FITNESS (Fine Tune on New and Similar Samples)

Anomaly detection on data streams with drifts and outliers

Abishek Sankararaman, Balakrishnan (Murali) Narayanaswamy, Vikramank Singh, Zhao Song

Amazon Web Services, AI Labs,
Santa Clara CA

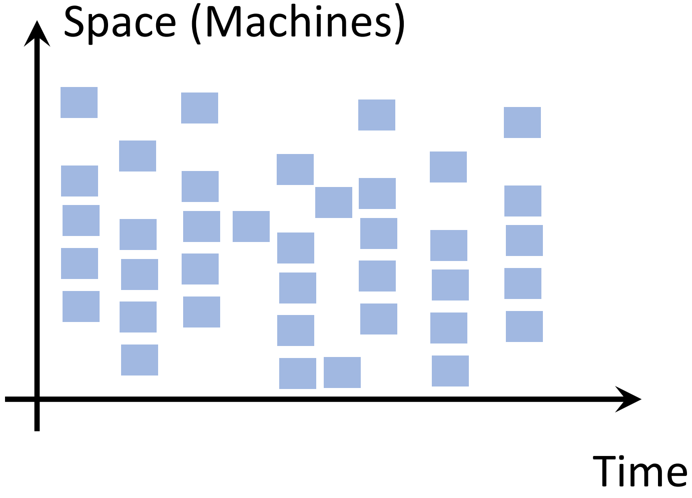
ICML 2022

Online Anomaly Detection

Technological developments has led to a surge in real-time data

IoT, Sensors, Machine health, Cloud Computing

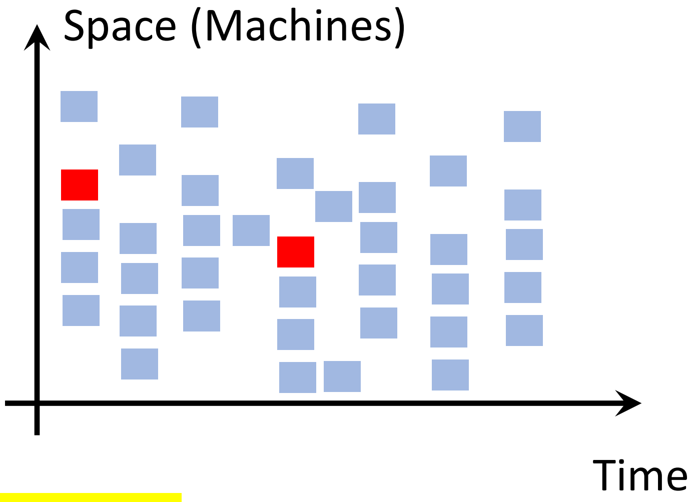
Complex data that humans alone cannot understand, manage, monitor and fix!



Online Anomaly Detection

Technological developments has led to a surge in real-time data

IoT, Sensors, Machine health, Cloud Computing



Complex data that humans alone cannot understand, manage, monitor and fix!

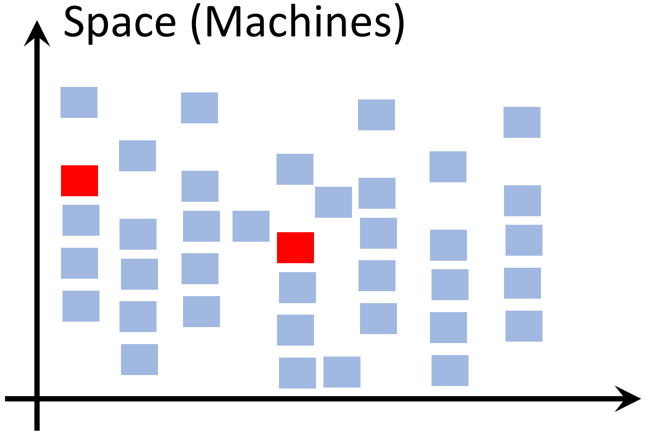
Anomaly detection : Important sub-routine for many monitoring and control applications

Performance monitoring, Security monitoring, Capacity provisioning

Online Anomaly Detection

Technological developments has led to a surge in real-time data

IoT, Sensors, Machine health, Cloud Computing



Complex data that humans alone cannot understand, manage, monitor and fix!

Anomaly detection : Important sub-routine for many monitoring and control applications

Performance monitoring, Security monitoring, Capacity provisioning

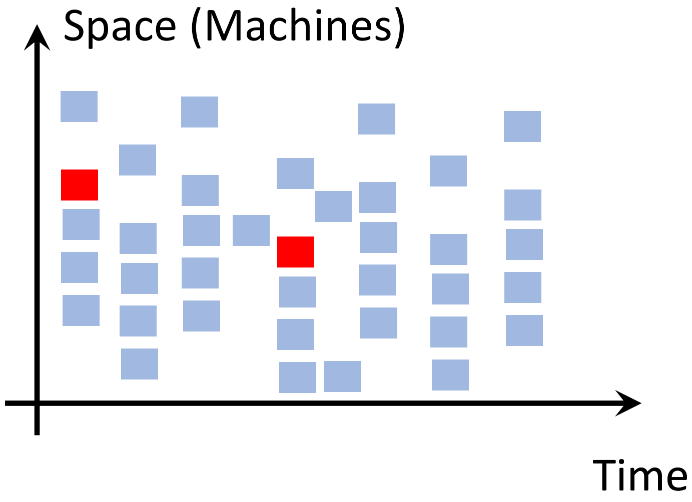
Application Challenges:

- Real-time data generation and decision

Online Anomaly Detection

Technological developments has led to a surge in real-time data

IoT, Sensors, Machine health, Cloud Computing



Complex data that humans alone cannot understand, manage, monitor and fix!

