Lipschitz Continuous Autoencoders in Application to Anomaly Detection

Young-geun Kim
Department of Statistics
Seoul National University

Yongchan Kwon
Department of Statistics
Seoul National University

Hyunwoong Chang
Department of Statistics
Texas A&M University

Myunghee Cho Paik
Department of Statistics
Seoul National University

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, build a theoretical framework of anomaly detection algorithms equipped with an objective function and a hypothesis space, and establish a desirable property of the anomaly detection algorithm, namely, *admissibility*. *Admissibility* implies that optimal autoencoders for normal data yield a larger reconstruction error for anomalous data than that for normal data on average. We then propose a class of *admissible* anomaly detection algorithms equipped with an integral probability metric-based objective function and a class of autoencoders, *Lipschitz continuous autoencoders*. The proposed algorithm for Wasserstein distance is implemented by minimizing an approximated Wasserstein distance with a penalty to enforce Lipschitz continuity with respect to Wasserstein distance. Through ablation studies, we demonstrate the efficacy of enforcing Lipschitz continuity of the proposed method. The proposed method is shown to be more effective in detecting anomalies than existing methods via applications to network traffic and image datasets [1].

[1]The work was performed when all authors belong to the department of Statistics, Seoul National University.

Proceedings of the 23[rd]International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

1 INTRODUCTION

Anomaly detection is the problem of identifying observations that deviate from the majority of the data in the absence of labeled data (Chandola et al., 2009; Hodge and Austin, 2004; Markou and Singh, 2003). To identify alarming situations, anomaly detection has been applied in fraud detection (Chan and Stolfo, 1998), medical diagnosis (Kononenko, 2001), network security (Garcia-Teodoro et al., 2009), and visual surveillance (Hu et al., 2004).

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 use only normal data to train a model and construct anomaly scores utilizing the trained model (Chandola et al., 2009; Khan and Madden, 2014). These methods assume that all training data belong to the normal class, and estimate the support (Schölkopf et al., 2001) or likelihood function of observations (Laxhammar et al., 2009). The data that are far from the support or have low values of likelihood are then classified as anomalies.

One approach to handle anomaly problems is to learn autoencoders minimizing the expected reconstruction error over the distribution of normal data, and utilize reconstruction error-based anomaly scores. In general, common elements of anomaly detection algorithms include specifying a hypothesis space of models with an objective function to obtain its optimizer and computing an anomaly score based on the objective function value of a test datum evaluated at the optimizer. With a similar focus, many deep neural network-based anomaly detection methods have been proposed, utilizing various objective functions to optimize neural networks with normal data alone (Chalapathy and Chawla, 2019). Details on deep learning-based anomaly detection algorithms are in Section 2. 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.

In this work, we formalize and build a class of anomaly detection algorithms fully identified by an objective function and a hypothesis space. Anomaly score can be defined as the contribution of a test datum to the objective function evaluated at the optimizer, which is a by-product once the objective function and the hypothesis space are determined. We show that given the cost-based objective function and optimizer, the anomaly score can be interpreted as an influence-like function used in robust 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 algorithms, namely *admissibility*. This property in words implies that the expected cost for anomalous data through the autoencoders optimized with only normal data is larger than the expected cost for the normal data. Existing methods are based on the premise that *admissibility* holds, but we show that this property does not hold in general. We then propose anomaly detection algorithms equipped with a new class of objective function based on integral probability metric (IPM) and a new class of hypothesis space for autoencoders, namely, *Lipschitz continuous autoencoders*. The proposed algorithm is based on the Lipschitz continuity of autoencoders with respect to (w.r.t.) IPM to guarantee *admissibility*. Besides, the anomaly score of proposed methods preserves rankings of the distance of each datum from the distribution of normal data and reflects the rate of change of the objective function by each datum. In particular, we take Wasserstein distance and provide a specific algorithm enforcing Lipschitz-continuity w.r.t. Wasserstein distance. Our contribution consists of four elements as follows.

- We build a theoretical framework for anomaly detection algorithms and characterize a desirable property of the algorithms, namely, *admissibility*. *Admissibility* implies that the expected cost for anomalous data evaluated at an optimizer for normal data is higher than the expected cost for the normal data. (Sections 3.1 and 3.2)
- We propose a class of autoencoders, *Lipschitz continuous autoencoders*. Anomaly detection algorithms equipped with an IPM-based objective function and the proposed class of autoencoders are *admissible*. (Section 3.3)
- We implement the proposed algorithm for Wasserstein distance by enforcing Lipschitz continuity of autoencoders w.r.t. Wasserstein distance. (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 *admissible* anomaly detection algorithms via Lipschitz continuous autoencoders. Section 4 demonstrates the application of the proposed method on network traffic and image data and reports results from ablation studies. 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 the entire training data and then compute anomaly scores based on extracted features from the trained model.

Inspired by support vector data description (SVDD) (Tax and Duin, 2004), deep SVDD (Ruff et al., 2018) has been proposed, replacing kernel feature mapping with neural networks mapping. The objective function of deep SVDD is the expected distance of extracted features from the centroid of the normal data cluster. Though deep SVDD is motivated by support vector algorithms, outputs of neural networks do not directly relate to the kernel, and the performance on CIFAR-10 (Krizhevsky and Hinton, 2009) is similar to kernel density estimation (Parzen, 1962) which often does not work well in high-dimensional cases.

