

Privacy Preservation through Random Non-linear Data Distortion

Abstract

Consider a scenario in which the data owner has some private/sensitive data and wants a data miner to access it for studying “important” patterns without revealing the sensitive information. Privacy preserving data mining aims to solve this problem by randomly transforming (distorting) the data prior to its release. Previous work only considered the case of linear distortions — additive, multiplicative or a combination of both — for studying the usefulness of the distorted output and the privacy preserved. In this paper, we consider a general class of potentially non-linear transformations of the data. We develop bounds on the expected accuracy of our non-linear distortion and also quantify privacy by using standard definitions. We show how our general transformation can be used in practice for two specific problem instances: a linear model and a popular non-linear model *viz.* neural network. The paper presents a thorough theoretical analysis of the transformation and possible applications. Experiments conducted on real-life datasets demonstrate the effectiveness of the approach.

1 Introduction

The first part of the paper talks about distance and the second part talks about privacy.

2 Related Work

Data perturbation-based privacy preserving techniques perturb data elements or attributes directly by either additive noise, multiplicative noise or a combination of both. They all rely on the fundamental property that the randomized dataset may not reveal private data while still allowing data analysis to be performed on them. We discuss each of them in more details in this section.

Given a data set \mathcal{D} , Agrawal and Srikant [1] proposed a technique of generating a perturbed dataset \mathcal{D}^* by using additive noise *i.e.* $\mathcal{D}^* = \mathcal{D} + \mathcal{R}$, where the entries of \mathcal{R} are i.i.d. samples from a zero mean unit variance gaussian distribution. Kargupta *et al.* [6] questioned the use of random additive noise and pointed out that additive noise can be easily filtered out using spectral filtering techniques thereby leading to privacy breach of the data.

Due to the potential drawback of additive perturbations, several types of multiplicative perturbation techniques have been proposed in the literature. Kim and Winkler [7] proposed one such perturbation technique which multiplies a random number generated from a truncated Gaussian distribution of mean one and small variance to each data point *i.e.* $\mathcal{D}^* = \mathcal{D} \times \mathcal{R}$, where the matrix multiplication is carried out element-wise. Geometrically, such a perturbation scheme is no more than multiplying the data by a rotation matrix and

hence it might be possible to breach the privacy if one can estimate the rotation matrix. One such attack technique has been discussed by Liu *et al.* [8] which uses a sample of the input and output to derive approximations on the estimate of the rotation matrix.

A closely related but different technique uses random data projection to preserve the privacy. In this technique, the data is projected into a random subspace using either orthogonal matrices (*e.g.* DCT/DFT as done by Mukherjee *et al.* [10]) or pseudo-random matrices (as done by Liu *et al.* [9]). It can be shown that using such transformations, the euclidean distance among any pairs of tuples is preserved and hence, many distance-based data mining techniques can be applied. Moreover, the privacy of the projection scheme can be quantified using the number of columns of the projection matrix. Figure 1 shows the distribution of the error as a function of the output dimension.

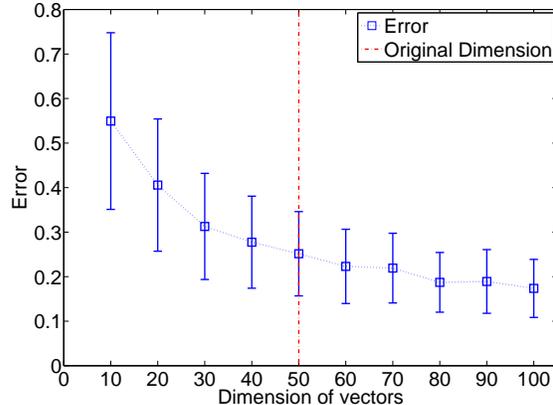


Figure 1: This graph shows the variation of error in estimating the inner product between two arbitrary vectors vs. the dimension of the output vector. The output is generated by randomly projecting the input in the subspace shown by points on the x -axis. The dimension of the input vectors are 50 as shown by dotted line. The y -axis refers to the error. The squares to the left of this line refers to dimensionality reduction and to the right refers to dimensionality inflation. Each point in the graph is an average of 100 independent trials.

In a more recent study, Chen *et al.* [3] proposed a combination of these techniques: $\mathcal{D}^* = \mathcal{T} + \mathcal{R} \times$

$\mathcal{D} + \mathcal{N}$, where \mathcal{T} is a random translation matrix, \mathcal{R} is a random rotation matrix and \mathcal{N} is a noise matrix. The paper further shows how to break this transformation in practice using a linear regression technique when the attacker knows a set of input-output pairs. However, the success of this attack depends on the variance of the matrices. The paper further defines a privacy measure known as *variance of difference (VoD)* which measures the difference of the covariance matrix between each column of \mathcal{D}^* and \mathcal{D} . We discuss this in more details later.

Data perturbation for categorical attributes have also been proposed by Warner [12] and [4]. Evfimevski *et al.* proposed the γ -amplification model [5] to bound the amount of privacy breach in categorical datasets.

3 Background

In this section we present the notations, the problem definition and an overview of the approach.