Anomaly detection : Important sub-routine for many monitoring and control applications

Performance monitoring, Security monitoring, Capacity provisioning

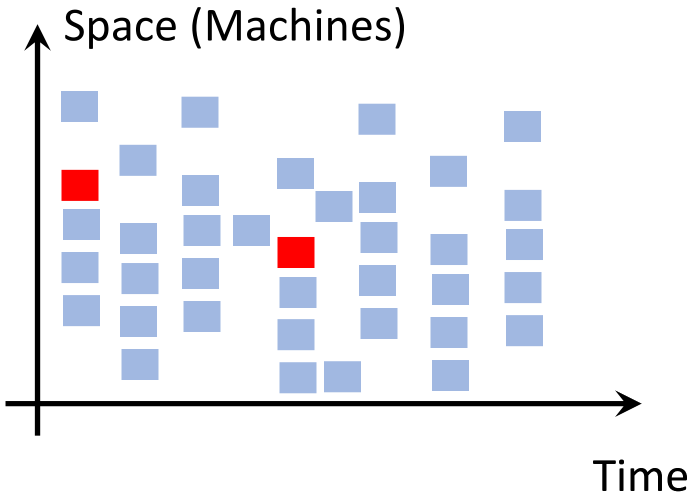
Application Challenges:

- Real-time data generation and decision
- Unknown data distribution

Online Anomaly Detection

Technological developments has led to a surge in real-time data

IoT, Sensors, Machine health, Cloud Computing



Complex data that humans alone cannot understand, manage, monitor and fix!

Anomaly detection : Important sub-routine for many monitoring and control applications

Performance monitoring, Security monitoring, Capacity provisioning

<u>Application Challenges</u>:

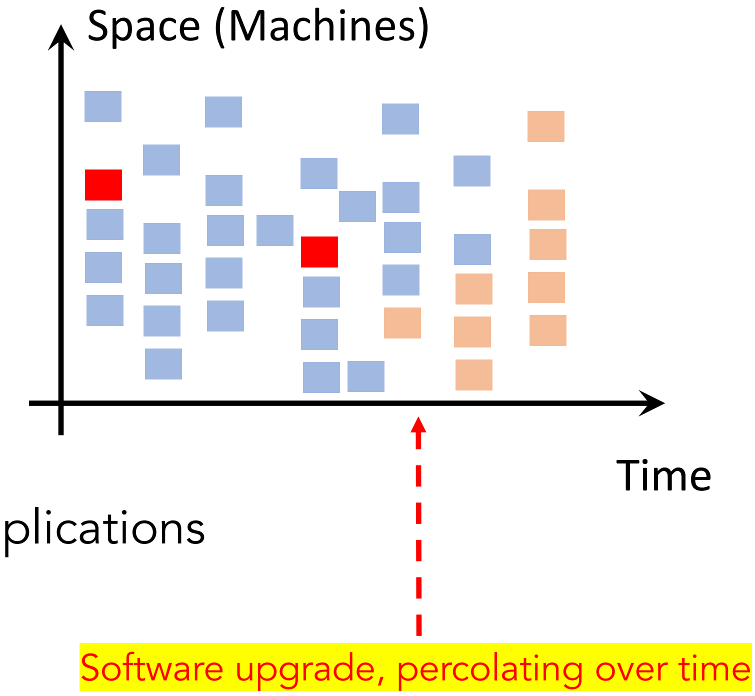
- Real-time data generation and decision
- Unknown data distribution
- Distributions can and will change with time

Online Anomaly Detection

Technological developments has led to a surge in real-time data

IoT, Sensors, Machine health, Cloud Computing

Complex data that humans alone cannot understand, manage, monitor and fix!



Anomaly detection : Important sub-routine for many monitoring and control applications

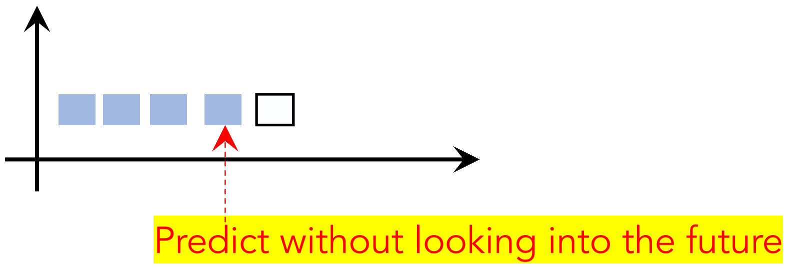
Performance monitoring, Security monitoring, Capacity provisioning

Application Challenges:

- Real-time data generation and decision
- Unknown data distribution
- Distributions can and will change with time

Desiderata for anomaly detection

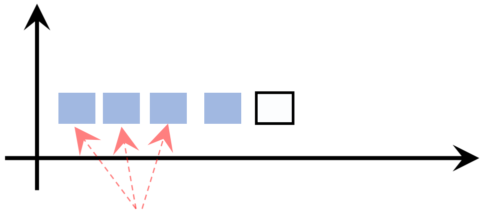
1. Streaming



2. Limited supervision
3. Adaptive to distribution shifts – gradual and sudden
4. Robust to a few outliers/adversarialy corrupted points
5. Competitive with offline methods if the data-stream is “nice and stationary”

Desiderata for anomaly detection

1. Streaming
2. Limited supervision
3. Adaptive to distribution shifts – gradual and sudden
4. Robust to a few outliers/adversarialy corrupted points
5. Competitive with offline methods if the data-stream is “nice and stationary”

