Lipschitz Continuous Autoencoders in Application to Anomaly Detection

Young-geun Kim
Department of Statistics
Seoul National University
Seoul 08826, Republic of Korea
cib009@snu.ac.kr

Yongchan Kwon
Department of Statistics
Seoul National University
Seoul 08826, Republic of Korea
ykwon0407@snu.ac.kr

Hyunwoong Chang
Department of Statistics
Seoul National University
Seoul 08826, Republic of Korea
darcy7888@snu.ac.kr

Myunghee Cho Paik
Department of Statistics
Seoul National University
Seoul 08826, Republic of Korea
myungheechopaik@snu.ac.kr

Abstract
Anomaly detection is the task of finding abnormal data that are distinct from normal behavior. Current deep learning-based anomaly detection methods train neural networks with normal data alone and calculate anomaly scores based on the trained model. In this work, we formalize current practices and build a theoretical framework of anomaly detection methods equipped with a loss function and a function space of the specified models. We establish a desirable property of the framework, namely, admissibility. Admissibility implies that the expected cost for anomalous data through the autoencoders optimized using normal data is larger than that for normal data. We then propose an integral probability metric (IPM)-based loss function and a class of autoencoders, Lipschitz continuous autoencoders with respect to IPM and prove that the proposed framework is admissible. Based on the constructed framework, we propose an admissible algorithm by minimizing an approximated Wasserstein distance with a penalty to enforce Lipschitz continuity. We demonstrate the efficacy of the proposed method with network traffic dataset and many image datasets including face image dataset.

1 Introduction
Anomaly detection is the problem of identifying observations that deviate from the majority of the data in the absence of labeled data [4, 8, 19]. To identify alarming situations, anomaly detection has been applied in fraud detection [3], medical diagnosis [12], network security [6], and visual surveillance [9].

The goal of anomaly detection is to construct a classifier that distinguishes abnormal data from normal data. Due to a lack of labeled abnormal observations, many anomaly detection algorithms have trained a model and constructed anomaly score based on the trained model using normal data alone [4, 10]. These methods assume that training data consist of only normal class, and estimate the support [26] or likelihood function of observations [14]. The data that are far from the support or have low values of likelihood are considered as anomalies. With a similar focus, many deep neural network-based anomaly detection methods have been proposed recently, utilizing various loss functions to optimize neural networks with normal data [2]. Details on deep learning-based anomaly scores are in Section

Preprint. Under review.
These approaches have intuitive appeal since the features from anomalies should show different behavior from those from normal, but their theoretical properties have not been explored.

One approach to handle anomaly problem is to train autoencoders minimizing the reconstruction error on the distribution of normal data, and construct an anomaly score based on a distance between the test input and the reconstructed value through the trained autoencoders. In fact, common elements of many existing anomaly detection methods include specifying and training a model with a loss function to obtain its optimizer, and computing an anomaly score from a derived contribution of a datum to the loss function. We formalize and build a class of the anomaly detection framework fully identified by a loss function and a function space of the specified models. Anomaly score can be defined as the contribution of a test datum to the loss function evaluated at the optimizer, which is a by-product once the loss function and the function space are determined. We show that given the cost-based loss function and optimizer, the anomaly score can be interpreted as an influence-like function used in statistics, a Gateaux derivative of the expected cost perturbed in the direction of a test datum. Using the constructed framework, we characterize a desirable property of anomaly detection methods, namely admissibility. This property in words implies that the expected cost for anomalous data through the autoencoders optimized using the normal data is larger than the expected cost for the normal data. Although not stated explicitly, existing methods are based on the premise that this property holds. We show that admissibility property does not hold in general. We propose anomaly detection methods with a new class of loss function based on integral probability metric (IPM) and a new class of a function space for autoencoders, namely, Lipschitz continuous autoencoders w.r.t. IPM. We prove that the proposed methods are guaranteed to be admissible. In particular, we take the loss function as Wasserstein distance and provide a specific algorithm enforcing Lipschitz-continuity w.r.t. IPM to guarantee admissibility.

Our contribution consists of four elements as follows.