3.1 Notations Let $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ be an input data vector. Let $\mathbf{x}^* = (x_1^*, x_2^*, \dots, x_p^*) \in \mathbb{R}^p$ be the corresponding output generated according to some transformation $\mathcal{T} : \mathbb{R}^n \rightarrow \mathbb{R}^p$. In this paper we study a very general form of \mathcal{T} :

$$(3.1) \quad \mathbf{x}^* = \mathcal{T}(\mathbf{x}) = \mathbf{B} + \mathbf{Q} \times f(\mathbf{A} + \mathbf{W}\mathbf{x})$$

where $f : \mathbb{R} \rightarrow \mathbb{R}$ is a function which¹

1. acts element-wise on its argument,
2. is continuous in the real line \mathbb{R} ,
3. bounded on all bounded intervals, and
4. $f(x) = O(e^{\alpha|x|})$ as $|x| \rightarrow \infty$ where $\alpha \in \mathbb{R}$ is a constant.

$[\mathbf{B}]_{p \times 1}$, $[\mathbf{Q}]_{p \times m}$, $[\mathbf{A}]_{m \times 1}$, and $[\mathbf{W}]_{m \times n}$ are matrices (with dimensions shown) whose entries b_{ij} , q_{ij} , a_{ij} , and w_{ij} are each independently drawn from normal distributions with mean zero and standard deviations σ_b , σ_q , σ_a , and σ_w respectively *e.g.* $w_{ij} \sim \mathcal{N}(0, \sigma_w)$. The normal distribution assumption for generating random matrices is not new and has been proposed by several authors [3][9]. Special cases of \mathcal{T} can be instantiated by choosing specific instances of f two of which we discuss in Section 6. $E(\cdot)$ denotes the mean of a random variable and $\sigma^2(\cdot)$ denotes its variance. The inner product between two vectors \mathbf{x} and \mathbf{y} is denoted by $\mathbf{x} \cdot \mathbf{y}$.

¹These are sufficient but by no means necessary conditions, which are in place to ensure the existence of the improper integrals that we later derive.

3.2 Problem Definition In this paper we analyze the relation between the input data vectors and their corresponding outputs under the transformation \mathcal{T} . While such a relationship can be studied in many different ways, we focus on the *inner product* between the input and the output. Inner product is an important primitive which can be used for many advanced data mining tasks such as distance computation, clustering, classification and more. Specifically, we try to gain insight into the following problem.

Problem Statement: Given two vectors $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ and $\mathbf{y} = (y_1, y_2, \dots, y_n) \in \mathbb{R}^n$, let $\mathbf{x}^* = \mathcal{T}(\mathbf{x}) = (x_1^*, x_2^*, \dots, x_p^*) \in \mathbb{R}^p$ and $\mathbf{y}^* = \mathcal{T}(\mathbf{y}) = (y_1^*, y_2^*, \dots, y_p^*) \in \mathbb{R}^p$ be the corresponding output vectors. Since \mathbf{x}^* and \mathbf{y}^* are random transformations of their parent vectors, we analyze the relationship between $\mathbf{x} \cdot \mathbf{y}$ and $\mathbf{x}^* \cdot \mathbf{y}^*$. Our study in this paper focuses on

1. understanding the **accuracy** of \mathcal{T} in preserving distances *i.e.* studying the properties of $E[\mathbf{x}^* \cdot \mathbf{y}^*]$, and
2. **analyzing** the privacy-preserving properties of \mathcal{T} .

3.3 Overview of Approach

4 Non-linear Data Distortion

In this section we present our data distortion method using a potentially non-linear transformation. Later we will analyze two special cases of this method: (1) $f = \tanh$ function which corresponds to the non-linear function used in neural networks, and (2) f is an identity function which corresponds to a linear transformation using \mathcal{T} .

In the next subsection we introduce the mechanism of this transformation and then show its distance-preserving properties.

4.1 Mechanism Let $[\mathbf{D}]_{m \times n}$ be a data set owned by Alice in which there are m instances each of dimensionality n . Alice wants Mark (a data miner) to grant access to this dataset. However, she does not want Mark to look at the raw data. So for every vector $\mathbf{x} \in \mathbf{D}$, Alice generates a new tuple $\mathbf{x}^* \in \mathbf{D}^*$ according to the following transformation:

$$(4.2) \quad \mathbf{x}^* = \mathbf{B} + \mathbf{Q} \times f(\mathbf{A} + \mathbf{W}\mathbf{x})$$

where $\mathbf{B}, \mathbf{Q}, \mathbf{A}$ and \mathbf{W} are all mean zero and constant variance gaussian i.i.d. random matrices as defined in Section 3.1. Figure 2 shows sample input data. Figure 3 shows the perturbation achieved by the transformation in different trials for the same input.

In the next subsection we discuss how the inner product between two input vectors is related to their transformed counterpart.

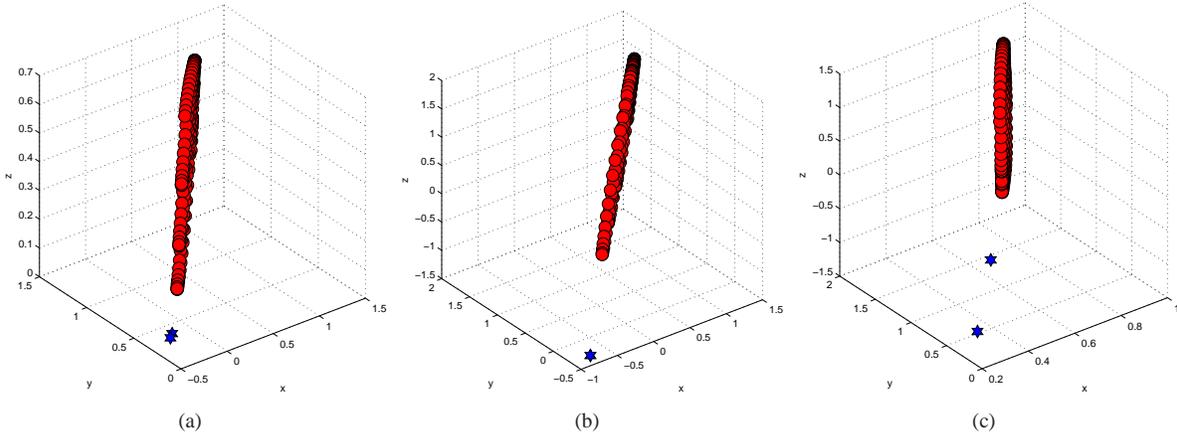


Figure 3: Sample outputs for the helix data set. Different plots show the different outputs achieved in different trials with the same input.