Approaches based on generative adversarial networks (GANs) (Goodfellow et al., 2014) have been proposed. The generator minimizes the Jensen-Shannon divergence between distributions of normal data and generated data. Anomaly detection with generative adversarial networks (AnoGANs) is a state-of-the-art GAN-based anomaly detection method, utilizing reconstruction error-based anomaly scores (Schlegl et al., 2017). The performance on a clinical image was prominent, but calculating reconstruction errors requires solving an optimization problem for every test datum to find the latent code. Zenati et al. (2018) proposed adversarially learned anomaly detection (ALAD) based on adversarially learned inference with conditional entropy (Li et al., 2017), a kind of GANs including encoder networks which directly map a datum to latent code. ALAD considers the composition of en-

coder and generator as autoencoders, and introduces a penalty term to enforce a better reconstruction. In applications to KDD99 (Lichman, 2013) 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 (Sakurada and Yairi, 2014; Xu et al., 2015; Zhou and Paffenroth, 2017). Instead of the reconstruction error, An and Cho (2015) proposed to use the reconstruction probability from variational autoencoders (Kingma and Welling, 2013) as anomaly scores. Deep autoencoding Gaussian mixture model (DAGMM) is a state-of-the-art autoencoder-based anomaly detection model (Zong et al., 2018), invoking the Gaussian mixture assumption of the vector that consists of representations. DAGMM utilizes a likelihood-based energy function to identify anomalies. Performances on the KDD99 and other benchmark image datasets were not on a par with ALAD.

These approaches are based on the premise that the objective function of abnormal data evaluated at optimizers for normal data will be higher than the objective function 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, we characterize a desirable property called *admissibility* in Section 3.2. Section 3.3 presents the proposed anomaly detection algorithms with *Lipschitz continuous autoencoder* (Definition 2), and shows that Lipschitz continuity on the autoencoders w.r.t distribution metric can guarantee *admissibility* (Theorem 1). Throughout this section, we discuss Lipschitz continuity in three different contexts. To avoid confusion, we provide basic notations and clarify the differences upfront.

We denote random variables for input data, a compact domain of input data, and the set of all Borel probability measures defined on the domain by X, $\mathcal{X}$, and $\Pi_{\mathcal{X}}$, respectively. The distributions of normal and abnormal data are denoted by $\mathbb{P}_X^{(0)}$ and $\mathbb{P}_X^{(1)}$, respectively, and the Dirac delta function that gives all mass to $x \in \mathcal{X}$ is denoted by δ_x. The push-forward operation transferring a probability measure with a function h is denoted by $h\#$. With this notation, for a given random variable A following $\mathbb{P}_A$, $h\#\mathbb{P}_A$ is the distribution of $h(A)$. For given two distributions $\mathbb{P}$ and $\mathbb{Q}$, and a class of functions $\mathcal{F}$, we denote IPM by $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) := \sup_{f \in \mathcal{F}} |\mathbb{E}_{X\sim\mathbb{P}} f(X) - \mathbb{E}_{X\sim\mathbb{Q}} f(X)|$.

For given metric d on $\mathcal{X}$ and positive real number K,

1. We call a function $h : \mathcal{X} \to \mathcal{X}$ is K-Lipschitz continuous w.r.t. d if $d(h(x), h(y)) \leq K d(x, y)$ for all $x, y \in \mathcal{X}$.
2. Wasserstein distance is $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})$ 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. We call a push-forward operation $h\# : \Pi_{\mathcal{X}} \to \Pi_{\mathcal{X}}$ is K-Lipschitz continuous w.r.t. $\gamma_{\mathcal{F}}$ if $\gamma_{\mathcal{F}}(h\#\mathbb{P}, h\#\mathbb{Q}) \leq K\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})$ for all $\mathbb{P}, \mathbb{Q} \in \Pi_{\mathcal{X}}$.

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. (2018) 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

In this subsection, we build a theoretical framework of anomaly detection algorithms. The goal of anomaly detection is to sort out anomalies from normal data. Many existing anomaly detection algorithms build a model, train the model with a specific objective function using the normal data alone, and construct anomaly scores based on the trained model. Common elements of anomaly detection algorithms include the objective function and the hypothesis space. Table 1 shows some examples. Except for variations due to regularization in the objective function, the listed methods use the specified objective functions and hypothesis spaces to derive the anomaly scores.

We formalize current practices as follows. The hypothesis space is denoted by $\mathcal{H}$ and the objective function is denoted by $T : \Pi_{\mathcal{X}} \times \mathcal{H} \to \mathbb{R}$, where $\mathbb{R}$ is the set of all real numbers. First, objective functions utilized in anomaly detection methods are expressed as $T(\mathbb{P}_X^{(0)}, h)$ where $h \in \mathcal{H}$. The next step is to find $h^{(0)}$ that minimizes $T(\mathbb{P}_X^{(0)}, h)$. The final step is to construct the anomaly score for a test datum x_0 by utilizing $h^{(0)}$. In many cases, anomaly score is $T(\delta_{x_0}, h^{(0)})$, the objective function value at the test datum evaluated at $h^{(0)}$, so anomaly detection procedures can be fully characterized by T and $\mathcal{H}$.