- We build a theoretical framework of anomaly detection methods and characterize a desirable property of anomaly detection framework, namely, admissibility. Admissibility implies that the expected cost for anomalous data through the autoencoders optimized using the normal data is larger than the expected cost for the normal data. (Sections 3.1 and 3.2)
- We propose a class of autoencoders, Lipschitz continuous autoencoders with respect to IPM and prove that the anomaly detection methods equipped with an IPM-based loss function and the proposed class of autoencoders are admissible. (Section 3.3)
- We propose an anomaly detection algorithm with an approximated Wasserstein distance as a loss function while enforcing Lipschitz continuity of autoencoders with respect to IPM. (Section 3.4)
- We demonstrate that the proposed method outperforms existing alternatives in many applications including network security and image recognition-based anomaly detection problems. (Section 4)

The remainder of the paper is organized as follows. In Section 2, we review related works. Section 3 provides the proposed method including an admissible anomaly detection framework via Lipschitz continuous autoencoders. Section 4 demonstrates the application of proposed method on network traffic and image data. All proofs of examples, propositions, and theorems are provided in Appendix A of the supplementary material.

2 Related works

Many deep learning-based anomaly detection methods have been proposed, including support vector-based, generative adversarial network-based, and autoencoder-based approaches. These methods first train neural networks using labeled-normal and the entire training data and then compute anomaly scores based on extracted features from the trained model.

Inspired by support vector data description (SVDD) [29], deep SVDD [23] has been proposed, replacing kernel feature mapping with neural networks mapping. The loss function of deep SVDD is the average distance of extracted features from the centroid of the normal data cluster. Though deep SVDD is motivated from support vector algorithms, outputs of neural networks do not directly relate to the kernel and the performance on CIFAR-10 [13] is similar to kernel density estimation [21] which often does not work well in high-dimensional cases.
Various autoencoder-based anomaly detection methods have been proposed, minimizing the expected AnoGANs. Approaches based on generative adversarial networks (GANs) \cite{7} have been proposed, using a generator to augment normal data and discriminator to extract features \cite{33}. The discriminator minimizes the cross-entropy loss function and the generator minimizes the distance between the distribution of real data and that of generated data, using only normal data. Anomaly detection with generative adversarial networks (AnoGANs) is a state-of-the-art GAN-based anomaly detection method, using generator and discriminator at the same time to measure abnormality \cite{25}. The performance on a clinical image was prominent, but AnoGANs need to solve an optimization problem for every test datum to find the latent code generating the most similar counterfeit. Zenati et al. \cite{34} proposed adversarially learned anomaly detection (ALAD) based on adversarially learned inference with conditional entropy \cite{16}, a kind of GANs including encoder networks which directly map a datum to latent code. In applications to KDD99 \cite{17} and CIFAR-10, ALAD outperformed AnoGANs.

Various autoencoder-based anomaly detection methods have been proposed, minimizing the expected reconstruction error of normal data. Anomaly scores are based on representations and the reconstruction error \cite{24,32,35}. Instead of the reconstruction error, An and Cho \cite{1} proposed to use the reconstruction probability from variational autoencoders \cite{11} as anomaly scores. Deep autoencoding Gaussian mixture model (DAGMM) is a state-of-the-art autoencoder-based anomaly detection model \cite{36}, enforcing the Gaussian mixture assumption of the vector that consists of representations and the distance between input and reconstructed data. DAGMM utilizes a likelihood-based energy function to identify anomalies. The performance on the KDD99 and other benchmark image dataset was not on a par with ALAD.

These approaches are based on the premise that the expected cost of abnormal data over the distribution of normal data will be larger than the expected cost of normal data but their properties have not been formally studied.

### 3 Proposed method

In Section 3.1, we formulate a framework for anomaly detection methods. Using the constructed framework in Section 3.1, we characterize a desirable property called admissibility shown in Definition 1. It turns out that to satisfy admissibility, we need to impose Lipschitz continuity on the autoencoders w.r.t distribution metric (Definition 2). Theorem 1 shows that the proposed method satisfies admissibility. Throughout this section, we discuss Lipschitz continuity in three different contexts. To avoid confusion, we clarify the differences upfront.

Let $\mathcal{X}$ be a compact set, $d$ be a metric on $\mathcal{X}$, and $h\#$ be the push-forward operation transferring a probability measure with a function $h$. For given two distributions $P$ and $Q$, and a class of functions $\mathcal{F}$, we denote IPM by $\gamma_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} |E_{X \sim P} f(X) - E_{X \sim Q} f(X)|$.