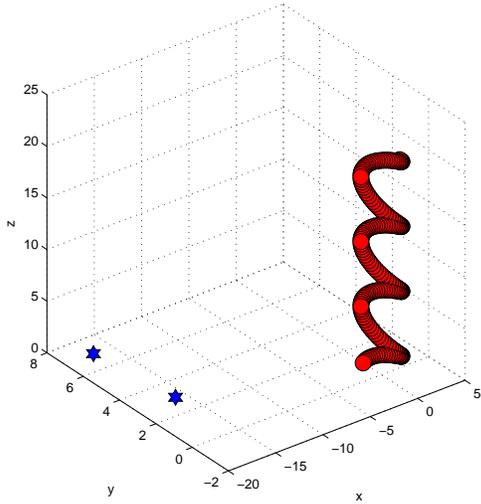


Figure 2: Sample input data set.

4.2 Derivation of $E[\mathbf{x}^* \cdot \mathbf{y}^*]$ In this section we show how $E[\mathbf{x}^* \cdot \mathbf{y}^*]$ can be evaluated.

Note that,

$$\begin{aligned}
 E[\mathbf{x}^* \cdot \mathbf{y}^*] &= E[x_1^* y_1^* + x_2^* y_2^* + \dots + x_p^* y_p^*] \\
 &= E[x_1^* y_1^*] + E[x_2^* y_2^*] + \dots + E[x_p^* y_p^*] \\
 (4.3) \quad &= pE[x_i^* y_i^*]
 \end{aligned}$$

where i is arbitrary. Further, letting \mathbf{w}_i denote the i -th row

of \mathbf{W} , we have

$$\begin{aligned}
 x_i^* y_i^* &= \left[b_i + \sum_{\ell=1}^m q_{i\ell} f(a_\ell + \mathbf{w}_\ell \mathbf{x}) \right] \\
 &\cdot \left[b_i + \sum_{\ell=1}^m q_{i\ell} f(a_\ell + \mathbf{w}_\ell \mathbf{y}) \right]
 \end{aligned}$$

In taking the expected value of the above expression, one need only consider those terms that are not linear in both $q_{i\ell}$ and b_i . All other terms evaluate to zero under the expected value operator by the independence of the random variables concerned and their property of having mean zero. Hence,

$$\begin{aligned}
 E[x_i^* y_i^*] &= E \left[b_i^2 + \sum_{\ell=1}^m q_{i\ell}^2 f(a_\ell + \mathbf{w}_\ell \mathbf{x}) f(a_\ell + \mathbf{w}_\ell \mathbf{y}) \right] \\
 &= E[b_i^2] + mE[q_{i\ell}^2] E[f(a_\ell + \mathbf{w}_\ell \mathbf{x}) f(a_\ell + \mathbf{w}_\ell \mathbf{y})] \\
 (4.4) \quad &= \sigma_b^2 + m\sigma_q^2 E[f(a_i + \mathbf{w}_i \mathbf{x}) f(a_i + \mathbf{w}_i \mathbf{y})]
 \end{aligned}$$

where i and ℓ are interchangeable. So it suffices to find $E[f(a_i + \mathbf{w}_i \mathbf{x}) f(a_i + \mathbf{w}_i \mathbf{y})]$ where i is arbitrary. Below we define two vectors $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ which aid in finding the expected value.

DEFINITION 4.1. Let $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ be $(p+1)$ -dimensional vectors defined as follows:

$$(4.5) \quad \hat{\mathbf{x}} = (\sigma_w \mathbf{x}, \sigma_a)$$

$$(4.6) \quad \hat{\mathbf{y}} = (\sigma_w \mathbf{y}, \sigma_a)$$

where σ_w and σ_a are the variances of \mathbf{W} and \mathbf{A} respectively and, \mathbf{x} and \mathbf{y} are the inputs.

Now let

$$\begin{aligned} X &= a_i + \mathbf{w}_i \mathbf{x} \\ Y &= a_i + \mathbf{w}_i \mathbf{y} \end{aligned}$$

be two random variables. Now X and Y are linear combinations of normally distributed random variables; hence they themselves are gaussian random vectors. Thus, it is easy to verify that

$$\begin{aligned} X &\sim N(0, \|\hat{\mathbf{x}}\|^2) \\ Y &\sim N(0, \|\hat{\mathbf{y}}\|^2) \end{aligned}$$

Combining Equations 4.3 and 4.4, we can write:

$$(4.7) \quad E[\mathbf{x}^* \cdot \mathbf{y}^*] = p \{ \sigma_b^2 + m \sigma_q^2 E[f(X)f(Y)] \}$$