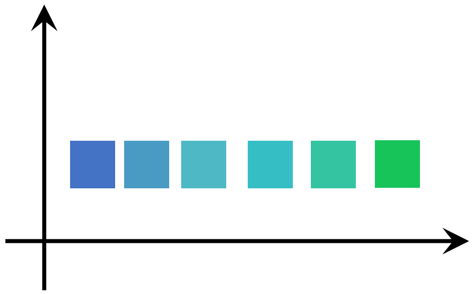


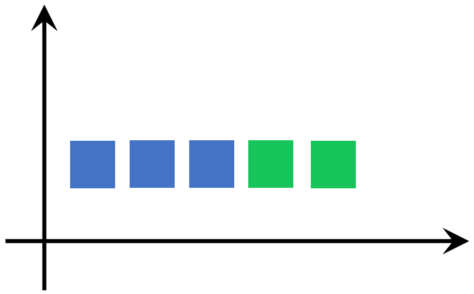
Limited feedback on past predictions

Desiderata for anomaly detection

1. Streaming
2. Limited supervision
3. Adaptive to distribution shifts – gradual and sudden
4. Robust to a few outliers/adversarialy corrupted points
5. Competitive with offline methods if the data-stream is "nice and stationary"

Gradual and sudden shifts





Desiderata for anomaly detection

1. Streaming

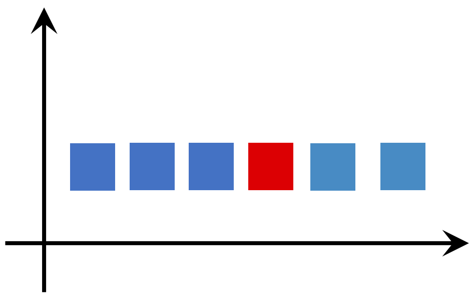
2. Limited supervision

3. Adaptive to distribution shifts – gradual and sudden

4. Robust to a few outliers/adversarialy corrupted points

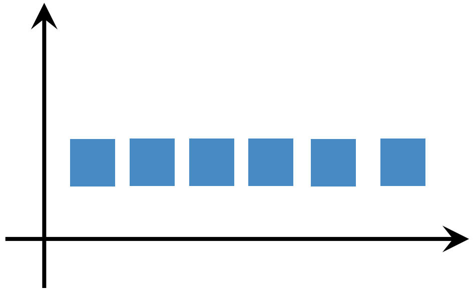
5. Competitive with offline methods if the data-stream is "nice and stationary"

Distinguish corruptions from "new normal".



Desiderata for anomaly detection

1. Streaming
2. Limited supervision
3. Adaptive to distribution shifts – gradual and sudden
4. Robust to a few outliers/adversarialy corrupted points
5. Competitive with offline methods if the data-stream is "nice and stationary"



Desiderata for anomaly detection

1. Streaming

2. Limited supervision

3. Adaptive to distribution shifts – gradual and sudden

4. Robust to a few outliers/adversarialy corrupted points

5. Competitive with offline methods if the data-stream is “nice and stationary”

- All existing methods achieve some, but not all these desiderata

Main Contributions

1. Statistical formulation of desiderata
    - Identify problem complexity parameters
    - Lower bounds

2. Prove the desiderata is a non-trivial benchmark
    - Not achieved by obvious algorithms
        - Fixed window sliding
        - Ignoring learning from samples predicted to be an anomaly

3. In the case when the data stream is Gaussian distributed, we propose FITNESS : GAUSSIAN that provably achieves the desiderata

4. For the general case, we propose FITNESS : GENERAL,
    - AD Model-agnostic
    - Flexible : takes a batch AD model and converts it to an online version that satisfying desiderata.

Main Contributions

1. Statistical formulation of desiderata
    - Identify problem complexity parameters
    - Lower bounds

2. Prove the desiderata is a non-trivial benchmark
    - Not achieved by obvious algorithms
        - Fixed window sliding
        - Ignoring learning from samples predicted to be an anomaly

3. In the case when the data stream is Gaussian distributed, we propose FITNESS : GAUSSIAN that provably achieves the desiderata

4. For the general case, we propose FITNESS : GENERAL,
    - AD Model-agnostic
    - Flexible : takes a batch AD model and converts it to an online version that satisfying desiderata.

Main Contributions

1. Statistical formulation of desiderata
    - Identify problem complexity parameters
    - Lower bounds

2. Prove the desiderata is a non-trivial benchmark
    - Not achieved by obvious algorithms
        - Fixed window sliding
        - Ignoring learning from samples predicted to be an anomaly

3. In the case when the data stream is Gaussian distributed, we propose FITNESS : GAUSSIAN that provably achieves the desiderata

4. For the general case, we propose FITNESS : GENERAL,
    - AD Model-agnostic
    - Flexible : takes a batch AD model and converts it to an online version that satisfying desiderata.

Main Contributions

1. Statistical formulation of desiderata
   - Identify problem complexity parameters
   - Lower bounds

2. Prove the desiderata is a non-trivial benchmark
   - Not achieved by obvious algorithms
     - Fixed window sliding
     - Ignoring learning from samples predicted to be an anomaly

3. In the case when the data stream is Gaussian distributed, we propose FITNESS : GAUSSIAN that provably achieves the desiderata

4. For the general case, we propose FITNESS : GENERAL,
   - AD Model-agnostic
   - Flexible : takes a batch AD model and converts it to an online version that satisfying desiderata.