1. We call a function $h : \mathcal{X} \rightarrow \mathcal{X}$ is $K$-Lipschitz continuous w.r.t. $d$ if $d(h(x), h(y)) \leq Kd(x, y)$ for all $x, y \in \mathcal{X}$.
2. Wasserstein distance is $\gamma_{\mathcal{F}}(P, Q)$ when $\mathcal{F}$ is $\mathcal{F}_d$, a set of 1-Lipschitz continuous functions w.r.t. $d$, $\mathcal{F}_d := \{ f \mid d(f(x), f(y)) \leq d(x, y) \text{ for all } x, y \in \mathcal{X} \}$.
3. For a given function $h : \mathcal{X} \rightarrow \mathcal{X}$, we call a push-forward operation $h\#$ is $K$-Lipschitz continuous w.r.t. $\gamma_{\mathcal{F}}$ if $\gamma_{\mathcal{F}}(h\#P_A, h\#P_B) \leq \gamma_{\mathcal{F}}(P_A, P_B)$ for any two random variables $A$ and $B$ defined on $\mathcal{X}$.

The first two are familiar ones to describe $K$-Lipschitz continuity of $h$ w.r.t. a metric $d$ on $\mathcal{X}$ and to define Wasserstein distance in dual form as an objective function in Section 3.4. The third is $K$-Lipschitz continuity w.r.t. IPM, to describe the Lipschitz continuous autoencoders in Definition 2. Patrini et al. \cite{22} defined the Lipschitz continuity w.r.t. Wasserstein distance. In this work, $K$-Lipschitz continuity w.r.t. $\gamma_{\mathcal{F}}$ is newly defined and shown to be required to achieve admissibility.

#### 3.1 Formulation of anomaly detection methods

The goal of anomaly detection is to sort out anomalies from normal data. To achieve this, many existing anomaly detection methods build a model, train with a specific loss function using the normal data alone, and construct anomaly score based on the model. We formalize these procedures...
as follows. We denote the input data for the model, domain of input data, and the distribution of normal data by $X$, $\mathcal{X}$, and $P_X^{(0)}$, respectively. The Dirac delta function that gives all mass to $x$, the set of all Borel probability measures defined on $\mathcal{X}$, and the set of all real numbers are denoted by $\delta_x$, $\Pi_\mathcal{X}$, and $\mathbb{R}$, respectively. First, loss functions utilized in anomaly detection methods are expressed as $T(P_X^{(0)}; h)$ where $T : \Pi_\mathcal{X} \times \mathcal{H} \to \mathbb{R}$ is a functional of a probability measure and a function in $\mathcal{H}$. Many existing anomaly detection methods use loss functions that can be expressed as $T_c(P_X^{(0)}; h) := \int c(x; h) dP_X^{(0)}(x)$ where $c$ is a cost function. Examples of cost functions and $\mathcal{H}$ are the hinge loss with reproducing kernel Hilbert space [26], negative log-likelihood with a class of probability density functions [14], and reconstruction error with a class of autoencoders [24, 32, 35].

The next step is to construct the anomaly score for a test datum $x_0$ by utilizing $h^{(0)}$. In many cases using $T_c(P_X^{(0)}; h)$ as a loss function, anomaly score is $c(x_0; h^{(0)})$, which is $T_c(\delta_{x_0}; h^{(0)})$ at $x = x_0$. We show in the following proposition that given the loss function expressed as $T_c$ and optimizer $h^{(0)}$, anomaly score, $T_c(\delta_{x_0}; h^{(0)})$, can be interpreted as rate of change of the expected loss in the direction of a test datum, a limit of which is known as an influence function in robust statistics. The influence function is a kind of Gateaux derivative and has played an important role in robust statistics that quantifies the effect of a datum to a probability measure [5].

**Proposition 1.** For any $x$, cost function $c$, $\nu \in (0, 1]$ and $h \in \mathcal{H}$,

$$T_c(\delta_x; h) = \frac{T_c(\nu \delta_x + (1 - \nu)P_X^{(0)}; h) - T_c(P_X^{(0)}; h)}{\nu} + T_c(P_X^{(0)}; h).$$

In this regard, for anomaly detection methods using $T(P_X^{(0)}; h)$ as a loss function, a natural choice of anomaly score is $T(\delta_{x_0}; h^{(0)})$. Using $T(\delta_{x_0}; h^{(0)})$ as anomaly scores, anomaly detection procedures can be fully characterized by $T$ and $\mathcal{H}$. In the next section we examine what property is desirable for anomaly detection framework defined by $T(P_X^{(0)}; h)$ and $\mathcal{H}$.