Table 1: Objective functions and hypothesis spaces of various anomaly detection methods.

Method	Objective function	Hypothesis space
Laxhammar et al. (2009)	Expected negative log-likelihood	Probability density functions
Ruff et al. (2018)	Expected distance from the centroid	Neural networks
Sakurada and Yairi (2014) Xu et al. (2015)	Expected reconstruction error	Autoencoders
Zenati et al. (2018)	Jensen-Shannon divergence	Pairs of encoder, discriminator, and generator.

With the constructed framework, anomaly scores in current practices can be interpreted as the rate of change of the objective function in the direction of a test datum. Many anomaly detection algorithms use objective functions expressed as $T_c(\mathbb{P}_X^{(0)}, h) := \int_{\mathcal{X}} c(x,h) d\mathbb{P}_X^{(0)}(x)$ where c is a cost function such as negative log-likelihood, distance from the centroid of the data cluster, and reconstruction error. We show in the following proposition that given the objective function T_c and optimizer $h^{(0)}$, anomaly score, $T_c(\delta_x, h^{(0)})$ reflects the rate of change of the expected cost in the direction of a test datum, a limit of which is known as an influence function in robust statistics (Cook and Weisberg, 1980). Influence function is a kind of Gateaux derivative and quantifies the effect of a datum to a functional whose input is a probability measure.

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

$$T_c(\delta_x, h) = \frac{T_c(\nu\delta_x + (1-\nu)\mathbb{P}_X^{(0)}, h) - T_c(\mathbb{P}_X^{(0)}, h)}{\nu} + T_c(\mathbb{P}_X^{(0)}, h).$$

That is, the anomaly score of algorithms identified by $T_c(\mathbb{P}_X^{(0)}, h)$ and $\mathcal{H}$ can be decomposed by $T_c(\mathbb{P}_X^{(0)}, h)$ and the rate of change of the expected cost in the direction of δ_x.

3.2 Admissibility: A Desirable Property Of Anomaly Detection Algorithm

In this subsection, we characterize a desirable property of $T(\mathbb{P}_X^{(0)}, h)$ and $\mathcal{H}$, thus of anomaly detection methods, namely, *admissibility*. Admissibility implies that optimal autoencoders for normal data yield a larger reconstruction error for anomalous data than that for normal data on average. We give heuristic and theoretical motivations of admissibility. Heuristically, current practices find h to minimize the objective function for normal data, construct anomaly scores using the values obtained from the training, and declare anomaly if the anomaly score of respective method is large. Since anomaly scores can be viewed as a contribution of a datum to the objective function, the property of anomaly scores being larger for anomalous datum should be reflected in the characteristic of the objective function and hypothesis space. This requirement is one form of admissibility (Proposition 3). Theoretically, for a special case presented in Proposition 2, admissibility, formally defined in Definition 1, holds. We denote Kullback-Leibler divergence between distributions $\mathbb{P}$ and $\mathbb{Q}$ by $\mathcal{D}_{\mathrm{KL}}(\mathbb{P}||\mathbb{Q}) := \int_{\mathcal{X}} \log\big((d\mathbb{P}(x)/dx)/(d\mathbb{Q}(x)/dx)\big) d\mathbb{P}(x)$, and Shannon entropy of a distribution $\mathbb{P}$ by $S(\mathbb{P}) := \int_{\mathcal{X}} -\log(d\mathbb{P}(x)/dx) d\mathbb{P}(x)$. Then, the expected negative log-likelihood evaluated with a distribution $\mathbb{P}$ and a probability density function h can be expressed as $T_L(\mathbb{P}, h) = \mathcal{D}_{\mathrm{KL}}(\mathbb{P}||H) + S(\mathbb{P})$ where H a distribution function associated to h. We also denote $\mathbb{P}_X^{(\nu)} := (1-\nu)\mathbb{P}_X^{(0)} + \nu\mathbb{P}_X^{(1)}$.

Proposition 2. *Let T_L be the expected negative log-likelihood and $\mathcal{H}$ be the set of all probability density functions defined on $\mathcal{X}$. If $S(\mathbb{P}_X^{(1)}) \geq S(\mathbb{P}_X^{(0)})$, then for any $\nu \in (0,1)$ satisfying $\mathcal{D}_{KL}(\mathbb{P}_X^{(1)}||\mathbb{P}_X^{(\nu)}) > \mathcal{D}_{KL}(\mathbb{P}_X^{(0)}||\mathbb{P}_X^{(\nu)})$, we have*

$$T_L(\mathbb{P}_X^{(\nu)}, h^{(0)}) > T_L(\mathbb{P}_X^{(0)}, h^{(0)}).$$

Roughly speaking, when normal data are clustered so that $S(\mathbb{P}_X^{(1)}) \geq S(\mathbb{P}_X^{(0)})$, and the contamination proportion is small enough so that $\mathcal{D}_{\mathrm{KL}}(\mathbb{P}_X^{(1)}||\mathbb{P}_X^{(\nu)}) > \mathcal{D}_{\mathrm{KL}}(\mathbb{P}_X^{(0)}||\mathbb{P}_X^{(\nu)})$, then $T_L(\mathbb{P}_X^{(\nu)}, h^{(0)}) > T_L(\mathbb{P}_X^{(0)}, h^{(0)})$. Motivated by Proposition 2, we present the formal definition of *admissibility* as follows.

