

# Differentially-Private Logistic Regression for Detecting Multiple-SNP Association in GWAS Databases

Fei Yu<sup>1,\*</sup>, Michal Rybar<sup>2</sup>, Caroline Uhler<sup>2,\*\*</sup>, and Stephen E. Fienberg<sup>1</sup>

<sup>1</sup> Carnegie Mellon University, Pittsburgh, PA 15213, USA  
{feiy, fienberg}@stat.cmu.edu

<sup>2</sup> Institute of Science and Technology Austria, Am Campus 1,  
3400 Klosterneuburg, Austria  
{michal.rybar, caroline.uhler}@ist.ac.at

**Abstract.** Following the publication of an attack on genome-wide association studies (GWAS) data proposed by Homer et al., considerable attention has been given to developing methods for releasing GWAS data in a privacy-preserving way. Here, we develop an end-to-end differentially private method for solving regression problems with convex penalty functions and selecting the penalty parameters by cross-validation. In particular, we focus on penalized logistic regression with elastic-net regularization, a method widely used to in GWAS analyses to identify disease-causing genes. We show how a differentially private procedure for penalized logistic regression with elastic-net regularization can be applied to the analysis of GWAS data and evaluate our method's performance.

**Keywords:** Differential privacy, genome-wide association studies (GWAS), logistic regression, elastic-net, ridge regression, lasso, cross-validation, single nucleotide polymorphism (SNP).

## 1 Introduction

### 1.1 Genetic Data Privacy Background

The goal of a genome-wide association study (GWAS) is to identify genetic variations associated with a disease. Typical GWAS databases contain information on hundreds of thousands of single nucleotide polymorphisms (SNPs) from thousands of individuals. The aim of GWAS is to find associations between SNPs and a certain phenotype, such as a disease. A particular phenotype is usually the result of complex relationships between multiple SNPs, making GWAS a very high-dimensional problem.

---

\* This research was partially supported by NSF Awards EMSW21-RTG and BCS-0941518 to the Department of Statistics at Carnegie Mellon University.

\*\* This research was partially performed while visiting the Simons Institute for the Theory of Computing.

Recently, penalized regression approaches have been applied to GWAS to overcome the challenges caused by the high-dimensional nature of the data. A popular approach consists of a two-step procedure. In the first step, all SNPs are screened and a subset is selected based on a simple  $\chi^2$ -test for association between each single SNP and the phenotype. In the second step, the selected subset of SNPs is tested for multiple-SNP association using penalized logistic regression. Elastic-net regularization, which imposes a combination of  $\ell_1$  and ridge penalties, has been shown to be a competitive method for GWAS (e.g. [1, 2]).

For many years, researchers believed that releasing statistics of SNPs aggregated from thousands of individuals would not compromise the participants' privacy. Such a belief came under challenge with the publication of an attack proposed by Homer et al. [3]. This publication drew widespread attention. As a consequence, NIH removed all aggregate SNP data from open-access databases [4, 5] and instituted an elaborate approval process for gaining access to aggregate genetic data. This NIH action in turn spurred interest in the development of methods for confidentiality protection of GWAS databases.

## 1.2 Differentially Private Methods for Solving Regression Problems

The approach of differential privacy, introduced by the cryptographic community (e.g. Dwork et al. [6]), provides privacy guarantees that protect GWAS databases against arbitrary external information. Building on such notion, Uhler et al. [7], Johnson and Shmatikov [8], and Yu et al. [9] proposed new methods for selecting a subset of SNPs in a differentially-private manner. These approaches enable us to perform the first step in the two-step procedure for identifying the relevant SNPs in a GWAS without compromising the study participants' privacy. The second step of the two-step procedure would involve performing penalized logistic regression with elastic-net regularization ( $l_1$  and  $l_2$  penalties) on the selected subset of SNPs in a differentially private manner. Kifer et al. [10] proposed an objective function perturbation mechanism that releases the coefficients of a convex risk minimization problem with convex penalties and satisfies differential privacy. We can use this method to perform logistic regression with elastic-net regularization in a differentially private way.

The performance of penalized logistic regression approaches depends heavily on the choice of regularization parameters. Selection of these regularization parameters is usually done via cross-validation. Chaudhuri and Vinterbo [11] proposed a differentially-private procedure for choosing the regularization parameters based on a stability argument. However, the method proposed by Chaudhuri and Vinterbo [11] only works on differentiable penalty functions, such as the  $\ell_2$  penalty, and it cannot be applied to elastic-net regularization or lasso.

In Section 2, we extend the stability-based method for selecting the regularization parameters developed by Chaudhuri and Vinterbo [11] so that it is applicable to any convex penalty function, including the elastic-net penalty. By combining this new result and the objective function perturbation mechanism proposed by Kifer et al. [10], we are able to carry out a privacy-preserving penal-

ized logistic regression analysis. In Section 3, we demonstrate how to implement the full objective function perturbation mechanism with cross-validation based on the results by Chaudhuri et al. [12] and Kifer et al. [10]. In particular, we provide the exact form of the random noise used in the objective function perturbation mechanism. Furthermore, we show that, under a slightly stronger condition, we can perturb the objective function by an alternative form of noise—the multivariate Laplace noise—and thereby obtain more accurate results. In Section 4, we show how to apply our results to develop an end-to-end differentially private penalized logistic regression method with elastic-net penalty and cross-validation for the selection of the penalty parameters. Finally, in Section 5, we demonstrate how well this end-to-end differentially private method performs on a GWAS data set.

## 2 Differentially-Private Penalized Regression

We start by reviewing the concept of differential privacy. Let  $\mathcal{D}$  denote the set of all data sets. Let  $D, D' \in \mathcal{D}$  denote two data sets that differ in one individual only. We denote this by  $D \sim D'$ .

