
relationship between n_{Sub}/n and γ. In Fig. 4(c), we show the relationship between the prediction accuracy on the test data and the subsampling proportion n_{Sub}/n. The conclusions are similar to those of the previous simulation, which demonstrate the effectiveness of the LUS method under the marginally balanced (but conditionally imbalanced) case. Fig. 4(c) suggests that we only need about 20% of the full data to achieve the same prediction accuracy as that of the MLE estimator using full-data.

6.3 Connect-4 Data: Marginal Imbalance

We compare different methods on the Connect-4 dataset[1], which is a two-player connection game in which the players drop discs to a vertical grid, and the player who firstly connects four of the discs wins the game. The input contains legal moves in the game and the label is the game status for the first player, i.e., win, loss or draw. Hence, there are $K = 3$ classes. The number of data points in each class is 16635, 6449, and 44473, respectively. Hence, this data is marginally imbalanced and there are 67557 total data points in the entire dataset. Each data point is represented by 126 binary features, i.e., either 0 or 1. We preprocess the data by removing the extreme sparse features for which more than 98% of the samples have zero values, since these features barely contributes to the model. The resulting $d = 85$ features are used for estimation. Hence, there will be $(d + 1)(K - 1) = 172$ coordinates. Following [8], we first randomly generate 100 subsets, each of which contains 20,000 uniform samples, to obtain 100 independent 'full' datasets. Then, we can calculate the variance of the full-sample MLE. The experiments are repeated 100 times to obtain the variance of each estimator. A pilot is obtained before the repetitions based on 4000 uniformly selected data points (20% of the full data) from the original population.

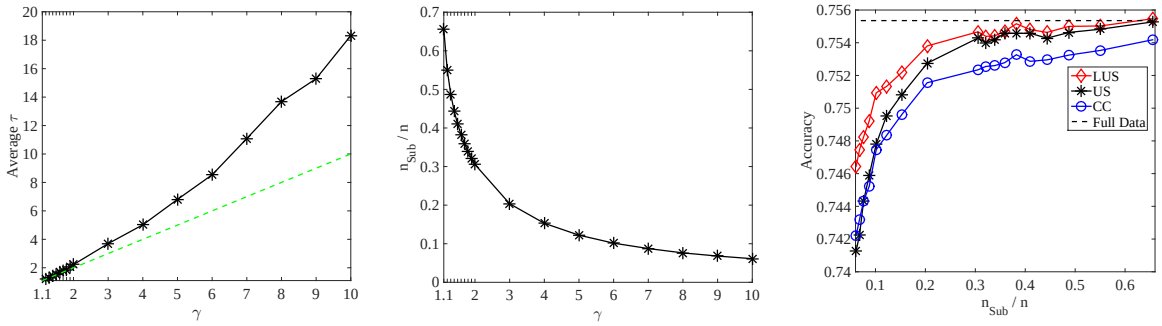
We test different values of $\gamma = \{1.1, 1.2, \cdots, 1.9, 2, 3, \cdots, 9, 10\}$ to examine the variance of different estimators. The plots of the τ value for each coordinate under $\gamma = 1.1, 2$ and 3 are reported in Fig. 5. The average value of τ for all the coordinates against γ is plotted in Fig. 6(a). From Fig. 5 and 6(a), the τ value approximately equals γ, which matches our theoretical results again. Fig. 6(b) shows how the subsample proportion n_{Sub}/n changes with γ varying. In Fig. 6(c), we show the relationship between the prediction accuracy on the full data and the subsample proportion. Similar to the simulation examples, we observe that when the subsampling proportion decreases, the LUS method shows much slower degradation compared to the US and CC methods. The experimental results suggest that we can use about 60% of the full data to substitute the full dataset with LUS.

[1] `http://archive.ics.uci.edu/ml/datasets/Connect-4`



(a) $\gamma = 1.1$ (b) $\gamma = 2$ (c) $\gamma = 3$

Figure 5: The τ value for each coordinate under different values of γ. $\tau = \mathrm{Var}(\hat{\theta}_{sub})/\mathrm{Var}(\hat{\theta}_{full})$.



(a) Average τ v.s. γ (b) n_{Sub}/n v.s. γ (c) Accuracy v.s. n_{Sub}/n

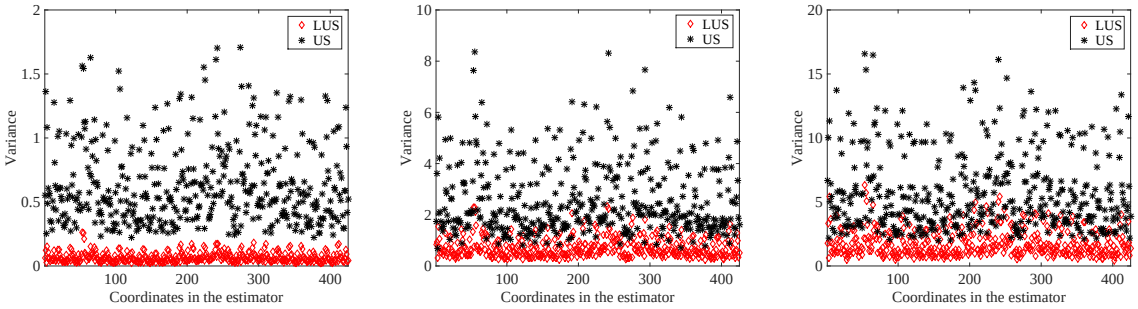
Figure 6: Plots in connect-4 data.

6.4 Letter Dataset: Marginal Balance

In this section, we test different methods on the letter data[2]. The goal is to recognize a number of 16-pixel character images as one of the 26 capital letters in the English alphabet. Therefore, the dataset contains $K = 26$ classes, $d = 16$ features, and $(d+1)(K-1) = 402$ coordinates in the estimator. There are totally $20,000$ data points with equal number of samples in each class, i.e., the data is marginally balanced. Therefore, the CC method is the same as the US method and thus we omit it for simplicity. Different from the connect-4 data, we do not generate multiple independent 'full' datasets by downsampling from the entire population, since the size of this dataset is relatively small compared to the number of features and number of classes. Hence we will not compute the variance of the full-sample MLE and we directly compare the variances of the LUS and US methods. 4000 data points (20% of the full data) are uniformly chosen to obtain the pilot before 100 random repetitions. Other settings are identical to those in the connect-4 experiment.