Related Work

Unsupervised Online Anomaly Detection : several algorithms have been proposed over the years

MEMSTREAM [Bhatia et al., '21] , SketchDetect [Huang et al., '15] : Discards samples that appear anomalous at the moment of arrival

KitSune [Mirsky et al., '18], xSTREAM [Manzoor et al., 2018], StreamIF [Ding et al., '13] : Fixed sliding window methods

DiLOF [Na et al., '18], RSHash [Sathe et al., '16], RCF [Guha et al,. '16], IF [Liu et al., '08], EIF [Harari et al., '19] : Offline methods

Continual Learning : Adaptivity demonstrated to drifts only in one-dimensional setting

[Lu et al., '18], [Gupta et al., '13], [Bifet et.al., '07], [Bifet et.al., '09]

Online supervised learning : These methodologies do not apply to unsupervised streams

[Chu et al., '04], [Defazio et al., '14], DYNASAGA [Daneshmand et al., '16] ,DriftSurf [Tahmasbi et al., '21]

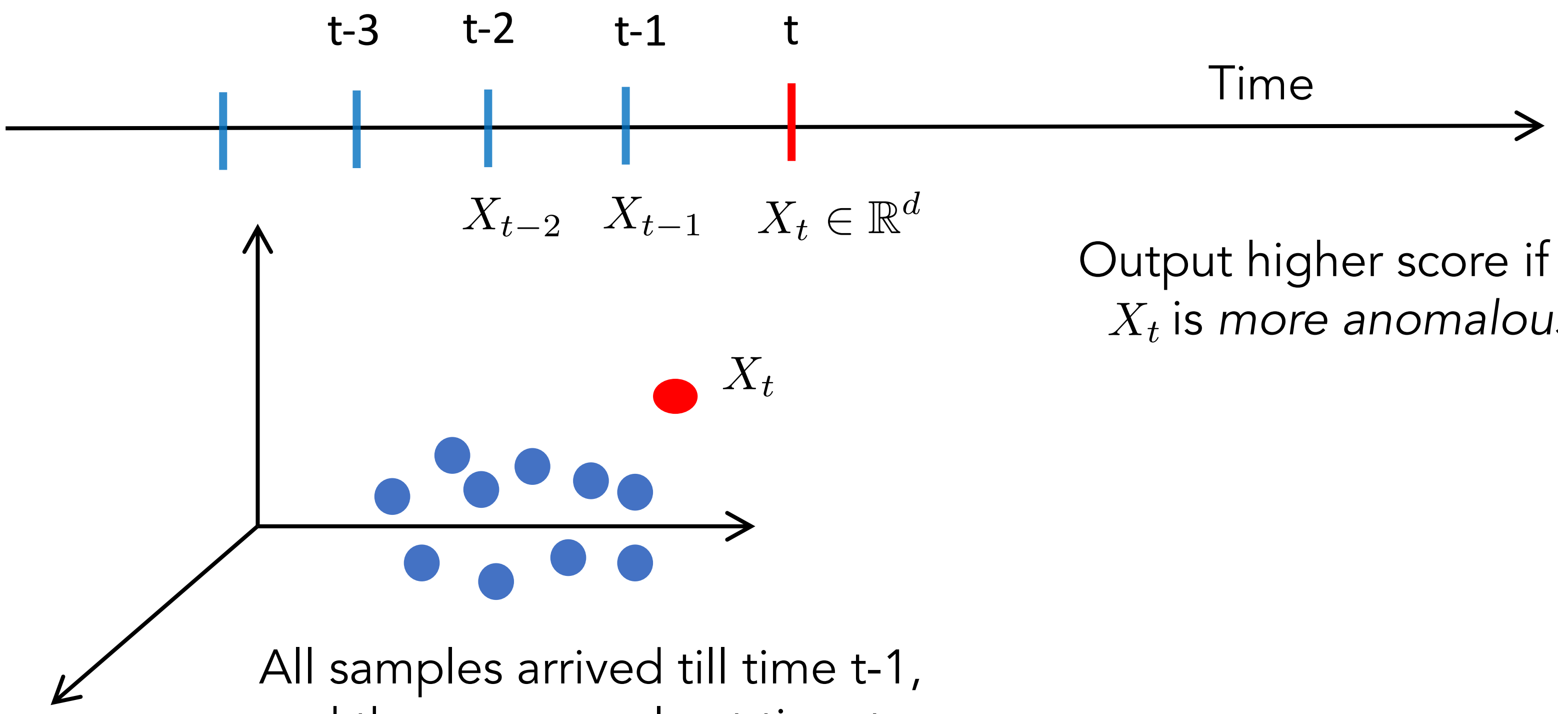
Robust Learning : Only works in the offline case

[Diakonikolas et al., '17],[Diakonikolas et al., '18], [Cheng et al., '19],[Cheng et al., '20]

Statistical Problem Formulation

Problem Statement

At each time t = 1,2,…, given a vector $X_t \in \mathbb{R}^d$ as input, output an *anomaly score* $S_t \in \mathbb{R}$



Output higher score if the input sample X_t is *more anomalous*

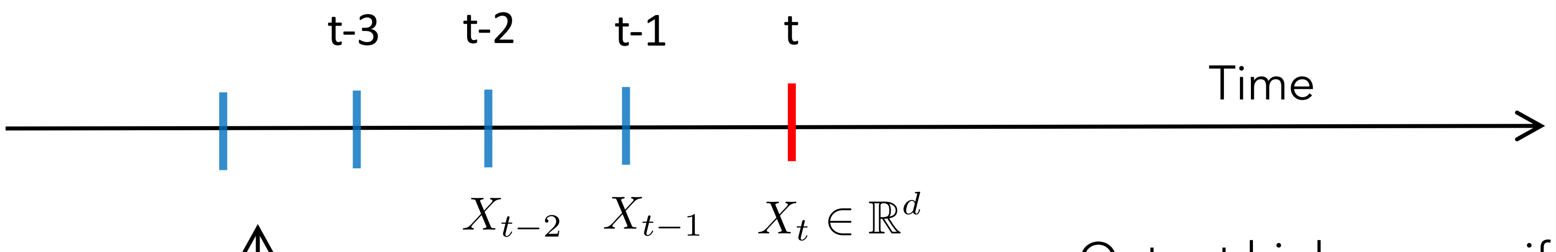
All samples arrived till time t-1, and the new sample at time t