**Definition 1 (differential privacy).** *A randomized mechanism  $\mathcal{K}$  is  $\epsilon$ -differentially private if, for all  $D \sim D'$  and for any measurable set  $S \subset \mathbb{R}$ ,*

$$\frac{\mathbb{P}(\mathcal{K}(D) \in S)}{\mathbb{P}(\mathcal{K}(D') \in S)} \leq e^\epsilon.$$

$\mathcal{K}$  is  $(\epsilon, \delta)$ -differentially private if, for all  $D \sim D'$  and for any measurable set  $S \subset \mathbb{R}$ ,

$$\mathbb{P}(\mathcal{K}(D) \in S) \leq e^\epsilon \mathbb{P}(\mathcal{K}(D') \in S) + \delta.$$

Let  $l : \mathbb{R}^s \times \mathcal{D} \rightarrow \mathbb{R}$  denote the loss function,  $r : \mathbb{R}^s \rightarrow \mathbb{R}$  a regularization function, and  $h : \mathbb{R}^s \times \mathcal{D} \rightarrow \mathbb{R}$  the validation function. Let  $T \in \mathcal{D}^n$  be a training data set of size  $n$  drawn from  $\mathcal{D}$  and  $V \in \mathcal{D}^m$  a validation data set of size  $m$  also drawn from  $\mathcal{D}$ . Let  $b \in \mathbb{R}^s$  denote the noise used to perturb the regularized loss function. Then we denote by  $\mathcal{T}(\lambda, \epsilon; T, l, r, b)$  the differentially private procedure to produce parameter estimates from the training data  $T$  given the regularization parameter  $\lambda$ , the privacy budget  $\epsilon > 0$ , the loss function  $l$ , the regularization function  $r$ , and the random noise  $b$ . We score a vector of regression coefficients resulting from the random procedure  $\mathcal{T}(\lambda, \epsilon; T, l, r, b)$  using the validation data  $V$  and the validation score function  $q(\theta, V) = -\frac{1}{m} \sum_{d \in V} h(\theta; d)$ .

**Definition 2 ( $(\beta_1, \beta_2, \delta)$ -stability. Chaudhuri and Vinterbo [11]).** *A validation score function  $q$  is said to be  $(\beta_1, \beta_2, \delta)$ -stable with respect to a training procedure  $\mathcal{T}$ , the candidate regularization parameters  $\Lambda$ , and the privacy budget  $\epsilon$ , if there exists  $E \subset \mathbb{R}^s$  such that  $\mathbb{P}(b \in E) \geq 1 - \delta$ , and when  $b \in E$ , the following conditions hold:*

1. **Training stability:** for all  $\lambda \in \Lambda$ , for all validation data sets  $V \in \mathcal{D}^m$ , and all training dataset  $T, T' \in \mathcal{D}^n$  with  $T \sim T'$ ,

$$|q(\mathcal{T}(\lambda, \epsilon; T, l, r, b), V) - q(\mathcal{T}(\lambda, \epsilon; T', l, r, b), V)| \leq \frac{\beta_1}{n}.$$

2. **Validation stability:** for all  $\lambda \in \Lambda$ , for all training data sets  $T \in \mathcal{D}^n$ , and all validation data sets  $V, V' \in \mathcal{D}^m$  with  $V \sim V'$ ,

$$|q(\mathcal{T}(\lambda, \epsilon; T, l, r, b), V) - q(\mathcal{T}(\lambda, \epsilon; T, l, r, b), V')| \leq \frac{\beta_2}{m}.$$

Chaudhuri and Vinterbo [11] gave conditions under which a validation score function is  $(\beta_1, \beta_2, \delta)$ -stable when the regularization function is differentiable and showed that as long as the validation score function  $q$  is  $(\beta_1, \beta_2, \delta)$ -stable for some  $\beta_1, \beta_2, \delta > 0$  with respect to the procedure  $\mathcal{T}$ , candidate regularization parameters  $\Lambda$ , and privacy budget  $\epsilon$ , we can choose the best regularization parameter in a differentially private manner using Algorithm 1 and Algorithm 2 in Chaudhuri and Vinterbo [11]. In Theorem 3, we specify the conditions under which a validation score function is  $(\beta_1, \beta_2, \delta)$ -stable for a general convex regularization function.

In the following, we combine the regularization function and the regularization parameters to form a vector of candidate regularization functions  $r = (r_1, \dots, r_t)$ . Then, selecting the regularization parameters is equivalent to selecting a linear combination of  $r_i$ 's in  $r$ .

**Theorem 3.** Let  $r = (r_1, \dots, r_t)$  be a vector of convex regularization functions with  $r_i : \mathbb{R}^s \rightarrow \mathbb{R}$  that are minimized at 0. Let  $\Lambda = \{\lambda_1, \dots, \lambda_k\}$  be a collection of regularization vectors, where  $\lambda_i$  is a  $t$ -dimensional vector of 0's and 1's. We denote by  $c_{min} := \sup_c \{\forall \lambda \in \Lambda, \lambda^T r \text{ is } c\text{-strong convex}\}$ . Let  $h(\theta; d)$  be a validation score that is non-negative and  $\kappa$ -Lipschitz in  $\theta$ . We denote  $\max_{d \in \mathcal{D}, \theta \in \mathbb{R}^s} h(\theta; d)$  by  $h^*$ . In addition, let  $l(\theta; d)$  be a convex loss function that is  $\gamma$ -Lipschitz in  $\theta$ . Finally, let  $\xi \in \mathbb{R}$  such that  $\mathbb{P}(\|b\|_2 > \xi) \leq \delta/k$  for some  $\delta \in (0, 1)$ . Then the validation score  $q(\theta, V) = -\frac{1}{m} \sum_{d \in V} h(\theta; d)$  is  $(\beta_1, \beta_2, \delta/k)$ -stable with respect to  $\mathcal{T}$ ,  $\epsilon$  and  $\Lambda$ , where

$$\mathcal{T}(\lambda, \epsilon; T, l, r, b) := \arg \min_{\theta} L(\theta; \lambda, \epsilon),$$

with

$$L(\theta; \lambda, \epsilon) = \frac{1}{n} \sum_{d \in T} l(\theta; d) + \lambda^T r(\theta) + \frac{\max\{0, c^* - c_{min}\}}{2} \|\theta\|_2^2 + \frac{\varphi}{\epsilon n} b^T \theta,$$

$$\beta_1 = \frac{2\gamma\kappa}{\max\{c^*, c_{min}\}}, \quad \beta_2 = \min \left\{ h^*, \frac{\kappa}{\max\{c^*, c_{min}\}} \left( \gamma + \frac{\varphi\xi}{\epsilon n} \right) \right\}.$$

*Proof.* See A.1. □

Note that choosing  $r(\theta) = \left( \frac{\lambda_1}{2} \|\theta\|_2^2, \dots, \frac{\lambda_k}{2} \|\theta\|_2^2 \right)$ , with  $\Lambda = \{e_1, \dots, e_k\}$ , where  $e_i$  is a  $k$ -dimensional vector that is 1 in the  $i$ th entry and 0 everywhere

else, results in Theorem 4 in Chaudhuri and Vinterbo [11]. Thus, Theorem 3 generalizes Theorem 4 in Chaudhuri and Vinterbo [11].