### 3.2 Admissibility: Desirable property of anomaly detection framework

Using the constructed framework in Section 3.1, we characterize a desirable property of $T(P_X^{(0)}; h)$ and $\mathcal{H}$, thus of anomaly detection methods, namely, admissibility. This property implies that the loss for anomalous data through the autoencoders optimized using the normal data is larger than the loss for normal data. Denoting by $P_X^{(1)}$ the distribution where anomalies come from, the formal definition of admissibility follows.

**Definition 1.** Anomaly detection framework equipped with $T$ and $\mathcal{H}$ is said to be admissible when $T(\nu P_X^{(1)} + (1 - \nu)P_X^{(0)}; h^{(0)}) > T(P_X^{(0)}; h^{(0)})$ for some $\nu \in (0, 1]$ and $h^{(0)} \in \mathcal{H}$ satisfying $T(P_X^{(0)}; h^{(0)}) \leq T(P_X^{(0)}; h)$ for all $h \in \mathcal{H}$.

That is, an anomaly detection framework is admissible if the loss function evaluated at $h^{(0)}$ for some contaminated data is larger than that for uncontaminated data. For the subclass of anomaly detection method equipped with $T_c$, the following proposition shows the condition for admissibility, that is, the loss function for abnormal data is larger than that for normal data.

**Proposition 2.** The anomaly detection equipped with $T_c$ and $\mathcal{H}$ is admissible if and only if $T_c(P_X^{(1)}; h^{(0)}) > T_c(P_X^{(0)}; h^{(0)})$ for any $h^{(0)} \in \mathcal{H}$ satisfying $T_c(P_X^{(0)}; h^{(0)}) \leq T_c(P_X^{(0)}; h)$ for all $h \in \mathcal{H}$.

Although admissibility seems to be a natural property that any anomaly detection can enjoy, it is not guaranteed since the description of $h^{(0)}$ is about the loss function evaluated with other $h \in \mathcal{H}$ on the fixed probability distribution, $P_X^{(0)}$. Below we present a simple example of anomaly detection methods equipped with $T_c(P_X^{(0)}; h)$ and a set $\mathcal{H}$ of all autoencoders, where the admissibility may not hold.

**Example 1.** Let $P_X^{(0)}$ and $P_X^{(1)}$ be joint distributions of two independent Gaussian random variables, denoted by $X^{(0)} \sim N_2(0_2, I_2)$ and $X^{(1)} \sim N_2(\mu, I_2)$, respectively, where $0_2 = (0, 0)^T$, $I_2$ is the $2 \times 2$ identity matrix, and $\mu \in \mathbb{R}^2$. Let $\mathcal{H}$ be the set of all autoencoders that consist of input layer with
We then define an IPM-based loss function given by

\[ \text{loss}_\text{IPM}(h) = \gamma_{\mathcal{F}}(\mathbb{P}, h\#), \]

where \( h \) is a function from \( \mathcal{X} \) to \( \mathcal{X} \), \( \mathcal{F} \) is a set of functions, \( \mathbb{P} \) is a probability measure, and \( h\# \) denotes the push-forward measure obtained by transferring \( \mathbb{P} \) with \( h \). For any random variable \( X \) from \( \mathcal{F} \), \( h\#\mathbb{P}_X \) is the distribution of \( h(X) \). We define the Lipschitz continuous autoencoder as follows.

**Definition 2.** An autoencoder \( h : \mathcal{X} \to \mathcal{X} \) is said to be K-Lipschitz continuous w.r.t. \( \gamma_{\mathcal{F}} \) if \( h\# : \Pi \mathcal{X} \to \Pi \mathcal{X} \) is K-Lipschitz continuous w.r.t. \( \gamma_{\mathcal{F}} \), which can be expressed as \( \gamma_{\mathcal{F}}(h\# \mathbb{P}, h\# \mathbb{Q}) \leq K \gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) \) for any two probability measures \( \mathbb{P} \) and \( \mathbb{Q} \) defined on \( \mathcal{X} \).