The last equation shows that the expected inner product can be evaluated using the joint probability distribution between X and Y . Further, it can be shown that since X and Y are gaussian random variables, the joint probability distribution is actually a bivariate gaussian distribution $g_{X,Y}(x, y)$:

$$g_{X,Y}(x, y) = \frac{1}{2\pi \|\hat{\mathbf{x}}\| \|\hat{\mathbf{y}}\| \sqrt{1 - \rho_{X,Y}^2}} \times \exp \left[-\frac{1}{2(1 - \rho_{X,Y}^2)} \left(\frac{x^2}{\|\hat{\mathbf{x}}\|^2} + \frac{y^2}{\|\hat{\mathbf{y}}\|^2} - \frac{2\rho_{X,Y}xy}{\|\hat{\mathbf{x}}\| \|\hat{\mathbf{y}}\|} \right) \right]$$

where for this form to be valid $\|\hat{\mathbf{x}}\|$ and $\|\hat{\mathbf{y}}\|$ must be nonzero and $\rho_{X,Y}$, the correlation coefficient of X and Y , must not be ± 1 . Unless otherwise stated, from now on we will assume that

- $\|\hat{\mathbf{x}}\| > 0, \|\hat{\mathbf{y}}\| > 0$, and
- $\rho_{X,Y} \neq \pm 1$

Note that these conditions are equivalent to $|\hat{\mathbf{x}} \cdot \hat{\mathbf{y}}| < \|\hat{\mathbf{x}}\| \|\hat{\mathbf{y}}\|$. $\rho_{X,Y}$ can be defined in terms of $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ as:

$$(4.8) \quad \rho_{X,Y} = \frac{\hat{\mathbf{x}} \cdot \hat{\mathbf{y}}}{\|\hat{\mathbf{x}}\| \|\hat{\mathbf{y}}\|}$$

Finally, we can write,

$$E[f(X)f(Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x)f(y)g_{X,Y}(x, y)dx dy$$

Note that $E[f(X)f(Y)]$ can be difficult if not impossible to solve explicitly and in full generality, depending on the choice of f . However, given f , the above integrals can be approximated numerically for instances of \mathbf{x} and \mathbf{y} in such a way that scales very well with the input dimension, n , which enters into the (trivial) computations of $\|\hat{\mathbf{x}}\|, \|\hat{\mathbf{y}}\|$, and $\hat{\mathbf{x}} \cdot \hat{\mathbf{y}}$ alone. Using such an approximation, $E[f(X)f(Y)]_{\text{approx}}$,

one can obtain a numerical approximation of $E[\mathbf{x}^* \cdot \mathbf{y}^*]$ (refer to Equation 4.7). However, the approximation becomes less accurate the larger p, m , and σ_q are. Since f is bounded, it is true that $E[f(X)f(Y)]$ is convergent. Putting it all together, we can write:

$$(4.9) \quad \begin{aligned} E[\mathbf{x}^* \cdot \mathbf{y}^*] &= p\sigma_b^2 \\ &+ pm\sigma_q^2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x)f(y)g_{X,Y}(x, y)dx dy \end{aligned}$$

Next, we state some properties of $E[f(X)f(Y)]$:

- **Case 1:** if $\hat{\mathbf{x}} \cdot \hat{\mathbf{y}} = 0$:
 - This implies that X and Y are independent (since X and Y are gaussian vectors). Hence $E[f(X)f(Y)] = E[f(X)]E[f(Y)]$.
- **Case 2:** if $\hat{\mathbf{x}} \cdot \hat{\mathbf{y}} < 0$ or $\hat{\mathbf{x}} \cdot \hat{\mathbf{y}} > 0$:
 - With the help of the additional assumption that f is an odd function, it can be shown using the expression for $g_{X,Y}(x, y)$ that

$$E[f(X)f(Y)] < 0 \text{ or } E[f(X)f(Y)] > 0.$$

Since computing $E[f(X)f(Y)]$ is difficult to find in full generality, in the next section we develop a bound on $E[f(X)f(Y)]$ and analyze its properties.

5 Bounds on $E[f(X)f(Y)]$

The improper integral for $E[f(X)f(Y)]$ (Equation 4.9) remains intractable without further knowledge of f . In the absence of an explicit antiderivative, given f one can generate a table of values for $E[f(X)f(Y)]_{\text{approx}}$ obtained by numerical integration for a number of choices of $\|\hat{\mathbf{x}}\|, \|\hat{\mathbf{y}}\|$, and $\hat{\mathbf{x}} \cdot \hat{\mathbf{y}}$. In order to develop a bound on $E[f(X)f(Y)]$, we use the following lemma (proof omitted).

$$\text{LEMMA 5.1. } |E[f(X)f(Y)]| \leq \sqrt{E[f^2(X)]E[f^2(Y)]}$$

The following lemma (Lemma 5.2) shows the bound on $E[f(X)f(Y)]$.

LEMMA 5.2. *Let $X, Y, \hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ be as defined in the previous sections. It can be shown that,*

$$\begin{aligned} |E[f(X)f(Y)]| &\leq \sqrt{\left(\int_{-\infty}^{\infty} f^2(x) \cdot \frac{e^{-x^2/(2\|\hat{\mathbf{x}}\|^2)}}{\sqrt{2\pi}\|\hat{\mathbf{x}}\|} dx \right)} \\ &\times \sqrt{\left(\int_{-\infty}^{\infty} f^2(y) \cdot \frac{e^{-y^2/(2\|\hat{\mathbf{y}}\|^2)}}{\sqrt{2\pi}\|\hat{\mathbf{y}}\|} dy \right)} \end{aligned}$$

Proof.

