TRANSACTIONS ON DATA PRIVACY 10 (2017) 203–235

Instance-Based Learning with ℓ-diversity*

Koray Mancuhan, Chris Clifton****

**Purdue University, 305 N. University St., West Lafayette IN, 47907, United States.

E-mail: kmancuha@purdue.edu, clifton@cs.purdue.edu

Received 28 February 2017; received in revised form 13 June 2017; accepted 9 October 2017

Abstract. Corporations are retaining ever-larger corpuses of personal data; the frequency or breaches and corresponding privacy impact have been rising accordingly. One way to mitigate this risk is through use of anonymized data, limiting the exposure of individual data to only where it is absolutely needed. This would seem particularly appropriate for data mining, where the goal is generalizable knowledge rather than data on specific individuals. In practice, corporate data miners often insist on original data, for fear that they might "miss something" with anonymized or differentially private approaches. This paper provides a theoretical justification for the use of anonymized data. Specifically, we show that a k-nearest neighbor classifier trained on anatomized data satisfying ℓ-diversity should be expected to do as well as on the original data. Anatomized data preserves all attribute values, but introduces uncertainty in the mapping between identifying and sensitive values, thus satisfying ℓ-diversity. The theoretical effectiveness of the proposed approach is validated using several publicly available datasets, showing that we outperform the state of the art for nearest neighbor classification using training data protected by k-anonymity, and are comparable to learning on the original data.

Keywords. ℓ-diversity, k-nearest neighbor, non-parametric models, machine learning

1 Introduction

Many privacy definitions have been proposed based on generalizing/suppressing data (ℓ-diversity[1], k-anonymity [2, 3], t-closeness [4], δ-presence [5], (α,k)-anonymity [6]). Other alternatives include value swapping [7], distortion [8], randomization [9], and noise addition (e.g., differential privacy [10]). Generalization consists of replacing identifying attribute values with a less specific version [3]. Suppression can be viewed as the ultimate generalization, replacing the identifying value with an "any" value [3]. Generalization has the advantage of preserving truth, but a less specific truth that reduces utility of the published data.

Xiao and Tao proposed anatomization as a method to enforce ℓ-diversity while preserving specific data values [11]. Anatomization splits instances across two tables, one containing identifying information and the other containing private information. The more general approach of fragmentation [12] divides a given dataset's attributes into two sets of attributes

*Supported by the Northrop Grumman Cybersecurity Research Consortium

(2 partitions) such that an encryption mechanism avoids associations between two different small partitions. Vimercati et al. extend fragmentation to multiple partitions [13], and Tamas et al. propose an extension that deals with multiple sensitive attributes [14]. The main advantage of anatomization/fragmentation is that it preserves the original values of data; the uncertainty is only in the mapping between individuals and sensitive values.

We show that this additional information has real value. First, we demonstrate that in theory, learning from anatomized data can be as good as learning from the raw data. We then demonstrate empirically that learning from anatomized data beats learning from generalization-based anonymization.

This paper looks only at instance-based learning, specifically k-nearest neighbor classifier (k-NN). This focus was chosen because we have solid theoretical results on the limits of learning using k-NN, allowing us to compare theoretical bounds on learning from anatomized data with learning from the underlying unprotected data. We demonstrate this for a simple approach of anatomizing the data; we simply consider all possible mappings of individuals to sensitive values as equally likely.

There is concern that anatomization is vulnerable to several attacks [15, 16, 17]. While this can be an issue, *any* method that provides meaningful utility fails to provide perfect privacy against a sufficiently strong adversary [18, 10]. Introducing uncertainty into the anonymization process reduces the risk of many attacks, e.g., minimality [19, 20]. Our theoretical analysis holds for any assignment of items to anatomy groups, including a random assignment, which provides a high degree of robustness against minimality and correlation-based attacks. While this does not eliminate privacy risk, if the alternative is to use the original data, we show that anatomy provides comparable utility while reducing the privacy risk. This paper has the following key contributions:

1. We define a classification task on anatomized data without violating the random worlds assumption. A violating classification task would be the prediction of sensitive attribute, a task that was found to be #P-complete by Kifer [15].

2. To our best knowledge, this is the first paper in the privacy community that studies the theoretical effect of running k-NN using anatomized training data. We show the anatomization effect for the error rate bounds and the convergence rate when the test data is neither anonymized nor anatomized. Inan et al. already give a practical applications of such a learning scenario [21].

3. We show the generalization error for any non-parametric classifier using the anatomized training data.

4. We compare the k-NN classifier trained on the anatomized data with the k-NN classifier trained on the unprotected data. In case of nearest neighbor classifier (1-NN), we also make an additional comparison to a generalization based learning scheme [21].

5. We last give an analytical comparison of the convergence rate of 1-NN classifier between ℓ-diversity and original data.

We next summarize related work and define the problem statement. We then give necessary definitions and notations. Section 4 proposes the heuristic algorithm and gives theoretical results. Experimental results are presented in section 5. Section 6 summarizes the work and gives future directions.

2 Related Work and Problem Statement

There have been studies on how to train classifiers on anonymized data. The existing work can be grouped into three categories: generalization based classification, distribution reconstruction based classification, and differential privacy based classification.

In generalization based classification, nearest neighbor classification using generalized data was investigated by Martin [22]. Nested generalization and non-nested hyperrectangles were used to generalize the data from which the nearest neighbor classifiers were trained. Iyengar suggested using a classification metric so as to find the optimum generalization. Then, a C4.5 decision tree classifier was trained from the optimally generalized training data [23]. Zhang et al. studied Naïve Bayes using partially specified training data [24], proposing a conditional likehoods computation algorithm exploring the instance space of attribute-value generalization taxonomies. Inan et al. proposed nearest neighbor and support vector machine classifiers using anonymized training data that satisfy k-anonymity. Taylor approximation was used to estimate the dot product, Euclidean distance and kernel function from generalized data [21].

In distribution reconstruction based classification, Agrawal et al. proposed an iterative distribution reconstruction algorithm for distorted training data from which a C4.5 decision tree classifier was trained [25]. Fung et al. gave a top-down specialization method (TDS) for anonymization so that the anonymized data allows accurate decision trees. A new scoring function was proposed for the calculation of decision tree splits from the compressed training data [26]. Dowd et al. studied C4.5 decision tree learning from training data perturbed by random substitutions. A matrix based distribution reconstruction algorithm was applied on the perturbed training data from which an accurate C4.5 decision tree classifier was learned [27].

In differential privacy based classification, Rubinstein et al. studied the kernels of support vector machine in the differential privacy and showed the trade-off between privacy level and the data utility. They analyzed finite and infinite dimensional kernels in function of the approximation error under differential privacy [28]. Lin at al. studied training support vector classification for outsourced data. Random transformation was applied on the training set so that the cloud server could compute the accurate model without knowing what the actual values were [29]. Jain et al. studied the support vector machine kernels in the differential privacy setting. They proposed differentially private mechanisms to train support vector machines for interactive, semi-interactive and non-interactive learning scenarios, providing theoretical analysis of the proposed approaches [30].

None of the earlier work has provided a classifier directly applicable to anatomized training data. Such a classifier requires specific theoretical and experimental analysis, because anatomized training data provides additional detail that has the potential to improve learning; but also additional uncertainty that must be dealt with. Furthermore, most of the previous work didn't justify theoretically why the proposed heuristics let classifiers generalize well. Therefore, this paper studies Problem 1:

Problem 1. Define a heuristic to train a k-nearest neighbor classifier on anatomized data without violating ℓ-diversity while using the sensitive information, with a theoretical guarantee of good generalization under reasonable assumptions.

3 Definitions and Notations

The first four definitions restate standard definitions of unprotected data and attribute types.

Definition 2. A dataset D is called a *person specific dataset* for population P if each instance $x_i \in D$ belongs to a unique individual $p \in P$. The person specific dataset has the schema in (1)

$$(C, A_1, ..., A_d, A_s) \tag{1}$$

where C is the class attribute, $A_i, ..., A_d$ are *identifying* attributes, and A_s is the *sensitive attribute*. Quasi-identifying and sensitive attributes are defined below.

The person specific dataset will be called the original training data in this paper. Next, we define the types of attributes.

Definition 3. A set of attributes are called *direct identifying attributes* if they let an adversary associate an instance $x_i \in D$ to a unique individual $p \in P$ without any background knowledge.

Definition 4. A set of attributes are called *quasi-identifying attributes* if there is background knowledge available to the adversary that associates the quasi-identifying attributes with a unique individual $p \in P$.

We include both direct and quasi-identifying attributes under the name identifying attributes. First name, last name and social security number (SSN) are common examples of direct identifying attributes. Direct identifying attributes are generally not useful in machine learning as unique identifiers would not generalize. Hence, the schema in (1) does not include any direct identifying attribute. They can be suppressed but other information can be both useful and assist in identifying individuals. Some common examples of quasi-identifying attributes are age, postal code, and occupation. These are assumed to be public knowledge and thus need not be protected. Thus, the quasi-identifying attributes are included in schema (1) as $A_1, ..., A_d$. Next, we will give the second type of attribute.

Definition 5. An attribute A_s of D is called a *sensitive attribute* if we should protect against adversaries correctly inferring the value for any individual $x_i \in D$.

Patient disease and individual income are common examples of sensitive attributes. Unique individuals $p \in P$ typically don't want these sensitive information to be revealed to individuals without a direct need to know that information. Provided an instance $x_i \in D$, the *class label* is denoted by $x_i.c$. We don't consider the case where c is sensitive, as this would make the purpose of classification to violate privacy. c is neither sensitive nor identifying in this paper, although our analysis holds for c being an identifying attribute.

Given the former definitions, we will next define the anonymized training data following the definition of k-anonymity [3].

Definition 6. A training dataset D that satisfies the following conditions is said to be *anonymized training data* D_k [3]:

1. The training data D_k does not contain any unique identifying attributes.
2. Every instance $x_i \in D_k$ is indistinguishable from at least $(k-1)$ other instances in D_k with respect to its quasi-identifying attributes.

Anatomy satisfies a slightly weaker definition; the indistinguishability applies only to sensitive data. This will be captured in Definitions 9-11.

In this paper, we assume that the anonymized training data D_k is created according to a *generalization* based data publishing method. We next define the *comparison classifiers*.

Definition 7. A k-nearest neighbor classifier (k-NN) that is trained on the anonymized training data D_k is called *the anonymized k-NN*.

Definition 8. A k-nearest neighbor classifier (k-NN) that is trained on the original training data D is called *the original k-NN*.

The theoretical aspects of comparison classifiers are out of the scope of this paper. We will remind the theoretical analysis of the original k-NN in the end of this section [31].

We go further from Definition 6, requiring that there must be multiple possible sensitive values that could be linked to an individual. ***The proposed algorithms will be centered around the following definitions***. This new requirement uses the definition of *groups* [11].

Definition 9. A *group* G_j is a subset of instances in original training data D such that $D = \cup_{j=1}^{m} G_j$, and for any pair (G_{j_1}, G_{j_2}) where $1 \leq j_1 \neq j_2 \leq m$, $G_{j_1} \cap G_{j_2} = \emptyset$.

Next, we define the concept of ℓ-diversity or ℓ-diverse (multiple possible sensitive values) for all the groups in the original training data D regarding to Xiao et al. [11].

Definition 10. A set of groups is *ℓ-diverse* if and only if $\forall\ G_j, v \in \Pi_{A_s}(G_j); \frac{freq(v,G_j)}{|G_j|} \leq \frac{1}{\ell}$ where A_s is the sensitive attribute in D, $\Pi_{A_s}(*)$ is the database A_s projection operation on original training data $*$ (or on data table in the database community), $freq(v, G_j)$ is the frequency of v in G_j and $|G_j|$ is the number of instances in G_j.

We extend the data publishing method *anatomization* that is originally based on ℓ-diverse groups by Xiao et al. [11]. We should note that Machanavajjhala et al. proposes the ℓ-diversity standard before Xiao et al. using the term blocks instead of groups [1, 11]. As we are following Xiao et al., we will be using groups.

Definition 11. Given an original training data D partitioned in m ℓ-diverse groups according to Definition 10, *anatomization* produces an *identifying table* IT and a *sensitive table* ST as follows. IT has schema

$$(C, A_1, ..., A_d, GID)$$

including the class attribute, the quasi-identifying attributes $A_i \in IT$ for $1 \leq i \leq d$, and the *group id* GID of the group G_j. For each group $G_j \in D$ and each instance $x_i \in G_j$, IT has an instance x_i of the form:

$$(x_i.c, x_i.a_1, ..., x_i.a_d, j)$$

ST has schema

$$(GID, A_s)$$

where A_s is the sensitive attribute in D and GID is the group id of the group G_j. For each group $G_j \in D$ and each instance $x_i \in G_j$, ST has an instance of the form:

$$(j, x_i.a_s)$$

Unlike Machanavajjhala et al., the data publishing scheme here doesn't require the instances in a group to generalize to the same value. This comes from Xiao et al.'s definition of anatomization. Unlike Xiao et al., our definition of anatomization doesn't include the count of sensitive attributes within a group as this statistic wouldn't form an interesting feature from the machine learning perspective.

The IT table includes only the quasi-identifying and class attributes. We assume that direct identifying attributes are removed before creating the IT and ST tables. We have the following observation from Definition 11 to train a classifier: *every instance $X_i \in IT$ can be matched to the instances $X_j \in ST$ within the same group using the common attribute GID in both data table.* This observation yields the *anatomized training data*.

Definition 12. Given two data tables IT and ST resulting from the anatomization on original training data D, the *anatomized training data* D_A is

$$D_A = \Pi_{IT.C,IT.A_1,\cdots IT.A_d,ST.A_s}(\ IT \bowtie\ ST)$$

where $\bowtie$ is the database inner join operation with respect to the condition $IT.GID = ST.GID$ and $\Pi(*)$ is the database projection operation on training data *.

Anatomized training data shows one of the most naïve data preprocessing approaches. Another one is ignoring the sensitive attribute in ST table.