Definition 1. *Anomaly detection algorithm equipped with an objective function T and a hypothesis space $\mathcal{H}$ is said to be admissible when $T(\mathbb{P}_X^{(\nu)}, h^{(0)}) > T(\mathbb{P}_X^{(0)}, h^{(0)})$ for some $\nu \in (0,1]$ and any $h^{(0)} \in \mathcal{H}$ satisfying $T(\mathbb{P}_X^{(0)}, h^{(0)}) \leq T(\mathbb{P}_X^{(0)}, h)$ for all $h \in \mathcal{H}$.*

That is, an anomaly detection algorithm is *admissible* if the objective function evaluated at $h^{(0)}$ for some contaminated data is larger than that for uncontaminated data. For the subclass of anomaly detection algorithm equipped with T_c, the following proposition shows the

condition for admissibility, that is, the objective function for abnormal data is larger than that for normal data.

Proposition 3. *The anomaly detection algorithm equipped with T_c and $\mathcal{H}$ is admissible if and only if $T_c(\mathbb{P}_X^{(1)}, h^{(0)}) > T_c(\mathbb{P}_X^{(0)}, h^{(0)})$ for any $h^{(0)} \in \mathcal{H}$ satisfying $T_c(\mathbb{P}_X^{(0)}, h^{(0)}) \leq T_c(\mathbb{P}_X^{(0)}, h)$ for all $h \in \mathcal{H}$.*

Although admissibility seems to be a natural property that any anomaly detection algorithm can enjoy, it is not guaranteed. Below we present a simple example of anomaly detection methods equipped with $T_c(\mathbb{P}_X^{(0)}, h)$ and a set of autoencoders $\mathcal{H}$, where the admissibility may not hold.

Example 1. Let $\mathbb{P}_X^{(0)}$ and $\mathbb{P}_X^{(1)}$ be distributions of two dimensional 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 2×2 identity matrix, and $\mu \in \mathbb{R}^2$. Let $\mathcal{H}$ be the set of all autoencoders that consist of an input layer with two nodes, one hidden layer with one node, and an output layer with two nodes, without any activation function. When c is squared L_2-norm, the anomaly detection algorithm equipped with T_c and $\mathcal{H}$ is *not admissible* for all μ.

3.3 Admissible Anomaly Detection via Lipschitz Continuous Autoencoders

In this subsection, we propose admissible anomaly detection algorithms equipped with a new objective function and a new hypothesis space of autoencoders. Hereafter we focus on $h : \mathcal{X} \to \mathcal{X}$ and call h autoencoders for convenience. Both proposed objective function and hypothesis space are based on IPM, a family of distribution metrics based on the comparison of integrals w.r.t. probability measures (Müller, 1997). IPM includes many novel metrics such as Wasserstein distance and maximum mean discrepancy (Sriperumbudur et al., 2009). Again, for a given class of functions $\mathcal{F}$, the IPM is denoted by $\gamma_{\mathcal{F}}$. The theoretical strength of the proposed algorithm consists of three parts; The proposed algorithm is admissible (Theorem 1), the anomaly score has the rank-preserving property as described in Theorem 2, and the anomaly score reflects an influence-like function as in the cost-based objective function (Proposition 4).

We first describe a special kind of hypothesis space for autoencoders, namely, *Lipschitz continuous autoencoders*. The notion of Lipschitz continuous autoencoders is based on the Lipschitz continuity in a metric space $(\Pi_{\mathcal{X}}, \gamma_{\mathcal{F}})$. Now, 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 $\mathbb{P}, \mathbb{Q} \in \Pi_{\mathcal{X}}$.*

We denote the set of all K-Lipschitz continuous autoencoders w.r.t. $\gamma_{\mathcal{F}}$ by $\mathcal{H}_{\mathcal{F}}^{(K)}$. Again, we emphasize that usual Lipschitz continuity w.r.t d is different from Lipschitz continuity w.r.t $\gamma_{\mathcal{F}}$. Lipschitz continuity w.r.t Wasserstein distance is defined in Patrini et al. (2018) but w.r.t. $\gamma_{\mathcal{F}}$ has not been defined to the best of the authors' knowledge.

We then define an IPM-based objective function, a distance between the distribution of input data and that of reconstructed data measured by $\gamma_{\mathcal{F}}$, expressed as

$$T_{\mathcal{F}}(\mathbb{P}, h) := \gamma_{\mathcal{F}}(\mathbb{P}, h\#\mathbb{P}). \tag{1}$$

Theorem 1 shows that with a properly chosen h in $\mathcal{H}_{\mathcal{F}}^{(K)}$, the anomaly detection algorithm equipped with $T_{\mathcal{F}}(\mathbb{P}, h)$ and $\mathcal{H}_{\mathcal{F}}^{(K)}$ is *admissible*.

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

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

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*. A lower bound for the difference is proportional to $\gamma_{\mathcal{F}}(\mathbb{P}_X^{(0)}, \mathbb{P}_X^{(1)})$ and reflects the level of contamination ν.