We denote the set of all K-Lipschitz continuous autoencoders w.r.t. \( \gamma_{\mathcal{F}} \) by \( \mathcal{H}_{\gamma_{\mathcal{F}}}^{(K)} \). We emphasize that usual Lipschitz continuity w.r.t. \( d \) is different from Lipschitz continuity w.r.t. \( \gamma_{\mathcal{F}} \). The Lipschitz continuity of an autoencoder w.r.t. \( \gamma_{\mathcal{F}} \) is the Lipschitz continuity of a push-forward operation transferring \( \mathbb{P}_A \) to \( \mathbb{P}_{h(A)} \) for any \( A \) defined on \( \mathcal{X} \). Lipschitz continuity w.r.t. \( \gamma_{\mathcal{F}} \) has not been defined to the best of the authors’ knowledge. By imposing \( K \)-Lipschitz continuity w.r.t. \( \gamma_{\mathcal{F}} \) with \( K \) less than 1, the distance between the distributions of reconstructed normal and reconstructed contaminated data is smaller than the distance between the distributions of normal and contaminated data.

We then define an IPM-based loss function given by

\[ T_{\mathcal{F}}(\mathbb{P}; h) := \gamma_{\mathcal{F}}(\mathbb{P}, h\#), \]

i.e., \( T_{\mathcal{F}}(\mathbb{P}; h) \) is a distance between the distribution of input data and that of reconstructed data. Theorem 1 shows that with a properly chosen \( h \) in \( \mathcal{H}_{\gamma_{\mathcal{F}}}^{(K)} \), the anomaly detection framework equipped with \( T_{\mathcal{F}}(\mathbb{P}; h) \) is admissible.

**Theorem 1.** If there is \( h \in \mathcal{H}_{\gamma_{\mathcal{F}}}^{(K)} \) satisfying \( T_{\mathcal{F}}(\mathbb{P}_X^{(0)}; h) < \epsilon \gamma_{\mathcal{F}}(\mathbb{P}_X^{(0)}, \mathbb{P}_X^{(1)}) \) for some \( \epsilon < (1 - K)/2 \), the anomaly detection framework equipped with \( T_{\mathcal{F}} \) and \( \mathcal{H}_{\gamma_{\mathcal{F}}}^{(K)} \) is admissible. In addition, for all \( \nu \in (0, 1] \),

\[ T_{\mathcal{F}}(\mathbb{P}_X^{(\nu)}; h) > T_{\mathcal{F}}(\mathbb{P}_X^{(0)}; h) + (\nu(1 - K) - 2\epsilon) \gamma_{\mathcal{F}}(\mathbb{P}_X^{(0)}, \mathbb{P}_X^{(1)}), \]

where \( \mathbb{P}_X^{(\nu)} = \nu \mathbb{P}_X^{(1)} + (1 - \nu) \mathbb{P}_X^{(0)} \).

A proof is given in Appendix A. Theorem 1 states that with \( h \) that is \( K \)-Lipschitz continuous w.r.t. \( \gamma_{\mathcal{F}} \), \( T_{\mathcal{F}}(\mathbb{P}_X^{(0)}; h) \) is smaller than \( T_{\mathcal{F}}(\mathbb{P}_X^{(\nu)}; h) \), which is the definition of admissibility. The difference between the two is proportional to \( \gamma_{\mathcal{F}}(\mathbb{P}_X^{(0)}, \mathbb{P}_X^{(1)}) \), which reflects the level of contamination.

**Example 2.** (Revisit Example 1) Let \( \mathbb{P}_X^{(0)} \) and \( \mathbb{P}_X^{(1)} \) be distributions defined in Example 1. Let \( \mathcal{H}_{\gamma_{\mathcal{F}}}^{(K)} \) be the all \( K \)-Lipschitz continuous autoencoders that consist of input layer with two nodes, one hidden layer with one node, and output layer with two nodes, without any activation function. When \( d \) is Euclidean distance and \( \mathcal{F} = \mathcal{F}_d \), the anomaly detection framework equipped with \( T_{\mathcal{F}} \) and \( \mathcal{H}_{\gamma_{\mathcal{F}}}^{(K)} \) is admissible if \( \|\mu\|_2 > 4\sqrt{1 + (1 - K)^2}/(1 - K) \) where \( \|\cdot\|_p \) denotes the \( L_p \)-norm.
Following the scheme presented in Section 3.1, the anomaly score is $T_X(\delta_x; h(0)) = \gamma_X(\delta_x, h(0), h(0))$, where $h(0) \in H^{(K)}$ minimizes $T_X(P_X(0); h)$. In the following theorem, we describe a property of the proposed anomaly score, $T_X(\delta_x; h(0))$. Theorem 2 roughly states that if $\gamma_X(\delta_x, P_X(0))$ is larger than $\gamma_X(\delta_x, P_X(0))$ with the margin proportional to $\gamma_X(P_X(0), P_X(1))$, that is if $\delta_x$ is closer to $P_X$ than $\delta_x$, the anomaly score for $x'$ is smaller than that for $x$.