The term  $\frac{\max\{0, c^* - c_{\min}\}}{2} \|\theta\|_2^2$  in Theorem 3 ensures that  $L(\theta; \lambda, \epsilon)$  is at least  $c^*$ -strongly convex. This is an essential condition for ensuring that our objective function perturbation algorithm (Algorithm 1) is differentially private. The value of  $\xi$  in Theorem 3 depends on the distribution of the perturbation noise  $b$ . In Section 3, we analyze two different distributions for the perturbation noise.

### 3 Distributions for the Perturbation Noise

Chaudhuri et al. [12] and Kifer et al. [10] showed that using perturbation noise  $B_2$  with density function

$$f_{B_2}(b) \propto \exp\left(-\frac{\|b\|_2}{2}\right)$$

in the procedure  $\mathcal{T}(\lambda, \epsilon; T, l, r, B_2)$  produces  $\epsilon$ -differentially private parameter estimates. In this section, we describe an efficient method for generating such perturbation noise. Furthermore, we show that under slightly stronger conditions the procedure  $\mathcal{T}(\lambda, \epsilon; T, l, r, B_1)$  is differentially private when we use perturbation noise  $B_1$  with density function

$$f_{B_1}(b) \propto \exp\left(-\frac{\|b\|_1}{2}\right),$$

which is simpler to generate than perturbation noise of the form  $B_2$ .

**Proposition 4.** *The random variable  $X = \frac{W}{\|W\|_2} Y$ , where  $W \sim \mathcal{N}(0, I_s)$  and  $Y \sim \chi^2(2s)$ , has density function  $f_X(x) \propto \exp\left(-\frac{\|x\|_2}{2}\right)$ .*

*Proof.* See Appendix A.3. □

This result shows that  $B_2 \sim \frac{W_s}{\|W_s\|_2} Y_{2s}$ , with  $W_s \sim \mathcal{N}(0, I_s)$  and  $Y_{2s} \sim \chi^2(2s)$ . On the other hand,  $B_1$  can be viewed as the joint distribution of  $s$  independent Laplace random variables with mean = 0 and scale = 2. In order to specify the stability parameter  $\beta_2$  in Theorem 3, we need to find  $\xi \in \mathbb{R}$  such that  $P(\|b\|_2 \geq \xi) \leq \delta/k$ . The following propositions enable us to find  $\xi$  for the perturbation noise  $B_1$  and  $B_2$ .

**Proposition 5.**  $\mathbb{P}(\|B_1\|_1 \geq 2s \log(sk/\delta)) \leq \delta/k$ .

*Proof.* See Lemma 17 in Chaudhuri and Vinterbo [11]. □

**Proposition 6.**  $\mathbb{P}\left(\|B_2\|_2 \geq \left(\sqrt{s} + \sqrt{\log(k/\delta)}\right)^2 + \log(k/\delta)\right) \leq \delta/k$ .

*Proof.* Note that  $\|B_2\|_2 = \left\| \frac{W_s}{\|W_s\|_2} Y_{2s} \right\|_2 = Y_{2s}$ , where  $Y_{2s} \sim \chi^2(2s)$ . The proof is completed by invoking Lemma 1 in Laurent and Massart [13]. □

Because  $P(\|B_1\|_1 \geq \xi) \geq P(\|B_1\|_2 \geq \xi)$ , Proposition 5 and Proposition 6 enable us to find  $\xi \in \mathbb{R}$  such that  $P(\|b\|_2 \geq \xi) \leq \delta/k$ . When the density function of  $b$  is  $f(b) \propto \exp\left(\frac{\|b\|_1}{2}\right)$ , then by Proposition 5,  $\xi = 2s \log(sk/\delta)$ . When the density function of  $b$  is  $f(b) \propto \exp\left(\frac{\|b\|_2}{2}\right)$ , then by Proposition 6,  $\xi = \left(\sqrt{s} + \sqrt{\log(k/\delta)}\right)^2 + \log(k/\delta)$ .

Algorithm 1 below is a reformulation of Algorithm 1 in Kifer et al. [10], i.e., the differentially private objective function optimization algorithm, and it incorporates the alternative perturbation noise. The objective function is formulated in such a way that it is compatible with the regularization parameter selection procedure described in Theorem 3.

---

**Algorithm 1.** Generalized Objective Perturbation Mechanism

---

**Input:** Dataset  $D = \{d_1, \dots, d_n\}$ ; a convex domain  $\Theta \subset \mathbb{R}^s$ ; privacy parameter  $\epsilon$ ;  $\lambda$ -strongly convex regularizer  $r$ ; convex loss function  $l(\theta; d)$  with rank-1 continuous Hessian  $\nabla^2 l(\theta; d)$ , an upper bound  $c$  on the maximal singular value of  $\nabla^2 l(\theta; d)$  and upper bounds  $\kappa_j$  on  $\|\nabla l(\theta; d)\|_j$  for  $j \in \{1, 2\}$  that hold for all  $d \in D$  and all  $\theta \in \Theta$ . It is also required that  $\varphi \geq 2\kappa_j$  and  $\lambda \geq \frac{c}{n(e^{\epsilon/4} - 1)}$ .

**Output:** A differentially-private parameter vector  $\theta^*$ .

1. Sample  $b \in \mathbb{R}^s$  according to noise distribution  $B_j$ ,  $j \in \{1, 2\}$ .
2. **return**  $\theta^* = \arg \min_{\theta} L(\theta; D, \lambda, b)$ , where

$$L(\theta; D, \lambda, b) = \frac{1}{n} \sum_{d \in D} l(\theta; d) + r(\theta) + \frac{\varphi}{\epsilon n} b^T \theta.$$


---

**Theorem 7.** *Algorithm 1 is  $\epsilon$ -differentially private.*

*Proof.* See A.2. □

**3.1 Comparison of the Performance of Algorithm 1 under Different Noise Distributions**

Note that we can always upper bound  $\|\nabla l(\theta; d)\|_2$  by  $\|\nabla l(\theta; d)\|_1$  and hence  $\kappa_2 \leq \kappa_1$  in Algorithm 1. However, as we show in this section, results from Algorithm 1 are more accurate when sampling noise from  $B_1$  compared to  $B_2$ . To compare the performance of Algorithm 1 under noise sampled from  $B_1$  and  $B_2$ , we follow the algorithm performance analysis in Chaudhuri et al. [12] and analyze  $\mathbb{P}(J(\theta_b) - J(\theta^*) > c)$ , where

