**Problem Statement**

- Identify unknown, anomalous events that deviate from a training set of normal events.

- Examples: fault detection in mission-critical systems, quality control in manufacturing, medical diagnosis.

**Statistical framework:**
- Training sample \( X_T = \{X_1, \ldots, X_T\} \)
- Nominal density \( X \sim f_0 \)
- Test sample \( X \sim f \)
- Anomaly detection problem: Test \( H_0: f = f_0 \) versus \( H_1: \{1 - \delta\} f_0 + \delta f \)

**Previous approaches based on:**
- Density \([5]\): Mass, Fréchet
- Distance: ORCA \([6]\)
- Minimum volume sets: Hero \([1]\), K-LPE \([2]\)

**MV set approach**

- Fix false alarm rate \( \alpha \)
- Seek acceptance region \( A \) with minimum volume \( \text{MV} \) or equivalently minimum entropy \( \text{ME} \) that satisfies

\[
Pr(X \in A | f_0) \geq 1 - \alpha
\]

**MV anomaly detection:** If test event falls in MV set, classify as normal; otherwise as anomalous

- Uniformly most powerful test for \( f \) = uniform.

- **MV set estimation** involves approximation of high dimensional quantities \([3, 4]\); level sets of

\[
f_\theta(x) = \sum_{k} N_k(x - X_k)
\]

**GEM principle**

- **Geometric entropy minimization (GEM) principle** \([5]\) circumscribes MV set estimation.

- Asymptotically consistent in recovering the p-value \((1)\) of the test point.

- Let \( X_{T+1} \) denote one of the \( (\mathbb{T}+1) \) K point subsets of \( X_T \times X \).

- **K-LNN** acceptance region

\[
A = \{X_T | d_{kNN}(X_T, X_{T+1}) > \epsilon \}
\]

- Declare \( X \) to be an anomaly if \( X \notin A \). False alarm rate converges to \( \alpha \) \( \rightarrow \alpha \) (\( T \rightarrow 1 \) + 1) and runtime is \( O(\mathbb{T}_X)\log K \).

**RELATION TO GEM**

- \([N,M]\) partition of \( X_T \): \( card(X_T) = N \) and \( card(X_a) = M = T - N \)
- Let \( X_{T+1} \) denote one of the \( (\mathbb{T}+1) \) K point subsets of \( X_T \).
- Construct bipartite k-NN graph from \( X_{T+1} \) to \( X_T \).
- **BP-kNNG** acceptance region:

\[
A = \{X_T | d_{kNN}(X_T, X_{T+1}) > \epsilon \}
\]

- Declare \( A \) to be an anomaly if \( X \notin A \). By GEM principle false alarm rate converges to \( \alpha \) \( \rightarrow \alpha \) (\( \mathbb{T}/N + 1 \)).

- Equivalence We can equivalently determine with ranking. Order \( X_T \times X_T \) according to

\[
d_{kNN}(X_T, X_T) \leq \cdots \leq d_{kNN}(X_T, X_T) \leq d_{kNN}(X_T, X_T)
\]

- Set \( A = \{X_T | X_T \in X_{T+1}\} \).

- **Computational savings:**

  1. Partitioning approach breaks down combinatorial problem into ranking problem

  2. Sufficient to construct bipartite graph once on the entire set of nominal data and queries

- For \( T \) queries, runtime per test instance is \( O(T^2/T + 1) \). When \( T = O(\mathbb{T}) \), the runtime is \( O(\mathbb{T}) \), linear rather than quadratic.

**GEM**

- **True p-value**

\[
Pr(X \in A | f_0) = \int_{f(X) \geq f_\theta(x)} f_\theta(dx) \]

- **Empirical p-value**

\[
p_{\text{emp}}(X) = \frac{\sum_{i \in X_T} \text{card}(X_{Ti} \cup X_T)}{\text{card}(X_T)}
\]

- Asymptotic consistency: \( \mathbb{T} \rightarrow \infty \rightarrow N \rightarrow \infty \)

\[
E[\text{emp}(X) - p(X)] = O(\sqrt{N}/(K\mathbb{T})^{1/2}) + 1/4
\]

- **Optimal parameter choice** w.r.t MSE for \( \mathbb{T} \) fixed

\[
N = \frac{\mathbb{T} \times \text{emp}(X)}{M} = M - N = O(T)
\]

- **False alarm rate** as ---

**ROC Comparison**

**False alarm rate comparison**

**Runtime comparison**

**Conclusions**

- **BP-kNNG** is based on GEM principle \([1]\) for MV set-based anomaly detection.

- **BP-kNNG** inherits theoretical optimality properties of GEM, including asymptotic consistency, unlike L10-kNNG.

- Bipartite construction reduces combinatorial problem to ranking problem.

- Consequence: Runtime of **BP-kNNG** is significantly better in comparison to **K-LPE**, **L10-kNNG** and **K-LPE**.

- Comparison to state of the art anomaly detection methods.

- **Detect anomalies** at desired false alarm rates.

**References**


