Theoretical Analysis

Assumption 1. *Outliers are sparser distributed than inliers.*

This assumption and its variants have been used in many previous studies. For example, kNN assumes that inliers' proximity is denser than that of outliers, and LOF detects local outliers based on the difference between points and their nearest neighbors.

Assumption 2. *The distribution of outliers has no overlaps with inliers.*

This assumption is essential for real-world applications since, in the absence of it, classifiers would be unable to differentiate between overlapped inliers and outliers.

Proposition 1. [Smoothness Prior] Given a distance function d, $\forall x_i, x_j \in \mathbb{R}^{dim}$, if $d\left(x_i, x_j\right) < \varepsilon$, $\forall \varepsilon$, $\left|f\left(x_i\right) - f\left(x_j\right)\right| < M\varepsilon$. $f(\cdot)$ is the prediction function under the principle of Smoothness Prior and $f(x)$ denotes the anomaly score of the sample x, and M is a positive value subject to the Lipschitz condition.

This proposition specifies that a learned classifier should follow the Smoothness Prior with the changing rate $f(x)$ across the whole value space below a certain threshold.

Lemma 1. A set of negative samples can be constructed such that the ratio of positive samples (inliers and outliers) to negative samples (uniform noise) is $1:n$. This ensures that the density of the noise is large than the density of outliers, and under our assumption, the density of inliers $\rho(x_i) \to +\infty, x_i \in X_n$. Thus, n is a finite value subject to the outlier density.

Proof. For $\forall x_{j1}, x_{j2} \in X_o, \forall x_{i1}, x_{i2} \in X_n$ we have $d(x_{j1}, x_{j2}) > 4\sqrt{dim}d(x_{i1}, x_{i2})$ where dim is the space dimension for the dataset and the 4 is a scaling factor. We let $D = \max_{i1,i2} d(x_{i1}, x_{i2})$, S be the dataset space, and $\rho(\cdot)$ be the density function. $\rho(x) = \max_y \frac{C(N(x,d(x,y)),x)}{Sqr(N(x,d(x,y)))}$, where x,y come from the same dataset, $N(x, d_x) = \{z | d(x, z) \leq d_x, z \in S\}$, $C(N, x)$ means the number of the data which has the same tag as x and is in the subset N, $Sqr(N)$ means the volume of the subset N. Construct the noise following a uniform distribution, in which the distance between two adjacent points is $4D$, we have:

$\min_{j1,k1} d(x_{j1}, x_{k1}) < \frac{\sqrt{dim}}{2} * 4D < 4\sqrt{dim}D < \min_{j1,j2} d(x_{j1}, x_{j2})$,

where $x_{j1}, x_{j2} \in X_o, x_{k1} \in X^-$. This indicates that noise is distributed near the outlier instead of the outlier and $\forall x_j \in X_o, \forall x_k \in X^-$, $\rho(x_k) > \rho(x_j)$. Given that $\max_{x \in S} \min_{x_{k1} \in X^-} d(x, x_{k1}) = 2\sqrt{dim}D$, we can generate the noise data to gurantee $\exists d(x_{i1}, x_{k1}) \geq 2D$. Then:

$\min_{i1,k1} d(x_{i1}, x_{k1}) > 2D - \max_{i1,i2} d(x_{i1}, x_{i2}) = \max_{i1,i2} d(x_{i1}, x_{i2})$.

It means that $\forall x_i \in X_n, \forall x_k \in X^-$, $\rho(x_k) < \rho(x_i)$. Thus, $\rho(x_j) < \rho(x_k) < \rho(x_i)$. In other words, it is always possible to generate noise that has a density between that of the inliers and outliers.

Proposition 2. [Sample Distances] Due to the sparsity of outliers, for $\forall x_i \in X_n, x_j \in X_o$, we can construct a set of uniform noise X^-, so as $\forall x_{j1}, x_{j2} \in X_o$, $\exists x_k, x_{k1}, x_{k2} \in X^-$, then, $d(x_i, x_j) \gg d(x_i, x_k), d(x_{j1}, x_{j2}) \gg d(x_{k1}, x_{k2})$.