Statistical Problem Formulation

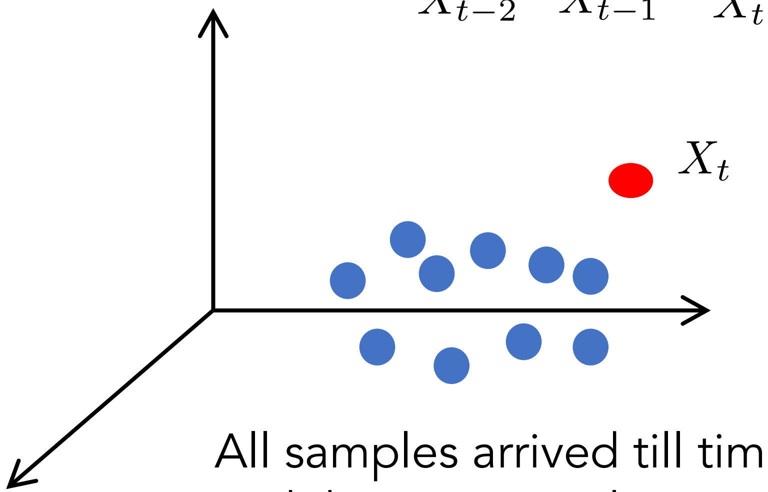
Problem Statement

At each time t = 1,2,…, ~~given a vector $X_t \in \mathbb{R}^d$~~ as input, output an *anomaly score* $S_t \in \mathbb{R}$

$X_t \sim \mathcal{D}_t$ is sampled independently from an *unknown* distribution $\mathcal{D}_t \in \mathcal{F}$ from a *known* set $\mathcal{F}$



Output higher score if the input sample X_t is *more anomalous*



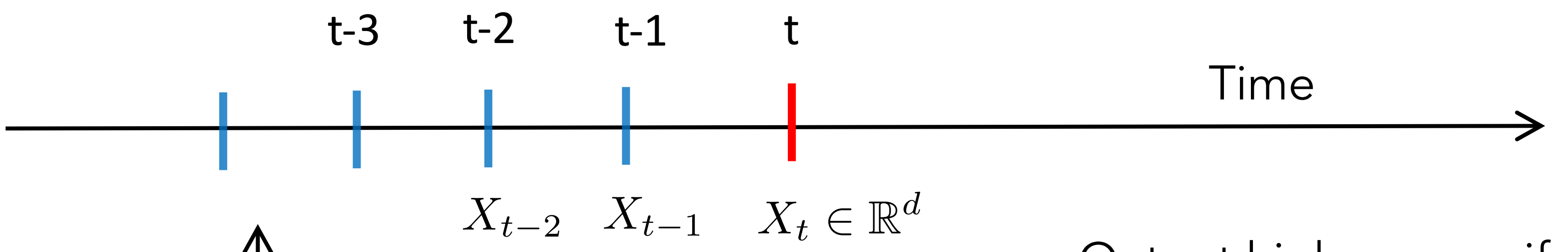
All samples arrived till time t-1, and the new sample at time t

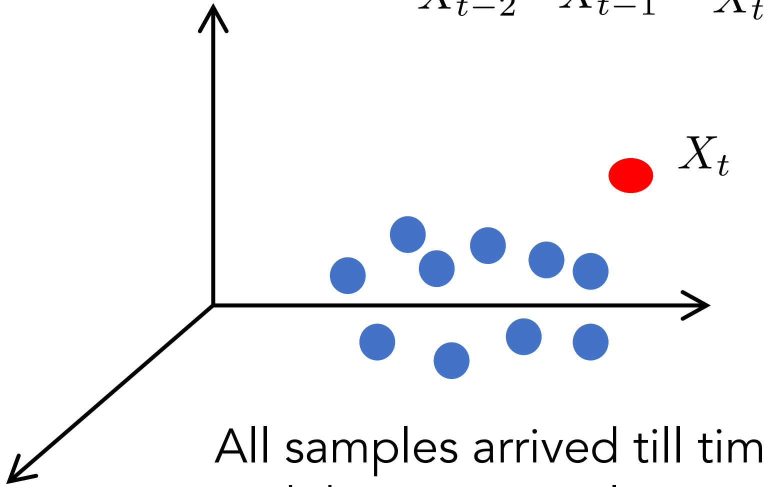
Statistical Problem Formulation

Problem Statement

At each time t = 1,2,…, ~~given a vector $X_t \in \mathbb{R}^d$~~ as input, output an *anomaly score* $S_t \in \mathbb{R}$

$X_t \sim \mathcal{D}_t$ is sampled independently from an *unknown* distribution $\mathcal{D}_t \in \mathcal{F}$ from a *known* set $\mathcal{F}$





All samples arrived till time t-1, and the new sample at time t

Output higher score if the input sample X_t is *more anomalous*

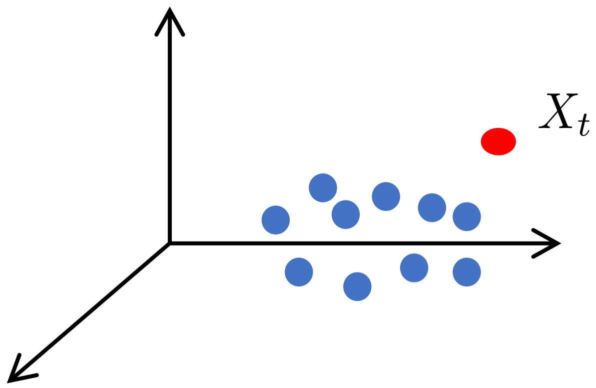
Anomaly Score ------> $S_t := g(\theta_t, X_t)$