$$J(\theta) = \frac{1}{n} \sum_{d \in D} l(\theta; d) + r(\theta)$$

with  $l$  and  $r$  as defined in Algorithm 1,  $\theta^* = \arg \min_{\theta} J(\theta)$ , and  $\theta_b = \arg \min_{\theta} [J(\theta) + \frac{\varphi}{\epsilon n} b^T \theta] = \arg \min_{\theta} L(\theta; b)$ . That is,  $J(\theta_b) - J(\theta^*)$  measures how much the objective function deviates from the optimum due to the added noise. Given random noise  $b \in \mathbb{R}^s$ ,  $J(\theta_b) + \frac{\varphi}{\epsilon n} b^T \theta_b \leq J(\theta^*) + \frac{\varphi}{\epsilon n} b^T \theta^*$ . Hence,  $J(\theta_b) - J(\theta^*) \leq \frac{\varphi}{\epsilon n} b^T (\theta^* - \theta_b) \leq \frac{\varphi}{\epsilon n} \|b\|_2 \|\theta^* - \theta_b\|_2$ . Let  $E$  denote the event that  $\{\|b\|_2 \leq \xi\}$ , where  $\xi = \frac{\epsilon n}{\varphi} \sqrt{\lambda c}$ . When  $E$  holds, then  $\frac{\varphi}{\epsilon n} b^T \theta$  is  $\frac{\varphi \xi}{\epsilon n}$ -Lipschitz. Hence, with  $G(\theta) = J(\theta)$   $\lambda$ -strongly convex,  $g_1(\theta) = \frac{\varphi}{\epsilon n} b^T \theta$  and  $g_2 = 0$ , we can invoke Lemma 8 to obtain  $\|\theta^* - \theta_b\|_2 \leq \frac{\varphi \xi}{\lambda \epsilon n}$ . Therefore, when  $E$  holds, then

$$J(\theta_b) - J(\theta^*) \leq \frac{\varphi}{\epsilon n} \|b\|_2 \|\theta^* - \theta_b\|_2 \leq \frac{\varphi}{\epsilon n} \xi \frac{\varphi \xi}{\lambda \epsilon n} = c.$$

Thus  $\mathbb{P}(J(\theta_b) - J(\theta^*) > c) \leq 1 - \mathbb{P}(E) = \mathbb{P}(\|b\|_2 > \xi)$  when the random noise  $b$  is sampled from  $B_1$  or  $B_2$ .  $\|B_1\|_1$  is the sum of  $s$  independent exponential random variables with mean = 2 and thus  $\|B_1\|_1 \sim \text{Gamma}(s, 2)$ . On the other hand,  $\|B_2\|_2 \sim \chi^2(2s)$ . But in fact  $\chi^2(2s) \sim \text{Gamma}(s, 2)$ . Therefore,  $\mathbb{P}(\|B_1\|_2 > \xi) \leq \mathbb{P}(\|B_1\|_1 > \xi) = \mathbb{P}(\|B_2\|_2 > \xi)$ . Thus, sampling the noise from  $B_1$  in Algorithm 1 produces more accurate results.

## 4 Application to Logistic Regression with Elastic-Net Regularization

In this section we show how to apply the results from the previous section to penalized logistic regression. The logistic loss function  $l(\theta; x, y)$  is given by

$$l(\theta; x, y) = \log(1 + \exp(-y \theta^T x)),$$

where  $y \in \{-1, 1\}$ . The first and second derivatives with respect to  $\theta$  are

$$\begin{aligned} \nabla l(\theta; x, y) &= -\frac{1}{1 + \exp(y \theta^T x)} y x \\ \nabla^2 l(\theta; x, y) &= \frac{1}{1 + \exp(-y \theta^T x)} \frac{1}{1 + \exp(y \theta^T x)} x x^T. \end{aligned}$$

It can easily be seen that the logistic loss function satisfies the following properties: (i)  $l(\theta; x, y)$  is convex; (ii)  $\nabla^2 l(\theta; x, y)$  is continuous; and (iii)  $\nabla^2 l(\theta; x, y)$  is a rank-1 matrix.

We denote by  $\|M\|_1$  the nuclear norm of the matrix  $M$  and we choose  $\kappa$  such that  $\|x\|_j \leq \kappa$  for all  $x$ , where  $j \in \{1, 2\}$ . Then

$$\begin{aligned} \|\nabla^2 l(\theta; x, y)\|_1 &\leq \|x x^T\|_1 = \|x\|_2^2 \leq \|x\|_j^2 \leq \kappa^2, \quad \text{for } j \in \{1, 2\}, \\ \|\nabla l(\theta; x, y)\|_j &\leq \|x\|_j \leq \kappa, \end{aligned}$$

Thus we can apply Algorithm 1 to output differentially private coefficients for logistic regression with elastic-net regularization. Moreover, the logistic loss function satisfies the conditions in Theorem 3 because  $l(\theta; x, y)$  is Lipschitz: There exists a parameter  $\theta$  such that

$$|l(\theta_1; x, y) - l(\theta_2; x, y)| \leq \|\nabla l(\theta; x, y)\|_2 \|\theta_1 - \theta_2\|_2 \leq \kappa \|\theta_1 - \theta_2\|_2.$$

Thus we can apply the stability argument in Theorem 3 to select the best regularization parameters in a differentially private way. In Section 5 we show how well this method performs on a GWAS data set.

## 5 Application to GWAS Data

We now evaluate the performance of the proposed method based on a GWAS data set. We analyze a binary phenotype such as a disease. Each SNP can take the values 0, 1, or 2. This represents the number of minor alleles at that site. A large SNP data set is freely available from the HapMap project<sup>1</sup>. It consists of SNP data from 4 populations of 45 to 90 individuals each, but does not contain any phenotypic information about the individuals. HAP-SAMPLE [14] can be used to generate SNP genotypes for cases and controls by resampling from HapMap. This ensures that the simulated data show linkage disequilibrium (i.e., correlations among SNPs) and minor allele frequencies similar to real data.

For our analysis we use the simulations from Malaspinas and Uhler [15]. The simulated data sets consist of 400 cases and 400 controls each with about 10,000 SNPs per individual (SNPs were typed with the Affymetrix CHIP on chromosome 9 and chromosome 13 of the Phase I/II HapMap data). For each data set two SNPs with a given minor allele frequency (MAF) were chosen to be causative. We will analyze the results for minor allele frequency (MAF) = 0.25. The simulations were performed under the multiplicative effects model: Denoting the two causative SNPs by  $X$  and  $Y$  and the disease status by  $D$  (i.e.,  $X, Y \in \{0, 1, 2\}$  and  $D \in \{-1, 1\}$ , where 1 describes the diseased state), then the multiplicative effects model can be defined through the odds of having a disease:

$$\frac{\mathbb{P}(D = 1 \mid X, Y)}{\mathbb{P}(D = -1 \mid X, Y)} = \epsilon \alpha^X \beta^Y \delta^{XY}.$$

This model corresponds to a log-linear model with interaction between the two SNPs. For our simulations we chose  $\epsilon = 0.64$ ,  $\alpha = \beta = 0.91$  and  $\delta = 2.73$ . This results in a sample disease prevalence of 0.5 and effect size of 1, which are typical values for association studies. See Malaspinas and Uhler [15] for more details.