Definition 13. Given two data tables IT and ST resulting from the anatomization on original training data D, the *identifying training data* D_{id} is

$$D_{id} = \Pi_{IT.C,IT.A_1,\cdots IT.A_d}(\ IT\ )$$

where $\Pi(*)$ is the database projection operation on training data *.

Identifying training data doesn't use all the information available in the published data (and would likely lead to users insisting on having the original data.) The naïve training method of Definition 12, on the other hand, is both computationally costly (a factor of ℓ increase in size) and noisy: for every true instance, there are $\ell - 1$ incorrect instances that might cause incorrect classification. However, if the test instance comes from the original data's distribution and the original training data is sufficiently large; then the naïve training method of Definition 12 is expected to work well. In this paper, we propose the following classifier based on this naïve training method.

Definition 14. A k-nearest neighbor (k-NN) classifier that is trained on the anatomized training data D_A is called *the anatomized k-NN classifier*.

We will later show that such a naïve classifier will have very interesting theoretical properties. Having $\ell - 1$ incorrect instances for every true instance does not always result in a bad classifier if the test instance to classify does not have uncertainty between the identifying and the sensitive information. In fact, it is theoretically possible that anatomized k-NN classifier outperforms the original k-NN despite the uniform noise of ℓ-diversity! Next will be the final definition in this section.

Definition 15. A k-nearest neighbor (k-NN) classifier that is trained on only the identifying training data D_{id} is called *the identifying k-NN classifier*.

Now we are giving the notations that will be used in the rest of this paper. D will denote the original training data whereas D_A will denote the anatomized training data. D has

N instances and D_A has $N\ell$ instances from definition 12. All instances are i.i.d whether they are in training or test data. The total number of attributes are assumed to be $d+1$ (d identifying attributes and 1 sensitive attribute.) For the sake of simplicity, A_{id} will denote the identifying attributes $A_1 \cdots A_d \in IT$. T stands for a test data which is not processed by any anatomization and generalization method. x will be an instance of the test data T. $d(U,V)$ is the quadratic distance metric for a pair of instances U and V in metric space $M \subset \mathbb{R}^{d+1}$. $X'_N(k)$ denotes the set of k number of nearest neighbors of x in D that the original k-NN classifier uses while $X'_{N\ell}(k)$ denotes the set of k number of nearest neighbors of x in D_A that the anatomized k-NN classifier uses. x_i will interchangeably be an instance of D or D_A and x_j will interchangeably be an instance of $X'_N(k)$ or $X'_{N\ell}(k)$. In case of $k=1$, we will use X'_N and $X'_{N\ell}$ for the nearest neighbors in D and D_A. X is the random variable with probability distribution $P(x)$ from which x and x_i are drawn. Training and test instances will be column vectors in format of $(a_1, ..., a_d, a_s)^T$. C is the class attribute in D and D_A with binary labels 1 and 2.

We are also recalling the notations from the statistical pattern recognition that will be used in the theoretical analysis [32]. In the probability notations and proofs, we will abuse x to represent $X = x$. Given the training data D and the class label i; $q_i(x)$, $P_i(x)$ and P_i stand for the posterior probability, the likelihood probability and the prior probability respectively. If the anatomized training data D_A is used, $q_{A_i}(x)$, $P_{A_i}(x)$ and P_{A_i} are the symmetric definitions for the class label i. When $\forall\, x \in T$ hold, the probability of classifying x incorrectly will be referred by *error rate*. $R(X'_N(k), x)$ is the error rate when $x \in T$ is classified using $X'_N(k)$. If $X'_{N\ell}(k)$ is used to classify x, $R_A(X'_{N\ell}(k), x)$ will be the error rate. When $\forall\, x_j \in X'_N(k); x_j \cong x$ hold, we denote the error rate by $R^k(x)$ in (2).

$$\begin{aligned} R^k(x) = & \sum_{i=1}^{k+1/2} \frac{1}{i} \binom{2i-2}{i-1} [q_1(x)q_2(x)]^i \\ & + \frac{1}{2} \binom{k+1}{k+1/2} [q_1(x)q_2(x)]^{k+1/2} \end{aligned} \tag{2}$$

$R^k_A(x)$ is the error rate when $\forall\, x_j \in X'_{N\ell}(k); x_j \cong x$ holds. $R^k_A(x)$ can trivially be derived from (2) by substituting $q_i(x)$ with $q_{A_i}(x)$. For all $x \in T$, the smallest error rate that could be obtained from the best possible classifier will be referred by *Bayesian error* [32]. The Bayesian errors given x are denoted by $R^*(x)$ and $R^*_A(x)$ when $\forall\, x_j \in X'_N(k); x_j \cong x$ and $\forall\, x_j \in X'_{N\ell}(k); x_j \cong x$ hold respectively. (3) computes $R^*(x)$.

$$\begin{aligned} R^*(x) &= min\{q_1(x), q_2(x)\} \\ &\cong \sum_{i=1}^{\infty} \frac{1}{i} \binom{2i-2}{i-1} [q_1(x)q_2(x)]^i \end{aligned} \tag{3}$$

$R^*_A(x)$ can trivially be derived again from (3) by substituting $q_i(x)$ with $q_{A_i}(x)$. R^k and R^k_A, which are the respective expectations of $R^k(x)$ ($E\{R^k(x)\}$) and $R^k_A(x)$ ($E\{R^k_A(x)\}$) regarding to X, will stand for the error rate of original k-NN and anatomized k-NN classifiers respectively. R^* and R^*_A, which are the respective expectations of $R^*(x)$ ($E\{R^*(x)\}$) and $R^*_A(x)$ ($E\{R^*_A(x)\}$) regarding to X, will stand for the Bayesian errors of original training data and anatomized training data respectively. We will denote $R^1(x)$ and $R^1_A(x)$ by $R(x)$ and $R_A(x)$ for convenience. Similarly, R and R_A will denote R^1 and R^1_A.

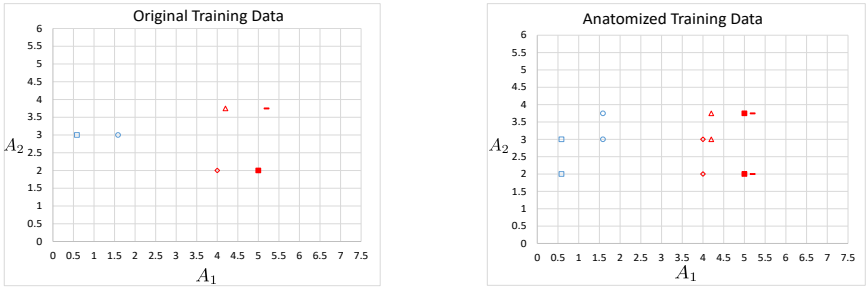
We finish this section with the assumptions in the theoretical analysis. We assume that all the training data has a smooth probability distribution. Although anatomization requires

a discrete probability distribution for the sensitive attribute A_s, such smoothness violation is negligible since the original k-NN classifier is known to fit well on discrete training data [33]. The sensitive attribute A_s is assumed to be non-binary. The binary case would be meaningless for ℓ-diversity since ℓ can only be 2, implying a coin flip privacy guarantee on the sensitive attribute. The anatomized k-NN is assumed to have odd k values, because even k values encompass the tie cases among $X'_N(k)$ and $X'_{N\ell}(k)$ that make the bounds ambiguous and complicated [32]. All instances are assumed to be in a separable metric space $M \subset \mathbb{R}^{d+1}$ following Cover et al., Devroye et al. and Fukunaga et al. [32, 34, 35]. Last, we will assume that the original training data D satisfies ℓ-diversity condition and that every group has ℓ instances in the creation of anatomized training data D_A from the original training data D.

4 Anatomized k-NN Classifier

4.1 Illustration

We will illustrate the anatomized k-NN classifier through the example in Figure 1. Although the example is for an anatomized 3-NN classifier, the procedure is general for any "k" value (cf. Definition 14.)



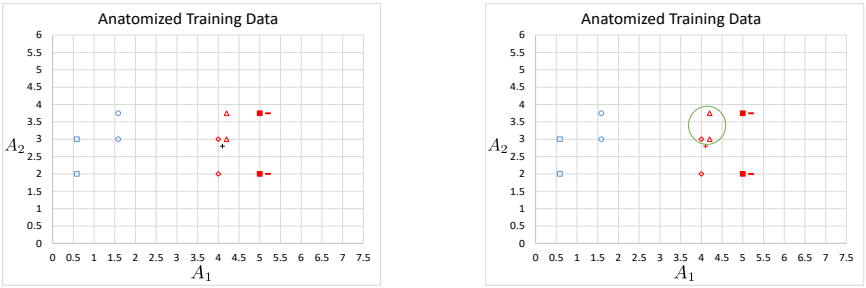
(a) Original Training Data (b) Anatomized Training Data

Figure 1: Toy Example of Training Data with Two Attributes A_1 and A_2

.

Figure 1a shows the original training data with six instances: two instances of a blue class (on the left side) and four of a red class (on the right side), with two attributes A_1 and A_2. Here, every instance has a different shape color since it belongs to a unique individual with unique A_1 value. Figure 1b shows the anatomized training data created from $IT(A_1, GID)$ and $ST(GID, A_2)$ when $\ell = 2$ (cf. Definitions 11 and 12.) It has 12 instances in total.

Figure 2 shows the anatomized 3-NN classifier in the toy example. To classify an unlabeled test instance (black $+$ in Figure 2a), the anatomized 3-NN classifier finds 3 instances from the anatomized training data that are closest to the test instance. In Figure 2b, the 3 points that are located in the green circle are the 3 closest points to the test instance. Since the test instance's true class label is red, we obtain the true classification despite using 2 noisy instances (incorrect matches) and 1 true instance (correct match.) Since the training data is separable in the toy example and the test instance comes from the subspace where the red instances are clustered, the correct classification is obtained despite the distortion of ℓ-diversity.



(a) Unlabeled Test Instance (b) Labeled Test Instance

Figure 2: Anatomized 3-NN Classifier in Toy Example

Note that the ℓ-diversity's distortion might be much more critical in practice than our toy example's. For example, such distortion could transform the linearly separable original training data to non-linearly separable anatomized training data. This could hence result in incorrect classifications as the distortion changes the likelihood probabilities that define the generalization error of any non-parametric classifier. *As long as the ℓ-diversity's distortion on the likelihood probabilities is minimized, the generalization ability of anatomized k-NN is more likely to be same as original k-NN's.* Before elaborating on the theoretical aspects of ℓ-diversity's perturbation, we will discuss the implementation details and complexity, and the preservation of ℓ-diversity.

4.2 Implementation

The anatomized k-NN classifier has two phases of implementation. First is the creation of the anatomized training data whereas the second phase is the implementation of original k-NN classifier to train on the anatomized training data. The first phase is the inner join operation between the IT and ST tables and the projection operation on the former inner join result (cf. Definition 12.) The inner join operation can be implemented in $O(N\ logN)$ time whereas the projection operation can be implemented in $O(N)$. Hence, the anatomized training data creation can be implemented in $O(N\ logN)$ time [36]. The second phase is the k-NN classifier implementation. The naïve approach of finding the exact k neighbors requires going through the entire anatomized training data which takes $O(N\ell\ \ (d+1))$ time. LSH based algorithms, on the other hand, can compute the c-approximate k nearest neighbors in $k\ \frac{1}{c^2} + O(k\ log[\frac{log\ N\ell}{log^{\frac{1}{3}} N\ell}])$ time [37].

4.3 Privacy Preservation

Correct or incorrect classification from the anatomized k-NN classifier doesn't help an adversary learn additional information about the sensitive attribute's distribution within a group. It is theoretically possible to have a correct classification from the k closest training instances even if this are all incorrect matches. In addition, the original training data might be linearly inseparable whereas the anatomized training data could be linearly separable. Anatomized k-NN classifier just uses k "plausible" instances in the anatomized training data which is more biased according to ℓ-diversity's distortion (cf. Section 4.4). Thus, the anatomized k-NN classifier preserves ℓ-diversity.

4.4 Theoretical Analysis

We will analyze the anatomized k-NN classifier in terms of the asymptotic error rate bounds (infinite size training data), error rate convergence on the finite size training data, and the generalization error. The analysis will follow Cover et al. and Fukunaga [32, 34].

The convergence analysis will be limited to anatomized 1-NN classifier since the original k-NN classifier is analyzed only for 1 and 2 neighbors. The analysis of 2 neighbors is omitted here since it directly follows the 1 neighbor case.

The generalization error will follow the non-parametric density estimation classifier to keep the discussion easy to follow. The analysis result is general for k-NN classifiers with any k value as well [32].

4.4.1 Asymptotic Error Rate Bounds

We will first show the error bounds for the anatomized 1-NN classifier. We will then discuss the extension to the anatomized k-NN classifier for all odd $k > 1$. Corollary 16 expresses formally the convergence of the nearest neighbor which is critical for the error bounds of the anatomized 1-NN classifier.

Corollary 16. *Let $x \in T$ and $x_1, \cdots, x_N \in D$ be i.i.d instances taking values separable in any metric space $M \subset \mathbb{R}^{d+1}$. Let X'_N be the nearest neighbor of x in D. Then, $\lim_{N\to\infty} X'_N = x$ with probability one [34].*

Proof. Let $S_x(r) = \{\bar{x} \in M : d(x, \bar{x}) \leq r\}$ be the sphere with radius $r > 0$ centered at x. Let's consider that x has a sphere $S_x(r)$ with non-zero probability. Therefore, for any radius $\delta > 0$;

$$\lim_{N\to\infty} P\{\min_{i=1,\cdots,N} d(x_i, x) \geq \delta\} = \lim_{N\to\infty} [1 - P(S_x(\delta))]^N = 0 \tag{4}$$

□

Basically, Corollary 16 says that if the original training data has an infinite number of instances, it is guaranteed to find the nearest neighbor of a test instance that is drawn from the same probability distribution.