$$\begin{aligned}
E[f^2(X)] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f^2(x) g_{X,Y}(x,y) dx dy \\
&= \int_{-\infty}^{\infty} f^2(x) dx \int_{-\infty}^{\infty} g_{X,Y}(x,y) dy \\
&= \int_{-\infty}^{\infty} f^2(x) g_X(x) dx \\
&= \int_{-\infty}^{\infty} f^2(x) \cdot \frac{e^{-x^2/(2\|\hat{\mathbf{x}}\|^2)}}{\sqrt{2\pi}\|\hat{\mathbf{x}}\|} dx
\end{aligned}$$

if X is not degenerate. Similarly,

$$E[f^2(Y)] = \int_{-\infty}^{\infty} f^2(y) \cdot \frac{e^{-y^2/(2\|\hat{\mathbf{y}}\|^2)}}{\sqrt{2\pi}\|\hat{\mathbf{y}}\|} dy$$

Therefore, the bound on $|E[f(X)f(Y)]|$ can be written as:

$$\begin{aligned}
|E[f(X)f(Y)]| &\leq \sqrt{\left(\int_{-\infty}^{\infty} f^2(x) \cdot \frac{e^{-x^2/(2\|\hat{\mathbf{x}}\|^2)}}{\sqrt{2\pi}\|\hat{\mathbf{x}}\|} dx\right)} \\
(5.10) \quad &\times \sqrt{\left(\int_{-\infty}^{\infty} f^2(y) \cdot \frac{e^{-y^2/(2\|\hat{\mathbf{y}}\|^2)}}{\sqrt{2\pi}\|\hat{\mathbf{y}}\|} dy\right)}
\end{aligned}$$

■

5.1 Variance Analysis In practice, given two input vectors, it is difficult to run the transformation for many independent trials and then take the average inner products of the output vectors. In this section we derive bounds on the variance of the estimated inner product, in order to quantify the error injected for a single run of the transformation.

LEMMA 5.3. *Let $X = a_i + \mathbf{w}_i \mathbf{x}$ and $Y = a_i + \mathbf{w}_i \mathbf{y}$ be two random variables where a_i is an arbitrary entry of $[\mathbf{A}]_{m \times 1}$, \mathbf{w}_i is an arbitrary row of $[\mathbf{W}]_{m \times n}$, p is the dimension of the output space, and \mathbf{x} and \mathbf{y} are the inputs. The variance of the inner product between the output vectors \mathbf{x}^* and \mathbf{y}^* can be written as:*

$$\begin{aligned}
\sigma_{\langle \mathbf{x}^*, \mathbf{y}^* \rangle}^2 &= 2p\sigma_b^4 + pm\sigma_b^2\sigma_q^2(E[f(Y)^2] + E[f(X)^2]) \\
&\quad + pm\sigma_q^4(3pE[f(X)^2 f(Y)^2] - pE[f(X)f(Y)]^2 \\
&\quad + (m-1)E[f(X)^2]E[f(Y)^2])
\end{aligned}$$

Proof. The proof is extremely algebra intensive. We omit it here due to shortage of space and plan to report it in an extended version of this paper.

What does this expression tell us?. Hillol's expression tells us that variance decreases as the size of the projection matrix increases. Can we have a similar result here?

6 Special cases

In this section we study two special cases of the general transformation \mathcal{T} : (1) when f is a sigmoid or tanh function which has been used as a popular choice for non-linear mapping, and (2) when f is an identity function making the resulting \mathcal{T} linear.

6.1 $f = \tanh$ function In this section we analyze the properties of $E[\mathbf{x}^* \cdot \mathbf{y}^*]$ when f is a sigmoid or hyperbolic tangent (\tanh) function. Our choice of $f = \tanh$ is not arbitrary; it makes transformation \mathcal{T} resemble that of a two-layer neural network, a tool widely used in data mining and machine learning for learning non-linear relationships from the data. With such a substitution, \mathcal{T} takes the following form:

$$\begin{aligned}
\mathbf{H}(\mathbf{x}) &= \tanh(\mathbf{A} + \mathbf{W}\mathbf{x}) \\
\mathbf{x}^* &= \mathbf{B} + \mathbf{Q}\mathbf{H}(\mathbf{x})
\end{aligned}$$

However, for the results here to describe such a trained neural network, one must assume that the weights are indeed independent and normally distributed with mean zero. Weights are assumed to be normal in many research as shown in [2] and [11].

Even with the substitution $f(x) = \tanh(x)$ in Equation 4.9, evaluation of $E[\tanh(X)\tanh(Y)]$ in closed form is still intractable. Hence we use the bound presented in Lemma 5.2 to gain insight into $E[\tanh(X)\tanh(Y)]$. Let us first evaluate $E[\tanh^2(X)]$. By definition,

$$E[\tanh^2(X)] = \int_{-\infty}^{\infty} \tanh^2(x) \cdot \frac{e^{-x^2/(2\|\hat{\mathbf{x}}\|^2)}}{\sqrt{2\pi}\|\hat{\mathbf{x}}\|} dx$$

Unfortunately, an anti-derivative does not exist for this function. So we approximate the \tanh function with a linear function that takes on the values -1 and 1 far to the left and right of the origin, respectively, and has a slope of constant, positive value in between. For simplicity we make this slope tangent to the slope of the f function at the origin, which means the slope of our approximation to be 1 over $[-1, 1]$ and zero otherwise. Letting $\Psi(X)$ denote the approximating function,

$$\tanh(X) \approx \Psi(X) = -1 \cdot \chi_{(-\infty, -1)} + x \cdot \chi_{[-1, 1]} + 1 \cdot \chi_{(1, \infty)}$$

where χ is the indicator function. Figure 4 shows the original \tanh function, the approximation to it and the step function.

