Semi-parallel Logistic Regression for GWAS on Encrypted Data

Miran Kim[1], Yongsoo Song[2], Baiyu Li[3], and Daniele Micciancio[3]

[1] University of Texas, Health Science Center at Houston, Houston, USA
`Miran.Kim@uth.tmc.edu`
[2] Microsoft Research, Redmond, USA
`Yogsoo.Song@microsoft.com`
[3] University of California, San Diego, USA
{`baiyu,daniele`}`@cs.ucsd.edu`

Abstract. The sharing of biomedical data is crucial to enable scientific discoveries across institutions and improve health care. For example, genome-wide association studies (GWAS) based on a large number of samples can identify disease-causing genetic variants. The privacy concern, however, has become a major hurdle for data management and utilization. Homomorphic encryption is one of the most powerful cryptographic primitives which can address the privacy and security issues. It supports the computation on encrypted data so that we can aggregate data and perform an arbitrary computation on an untrusted cloud environment without the leakage of sensitive information.

This paper presents a secure outsourcing solution to assess logistic regression models for quantitative traits to test their associations with genotypes. We adapt the semi-parallel training method by Sikorska et al., which builds a logistic regression model for covariates, followed by one-step parallelizable regressions on all individual single nucleotide polymorphisms (SNPs). In addition, we modify our underlying approximate homomorphic encryption scheme for performance improvement.

We evaluate the performance of our solution through experiments on real-world dataset. It achieves the best performance of homomorphic encryption system for GWAS analysis in terms of both complexity and accuracy. For example, given a dataset consisting of 245 samples, each of which has 10643 SNPs and 3 covariates, our algorithm takes about 41 seconds to perform logistic regression based genome wide association analysis over encryption. We demonstrate the feasibility and scalability of our solution.

Keywords: Homomorphic encryption · Genome-wide association studies · Logistic regression.

1 Introduction

Since National Institutes of Health (NIH) released the Gemonic Data Sharing policy allowing the use of cloud computing services for storage and analysis of controlled-access data [2], we are getting more challenge to ensure security and privacy of data in cloud computing systems. In the United States, the Health Insurance Portability and Accountability Act regulates medical care data sharing [28]. A community effort has been made to protect the privacy of genomic data, for example, iDASH (integrating Data for Analysis, Anonymization, Sharing) has hosted secure genome analysis competition for the past 5 years. This contest encourages cryptography experts to develop practical yet rigorous solutions for privacy preserving genomic data analysis. As a result, we could demonstrate the feasibility of secure genome data analysis using various cryptographic primitives such as homomorphic encryption (HE), differential privacy, multi-party computation, and software guard extension. In particular, HE has emerged as one of the promising solutions for secure outsourced computation over genome data in practical biomedical applications [11, 20, 21, 3].

1.1 Summary of Results

In this work, we provide a solution for the second track of iDASH 2018 competition, which aims to develop a method for outsourcing computation of Genome Wide Association Studies (GWAS) on homomorphically encrypted data. We propose a practical protocol to assess logistic regression model to compute p-values of different Single Nucleotide Polymorphisms (SNPs). We investigate the association of genotypes and phenotypes by adjusting the models on the basis of covariates. The results will be used for identifying genetic variants that are statistically correlated with phenotypes of interest.

One year ago, participants of the third task in iDASH 2017 competition were challenged to train a single logistic regression model on encrypted data. Although significant performance improvements over existing solutions have been demonstrated [19, 7], it is still computationally intensive to perform logistic regression based GWAS. A straightforward implementation would require building one model for each SNP, incurring a high performance overhead of secure computation. This motivates the use of the semi-parallel algorithm, which was previously discussed in [26, 27]. Following the approach, our algorithm proceeds in two steps: (1) construct a logistic regression model by applying the gradient descent method of [19] while taking only the covariates into account, (2) compute the regression parameters of logistic regression the corresponding to SNPs with one additional update of Newtons method. This approach enables us to obtain logistic regression based models for thousands of SNPs all in one.

Our solution is based on a homomorphic scheme by Cheon et al. [8] with native support for approximate computation over the real numbers. There has been a significant improvement of performance in [7] by exploiting the Residue Number System (RNS). The authors proposed a different way of performing homomorphic operations which do not require any RNS conversions. In this paper, we present a new variant of this approximate HE scheme which provides advantages in parameter setting and complexity. Specifically, we modify the key-switching procedure which is a core operation in homomorphic multiplication or permutation. The previous method [8, 7] is based on the raising modulus idea suggested in [15], which significantly reduces the size of public key but constrains a large parameter set. We combine this idea with decomposition techniques to use a smaller parameter and thereby achieve a significant performance gain while guaranteeing the same security level. We further leverage efficient packing techniques and parallelization approaches to reduce the required storage space and optimize computational time. Consequently, we could scale up on large-scale association studies with a high accuracy.

1.2 Related Works

There are a number of recent research articles on HE-based machine learning applications. Kim et al. presented the first secure outsourcing method to train a logistic regression model on encrypted data [22] and the follow-up showed remarkably good performance with real data [19, 7]. For example, the training of a logistic regression model took about 3.6 minutes on encrypted data consisting of 1579 samples and 18 features. A slightly different approach is taken in [6], where the authors use Gentry's bootstrapping technique in fully homomorphic encryption so that their solution can run for an arbitrary number of iterations of gradient descent algorithm.

2 Background

The binary logarithm will be simply denoted by $\log(\cdot)$. We denote vectors in bold, e.g. $\mathbf{a}$, and matrices in upper-case bold, e.g. $\mathbf{A}$. For an $n \times m$ matrix $\mathbf{A}$, we use $\mathbf{A}_i$ to denote the i-th row of $\mathbf{A}$, and $\mathbf{a}_j$ the j-th column of $\mathbf{A}$. For a $d_1 \times d$ matrix $\mathbf{A}_1$ and a $d_2 \times d$ matrix $\mathbf{A}_2$, $(\mathbf{A}_1; \mathbf{A}_2)$ denotes the $(d_1 + d_2) \times d$ matrix obtained by concatenating two matrices in a vertical direction. If two matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ have the same number of rows, $(\mathbf{A}_1 | \mathbf{A}_2)$ denotes a matrix formed by horizontal concatenation. We let λ denote the security parameter throughout the paper: all known valid attacks against the cryptographic scheme under scope should take $\Omega(2^\lambda)$ bit operations.

2.1 Logistic Regression

Logistic regression is a widely used statistical model when the response variable is categorical with two possible outcomes [12, 30]. In particular, it is very popular in biomedical informatics research and serve as the foundation of many risk calculators [29, 18, 14, 17].

Let the observed phenotype be given as a vector $\mathbf{y} \in \{\pm 1\}^n$ of length n, the states of p SNPs as the $n \times p$ matrix $\mathbf{S}$, and the states of k covariates as the $n \times k$ matrix $\mathbf{X}$. Suppose that an intercept is included in the matrix of covariates, that is, $\mathbf{X}$ contains a column of ones. For convenience, let $\mathbf{u}_i = (\mathbf{X}_i, s_{ij})$ for $i = 1, \ldots, n$. For each $j \in [p]$, logistic regression aims to find an optimal vector $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$ which maximizes the likelihood estimator

$$\prod_{i=1}^{n} \Pr[y_i | \mathbf{u}_i] = \prod_{i=1}^{n} \sigma(-y_i \cdot \mathbf{u}_i^T \boldsymbol{\beta}),$$

where $\sigma(x) = 1/(1 + \exp(-x))$ is the sigmoid function, or equivalently minimizes the loss function, defined as the negative log-likelihood:

$$L(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-y_i \cdot \mathbf{u}_i^T \boldsymbol{\beta})).$$

Note that $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathbf{X}} | \beta_j)$ depends on the index j, and we are particularly interested in the last component β_j that corresponds to the j-th SNP.

There is no closed form formula for the regression coefficients that minimizes the loss function. Instead, we employ an iterative process: We can begin with some initial guess for the parameters and then repeatedly update them to make the loss smaller until the process converges. Specifically, the gradient descent (GD) takes a step in the direction of the steepest decrease of L. The method of GD can face a problem of zig-zagging along a local optima and this behavior of the method becomes typical if it increases the number of variables of an objective function. We can employ Nesterov's accelerated gradient [23] to address this phenomenon, which uses moving average on the update vector and evaluates the gradient at this *looked-ahead* position.