Obviously, the nearest neighbor could have either a correct or an incorrect sensitive attribute if we use the anatomized training data instead of the original training data. Although the sensitive attribute value could change the specific instance which becomes the nearest neighbor in terms of the distance, there would still be a nearest neighbor if the original has an infinite number of instances. Finding the nearest neighbor from the anatomized training data is equivalent to finding the nearest neighbor from D where one attribute value is swapped in the multivariate distribution. Next, Theorem 17 shows the error bounds of the anatomized 1-NN classifier under this assumption from Corollary 16.

Theorem 17. *Let $M \subset \mathbb{R}^{d+1}$ be a metric space. Let $P_{A_1}(x)$ and $P_{A_2}(x)$ be the likelihood probabilities of x such that $P_A(x) = P_{A_1}P_{A_1}(x) + P_{A_2}P_{A_2}(x)$ with class priors P_{A_1} and P_{A_2}. Last, let's assume that x is either a point of non-zero probability measure or a continuity point of $P_{A_1}(x)$ or $P_{A_2}(x)$. Then, the nearest neighbor has the probability of error R_A with the bounds*

$$R^*_A \leq R_A \leq 2R^*_A \tag{5}$$

where R^*_A denotes the Bayesian error when the anatomized training data D_A is used.

We now give a sketch of proof for Theorem 17. Let $R_A(X'_{N\ell}, x)$ denote the probability of error for a pair of instances $x \in T$ and $X'_{N\ell} \in D_A$. Since Corollary 16 shows that $\lim_{N\to\infty} X'_{N\ell} = x$ always hold, (6) is derived from (2) by substituting k with 1 and $q_i(x)$ with $q_{A_i}(x)$.

$$\lim_{N\to\infty} R_A(X'_{N\ell}, x) \cong R_A(x) = 2q_{A_1}(x)q_{A_2}(x) \tag{6}$$

The rest of the derivation follows Cover et al. using (2), (3) [34].

Extending (5) from the anatomized 1-NN classifier to the anatomized k-NN classifier for all odd $k > 1$ follows the steps in Corollary 16 and Theorem 17. The key is to show that $\lim_{N\to\infty} x_j = x$ holds for all $x_j \in X'_{N\ell}(k)$. The rest is to derive an expression of $R^k_A(x)$ as in (6) for all odd $k > 1$ and show that $R^k_A(x)$ is always less than $2R^*_A$ and $R^{k-2}_A(x)$. This can be derived following the analysis of original k-NN classifier in Fukunaga [32]. The anatomized k-NN classifier has the bound (7)

$$R^*_A \leq \cdots \leq R^5_A \leq R^3_A \leq R_A \leq 2R^*_A \tag{7}$$

for all odd $k > 1$.

Note that the Bayesian errors R^*_A and R^* are not always same due to the ℓ-diverse groups of the anatomization. The ℓ-diverse groups cause new likelihood $P_{A_i}(x)$ and eventually posterior probabilities $q_{A_i}(x)$. R^*_A thus differs from (3), because (3) uses $q_i(x)$ instead of $q_{A_i}(x)$. We will return back to this in the generalization error discussion.

4.4.2 Error Rate Convergence on Finite Size Training Data

We now discuss the error rate of the anatomized 1-NN classifier when the size of anatomized training data is finite. This error rate will let us derive the convergence rate to the Bayesian error for the anatomized 1-NN classifier. The discussion here won't be generalized to the anatomized k-NN classifier since the finite size training data performance of k-NN classifiers are not generalized to 3 or more neighbors in the pattern recognition literature [35, 32]. Theorem 18 extends the analysis of Fukunaga and Fukunaga et al. [32, 38].

Theorem 18. *Let $M \subset \mathbb{R}^{d+1}$ be a metric space. Let's assume that the original training data D satisfies ℓ-diversity condition and that every group has ℓ instances in the creation of anatomized training data D_A from the original training data D. Let $P_A(x)$ and $P(x)$ be the smooth density functions of x. Let $P_{A_1}(x)$ and $P_{A_2}(x)$ be the class likelihood density functions of x. Let P_{A_1} and P_{A_2} be the class priors such that $P_A(x) = P_{A_1}P_{A_1}(x) + P_{A_2}P_{A_2}(x)$. Let $q_{A_1}(x)$ and $q_{A_2}(x)$ be the smooth posterior probability densities such that $q_{A_1}(x) + q_{A_2}(x) = 1$ and $N\ell \to \infty$. Let $q_{A_1}(X'_{N\ell})$ and $q_{A_2}(X'_{N\ell})$ be the smooth posterior probability densities such that $q_{A_1}(X'_{N\ell}) + q_{A_2}(X'_{N\ell}) = 1$ and $N\ell \nrightarrow \infty$. Let $\delta > 0$ be the difference between $q_{A_i}(x)$ and $q_{A_i}(X'_{N\ell})$ for class labels $i = \{1, 2\}$. Let $d(X'_{N\ell}, x)$ be the quadratic distance with matrix A and ρ be the calculated value of $d(X'_{N\ell}, x)$. Let R_A be the error rate of the anatomized 1-NN classifier when $N\ell \to \infty$. Last, let R_{A_N} be the error rate of the anatomized 1-NN classifier when $N\ell \nrightarrow \infty$. Then,*

$$R_{A_N} \cong R_A + \beta \frac{1}{(N\ell)^{\frac{2}{d+1}}} E_X\{|A|^{-\frac{1}{d+1}} tr\{AB(x)\}\} \tag{8}$$

where β is

$$\beta = \frac{\Gamma^{\frac{2}{d+1}}(\frac{d+3}{2})\Gamma(\frac{2}{d+1}+1)}{\pi(d+1)} \tag{9}$$

and $B(x)$ is

$$\begin{aligned} B(x) = P_A^{-\frac{2}{d+1}}(x)[q_{A_2}(x) - q_{A_1}(x)] \\ \times [\frac{1}{2}\nabla^2 q_{A_1}(x) + P_A^{-1}(x)\nabla P_A(x)\nabla^T q_{A_1}(x)] \end{aligned} \tag{10}$$

Proof. We first define $q_{A_i}(X'_{N\ell})$ in function of $q_{A_i}(x)$ and δ.

$$q_{A_1}(X'_{N\ell}) = q_{A_1}(x) + \delta \tag{11}$$
$$q_{A_2}(X'_{N\ell}) = q_{A_2}(x) - \delta \tag{12}$$

R_{A_N} is written in function of R_A and δ using (11) and (12) in (13)

$$\begin{aligned} R_{A_N} &= E\{q_{A_1}(x)(q_{A_2}(x) - \delta) + q_{A_2}(x)(q_{A_1}(x) + \delta)\} \\ &= R_A + E[[q_{A_2}(x) - q_{A_1}(x)]\delta] \end{aligned} \tag{13}$$

where $E[[q_{A_2}(x) - q_{A_1}(x)]\delta]$ is (14)

$$\begin{aligned} E\{(q_{A_2}(x) - q_{A_1}(x))\delta)\} &= E_x\{E_\rho\{E_{X'_{N\ell}}\{[q_{A_2}(x) - q_{A_1}(x)]\delta|\rho, x\}|x\}\} \\ &= E_x\{[q_{A_2}(x) - q_{A_1}(x)]E_\rho\{E_{X'_{N\ell}}\{\delta|\rho, x\}|x\}\}. \end{aligned} \tag{14}$$

Following Fukunaga, last line of (14) requires 3-step expectation calculation. Step 1 gives

$$E_{X'_{N\ell}}\{\delta|\rho, x\} \approxeq \frac{\rho^2}{d+1} \times tr\{A[\frac{1}{2}\nabla^2 q_{A_1}(x) + P_A^{-1}(x)\nabla P_A(x)\nabla^T q_{A_1}(x)]\} \tag{15}$$

Step 2 uses (15) to calculate $E_\rho\{E_{X'_{N\ell}}\{\delta|\rho, x\}|x\}$. This eventually requires the computation of $E\{\rho^2\}$. Although the probability distribution of ρ is unknown, the probability distribution of local region u around test instance $X \in T$ including the nearest neighbor $X'_{N\ell}$ is known. We therefore need to formulate ρ^2 in function of u^2. The derivation of u as a function of ρ is given in (16).

$$u \approxeq p(x)\frac{\pi^{\frac{d+1}{2}}}{\Gamma(\frac{d+3}{2})}\rho^{d+1}|A|^{\frac{1}{2}} \tag{16}$$

Rewriting (16) and taking the expectation of both sides result in (17).

$$E\{\rho^2\} = \frac{\Gamma^{\frac{2}{d+1}}(\frac{d+3}{2})}{p^{\frac{2}{d+1}}(x)\pi|A|^{\frac{1}{d+1}}}E\{u^{\frac{2}{d+1}}\} \tag{17}$$

(18) computes $E\{u^{\frac{2}{n}}\}$.

$$\int_0^1 u^{\frac{2}{d+1}}P_u(u)du = N\ell\int_0^1 u^{\frac{2}{d+1}}(1-u)^{N\ell-1}du = \frac{\Gamma(\frac{2}{d+1}+1)\Gamma(N\ell+1)}{\Gamma(N\ell+\frac{2}{d+1}+1)} \tag{18}$$

Replacing the result of (18) in (17) gives (19).

$$E\{\rho^2\} = \frac{\Gamma^{\frac{2}{d+1}}(\frac{d+3}{2})\Gamma(\frac{2}{d+1}+1)}{p^{\frac{2}{d+1}}(x)\pi|A|^{\frac{1}{d+1}}}\frac{\Gamma(N\ell+1)}{\Gamma(N\ell+\frac{2}{d+1}+1)} \tag{19}$$

Assuming N and n have values large enough, $\frac{\Gamma(N\ell+1)}{\Gamma(N\ell+\frac{2}{n}+1)}$ is approximated in (20).

$$\frac{\Gamma(N\ell+1)}{\Gamma(N\ell+\frac{2}{n}+1)} = \frac{N\ell}{N\ell+\frac{2}{n}} \times \frac{\Gamma(N\ell)}{\Gamma(N\ell+\frac{2}{n})} \approxeq \frac{1}{(N\ell)^{\frac{2}{n}}} \tag{20}$$

Replacing the result of (20) in (19) results in (21).

$$E\{\rho^2\} \approxeq \frac{\Gamma^{\frac{2}{d+1}}(\frac{d+3}{2})\Gamma(\frac{2}{d+1}+1)}{p^{\frac{2}{d+1}}(x)\pi|A|^{\frac{1}{d+1}}} \frac{1}{(N\ell)^{\frac{2}{d+1}}} \tag{21}$$

Using (21) in $E_\rho\{E_{X'_{N\ell}}\{\delta|\rho,x\}|x\}$ results in

$$\begin{aligned} E_\rho\{E_{X'_{N\ell}}\{\delta|\rho,x\}|\, x\} \approxeq \beta \frac{1}{(N\ell)^{\frac{2}{d+1}}} |A|^{-\frac{1}{d+1}} \times tr\{A\, P_A^{-\frac{2}{d+1}}(x) \\ \times [\frac{1}{2}\nabla^2 q_{A_1}(x) + P_A^{-1}(x)\nabla P_A(x)\nabla^T q_{A_1}(x)]\} \end{aligned} \tag{22}$$

where β is (23).

$$\beta = \frac{\Gamma^{\frac{2}{d+1}}(\frac{d+3}{2})\Gamma(\frac{2}{d+1}+1)}{\pi(d+1)} \tag{23}$$

Step 3 uses (22) to calculate the last line of (14). Rewriting results in (24)

$$E_x\{[q_{A_2}(x) - q_{A_1}(x)]\, E_\rho\{E_{X'_{N\ell}}\{\delta|\rho,x\}|x\}\} \approxeq \beta \frac{1}{(N\ell)^{\frac{2}{d+1}}} E_x\{|A|^{-\frac{1}{d+1}} tr\{AB(x)\}\} \tag{24}$$

where $B(x)$ is (25)

$$B(x) = P_A^{-\frac{2}{d+1}}(x)[q_{A_2}(x) - q_{A_1}(x)] \times [\frac{1}{2}\nabla^2 q_{A_1}(x) + P_A^{-1}(x)\nabla P_A(x)\nabla^T q_{A_1}(x)]. \tag{25}$$

Replacing (24) in (13) and rewriting (13) yields (8). □

From Theorem 23, we see that the anatomized 1-NN classifier has a faster convergence rate than the original 1-NN classifier's ($O(\frac{1}{N\ell})$ vs $O(\frac{1}{N})$.) This is a surprising result despite the ℓ-diversity condition. However, we still don't know what kind of error rate (R_A) the anatomized 1-NN classifier is converging to. Formally, we need to compare the bounds of error rate R_A to the bounds of error rate R. The generalization error analysis will elaborate this comparison through non-parametric density estimation classifier.

4.4.3 Generalization Error Analysis

In pattern recognition literature, the generalization ability of any classifier is defined through the classifier's Bayesian error estimation ability [32, 35, 39]. This is reasonable for k-NN classifiers as well since the error rate of original or anatomized k-NN classifier is bounded by the Bayesian errors (See (7) for anatomized k-NN.)

In this section, the Bayesian error will be estimated for binary classification using non-parametric density estimation classifier. Parzen density estimation will be used with mixed kernel function [32]. This approach is chosen because its derivation is easier and more readable than the k-NN density estimation's one. The analysis, which follows Fukunaga [32] and Fukunaga et al. [40], is general enough for any non-parametric density estimation classifier including k-NN [32]. The multi-label classification is ignored since its theoretical work is limited for the original training data [39]. We first give three Axioms and a Lemma.