**Theorem 2.** If $T_X(P_X(0); h(0)) < \epsilon \gamma_X(P_X(0), P_X(1))$ for some positive $\epsilon$ and two data $x$ and $x'$ satisfy $\gamma_X(\delta_x, P_X(0)) > (1 + K)/(1 - K) \gamma_X(\delta_x, P_X(0)) + (2e/(1 - K)) \gamma_X(P_X(0), P_X(1))$, then $T_X(\delta_x; h(0)) > T_X(\delta_x; h(0))$.

### 3.4 Anomaly detection with loss function and Lipschitz continuous autoencoders with respect to Wasserstein distance

In this section, we provide a specific anomaly detection algorithm when $\gamma_X$ is the 1-Wasserstein distance, i.e., $\gamma_X(P_X(0), h(0)) := \max_{f \in \mathcal{F}_d} \|E_{X \sim P_X(0)} f(X) - E_{X \sim h(0)} f(X)\|_1$. Again, $\mathcal{F}_d$ is the set of all 1-Lipschitz continuous functions w.r.t. $d$ to define Wasserstein distance, and $h$ is a $K$-Lipschitz continuous autoencoder w.r.t. $\gamma_X$.

The proposed algorithm utilizes an autoencoder $h$ that minimizes $T_{\mathcal{F}_d}(P_X(0); h) := \max_{f \in \mathcal{F}_d} \|E_{X \sim P_X(0)} f(X) - E_{X \sim h(0)} f(X)\|_1$ under the constraint that $h$ is in $H^{(K)}_{\mathcal{F}_d}$, a set of $K$-Lipschitz continuous autoencoders w.r.t. Wasserstein distance. This procedure is admissible. To enforce the $K$-Lipschitz continuity w.r.t. Wasserstein distance, we employ the Lemma A.1 of Patrini et al. [22] that links the Lipschitz continuity w.r.t. metric on $\mathcal{X}$ and Lipschitz continuity w.r.t. Wasserstein distance.

**Lemma 3.** (Lemma A.1 of Patrini et al. [22]) For any autoencoder $h : \mathcal{X} \rightarrow \mathcal{X}$ that is $K$-Lipschitz continuous w.r.t. $d$, $h$ is a $K$-Lipschitz continuous autoencoder w.r.t. $\gamma_X$.

Motivated by the Lemma 3, we propose to build autoencoders that minimize $T_{\mathcal{F}_d}(P_X(0); h)$ with a penalty term enforcing $K$-Lipschitz continuity of $h$ w.r.t. $d$. To handle the intractability of $T_{\mathcal{F}_d}(P_X(0); h)$, we employ the approach of Tolstikhin et al. [30] based on the primal form of Wasserstein distance, $\inf_{\pi \in \Pi(P, Q)} E_{(A, B) \sim \pi} d(A, B)$ where $\Pi(P, Q)$ is the set of all couplings of $P$ and $Q$. We minimize an approximated primal form of the Wasserstein distance. Let $P_Z$ be a user-specified prior distribution defined on $Z$, the space of low-dimensional representation. We denote encoder and decoder of $h$ by $h_{\text{Enc}} : \mathcal{X} \rightarrow Z$ and $h_{\text{Dec}} : Z \rightarrow \mathcal{X}$, respectively. Using the primal form of Wasserstein distance, an approximation of $T_{\mathcal{F}_d}(P_X(0); h)$ with penalty term enforcing Lipschitz continuity is

$$T_{\mathcal{F}_d}(P_X(0); h) + \phi \int_{\mathcal{X}} \int_{\mathcal{X}} \max \left( \frac{d(h(x_1), h(x_2))^2}{d(x_1, x_2)^2} - K^2, 0 \right) dP_X(0)(x_1)dP_X(0)(x_2)$$

$$\approx \int_{\mathcal{X}} d(x, h(x)) dP_X(0)(x) + \lambda \text{MMD}(P_Z, h_{\text{Enc}}(P_X(0)))$$

$$+ \phi \int_{\mathcal{X}} \int_{\mathcal{X}} \max \left( \frac{d(h(x_1), h(x_2))^2}{d(x_1, x_2)^2} - K^2, 0 \right) dP_X(0)(x_1)dP_X(0)(x_2),$$