2.2 Newton's Method

We can alternatively use Newton algorithm to estimate parameters [24]. It can be achieved by calculating the first and the second derivatives of the loss function, followed by the update:

$$\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} - (\nabla_{\boldsymbol{\beta}}^2 L(\boldsymbol{\beta}))^{-1} \cdot \nabla_{\boldsymbol{\beta}} L(\boldsymbol{\beta}).$$

Let $p_i = \sigma(\mathbf{u}_i^T \boldsymbol{\beta})$ for $i \in [n]$; then p_i represents the probability of success for each sample. We see that

$$\nabla_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) = \mathbf{U}^T (\mathbf{y} - \mathbf{p}),\ \nabla_{\boldsymbol{\beta}}^2 L(\boldsymbol{\beta}) = -\mathbf{U}^T \mathbf{W} \mathbf{U}$$

where $\mathbf{U}$ is an $n \times (k+1)$ regressor matrix whose i-th row contains the variables $\mathbf{u}_i$, $\mathbf{p} = (p_i)_{i=1}^n$ is a column vector of the estimated probabilities p_i, and $\mathbf{W}$ is a diagonal weighting matrix with elements $w_i = p_i(1-p_i)$. Then the above update formula can be rewritten as

$$\begin{aligned} \boldsymbol{\beta} &\leftarrow (\mathbf{U}^T\mathbf{W}\mathbf{U})^{-1} \cdot \mathbf{U}^T\mathbf{W} \cdot (\mathbf{U}\boldsymbol{\beta} + \mathbf{W}^{-1}(\mathbf{y}-\mathbf{p})) \\ &= (\mathbf{U}^T\mathbf{W}\mathbf{U})^{-1} \cdot \mathbf{U}^T\mathbf{W}\mathbf{z}. \end{aligned} \tag{1}$$

where $\mathbf{z} = \mathbf{U}\boldsymbol{\beta} + \mathbf{W}^{-1}(\mathbf{y}-\mathbf{p})$. Here, the vector $\mathbf{z}$ is known as the *working response*. This method is also called *Iteratively Reweighted Least Squares*. More details can be found in [24]. On the other hand, the Fisher information $\mathbf{U}^T\mathbf{W}\mathbf{U}$ can be partitioned into a block form:

$$\begin{array}{r@{}c} \begin{array}{r} k\left\{\vphantom{\begin{matrix}a\\a\\a\end{matrix}}\right. \\ 1\{ \end{array} & \left[\begin{array}{c:c} \quad\mathbf{A}\quad & \mathbf{b} \\ \hdashline \mathbf{b}^T & c \end{array}\right] \\ & \begin{array}{cc} \underbrace{\qquad\quad}_{k} & \underbrace{\;\;}_{1} \end{array} \end{array}$$

where $\mathbf{A} = \mathbf{X}^T\mathbf{W}\mathbf{X}$, $\mathbf{s}_j = (s_{ij})_{i=1}^n$ is a column vector of all samples of the j-th SNP, $\mathbf{b} = \mathbf{X}^T\mathbf{W}\mathbf{s}_j$, and $c = \mathbf{s}_j^T\mathbf{W}\mathbf{s}_j$. Then the inverse of $\mathbf{U}^T\mathbf{W}\mathbf{U}$ is

$$\left[\begin{array}{c:c} * & * \\ \hdashline -\frac{1}{t}\mathbf{b}^T\mathbf{A}^{-1} & \frac{1}{t} \end{array}\right]$$

where $t = c - \mathbf{b}^T\mathbf{A}^{-1}\mathbf{b}$. Therefore, the estimated SNP effect β_j and the variance for the estimation are computed by

$$\begin{aligned} \beta_j &= -\frac{1}{t} \cdot (\mathbf{b}^T\mathbf{A}^{-1}) \cdot (\mathbf{X}^T\mathbf{W}\mathbf{z}) + \frac{1}{t} \cdot (\mathbf{s}_j^T\mathbf{W}\mathbf{z}) \\ &= \frac{|\mathbf{A}| \cdot \mathbf{s}_j^T\mathbf{W}\mathbf{z} - \mathbf{b}^T \cdot \mathtt{adj}(\mathbf{A}) \cdot (\mathbf{X}^T\mathbf{W}\mathbf{z})}{|\mathbf{A}| \cdot c - \mathbf{b}^T \cdot \mathtt{adj}(\mathbf{A}) \cdot \mathbf{b}}, \qquad (2) \\ \mathtt{var}_j &= \frac{1}{c - \mathbf{b}^T \cdot \mathbf{A}^{-1} \cdot \mathbf{b}}, \qquad (3) \end{aligned}$$

where $\mathtt{adj}(\mathbf{A})$ denotes the adjugate matrix and $|\mathbf{A}|$ the determinant of $\mathbf{A}$.

3 Full RNS Variant of HEAAN, Revisited

In this paper, we apply the full RNS variant of the HEAAN scheme [8], called RNS-HEAAN [7], for efficient arithmetic over the real numbers. In addition, we modify some algorithms to meet our goals.

The previous RNS-HEAAN scheme uses some approximate modulus switching algorithms for the key-switching procedure. The evaluation key should have a much larger modulus compared to encrypted data due to multiplicative noise. In this work, we developed and implemented a new key-switching algorithm which provides a trade-off between complexity and parameter. Our new key-switching process requires more Number Theoretic Transformation (NTT) conversions, but the HE parameters such as the ring dimension N can be reduced while keeping the same security level. In particular, our method is more efficient than the previous one when the depth of a circuit to be evaluated is small.

The following is a simple description of `RNS-HEAAN` based on the ring learning with errors (RLWE) problem. Let $R = \mathbb{Z}[X]/(X^N + 1)$ be a cyclotomic ring for a power-of-two integer N. An ordinary ciphertext of `RNS-HEAAN` can be represented as a linear polynomial $c(Y) = c_0 + c_1 \cdot Y$ over the ring R_Q where Q denotes the ciphertext modulus and $R_Q = R \pmod{Q}$ is the residue ring modulo Q.

- $\underline{\texttt{Setup}(q, L, \eta; 1^\lambda)}$. Given a base integer module q, a maximum level L of computation, a bit precision η, and a security parameter λ, the `Setup` algorithm generates the following parameters:

  - Choose a basis $\mathcal{D} = \{p_0, q_1, \ldots, q_L\}$ such that $q_i/q \in (1 - 2^{-\eta}, 1 + 2^{-\eta})$ for $1 \le i \le L$. We write $Q_\ell = \prod_{i=1}^{\ell} q_i$ for $1 \le \ell \le L$.
  - Choose a power-of-two integer N.
  - Choose a secret key distribution χ_{key}, an encryption key distribution χ_{err}, and an error distribution χ_{enc} over R.

We always use the RNS form with respect to the basis $\{p_0, q_0, \ldots, q_\ell\}$ (or its sub-basis) to represent polynomials in our scheme. For example, an element $a(X)$ of R_{Q_ℓ} is identified with the tuple $(a_0, a_1, \ldots, a_\ell) \in \prod_{i=0}^{\ell} R_{q_i}$ where $a_i = a \pmod{q_i}$. We point out that all algorithms in our scheme are RNS-friendly so that we do not have to perform any RNS conversions.

The main difference of our scheme from previous work [7] is that the key-switching procedure is based on both the decomposition and modulus raising techniques. The use of decomposition allows us to use a smaller parameter, but its complexity may be increased when the level of HE scheme is large. However, we realize that the GWAS analysis does not require a huge depth, so this new key-switching technique is beneficial to obtain a better performance in this specific application. The generation of switching key and key-switching algorithms are described as follows.

- $\underline{\texttt{KSGen}(s_1, s_2)}$. Given two secret polynomials $s_1, s_2 \in R$, sample $\tilde{a}_i(X) \leftarrow U(R_{p_0 \cdot Q_L})$ and errors $\tilde{e}_i \leftarrow \chi_{err}$ for $1 \le i \le L$. Output the switching key $\mathsf{swk} = \{\mathsf{swk}_i = (\tilde{b}_i, \tilde{a}_i)\}_{1 \le i \le L} \in \left(R^2_{p_0 Q_L}\right)^L$ where $\tilde{b}_i = -\tilde{a} \cdot s_2 + \tilde{e} + p_0 B_i \cdot s_1 \pmod{p_0 \cdot Q_L}$ for the integer $B_i \in \mathbb{Z}_{Q_L}$ such that $B_i = 1 \pmod{p_i}$ and $B_j = 0 \pmod{p_j}$ for all $j \neq i$.
- $\underline{\texttt{KeySwitch}_{\mathsf{swk}}(\mathsf{ct})}$. For $\mathsf{ct} = (c_0, c_1) \in R^2_{Q_\ell}$, let $c_{1i} = c_1 \pmod{q_i}$ for $1 \le i \le \ell$. We first compute

$$\tilde{\mathsf{ct}} = \sum_{i=1}^{\ell} c_{1i} \cdot \mathsf{swk}_i \pmod{p_0 Q_\ell},$$

then return the ciphertext $\mathsf{ct}' = (c_0, 0) + \lfloor p_0^{-1} \cdot \tilde{\mathsf{ct}} \rceil \pmod{Q_\ell}$.

The idea of key-switching procedure is used to relinearize a ciphertext in homomorphic multiplication algorithm below. All other algorithms including key generation, encryption and decryption are exactly same as the previous RNS-based scheme.

- $\underline{\texttt{KeyGen}(1^\lambda)}$.

  - Sample $s \leftarrow \chi_{key}$ and set the secret key as $\mathsf{sk} = (1, s)$.

- Sample $a \leftarrow U(R_{Q_L})$ and $e \leftarrow \chi_{err}$. Set the public key pk as $\mathsf{pk} = (b, a) \in R_{Q_L}^2$ where $b = -a \cdot s + e \pmod{Q_L}$.
- Set the evaluation key as $\mathsf{evk} \leftarrow \mathtt{KSGen}(s^2, s)$.