Axiom 19. *Given the anatomized training data D_A and the training data D; let P_{A_i} and P_i be the class priors for class labels $i = \{1, 2\}$. Assume that D satisfies the ℓ-diversity condition and that every group has ℓ instances in the creation of anatomized training data D_A from the original training data D. Then, $P_i = P_{A_i}$.*

Axiom 20. *Given the anatomized training data D_A and the training data D; let $P_A(X.A_{id})$ and $P(X.A_{id})$ be the smooth joint densities of identifying attributes A_{id}. Assume that D satisfies ℓ-diversity condition and that every group has ℓ instances in the creation of anatomized training data D_A from the original training data D. Then, $P(X.A_{id}) = P_A(X.A_{id})$.*

Axiom 21. *Given the anatomized training data D_A and the training data D; let $P_A(X.A_s)$ and $P(X.A_s)$ be the smooth densities of sensitive attribute A_s. Assume that D satisfies ℓ-diversity condition and that every group has ℓ instances in the creation of anatomized training data D_A from the original training data D. Then, $P(X.A_s) = P_A(X.A_s)$.*

Axioms 19, 20 and 21 are obvious due to the following: *provided a sample of size N drawn from a probability distribution P, repeating every instance for fixed $\ell > 0$ times and obtaining a sample of size $N\ell$ does not change the probability distribution P. The estimated parameters $\widehat{\mu}$ and $\widehat{\sigma^2}$ of distribution P remain same as long as there is no suppression.*

Lemma 22. *Given the anatomized training data D_A and the training data D, let identifying attributes A_{id} and the sensitive attribute A_s be independent. Let's assume that D satisfies ℓ-diversity condition and that every group has ℓ instances in the creation of anatomized training data D_A from the original training data D. Then, $P_A(x) = P(x)$ is always true under the axioms 20 and 21.*

Using Axioms 20 and 21, the proof of Lemma 22 is straightforward. Lemma 22 and Axioms 1-to-3 yield the Theorem 23. Using Lemma 22, we will assume that $R^*_A = R^*$ holds asymptotically for Bayesian errors.

Theorem 23. *Let $M \subset \mathbb{R}^{d+1}$ be a metric space. Provided anatomized training data D_A, let $P_{A_1}(x)$ and $P_{A_2}(x)$ be the class likelihood probability density functions of X. Let P_{A_1} and P_{A_2} be the class priors. Let $P_A(x)$ be the smooth density function of X such that $P_A(x) = P_{A_1}P_{A_1}(x)+P_{A_2}P_{A_2}(x)$. Provided original training data D, let $P_1(x)$ and $P_2(x)$ be the class likelihood probability density functions of X. Let P_1 and P_2 be the class priors. Let $P(x)$ be the smooth density function of X such that $P(x) = P_1P_1(x) + P_2P_2(x)$. Let $h_A(x) = -ln(\frac{P_{A_1}(x)}{P_{A_2}(x)})$ and $h(x) = -ln(\frac{P_1(x)}{P_2(x)})$ be the density classifiers with biases $\Delta h_A(x)$ and $\Delta h(x)$ respectively. Let $t = ln(\frac{P_{A_1}}{P_{A_2}}) = ln(\frac{P_1}{P_2})$ be the decision threshold with threshold bias Δt. Let $\epsilon_A > 0$ be the small changes on $P_1(x)$ and $P_2(x)$ resulting in $P_{A_1}(x)$ and $P_{A_2}(x)$; and $\widehat{R}^*_A$, $\widehat{R}^*$ be the Bayesian error estimations with respective biases ΔR^*_A, ΔR^*. Let $\widehat{P}_{A_i}(x)$ and $\widehat{P}_i(x)$ be the Parzen density estimations of likelihood densities; and $K(*)$ be the kernel function for original training data D with shape matrix A and size/volume parameter r [32]. Last, let's assume the following:*

1. *A_{id} and A_s are independent in the original training data D and the anatomized training data D_A (independence in joint distributions, not for the distributions with respect to class labels $\{1, 2\}$.*
2. *$R^*_A = R^*$ hold.*
3. *$\Delta t < 1$.*
4. *The original training data D satisfies ℓ-diversity condition and every group has ℓ instances in the creation of anatomized training data D_A from the original training data D.*

Therefore,

$$\begin{aligned}\widehat{R}^*_A \cong R^* + a_1 r^2 + a_2 r^4 + a_3 \frac{r^{-(d+1)}}{N} \\ + \epsilon_A a_4 r^2 + \epsilon_A a_5 r^4 - \epsilon_A a_6 \frac{r^{-(d+1)}}{N}\end{aligned} \tag{26}$$

where a_i is an integration term. $a_1 r^2$, $a_2 r^4$, $\epsilon_A a_4 r^2$ and $\epsilon_A a_5 r^4$ are the bias terms while $a_3 \frac{r^{-(d+1)}}{N}$ and $\epsilon_A a_6 \frac{r^{-(d+1)}}{N}$ are the variance terms.

In the proof, we will use Taylor approximations up to the second order and the negligible terms will be ignored for convenience throughout the derivation. We will abuse x to denote random variable X.

Proof. Let's rewrite $P_{A_1}(x)$ and $P_{A_2}(x)$ by $P_1(x) + \epsilon_1$ and $P_2(x) - \epsilon_2$ where ϵ_i is a small change for any class label $i = \{1, 2\}$ and $\epsilon_1 \neq \epsilon_2$. We first write $P_A(x)$.

$$\begin{aligned}P_A(x) &= P_1 P_{A_1}(x) + P_2 P_{A_2}(x) \\ &= P_1[P_1(x) + \epsilon_1] + P_2[P_2(x) - \epsilon_2] \\ &= P_1 P_1(x) + P_2 P_2(x) + P_1\epsilon_1 - P_2\epsilon_2 \\ &= P(x) + P_1\epsilon_1 - P_2\epsilon_2\end{aligned} \tag{27}$$

In the last line of (27), lemma 22 tells that $P_A(x) = P(x)$ always holds under the theorem's independence assumptions. Thus, (28) is valid

$$\epsilon_2 = \frac{P_1}{P_2}\epsilon_1 = e^t \epsilon_1 \tag{28}$$

when $t = ln(P_1/P_2)$ [32]. Let ϵ_A stand for ϵ_1 in the remainder of the text. We then write $P_{A_1}(x) = P_1(x) + \epsilon_A$ and $P_{A_2}(x) = P_2(x) - e^t\epsilon_A$ for the class likelihoods of the anatomized training data D_A. We will derive in detail the approximations for the class 1 and omit the details for class 2 since both derivations are symmetric.

Next is the approximation of $E\{\widehat{P}_{A_i}(x)\}$ in function of $P_i(x)$ and ϵ_A. (29) approximates $E\{\widehat{P}_{A_i}(x)\}$ using convolution and $\int K(x)dx = 1$.

$$\begin{aligned}E\{\widehat{P}_{A_1}(x)\} &= \int P_{A_1}(x)K_1(x-y)dy = \int [P_1(y) + \epsilon_A]K_1(x-y)dy \\ &= \int P_1(y)K_1(x-y)dy + \epsilon_A \int K_1(x-y)dy \\ &= P_1(x) * K_1(x) + \epsilon_A \\ &\approxeq P_1(x) + P_1(x)\frac{1}{2}\alpha_1(x)r^2 + \epsilon_A\end{aligned} \tag{29}$$

In (29), $\alpha_i(x)$ is $tr\{\frac{\nabla^2 P_i(x)}{P_i(x)}A\}$. Through a similar approach, we have

$$E\{\widehat{P}_{A_2}(x)\} \approxeq P_2(x) + P_2(x)\frac{1}{2}\alpha_2(x)r^2 - e^t\epsilon_A \tag{30}$$

According to Fukunaga, variance is

$$Var\{\widehat{P}_{A_1}(x)\} = \frac{1}{N}[P_{A_1}(x) * K_1^2(x) - E^2\{\widehat{P}_{A_1}(x)\}]. \tag{31}$$

We will only use the first order terms to keep the calculation tractable. $E^2\{\widehat{P}_{A_1}(x)\}$ is (32)

$$[E\{\widehat{P}_1(x)+\epsilon_A]^2 \approxeq [P_1(x)+\epsilon_A]^2 \approxeq P_1(x)[P_1(x)+2\epsilon_A] \tag{32}$$

and $P_{A_1}(x) * K_1^2(x)$ is (33) using $w_1 = \int K^2(x)dx$.

$$\begin{aligned}\int P_{A_1}(y)K_1^2(x-y)dy &= \int [P_1(y)+\epsilon_A]K_1^2(x-y)dy \\ &= \int P_1(y)K_1^2(x-y)dy + \epsilon_A \int K_1^2(x-y)dy \\ &= P_1(x) * K_1^2(x) + \epsilon_1 w_1 \\ &\approxeq w_1 P_1(x) + \epsilon_A w_1\end{aligned} \tag{33}$$

Replacing (32) and the last line of (33) in (31) results in (34).

$$\begin{aligned}Var\{\widehat{P}_{A_1}(x)\} &\approxeq \frac{1}{N}[[w_1P_1(x)+\epsilon_A w_1] - [P_1(x)[P_1(x)+2\epsilon_A]]] \\ &= \frac{1}{N}[w_1P_1(x) - P_1^2(x) + \epsilon_A[w_1 - 2P_1(x)]] \\ &= \frac{1}{N}[w_1P_1(x) - P_1^2(x)] + \frac{\epsilon_A}{N}[w_1 - 2P_1(x)]\end{aligned} \tag{34}$$

The approximation of $P_{A_2}(x) * K_2^2(x)$ and $E^2\{\widehat{P}_{A_2}(x)\}$yields (35).

$$Var\{\widehat{P}_{A_2}(x)\} \approxeq \frac{1}{N}[w_2P_2(x) - P_2^2(x)] - \frac{\epsilon_A e^t}{N}[w_2 - 2P_2(x)] \tag{35}$$

According to Fukunaga, the bias ΔR_A^* of the Bayesian error estimation is (36)

$$\begin{aligned}E[\Delta R_A^*] \approxeq \frac{1}{2\pi}\int\int E[\Delta h_A(x) + \frac{(j\omega)}{2}\Delta h_A^2(x)]e^{j\omega h_A(x)} \\ \times [P_{A_1}P_{A_1}(x) - P_{A_2}P_{A_2}(x)]d\omega dx.\end{aligned} \tag{36}$$

(36) requires the approximations of the expected decision function biases $E\{\Delta h_A(x)\}$ and $E\{\Delta h_A^2(x)\}$. As the approximations are a function of $E\{\frac{\Delta P_{A_i}(x)}{P_{A_i}(x)}\}$ and $E\{(\frac{\Delta P_{A_i}(x)}{P_{A_i}(x)})^2\}$, we will next derive them using terms up to second order. $E\{\frac{\Delta P_{A_1}(x)}{P_{A_1}(x)}\}$ is approximately the last line of (37) using $\frac{\epsilon_A}{P_1(x)} < 1$ and the first order Taylor series.

$$\begin{aligned}E\{\frac{\Delta P_{A_1}(x)}{P_{A_1}(x)}\} &= \frac{1}{P_{A_1}(x)}E\{\widehat{P}_{A_1}\} - 1 \\ &\approxeq \frac{1}{P_1(x)+\epsilon_A}[P_1(x) + P_1(x)\frac{1}{2}\alpha_1(x)r^2 + \epsilon_A] - 1 \\ &= \frac{1}{1+\frac{\epsilon_A}{P_1(x)}} + \frac{1}{1+\frac{\epsilon_A}{P_1(x)}}\frac{1}{2}\alpha_1(x)r^2 + \frac{1}{1+\frac{P_1(x)}{\epsilon_A}} - 1 \\ &\approxeq \frac{1}{2}\alpha_1(x)r^2 - \frac{\epsilon_A}{P_1(x)}[\frac{1}{2}\alpha_1(x)r^2 + 1 - \frac{1}{1+\frac{\epsilon_A}{P_1(x)}}] \\ &\approxeq \frac{1}{2}\alpha_1(x)r^2 - \frac{\epsilon_A}{P_1(x)}[\frac{1}{2}\alpha_1(x)r^2 + 1 - (1 - \frac{\epsilon_A}{P_1(x)})] \\ &= \frac{1}{2}\alpha_1(x)r^2 - \epsilon_A\frac{1}{2}\frac{\alpha_1(x)}{P_1(x)}r^2\end{aligned} \tag{37}$$

Similarly $E\{\frac{\Delta P_{A_2}(x)}{P_{A_2}(x)}\}$ is (38).

$$E\{\frac{\Delta P_{A_2}(x)}{P_{A_2}(x)}\} \approxeq \frac{1}{2}\alpha_2(x)r^2 + e^t\epsilon_A\frac{1}{2}\frac{\alpha_2(x)}{P_2(x)}r^2 \tag{38}$$

$E\{(\frac{\Delta P_{A_1}(x)}{P_{A_1}(x)})^2\}$ is (39)

$$E\{(\frac{\Delta P_{A_1}(x)}{P_{A_1}(x)})^2\} = \frac{1}{P^2_{A_1}(x)}Var\{\widehat{P}_{A_1}(x)\} + \frac{1}{P^2_{A_1}(x)}E^2\{\Delta P_{A_1}(x)\} \tag{39}$$

that requires the approximation of $\frac{1}{P^2_{A_1}(x)}Var\{\widehat{P}_{A_1}(x)\}$ and $\frac{1}{P^2_{A_1}(x)}E^2\{\Delta P_{A_1}(x)\}$.
$\frac{1}{P^2_{A_1}(x)}Var\{\widehat{P}_{A_1}(x)\}$ is approximated in (40) using $(\frac{P_1(x)}{P_1(x)+\epsilon_A})^2 \approxeq 1-\frac{2\epsilon_A}{P_1(x)}$ and $w_1 = s_1 r^{d+1}$

$$\begin{aligned}
\frac{1}{P^2_{A_1}(x)}Var\{\widehat{P}_{A_1}(x)\} &\approxeq \frac{1}{N(P_1(x)+\epsilon_A)^2}[w_1P_1(x) - P_1^2(x)] + \\
&\qquad \epsilon_A\frac{1}{N(P_1(x)+\epsilon_A)^2}[w_1 - 2P_1(x)] \\
&= \frac{P_1^2(x)}{(P_1(x)+\epsilon_A)^2}[\frac{w_1}{NP_1(x)} + \frac{\epsilon_A w_1}{NP_1^2(x)} - \frac{2\epsilon_A}{NP_1(x)}] \\
&\approxeq [1-\frac{2\epsilon_A}{P_1(x)}][\frac{w_1}{NP_1(x)} + \frac{\epsilon_A w_1}{NP_1^2(x)} - \frac{2\epsilon_A}{NP_1(x)}] \\
&\approxeq \frac{s_1}{NP_1(x)r^{d+1}} - \epsilon_A[\frac{s_1}{NP_1^2(x)r^{d+1}} + \frac{2}{NP_1(x)}] \\
&\approxeq \frac{s_1}{NP_1(x)r^{d+1}} - \epsilon_A\frac{s_1}{NP_1^2(x)r^{d+1}}
\end{aligned} \tag{40}$$

whereas $\frac{1}{P^2_{A_1}(x)}E^2\{\Delta P_{A_1}(x)\}$ is approximated in (41).