where MMD denote the maximum mean discrepancy, $\lambda$ and $\phi$ are hyperparameters. In the right-hand side of (2), the first two terms appear because we use primal form with a constraint for the encoded values as in Tolstikhin et al. [30], and the final term is for the $K$-Lipschitz continuity of $h$. In implementation, we use $\int_{\mathcal{X}} d(x, h(x))^2 dP_X(0)(x)$, a common choice for reconstruction error, instead of the first term in the right-hand side of (2). This enforces to minimize $\int_{\mathcal{X}} d(x, h(x))^2 dP_X(0)(x)$ by Jensen’s inequality. The Algorithm 1 presents the process of training $K$-Lipschitz continuous autoencoders w.r.t. $\gamma_X$ where $d$ is Euclidean distance.

After training the $K$-Lipschitz continuous autoencoder $h(0)$, for a given test datum $x$, the anomaly score is $T_{\mathcal{F}_d}(\delta_x; h(0)) = d(x, h(0)(x))$. We propose to declare $x$ to be abnormal datum when $T_{\mathcal{F}_d}(\delta_x; h(0))$ is larger than a preset threshold.
We set normal and abnormal classes as follows. In KDD99 dataset, since "attack" flow is the majority, we treat the "normal" flow as the normal class, as in Zong et al. [36] and Zenati et al. [34]. In MNIST and Fashion-MNIST, we employ one-class classification setup [23, 34]. For each class, we set the "normal" class to normal and all others classes to abnormal. In CelebA, we set images with and without glasses to be abnormal and normal, respectively, to evaluate the ability to detect unexpected objects. Since wearing glasses highly depends on gender, only images of male celebrities are used.

Algorithm 1: Learning $K$-Lipschitz continuous autoencoder w.r.t. $\gamma_{f_d}$ where $d$ is Euclidean distance.

**Input:** Training dataset $X$, prior distribution $P_Z$, batch size $B$, positive definite kernel $k$, encoder $h_{Enc}(\cdot; w_{Enc})$, decoder $h_{Dec}(\cdot; w_{Dec})$, and hyperparameters $\lambda > 0$, $\phi > 0$, and $0 < K < 1$.

**Output:** A $K$-Lipschitz continuous autoencoder w.r.t. $\gamma_{f_d}$.

1: Initialize $(w_{Enc}, w_{Dec})$.
2: while $(w_{Enc}, w_{Dec})$ not converges:
3: Sample $x_1, \ldots, x_B$ from $X$.
4: Sample $z_1, \ldots, z_B$ following $P_Z$.
5: $\tilde{z}_i \leftarrow h_{Enc}(x_i; w_{Enc})$ for $i = 1, \ldots, B$.
6: $\tilde{x}_i \leftarrow h_{Dec}(\tilde{z}_i; w_{Dec})$ for $i = 1, \ldots, B$.
7: $\text{ReconError} \leftarrow B^{-1} \sum_{i=1}^{B} ||x_i - \tilde{x}_i||^2_2$.
8: $\text{LipschitzPenalty} \leftarrow B^{-1}(B - 1)^{-1} \sum_{i \neq j}^{B} \max(||x_i - x_j||^2_2/||x_i - x_j||^2_2 - K^2, 0)$.
9: $(w_{Enc}, w_{Dec}) \leftarrow \text{Adam}(\nabla_{w_{Enc}, w_{Dec}}\text{ReconError} + \lambda \text{MMD} + \phi \text{LipschitzPenalty})$.

4 Experiments

We demonstrate the efficacy of the proposed method with two experiments: (i) a performance comparison among the anomaly detection algorithms, (ii) an illustration of the proposed method. In the first experiment, we consider the two cases where training dataset is either uncontaminated or contaminated. Second, we visualize the effect of the proposed autoencoders on anomalies by comparing abnormal images and their reconstructions.

4.1 Dataset description

We conduct experiments with four datasets; KDD99 [17], MNIST [15], Fashion-MNIST [31], and CelebA [18]. KDD99 is a large-scale network-traffic dataset, widely used benchmark dataset in anomaly detection [28, 34, 36]. MNIST and Fashion-MNIST are image datasets commonly used to evaluate the anomaly detection performance on image recognition-based anomaly detection [23, 34]. CelebA is a face image dataset, and used to evaluate the applicability in face recognition-based anomaly detection. Detailed description about the datasets is attached in Appendix B.1.