In the first step, we screen all SNPs and select a subset of SNPs with the highest  $\chi^2$ -scores based on a simple  $\chi^2$ -test for association between each single SNP and the phenotype. Various approaches for performing the screening in a differentially private manner were discussed and analyzed in Uhler et al. [7], Johnson and Shmatikov [8], and Yu et al. [9]; We concentrated on the second step and did not employ the differentially private screening approaches in this paper. The second step of the two-step procedure consists of performing penalized logistic regression with elastic-net regularization on the selected subset of SNPs and choosing the best regularization parameters in a differentially private manner. In the following, we analyze the statistical utility of the second step and show how accurately our end-to-end differentially private penalized logistic regression method is able to detect the causative SNPs and their interaction.

<sup>1</sup> <http://hapmap.ncbi.nlm.nih.gov/>

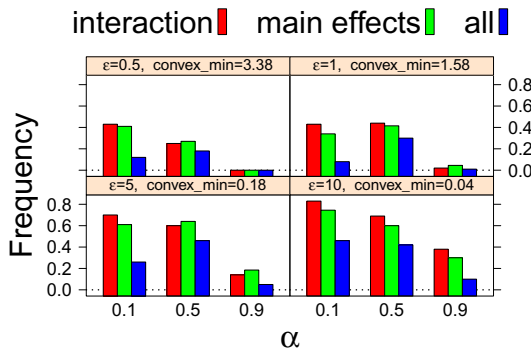


The elastic-net penalty function has the form  $\frac{1}{2}\lambda(1 - \alpha)\ell_2 + \lambda\alpha\ell_1$ , where  $\alpha$  controls the sparsity of the resulting model and  $\lambda$  controls the extent to which the elastic-net penalty affects the loss function. In the simulation, we apply a threshold criterion to the terms in the model so that we exclude from the model the  $i$ th term if its regression coefficient,  $\theta_i$ , satisfies  $|\theta_i|/\max_i\{|\theta_i|\} < r$ , where  $\max_i\{|\theta_i|\}$  is the largest coefficient in absolute value and  $r$  is a thresholding ratio, which we set to 0.01.

In our experiments, we selected  $M = 5$  SNPs with the highest  $\chi^2$ -scores, which include the two causative SNPs, for further analysis. We denote by  $\epsilon$  the privacy budget, by  $\alpha$  the sparsity parameter in the elastic-net penalty function, and by “*convex\_min*” the condition of strong convexity imposed on the objective function (see Theorem 3). Note that *convex\_min* is a function of  $M$  and  $\epsilon$ . For elastic-net with  $\alpha$  fixed, we need the smallest candidate parameter  $\lambda_{min} \geq \text{convex\_min}/(1 - \alpha)$ .

In Figure 1, we analyze the sensitivity of our method. For different sparsity parameters  $\alpha$  and different privacy budgets  $\epsilon$ , which determine *convex\_min* given a fixed  $M$ , we show how often, out of 100 simulations each, our algorithm recovered the interaction term (leftmost bar in red), the main effects scaled by a factor of 1/2 to account for the two main effects (middle bar in green) and all effects, i.e. the interaction effect and the two main effects (rightmost bar in blue). As the privacy budget  $\epsilon$  increases, the amount of noise added to the regression problem decreases, and hence the frequency of selecting the correct effects in the regression analysis increases. The plots also show that as the sparsity parameter  $\alpha$  increases, the frequency of selecting the correct terms decreases.

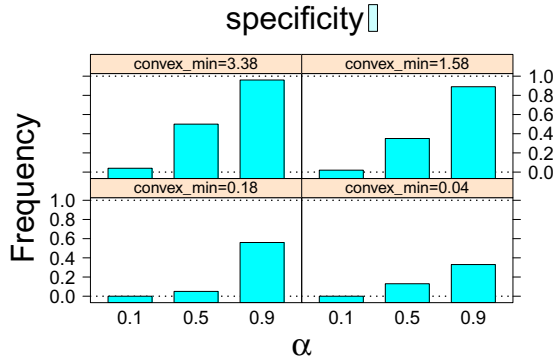
In Figure 2 we analyze the specificity of our method. For different sparsity parameters  $\alpha$  and different strong convexity conditions *convex\_min*, we show how



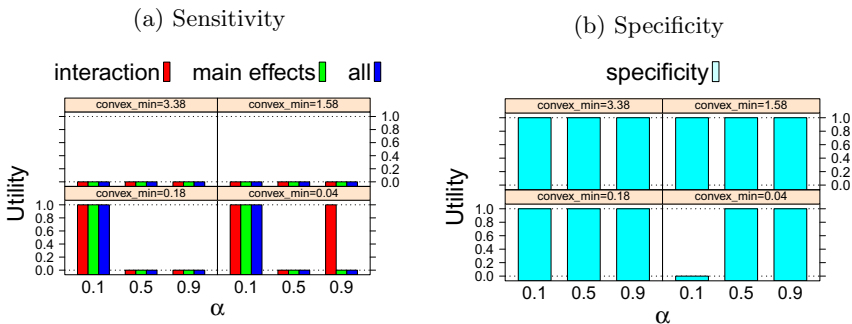
**Fig. 1.** Sensitivity analysis for different sparsity parameters  $\alpha$ , privacy budgets  $\epsilon$ , and strong convexity conditions *convex\_min* when the top 5 SNPs are used for the analysis: the red (leftmost) bar shows how often, out of 100 simulations each, the algorithm recovered the interaction term, the green (middle) bar corresponds to the main effects scaled by a factor of 1/2 and the rightmost (blue) bar corresponds to all effects, i.e. 2 main effects and 1 interaction effect

often, out of 100 simulations each, our algorithm did not include any additional effects in the selected model. As  $\alpha$  increases, the selected model becomes sparser and the algorithm is hence less likely to wrongly include additional effects. We also observe that as *convex\_min* decreases, the specificity increases. This can be explained by how we choose the candidate parameters  $\lambda$ , namely as multiples of the smallest allowed value for  $\lambda$ , which is  $\text{convex\_min}/(1 - \alpha)$ . When  $\lambda$  is small, the effect of the penalty terms diminishes, and we are essentially performing a regular logistic regression, which does not produce sparse models.