$$\begin{aligned}
\frac{1}{P^2_{A_1}(x)}E^2\{\Delta P_{A_1}(x)\} &\approxeq (\frac{1}{2}\alpha_1(x)r^2 - \epsilon_A\frac{1}{2}\frac{\alpha_1(x)}{P_1(x)}r^2)^2 \\
&\approxeq \frac{1}{4}\alpha_1^2(x)r^4 - \alpha_1(x)r^2\epsilon_A\frac{1}{2}\frac{\alpha_1(x)}{P_1(x)}r^2 \\
&= \frac{1}{4}\alpha_1^2(x)r^4 - \epsilon_A\frac{1}{2}\frac{\alpha_1^2(x)}{P_1(x)}r^4
\end{aligned} \tag{41}$$

Replacing the last lines of (40) and (41) in (39) results in (42).

$$\begin{aligned}
E\{(\frac{\Delta P_{A_1}(x)}{P_{A_1}(x)})^2\} \approxeq \frac{s_1}{NP_1(x)r^{d+1}} &- \epsilon_A\frac{s_1}{NP_1^2(x)r^{d+1}} \\
&+ \frac{1}{4}\alpha_1^2(x)r^4 - \epsilon_A\frac{1}{2}\frac{\alpha_1^2(x)}{P_1(x)}r^4
\end{aligned} \tag{42}$$

The same derivation for $E\{(\frac{\Delta P_{A_2}(x)}{P_{A_2}(x)})^2\}$ results in (43).

$$\begin{aligned}
E\{(\frac{\Delta P_{A_2}(x)}{P_{A_2}(x)})^2\} \approxeq \frac{s_2}{NP_2(x)r^{d+1}} &+ e^t\epsilon_A\frac{s_2}{NP_2^2(x)r^{d+1}} \\
&+ \frac{1}{4}\alpha_2^2(x)r^4 + e^t\epsilon_A\frac{1}{2}\frac{\alpha_2^2(x)}{P_2(x)}r^4
\end{aligned} \tag{43}$$

According to Fukunaga, $E\{\Delta h_A(x)\}$ is approximately (44)

$$E\{\frac{\Delta P_{A_2}(x)}{P_{A_2}(x)}\} - \frac{1}{2}E\{(\frac{\Delta P_{A_2}(x)}{P_{A_2}(x)})^2\} - E\{\frac{\Delta P_{A_1}(x)}{P_{A_1}(x)}\} + \frac{1}{2}E\{(\frac{\Delta P_{A_1}(x)}{P_{A_1}(x)})^2\} - \Delta t \tag{44}$$

and $E\{\Delta h_A^2(x)\}$ is approximately (45) using the second order Taylor approximation.

$$\begin{aligned} &E\{(\frac{\Delta P_{A_2}(x)}{P_{A_2}(x)})^2\} + E\{(\frac{\Delta P_{A_1}(x)}{P_{A_1}(x)})^2\} - 2E\{(\frac{\Delta P_{A_1}(x)}{P_{A_1}(x)})(\frac{\Delta P_{A_2}(x)}{P_{A_2}(x)})\} + \Delta t^2 \\ &- 2\Delta t[E\{\frac{\Delta P_{A_2}(x)}{P_{A_2}(x)}\} - \frac{1}{2}E\{(\frac{\Delta P_{A_2}(x)}{P_{A_2}(x)})^2\} - E\{\frac{\Delta P_{A_1}(x)}{P_{A_1}(x)}\} + \frac{1}{2}E\{(\frac{\Delta P_{A_1}(x)}{P_{A_1}(x)})^2\}] \end{aligned} \tag{45}$$

Replacing the result of (37), (38), (42) and (43) in (44) and rewriting yield (46).

$$\begin{aligned} E\{\Delta h_A(x)\} &\approxeq \frac{r^2}{2}[\alpha_2(x) - \alpha_1(x)] + \frac{r^4}{8}(\alpha_1^2(x) - \alpha_2^2(x)) \\ &+ \frac{r^{-(d+1)}}{2N}[\frac{s_1}{P_1(x)} - \frac{s_2}{P_2(x)}] + \epsilon_A\frac{r^2}{2}[\frac{\alpha_1(x)}{P_1(x)} + e^t\frac{\alpha_2(x)}{P_2(x)}] \\ &- \epsilon_A\frac{r^4}{4}[\frac{\alpha_1^2(x)}{P_1(x)} + e^t\frac{\alpha_2^2(x)}{P_2(x)}] - \epsilon_A\frac{r^{-(d+1)}}{2N}[\frac{s_1}{P_1^2(x)} + e^t\frac{s_2}{P_2^2(x)}] \\ &= E\{\Delta h(x)\} + \epsilon_A\frac{r^2}{2}[\frac{\alpha_1(x)}{P_1(x)} + e^t\frac{\alpha_2(x)}{P_2(x)}] \\ &- \epsilon_A\frac{r^4}{4}[\frac{\alpha_1^2(x)}{P_1(x)} + e^t\frac{\alpha_2^2(x)}{P_2(x)}] - \epsilon_A\frac{r^{-(d+1)}}{2N}[\frac{s_1}{P_1^2(x)} + e^t\frac{s_2}{P_2^2(x)}] \end{aligned} \tag{46}$$

Note that the first three terms of approximation are $E\{\Delta h(x)\}$ according to Fukunaga [32] and the remaining terms with ϵ_A are the effect of ℓ-diversity. Replacing the result of (37), (38), (42) and (43) in (45) and rewriting yield (47).

$$\begin{aligned} E\{\Delta h_A^2(x)\} &\approxeq [\frac{1}{2}r^2(\alpha_2(x) - \alpha_1(x)) - \Delta t]^2 - \frac{\Delta t}{4}r^4(\alpha_1^2(x) - \alpha_2^2(x)) \\ &+ \frac{r^{-(d+1)}}{N}[\frac{s_1(1-\Delta t)}{P_1(x)} - \frac{s_2(1+\Delta t)}{P_2(x)}] - \epsilon_A\,\Delta t\,r^2[\frac{\alpha_1(x)}{P_1(x)} + e^t\frac{\alpha_2(x)}{P_2(x)}] \\ &+ \epsilon_A\frac{r^4}{2}[\frac{\alpha_1(x)\alpha_2(x)}{P_1(x)} - e^t\frac{\alpha_1(x)\alpha_2(x)}{P_2(x)}] \\ &- \epsilon_A\frac{r^4}{2}[\frac{\alpha_1^2(x)(1-\Delta t)}{P_1(x)} - e^t\frac{\alpha_2^2(x)(1+\Delta t)}{P_2(x)}] \\ &- \epsilon_A\frac{r^{-(d+1)}}{N}[\frac{(1-\Delta t)s_1}{P_1^2(x)} + e^t\frac{(1+\Delta t)s_2}{P_2^2(x)}] \\ &= E\{\Delta h^2(x)\} \\ &- \epsilon_A\,\Delta t\,r^2[\frac{\alpha_1(x)}{P_1(x)} + e^t\frac{\alpha_2(x)}{P_2(x)}] + \epsilon_A\frac{r^4}{2}[\frac{\alpha_1(x)\alpha_2(x)}{P_1(x)} - e^t\frac{\alpha_1(x)\alpha_2(x)}{P_2(x)}] \\ &- \epsilon_A\frac{r^4}{2}[\frac{\alpha_1^2(x)(1-\Delta t)}{P_1(x)} - e^t\frac{\alpha_2^2(x)(1+\Delta t)}{P_2(x)}] \\ &- \epsilon_A\frac{r^{-(d+1)}}{N}[\frac{(1-\Delta t)s_1}{P_1^2(x)} + e^t\frac{(1+\Delta t)s_2}{P_2^2(x)}] \end{aligned} \tag{47}$$

The first three terms of the approximation are $E\{\Delta h^2(x)\}$ according to Fukunaga [32] and the remaining terms with ϵ_A are again the effect of ℓ-diversity. Plugging the results of (46) and (47) in (36) and rewriting (36) give (26) where each a_i stands for an integration term. □

(26) shows that the anatomized training data D_A reduces the variance term of the non-parametric density classifier that estimates the Bayesian error. This explains the faster convergence of the anatomized k-NN classifier that was derived in the previous section. Given the original training data of finite size N, using the anatomized training data of finite size $N\ell$ reduces the search space of possible models. This means that the anatomized k-NN classifier considers less options of probabilistic models than the original k-NN classifier.

However, the bias term is increased which makes the non-parametric density classifier more susceptible to underfitting. In overall, the non-parametric density classifier on the anatomized training data has a shifted bias-variance trade-off relative to the non-parametric models on the original training data. The distortion of ℓ-diversity in the anatomization yields the following conclusion about the generalization ability of k-NN:

1. If the original k-NN classifier overfits, the anatomized k-NN classifier suffers less from overfitting provided that both classifiers have the same k hyper-parameter. In this case, the anatomized k-NN always generalizes better than the original k-NN.

2. If the original k-NN classifier fits well (optimum bias-variance tradeoff), the anatomized k-NN classifier always underfits provided that both classifiers have the same k hyper-parameter. In this case, the original k-NN generalizes better than the anatomized k-NN.

3. If the original k-NN classifier underfits, the anatomized k-NN classifier suffers more from underfitting provided that both classifiers have the same k hyper-parameter. In this case, the original k-NN generalizes better than the anatomized k-NN as well.

The key result to keep in mind is this: *ℓ-diversity regularizes the original k-NN classifier while it provides privacy*.

5 Experiments and Results

5.1 Prerequisites

5.1.1 Datasets

We tested our algorithm on the adult, IPUMS and marketing datasets of the UCI data repository [41] and the fatality dataset of the Keel data repository [42]:

1. **Adult:** The adult dataset is drawn from 1994 census data of the United States [41]. It is composed of 45222 instances after the removal of instances with missing values. The binary classification task is to predict whether a person's adjusted gross income is $\leq 50K$ or $> 50K$. The attribute *final weight* is ignored. *Education* is treated as the sensitive attribute in the experiments. The quasi-identifying attributes are *age*, *workclass*, *maritalstatus*, *occupation*, *race*, *sex*, *capitalgain*, *capitalloss*, *hoursperweek* and *nativecountry*. The class attribute is *income*.

2. **IPUMS:** This data is drawn from the 1970, 1980 and 1990 census data of the Los Angeles and Long Beach areas [41]. It has 233584 instances in total. We pick the 10 attributes that are included in the adult data. The binary classification task is to predict whether a person's total income is $\leq 50K$ or $> 50K$. The classifiers are expected to show a different behavior from the former adult data since the population (and to some extent, classification task, as it is total income rather than adjusted gross income) are different. *Educrec* is treated as the sensitive attribute in the experiments. The quasi-identifying attributes are *age, sex, raceg, marst, occ1950, classwkg, hrswork2, migplac5* and *vetstat*. The class attribute is *bintotinc*, a binary attribute that we created from the totinc (total income) of the original dataset based on the former binary classification task.

3. **Marketing Data:** This data is drawn from a phone based marketing campaign of a Portuguese banking institution for long term deposits [41]. We created the following binary classification task which is linearly separable: "among all the people who didn't submit a long term deposit, predict whether a person has a housing loan or not". We performed the following preprocessing using Weka filters [43]: 1) pick 39922 instances who didn't make a long term deposit 2) choose four attributes *job*, *day*, *month* and *age* using the correlation with the class attribute *housing*. Discretized *age* is treated as the sensitive attribute whereas the quasi-identifying attributes are *job*, *day* and *month*.

4. **Fatality Data:** This data is a U.S. National Center for Statistics and Analysis compilation of 2001 car accidents. The original class attribute *injury_severity* has eight labels indicating the level of injury suffered [42]. We create the binary attribute *is_injured* with values "Injured" and "No_Injury" in the following way: 1) remove the instances with labels "Injured_Severity_Unknown", "Died_Prior_to_Accident", "Unknown" and "Possible_Injury' from the original data. This results in 91085 instances 2) label "Injured" the instances with labels "Nonincapaciting_Evident_Injury", "Incapaciting_Injury" and "Fatal_Injury". No feature selection is applied on this dataset. *police_reported_alcohol_involvement* is treated as the sensitive attribute. The remaining attributes in the data catalog are the quasi-identifying attributes [42]

5.1.2 Privacy Setup

The anatomization was done according to Xiao et al.'s bucketization algorithm [11]. When the ℓ-diversity condition is not satisfied, the instances were divided into groups of size ℓ according to the original bucketization algorithm. Leftover instances were suppressed (not used in training models.) Although this violates the assumption in the theoretical analysis, we believe such experiments will still be useful to show whether the theoretical analysis under the existing assumption is verifiable or not.

Anonymized training data was created for the adult dataset. The anonymized k-NN is not included for other datasets since Inan et al. provided generalization hierarchies only in the adult dataset [21]. Hence, we used Inan et al.'s value generalization hierarchies in the experiments. The privacy parameters were $k = \ell$ for k-anonymity and ℓ-diversity to compare the classifiers using same group sizes in training data.