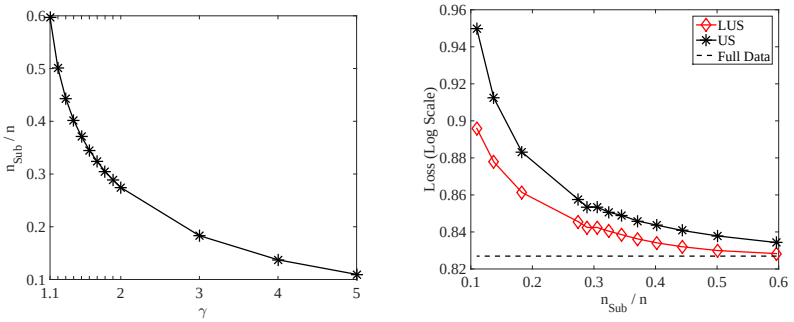
The variances of the coordinates under $\gamma = 1.1, 2$ and 3 are reported in Fig. 7. Again, the LUS method is superior. Fig. 8(a) plots the subsample proportion against γ. In this dataset, the change of the subsample size n_{Sub} does not have a sufficiently large impact on the prediction accuracy for both US and LUS methods. This is because the change affects the predicted probability $\hat{p}(\boldsymbol{x}, k)$ for each class k given $\boldsymbol{x}$, but does not change the final decision of the predicted label $\arg\max_k(\hat{p}(\boldsymbol{x}, k))$

[2] `http://archive.ics.uci.edu/ml/datasets/Letter+Recognition`



(a) $\gamma = 1.1$ (b) $\gamma = 2$ (c) $\gamma = 3$

Figure 7: Variance for each coordinate under different values of γ.



(a) n_{Sub}/n v.s. γ (b) Loss (log scale) v.s. n_{Sub}/n

Figure 8: Plots in letter data.

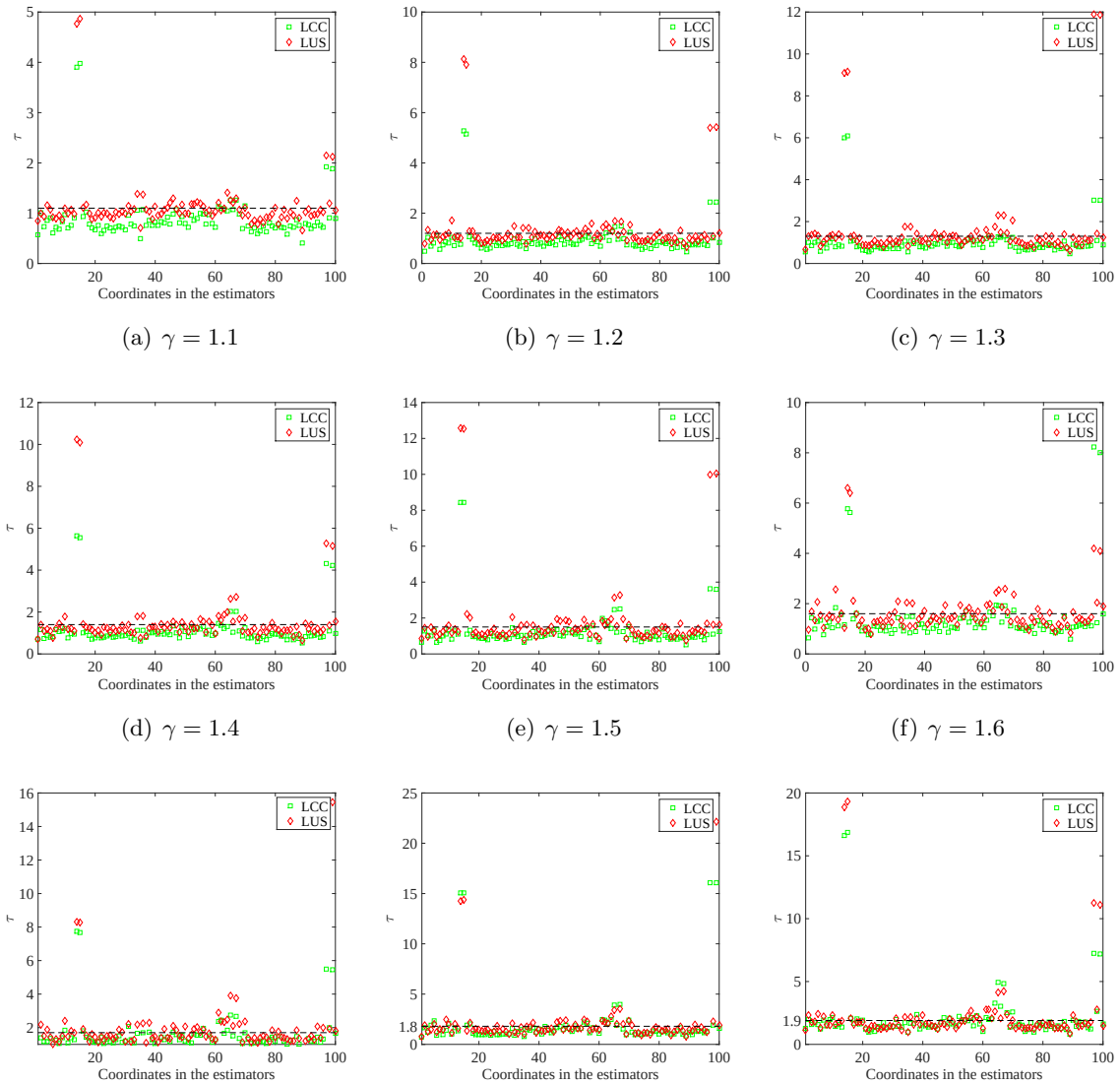
as much. In order to provide a more clear visualization of the prediction performance, we use the log-loss objective value on the full dataset. Fig. 8(b) gives the relationship between the loss function and the subsampling proportion. It shows that the LUS method consistently outperforms the US method, and it suggests that we can use about half of the data in LUS to achieve about the same performance as the full data.