- $\underline{\mathtt{Enc}_{\mathsf{pk}}(m)}$. Given $m \in R$, sample $v \leftarrow \chi_{enc}$ and $e_0, e_1 \leftarrow \chi_{err}$. Output the ciphertext $\mathsf{ct} = v \cdot \mathsf{pk} + (m + e_0, e_1) \pmod{Q_L}$.
- $\underline{\mathtt{Dec}_{\mathsf{sk}}(\mathsf{ct})}$. For given ciphertext $\mathsf{ct} \in R_{Q_\ell}^2$, output $\langle \mathsf{ct}^{(0)}, \mathsf{sk} \rangle \pmod{q_0}$.
- $\underline{\mathtt{Add}(\mathsf{ct}, \mathsf{ct}')}$. Given two ciphertexts $\mathsf{ct}, \mathsf{ct}' \in R_{Q_\ell}^2$, output the ciphertext $\mathsf{ct}_{\mathsf{add}} = \mathsf{ct} + \mathsf{ct}' \pmod{Q_\ell}$.
- $\underline{\mathtt{Mult}_{\mathsf{evk}}(\mathsf{ct}, \mathsf{ct}')}$. For two ciphertexts $\mathsf{ct} = (c_0, c_1)$ and $\mathsf{ct}' = (c_0', c_1')$, compute

$$d_0 = c_0 c_0', \;\; d_1 = c_0 c_1' + c_0' c_1, \;\; d_2 = c_1 c_1' \pmod{Q_\ell}.$$

Let $c_{2i} = d_2 \pmod{q_i}$ for $1 \leq i \leq \ell$, and compute

$$\tilde{\mathsf{ct}} = \sum_{i=1}^{\ell} d_{2i} \cdot \mathsf{evk}_i \pmod{p_0 Q_\ell}.$$

Output the ciphertext $\mathsf{ct}' = (c_0, c_1) + \lfloor p_0^{-1} \cdot \tilde{\mathsf{ct}} \rceil \pmod{Q_\ell}$.

Finally, RNS-HEAAN provides the rescaling operation to round messages over encryption, thereby enabling to control the magnitude of messages during computation.

- $\underline{\mathtt{ReScale}(\mathsf{ct})}$. For given $\mathsf{ct} \in \mathcal{R}_{Q_\ell}^2$, return the ciphertext $\mathsf{ct}' = \lfloor q_\ell^{-1} \cdot \mathsf{ct} \rceil \pmod{Q_{\ell-1}}$.

We usually rescale the encrypted message after multiplication as we round-off the significant digits after multiplication in plain fixed/floating point computation. In the next section, we assume that the rescaling procedure is included in homomorphic multiplications for simpler description, but a rigorous analysis about level consumption will be provided later in the parameter setting section.

As in the original HEAAN scheme, the native plaintext space can be understood as an $N/2$-dimensional complex vector space (each vector component is called a *plaintext slot*). The addition and multiplication in R correspond to the component-wise addition and multiplication on plaintext slots. Furthermore, it provides an operation that shifts the plaintext vector over encryption: For a ciphertext ct encrypting a plaintext vector $(m_1, \ldots, m_\ell) \in \mathbb{R}^\ell$, we could obtain an encryption of a shifted vector $(m_{r+1}, \ldots, m_\ell, m_1, \ldots, m_r)$. Let us denote such operation by $\mathsf{Rot}(\mathsf{ct}; r)$. For more detailed, we refer the reader to [7]. In the rest of this paper, we let $N_2 = N/2$ and denote by $\mathtt{E}(\cdot)$ the encryption function for convenience.

4 Our Method

4.1 Database Encoding

As noted before, the learning data are recorded into an $n \times k$ matrix $\mathbf{X}$ of covariates, an $n \times p$ binary matrix $\mathbf{S} = (s_{ij})$ of all the SNP data, and an n-dimensional binary column vector $\mathbf{y}$ of the dependent variable. In large-scale GWAS, the number of parameters of SNPs, p can be in the thousands, so we split the SNP data into several N_2-dimensional vectors, encrypt them, and send the resulting ciphertexts to the server. For simplicity, we assume in the following discussion that each row of $\mathbf{S}$ is encrypted into a single ciphertext. More specifically, for $1 \leq i \leq n$ and for $1 \leq \ell \leq k$, we encrypt

$$\mathtt{E}(x_{i\ell}\mathbf{S}_i) = \mathtt{E}(x_{i\ell}s_{i1}, \ldots, x_{i\ell}s_{ip}).$$

As we mentioned earlier, we add a column of ones to $\mathbf{X}$ to allow for an intercept in the regression; that is, we assume $x_{i1} = 1$ for all $1 \leq i \leq n$. So, when $\ell = 1$, the ciphertext $\mathtt{E}(x_{i1}\mathbf{S}_i)$ encrypts exactly the i-th SNP sample.

Next, consider the matrix $\mathbf{y}^T\mathbf{X} \in \mathbb{R}^{n \times k}$ defined as

$$\begin{aligned}\mathbf{y}^T\mathbf{X} &= \left[y_1\mathbf{X}_1; \cdots; y_n\mathbf{X}_n\right] \\ &= \begin{bmatrix} y_1x_{11} & y_1x_{12} & \cdots & y_1x_{1k} \\ y_2x_{21} & y_2x_{22} & \cdots & y_2x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ y_nx_{n1} & y_nx_{n2} & \cdots & y_nx_{nk} \end{bmatrix}.\end{aligned}$$

For simplicity, we assume that n and k are power-of-two integers satisfying $\log n + \log k \leq \log(N_2)$. Kim et al. [19] suggested an efficient encoding map to encode the whole matrix $\mathbf{y}^T\mathbf{X}$ in a single ciphertext in a row-by-row manner. Specifically, we will identify this matrix with a vector in $\mathbb{R}^{n \cdot k}$, that is,

$$\begin{aligned}\mathbf{y}^T\mathbf{X} &\mapsto (y_1\mathbf{X}_1|\cdots|y_n\mathbf{X}_n) \\ &= (y_1x_{11}, \ldots, y_1x_{1k}, \ldots, y_nx_{n1}, \ldots, y_nx_{nk})\end{aligned}$$

Similarly, we identify the matrix $\mathbf{X}$ with a vector in $\mathbb{R}^{n \cdot k}$ as follows:

$$\begin{aligned}\mathbf{X} &\mapsto (\mathbf{X}_1|\cdots|\mathbf{X}_n) \\ &= (x_{11}, \ldots, x_{1k}, \ldots, x_{n1}, \ldots, x_{nk}).\end{aligned}$$

For an efficient implementation, we can make $N_2/(k \cdot n)$ copies of each component of $\mathbf{y}^T\mathbf{X}$ and $\mathbf{X}$ to encode them into fully packed plaintext slots. For example, we can generate the encryption of $\mathbf{y}^T\mathbf{X}$ as

$$\mathtt{E}(\mathbf{y}^T\mathbf{X}) = \mathtt{E}\left(y_1\mathbf{X}_1^{(N_2/(k\cdot n))}|\cdots|y_n\mathbf{X}_n^{(N_2/(k\cdot n))}\right),$$

where $y_i\mathbf{X}_i^{(N_2/(k\cdot n))}$ denotes an array containing $N_2/(k \cdot n)$ copies of $y_i\mathbf{X}_i$. In the case of the target vector $\mathbf{y}$, we make N_2/n copies of each entry, so that the encoding aligns y_i with each copies of $y_i\mathbf{X}_i$ and $\mathbf{X}_i$ in the ciphertexts. Let us denote the generated ciphertext by $\mathtt{E}(\mathbf{y})$.

Finally, we now consider how to encrypt a covariance matrix $\mathbf{X}^T\mathbf{X}$ which can be used for computing the adjugate matrix and determinant of $\mathbf{A} = \mathbf{X}^T\mathbf{W}\mathbf{X}$. The adjugate $\mathtt{adj}(\mathbf{A})$ is a $k \times k$ matrix whose entries are defined as

$$\mathtt{adj}(\mathbf{A})_{j\ell} := (-1)^{j+\ell} \cdot |\hat{\mathbf{A}}_{\ell j}|$$

for $1 \leq j, \ell \leq k$, where $|\hat{\mathbf{A}}_{\ell j}|$ is the determinant of $\hat{\mathbf{A}}_{\ell j}$. Here, $\hat{\mathbf{A}}_{\ell j}$ is a $(k - 1) \times (k - 1)$ submatrix obtained by removing the j-th column and ℓ-th row from $\mathbf{A}$. For example, when $k = 4$, the determinant $|\hat{\mathbf{A}}_{11}|$ is computed by $a_{22}(a_{33}a_{44}-a_{34}a_{43})+a_{23}(a_{34}a_{42}-a_{32}a_{44})+a_{24}(a_{32}a_{43}-a_{33}a_{42})$, which can be rewritten as a component-wise product of three vectors

$$\begin{aligned}\mathbf{A}_{1,1,1} &= (a_{22}, -a_{22}, a_{23}, -a_{23}, a_{24}, -a_{24}), \\ \mathbf{A}_{1,1,2} &= (a_{33}, -a_{34}, a_{34}, -a_{32}, a_{32}, -a_{33}), \\ \mathbf{A}_{1,1,3} &= (a_{44}, -a_{43}, a_{42}, -a_{44}, a_{43}, -a_{42}).\end{aligned}$$

Algorithm 1 $\texttt{AllSum}(\mathsf{ct}, \psi, \alpha)$

```
Input: $\mathsf{ct}$, input ciphertext, the unit initial amount by which the ciphertext shifts $\psi$, the number
    of the repeated doubling $\alpha$
for $i = 0, 1, \ldots, \log \alpha - 1$ do
    Compute $\mathsf{ct} \leftarrow \mathsf{Add}(\mathsf{ct}, \mathsf{Rot}(\mathsf{ct}; \psi \cdot 2^i))$
end for
return $\mathsf{ct}$
```

In general, we can consider $(k-1)!$-dimensional vectors $\mathbf{A}_{j,\ell,1}, \mathbf{A}_{j,\ell,2}, \ldots, \mathbf{A}_{j,\ell,(k-1)}$ that can be used to compute $|\hat{\mathbf{A}}_{\ell j}|$. To do so, for each $i \in [n]$, we first pre-compute the i-th covariance matrix $\mathbf{X}_i^T \mathbf{X}_i \in \mathbb{R}^{k \times k}$ and generate the corresponding vector $(\mathbf{X}_i^T \mathbf{X}_i)_{j,\ell,t}$ for $1 \leq j \leq \ell \leq k$ and $1 \leq t \leq k-1$. Suppose that $N_2 \geq n \cdot (k-1)!$. Let $\phi = N_2/(n \cdot (k-1)!)$, and we encrypt the following concatenated vector

$$\Sigma_{j,\ell,t} = \left((\mathbf{X}_1^T \mathbf{X}_1)_{j,\ell,t}^{(\phi)} \mid \ \ldots \ |(\mathbf{X}_n^T \mathbf{X}_n)_{j,\ell,t}^{(\phi)} \right).$$

We denote the resulting ciphertext by $\mathtt{E}(\Sigma_{j,\ell,t})$.