Model Class (Autoencoder, Isolation Forest)

Parameters (Network Weights, Tree Splits)

Statistical Problem Formulation : Notations

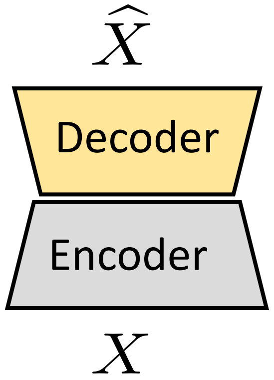
Notation	**Meaning**
Θ	The set of all possible parameters
$g(\cdot,\cdot) : \Theta \times \mathbb{R}^d \to \mathbb{R}$	A family of anomaly scoring functions
$g(\theta, X)$	Anomaly score given by model θ on input X
$\mathcal{F}$	Family of probability distributions from which the each data point X is sampled from



Statistical Problem Formulation : Notations

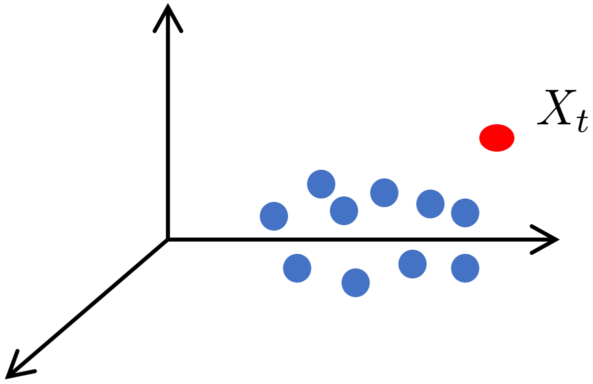
Notation	Meaning
Θ	The set of all possible parameters
$g(\cdot,\cdot) : \Theta \times \mathbb{R}^d \to \mathbb{R}$	A family of anomaly scoring functions
$g(\theta, X)$	Anomaly score given by model θ on input X
$\mathcal{F}$	Family of probability distributions from which the each data point X is sampled from

Example – Autoencoder as an anomaly scoring function



Θ The set of all possible weights of a *fixed* architecture

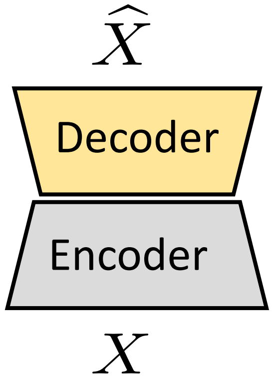
$g(\theta, X) := \|X - \widehat{X}\|$ Reconstruction error is the anomaly score



Statistical Problem Formulation : Notations

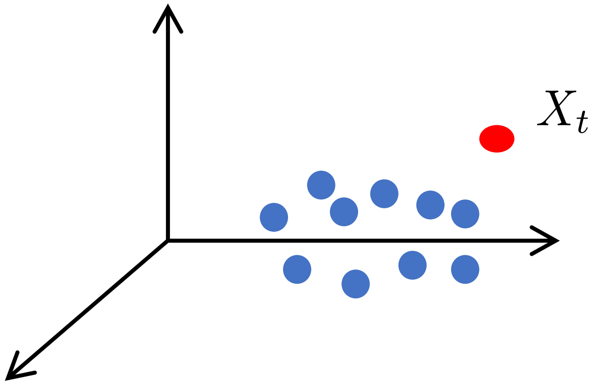
Notation	**Meaning**
Θ	The set of all possible parameters
$g(\cdot,\cdot) : \Theta \times \mathbb{R}^d \to \mathbb{R}$	A family of anomaly scoring functions
$g(\theta, X)$	Anomaly score given by model θ on input X
$\mathcal{F}$	Family of probability distributions from which the each data point X is sampled from

Example – Autoencoder as an anomaly scoring function



Θ The set of all possible weights of a *fixed* architecture

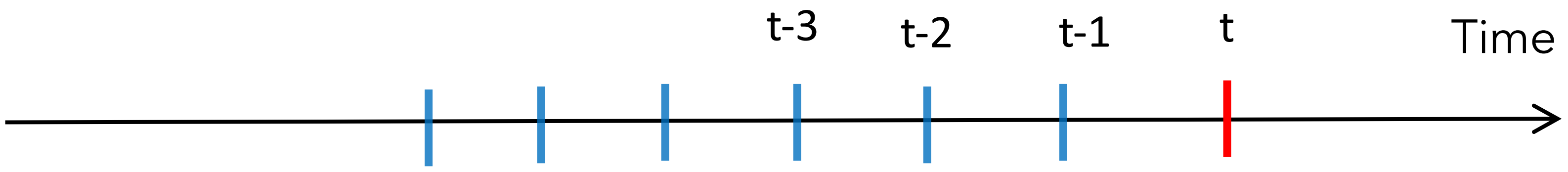
$g(\theta, X) := \|X - \widehat{X}\|$ Reconstruction error is the anomaly score



Goal : Given g(. , .) and $\mathcal{F}$, how to choose the parameter θ at each time, in an online fashion