Anonymized and anatomized training data had the same identifying and sensitive attributes. The sensitive attributes were chosen such that the ℓ-diversity is satisfied for at least $\ell = 2$.

5.1.3 Model Evaluation Setup

Weka's IBk class was used to train k-NN classifiers on the original, identifying and anatomized training data [43]. The anatomized training data is created from the IT and ST tables using the merging and dropping functions of Pandas [44].

10-fold cross validation was used for evaluation of the error rate bounds and generalization ability (we use boxplots to show variance across the different folds), and the error rate was used as the evaluation metric. The comparison includes anatomized k-NN, original k-NN and identifying k-NN. The comparison on adult dataset also includes anonymized k-NN due to the privacy setup in the previous section. The error rates of anatomized and original k-NN are compared using the Student t-test. Other models are not included in Student t-test, because Theorem 23 covers only anatomized and original k-NN.

10-fold cross validation was also used for evaluation of the 1-NN classifier's convergence rate. The error rate was again used for the evaluation metric. However, we trained the anatomized and original 1-NN classifiers incrementally at each iteration of the cross validation. In a given iteration, the training set was divided into 9 partitions and the models are trained 9 times. The training started from the first partition and continued further by adding a partition at a time. The average error rates are computed over 10 different error rate values for a given training set size in the analysis (cf. Section 5.2.6.)

5.2 Analysis of Results

While Figures 3 through 8 show the boxplots of error rates for k-NN classifiers, Figure 9 shows the lines of convergence rate for the error rates of original and anatomized 1-NN classifiers. In Figures 3 to 8, "Org." and "Id." labels will stand for original k-NN and identifying k-NN respectively. The anatomized and anonymized k-NN will be represented by their respective privacy parameters (L for ℓ and k for k.) Our analysis have four observation aspects:

1. **Comparison between the anatomized and the original k-NN:** From Theorem 23, there are two possibilities for the error rates. In first possibility, the anatomized k-NN classifier is effected from overfitting less than the original k-NN classifier. In this case, we expect that the anatomized k-NN has smaller error rate than the original k-NN on average. In second possibility, the anatomized k-NN suffers from underfitting while the original k-NN classifier either fits well or suffers less from underfitting. In this case, the anatomized k-NN's error rate is expected to be greater than the original k-NN's. Increasing the ℓ parameter would result in the increase of distortion (ϵ_A in (26)). From Theorem 23, the error rate expectation in the former two possibilities would still be valid in function of the increase in ℓ unless there is suppression. If some instances are suppressed to create the anatomized training data, then the theoretical analysis of the anatomized k-NN classifier would be invalid because the assumption of Theorem 23 is violated.

2. **Comparison between anatomized and identifying k-NN:** In the first case, the identifying k-NN is likely to outperform the anatomized k-NN if the sensitive attribute is a bad predictor of the class attribute in the original training data. The sensitive attribute changes the likelihood probability in density based decision function and the model either overfits by the increase in variance or underfits by the increase in the bias. The anatomized k-NN therefore estimates a model of the original training data that is not likely to generalize well. In the second case, the anatomized k-NN are likely to

outperform the identifying k-NN if the sensitive attribute is a good predictor of the class attribute in the original training data. The sensitive attribute changes again the likelihood probability in density based decision function and the anatomized k-NN classifier catches a better tradeoff between the bias and the variance terms. The anatomized k-NN classifier would avoid the potential overfitting or underfitting that the identifying k-NN could have.

3. **Comparison between the anatomized and the anonymized k-NN:** The anatomized k-NN are expected to outperform the anonymized k-NN because anatomization preserves the original values for all the attributes. The generalization based k-anonymity, on the other hand, distorts most of the original attribute values [21].

4. **Comparison of the anatomized and original 1-NN in the convergence rates:** From Theorem 18, we expect that the anatomized 1-NN will converge faster than the original 1-NN classifier to the lowest possible error rate if there is no suppression in the creation of the anatomized training data. Suppression would again violate the assumption of Theorem 18.

We should note that it we are comparing an anatomy method satisfying ℓ-diversity against a weaker k-anonymity requirement for the generalization-based approach. While we don't specifically use an ℓ-diversity based generalization [1], such a method would be expected to generalize more and give worse results. The generalization-based k-anonymization we use already produces 2-diverse datasets. More than 44000 instances in the adult dataset are 11-diverse for k-values 2 to 5. As 97% of the generalization based k-anonymized adult dataset satisfy ℓ-diversity, the results should be similar with generalization based ℓ-diversity.

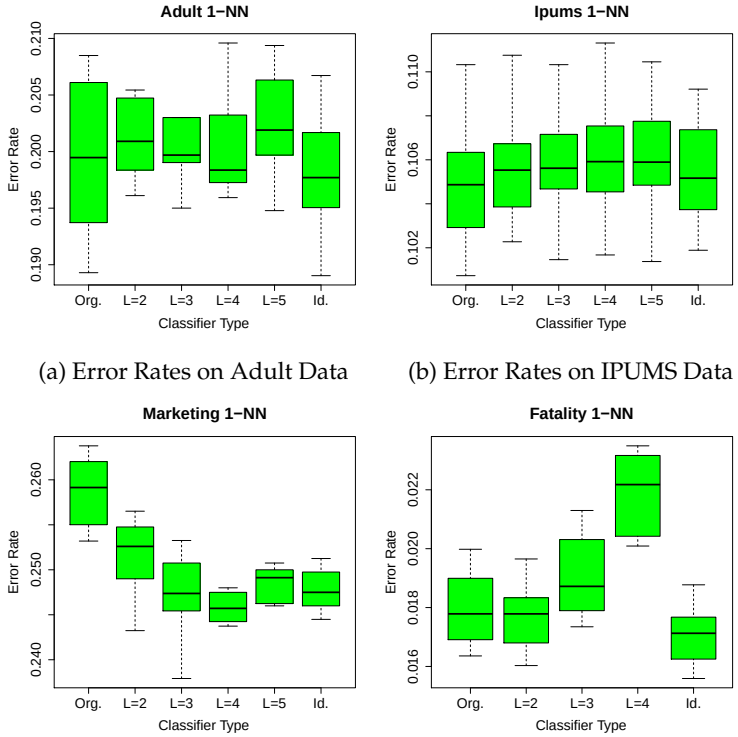
We are now presenting the analysis of error rates in first three aspects for all the datasets. We will then present the analysis of convergence rate for anatomized 1-NN in the fourth aspect.

5.2.1 Analysis of Error Rates for 1-NN

Figure 3 shows the error rates for the original, anatomized and identifying 1-NN classifiers that are tested on the four datasets.

In Figure 3a, the original 1-NN classifier outperforms the anatomized 1-NN classifier when ℓ is 2 and 3. The anatomized 1-NN classifier hence underfits more than the original 1-NN due to the distortion of ℓ-diversity. In the second aspect, the sensitive attribute is a bad predictor of the class attribute because the average error rate of the identifying 1-NN classifier is less than the original 1-NN's. The anatomized 1-NN classifier's error rates are greater than both the original and identifying 1-NN classifier's because the anatomized 1-NN classifiers are estimating an original 1-NN classifier that doesn't fit well. The former claims hold when ℓ is 4 and 5 despite suppression.

In Figure 3b, the original 1-NN outperforms the anatomized 1-NN when ℓ is 2-to-4. Theorem 23 concludes that the anatomized 1-NN classifier underfits more than the original 1-NN classifier. In the second aspect, the sensitive attribute is a bad predictor of the class attribute because the average error rate of the identifying and original 1-NN classifiers are almost same. The anatomized 1-NN classifier's error rates thus are greater than both the original and identifying 1-NN classifier's because the anatomized 1-NN classifiers are estimating an original 1-NN classifier that doesn't fit well. When ℓ is 5, the former claims hold although the assumptions of our theoretical analysis is violated due to suppression.



(a) Error Rates on Adult Data (b) Error Rates on IPUMS Data

(c) Error Rates on Marketing Data (d) Error Rates on Fatality Data

Figure 3: 10 Cross Validation Errors Rates for 1-NN Classifiers

In Figure 3c, the original 1-NN classifier has higher error rates than the anatomized 1-NN classifiers when ℓ is 2-to-5. From Theorem 23, the original 1-NN classifier suffers from overfitting whereas the anatomized 1-NN classifier catches a better bias variance tradeoff. Note that when ℓ is increased from 4 to 5, the error rates start increasing. Increase in ℓ causes here the increase in the bias such that the anatomized 1-NN classifier starts underfitting. In the second aspect, the identifying 1-NN classifier has a much lower error rate than the original 1-NN classifier has. This means that the sensitive attribute is a bad predictor of the class attribute. Since sensitive attribute is a bad predictor, the anatomized 1-NN classifier approximates a bad classifier of the original training data which increases its generalization error relative to the identifying 1-NN classifier.

In Figure 3d, the original 1-NN classifier has higher error rates than the anatomized 1-NN classifiers when ℓ is 2. From Theorem 23, the original 1-NN classifier suffers from overfitting whereas the anatomized 1-NN classifier captures a better bias variance tradeoff. When ℓ is 3 or 4, the assumptions of the theoretical analysis is violated due to suppression. The general conclusion here would be the significant distortion of the likelihood probabilities which increases the bias term (underfitting). In the second aspect, the identifying 1-NN classifier's error rates are less than the original 1-NN classifier's. The sensitive attribute hence is a bad predictor of the class attribute and the anatomized 1-NN classifier approximates a bad classifier which increases its generalization error relative to identifying 1-NN.

Figure 4 gives the boxplots of error rates for anonymized k-NN classifiers in addition to

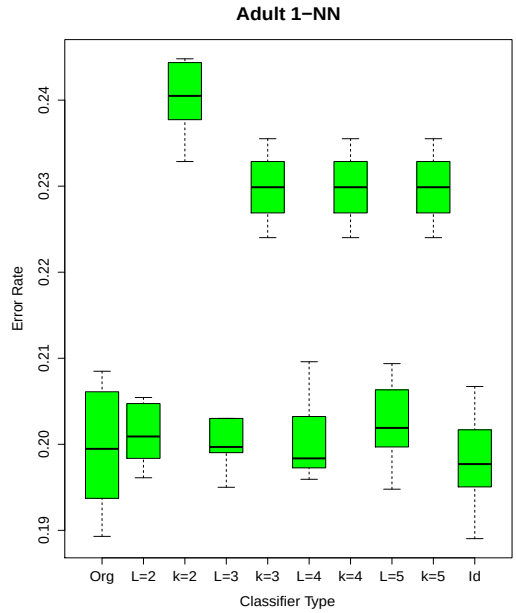
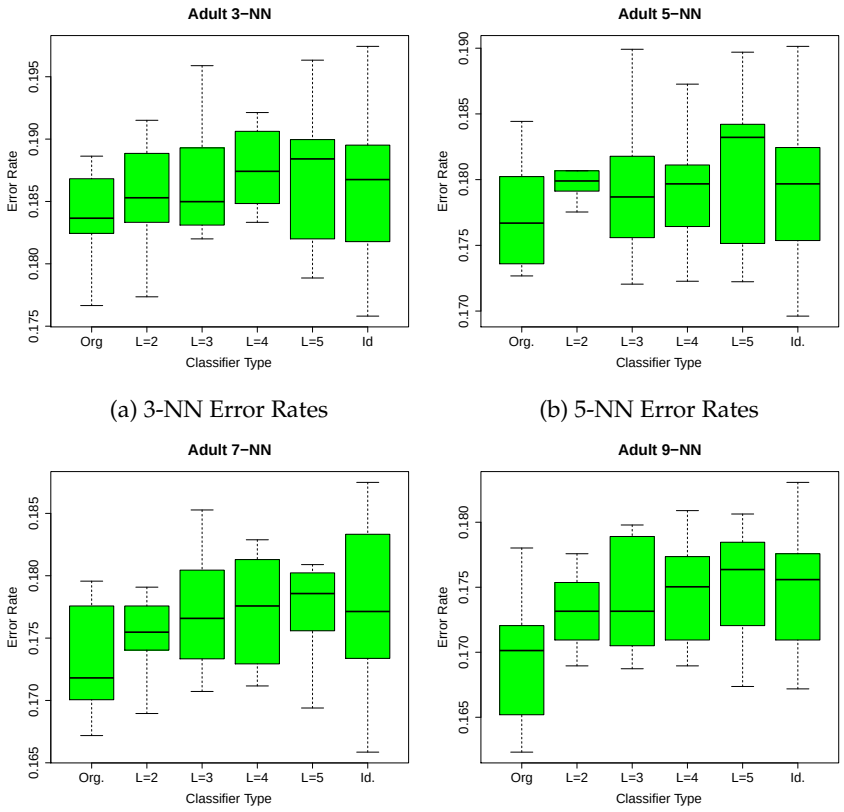


Figure 4: 1-NN Error Rates on Adult (k for k and L for ℓ)

original, identifying and anatomized 1-NN. In the third aspect, we have the expected result for the anonymized 1-NN classifiers. Their error rates are greater than the anatomized 1-NN classifiers. Due to generalization of the identifying attribute values, the utility loss of the anonymized training data is more than the anatomized training data's. In the context of anonymized 1-NN, note that the error rates are decreased when k is increased. Increasing k for the anonymized 1-NN classifier is same as increasing the number of neighbors for the k-NN classifier.