In Figure 3, we plotted the results of non-private penalized logistic regression with elastic-net penalty to contrast Figure 1 and Figure 2. The results of the non-private penalized logistic regression is indirectly related to  $\epsilon$  because the choice of the smallest regularization parameter  $\lambda$  is bounded below by  $\text{convex\_min}/(1 - \alpha)$  and *convex\_min* is a function of  $\epsilon$ . We can observe from Figure 3 that when the regularization parameter  $\lambda$  is large (i.e.,  $\text{convex\_min} \geq 1.58$ ), the regression analysis screens out all effects. Hence, the sensitivity is 0 and the specificity is 1. When  $\lambda$  is small (i.e.,  $\text{convex\_min} \leq 0.18$ ), the amount of regularization



**Fig. 2.** Specificity analysis for different sparsity parameters  $\alpha$  and strong convexity conditions *convex\_min*: the plot shows how often, out of 100 simulations each, our algorithm did not include any additional effects in the selected model



**Fig. 3.** Results of non-private logistic regression with elastic-net penalty. Figure 3a and Figure 3b would be compared with Figure 1 and Figure 2, respectively.

also becomes marginal, and we begin to see that the sensitivity increases but the specificity decreases. Figure 3 shows that we can identify the correct model when  $\alpha = 0.1$  and *convex\_min* = 0.18. In contrast, when we use the same  $\alpha$  and *convex\_min* for differentially private regressions, Figure 1 shows that we can obtain a good sensitivity result, but Figure 2 shows that the specificity result for this choice is poor.

## 6 Conclusions

Various papers have argued that it is possible to use aggregate genomic data to compromise the privacy of individual-level information collected in GWAS databases. In this paper, we respond to these attacks by proposing a new method to release regression coefficients from association studies that satisfy differential privacy and hence come with privacy guarantees against arbitrary external information.

By extending the approaches in Chaudhuri and Vinterbo [11] and Kifer et al. [10] we developed an end-to-end differentially private procedure for solving regression problems with convex penalty functions including selecting the penalty parameters by cross-validation. We also provided the exact form of the random noise used in the objective function perturbation mechanism and showed that the perturbation noise can be efficiently sampled.

As a special case of a regression problem, we focused on penalized logistic regression with elastic-net regularization, a method widely used to perform GWAS analyses and identify disease-causing genes. Our simulation results in Section 5 showed that our method is applicable to GWAS data sets and enables us to perform data analysis that preserves privacy and utility. The risk-utility analysis about the tradeoff between privacy ( $\epsilon$ ) and utility (correctly identifying the causative SNPs) helps us decide on the appropriate level of privacy guarantee for the released data. We hope that approaches such as those described in this paper will allow the release of more information from GWAS going forward and allay the privacy concerns that others have voiced over the past decade.

## References

- [1] Austin, E., Pan, W., Shen, X.: Penalized regression and risk prediction in genome-wide association studies. *Statistical Analysis and Data Mining* 6(4) (August 2013)
- [2] Cho, S., et al.: Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis. *BMC Proceedings* 3(suppl. 7), S25 (2009)
- [3] Homer, N., et al.: Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP geno-typing microarrays. *PLoS Genetics* 4(8), e1000167 (2008)
- [4] Couzin, J.: Whole-genome data not anonymous, challenging assumptions. *Science* 321(5894), 1278 (2008)
- [5] Zerhouni, E.A., Nabel, E.G.: Protecting aggregate genomic data. *Science* 322(5898), 44 (2008)

- [6] Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) TCC 2006. LNCS, vol. 3876, pp. 265–284. Springer, Heidelberg (2006)
- [7] Uhler, C., Slavkovic, A.B., Fienberg, S.E.: Privacy-preserving data sharing for genome-wide association studies. *Journal of Privacy and Confidentiality* 5(1), 137–166 (2013)
- [8] Johnson, A., Shmatikov, V.: Privacy-preserving data exploration in genome-wide association studies. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1079–1087 (2013)
- [9] Yu, F., et al.: Scalable Privacy-Preserving Data Sharing Methodology for Genome-Wide Association Studies. *Journal of Biomedical Informatics* (February 2014)
- [10] Kifer, D., Smith, A., Thakurta, A.: Private convex empirical risk minimization and high-dimensional regression. *Proceedings of Journal of Machine Learning Research - Proceedings Track* 23, 25.1–25.40 (2012)
- [11] Chaudhuri, K., Vinterbo, S.A.: A stability-based validation procedure for differentially private machine learning. In: *Advances in Neural Information Processing Systems*, pp. 1–19 (2013)
- [12] Chaudhuri, K., Monteleoni, C., Sarwate, A.D.: Differentially private empirical risk minimization. *JMLR* 12(7), 1069–1109 (2011)
- [13] Laurent, B., Massart, P.: Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics* 28(5), 1302–1338 (2000)
- [14] Wright, F.A., et al.: Simulating association studies: a data-based resampling method for candidate regions or whole genome scans. *Bioinformatics* 23(19), 2581–2588 (2007)
- [15] Malaspinas, A.S., Uhler, C.: Detecting epistasis via Markov bases. *Journal of Algebraic Statistics* 2(1), 36–53 (2010)
- [16] Gómez, E., Gomez-Villegas, M.A., Marín, J.M.: A multivariate generalization of the power exponential family of distributions. *Communications in Statistics - Theory and Methods* 27(3), 589–600 (1998)

## A Proofs

### A.1 Proof of Theorem 3

**Lemma 8.** *Let  $G$ ,  $g_1$ , and  $g_2$  be vector-valued continuous functions. Suppose that  $G$  is  $\lambda$ -strongly convex,  $g_1$  is convex and  $\gamma_1$ -Lipschitz, and  $g_2$  is convex and  $\gamma_2$ -Lipschitz. If  $f_1 = \arg \min_f (G + g_1)(f)$  and  $f_2 = \arg \min_f (G + g_2)(f)$ , then  $\|f_1 - f_2\|_2 \leq (\gamma_1 + \gamma_2)/\lambda$ .*