6.5 Web Spam Data: Binary Classification

In this section, we compare the LUS method with the LCC method on the Web Spam data[3], which is a binary classification problem used in [8] to evaluate the LCC method. Since the comparison among LCC, US and CC on this data has been reported in [8], we do not repeat them here and focus on the comparison between LUS and LCC. This data contains 350,000 web pages and about 60% of them are web spams. This data set is approximately marginally balanced, but it has been shown to have strong conditional imbalance in [8]. Here, we adopt the same settings as described in [8] to compare the LUS and the LCC methods. That is, we select 99 features which appear in at least 200 documents, and the features are log-transformed. 10% of the observations are uniformly selected to obtain a pilot estimator. Since we only have a single data set, we follow [8] to subsample 100 datasets, each of which contains 100,000 data points, as 100 independent 'full' datasets, and then repeat the experiments 100 times.

[3] `http://www.cc.gatech.edu/projects/doi/WebbSpamCorpus.html`

Observe that when $\gamma \geq 2$ (or $\alpha = \frac{1}{\gamma - 1} \leq 1$ for LCC), LUS and LCC methods are equivalent to each other. Therefore we only focus on the case of $\gamma < 2$. Similar to previous experiments, we test different values of $\gamma = \{1.1, 1.2, \cdots, 1.9\}$ and accordingly set $\alpha = \{10, 5, \cdots, \frac{10}{9}\}$, so that the two methods have the same asymptotic variance. Then, we will compare the number of subsampled data points to see which method is more effective in terms of subsampled data size n_{Sub}.



(a) $\gamma = 1.1$ (b) $\gamma = 1.2$ (c) $\gamma = 1.3$

(d) $\gamma = 1.4$ (e) $\gamma = 1.5$ (f) $\gamma = 1.6$

(g) $\gamma = 1.7$ (h) $\gamma = 1.8$ (i) $\gamma = 1.9$

Figure 9: The τ value for each coordinate under different values of γ. $\tau = \text{Var}(\hat{\theta}_{sub})/\text{Var}(\hat{\theta}_{full})$.

Fig. 9 plots the τ values for different choices of γ. Just as expected from the theoretical results, both LUS and LCC methods have the same variance that is γ (or $(1 + \frac{1}{\alpha})$) times variance of the full-sample MLE. Next, we compare the subsampling proportion of the methods when γ changes in Fig. 10. It shows that the LUS method consistently subsamples a much smaller number of data points to achieve the same variance in Fig. 9 parameterized by γ. This demonstrates that the new formulation in LUS is not only theoretically better justified, but also more effective than LCC in

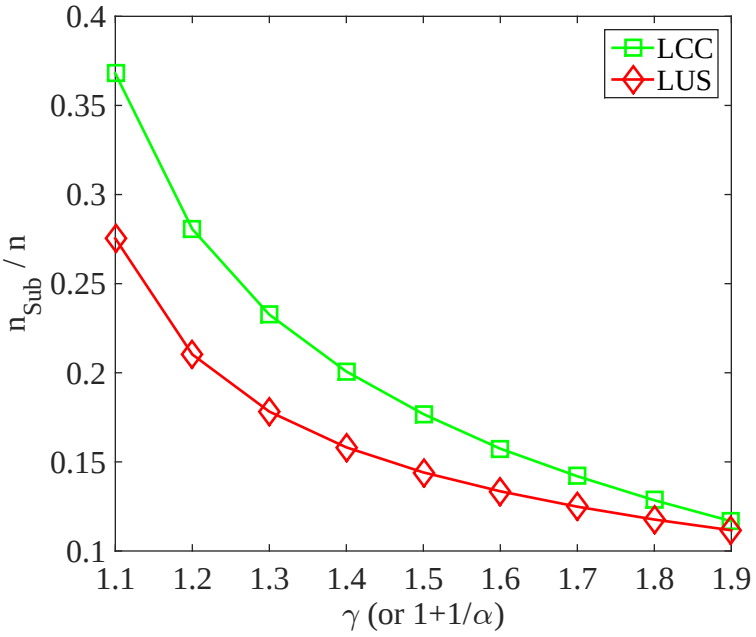


Figure 10: Plot of n_{Sub}/n v.s. γ (or $1 + \frac{1}{\alpha}$) in webspam data.

practice (for the case of $\gamma \leq 2$). Similar to the letter dataset, the prediction performance under different settings does not vary much, and a very small number of subsample is sufficient to obtain a good prediction performance. Therefore we do not include the comparison here.

7 Conclusion

This paper introduced a general subsampling method for solving large-scale logistic regression problems. We investigated the asymptotic variance of the proposed estimator. Based on the theoretical analysis, we proposed an effective sampling strategy called Local Uncertainty Sampling to achieve any given level of desired variance. We proved that the method always achieves lower variance than random subsampling for a given expected sample size, and the improvement may be significant under the favorable condition of strong conditional imbalance. Therefore the method can effectively accelerate the computation of large-scale logistic regression in practice. Experiments on synthetic and real-world datasets are provided to illustrate the effectiveness of the proposed method. The empirical studies confirm the theory, and demonstrate that the local uncertainty sampling method outperforms the uniform sampling, case-control sampling and the local case-control sampling methods under various settings. By using the proposed method, we are able to select a very small subset of the original data to achieve the same performance as that of the full dataset, which provides an effective mean for big data computation under limited resources.

Acknowledgment

This research is partially supported by NSF IIS-1250985, NSF IIS-1407939, and NIH R01AI116744.

References

[1] Naoki Abe, Bianca Zadrozny, and John Langford. An iterative method for multi-class cost-sensitive learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 3–11. ACM, 2004.

[2] James A Anderson. Separate sample logistic discrimination. *Biometrika*, 59(1):19–35, 1972.

[3] Norman Breslow. Design and analysis of case-control studies. *Annual review of public health*, 3(1):29–54, 1982.

[4] Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Editorial: special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter*, 6(1):1–6, 2004.

[5] Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *Advances in Neural Information Processing Systems*, pages 442–450, 2010.

[6] Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. Sample selection bias correction theory. In *Algorithmic Learning Theory*, pages 38–53. Springer, 2008.

[7] Paramveer Dhillon, Yichao Lu, Dean P Foster, and Lyle Ungar. New subsampling algorithms for fast least squares regression. In *Advances in Neural Information Processing Systems*, pages 360–368, 2013.

[8] William Fithian and Trevor Hastie. Local case-control sampling: Efficient subsampling in imbalanced data sets. *Ann. Statist.*, 42:1693–1724, 2014.

[9] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *Knowledge and Data Engineering, IEEE Transactions on*, 21(9):1263–1284, 2009.

[10] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

[11] Hyun-Chul Kim, Shaoning Pang, Hong-Mo Je, Daijin Kim, and Sung Yang Bang. Pattern classification using support vector machine ensemble. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 2, pages 160–163. IEEE, 2002.