5.2.2 Analysis of Error Rates for k-NN on Adult Data

Figure 5 shows the results of multiple types of k-NN classifiers on the adult data for 3, 5, 7 and 9 neighbors (k). In the first aspect, we start discussion with $\ell = 2$ and $\ell = 3$ (no suppression cases.) The anatomized k-NN classifiers have higher error rates than the original k-NN classifiers. Since Theorem 23 tells that the anatomized training data increases the bias terms of the k-NN classifiers' generalization error while reducing the variance, the anatomized k-NN classifiers suffer from underfitting. Although the former Theorem's assumptions are violated in cases of $\ell = 4$ and $\ell = 5$, we see that the error rates of anatomized k-NN classifiers are still higher than the original k-NN classifiers'. We thus can assume that the same underfitting behavior continues. In the second aspect, the sensitive attribute is a good predictor of the class attribute because the original 3-NN, 5-NN, 7-NN and 9-NN classifiers outperform the identifying 3-NN, 5-NN, 7-NN and 9-NN classifiers in terms of the error rates. When $\ell = 2$ and $\ell = 3$, the anatomized k-NN classifier captures the bias variance tradeoff between the identifying and original k-NN classifiers for multiple values of k. The anatomized k-NN classifiers outperforms the identifying k-NN classifiers while it is outperformed by the original k-NN classifiers. Due to suppression, the former bias



(a) 3-NN Error Rates

(b) 5-NN Error Rates

(c) 7-NN Error Rates

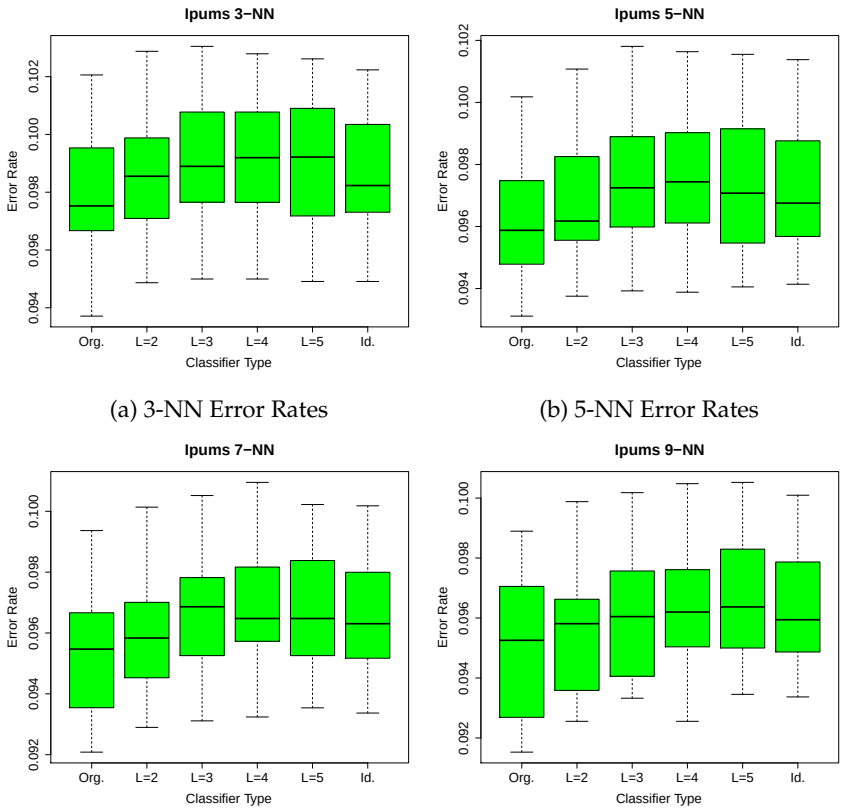
(d) 9-NN Error Rates

Figure 5: 10 Cross Validation Errors Rates of Adult Data

variance tradeoff conditions don't hold in cases of $\ell = 4$ and $\ell = 5$. The anatomized k-NN classifiers thus don't have the former tradeoff.

5.2.3 Analysis of Error Rates for k-NN on IPUMS Data

Figure 6 shows the results for multiple types of k-NN classifiers on the IPUMS data. The IPUMS data satisfy the ℓ-diversity condition when $\ell = 2$, $\ell = 3$ and $\ell = 4$, so the Theorem 23 and the bound 6 are expected to hold in here. In the first aspect, the anatomized 3-NN, 5-NN, 7-NN and 9-NN classifiers are outperformed by the original 3-NN, 5-NN, 7-NN and 9-NN classifiers. From Theorem 23, the increase in the bias terms yields the underfitting classifiers. In the second aspect, the sensitive attribute is good predictor of the class attribute because the original k-NN classifiers outperform the identifying k-NN classifier for multiple values of hyperparameter k. When $\ell = 3$ and $\ell = 4$, the anatomized k-NN classifiers surprisingly fail to capture the bias variance tradeoff between the original and the identifying k-NN classifiers. From Theorem 23, the increase in bias is way greater to capture bias variance tradeoff for the anatomized k-NN classifiers (cf. Figure 6.) The case of $\ell = 2$ is special. Note that the anatomized 5-NN, 7-NN and 9-NN classifiers capture the bias variance trade-off between the original and identifying ones, as expected (cf. Figures



(a) 3-NN Error Rates

(b) 5-NN Error Rates

(c) 7-NN Error Rates

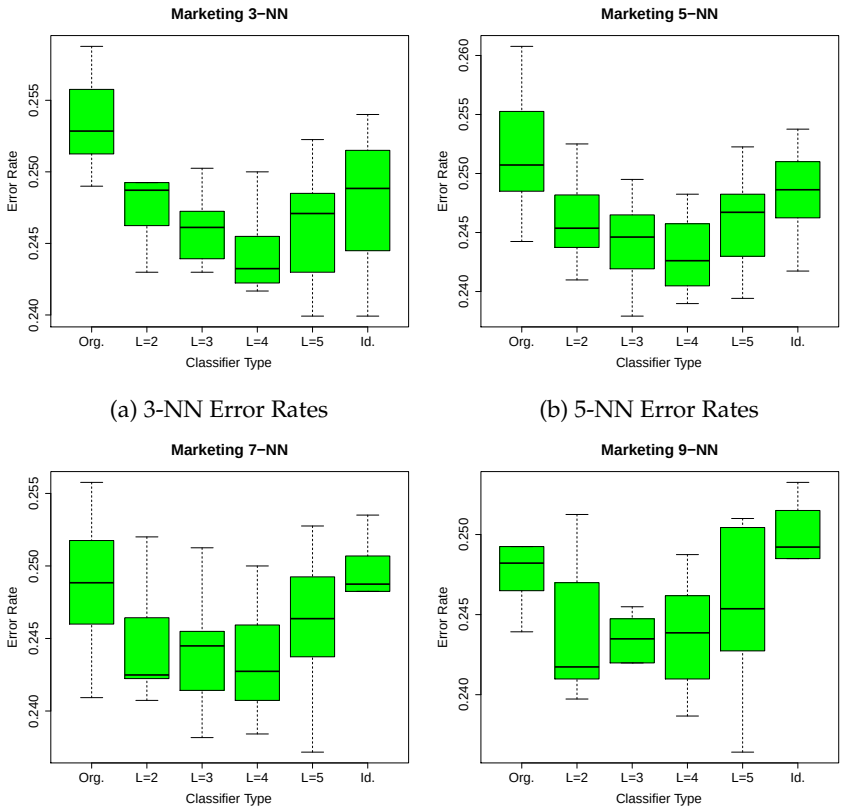
(d) 9-NN Error Rates

Figure 6: 10 Cross Validation Errors Rates of Ipums Data

6b, 6c and 6d.) For anatomized 3-NN classifier, the increase in bias is way greater than the decrease in variance. It thus fails to capture again the bias variance tradeoff between the original 3-NN and the identifying 3-NN classifiers (cf. Figure 6a.) Last, the anatomized 3-NN, 5-NN, 7-NN and 9-NN classifiers under $\ell = 5$ (suppression) show a similar trend to the anatomized 3-NN, 5-NN, 7-NN and 9-NN classifiers under $\ell = 3$ and $\ell = 4$ in terms of the first and second aspects.

5.2.4 Analysis of Error Rates for k-NN on Marketing Data

Figure 7 shows the result for multiple types of k-NN classifiers on the marketing data. The marketing data satisfy the ℓ-diversity condition when ℓ is 2-to-5. Theorem 23 and the bound 6 thus are expected to hold in all ℓ values. In the first aspect, the anatomized k-NN classifiers outperform the original k-NN classifiers for all combinations of ℓ and hyperparameter k. This shows that the distortion of ℓ-diversity reduces the generalization error by increasing the bias and reducing the variance of the original k-NN classifier (fixing the overfitting issue.) Note that when ℓ is increased from 4 to 5, the overfitting issue is fixed less because the increase in bias exceeded the good bias variance tradeoff and the model is directed to the underfitting case. In the second aspect, the sensitive attribute is a good predictor of



(a) 3-NN Error Rates

(b) 5-NN Error Rates

(c) 7-NN Error Rates

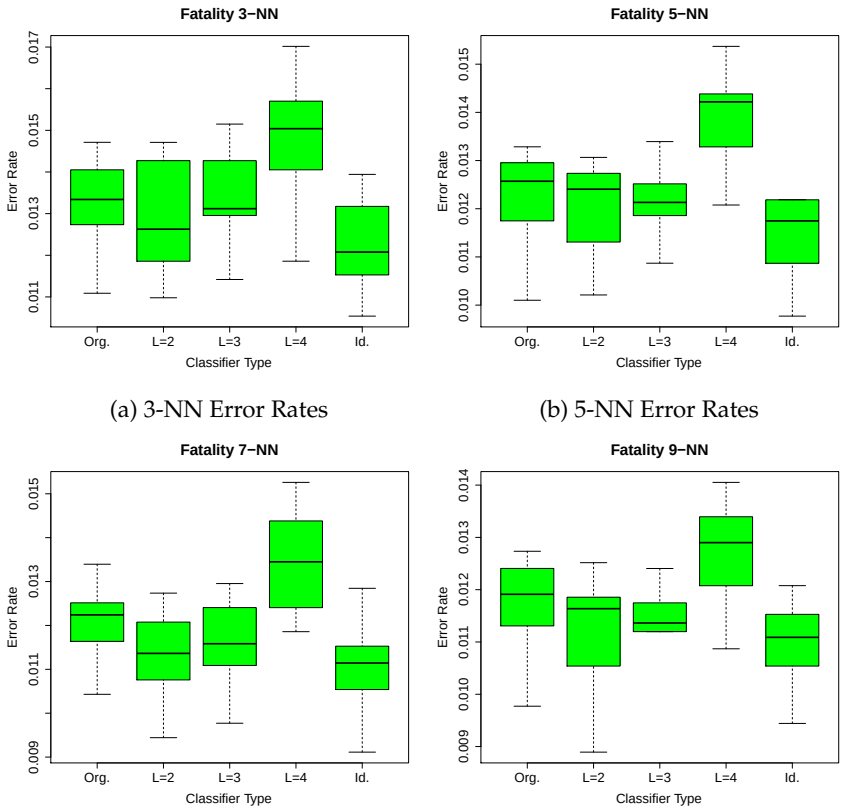
(d) 9-NN Error Rates

Figure 7: 10 Cross Validation Errors Rates of Marketing Data

7-NN and 9-NN classifiers since the original 7-NN and 9-NN's error rates are less than the identifying 7-NN and 9-NN's (cf. Figures 7c and 7d.) Surprisingly, the anatomized 7-NN and 9-NN's error rate are lower than both the original and the identifying 7-NN and 9-NN classifiers. The sensitive attribute, on the other hand, is a bad predictor of 3-NN and 5-NN classifiers since the original 3-NN and 5-NN's error rates are greater than the identifying 3-NN and 5-NN's (cf. Figures 7a and 7b.) Surprisingly, the anatomized 3-NN and 5-NN classifier have again lower error rates than both original and identifying 3-NN and 5-NN classifiers. The most plausible reason is the regularization effect of ℓ-diversity that results in the lowest generalization error of the anatomized k-NN classifiers for all ℓ-values. The distortion of ℓ-diversity increases the bias such that it fixes the overfitting issue of both the original and identifying k-NN classifiers.

5.2.5 Analysis of Error Rates for k-NN on Fatality Data

Figure 8 shows the result for multiple types of k-NN classifiers on the fatality data. As the fatality data satisfy the ℓ-diversity condition for $\ell = 2$, Theorem 23 thus is expected to hold in the first aspect. Anatomized k-NN classifiers outperform the original k-NN classifiers when $\ell = 2$. From Theorem 23, the increase in bias reduces the generalization error of



(a) 3-NN Error Rates

(b) 5-NN Error Rates

(c) 7-NN Error Rates

(d) 9-NN Error Rates

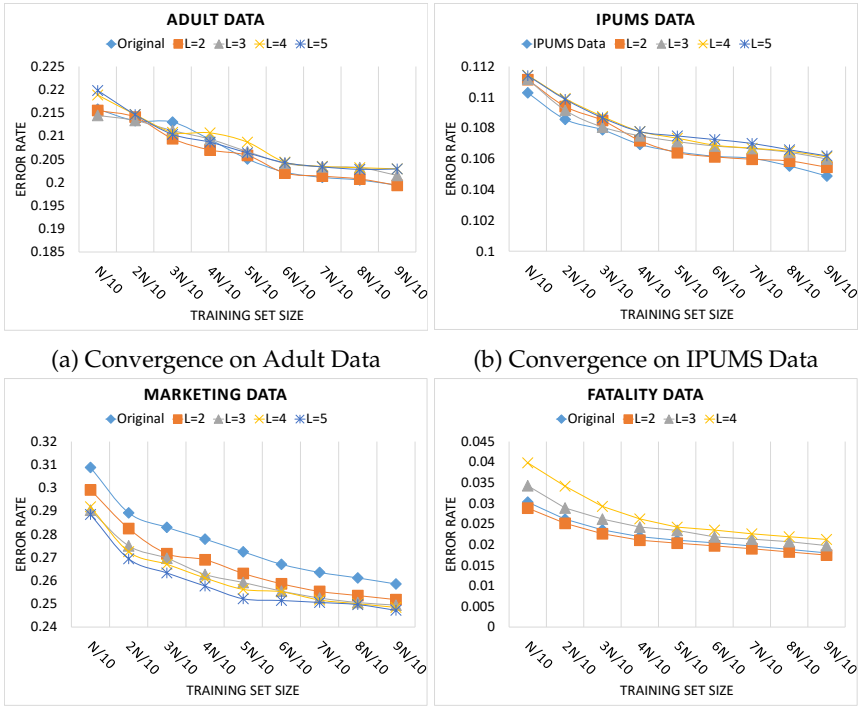
Figure 8: 10 Cross Validation Errors Rates of Fatality Data

the original k-NN classifiers which is overfitting to the original training data. In the second aspect, the sensitive attribute is a bad predictor of k-NN classifiers since the original k-NN's error rates are greater than the identifying k-NN's. Although the error rates of anatomized k-NN is less than the original k-NN's, its error rates are greater than the identifying k-NN's. This is expected since the anatomized k-NN is trying to capture the bias variance tradeoff between the original and the identifying k-NN classifiers. The anatomized k-NN classifier is estimating the original k-NN classifier's distribution which is not the best classifier in the existing data. Last, increasing ℓ to 3 and 4 increases the error rates of the anatomized k-NN classifiers. Due to suppression, the theoretical analysis do not hold here. The most plausible reason is that the reduction of the training data size results in the overfitting of the models. This would increase the generalization error.