*Proof (of Lemma 8).*  $G + g_1$  and  $G + g_2$  are  $\lambda$ -strongly convex because  $G$  is  $\lambda$ -strongly convex and  $g_1$  and  $g_2$  are convex. Then for  $j, k, w \in \{1, 2\}$ ,  $j \neq k$ ,

$$(G + g_w)(f_j) \geq (G + g_w)(f_k) + \partial(G + g_w)(f_k)^T (f_j - f_k) + \frac{\lambda}{2} \|f_j - f_k\|^2$$

where  $\partial(G + g_w)$  denotes the subgradient. We know that  $0 \in \partial(G + g_w)(f_w)$  because  $f_w$  minimizes  $G + g_w$ . Hence,

$$\begin{aligned} (G + g_2)(f_1) &\geq (G + g_2)(f_2) + \frac{\lambda}{2} \|f_1 - f_2\|_2^2, \\ (G + g_1)(f_2) &\geq (G + g_1)(f_1) + \frac{\lambda}{2} \|f_1 - f_2\|_2^2. \end{aligned}$$

By summing these two inequalities we obtain

$$(G + g_2)(f_1) + (G + g_1)(f_2) \geq (G + g_2)(f_2) + (G + g_1)(f_1) + \lambda \|f_1 - f_2\|_2^2$$

and hence  $[g_2(f_1) - g_2(f_2)] + [g_1(f_2) - g_2(f_1)] \geq \lambda \|f_1 - f_2\|_2^2$ . The fact that  $g_w$  is  $\gamma_w$ -Lipschitz implies that  $\left|g_2(f_1) - g_2(f_2)\right| + \left|g_1(f_2) - g_2(f_1)\right| \leq (\gamma_1 + \gamma_2) \|f_1 - f_2\|_2$  and hence

$$\begin{aligned} \lambda \|f_1 - f_2\|_2^2 &\leq [g_2(f_1) - g_2(f_2)] + [g_1(f_2) - g_2(f_1)] \\ &\leq \left|g_2(f_1) - g_2(f_2)\right| + \left|g_1(f_2) - g_2(f_1)\right| \leq (\gamma_1 + \gamma_2) \|f_1 - f_2\|_2. \end{aligned}$$

Therefore  $\|f_1 - f_2\|_2 \leq (\gamma_1 + \gamma_2)/\lambda$ .  $\square$

*Proof (of Theorem 3).* For notational convenience we assume that  $c_{\min} \geq c^*$  so that

$$L(\theta; T) = \frac{1}{n} \sum_{d \in T} l(\theta; d) + \lambda^T r(\theta) + \frac{\varphi}{\epsilon n} b^T \theta.$$

If  $c_{\min} < c^*$ , we can extend  $r$  to include  $r_{t+1}(\theta) = \frac{\max\{0, c^* - c_{\min}\}}{2} \|\theta\|_2^2$  and extend each  $\lambda \in A$  such that  $\lambda_{t+1} = 1$ . First, we show that  $|q(\theta^*(T), V) - q(\theta^*(T'), V)| \leq \beta_1/n$  for training sets  $T$  and  $T'$  that differ only by one record. Here,  $\theta^*(T) = \arg \min_{\theta} L(\theta; T)$ . Let  $d = T \setminus T'$ ,  $d' = T' \setminus T$ ,

$$\begin{aligned} G(\theta; T, T') &= \frac{1}{n} \sum_{d \in T \cap T'} l(\theta; d) + \lambda^T r(\theta) + \frac{\varphi}{\epsilon n} b^T \theta, \\ g_1(\theta; T, T') &= \frac{1}{n} l(\theta; d) \quad \text{and} \quad g_2(\theta; T, T') = \frac{1}{n} l(\theta; d'). \end{aligned}$$

Then  $G$  is  $c_{\min}$ -strongly convex, and  $g_1$  and  $g_2$  are convex and  $\gamma/n$ -Lipschitz. By Lemma 8,  $\|\theta^*(T) - \theta^*(T')\|_2 \leq \frac{2\gamma}{nc_{\min}}$ . Since  $h$  is  $\kappa$ -Lipschitz we obtain for any validation set  $V$ ,  $|q(\theta^*(T), V) - q(\theta^*(T'), V)| \leq \frac{2\gamma\kappa}{nc_{\min}}$ .

Second, we show that for all  $\lambda \in A$  and for all validation sets  $V$  and  $V'$  that differ in a single record,  $|q(\theta^*(T), V) - q(\theta^*(T'), V')| \leq \beta_2/m$ . Since  $h$  is non-negative,  $|q(\theta^*(T), V) - q(\theta^*(T'), V')| \leq h_{\max}/m$ , where  $h_{\max} = \sup_d h(\theta^*(T); d)$ . By definition,  $h_{\max} \leq h^*$ . Moreover, because  $h$  is  $\kappa$ -Lipschitz,  $h_{\max} \leq \kappa \|\theta^*(T)\|_2$ . So  $h_{\max} \leq \min\{h^*, \kappa \|\theta^*(T)\|_2\}$ . Now let  $E$  be the event that  $\|b\|_2 \leq \xi$ . Provided that  $E$  holds, we have  $|b^T \theta_1 - b^T \theta_2| \leq \|b\|_2 \|\theta_1 - \theta_2\|_2 \leq \xi \|\theta_1 - \theta_2\|_2$ . Let  $G(\theta) = \lambda^T r(\theta)$ ,  $g_1(\theta; T) = \frac{1}{n} \sum_{d \in T} l(\theta; d) + \frac{\varphi}{\epsilon n} b^T \theta$ , and  $g_2(\theta) = 0$ . Then  $G$  is  $c_{\min}$ -strongly convex,  $g_1$  is  $(\gamma + \frac{\varphi\xi}{\epsilon n})$ -Lipschitz, and  $g_2$  is 0-Lipschitz. Since  $G + g_2$  is minimized when  $\theta = 0$ , we obtain by invoking Lemma 8 that  $\|\theta^*(T)\|_2 = \|\theta^*(T) - 0\|_2 \leq \frac{1}{c_{\min}} (\gamma + \frac{\varphi\xi}{\epsilon n})$ . Therefore,  $|q(\theta^*(T), V) - q(\theta^*(T), V')| \leq \frac{1}{m} \min\left\{h^*, \frac{\kappa}{c_{\min}} (\gamma + \frac{\varphi\xi}{\epsilon n})\right\}$ .  $\square$

## A.2 Proof of Theorem 7

**Lemma 9.** *If  $A$  is of full rank and  $E$  has rank at most 2, then*

$$\frac{\det(A + E) - \det(A)}{\det(A)} = \lambda_1(A^{-1}E) + \lambda_2(A^{-1}E) + \lambda_1(A^{-1}E)\lambda_2(A^{-1}E),$$

where  $\lambda_j(Z)$  denotes the  $j$ -th eigenvalue of matrix  $Z$ .

*Proof (of Lemma 9).* See Lemma 10 in Chaudhuri et al. [12].