[12] Gary King and Langche Zeng. Logistic regression in rare events data. *Political Analysis*, 9(2):137–163, 2001.

[13] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[14] Nathan Mantel and William Haenszel. Statistical aspects of the analysis of data from retrospective studies. *J natl cancer inst*, 22(4):719–748, 1959.

[15] Paul Mineiro and Nikos Karampatziakis. Loss-proportional subsampling for subsequent erm. *arXiv preprint arXiv:1306.1840*, 2013.

[16] AJ Scott and CJ Wild. Fitting logistic regression models in stratified case-control studies. *Biometrics*, pages 497–510, 1991.

[17] Alastair Scott and Chris Wild. On the robustness of weighted methods for fitting models to case–control data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2):207–219, 2002.

[18] Alastair J Scott and CJ Wild. Fitting logistic models under case-control or choice based sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 170–182, 1986.

[19] Aik Choon Tan, David Gilbert, and Yves Deville. Multi-class protein fold classification using a new ensemble machine learning approach. *Genome Informatics*, 14:206–217, 2003.

[20] Achmad Widodo and Bo-Suk Yang. Support vector machine in machine condition monitoring and fault diagnosis. *Mechanical Systems and Signal Processing*, 21(6):2560–2574, 2007.

[21] Yu Xie and Charles F Manski. The logit model and response-based samples. *Sociological Methods & Research*, 17(3):283–302, 1989.

[22] Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the Twenty-first International Conference on Machine Learning*, page 114. ACM, 2004.

[23] Tong Zhang and F Oles. The value of unlabeled data for classification problems. In *Proceedings of the Seventeenth International Conference on Machine Learning,(Langley, P., ed.)*, pages 1191–1198. Citeseer, 2000.

A Proof of Theorem 4.1

The following lemma is useful in our analysis.

Lemma A.1. *For any norm $\|\cdot\|$ defined on the parameter space for $\boldsymbol{\Theta}$, assume the quantities $\|\nabla_{\boldsymbol{\theta}_k} f(\boldsymbol{x}, \boldsymbol{\theta}_k)\|$, $\|\nabla^2_{\boldsymbol{\theta}_k} f(\boldsymbol{x}, \boldsymbol{\theta}_k)\|$ and $\|\nabla^3_{\boldsymbol{\theta}_k} f(\boldsymbol{x}, \boldsymbol{\theta}_k)\|$ for $k = 1, \cdots, K-1$ are bounded. Then, for any compact set $\mathbb{S} \in \mathbb{R}^{p\times(K-1)}$, we have*

$$\begin{aligned}
&\sup_{\boldsymbol{\Theta}\in\mathbb{S}} |\hat{R}_n(\boldsymbol{\Theta}) - R(\boldsymbol{\Theta})| \xrightarrow{p} 0,\\
&\sup_{\boldsymbol{\Theta}\in\mathbb{S}} \|\nabla \hat{R}_n(\boldsymbol{\Theta}) - \nabla R(\boldsymbol{\Theta})\| \xrightarrow{p} 0,\\
&\sup_{\boldsymbol{\Theta}\in\mathbb{S}} \|\nabla^2 \hat{R}_n(\boldsymbol{\Theta}) - \nabla^2 R(\boldsymbol{\Theta})\| \xrightarrow{p} 0,
\end{aligned}$$

Proof. For fixed $\boldsymbol{\Theta}$, we define

$$\psi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\Theta}) = \sum_{k=1}^{K-1} y_k \cdot g(\boldsymbol{x}, \boldsymbol{\theta}_k) - \log\left(1 + \sum_{k=1}^{K-1} e^{g(\boldsymbol{x}, \boldsymbol{\theta}_k)}\right),$$

then we have $\hat{R}_n(\boldsymbol{\Theta}) = \frac{1}{n}\sum_{i=1}^n Z_i \psi(\boldsymbol{X}_i, \boldsymbol{y}_i, \boldsymbol{\Theta})$ and $\mathbb{E}_{\mathcal{A}}\Big[Z_i \psi(\boldsymbol{X}_i, \boldsymbol{y}_i, \boldsymbol{\Theta})\Big] = R(\boldsymbol{\Theta})$. By the Law of Large Numbers, we know that $\hat{R}_n(\boldsymbol{\Theta})$ converges point-wise to $R(\boldsymbol{\Theta})$ in probability.

According to the assumption, there exists a constant $M > 0$ such that

$$\|\nabla_{\boldsymbol{\Theta}} \psi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\Theta})\| \le \sum_{k=1}^{K-1} \left\| \left(y_k - \frac{e^{g(\boldsymbol{x}, \boldsymbol{\theta}_k)}}{1 + \sum_{k'=1}^{K-1} e^{g(\boldsymbol{x}, \boldsymbol{\theta}_{k'})}} \right) \nabla_{\boldsymbol{\theta}_k} f(\boldsymbol{x}, \boldsymbol{\theta}_k) \right\| \le \sum_{k=1}^{K-1} \|\nabla_{\boldsymbol{\theta}_k} f(\boldsymbol{x}, \boldsymbol{\theta}_k)\| \le M.$$

Given any $\epsilon > 0$, we may find a finite cover $\mathbb{S}_\epsilon = \{\boldsymbol{\Theta}_1, \cdots, \boldsymbol{\Theta}_L\} \subset \mathbb{S}$ so that for any $\boldsymbol{\Theta} \in \mathbb{S}$, there exists $\boldsymbol{\Theta}_j \in \mathbb{S}_\epsilon$ such that $|\psi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\Theta}) - \psi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\Theta}_j)| < \epsilon$. Since $\mathbb{S}_\epsilon$ is finite, as $n \to \infty$, $\sup_{\boldsymbol{\Theta} \in \mathbb{S}_\epsilon} |\hat{R}_n(\boldsymbol{\Theta}) - R(\boldsymbol{\Theta})|$ converges to 0 in probability. Therefore as $n \to \infty$, with probability 1, we have

$$\sup_{\boldsymbol{\Theta} \in \mathbb{S}} |\hat{R}_n(\boldsymbol{\Theta}) - R(\boldsymbol{\Theta})| < 2\epsilon + \sup_{\boldsymbol{\Theta} \in \mathbb{S}_\epsilon} |\hat{R}_n(\boldsymbol{\Theta}) - R(\boldsymbol{\Theta})| \to 2\epsilon.$$

Let $\epsilon \to 0$, we obtain the first bound. The second and the third bounds can be similarly obtained. □

We are now ready to prove Theorem 4.1.