5.2.6 Analysis of Convergence Rates

Figure 9 shows the error rates of original and anatomized 1-NN classifiers as a function of the increasing training set sizes.

In Figure 9a, we see that the anatomized 1-NN classifier converges faster than the original 1-NN classifier on the adult data when $\ell = 2$ and $\ell = 3$, as expected from Theorem 18.



(a) Convergence on Adult Data (b) Convergence on IPUMS Data

(c) Convergence on Marketing Data (d) Convergence on Fatality Data

Figure 9: Convergence of 1-NN Using Multiple Size Training Sets

The former ℓ values satisfy the ℓ-diversity condition, so there is no suppression and the assumptions of the Theorem 18 are not violated. Note that the anatomized 1-NN classifier converges slower than the original 1-NN classifier when the assumption of Theorem 18 is violated under $\ell = 4$ and $\ell = 5$ (due to suppression.)

In Figure 9b, we see that the anatomized 1-NN classifier converges faster than the original 1-NN classifier on the IPUMS data when $\ell = 2$, $\ell = 3$ and $\ell = 4$, as expected from Theorem 18. The former ℓ values satisfy the ℓ-diversity condition, so there is no suppression and the assumptions of the Theorem 18 are not violated. Note that the anatomized 1-NN classifier still converges faster than the original 1-NN classifier when the assumption of Theorem 18 is violated under $\ell = 5$. The most likely reason behind is that the number of instances in the IPUMS data is too large and the number of suppressed instances are negligible relative to its size.

In Figure 9c, we see that the anatomized 1-NN classifier converges faster than the original 1-NN classifier on the marketing data for all values of ℓ, as expected from Theorem 18. The marketing data satisfy the ℓ-diversity condition for ℓ values 2-to-5. Hence, the assumptions of the Theorem 18 are never violated in the experiments.

Last, Figure 9d shows that the anatomized 1-NN classifier converges faster than the original 1-NN classifier on the fatality data under $\ell = 2$. This is again expected from Theorem 18 since the fatality data satisfy the ℓ-diversity condition for $\ell = 2$. It is easy to notice that the convergence of anatomized 1-NN classifier is slower than the original 1-NN classifier under $\ell = 3$ and $\ell = 4$ due to suppression, as expected.

Classifier	Org. vs ℓ=2	Org. vs ℓ=3	Org. vs ℓ=4	Org. vs ℓ=5
1-NN	F	F	F	F
3-NN	F	P	P	P
5-NN	P	F	F	P
7-NN	F	P	P	P
9-NN	P	P	P	P

Table 1: Anatomized k-NN vs Original k-NN on Adult

Classifier	Org. vs ℓ=2	Org. vs ℓ=3	Org. vs ℓ=4	Org. vs ℓ=5
1-NN	P	P	P	P
3-NN	P	P	P	P
5-NN	P	P	P	P
7-NN	P	P	P	P
9-NN	P	P	P	P

Table 2: Anatomized k-NN vs Original k-NN on IPUMS

Classifier	Org. vs ℓ=2	Org. vs ℓ=3	Org. vs ℓ=4	Org. vs ℓ=5
1-NN	P	P	P	P
3-NN	P	P	P	P
5-NN	P	P	P	P
7-NN	F	F	F	F
9-NN	P	F	F	F

Table 3: Anatomized k-NN vs Original k-NN on Marketing

Classifier	Org. vs ℓ=2	Org. vs ℓ=3	Org. vs ℓ=4
1-NN	F	P	P
3-NN	F	F	P
5-NN	F	F	P
7-NN	P	F	P
9-NN	P	F	F

Table 4: Anatomized k-NN vs Original k-NN on Fatality

5.3 Student t-test for Anatomized k-NN versus Original k-NN

Tables 1, 2, 3 and 4 give the statistical test results for *confidence interval 0.95*. In all Tables, "P" stands for pass while "F" stands for fail. "N/A" stands for not applicable in cases where the domain size of sensitive attribute is less than the ℓ value. "Org." stand for the original k-NN whereas "ℓ" stand for the anatomized k-NN. Note that we do the test for original k-NN vs anatomized k-NN, because the Theorem 23's scope covers this analysis.

From the Tables, the combinations of ℓ and hyperparameter k give at least 1 statistically

significant comparison when there is no suppression in the creation of the anatomized training data. There is at least one statistically significant comparison when the theoretical analysis is supposed to hold under the ℓ-diversity condition.

6 Conclusion and Future Directions

This work demonstrates the feasibility of k-NN classification using training data protected by anatomization under ℓ-diversity. We show that the asymptotic error bounds are the same for anatomized data as for the original data. Perhaps surprisingly, the proposed 1-NN classifier has a faster convergence to the asymptotic error rate than the convergence of 1-NN classifier using the training data without anatomization. In addition, the analysis suggests that any non-parametric classifier using the anatomized training data has the variance term of generalization error that is less than the non-parametric classifiers' using the original training data. In contradiction, any non-parametric classifier using the anatomized training data has the bias terms of generalization error that are greater than the non-parametric classifiers' using the original training data. The anatomized training data thus pushes the optimum point of bias variance tradeoff towards the bias terms.

Experiments on multiple datasets confirm the theoretical convergence rates. These experiments also demonstrate that proposed k-NN on anatomized data approaches or even outperforms k-NN on original data. In particular, the experiments on well known Adult data show that 1-NN on anatomized data outperforms learning on data anonymized to the same anonymity levels using generalization.

References

[1] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. ℓ-diversity: Privacy beyond k-anonymity. *Proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE 2006)*, April 2006.

[2] Pierangela Samarati. Protecting respondents' identities in microdata release. *Journal of IEEE Transaction on Knowledge Data Engineering*, 13(6):1010–1027, 2001.

[3] Latanya Sweeney. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002.

[4] Ninghui Li and Tiancheng Li. t-closeness: Privacy beyond k-anonymity and ℓ-diversity. *Proceedings of the 23rd International Conference on Data Engineering (ICDE 2007)*, April 16–20 2007.

[5] Mehmet Ercan Nergiz and Christopher W. Clifton. δ-presence without complete world knowledge. *Journal of IEEE Transactions on Knowledge Data Engineering*, 22(6):868–883, 2010.

[6] Raymond Chi-Wing Wong, Jiuyong Li, Ada Wai-Chee Fu, and Ke Wang. (α, k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*, pages 754–759, 2006.

[7] Richard A. Moore, Jr. Controlled data-swapping techniques for masking public use microdata sets. Statistical Research Division Report Series RR 96-04, U.S. Bureau of the Census, Washington, DC., 1996.

[8] Rakesh Agrawal and Ramakrishnan Srikant. Privacy-preserving data mining. *Proceedings of the 2000 ACM SIGMOD Conference on Management of Data*, pages 439–450, May 14–19 2000.

[9] Alexandre Evfimievski, Johannes Gehrke, and Ramakrishnan Srikant. Limiting privacy breaches in privacy preserving data mining. *Proceedings of the 22nd ACM SIGACT-SIGMOD-*

SIGART Symposium on Principles of Database Systems (PODS 2003), pages 211–222, June 9–12 2003.

[10] Cynthia Dwork. Differential privacy. *33rd International Colloquium on Automata, Languages and Programming (ICALP 2006)*, pages 1–12, July 9–16 2006.

[11] Xiaokui Xiao and Yufei Tao. Anatomy: Simple and effective privacy preservation. *Proceedings of 32nd International Conference on Very Large Data Bases (VLDB 2006)*, September 12–15 2006.

[12] Valentina Ciriani, Sara Vimercati, Sabrina De Capitani di Vimercati, Sushil Jajodia, Stefano Paraboschi, and Pierangela Samarati. Combining fragmentation and encryption to protect privacy in data storage. *Journal of ACM Transactions on Information and System Security*, 13:22:1–22:33, July 2010.

[13] Sabrina De Capitani di Vimercati, Sara Foresti, Sushil Jajodia, Giovanni Livraga, Stefano Paraboschi, and Pierangela Samarati. Extending loose associations to multiple fragments. *Proceedings of 27th Annual IFIP WG 11.3 Working Conference on Data and Applications Security and Privacy (DBSec 2013)*, pages 1–16, 2013.

[14] Tamas Gal, Zhiyuan Chen, and Aryya Gangopadhyay. A privacy protection model for patient data with multiple sensitive attributes. *International Journal of Information Security and Privacy*, 2(3):28–44, 2008.

[15] Daniel Kifer. Attacks on privacy and definetti's theorem. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 127–138, June 29 –July 2 2009.

[16] Xianmang He, Yanghua Xiao, Yujia Li, Qing Wang, Wei Wang, and Baile Shi. Permutation anonymization: Improving anatomy for privacy preservation in data publication. *The 15th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) Workshops*, pages 111–123, 2011.

[17] Tiancheng Li, Ninghui Li, Jian Zhang, and Ian Molloy. Slicing: A new approach for privacy preserving data publishing. *Journal of IEEE Transactions on Knowledge and Data Engineering*, 24(3):561–574, 2012.

[18] Tiancheng Li and Ninghui Li. On the tradeoff between privacy and utility in data publishing. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 517–526, June 28–July 1 2009.

[19] Raymond Chi-Wing Wong, Ada Wai-Chee Fu, Ke Wang, and Jian Pei. Minimality attack in privacy preserving data publishing. *Proceedings of 33rd International Conference on Very Large Data Bases (VLDB 2007)*, pages 543–554, 2007.

[20] Graham Cormode, Ninghui Li, Tiancheng Li, and Divesh Srivastava. Minimizing minimality and maximizing utility: Analyzing method-based attacks on anonymized data. *Journal of the Very Large Data Bases Endowment*, 3(1):1045–1056, 2010.

[21] Ali Inan, Murat Kantarcioglu, and Elisa Bertino. Using anonymized data for classification. *Proceedings of the 2009 IEEE International Conference on Data Engineering (ICDE 2009)*, pages 429–440, 2009.

[22] Brent Martin. Instance-based learning : Nearest neighbor with generalization. Technical report, University of Waikato, Department of Computer Science, 1995.

[23] Vijay S. Iyengar. Transforming data to satisfy privacy constraints. *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 279–288, July 23–26 2002.

[24] Jun Zhang, Dae-Ki Kang, Adrian Silvescu, and Vasant Honavar. Learning accurate and concise naïve bayes classifiers from attribute value taxonomies and data. *Journal of Knowledge and Information Systems*, 9(2):157–179, 2006.

[25] Dakshi Agrawal and Charu C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. *Proceedings of the 20th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 247–255, May 21–23 2001.

[26] Benjamin C. M. Fung, Ke Wang, and Philip S. Yu. Top-down specialization for information and privacy preservation. *Proceedings of the 21st International Conference on Data Engineering (ICDE 2005)*, pages 205–216, 2005.

[27] Jim Dowd, Shouhuai Xu, and Weining Zhang. Privacy-preserving decision tree mining based on random substitutions. *Proceedings of International Conference on Emerging Trends in Information and Communication Security*, 2005.

[28] Benjamin IP Rubinstein, Peter L Bartlett, Ling Huang, and Nina Taft. Learning in a large function space: Privacy-preserving mechanisms for svm learning. *Journal of Privacy and Confidentiality*, 4(1):65–100, 2012.

[29] Keng-Pei Lin and Ming-Syan Chen. Privacy-preserving outsourcing support vector machines with random transformation. *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2010)*, pages 363–372, 2010.

[30] Prateek Jain and Abhradeep Thakurta. Differentially private learning with kernels. *Proceedings of the 30th International Conference on Machine Learning*, 28(3):118–126, 17–19 Jun 2013.

[31] Vladimir Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

[32] Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition (2nd Edition)*. Academic Press Professional, Inc., San Diego, CA, USA, 1990.

[33] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining, (1st Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005.

[34] Thomas M Cover and Peter E Hart. Nearest neighbor pattern classification. *Journal of IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[35] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

[36] Raghu Ramakrishnan and Johannes Gehrke. *Database management systems*. McGraw-Hill, 2000.

[37] Alexandr Andoni and Piotr Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 459–468, 2006.

[38] Keinosuke Fukunaga and Donald M Hummels. Bias of nearest neighbor error estimates. *Journal of IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(1):103–112, 1987.

[39] András Antos, Luc Devroye, and László Györfi. Lower bounds for bayes error estimation. *Journal of IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(7):643–645, 1999.

[40] Keinosuke Fukunaga and Donald M Hummels. Bayes error estimation using parzen and k-nn procedures. *Journal of IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(5):634–643, 1987.

[41] M. Lichman. UCI machine learning repository, 2013.

[42] J Alcalá, A Fernández, J Luengo, J Derrac, S García, L Sánchez, and F Herrera. Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic and Soft Computing*, 17:255–287, 2010.

[43] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco, CA, USA, October 1999.

[44] Wes McKinney. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. O'Reilly Media Inc., 2012.