4.2 Homomorphic Evaluation of Logistic Regression

The main idea of the semi-parallel logistic regression analysis [26, 27] is to assume that the probabilities predicted by a model without SNP will not change much once SNP is included to the model. We will follow their approach, where the first step is to construct a logistic regression model taking only the covariates into account, and the second step is to compute the model coefficients of the logistic regression corresponding to the SNP in a semi-parallel way.

We start with a useful aggregation operation across plaintext slots from the literature [16, 9, 10]. This algorithm is referred as $\texttt{AllSum}$, which is parameterized by integers ψ and α. See Algorithm 1 for an implementation. Let $\ell = \psi \cdot \alpha$. Given a ciphertext ct representing a plaintext vector $\mathbf{m} = (m_1, \ldots, m_\ell) \in \mathbb{R}^\ell$, the $\texttt{AllSum}$ algorithm outputs a ciphertext ct' encrypting

$$\begin{aligned} \mathbf{m}' = (&\sum_{j=0}^{\alpha-1} m_{\psi j+1}, \sum_{j=0}^{\alpha-1} m_{\psi j+2}, \ldots \sum_{j=0}^{\alpha-1} m_{\psi(j+1)}, \\ &\sum_{j=0}^{\alpha-1} m_{\psi j+1}, \sum_{j=0}^{\alpha-1} m_{\psi j+2}, \ldots \sum_{j=0}^{\alpha-1} m_{\psi(j+1)}, \ldots), \end{aligned}$$

i.e., $m'_i = \sum_{j=0}^{\alpha-1} m_{\psi j+i}$ for $1 \leq i \leq \psi$, and $m'_{\psi j+i} = m'_i$ for $1 \leq j \leq \alpha - 1$.

As mentioned before, our algorithm consists of two steps to perform the semi-parallel logistic regression training while taking as input the following ciphertexts: $\{\mathtt{E}(x_{i\ell}\mathbf{S}_i)\}$, $\mathtt{E}(\mathbf{y}^T\mathbf{X})$, $\mathtt{E}(\mathbf{X})$, $\mathtt{E}(\mathbf{y})$, and $\{\mathtt{E}(\Sigma_{j,\ell,t})\}$, for $1 \leq i \leq n$, $1 \leq j \leq \ell \leq k$, and $1 \leq t \leq k-1$.

4.2.1 Logistic Regression Model Training for Covariates

The best solution to train a logistic regression model from homomorphically encrypted dataset is to evaluate Nesterov's accelerated gradient descent method [19, 7]. We adapt their evaluation strategy to train a model for covariates. In particular, we employ a degree 3 approximation polynomial of the sigmoid function: $\sigma_3(x) = 0.5 + 0.15012x - 0.001593x^3$.

Step 0: For simplicity, let $\mathbf{v}_i = y_i \mathbf{X}_i$ and $\ell = N_2/(k \cdot n)$. Since the input ciphertext $\mathtt{E}(\mathbf{y}^T\mathbf{X})$ represents ℓ copies of $\mathbf{v}_i$, Step 6 in [19] outputs the following ciphertext that encrypts the same number of copies of the vectors $\sigma_3(\mathbf{v}_i^T \boldsymbol{\beta}_{\mathbf{X}}) \cdot \mathbf{v}_i$:

$$\mathtt{ct}_6 = \mathtt{E}\begin{bmatrix} \sigma_3(\mathbf{v}_1^T\boldsymbol{\beta}_{\mathbf{X}}) \cdot v_{11} & \cdots & \sigma_3(\mathbf{v}_1^T\boldsymbol{\beta}_{\mathbf{X}}) \cdot v_{1k} \\ \vdots & \ddots & \vdots \\ \sigma_3(\mathbf{v}_1^T\boldsymbol{\beta}_{\mathbf{X}}) \cdot v_{11} & \cdots & \sigma_3(\mathbf{v}_1^T\boldsymbol{\beta}_{\mathbf{X}}) \cdot v_{1k} \\ \vdots & \ddots & \vdots \\ \sigma_3(\mathbf{v}_n^T\boldsymbol{\beta}_{\mathbf{X}}) \cdot v_{n1} & \cdots & \sigma_3(\mathbf{v}_n^T\boldsymbol{\beta}_{\mathbf{X}}) \cdot v_{nk} \\ \vdots & \ddots & \vdots \\ \sigma_3(\mathbf{v}_n^T\boldsymbol{\beta}_{\mathbf{X}}) \cdot v_{n1} & \cdots & \sigma_3(\mathbf{v}_n^T\boldsymbol{\beta}_{\mathbf{X}}) \cdot v_{nk} \end{bmatrix}.$$

Then Step 7 in [19] is changed from $\mathtt{AllSum}(\mathtt{ct}_6, k, n)$ into $\mathtt{ct}_7 = \mathtt{AllSum}(\mathtt{ct}_6, N_2/n, n)$, so that the output ciphertext is as follows:

$$\mathtt{ct}_7 = \mathtt{E}\begin{bmatrix} \sum_i \sigma_3(\mathbf{v}_i^T\boldsymbol{\beta}_{\mathbf{X}}) \cdot v_{i1} \cdots \sum_i \sigma_3(\mathbf{v}_i^T\boldsymbol{\beta}_{\mathbf{X}}) \cdot v_{ik} \\ \sum_i \sigma_3(\mathbf{v}_i^T\boldsymbol{\beta}_{\mathbf{X}}) \cdot v_{i1} \cdots \sum_i \sigma_3(\mathbf{v}_i^T\boldsymbol{\beta}_{\mathbf{X}}) \cdot v_{ik} \\ \vdots \quad\quad \ddots \quad\quad \vdots \\ \sum_i \sigma_3(\mathbf{v}_i^T\boldsymbol{\beta}_{\mathbf{X}}) \cdot v_{i1} \cdots \sum_i \sigma_3(\mathbf{v}_i^T\boldsymbol{\beta}_{\mathbf{X}}) \cdot v_{ik} \end{bmatrix}.$$

In the end, the model parameters $\boldsymbol{\beta}_{\mathbf{X}}$ are encrypted as a ciphertext with fully-packed plaintext slots. More precisely, it yields encrypted model parameters $\mathtt{E}(\boldsymbol{\beta}_{\mathbf{X}})$ that represent a plaintext vector containing $N_2/k = \ell \cdot n$ copies of $\boldsymbol{\beta}_{\mathbf{X}}$ as follows:

$$\mathtt{E}(\boldsymbol{\beta}_{\mathbf{X}}) = \begin{bmatrix} \beta_{X1} & \beta_{X2} & \cdots & \beta_{Xk} \\ \beta_{X1} & \beta_{X2} & \cdots & \beta_{Xk} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{X1} & \beta_{X2} & \cdots & \beta_{Xk} \end{bmatrix}.$$

4.2.2 Parallel Logistic Regression Model Building for SNPs

Starting with $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathbf{X}}, 0) \in \mathbb{R}^{k+1}$, we will perform one step of Newton's method for regression with SNPs. This implies that the regression coefficients multiplied by the values of the predictor are $\mathbf{U}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}_{\mathbf{X}}$, so for all $i \in [n]$, if we let the predicted value be $\hat{y}_i = \mathbf{u}_i^T\boldsymbol{\beta}$, then we have $\hat{y}_i = \mathbf{x}_i^T\boldsymbol{\beta}_{\mathbf{X}}$. We note that

$$\begin{aligned} (\mathbf{W}\mathbf{z})_i &= w_i \cdot z_i \\ &= p_i(1-p_i) \cdot \left(\hat{y}_i + \frac{y_i - p_i}{p_i \cdot (1-p_i)}\right) \\ &= p_i(1-p_i) \cdot \hat{y}_i + (y_i - p_i). \end{aligned} \tag{4}$$

with $p_i = \sigma_3(\hat{y}_i)$. In the following, we describe how to securely evaluate these variables from the model parameters $\boldsymbol{\beta}_{\mathbf{X}}$. In the end, the server outputs the numerator and the denominator of Equation (2), denoted by $\beta_j^\star$ and $\beta_j^\dagger$.