Proof. For notational simplicity, we abbreviate the point-wise functions $f(\boldsymbol{x}, \boldsymbol{\theta}_k)$, $g(\boldsymbol{x}, \boldsymbol{\theta}_k)$, $p(\boldsymbol{x}, k) = \mathbb{P}_{\mathcal{D}}(C = k | \boldsymbol{X} = \boldsymbol{x})$, $a(\boldsymbol{x}, k)$ and $q(\boldsymbol{x})$ at $\boldsymbol{x}$ as f_k, g_k, p_k, a_k and q respectively. Moreover, denote $\nabla_i = \nabla_{\boldsymbol{\theta}_i} f(\boldsymbol{x}, \boldsymbol{\Theta}^*)$.

(1) Define $\boldsymbol{h} = \left(h_1, \cdots, h_{K-1}\right)^\top$ and the function

$$\tilde{R}(\boldsymbol{h}) = \mathbb{E}_{(\boldsymbol{X}, C, Z) \sim \mathcal{D}_{\mathcal{A}}} \; Z \left[\sum_{k=1}^{K-1} \mathbb{I}(C = k) h_k - \log \left(1 + \sum_{k=1}^{K-1} e^{h_k} \right) \right],$$

then $\boldsymbol{f}^* = \left(f_1^*, \cdots, f_{K-1}^*\right)^\top$ is the global optimizer of $\tilde{R}(\cdot)$, and we have $R(\boldsymbol{\Theta}) = \tilde{R}(\boldsymbol{f})$. Moreover,

$$\begin{aligned} R(\boldsymbol{\Theta}^*) - R(\boldsymbol{\Theta}) &= \mathbb{E}_{(\boldsymbol{X}, C) \sim \mathcal{D}} \; a(\boldsymbol{X}, C) \left[\sum_{k=1}^{K-1} Y_k (f_k^* - f_k) + \log \frac{1 + \sum_{k=1}^{K-1} e^{f_k}}{1 + \sum_{k=1}^{K-1} e^{f_k^*}} \right] \\ &= \mathbb{E}_{(\boldsymbol{X}, C) \sim \mathcal{D}} \; a(\boldsymbol{X}, C) \left[\sum_{k=1}^{K-1} \frac{e^{f_k^*}}{1 + \sum_{k=1}^{K-1} e^{f_k^*}} (f_k^* - f_k) + \log \frac{1 + \sum_{k=1}^{K-1} e^{f_k}}{1 + \sum_{k=1}^{K-1} e^{f_k^*}} \right]. \end{aligned}$$

Note that for the convex function $G(\boldsymbol{f}) = \log \left(1 + \sum_{k=1}^{K-1} e^{f_k}\right)$, its KL divergence for $\boldsymbol{f}$ and $\boldsymbol{f}^*$ is

$$\begin{aligned} \Delta(\boldsymbol{f}, \boldsymbol{f}^*) &= G(\boldsymbol{f}) - G(\boldsymbol{f}^*) - \nabla G(\boldsymbol{f}^*)^\top (\boldsymbol{f} - \boldsymbol{f}^*) \\ &= \log \frac{1 + \sum_{k=1}^{K-1} e^{f_k}}{1 + \sum_{k=1}^{K-1} e^{f_k^*}} + \sum_{k=1}^{K-1} \frac{e^{f_k^*}}{1 + \sum_{k=1}^{K-1} e^{f_k^*}} (f_k^* - f_k), \end{aligned}$$

and $\Delta(\boldsymbol{f}, \boldsymbol{f}^*) \geq 0$ with $\Delta(\boldsymbol{f}, \boldsymbol{f}^*) = 0$ only when $\boldsymbol{f} = \boldsymbol{f}^*$. Moreover, since $\boldsymbol{f} = (f_1, \cdots, f_K)^T$ is point-wise function on $\boldsymbol{x}$, then $\boldsymbol{f} = \boldsymbol{f}^*$ indicates that $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\Theta}) = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\Theta}^*)$ for any $\boldsymbol{x} \sim \mathcal{D}$, and thus we have $\boldsymbol{\Theta} = \boldsymbol{\Theta}^*$. This is due to the assumption of the Theorem, which says that the parameter space is compact and $\forall \boldsymbol{\Theta} \neq \boldsymbol{\Theta}^*$ we have $\mathbb{P}_{\mathcal{D}}(f(\boldsymbol{X}, \boldsymbol{\Theta}) \neq f(\boldsymbol{X}, \boldsymbol{\Theta}^*)) > 0$. Hence,

$$R(\boldsymbol{\Theta}^*) - R(\boldsymbol{\Theta}) = \mathbb{E}_{(\boldsymbol{X}, C) \sim \mathcal{D}} \; a(\boldsymbol{X}, C) \Delta(\boldsymbol{f}, \boldsymbol{f}^*),$$

and we have for any $\boldsymbol{\Theta} \neq \boldsymbol{\Theta}^*$, $R(\boldsymbol{\Theta}) < R(\boldsymbol{\Theta}^*)$. It follows that given any $\epsilon' > 0$, there exists $\epsilon > 0$ so that $R(\boldsymbol{\Theta}) \geq R(\boldsymbol{\Theta}^*) - 2\epsilon$ implies that $\|\boldsymbol{\Theta} - \boldsymbol{\Theta}'\| < \epsilon'$. Now according to Lemma A.1, given any $\delta > 0$, when $n \to \infty$, with probability larger than $1 - \delta$, we have

$$R(\hat{\boldsymbol{\Theta}}_{Sub}) \geq \hat{R}_n(\hat{\boldsymbol{\Theta}}_{Sub}) - \epsilon \geq \hat{R}_n(\boldsymbol{\Theta}^*) - \epsilon \geq R(\boldsymbol{\Theta}^*) - 2\epsilon.$$

This implies that $\|\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}}_{Sub}\| < \epsilon'$.