It is easy to see that,

$$\Psi(X)^2 = 1 \cdot \chi_{(-\infty, -1)} + x^2 \cdot \chi_{[-1, 1]} + 1 \cdot \chi_{(1, \infty)}$$

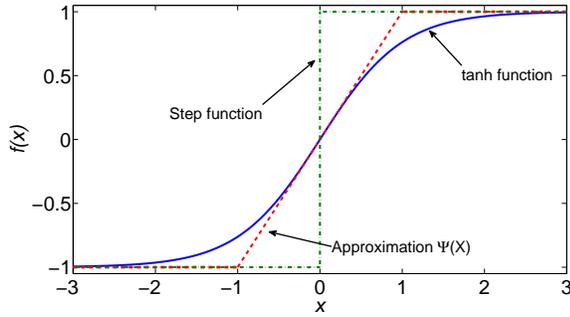


Figure 4: Hyperbolic tangent (tanh) function shown in bold. $\Psi(x)$ is the approximation to $\tanh(x)$. A step function is also shown.

Denoting $g_X(x)$ as the marginal distribution of X we get,

$$\begin{aligned}
E[\tanh^2(x)] &= \int_{-\infty}^{\infty} \tanh^2(x) \cdot g_X(x) dx \\
&< \int_{-\infty}^{\infty} \Psi(X)^2 \cdot g_X(x) dx \\
&= \int_{-\infty}^{-1} g_X(x) dx + \int_{-1}^1 x^2 \cdot g_X(x) dx \\
&\quad + \int_1^{\infty} g_X(x) dx \\
&= 2 \int_{-\infty}^{-1} g_X(x) dx + \int_{-1}^1 x^2 \cdot g_X(x) dx
\end{aligned}$$

$$\begin{aligned}
\text{Term 1} &= 2 \int_{-\infty}^{-1} \frac{e^{-x^2/(2\|\hat{\mathbf{x}}\|^2)}}{\sqrt{2\pi}\|\hat{\mathbf{x}}\|} dx \\
&= 2 \frac{1}{\sqrt{2\pi}\|\hat{\mathbf{x}}\|} \int_{-\infty}^{-1} e^{-x^2/(2\|\hat{\mathbf{x}}\|^2)} dx \\
&= 2 \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-\frac{1}{\|\hat{\mathbf{x}}\|}} e^{-y^2/2} dy \quad [\text{where } y = \frac{x}{\|\hat{\mathbf{x}}\|}] \\
&= 2\Phi\left(-\frac{1}{\|\hat{\mathbf{x}}\|}\right)
\end{aligned}$$

Now we evaluate Term 2. First we evaluate the following integral.

$$\begin{aligned}
\int x e^{-x^2/(2\|\hat{\mathbf{x}}\|^2)} dx &= \int \|\hat{\mathbf{x}}\|^2 e^{-y} dy \quad [\text{using } y = \frac{x^2}{2\|\hat{\mathbf{x}}\|^2}] \\
&= -\|\hat{\mathbf{x}}\|^2 e^{-x^2/(2\|\hat{\mathbf{x}}\|^2)} + c
\end{aligned}$$

Now to evaluate Term 2 (using the previous result),

$$\begin{aligned}
\text{Term 2} &= \frac{1}{\sqrt{2\pi}\|\hat{\mathbf{x}}\|} \left[\int_{-1}^1 x^2 \cdot e^{-x^2/(2\|\hat{\mathbf{x}}\|^2)} dx \right] \\
&= \frac{1}{\sqrt{2\pi}\|\hat{\mathbf{x}}\|} \left[x \cdot \int x \cdot e^{-x^2/(2\|\hat{\mathbf{x}}\|^2)} dx \right]_{-1}^1 \\
&\quad + \frac{1}{\sqrt{2\pi}\|\hat{\mathbf{x}}\|} \left[\int_{-1}^1 \|\hat{\mathbf{x}}\|^2 e^{-x^2/(2\|\hat{\mathbf{x}}\|^2)} dx \right] \\
&= \frac{-1}{\sqrt{2\pi}\|\hat{\mathbf{x}}\|} \left(\|\hat{\mathbf{x}}\|^2 x e^{-x^2/(2\|\hat{\mathbf{x}}\|^2)} \right)_{-1}^1 \\
&\quad + \frac{\|\hat{\mathbf{x}}\|}{\sqrt{2\pi}} \left[\int_{-1}^1 e^{-x^2/(2\|\hat{\mathbf{x}}\|^2)} dx \right] \\
&= \frac{-\|\hat{\mathbf{x}}\|}{\sqrt{2\pi}} \left(e^{-1/(2\|\hat{\mathbf{x}}\|^2)} + e^{-1/(2\|\hat{\mathbf{x}}\|^2)} \right) \\
&\quad + \|\hat{\mathbf{x}}\|^2 \left[\Phi\left(\frac{1}{\|\hat{\mathbf{x}}\|}\right) - \Phi\left(-\frac{1}{\|\hat{\mathbf{x}}\|}\right) \right]
\end{aligned}$$

Combining the results,

$$\begin{aligned}
E[\tanh^2(x)] &< 2\Phi\left(-\frac{1}{\|\hat{\mathbf{x}}\|}\right) + \|\hat{\mathbf{x}}\|^2 \left[\Phi\left(\frac{1}{\|\hat{\mathbf{x}}\|}\right) - \Phi\left(-\frac{1}{\|\hat{\mathbf{x}}\|}\right) \right] \\
&\quad - \frac{\|\hat{\mathbf{x}}\|}{\sqrt{2\pi}} \left(2e^{-1/(2\|\hat{\mathbf{x}}\|^2)} \right)
\end{aligned}$$

Using a similar argument, it can be shown that,

$$\begin{aligned}
E[\tanh^2(y)] &< 2\Phi\left(-\frac{1}{\|\hat{\mathbf{y}}\|}\right) + \|\hat{\mathbf{y}}\|^2 \left[\Phi\left(\frac{1}{\|\hat{\mathbf{y}}\|}\right) - \Phi\left(-\frac{1}{\|\hat{\mathbf{y}}\|}\right) \right] \\
&\quad - \frac{\|\hat{\mathbf{y}}\|}{\sqrt{2\pi}} \left(2e^{-1/(2\|\hat{\mathbf{y}}\|^2)} \right)
\end{aligned}$$

These results can now be combined to get the final form of the bound using Equation 5.10.