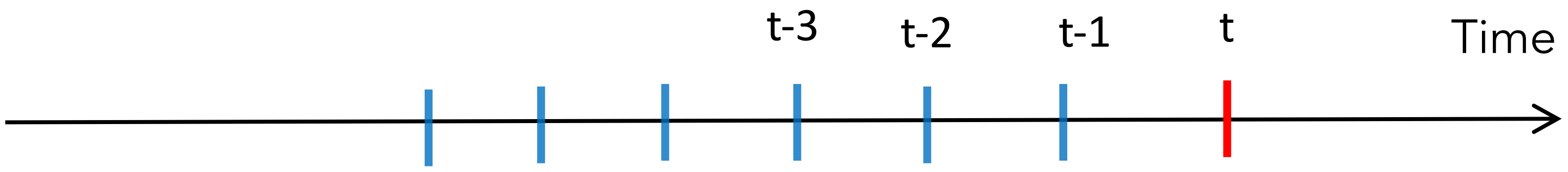
Statistical Problem Formulation : Setup

Sequential Interaction with an adversary



Statistical Problem Formulation : Setup

Sequential Interaction with an adversary

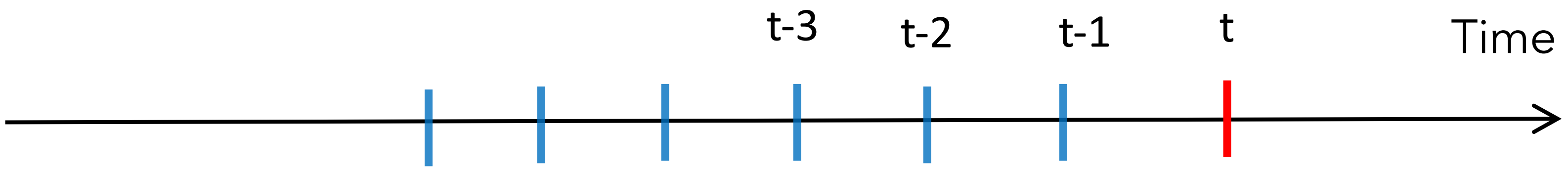


At each time instant t,

1. The adversary picks a distribution $\mathcal{D}_t \in \mathcal{F}$

Statistical Problem Formulation : Setup

Sequential Interaction with an adversary

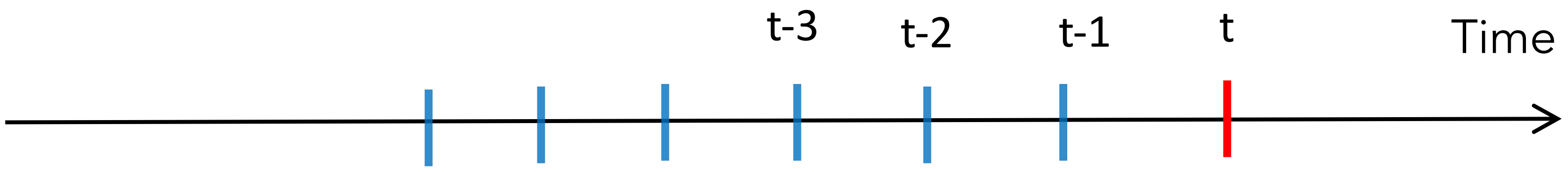


At each time instant t,

1. The adversary picks a distribution $\mathcal{D}_t \in \mathcal{F}$
2. The adversary then samples $\widetilde{X}_t \sim \mathcal{D}_t$ independently at random

Statistical Problem Formulation : Setup

Sequential Interaction with an adversary

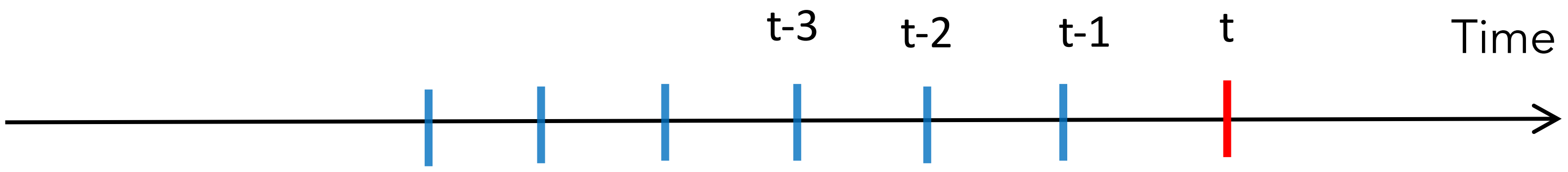


At each time instant t,

1. The adversary picks a distribution $\mathcal{D}_t \in \mathcal{F}$
2. The adversary then samples $\widetilde{X}_t \sim \mathcal{D}_t$ independently at random
3. The adversary chooses an arbitrary corruption c_t

Statistical Problem Formulation : Setup

Sequential Interaction with an adversary

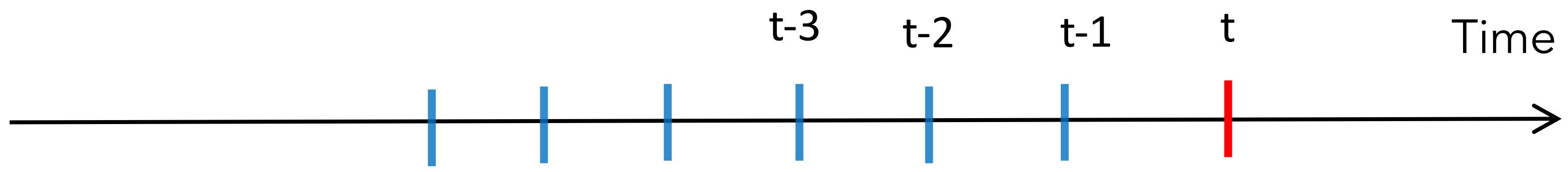


At each time instant t,

1. The adversary picks a distribution $\mathcal{D}_t \in \mathcal{F}$
2. The adversary then samples $\widetilde{X}_t \sim \mathcal{D}_t$ independently at random
3. The adversary chooses an arbitrary corruption c_t
4. Adversary reveals $X_t := \widetilde{X}_t + c_t$ to the AD Algorithm

Statistical Problem Formulation : Setup

Sequential Interaction with an adversary



At each time instant t,

1. The adversary picks a distribution $\mathcal{D}_t \in \mathcal{F}$
2. The adversary then samples $\widetilde{X}_t \sim \mathcal{D}_t$ independently at random
3. The adversary chooses an arbitrary corruption c_t
4. Adversary reveals $X_t := \widetilde{X}_t + c_t$ to the AD Algorithm
5. Subsequently, the AD algorithm depending on $(X_s)_{s \leq t}$ all inputs thus far,
   a) Picks an action $\theta_t \in \Theta$
   b) Outputs anomaly score $g(\theta_t, X_t)$

Statistical Problem Formulation - Objective

Anomaly Scores to be low for non-anomalous points and high for anomalous points

Statistical Problem Formulation - Objective

Anomaly Scores to be low for non-anomalous points and high for anomalous points

Which inputs are not anomalous ?