(2) Let $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1^\top, \cdots, \boldsymbol{\theta}_K^\top)^\top \in \mathbb{R}^{d(K-1)}$. By the Mean Value Theorem, we have

$$\sqrt{n}\left(\hat{\boldsymbol{\Theta}}_{Sub} - \boldsymbol{\Theta}^*\right) = -\nabla^2 \hat{R}_n(\bar{\boldsymbol{\Theta}})^{-1} \sqrt{n} \nabla \hat{R}_n(\boldsymbol{\Theta}^*), \tag{14}$$

where $\bar{\boldsymbol{\Theta}} = t\boldsymbol{\Theta}^* + (1-t)\hat{\boldsymbol{\Theta}}_{Sub}$ for some $t \in [0,1]$. Note that Lemma A.1 implies that $\nabla^2 \hat{R}_n(\bar{\boldsymbol{\Theta}})^{-1}$ converges to $\nabla^2 \hat{R}_n(\boldsymbol{\Theta}^*)^{-1}$ in probability; moreover, $\hat{\boldsymbol{\Theta}}_{Sub} \to \boldsymbol{\Theta}^*$ in probability and hence $\bar{\boldsymbol{\Theta}} \to \boldsymbol{\Theta}^*$ in probability. It follows that the limit distribution of $\sqrt{n}\left(\hat{\boldsymbol{\Theta}}_{Sub} - \boldsymbol{\Theta}^*\right)$ is given by

$$-\nabla^2 R(\boldsymbol{\Theta}^*)^{-1} \sqrt{n} \nabla \hat{R}_n(\boldsymbol{\Theta}^*).$$

Observe that $\sqrt{n}\nabla \hat{R}_n(\boldsymbol{\Theta}^*)$ is the sum of n i.i.d random vectors with mean $\mathbb{E}\sqrt{n}\nabla \hat{R}_n(\boldsymbol{\Theta}^*) = \sqrt{n}\mathbb{E}\nabla \hat{R}_n(\boldsymbol{\Theta}^*) = 0$ and the asymptotic variance of $\sqrt{n}\left(\hat{\boldsymbol{\Theta}}_{Sub} - \boldsymbol{\Theta}^*\right)$ is

$$\mathrm{Var}\left(\sqrt{n}(\hat{\boldsymbol{\Theta}}_{Sub} - \boldsymbol{\Theta}^*)\right) = \mathbb{Q}(\boldsymbol{\Theta}^*)^{-1} \mathrm{Var}\left(\sqrt{n}\nabla \hat{R}_n(\boldsymbol{\Theta}^*)\right) \mathbb{Q}(\boldsymbol{\Theta}^*)^{-1}, \tag{15}$$

where $\mathbb{Q}(\boldsymbol{\Theta}^*) = -\nabla^2 R(\boldsymbol{\Theta}^*)$. Now, we can derive an explicit formula as

$$\mathbb{Q}(\boldsymbol{\Theta}^*) = \begin{bmatrix} \mathbb{Q}_{11} & \mathbb{Q}_{12} & \cdots & \mathbb{Q}_{1,K-1} \\ \mathbb{Q}_{21} & \mathbb{Q}_{22} & \cdots & \mathbb{Q}_{2,K-1} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbb{Q}_{K-1,1} & \mathbb{Q}_{K-1,2} & \cdots & \mathbb{Q}_{K-1,K-1} \end{bmatrix},$$

where

$$\begin{aligned}
\mathbb{Q}_{jj} =& -\nabla^2_{jj} R(\boldsymbol{\Theta}^*) \\
=& \mathbb{E}_{(\boldsymbol{X},C,Z)\sim\mathcal{A}}\, Z \left(\frac{e^{g_j^*}}{1+\sum_{k=1}^{K-1} e^{g_k^*}}\right)\left(1 - \frac{e^{g_j^*}}{1+\sum_{k=1}^{K-1} e^{g_k^*}}\right) \cdot \nabla_j \nabla_j^\top \\
=& \mathbb{E}_{(\boldsymbol{X},C)\sim\mathcal{D}}\, a(\boldsymbol{X},C) \left(\frac{e^{g_j^*}}{1+\sum_{k=1}^{K-1} e^{g_k^*}}\right)\left(1 - \frac{e^{g_j^*}}{1+\sum_{k=1}^{K-1} e^{g_k^*}}\right) \cdot \nabla_j \nabla_j^\top \\
=& \mathbb{E}_{(\boldsymbol{X},C)\sim\mathcal{D}}\, a(\boldsymbol{X},C) \cdot \frac{a_j p_j \sum_{k\neq j}^{K} a_k p_k}{\left(\sum_{k=1}^{K} a_k p_k\right)^2} \cdot \nabla_j \nabla_j^\top \qquad \left(e^{g_k^*} = \frac{a_k p_k}{a_K p_K} \text{ based on Eq. (5)}\right) \\
=& \mathbb{E}_{\boldsymbol{X}\sim\mathcal{D}}\, \frac{a_j p_j \sum_{k\neq j}^{K} a_k p_k}{\sum_{k=1}^{K} a_k p_k} \cdot \nabla_j \nabla_j^\top,
\end{aligned} \tag{16}$$

and

$$\begin{aligned}
\mathbb{Q}_{ij} =& -\nabla^2_{ij} R(\boldsymbol{\Theta}^*) \\
=& -\mathbb{E}_{(\boldsymbol{X},C)\sim\mathcal{D}}\, a(\boldsymbol{X},C) \left(\frac{e^{g_i^*}}{1+\sum_{k=1}^{K-1} e^{g_k^*}}\right)\left(\frac{e^{g_j^*}}{1+\sum_{k=1}^{K-1} e^{g_k^*}}\right) \cdot \nabla_j \nabla_j^\top \\
=& -\mathbb{E}_{\boldsymbol{X}\sim\mathcal{D}}\, \frac{a_i p_i a_j p_j}{\sum_{k=1}^{K} a_k p_k} \cdot \nabla_j \nabla_j^\top.
\end{aligned} \tag{17}$$

This implies that we can rewrite $\mathbb{Q}(\boldsymbol{\Theta}^*)$ as

$$\mathbb{Q}(\boldsymbol{\Theta}^*) = \mathbb{E}_{\boldsymbol{X}\sim\mathcal{D}} \boldsymbol{\nabla}\mathbf{S}\boldsymbol{\nabla}^\top. \tag{18}$$

Next, we derive an explicit formula for $\text{Var}\left(\sqrt{n}\nabla\hat{R}_n(\boldsymbol{\Theta}^*)\right)$ as follows.