4.2 Experiment setting

We set normal and abnormal classes as follows. In KDD99 dataset, since "attack" flow is the majority, "normal" flow is treated as abnormal class, as in Zong et al. [36] and Zenati et al. [34]. In MNIST and Fashion-MNIST, we employ one-class classification setup [23, 34]. For each class, we set the class to normal and all others classes to abnormal. In CelebA, we set images with and without glasses to be abnormal and normal, respectively, to evaluate the ability to detect unexpected objects. Since wearing glasses highly depends on gender, only images of male celebrities are used.

The proportion of training, validation, and test set is 50%, 25%, and 25%, respectively, for KDD99 and CelebA, and 60%, 20%, and 20%, respectively, for MNIST and Fashion-MNIST. We control the proportion of abnormal data on training and validation sets by randomly removing some anomalies. The level of contamination is chosen from $\{0, 0.05\}$. We call experiments for proportions of 0% and 5% as uncontaminated training dataset and contaminated training dataset, respectively.

All the models used in experiments are trained in unsupervised fashion without any labels. The validation set is for monitoring the overfitting in terms of loss function. For each method, we report the performance evaluated with test data. Evaluation metrics of anomaly detection task are area under the receiver operating characteristic curve (AUC) and area under the precision-recall curve (AUPRC). We compare the proposed method with two state-of-the-arts anomaly detection methods, deep SVDD and ALAD discussed in Section 2.
4.3 Architecture

For the proposed method, we use autoencoders without any convolutional layers for KDD99, and convolutional autoencoders for image datasets. For other baseline methods, we employ the following policy: For given dataset and baseline method, if an experiment about the dataset is implemented in the published work of the method, then we use almost the same architecture in the published work. Otherwise, we use architectures similar to the our method or in the published work of the baseline method. Including specifics of architecture, implementation details are provided in Appendix B.2.

4.4 Results

Table 1: Average AUC in % with standard deviation (over 10 runs for KDD99 and 5 runs for CelebA dataset) of various methods.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Uncontaminated training dataset (0%)</th>
<th>Contaminated training dataset (5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>deep SVDD</td>
<td>ALAD</td>
</tr>
<tr>
<td>KDD99</td>
<td>93.12±13.6</td>
<td>97.86±1.6</td>
</tr>
<tr>
<td>CelebA</td>
<td>47.71±4.0</td>
<td>53.54±0.5</td>
</tr>
</tbody>
</table>

Table 2: Average AUPRC in % with standard deviation (over 10 runs for KDD99 and 5 runs for CelebA dataset) of various methods.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Uncontaminated training dataset (0%)</th>
<th>Contaminated training dataset (5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>deep SVDD</td>
<td>ALAD</td>
</tr>
<tr>
<td>KDD99</td>
<td>90.14±15.1</td>
<td>84.80±9.0</td>
</tr>
<tr>
<td>CelebA</td>
<td>12.45±1.2</td>
<td>14.63±0.5</td>
</tr>
</tbody>
</table>

Figure 1: Visualization of the effect of Lipschitz continuous autoencoders on anomalies. The first row presents abnormal images sampled from the test set of CelebA and the second row presents corresponding reconstruction.

Tables 1 and 2 show a comparison of AUC and AUPRC of deep SVDD, ALAD, and the proposed method in cases of the uncontaminated and contaminated training dataset. The proposed method achieves the best mean AUC and the best mean AUPRC in all cases with KDD99 and CelebA. In addition, the proposed method outperforms in most of the cases for MNIST and Fashion-MNIST dataset. Performances on MNIST and Fashion-MNIST are presented in Appendix C in the supplementary material.

We visualize the effect of Lipschitz continuous autoencoders on anomalies. Figure presents randomly sampled abnormal images and their reconstructed images by the proposed autoencoders from CelebA dataset. The reconstruction process removes the anomalous part (glasses) of abnormal data.

5 Conclusion

In this work, we formalize anomaly detection methods with a loss function and a function space of models, and characterize a desirable property of anomaly detection frameworks, admissibility. We then propose an admissible anomaly detection frameworks equipped with an IPM-based loss function and a class of Lipschitz continuous autoencoders. To detect anomalies, we propose an anomaly detection algorithm based on the proposed admissible anomaly detection framework. The proposed algorithm outperforms state-of-the-art anomaly detection methods on the three datasets.
References