*Proof (of Theorem 7).* Similar to the proof by Chaudhuri et al. [12], we show that if  $r$  is infinitely differentiable, then Algorithm 1 is  $\epsilon$ -differentially private. It then follows from the successive approximation method by Kifer et al. [10] that Algorithm 1 is still  $\epsilon$ -differentially private even if  $r$  is convex but not necessarily differentiable.

Let  $g$  denote the probability density function of the algorithm's output  $\theta^*$ . Our goal is to show that  $e^{-\epsilon} \leq \frac{g(\theta|D)}{g(\theta|D')} \leq e^\epsilon$ . Suppose that the Hessian of  $r$  is continuous. Because  $0 = \nabla L(\theta; D)$ , we have

$$T_D(\theta) := b = -\frac{\epsilon}{\varphi} \left[ \sum_{d \in D} \nabla l(\theta; d) + n \nabla r(\theta) \right] \quad \text{and} \quad \nabla T_D(\theta) = -\frac{\epsilon}{\varphi} \left[ \sum_{d \in D} \nabla^2 l(\theta; d) + n \nabla^2 r(\theta) \right].$$

$T_D$  is injective because  $L(\theta; D)$  is strongly convex. Also,  $T_D$  is continuously differentiable. Therefore,

$$\frac{g(\theta|D)}{g(\theta|D')} = \frac{f(T_D(\theta))}{f(T_{D'}(\theta))} \frac{|\det(\nabla T_D)(\theta)|}{|\det(\nabla T_{D'})(\theta)|},$$

where  $f$  is the density function of  $b$ .

We first consider  $\frac{|\det(\nabla T_D)(\theta)|}{|\det(\nabla T_{D'})(\theta)|}$ . Let  $A = -\frac{\epsilon}{\varphi} \nabla T_{D'}$ ,  $E = \nabla^2 l(\theta; D \setminus D') - \nabla^2 l(\theta; D' \setminus D)$ . Because  $l$  is convex and  $r$  is strongly convex,  $\nabla T_D(\theta)$  is positive definite. Hence,  $A$  has full rank. Also,  $E$  has rank at most 2 because  $\nabla^2 l(\theta; d)$  is a rank 1 matrix by assumption. By Lemma 9,

$$\frac{|\det(\nabla T_D(\theta))|}{|\det(\nabla T_{D'}(\theta))|} = \left| \frac{\det(A + E)}{\det(A)} \right| \leq 1 + s_1(A^{-1}E) + s_2(A^{-1}E) + s_1(A^{-1}E)s_2(A^{-1}E),$$

where  $s_i(M)$  denotes the  $i$ th largest singular value of  $M$ . Because  $r$  is  $\lambda$ -strongly convex, the smallest eigenvalue of  $A$  is at least  $n\lambda$ . So  $s_i(A^{-1}E) \leq \frac{s_i(E)}{n\lambda}$ . Because  $\|\nabla l(\theta; d)\|_j \leq \kappa$  for  $j \in \{1, 2\}$ , applying the triangle inequality to the nuclear norm yields  $s_1(E) + s_2(E) \leq \|\nabla^2 l(\theta; D \setminus D')\|_1 + \|\nabla^2 l(\theta; D' \setminus D)\|_1 \leq 2c$ . Therefore,  $s_1(A^{-1}E) s_2(A^{-1}E) \leq (\frac{c}{n\lambda})^2$ , and

$$\frac{|\det(\nabla T_D)(\theta)|}{|\det(\nabla T_{D'})(\theta)|} = \frac{|\det(A + E)|}{|\det(A)|} \leq \left( 1 + \frac{c}{n\lambda} \right)^2.$$

Now we consider  $\frac{f(T_D(\theta))}{f(T_{D'}(\theta))}$ . Since

$$\begin{aligned} \|T_D(\theta) - T_{D'}(\theta)\|_j &= \left( \frac{\epsilon}{\varphi} \right) \|\nabla l(\theta; D \setminus D') - \nabla l(\theta; D' \setminus D)\|_j \\ &\leq \left( \frac{\epsilon}{\varphi} \right) \left( \|\nabla l(\theta; D \setminus D')\|_j + \|\nabla l(\theta; D' \setminus D)\|_j \right) \leq \frac{2\kappa\epsilon}{\varphi}, \end{aligned}$$

we obtain  $\frac{f(T_D(\theta))}{f(T_{D'}(\theta))} = \exp\left(-\frac{\|T_D(\theta)\|_j}{2}\right) / \exp\left(\frac{-\|T_{D'}(\theta)\|_j}{2}\right) \leq \exp\left(\frac{\kappa\epsilon}{\varphi}\right)$ , and therefore,

$$\frac{f(T_D(\theta))}{f(T_{D'}(\theta))} \frac{|\det(\nabla T_D)(\theta)|}{|\det(\nabla T_{D'})(\theta)|} \leq \exp\left(\frac{\kappa\epsilon}{\varphi} + 2 \log\left(1 + \frac{c}{n\lambda}\right)\right) \leq e^\epsilon.$$

### A.3 Proof of Proposition 4

*Proof (of Proposition 4).* The distribution of  $X$  is a special case of an  $s$ -dimensional *power exponential distribution* as defined by Gómez et al. [16], namely  $X \sim PE_s(\mu, \Sigma, \beta)$  with  $\mu = (0, \dots, 0)^T$ ,  $\Sigma = \text{Id}_s$  and  $\beta = \frac{1}{2}$ . Gómez et al. [16] proved that if  $T \sim PE_s(\mu, \Sigma, \beta)$ , then  $T$  has the same distribution as  $\mu + YA^TZ$ , where  $Z$  is a random vector with uniform distribution on the unit sphere in  $\mathbb{R}^s$ ,  $Y$  is an absolutely continuous non-negative random variable, independent from  $Z$ , whose density function is

$$g(y) = \frac{s}{\Gamma\left(1 + \frac{s}{2\beta}\right) 2^{\frac{s}{2\beta}}} y^{s-1} \exp\left(-\frac{1}{2}y^{2\beta}\right) I_{(0,\infty)}(y),$$

and  $A \in \mathbb{R}^{s \times s}$  is a square matrix such that  $A^T A = \Sigma$ .

Note that for  $\beta = \frac{1}{2}$ , the distribution of  $Y$  boils down to a  $\chi^2$ -distribution with  $2s$  degrees of freedom. In addition, if  $W \sim \mathcal{N}(0, \text{Id}_s)$ , then  $W/\|W\|$  is uniformly distributed on the unit  $s$ -sphere. Finally, since  $\Sigma = \text{Id}_s$  we get that  $A = \text{Id}$ .  $\square$