$$\text{Var}\left(\sqrt{n}\nabla\hat{R}_n(\boldsymbol{\Theta}^*)\right) = \begin{bmatrix} \mathbb{V}_{11} & \mathbb{V}_{12} & \cdots & \mathbb{V}_{1,K-1} \\ \mathbb{V}_{21} & \mathbb{V}_{22} & \cdots & \mathbb{V}_{2,K-1} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbb{V}_{K-1,1} & \mathbb{V}_{K-1,2} & \cdots & \mathbb{V}_{K-1,K-1} \end{bmatrix}, \tag{19}$$

where

$$\begin{aligned}
&\mathbb{V}_{jj}\left(\sqrt{n}\nabla_{\boldsymbol{\theta}_j}\hat{R}_n(\boldsymbol{\Theta}^*), \sqrt{n}\nabla_{\boldsymbol{\theta}_j}\hat{R}_n(\boldsymbol{\Theta}^*)\right) \\
&= \mathbb{E}_{(\boldsymbol{X},C,Z)\sim\mathcal{A}}\ Z\left(\mathbb{I}(C=j) - \frac{e^{g_j^*}}{1+\sum_{k=1}^{K-1} e^{g_k^*}}\right)^2 \cdot \nabla_j\nabla_j^\top \qquad (Z^2 = Z) \qquad (20) \\
&= \mathbb{E}_{(\boldsymbol{X},C)\sim\mathcal{D}}\ a(\boldsymbol{X},C)\left(\mathbb{I}(C=j) - \frac{e^{g_j^*}}{1+\sum_{k=1}^{K-1} e^{g_k^*}}\right)^2 \cdot \nabla_j\nabla_j^\top \\
&= \mathbb{E}_{\boldsymbol{X}\sim\mathcal{D}}\left[\sum_{k\neq j}^{K} a_k p_k \cdot \left(\frac{a_j p_j}{\sum_{k=1}^K a_k p_k}\right)^2 + a_j p_j\left(1 - \frac{a_j p_j}{\sum_{k=1}^K a_k p_k}\right)^2\right] \cdot \nabla_j\nabla_j^\top \\
&= \mathbb{E}_{\boldsymbol{X}\sim\mathcal{D}} \frac{a_j p_j \sum_{k\neq j}^K a_k p_k}{\sum_{k=1}^K a_k p_k} \cdot \nabla_j\nabla_j^\top, \qquad (21)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{V}_{ij}\left(\sqrt{n}\nabla_{\boldsymbol{\theta}_i}\hat{R}_n(\boldsymbol{\Theta}^*), \sqrt{n}\nabla_{\boldsymbol{\theta}_j}\hat{R}_n(\boldsymbol{\Theta}^*)\right) \\
&= \mathbb{E}_{(\boldsymbol{X},C)\sim\mathcal{D}}\ a(\boldsymbol{X},C)\left(\mathbb{I}(C=i) - \frac{e^{g_i^*}}{1+\sum_{k=1}^{K-1} e^{g_k^*}}\right)\left(\mathbb{I}(C=j) - \frac{e^{g_j^*}}{1+\sum_{k=1}^{K-1} e^{g_k^*}}\right) \cdot \nabla_j\nabla_j^\top \\
&= \mathbb{E}_{\boldsymbol{X}\sim\mathcal{D}}\left[\sum_{k\neq i,j}^{K} a_k p_k \cdot \frac{a_i p_i a_j p_j}{\left(\sum_{k=1}^K a_k p_k\right)^2} - \frac{a_i p_i a_j p_j \sum_{k\neq i}^K a_k p_k}{\left(\sum_{k=1}^K a_k p_k\right)^2} - \frac{a_i p_i a_j p_j \sum_{k\neq j}^K a_k p_k}{\left(\sum_{k=1}^K a_k p_k\right)^2}\right] \cdot \nabla_j\nabla_j^\top \\
&= -\mathbb{E}_{\boldsymbol{X}\sim\mathcal{D}}\ \frac{a_i p_i a_j p_j}{\sum_{k=1}^K a_k p_k} \cdot \nabla_j\nabla_j^\top. \qquad (22)
\end{aligned}$$

This means that $\mathbb{Q}(\boldsymbol{\Theta}^*) = \text{Var}\left(\sqrt{n}\nabla\hat{R}_n(\boldsymbol{\Theta}^*)\right)$. Hence, we conclude

$$\sqrt{n}\left(\hat{\boldsymbol{\Theta}}_{Sub} - \boldsymbol{\Theta}^*\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \left[\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\boldsymbol{\nabla}\mathbf{S}\boldsymbol{\nabla}^\top\right]^{-1}\right).$$

This proves the desired result.

□

B Proof of Theorem 4.2

Proof. Observe that the quantity $\boldsymbol{\nabla}$ is independent of the sampling probability $a(\boldsymbol{x}, c)$, and hence the variance in Eq. (7) has dependence on the acceptance probability via the dependence on $\mathbf{S}$. Therefore to prove the desired variance bound, we only need to show that

$$\gamma \mathbf{S} \succeq \mathbf{S}_{full} = \gamma \mathbf{S}_{US:\frac{1}{\gamma}}.$$

In order to prove this inequality, we only need to verify that for any vector $\boldsymbol{\beta} \in \mathbb{R}^{K-1}$, $\boldsymbol{\beta}^\top(\gamma \mathbf{S} - \mathbf{S}_{full})\boldsymbol{\beta} \geq 0$. That is,

$$\gamma \sum_{i=1}^{K-1} a_i p_i \beta_i^2 - \frac{\gamma}{\sum_{i=1}^{K} a_i p_i} \left(\sum_{i=1}^{K-1} a_i p_i \beta_i \right)^2 - \sum_{i=1}^{K-1} p_i \beta_i^2 + \left(\sum_{i=1}^{K-1} p_i \beta_i \right)^2 \geq 0, \tag{23}$$

where $\sum_{i=1}^{K} p_i = 1$ and $0 \leq a_i \leq 1$.

(1) Consider the first case in Theorem 4.2, where $\gamma \geq 2q$. Given data point $\boldsymbol{x}$, we consider the following three cases.