This proposition suggests that noise needs to be interspersed between the inliers and outliers. The density of negative samples should be less than inliers and larger than outliers.

Lemma 2. Let $D = \max_j \left(\min_k d(x_j, x_k)\right)$, where $x_j \in X_o, x_k \in X^-$. Then, $\forall x_j \in X_o, f(x_j) \geq 1 - MD, MD < 1$.

Proof. If $\exists x_j \in X_o$, s.t. $f(x_j) < 1 - MD$, with smoothness prior (i.e., x_j, x_k are in a subspace with $d(x_j, x_k) < \epsilon$, thus, $f(x_j) \to f(x_k)$), $\exists x_k \in X^-, f(x_k) < 1 - MD + Md(x_j, x_{k^*}) < 1$. x_{k^*} is the closest of $x_k \in X^-$ to x_j. Thus, there is an optimal classification value $f^*(x)$, so that $f(x_k) \leq f^*(x_k), \forall x_k \in X^-$, and we further define $f^*(x)$ as:

$$f^*(x) = \begin{cases} 1, & \forall x \in X^-, \\ f(x), & \forall x \in X_n, \\ 1 - MD_x, & \forall x \in X_o. \end{cases} \tag{2-1}$$

$$\mathcal{L}_f = -\Big(\sum_{i=0}^{|X|} \log(1 - f(x_i)) + \sum_{k=0}^{|X^-|} \log f(x_k)\Big) \quad x_i \in X_n \cap X_o, x_k \in X^- \tag{2-2}$$

Here, $D_x = \min_k d(x_k, x), x_k \in X^-$. According to Equ. 2-2(Our loss function), we define C as a sample set that belongs to the same subspace as x_j and $\forall x \in C, x \in X^-$.

Let $g(x) = -\big(\log\big(1 - (x - \delta)\big) + \log(x)\big), 0 < \delta < 1$, because $g(1) < g(x)$, $\forall \delta \leq x < 1$. With Proposition 2, one outlier has little effect on another outlier. Thus, we can only care about one outlier and noise samples around that outlier. We use $\mathcal{L}_{f(x_j)}$

to represent the loss function, where x_j is the outlier:

$$\begin{aligned}
\mathcal{L}_{f(x_j)} &= -\big(\sum_{k}^{|C|} \log f(x_k) + \log(1 - f(x_j))\big) \\
&\geq -\big(\log(1 - f(x_j)) + \log f(x_{k*}) + \sum_{k}^{|C|\backslash\{x_{k*}\}} \log f^*(x_k)\big) \\
&\geq -\big(\log(1 - f^*(x_j)) + \log f^*(x_{k*}) + \sum_{k}^{|C|\backslash\{x_{k*}\}} \log f^*(x_k)\big) \\
&= \mathcal{L}_{f^*(x_j)},
\end{aligned} \tag{2-3}$$

where $f^*(x_k) = 1, f^*(x_j) = 1 - \min_k Md(x_j, x_k), x_k \in X^-, x_j \in X_o$. Therefore, if there exists $\exists x_j \in X_o$, s.t. $f(x_j) > MD$, it is theoretically possible to find $f^*(x_j)$ that minimizes the loss and thus complete the proof.

Theorem 1. Each predicted value of the outlier is higher than each predicted value of the inlier.

Proof. Due to the high density of inliers, $\forall x_i \in X_n$, when $\rho(x_i) \to +\infty$, $f(x_i) \to 0$. There must exist ρ_0, s.t. $\forall x_i \in X_n$, then, $f(x_i) < \tau$. According to **Lemma 1**, when $M < \frac{1}{2D}, \tau = 0.1$, we may have a boundary value λ, so that $\forall x_i \in X_n, x_j \in X_o, f(x_j) > \lambda > f(x_i)$, and we complete the proof.