Example 2. (Revisit Example 1) Let $\mathbb{P}_X^{(0)}$ and $\mathbb{P}_X^{(1)}$ be distributions defined in Example 1. Let $\mathcal{H}_{\mathcal{F}}^{(K)}$ be the all K-Lipschitz continuous autoencoders that consist of an input layer with two nodes, one hidden layer with one node, and an output layer with two nodes, without any activation function. When d is Euclidean distance and $\mathcal{F}$ is $\mathcal{F}_d$, the anomaly detection algorithm equipped with $T_{\mathcal{F}}$ and $\mathcal{H}_{\mathcal{F}}^{(K)}$ is admissible if $||\mu||_2 > 4\sqrt{1+(1-K)^2}/(1-K)$ where $||\cdot||_p$ denotes the L_p-norm.

In contrast to Example 1 where plain autoencoders are applied, admissibility is gained by choosing K-Lipschitz continuous autoencoders.

Following the scheme presented in Section 3.1, the anomaly score is $T_{\mathcal{F}}(\delta_x, h^{(0)}) = \gamma_{\mathcal{F}}(\delta_x, h^{(0)}\#\delta_x)$, where $h^{(0)} \in \mathcal{H}_{\mathcal{F}}^{(K)}$ minimizes $T_{\mathcal{F}}(\mathbb{P}_X^{(0)}, h)$. In the following theorem, we describe properties of the proposed

anomaly score, $T_{\mathcal{F}}(\delta_x, h^{(0)})$. Theorem 2 roughly states that if $\gamma_{\mathcal{F}}(\delta_x, \mathbb{P}_X^{(0)})$ is larger than $\gamma_{\mathcal{F}}(\delta_{x'}, \mathbb{P}_X^{(0)})$ with the margin $\gamma_{\mathcal{F}}(\mathbb{P}_X^{(0)}, \mathbb{P}_X^{(1)})$, that is if $\delta_{x'}$ is closer to $\mathbb{P}_X^{(0)}$ than δ_x, the anomaly score for x' is smaller than that for x.

Theorem 2. *Let $K \in (0,1)$ and $h \in \mathcal{H}_{\mathcal{F}}^{(K)}$ satisfy $T_{\mathcal{F}}(\mathbb{P}_X^{(0)}, h) < \epsilon\gamma_{\mathcal{F}}(\mathbb{P}_X^{(0)}, \mathbb{P}_X^{(1)})$ for some $\epsilon < (1-K)/2$. If two data x and x' satisfy $\gamma_{\mathcal{F}}(\delta_x, \mathbb{P}_X^{(0)}) > (1+K)/(1-K)\gamma_{\mathcal{F}}(\delta_{x'}, \mathbb{P}_X^{(0)}) + \gamma_{\mathcal{F}}(\mathbb{P}_X^{(0)}, \mathbb{P}_X^{(1)})$, then $T_{\mathcal{F}}(\delta_x, h^{(0)}) > T_{\mathcal{F}}(\delta_{x'}, h^{(0)})$.*

In addition, the following proposition shows that the anomaly score reflects the contribution of each data to the objective function as in cost-based objective function (Proposition 1).

Proposition 4. *For any x, class of functions $\mathcal{F}$, $\nu \in (0,1]$ and $h \in \mathcal{H}$,*

$$T_{\mathcal{F}}(\delta_x, h) \geq \frac{T_{\mathcal{F}}(\nu\delta_x + (1-\nu)\mathbb{P}_X^{(0)}, h) - T_{\mathcal{F}}(\mathbb{P}_X^{(0)}, h)}{\nu} + T_{\mathcal{F}}(\mathbb{P}_X^{(0)}, h).$$

That is, the anomaly score has a lower bound that increases as the rate of change of the objective function in the direction of datum increases.

3.4 Implementation of The Proposed Algorithm for Wasserstein Distance

In this subsection, we provide an implementation of proposed anomaly detection algorithms when $\gamma_{\mathcal{F}}$ is the 1-Wasserstein distance, *i.e.*, $\gamma_{\mathcal{F}_d}(\mathbb{P}_X^{(0)}, h\#\mathbb{P}_X^{(0)}) := \sup_{f \in \mathcal{F}_d} |\mathbb{E}_{X\sim\mathbb{P}_X^{(0)}} f(X) - \mathbb{E}_{X\sim h\#\mathbb{P}_X^{(0)}} f(X)|$. 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. $\gamma_{\mathcal{F}_d}$.

The proposed algorithm utilizes an autoencoder h that minimizes $T_{\mathcal{F}_d}(\mathbb{P}_X^{(0)}, h) := \gamma_{\mathcal{F}_d}(\mathbb{P}_X^{(0)}, h\#\mathbb{P}_X^{(0)})$ under the constraint that h is in $\mathcal{H}_{\mathcal{F}_d}^{(K)}$, 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. (2018) that links the Lipschitz continuity w.r.t. a metric on $\mathcal{X}$ and Lipschitz continuity w.r.t. Wasserstein distance.