Step 1: Let $\hat{\mathbf{y}} = (\hat{y}_i)_{i=1}^n$ be a column vector of the predicted values. The goal of this step is to generate its encryption. The public server first performs the homomorphic multiplication between two ciphertexts $\mathtt{E}(\boldsymbol{\beta}_{\mathbf{X}})$ and $\mathtt{E}(\mathbf{X})$, and then applies `AllSum` to the resulting ciphertext:

$$\mathtt{E}(\hat{\mathbf{y}}_\star) \leftarrow \mathtt{AllSum}(\mathtt{E}(\boldsymbol{\beta}_{\mathbf{X}}) \cdot \mathtt{E}(\mathbf{X}), 1, k). \tag{5}$$

The output ciphertext $\mathtt{E}(\hat{\mathbf{y}}_\star)$ encrypts the values $\hat{y}_i$ at the $(t \cdot k+1)$ positions for $(i-1) \cdot \ell \leq t < i \cdot \ell$ and some garbage values in the other entries, denoted by $\star$, i.e.,

$$\mathtt{E}(\hat{\mathbf{y}}_\star) = \mathtt{E}\begin{bmatrix} \hat{y}_1 & \star & \cdots & \star \\ \vdots & \vdots & \ddots & \vdots \\ \hat{y}_1 & \star & \cdots & \star \\ \vdots & \vdots & \ddots & \vdots \\ \hat{y}_n & \star & \cdots & \star \\ \vdots & \vdots & \ddots & \vdots \\ \hat{y}_n & \star & \cdots & \star \end{bmatrix}.$$

The server then performs a constant multiplication by c to annihilate the garbage values. The polynomial $c \leftarrow \mathtt{Encode}(C)$ is the encoding of the following matrix, where $\mathtt{Encode}(\cdot)$ is a standard procedure [8] to encode a real vector as a ring element in R:

$$C = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \end{bmatrix}.$$

The next step is to replicate the values $\hat{y}_i$ to other columns:

$$\mathtt{E}(\hat{\mathbf{y}}) \leftarrow \mathtt{AllSum}(\mathtt{CMult}(\mathtt{E}(\hat{\mathbf{y}}_\star); c), -1, k),$$

denoted by $\mathtt{CMult}(\cdot)$ a scalar multiplication. So the output ciphertext $\mathtt{E}(\hat{\mathbf{y}})$ has $N_2/n = \ell \cdot k$ copies of $\hat{y}_i$:

$$\mathtt{E}(\hat{\mathbf{y}}) = \mathtt{E}\begin{bmatrix} \hat{y}_1 & \hat{y}_1 & \cdots & \hat{y}_1 \\ \vdots & \vdots & \ddots & \vdots \\ \hat{y}_1 & \hat{y}_1 & \cdots & \hat{y}_1 \\ \vdots & \vdots & \ddots & \vdots \\ \hat{y}_n & \hat{y}_n & \cdots & \hat{y}_n \\ \vdots & \vdots & \ddots & \vdots \\ \hat{y}_n & \hat{y}_n & \cdots & \hat{y}_n \end{bmatrix}.$$

Step 2: This step is simply to evaluate the approximating polynomial of the sigmoid function by applying the pure SIMD additions and multiplications:

$$\mathtt{E}(\mathbf{p}) \leftarrow \sigma_3(\mathtt{E}(\hat{\mathbf{y}})).$$

Then the server securely compute the weights w_i and their multiplication with the working response vector $\mathbf{z}$ using Equation (4):

$$\begin{aligned} \mathtt{E}(\mathbf{w}) &\leftarrow \mathtt{E}(\mathbf{p}) \cdot (1 - \mathtt{E}(\mathbf{p})), \\ \mathtt{E}(\mathbf{Wz}) &\leftarrow \mathtt{E}(\mathbf{w}) \cdot \mathtt{E}(\hat{\mathbf{y}}) + (\mathtt{E}(\mathbf{y}) - \mathtt{E}(\mathbf{p})). \end{aligned} \tag{6}$$

Here the two output ciphertexts contain N_2/n copies of the values w_i and $w_i z_i$, respectively:

$$\mathtt{E}(\mathbf{w}) = \mathtt{E}\begin{bmatrix} w_1 & w_1 & \cdots & w_1 \\ \vdots & \vdots & \ddots & \vdots \\ w_1 & w_1 & \cdots & w_1 \\ \vdots & \vdots & \ddots & \vdots \\ w_n & w_n & \cdots & w_n \\ \vdots & \vdots & \ddots & \vdots \\ w_n & w_n & \cdots & w_n \end{bmatrix}, \quad \mathtt{E}(\mathbf{Wz}) = \mathtt{E}\begin{bmatrix} w_1z_1 & w_1z_1 & \cdots & w_1z_1 \\ \vdots & \vdots & \ddots & \vdots \\ w_1z_1 & w_1z_1 & \cdots & w_1z_1 \\ \vdots & \vdots & \ddots & \vdots \\ w_nz_n & w_nz_n & \cdots & w_nz_n \\ \vdots & \vdots & \ddots & \vdots \\ w_nz_n & w_nz_n & \cdots & w_nz_n \end{bmatrix}.$$

Step 3: The goal of this step is to generate trivial encryptions $\mathtt{E}(w_i)$ such that for $i \in [n]$, $\mathtt{E}(w_i)$ has w_i in all positions of its plaintext vector. We employ the hybrid algorithm of [16] for replication, denoted by $\mathtt{Replicate}(\cdot)$. The server outputs n ciphertexts

$$\{\mathtt{E}(w_i)\}_{1 \le i \le n} \leftarrow \mathtt{Replicate}(\mathtt{E}(\mathbf{w})).$$

Similarly, the server takes in the ciphertext $\mathtt{E}(\mathbf{Wz})$ and performs another replication operation:

$$\{\mathtt{E}(w_i z_i)\}_{1 \le i \le n} \leftarrow \mathtt{Replicate}(\mathtt{E}(\mathbf{Wz})).$$

Step 4: For all $j \in [p]$, we define the vector $\mathbf{b}_j = \mathbf{X}^T\mathbf{W}\mathbf{s}_j \in \mathbb{R}^k$ and denote the ℓ-th component of $\mathbf{b}_j$ by $b_{j\ell}$. We note that $b_{j\ell} = \mathbf{x}_\ell^T\mathbf{W}\mathbf{s}_j = \sum_{i=1}^n (x_{i\ell} \cdot w_i \cdot s_{ij})$, where $\mathbf{x}_\ell = (x_{i\ell})_{i=1}^n$ is the j-th column of the design matrix $\mathbf{X}$. Then for all $\ell \in [k]$, the server generates encryptions of the vectors $\mathbf{B}_\ell = \mathbf{x}_\ell^T\mathbf{W}\mathbf{S} = (b_{1\ell}, b_{2\ell}, \ldots, b_{p\ell})$ by computing

$$\mathtt{E}(\mathbf{B}_\ell) \leftarrow \sum_{i=1}^n \mathtt{E}(w_i) \cdot \mathtt{E}(x_{i\ell}\mathbf{S}_i). \tag{7}$$

On the other hand, since we add a column of ones to the matrix $\mathbf{X}$, we have $c_j = \mathbf{s}_j^T\mathbf{W}\mathbf{s}_j = \sum_{i=1}^n w_i \cdot s_{ij} = \sum_{i=1}^n x_{i1} \cdot w_i \cdot s_{ij} = b_{1j}$ for $j \in [p]$, which implies $\mathtt{E}(\mathbf{B}_1)$ can be understood as an encryption of $(c_1, c_2, \ldots, c_p)$.

Step 5: This step is to securely compute the values $\mathbf{s}_j^T\mathbf{Wz} = \sum_{i=1}^n s_{ij} \cdot w_i \cdot z_i$ for $j \in [p]$. Specifically, the server performs the following computation:

$$\mathtt{E}(\mathbf{s}_1^T\mathbf{Wz}, \ldots, \mathbf{s}_p^T\mathbf{Wz}) \leftarrow \sum_{i=1}^n \mathtt{E}(w_i z_i) \cdot \mathtt{E}(x_{i1}\mathbf{S}_i). \tag{8}$$

Step 6: The goal of this step is to securely compute the vector $\mathbf{X}^T\mathbf{Wz}$ such that the ℓ-th element is obtained by $\mathbf{x}_\ell^T\mathbf{Wz} = \sum_{i=1}^n (x_{i\ell} \cdot w_i \cdot z_i)$ for $\ell \in [k]$. The server first performs the pure SIMD multiplication between two ciphertexts $\mathtt{E}(\mathbf{X})$ and $\mathtt{E}(\mathbf{Wz})$:

$$\mathtt{E}(\mathbf{X} \odot \mathbf{Wz}) \leftarrow \mathtt{E}(\mathbf{X}) \cdot \mathtt{E}(\mathbf{Wz}). \tag{9}$$

Here, the output ciphertext $\mathtt{E}(\mathbf{X} \odot \mathbf{Wz})$ encrypts the values $x_{i\ell}w_iz_i$:

$$\mathtt{E}(\mathbf{X} \odot \mathbf{Wz}) = \mathtt{E}\begin{bmatrix} x_{11}w_1z_1 & x_{12}w_1z_1 & \cdots & x_{1k}w_1z_1 \\ \vdots & \vdots & \ddots & \vdots \\ x_{11}w_1z_1 & x_{12}w_1z_1 & \cdots & x_{1k}w_1z_1 \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1}w_nz_n & x_{n2}w_nz_n & \cdots & x_{nk}w_nz_n \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1}w_nz_n & x_{n2}w_nz_n & \cdots & x_{nk}w_nz_n \end{bmatrix}.$$