Figure 5 shows a plot of the bound $|E[\tanh(X) \tanh(Y)]|$ with variation of $\|\hat{\mathbf{x}}\|$ and $\|\hat{\mathbf{y}}\|$. It can be shown that the bound lies between 0 and 1. When both $\|\hat{\mathbf{x}}\|$ and $\|\hat{\mathbf{y}}\|$ are small *i.e.* close to the origin, we know that the expected inner product of their output should be close to 0 as well. Looking at the figure we see that this is indeed the case. So our bound is a good approximation when we are close to the origin and becomes crude as we move further away from the origin.

6.2 Linear Transformation The second transformation that we study in this section is a linear transformation. Linear transformations have been widely studied in the form of random projection, multiplicative perturbation [9][6][3] where the output is linearly dependent on the input:

$$\mathbf{x}^* = \mathbf{T} + \mathbf{R}\mathbf{x}$$

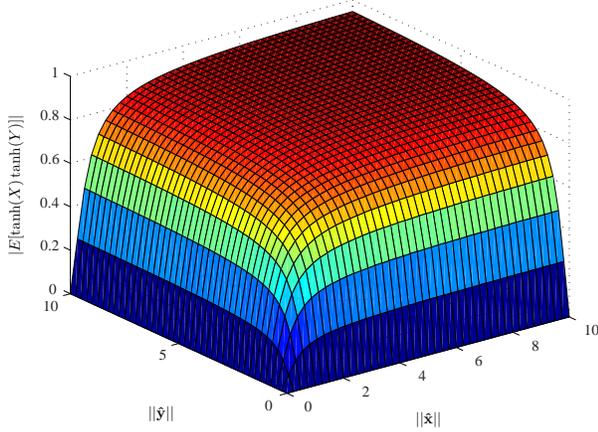


Figure 5: Plot of $|E[\tanh(X) \tanh(Y)]|$ vs. $\|\hat{\mathbf{x}}\|$ and $\|\hat{\mathbf{y}}\|$.

where T and R are random translation and rotation matrices. In order for our transformation \mathcal{T} to be linear, we assume that f is an identity function *i.e.* $f(x) = x, \forall x \in \mathbb{R}$. Unlike the previous section, in this section we show how a closed form expression for $E[\mathbf{x}^* \cdot \mathbf{y}^*]$ can be developed for such a transformation.

Using the definition of X and Y , it is easy to show that,

$$E[f(X)f(Y)] = E[XY] = \hat{\mathbf{x}} \cdot \hat{\mathbf{y}}$$

Since $\hat{\mathbf{x}} = \sigma_w(\hat{\mathbf{x}}, \sigma_a)$ and $\hat{\mathbf{y}} = \sigma_w(\hat{\mathbf{y}}, \sigma_a)$,

$$\hat{\mathbf{x}} \cdot \hat{\mathbf{y}} = \sigma_w^2(\mathbf{x} \cdot \mathbf{y}) + \sigma_a^2$$

Combining these results, we have:

$$\begin{aligned} E[\mathbf{x}^* \cdot \mathbf{y}^*] &= p\sigma_b^2 + pm\sigma_q^2 E[XY] \\ &= p\sigma_b^2 + pm\sigma_q^2 (\hat{\mathbf{x}} \cdot \hat{\mathbf{y}}) \\ &= p\sigma_b^2 + pm\sigma_a^2\sigma_q^2 + pm\sigma_q^2\sigma_w^2(\mathbf{x} \cdot \mathbf{y}) \end{aligned}$$

This equation shows that for a linear transformation, the inner product of the output vectors is proportional to the inner product of the input vectors. In other words, the distances are preserved on average (up to scaling and translation). This result is in line with what some other authors reported elsewhere [3][9].

Let us investigate the quality of the bound for this transformation. Substituting $f(x) = x$ and $f(y) = y$, in Equation 5.10, we see that the integrals are $E[x^2]$ and $E[y^2]$ respectively. Now, since $x \sim N(0, \|\hat{\mathbf{x}}\|^2)$ and $y \sim N(0, \|\hat{\mathbf{y}}\|^2)$, $E[x^2] = \|\hat{\mathbf{x}}\|^2$ and $E[y^2] = \|\hat{\mathbf{y}}\|^2$. Thus,

$$E_{est}[XY] \leq \|\hat{\mathbf{x}}\| \|\hat{\mathbf{y}}\|$$

where E_{est} denotes the estimated value of the expectation. Therefore we can write the following expression for the bound:

$$E[\mathbf{x}^* \cdot \mathbf{y}^*] \leq p\sigma_b^2 + pm\sigma_q^2 \|\hat{\mathbf{x}}\| \|\hat{\mathbf{y}}\|$$

where,

$$\begin{aligned} \|\hat{\mathbf{x}}\| &= \sqrt{\sigma_w^2(\|\mathbf{x}\|^2) + \sigma_a^2} \\ \|\hat{\mathbf{y}}\| &= \sqrt{\sigma_w^2(\|\mathbf{y}\|^2) + \sigma_a^2} \end{aligned}$$