Lemma 3. *(Lemma A.1 of Patrini et al. (2018)) For any autoencoder $h : \mathcal{X} \to \mathcal{X}$ that is K-Lipschitz continuous w.r.t. d, h is a K-Lipschitz continuous autoencoder w.r.t. $\gamma_{\mathcal{F}_d}$.*

By Lemma 3, Theorem 1 stays true when we replace $(T_{\mathcal{F}}, \mathcal{H}_{\mathcal{F}}^{(K)})$ with $(T_{\mathcal{F}_d}, \mathcal{H}_d^{(K)})$ where $\mathcal{H}_d^{(K)}$ is the set of Lipschitz continuous autoencoders w.r.t. d. Motivated by Lemma 3, we propose to build autoencoders that minimize $T_{\mathcal{F}_d}(\mathbb{P}_X^{(0)}, h)$ with a penalty term enforcing K-Lipschitz continuity of h w.r.t. d. To handle the intractability of $T_{\mathcal{F}_d}(\mathbb{P}_X^{(0)}, h)$, we employ the approach of Tolstikhin et al. (2017) based on the primal form of Wasserstein distance, $\inf_{\pi\in\Pi(\mathbb{P},\mathbb{Q})} \mathbb{E}_{(A,B)\sim\pi} d(A,B)$ where $\Pi(\mathbb{P},\mathbb{Q})$ is the set of all couplings of $\mathbb{P}$ and $\mathbb{Q}$. We minimize an approximated primal form of the Wasserstein distance. Let $\mathbb{P}_Z$ be a user-specified distribution defined on $\mathcal{Z}$, the space of low-dimensional representation. We denote encoder and decoder of h by $h_{\text{Enc}} : \mathcal{X} \to \mathcal{Z}$ and $h_{\text{Dec}} : \mathcal{Z} \to \mathcal{X}$, respectively. Using the primal form of Wasserstein distance, an approximation of $T_{\mathcal{F}_d}(\mathbb{P}_X^{(0)}, h)$ with penalty term $R(\mathbb{P}_X^{(0)}, K)$ that enforces Lipschitz continuity is

$$\begin{aligned} &T_{\mathcal{F}_d}(\mathbb{P}_X^{(0)}, h) + \phi R(\mathbb{P}_X^{(0)}, K) \\ &\approx \int_{\mathcal{X}} d(x, h(x)) d\mathbb{P}_X^{(0)}(x) + \lambda \text{MMD}(\mathbb{P}_Z, h_{\text{Enc}}\#\mathbb{P}_X^{(0)}) \\ &+ \phi\mathbb{E}_{(X,X')\sim\mathbb{P}_X^{(0)}\times\mathbb{P}_X^{(0)}} \max\big(\frac{d(h(X), h(X'))^2}{d(X,X')^2} - K^2, 0\big), \end{aligned} \tag{2}$$

where MMD denotes the maximum mean discrepancy, λ and ϕ are hyperparameters. The first two terms on the right-hand side of (2) appear because we use the primal form with a constraint for the encoded values as in Tolstikhin et al. (2017), and the final term is for the K-Lipschitz continuity of h. In implementation, we use $\int_{\mathcal{X}} d(x, h(x))^2 d\mathbb{P}_X^{(0)}(x)$, a common choice for reconstruction error, instead of the first term on the right-hand side of (2). This enforces to minimize $\int_{\mathcal{X}} d(x, h(x)) d\mathbb{P}_X^{(0)}(x)$ by Jensen's inequality. The Algorithm 1 presents the process of training K-Lipschitz continuous autoencoders w.r.t. $\gamma_{\mathcal{F}_d}$ where d is Euclidean distance. Here, Adam denotes the Adam optimizer (Kingma and Ba, 2014).

After training the K-Lipschitz continuous autoencoder $h^{(0)}$, for a given test datum x, the anomaly score is $T_{\mathcal{F}_d}(\delta_x, h^{(0)}) = d(x, h^{(0)}(x))$, the reconstruction error by the trained autoencoder. We propose to declare x to be abnormal datum when $T_{\mathcal{F}_d}(\delta_x, h^{(0)})$ is larger than a preset threshold such as a specific quantile of anomaly scores for normal data.

4 EXPERIMENTS

We demonstrate the efficacy of the proposed method with three experiments[2]: (i) illustration of the proposed method, (ii) an ablation study to evaluate the

[2]The implementation code is provided in https://github.com/kyg0910/Lipschitz-Continuous-Autoencoders-in-Application-to-Anomaly-Detection.

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

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

Output: A K-Lipschitz continuous autoencoder w.r.t. $\gamma_{\mathcal{F}_d}$.

Initialize $(w_{\text{Enc}}, w_{\text{Dec}})$.
while $(w_{\text{Enc}}, w_{\text{Dec}})$ not converges:
  Sample $x_1, ..., x_B$ from $\mathbb{X}$
  Sample $z_1, ..., z_B$ following $\mathbb{P}_Z$
  $\widetilde{z}_i \leftarrow h_{\text{Enc}}(x_i; w_{\text{Enc}})$ for $i = 1, \ldots, B$
  $\widetilde{x}_i \leftarrow h_{\text{Dec}}(\widetilde{z}_i; w_{\text{Dec}})$ for $i = 1, \ldots, B$
  ReconError $\leftarrow B^{-1} \sum_{i=1}^{B} ||x_i - \widetilde{x}_i||_2^2$
  MMD $\leftarrow B^{-2} \sum_{i,j=1}^{B} \big(k(z_i, z_j) - 2k(z_i, \widetilde{z}_j) + k(\widetilde{z}_i, \widetilde{z}_j)\big)$
  LipschitzPenalty $\leftarrow B^{-1}(B-1)^{-1} \sum_{l \neq j}^{B} \max(||\widetilde{x}_l - \widetilde{x}_j||_2^2 / ||x_l - x_j||_2^2 - K^2, 0)$
  $\big(w_{Enc}, w_{Dec}\big) \leftarrow$ Adam(ReconError + λMMD + ϕLipschitzPenalty)