Then the server aggregates the values in the same column to obtain a ciphertext encrypting $\mathbf{x}_\ell^T\mathbf{W}\mathbf{z}$:

$$\mathtt{E}(\mathbf{X}^T\mathbf{W}\mathbf{z}) \leftarrow \mathtt{AllSum}(\mathtt{E}(\mathbf{X}\odot\mathbf{W}\mathbf{z})), N_2/(k\cdot n), n).$$

Notice that this ciphertext contains the scalar $\mathbf{x}_\ell^T\mathbf{W}\mathbf{z}$ in every entry of the ℓ-th column, for $1 \le \ell \le k$:

$$\mathtt{E}(\mathbf{X}^T\mathbf{W}\mathbf{z}) = \mathtt{E}\begin{bmatrix} \mathbf{x}_1^T\mathbf{W}\mathbf{z} & \mathbf{x}_2^T\mathbf{W}\mathbf{z} & \cdots & \mathbf{x}_k^T\mathbf{W}\mathbf{z} \\ \mathbf{x}_1^T\mathbf{W}\mathbf{z} & \mathbf{x}_2^T\mathbf{W}\mathbf{z} & \cdots & \mathbf{x}_k^T\mathbf{W}\mathbf{z} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_1^T\mathbf{W}\mathbf{z} & \mathbf{x}_2^T\mathbf{W}\mathbf{z} & \cdots & \mathbf{x}_k^T\mathbf{W}\mathbf{z} \end{bmatrix}.$$

Finally, it outputs k ciphertexts, each encrypting $\mathbf{x}_\ell^T\mathbf{W}\mathbf{z}$ for $1 \le \ell \le k$, by applying the replication operation as follows:

$$\{\mathtt{E}(\mathbf{x}_\ell^T\mathbf{W}\mathbf{z})\}_{1\le\ell\le k} \leftarrow \mathtt{Replicate}(\mathtt{E}(\mathbf{X}^T\mathbf{W}\mathbf{z})).$$

Step 7: The goal of this step is to compute the encryptions of the adjugate matrix and the determinant of $\mathbf{A} = \mathbf{X}^T\mathbf{W}\mathbf{X}$. We note that

$$\mathbf{A}_{r,s,t} = \left(\sum_{i=1}^{n} w_i \cdot \mathbf{X}_i^T\mathbf{X}_i\right)_{r,s,t} = \sum_{i=1}^{n} w_i \cdot (\mathbf{X}_i^T\mathbf{X}_i)_{r,s,t} \tag{10}$$

for $1 \le r \le s \le k$ and $1 \le t \le k-1$. The server first multiplies the ciphertexts $\mathtt{E}(\Sigma_{r,s,t})$ with the ciphertext $\mathtt{E}(\mathbf{w})$ to obtain

$$\mathtt{E}(\Sigma'_{r,s,t}) \leftarrow \mathtt{E}(\mathbf{w}) \cdot \mathtt{E}(\Sigma_{r,s,t}). \tag{11}$$

Here, the ciphertext $\mathtt{E}(\Sigma'_{r,s,t})$ encrypts n vectors $w_i \cdot (\mathbf{X}_i^T\mathbf{X}_i)_{r,s,t}$ for $1 \le i \le n$. Then we apply $\mathtt{AllSum}$ to aggregate these vectors and obtain $\mathbf{A}_{r,s,t}$ as in Equation (10):

$$\mathtt{E}(\mathbf{A}_{r,s,t}) \leftarrow \mathtt{AllSum}(\mathtt{E}(\Sigma'_{r,s,t}), \phi, n).$$

Next, the server performs multiplications between the ciphertexts $\mathtt{E}(\mathbf{A}_{r,s,t})$ as follows:

$$\mathtt{E}(\Sigma_{r,s}) \leftarrow \prod_{t=1}^{k-1} \mathtt{E}(\mathbf{A}_{r,s,t}). \tag{12}$$

The adjugate matrix can be obtained by aggregating $(k-1)!$ number of values in $\mathtt{E}(\Sigma_{r,s})$:

$$\mathtt{E}(\mathtt{adj}(\mathbf{A})_{r,s}) \leftarrow \mathtt{AllSum}(\mathtt{E}(\Sigma_{r,s}), 1, (k-1)!).$$

In addition, the server computes

$$\mathtt{E}(x_{1r}\mathbf{W}) \leftarrow \mathtt{AllSum}(\mathtt{E}(x_{1r}) \cdot \mathtt{E}(\mathbf{w}), N_2/n, n)$$

for $1 \le r \le k$ and obtains a trivial encryption of the determinant of $\mathbf{A}$ as follows:

$$\mathtt{E}(|\mathbf{A}|) \leftarrow \sum_{r=1}^{k} \mathtt{E}(x_{1r}\mathbf{W}) \cdot \mathtt{E}(\mathtt{adj}(\mathbf{A})_{1r}).$$

Step 8: The final step is to securely compute the encryptions of $\mathtt{E}(\boldsymbol{\beta}^*)$ and $\mathtt{E}(\boldsymbol{\beta}_*)$ by pure SIMD additions and multiplications. We note that the multiplication of the vectors $\mathbf{B}_j$ from the left side and $(\mathbf{X}^T\mathbf{W}\mathbf{z})$ from the right side with the matrix $\mathtt{adj}(A)$ can be written as

$$\mathbf{B}_j^T \cdot \mathtt{adj}(A) \cdot (\mathbf{X}^T\mathbf{W}\mathbf{z}) = \sum_{r,s=1}^{k} b_{jr} \cdot (\mathtt{adj}(A))_{r,s} \cdot (\mathbf{X}^T\mathbf{W}\mathbf{z})_s$$

So the server evaluates the numerator of Equation (2) to get the encryption of $\boldsymbol{\beta}^*$:

$$\mathtt{E}(\boldsymbol{\beta}^*) \leftarrow \mathtt{E}(|\mathbf{A}|) \cdot \mathtt{E}(\mathbf{s}_1^T\mathbf{W}\mathbf{z}, \ldots, \mathbf{s}_p^T\mathbf{W}\mathbf{z}) - \sum_{r,s=1}^{k} \mathtt{E}(\mathbf{B}_r) \cdot \mathtt{E}(\mathtt{adj}(\mathbf{A})_{rs}) \cdot \mathtt{E}(\mathbf{x}_s^T\mathbf{W}\mathbf{z}). \tag{13}$$

Then the output ciphertext $\mathtt{E}(\boldsymbol{\beta}^*)$ encrypts the values $\boldsymbol{\beta}_j^*$ in a way that $\mathtt{E}(\boldsymbol{\beta}^*) = \mathtt{E}(\beta_1^*, \beta_2^*, \ldots, \beta_p^*)$. Similarly, we evaluate the denominator of Equation (2) to get an encryption of $\boldsymbol{\beta}^\dagger$:

$$\mathtt{E}(\boldsymbol{\beta}^\dagger) \leftarrow \mathtt{E}(|\mathbf{A}|) \cdot \mathtt{E}(c_1, c_2, \ldots, c_p) \; - \sum_{r,s=1}^{k} \mathtt{E}(\mathbf{B}_r) \cdot \mathtt{E}(\mathtt{adj}(\mathbf{A})_{rs}) \cdot \mathtt{E}(\mathbf{B}_s). \tag{14}$$

Hence, the output ciphertext $\mathtt{E}(\boldsymbol{\beta}^\dagger)$ represents the values $\boldsymbol{\beta}_j^\dagger$ in a way that $\mathtt{E}(\boldsymbol{\beta}^\dagger) = \mathtt{E}(\beta_1^\dagger, \beta_2^\dagger, \ldots, \beta_p^\dagger)$.

The server sends the resulting ciphertexts $\mathtt{E}(\boldsymbol{\beta}^*)$, $\mathtt{E}(\boldsymbol{\beta}^\dagger)$, and $\mathtt{E}(|\mathbf{A}|)$ to the authority who has the secret key of the underlying HE scheme. Afterwards, the authority decrypts the values and computes the test statistics by using the Wald z-test, which are defined by the coefficient estimates divided by the standard errors of the parameters:

$$\beta_j/\sqrt{\mathtt{var}_j} = \beta_j^*/\sqrt{|\mathbf{A}| \cdot \beta_j^\dagger}$$

for all $j \in [p]$. In the end, the p-values can be obtained by the textbook definition such as $2 \cdot (\mathtt{pnorm}(-\mathtt{abs}(\beta_j/\sqrt{\mathtt{var}_j})))$.

5 Results

In this section, we explain how to set the parameters and report the performance of our regression algorithms.

5.1 Dataset Description

The dataset provided by the iDASH competition organizers consists of 245 samples, partitioned into two groups by the condition of high cholesterol, 137 under control group and 108 under disease group. Each sample contains a binary phenotype along with 10643 SNPs and 3 covariates (age, weight, and height). This training data was extracted from Personal Genome Project [1]. The organizers changed the input size in terms of SNPs and cohort size and threshold of significance to test the scalability of submitted solutions.