(a) Assume that $q = p_K \geq 0.5$. Denote the left side of Eq. (23) by $F(a)$, and plug Eq. (9) into $F(a)$, we obtain

$$\sum_{i=1}^{K} a_i p_i = \frac{4}{\gamma} q(1-q),$$

and

$$\begin{aligned} F(a) &= (2q-1) \sum_{i=1}^{K-1} p_i \beta_i^2 - \left(\frac{q}{1-q} - 1 \right) \left(\sum_{i=1}^{K-1} p_i \beta_i \right)^2 \\ &\geq \frac{2q-1}{1-q} \left(\sum_{i=1}^{K-1} p_i \beta_i \right)^2 - \left(\frac{q}{1-q} - 1 \right) \left(\sum_{i=1}^{K-1} p_i \beta_i \right)^2 \\ &= 0, \end{aligned}$$

where the inequality is obtained by the Cauchy-Schwartz inequality that implies

$$\sum_{i=1}^{K-1} p_i \beta_i^2 \geq \frac{\left(\sum_{i=1}^{K-1} p_i \beta_i \right)^2}{\sum_{i=1}^{K-1} p_i}.$$

The equality holds if and only if $\beta_1 = \cdots = \beta_{K-1}$. Therefore, we conclude $F(a) \geq 0$.

(b) Assume that there exists some $j \neq K$ such that $q = p_j \geq 0.5$. By plugging Eq. (9) into $F(a)$, we have

$$\begin{aligned} F(a) &= 2q \left(\sum_{i \neq j}^{K-1} p_i \beta_i^2 + (1-q)\beta_j^2 \right) - \frac{q}{1-q} \left(\sum_{i \neq j}^{K-1} p_i \beta_i + (1-q)\beta_j \right)^2 - \sum_{i=1}^{K-1} p_i \beta_i^2 + \left(\sum_{i=1}^{K-1} p_i \beta_i \right)^2 \\ &= (2q-1) \sum_{i \neq j}^{K-1} p_i \beta_i^2 - \left(\frac{q}{1-q} - 1 \right) \left(\sum_{i \neq j}^{K-1} p_i \beta_i \right)^2 \\ &\geq 0. \end{aligned}$$

Again, the above inequality is obtained by the Cauchy-Schwartz inequality.

(c) Assume that $p_k < 0.5$ for all $k = 1, \cdots, K$ and hence $q = 0.5$. Under this case, we can immediately obtain $F(a) = 0$.

Combing (a), (b) and (c), we have shown that $F(a) \geq 0$ under the first case, where $\gamma \geq 2q$.

(2) Consider the second case in Theorem 4.2, where $1 \leq \gamma < 2q$. Given data point $\boldsymbol{x}$, we consider the following three cases.

(a) Assume that $q = p_K \geq 0.5$. By plugging Eq. (10) into $F(a)$, we have

$$\sum_{i=1}^{K} a_i p_i = \frac{\gamma(1-q)}{\gamma - q},$$

and

$$F(a) = (\gamma - 1) \sum_{i=1}^{K-1} p_i \beta_i^2 - \left(\frac{\gamma - q}{1-q} - 1 \right) \left(\sum_{i=1}^{K-1} p_i \beta_i \right)^2 \geq 0.$$

(b) Assume that there exists some $j \neq K$ such that $q = p_j \geq 0.5$. By plugging Eq. (10) into $F(a)$, we have

$$\begin{aligned} F(a) &= \gamma \left(\sum_{i \neq j}^{K-1} p_i \beta_i^2 + \frac{q(1-q)}{\gamma - q} \beta_j^2 \right) - \frac{\gamma - q}{1-q} \left(\sum_{i \neq j}^{K-1} p_i \beta_i + \frac{q(1-q)}{\gamma - q} \beta_j \right)^2 - \sum_{i=1}^{K-1} p_i \beta_i^2 + \left(\sum_{i=1}^{K-1} p_i \beta_i \right)^2 \\ &= (\gamma - 1) \sum_{i \neq j}^{K-1} p_i \beta_i^2 - \left(\frac{\gamma - q}{1-q} - 1 \right) \left(\sum_{i \neq j}^{K-1} p_i \beta_i \right)^2 \\ &\geq 0. \end{aligned}$$

(c) Assume that $p_k < 0.5$ for all $k = 1, \cdots, K$ and hence $q = 0.5$. Under this case, we can derive with the same way as above and obtain

$$F(a) = (\gamma - 1) \left[\sum_{i=1}^{K-1} p_i \beta_i^2 - \left(\sum_{i=1}^{K-1} p_i \beta_i \right)^2 \right] \geq 0.$$

Combing (a), (b) and (c), we have shown $F(a) \geq 0$ under the second case $1 \leq \gamma < 2q$.

The first case and the second case together imply that Eq. (23) holds, which proves $\gamma \mathbf{S} \succeq \mathbf{S}_{full}$. This leads to the desired variance bound in the theorem.

We now consider the expected sample size. Again, we consider two cases

(1) Consider the first case in Theorem 4.2, where $\gamma \geq 2q$: the conditional expectation of $a(\boldsymbol{x}, C)$ given $\boldsymbol{x}$ is

$$\bar{a}(\boldsymbol{x}) = \mathbb{E}_{C|\boldsymbol{x} \sim \mathcal{D}}\ a(\boldsymbol{x}, C) = \sum_{k=1}^{K} a_k p_k = \frac{4}{\gamma} q(1-q).$$

Therefore, the point-wise conditional expectation of the acceptance probability satisfies $\bar{a}(\boldsymbol{x}) \leq \frac{1}{\gamma}$.

(2) Consider the second case in Theorem 4.2, where $1 \leq \gamma < 2q$: the conditional expectation of $a(\boldsymbol{x}, C)$ given $\boldsymbol{x}$ is

$$\bar{a}(\boldsymbol{x}) = \mathbb{E}_{C|\boldsymbol{x}\sim\mathcal{D}}\ a(\boldsymbol{x}, C) = \sum_{k=1}^{K} a_k p_k = \frac{1}{\gamma}\frac{\gamma^2(1-q)}{\gamma - q}.$$

Note that $0.5 \leq q \leq 1$ and $1 \leq \gamma < 2q$. We have

$$\gamma\bar{a}(\boldsymbol{x}) - 1 = \frac{\gamma^2(1-q)}{\gamma - q} - 1 = \frac{(\gamma-1)(\gamma(1-q)-q)}{\gamma - q} < \frac{q(\gamma-1)(1-2q)}{\gamma - q} \leq 0.$$

Therefore, the point-wise conditional expectation of the acceptance probability satisfies $\bar{a}(\boldsymbol{x}) \leq \frac{1}{\gamma}$.

Combing both (1) and (2), we know that the expected number of accepted examples is

$$n_{Sub} = n\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\bar{a}(\boldsymbol{x}) \leq \frac{n}{\gamma}.$$

This completes the proof of Theorem 4.2. □