effect of Lipschitz continuity, and (iii) a comparison with the existing anomaly detection algorithms. In the first experiment, we visualize the effect of the proposed autoencoders on anomalies by comparing abnormal images and their reconstructions. For the second, we compare anomaly detection performances from various levels of regularization in (2). For the third, we compare performances of deep SVDD, ALAD, and the proposed method. We also consider the case where the training dataset is contaminated by anomalies to validate the robustness of algorithms.

4.1 Dataset Description

We conduct experiments with four datasets; KDD99 (Lichman, 2013), MNIST (LeCun, 1998), Fashion-MNIST (Xiao et al., 2017), and CelebA (Liu et al., 2015). KDD99 is a large-scale network-traffic dataset, widely used benchmark dataset in anomaly detection (Zenati et al., 2018; Zong et al., 2018). MNIST and Fashion-MNIST are image datasets commonly used to evaluate the anomaly detection performance on image recognition-based anomaly detection (Ruff et al., 2018; Zenati et al., 2018). CelebA is a face image dataset and used to evaluate the applicability in face recognition-based anomaly detection. A detailed description of the datasets is attached in Appendix B.1.

4.2 Experiment Setting

We set normal and abnormal classes as follows. In KDD99, since "attack" flow is the majority, "normal" flow is treated as the abnormal class, as in Zong et al. (2018) and Zenati et al. (2018). In MNIST and Fashion-MNIST, we employ a one-class classification setup (Ruff et al., 2018; Zenati et al., 2018). For each class, we set one class to normal and all other 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.

We compare the proposed method with two state-of-the-arts anomaly detection methods, deep SVDD and ALAD discussed in Section 2. As in deep SVDD and ALAD papers, all the models are trained in unsupervised fashion. The labels of the validation set were also not used, and only the objective function is used to avoid overfitting. For each method, we report the performance evaluated with test data. Evaluation metrics are the area under the receiver operating characteristic curve (AUC) and the area under the precision-recall curve (AUPRC).

4.3 Architecture

For the proposed method, we use plain autoencoder architectures for KDD99 and convolutional autoencoders architectures for image datasets. For other methods, we use almost the same architecture in published work. For a given dataset, all methods use similar size of networks in terms of the number of layers and parameters. Implementation details are provided in Appendix B.2.

4.4 Results

We visualize the effect of Lipschitz continuous autoencoders on anomalies. Figure 1 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.

Table 2 and 3 show AUCs and AUPRCs, respectively, of ablation study to evaluate the effect of Lipschitz continuity imposed on autoencoders. The penalty term to enforce Lipschitz continuity in (2) has hyper-

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 the corresponding reconstruction.

Table 2: Average AUCs on KDD99 of the proposed method for various ϕ and K are provided in % with standard deviation. The number of replication is 10.

ϕ	K			
	0.6	0.7	0.8	0.9
0	98.6±0.6	98.6±0.6	98.6±0.6	98.6±0.6
5	99.2±0.1	99.3±0.1	99.3±0.1	98.1±1.1
10	99.2±0.1	**99.4±0.1**	96.9±1.3	95.3±2.4
20	96.6±0.6	89.4±1.6	91.9±0.3	96.8±1.8

parameters K and ϕ, and the level of enforcement increases as ϕ increases or K decreases. The ablation study is conducted with the uncontaminated dataset. To control noises from the approximation part, λ is set to 0. The baseline model, where ϕ is 0, is a plain autoencoder without enforcing Lipschitz continuity. Compared with the baseline model, a moderate level of Lipschitz continuity significantly enhances performance. The mean AUC is increased from 98.6 to 99.4, and the mean AUPRC, from 93.0 to 96.4 due to Lipschitz continuity. The standard deviation of AUC decreased from 0.6 to 0.1, and of AUPRC, from 2.6 to 0.3 due to Lipschitz continuity. Additional results for $\lambda \in \{0, 5, 10\}$, $\phi \in \{0, 5, 10, 20\}$, and $K \in \{0.1, 0.2, \ldots, 0.9\}$ are given in Appendix C in the supplementary material. For every λ, an adequately chosen ϕ and K significantly increased AUC and AUPRC.

Table 4 shows 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. Besides, 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 D in the supplementary material.

Table 3: Average AUPRCs on KDD99 of the proposed method for various ϕ and K are provided in % with standard deviation. The number of replication is 10.

ϕ	K			
	0.6	0.7	0.8	0.9
0	93.0±2.6	93.0±2.6	93.0±2.6	93.0±2.6
5	95.4±0.8	96.2±0.8	95.8±0.8	92.8±2.3
10	95.6±0.6	**96.4±0.3**	89.6±2.4	88.0±4.4
20	86.5±1.9	73.8±3.8	81.0±0.8	89.9±3.1

Table 4: Average AUCs and AUPRCs of deep SVDD, ALAD, and the proposed method are provided in % with standard deviation. The number of replication is 20 for KDD99 and 5 for CelebA.