We may assume that the imputation and normalization are done in the clear prior to encryption. More precisely, we impute the missing covariate values with the sample mean of the observed covariates in the clear. We also center the covariates matrix $\mathbf{X}$ by subtracting the minimum from each column and dividing by a quantity proportional to the range:

$$x'_{ij} = \frac{x_{ij} - \tilde{m}_j}{\hat{m}_j - \tilde{m}_j},$$

where $\hat{m}_j = \mathtt{max}_i\{x_{ij}\}$ and $\tilde{m}_j = \mathtt{min}_i\{x_{ij}\}$.

5.2 Parameters Settings

We explain how to choose the parameter sets for building secure semi-parallel logistic regression model. We begin with a parameter L which determines the largest bitsize of a fresh ciphertext modulus. Since the plaintext space is a vector space of real numbers, we multiply a scale factor of p to plaintexts before encryption. It is a common practice to perform the rescaling operation by a factor of p on ciphertexts after each (constant) multiplication in order to preserve the precision of the plaintexts. This means that a ciphertext modulus is reduced by $\log p$ bits after each multiplication or we can say that a multiplication operation consumes one level.

Suppose that we start with $\mathbf{v}^{(0)} = \boldsymbol{\beta}_{\mathbf{X}}^{(0)} = \mathbf{0} \in \mathbb{R}^k$ and the input ciphertext $\mathtt{E}(\mathbf{y}^T\mathbf{X})$ is at level L. It follows from the parameter analysis of [19] that the ciphertext level of $\mathtt{E}(\boldsymbol{\beta}_{\mathbf{x}})$ after the evaluation of Nesterov's accelerated GD is $L - (4 \cdot (\textsc{IterNum} - 1) + 1)$ where $\textsc{IterNum}$ denotes the number of iterations of the GD algorithm. Similarly, we expect each of Steps 1 and 2 to consume two levels for computing the ciphertexts $\mathtt{E}(\hat{\mathbf{y}})$ and $\mathtt{E}(\mathbf{p})$. This means that $\mathtt{E}(\mathbf{p})$ is at level $L - (4 \cdot \textsc{IterNum} + 1)$; so we get

$$\begin{aligned} \mathtt{lvl}(\mathtt{E}(\mathbf{w})) &= L - (4 \cdot \textsc{IterNum} + 2), \\ \mathtt{lvl}(\mathtt{E}(\mathbf{Wz})) &= L - (4 \cdot \textsc{IterNum} + 3). \end{aligned}$$

We now consider the replication procedure in Step 3. Although the input vector $\mathbf{w} = (w_i)_{i=1}^n$ is fully packed into a single ciphertext (i.e., the length of the corresponding plaintext vector is N_2), it suffices to produce n number of ciphertexts, each of which represents an entry w_i across the entire array. As presented in Section 4.2 of [16], the replication procedure consists of two phases of computation. The first phase is to partition the entries in the input vector into size-2^s blocks and construct $n/2^s$ number of vectors consisting of the entries in the i-th block with replicated $N_2/2^s$ times. We use a simple replication operation $n/2^s$ times that applies multiplicative masking to extract the entry and then perform the `AllSum` operation to replicate them as in Step 1; its depth is just a single constant multiplication. The second phase is to recursively apply replication operations in a binary tree manner, such that in each stage we double the number of vectors while halving the number of distinct values in each vector; its depth is s constant multiplications. In total, we expect to consume $(s + 1)$ levels during the replication procedure; so we get

$$\begin{aligned} \mathtt{lvl}(\mathtt{E}(w_i)) &= L - (4 \cdot \textsc{IterNum} + s + 3), \\ \mathtt{lvl}(\mathtt{E}(w_i z_i)) &= L - (4 \cdot \textsc{IterNum} + s + 4). \end{aligned}$$

Later, Step 4 consumes one level from the level $\mathtt{lvl}(\mathtt{E}(w_i))$ for multiplication; so we have

$$\mathtt{lvl}(\mathtt{E}(\mathbf{B}_\ell)) = \mathtt{lvl}(\mathtt{E}(c_1, \ldots, c_p)) = L - (4 \cdot \textsc{IterNum} + s + 4).$$

In Step 5, it consumes one more level from the computation of $\mathtt{E}(w_i z_i)$; so we get

$$\mathtt{lvl}(\mathtt{E}(\mathbf{s}_1^T \mathbf{Wz}, \ldots, \mathbf{s}_p^T \mathbf{Wz})) = L - (4 \cdot \textsc{IterNum} + s + 5).$$

On the other hand, Step 6 requires one level of multiplication for the evaluation of the update formula (9); so we know

$$\mathtt{lvl}(\mathtt{E}(\mathbf{X} \odot \mathbf{Wz})) = \mathtt{lvl}(\mathtt{E}(\mathbf{Wz})) - 1 = L - (4 \cdot \textsc{IterNum} + 4).$$

As discussed above, the output ciphertexts $\mathtt{E}(\mathbf{x}_\ell^T \mathbf{W}\mathbf{z})$ consume $(s'+1)$ levels during the replication procedure such that $2^{s'}$ is the unit block size of the first step of the replication procedure; so we have

$$\mathtt{E}(\mathbf{x}_\ell^T \mathbf{W}\mathbf{z}) = \mathtt{lvl}(\mathtt{E}(\mathbf{X} \odot \mathbf{W}\mathbf{z})) - (s'+1) = L - (4 \cdot \textsc{IterNum} + s' + 5).$$

In Step 7, it requires one and $\log(k-1)$ levels of multiplications for the evaluation of the update formulas (11) and (12), respectively. If we let $\ell' = \mathtt{max}\{\mathtt{lvl}(\mathtt{E}(\mathbf{w})), \mathtt{lvl}(\mathtt{E}(\Sigma_{r,s,t}))\}$, then we have

$$\begin{aligned} \mathtt{lvl}(\mathtt{E}(\mathtt{adj}(\mathbf{A})_{rs}) &= \ell' - (1 + \log(k-1)), \\ \mathtt{lvl}(\mathtt{E}(|\mathbf{A}|)) &= \ell' - (2 + \log(k-1)). \end{aligned}$$

It follows from the update formulas (13) and (14) in Step 8 that it is enough to set as

$$\mathtt{lvl}(\mathtt{E}(\mathtt{adj}(\mathbf{A})_{rs})) = \mathtt{lvl}(\mathtt{E}(\mathbf{B}_\ell)) = 3$$

for obtaining the correct results.

In our implementation, we set $\textsc{IterNum} = 2$, $L = 19$, $s = 4$, and $s' = 0$. The encryption levels of data are set as follows:

- $\mathtt{lvl}(\mathtt{E}(\mathbf{y}^T \mathbf{X})) = L = 19$,
- $\mathtt{lvl}(\mathtt{E}(\mathbf{X})) = \mathtt{lvl}(\mathtt{E}(\boldsymbol{\beta}_\mathbf{X})) = 14$, from (5)
- $\mathtt{lvl}(\mathtt{E}(\mathbf{y})) = \mathtt{lvl}(\mathtt{E}(\mathbf{p})) = 10$, from (6),
- $\mathtt{lvl}(\mathtt{E}(x_{i\ell}\mathbf{S}_i)) = \mathtt{lvl}(\mathtt{E}(w_i)) = 4$, from (7),
- $\mathtt{lvl}(\mathtt{E}(\Sigma_{r,s,t})) = \mathtt{lvl}(\mathtt{E}(\mathtt{adj}(\mathbf{A}))) + 3 = 6$.

We use $\log p_0 \approx 60$, $\log q_0 \approx 51$, and $\log q_i \approx 43$ for $i = 1, \ldots, L$. Therefore, we derive a lower bound of the bit size of the largest RLWE modulus Q as

$$\begin{aligned} \log Q &= \log q_0 + (L-1) \cdot \log q_i + \log p_0 \\ &\approx 51 + 43 \cdot 18 + 60 = 885. \end{aligned}$$

We choose the secret key from the ternary distribution, which means to select uniformly at random from $\{-1, 0, 1\}$. The error is sampled from the discrete Gaussian distribution of standard deviation $\mathsf{stdev} = 3.2$. We follow the recommended parameters from the standardization workshop paper [5], thus providing at least 128-bits security level of our parameters. We summarize the parameters of our implementation in Table 1.

Table 1: Parameters

$\log N$	L	$\log p$	$\log q_0$	$\log p_0$	$\log Q$
15	19	43	51	60	885

5.3 Optimization Techniques

The standard method of homomorphic multiplication consists of two steps: raw multiplication and key-switching. The first step computes the product of two ciphertexts $\mathsf{ct}(Y) = c_0 + c_1 Y$ and $\mathsf{ct}'(Y) = c_0' + c_1' Y$ (as done in [4]), and returns a quadratic polynomial, called *extended ciphertext*, $\mathsf{ct}_{\mathsf{mult}} = c_0 c_0' + (c_0 c_1' + c_0' c_1)Y + c_1 c_1' Y^2$. This ciphertext can be viewed as an encryption of the product of plaintexts with the extended secret $(1, s, s^2)$. And then, the key-switching procedure transforms it into a normal (linear) ciphertext encrypting the same message with the secret key $(1, s)$.

We observe that the second step is much more expensive than the first one since it includes an evaluation of NTT (Fourier transformation over the modulo space), and that a simple arithmetic (e.g. linear operation) is allowed between extended ciphertexts. To reduce the complexity, we adapt the technique called *lazy key-switching*, which performs some arithmetic over extended ciphertexts instead of running the second step right after a raw multiplication. We get a normal ciphertext by performing only one key-switching operation after evaluating linear circuits over the extended ciphertexts. It can reduce the number of required key-switching algorithms as well as the total computational cost. For instance, if we add many terms after raw multiplications in the right hand side of the update (7) and apply key-switching to the output ciphertext, this takes only one key-switching rather than n. This lazy key-switching can also be applied to the update formulas (8), (13), and (14).