Figure 6 shows a plot of $E[\mathbf{x}^* \cdot \mathbf{y}^*]$ when θ , the angle between $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ varies. For all the figures, the circles show the true variation of $E[\mathbf{x}^* \cdot \mathbf{y}^*]$ vs. θ . The squares represent the bound. Note that for all the figures, the bound correctly represents the inner-product only when $\theta = 0, \pm 2\pi, \pm 4\pi, \dots$. The three figures demonstrate the effect on the output for three values of $\|\hat{\mathbf{x}}\|$ and $\|\hat{\mathbf{y}}\|$. As can be seen, the bound is a good approximation of the true value when $\|\hat{\mathbf{x}}\|$ and $\|\hat{\mathbf{y}}\|$ are small.

7 Privacy Analysis

8 Experimental Results

9 Conclusion

Acknowledgments

This work was supported by the IVHM project at NASA Ames Research Center. Mark Stefanski would also like to acknowledge the NASA USRP internship program.

References

- [1] R. Agrawal and R. Srikant. Privacy-preserving Data Mining. In *Proceedings of SIGMOD'00*, pages 439–450, Dallas, Texas, May 2000.
- [2] I. Bellido and E. Fiesler. Do Backpropagation Trained Neural Networks Have Normal Weight Distributions? In *Proceedings of ICANN'93*, pages 772–775, Amsterdam, Netherlands, September 1993.
- [3] K. Chen, G. Sun, and L. Liu. Towards Attack-Resilient Geometric Data Perturbation. In *Proceedings of SDM'08*, pages 78–89, 2008.
- [4] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy Preserving Mining of Association Rules. In *Proceedings of KDD'02*, pages 217–228, 2002.
- [5] Alexandre Evfimievski, Johannes Gehrke, and Ramakrishnan Srikant. Limiting Privacy Breaches in Privacy Preserving Data Mining. In *Proceedings of PODS'03*, pages 211–222, 2003.
- [6] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar. On the Privacy Preserving Properties of Random Data Perturbation Techniques. In *Proceedings of ICDM'03*, page 99, Melbourne, Florida, November 2003.
- [7] J. J. Kim and W. E. Winkler. Multiplicative Noise for Masking Continuous Data. Technical Report Statistics #2003-01, Statistical Research Division, U.S. Bureau of the Census, Washington D.C., April 2003.
- [8] K. Liu, C. Giannella, and H. Kargupta. An Attacker's View of Distance Preserving Maps for Privacy Preserving Data Mining. In *Proceedings of PKDD'06*, pages 297–308, Berlin, Germany, 2006.

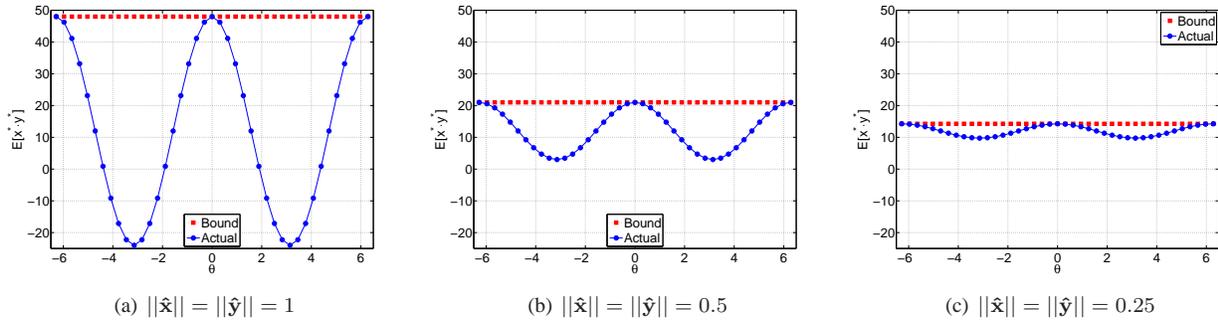


Figure 6: Variation of the output $E[\mathbf{x}^* \cdot \mathbf{y}^*]$ with respect to θ (in radians), the angle between $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$. Circles represent the true output and squares represent the bound. For all figures, the bound is independent of θ . For a fixed $\|\hat{\mathbf{x}}\|$ and $\|\hat{\mathbf{y}}\|$, actual output oscillates and equals the bound only at $\theta = 0, \pm 2\pi, \dots$. As $\|\hat{\mathbf{x}}\| \rightarrow 0$ and $\|\hat{\mathbf{y}}\| \rightarrow 0$, the actual and estimated value comes closer. The bound is very tight when $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ close to origin.

- [9] K. Liu, H. Kargupta, and J. Ryan. Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining. *IEEE TKDE*, 18(1):92–106, January 2006.
- [10] S. Mukherjee, Z. Chen, and A. Gangopadhyay. A Privacy-preserving Technique for Euclidean Distance-based Mining Algorithms using Fourier-related Transforms. *The VLDB Journal*, 15(4):293–315, 2006.
- [11] T. Szabó, L. Antoni, G. Horváth, and B. Fehér. A Full-Parallel Digital Implementation for Pre-Trained NNs. In *Proceedings of IJCNN'00-Volume 2*, page 2049, Como, Italy, July 2000.
- [12] S. Warner. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of American Statistical Association*, 65(63–69), 1965.