Dataset	Method	AUC	AUPRC
Uncontaminated training dataset (0%)			
KDD99	deep SVDD	98.7±2.4	95.6±3.5
	ALAD	98.2±1.3	86.6±7.1
	Proposed	**99.3±0.2**	**96.2±1.0**
CelebA	deep SVDD	54.9±7.2	16.6±4.6
	ALAD	53.9±0.7	14.8±0.6
	Proposed	**65.7±0.4**	**22.0±0.3**
Contaminated training dataset (5%)			
KDD99	deep SVDD	75.9±19.7	52.4±16.5
	ALAD	96.4±1.6	76.8±8.1
	Proposed	**97.7±0.2**	**80.8±1.6**
CelebA	deep SVDD	52.4±5.7	14.1±1.6
	ALAD	53.8±0.7	14.8±0.6
	Proposed	**62.3±0.6**	**18.8±0.5**

5 CONCLUSION

In this work, we formalize anomaly detection methods with an objective function and a hypothesis space and characterize a desirable property of anomaly detection algorithms, *admissibility*. We then propose *admissible* anomaly detection algorithms equipped with an IPM-based objective function and a class of autoencoders, *Lipschitz continuous autoencoders*. By implementing the proposed method for Wasserstein distance, we present an admissible anomaly detection algorithm. The proposed algorithm outperforms state-of-the-art anomaly detection methods on KDD99, MNIST, Fashion-MNIST, and CelebA datasets.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2017R1A2B4008956).

References

An, J. and Cho, S. (2015). Variational autoencoder based anomaly detection using reconstruction probability. *SNU Data Mining Center, Tech. Rep.*

Chalapathy, R. and Chawla, S. (2019). Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*.

Chan, P. K. and Stolfo, S. J. (1998). Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. In *KDD*, volume 98, pages 164–168.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15.

Cook, R. D. and Weisberg, S. (1980). Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics*, 22(4):495–508.

Garcia-Teodoro, P., Diaz-Verdejo, J., Maciá-Fernández, G., and Vázquez, E. (2009). Anomaly-based network intrusion detection: Techniques, systems and challenges. *computers & security*, 28(1-2):18–28.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Hodge, V. and Austin, J. (2004). A survey of outlier detection methodologies. *Artificial intelligence review*, 22(2):85–126.

Hu, W., Tan, T., Wang, L., and Maybank, S. (2004). A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 34(3):334–352.

Khan, S. S. and Madden, M. G. (2014). One-class classification: taxonomy of study and review of techniques. *The Knowledge Engineering Review*, 29(3):345–374.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23(1):89–109.

Krizhevsky, A. and Hinton, G. (2009). Learning multiple layers of features from tiny images. Technical report, Citeseer.

Laxhammar, R., Falkman, G., and Sviestins, E. (2009). Anomaly detection in sea traffic-a comparison of the gaussian mixture model and the kernel density estimator. In *2009 12th International Conference on Information Fusion*, pages 756–763. IEEE.

LeCun, Y. (1998). The mnist database of handwritten digits. *http://yann.lecun.com/exdb/mnist/*.

Li, C., Liu, H., Chen, C., Pu, Y., Chen, L., Henao, R., and Carin, L. (2017). Alice: Towards understanding adversarial learning for joint distribution matching. In *Advances in Neural Information Processing Systems*, pages 5495–5503.

Lichman, M. (2013). Uci machine learning repository. irvine, ca: University of california, school of information and computer science. *URL http://archive. ics. uci. edu/ml*.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738.

Markou, M. and Singh, S. (2003). Novelty detection: a reviewpart 1: statistical approaches. *Signal processing*, 83(12):2481–2497.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443.

Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076.

Patrini, G., Carioni, M., Forre, P., Bhargav, S., Welling, M., Berg, R. v. d., Genewein, T., and Nielsen, F. (2018). Sinkhorn autoencoders. *arXiv preprint arXiv:1810.01118*.

Ruff, L., Görnitz, N., Deecke, L., Siddiqui, S. A., Vandermeulen, R., Binder, A., Müller, E., and Kloft, M. (2018). Deep one-class classification. In *International Conference on Machine Learning*, pages 4390–4399.

Sakurada, M. and Yairi, T. (2014). Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, page 4. ACM.

Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., and Langs, G. (2017). Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging*, pages 146–157. Springer.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. (2001). Estimating the

support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. (2009). On integral probability metrics,\phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*.

Tax, D. M. and Duin, R. P. (2004). Support vector data description. *Machine learning*, 54(1):45–66.

Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. (2017). Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*.

Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Xu, D., Ricci, E., Yan, Y., Song, J., and Sebe, N. (2015). Learning deep representations of appearance and motion for anomalous event detection. *arXiv preprint arXiv:1510.01553*.

Zenati, H., Romain, M., Foo, C.-S., Lecouat, B., and Chandrasekhar, V. (2018). Adversarially learned anomaly detection. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 727–736. IEEE.

Zhou, C. and Paffenroth, R. C. (2017). Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 665–674. ACM.

Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., and Chen, H. (2018). Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *In Proceedings of the 6th International Conference on Learning Representations*.