5.4 Performance Results

We present our implementation results using the proposed techniques. All the experiments were performed on an iMac with an Intel Core i5 running with 4 cores rated at 3.8 GHz.

In Table 2, we evaluated our model's performance based on the average running time and the memory usages in the key generation, encryption, evaluation, and decryption procedures.

Table 2: Timing results for iDASH dataset

Stage	Timing	Memory usage
Key Generation	5.071 s	3.583 GB
Encryption	5.497 s	6.669 GB
Training with covariates	7.093 s	7.176 GB
Training with SNPs	33.737 s	9.786 GB
Total evaluation	**40.830 s**	-
Decryption	0.023 s	-

We achieved very high level of accuracy in the final output (after decryption). The type-I (false positive) and type-II (false negative) errors of the output of our solution are very small when comparing to both the semi-parallel model and the gold standard (full logistic regression) model with respect to various p-value cut-off thresholds. See Figures 1 and 2 for comparisons against these two plain models with a cut-off of 10^{-5}. To better compare the estimated p-values (above or below certain cut-offs) on the encrypted model against the plaintext one (semi-parallel GWAS), we measured F_1-scores on the p-values obtained from our solution against the two plain models. The resulting F_1-scores are very close to 1 across all cases with different cut-offs (10^{-2} to 10^{-5}), which are shown in Table 3.

Fig. 1: Comparison with the semi-parallel model (p-value cutoff: 10^{-5})

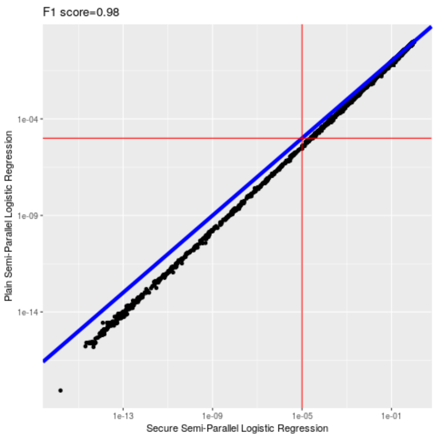

Fig. 2: Comparison with the gold standard model (p-value cutoff: 10^{-5})

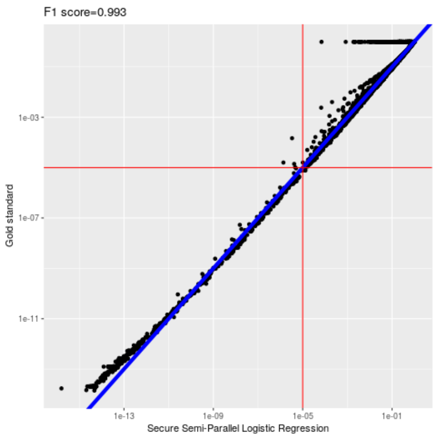

Table 3: F_1-Scores on different models

Cut-off	v.s. Plain semi-parallel model	v.s. Gold standard
10^{-2}	0.9786	0.9831
10^{-3}	0.9710	0.9855
10^{-4}	0.9676	0.9839
10^{-5}	0.9804	0.9934

We also conducted the DeLong's test [13, 25] to validate our solution against the semi-parallel model. Specifically, we drawn at uniformly random about 10% of the total SNP test data and transformed the corresponding p-values to 0-1 labels according to the cut-off threshold; then we constructed the ROC (Receiver Operating Characteristic) curves for these labels and performed the DeLong's test to compare the AUCs (Area Under the Curve) of these curves. Such test was repeated 10 times to obtain the mean and the standard deviation of the p-values of the test. The results are shown in Table 4.

Table 4: DeLong's Test for AUCs of our solution against the plain semi-parallel model

Cut-off	Mean and stdev of the test results
10^{-2}	0.4038 ± 0.3001
10^{-3}	0.5357 ± 0.2704
10^{-4}	0.6404 ± 0.2638
10^{-5}	0.8959 ± 0.2195

6 Discussion and Conclusion

One constraint in our approach is that the matrix inverse can be computed in an efficient way when the input dimension is small. In modern GWAS, it is common to include covariates to account for such factors as gender, age, other clinical variables and population structure. A significant challenge in performing efficient secure GWAS on this generalized model is to handle matrix inversion for larger ones.

In this paper, we showed a state-of-the-art performance of secure logistic regression model training for GWAS. We have demonstrated the feasibility and scalability of our model in speed and memory consumption. We expect that the performance can be improved if the underlying HE scheme is rewritten with optimized code.

References

1. Personal genome project. `https://www.personalgenomes.org/us`.
2. NIH genomic data sharing. `https://osp.od.nih.gov/scientific-sharing/genomic-data-sharing/`, 2015.
3. C. Bonte, E. Makri, A. Ardeshirdavani, J. Simm, Y. Moreau, and F. Vercauteren. Towards practical privacy-preserving genome-wide association study. *BMC bioinformatics*, 19(1):537, 2018.
4. Z. Brakerski and V. Vaikuntanathan. Fully homomorphic encryption from Ring-LWE and security for key dependent messages. In *Advances in Cryptology–CRYPTO 2011*, pages 505–524. 2011.
5. M. Chase, H. Chen, J. Ding, S. Goldwasser, S. Gorbunov, J. Hoffstein, K. Lauter, S. Lokam, D. Moody, T. Morrison, A. Sahai, and V. Vaikuntanathan. Security of homomorphic encryption. Technical report, HomomorphicEncryption.org, Redmond WA, USA, July 2017.
6. H. Chen, R. Gilad-Bachrach, K. Han, Z. Huang, A. Jalali, K. Laine, and K. Lauter. Logistic regression over encrypted data from fully homomorphic encryption. *BMC medical genomics*, 11(4):81, 2018.
7. J. H. Cheon, K. Han, A. Kim, M. Kim, and Y. Song. A full RNS variant of approximate homomorphic encryption. In *International Conference on Selected Areas in Cryptography*. Springer, 2018.
8. J. H. Cheon, A. Kim, M. Kim, and Y. Song. Homomorphic encryption for arithmetic of approximate numbers. In *Advances in Cryptology–ASIACRYPT 2017*, pages 409–437. Springer, 2017.
9. J. H. Cheon, M. Kim, and M. Kim. Search-and-compute on encrypted data. In *International Conference on Financial Cryptography and Data Security*, pages 142–159. Springer, 2015.
10. J. H. Cheon, M. Kim, and M. Kim. Optimized search-and-compute circuits and their application to query evaluation on encrypted data. *IEEE Transactions on Information Forensics and Security*, 11(1):188–199, 2016.
11. J. H. Cheon, M. Kim, and K. Lauter. Homomorphic computation of edit distance. In *International Conference on Financial Cryptography and Data Security*, pages 194–212. Springer, 2015.
12. D. R. Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 215–242, 1958.
13. E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845, 1988.
14. D. A. Freedman. Statistical models: theory and practice, 2009.
15. C. Gentry, S. Halevi, and N. P. Smart. Homomorphic evaluation of the AES circuit. In *Advances in Cryptology–CRYPTO 2012*, pages 850–867. 2012.
16. S. Halevi and V. Shoup. Algorithms in HElib. In *International Cryptology Conference*, pages 554–571. Springer, 2014.
17. C. W. Hug and P. Szolovits. ICU acuity: real-time models versus daily models. In *AMIA annual symposium proceedings*, volume 2009, page 260. American Medical Informatics Association, 2009.
18. P. S. Kamath and W. R. Kim. The model for end-stage liver disease (meld). *Hepatology*, 45(3):797–805, 2007.
19. A. Kim, Y. Song, M. Kim, K. Lee, and J. H. Cheon. Logistic regression model training based on the approximate homomorphic encryption. *BMC medical genomics*, 11(4):83, 2018.
20. M. Kim and K. Lauter. Private genome analysis through homomorphic encryption. *BMC medical informatics and decision making*, 15(Suppl 5):S3, 2015.
21. M. Kim, Y. Song, and J. H. Cheon. Secure searching of biomarkers through hybrid homomorphic encryption scheme. *BMC medical genomics*, 10(2):42, 2017.

22. M. Kim, Y. Song, S. Wang, Y. Xia, and X. Jiang. Secure logistic regression based on homomorphic encryption: design and evaluation. *JMIR medical informatics*, 6(2), 2018.
23. Y. Nesterov. A method of solving a convex programming problem with convergence rate o (1/k2). In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983.
24. C. Robert. Machine learning, a probabilistic perspective, 2014.
25. X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12(1):77, Mar 2011.
26. A. A. Shabalin. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, 2012.
27. K. Sikorska, E. Lesaffre, P. F. Groenen, and P. H. Eilers. GWAS on your notebook: fast semi-parallel linear and logistic regression for genome-wide association studies. *BMC bioinformatics*, 14(1):166, 2013.
28. J. J. Trinckes and Jr. The definitive guide to complying with the HIPAA/HITECH privacy and security rules, 3 Dec. 2012.
29. J. Truett, J. Cornfield, and W. Kannel. A multivariate analysis of the risk of coronary heart disease in framingham. *Journal of chronic diseases*, 20(7):511–524, 1967.
30. S. H. Walker and D. B. Duncan. Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54(1-2):167–179, 1967.