If no adversarial corruption, i.e., $c_t = 0$ => Sample is benign and not an anomaly

Statistical Problem Formulation - Objective

Anomaly Scores to be low for non-anomalous points and high for anomalous points

Which inputs are not anomalous ?

If no adversarial corruption, i.e., $c_t = 0$ => Sample is benign and not an anomaly

Desired Output on non-anomalous points ?

Score point X_t with parameters θ_t "close" to some $\arg\min_{\theta \in \Theta} \mathbb{E}_{X \sim \mathcal{D}_t}[g(\theta, X)]$

Statistical Problem Formulation - Objective

Anomaly Scores to be low for non-anomalous points and high for anomalous points

Which inputs are not anomalous ?

If no adversarial corruption, i.e., $c_t = 0$ => Sample is benign and not an anomaly

Desired Output on non-anomalous points ?

Score point X_t with parameters θ_t "close" to some $\arg\min_{\theta \in \Theta} \mathbb{E}_{X \sim \mathcal{D}_t}[g(\theta, X)]$

Performance Measure

Closeness measured by $\mathcal{L}(\cdot, \cdot) : \Theta \times \Theta \to \mathbb{R}_+$, a loss function.

$\mathcal{L}(\theta_1, \theta_2)$ measures difference between functions $g(\theta_1, \cdot)$ and $g(\theta_2, \cdot)$

Statistical Problem Formulation - Objective

Anomaly Scores to be low for non-anomalous points and high for anomalous points

Which inputs are not anomalous ?

If no adversarial corruption, i.e., $c_t = 0$ => Sample is benign and not an anomaly

Desired Output on non-anomalous points ?

Score point X_t with parameters θ_t "close" to some $\arg\min_{\theta \in \Theta} \mathbb{E}_{X \sim \mathcal{D}_t}[g(\theta, X)]$

Performance Measure

Closeness measured by $\mathcal{L}(\cdot, \cdot) : \Theta \times \Theta \to \mathbb{R}_+$, a loss function.

$\mathcal{L}(\theta_1, \theta_2)$ measures difference between functions $g(\theta_1, \cdot)$ and $g(\theta_2, \cdot)$

Example: For the autoencoder model, the L2 norm between the weights $||\theta_1 - \theta_2||$ is a "good measure" of AD performance deviation between models θ_1 and θ_2

[Kim et. al. '20] prove this measure to be valid for any Lipschitz model g(. , .).

Statistical Problem Formulation - Regret

Define the *instantaneous regret* of the AD algorithm at time t, denoted by r_t as

$$r_t := \inf\{\mathcal{L}(\theta_t, \theta^*), \theta^* \in \arg\min_{\theta \in \Theta} \mathbb{E}_{X_t \sim \mathcal{D}_t}[g(\theta, X_t)]\}$$

How far is the model used at time t from the optimal possible model

Statistical Problem Formulation - Regret

Define the *instantaneous regret* of the AD algorithm at time t, denoted by r_t as

$$r_t := \inf\{\mathcal{L}(\theta_t, \theta^*), \theta^* \in \arg\min_{\theta \in \Theta} \mathbb{E}_{X_t \sim \mathcal{D}_t}[g(\theta, X_t)]\}$$

How far is the model used at time t from the optimal possible model

Regret

$$R_T := \sum_{t=1}^{T} \mathbf{1}(c_t = 0) r_t$$

Total cumulative regret on non-anomalous points

Statistical Problem Formulation - Regret

Define the *instantaneous regret* of the AD algorithm at time t, denoted by r_t as

$$r_t := \inf\{\mathcal{L}(\theta_t, \theta^*), \theta^* \in \arg\min_{\theta \in \Theta} \mathbb{E}_{X_t \sim \mathcal{D}_t}[g(\theta, X_t)]\}$$

How far is the model used at time t from the optimal possible model

Regret

$$R_T := \sum_{t=1}^{T} \mathbf{1}(c_t = 0) r_t$$

Total cumulative regret on non-anomalous points

Central Design Question

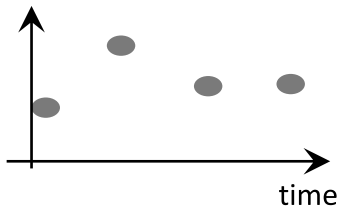
Can an algorithm be designed such that regret is small, whenever the adversary is constrained to place only a "small number" of anomalies and "small amount" of distribution drift ?

Statistical Problem Formulation - Desiderata

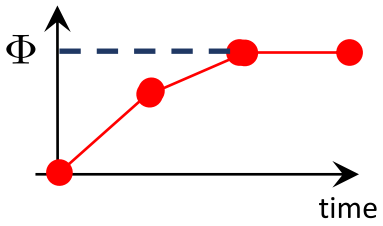
Measuring Drift

$$\Phi := \sum_{t=2}^{T} \text{Total-Variation}(\mathcal{D}_{t-1}, \mathcal{D}_t)$$

Schematic of distribution



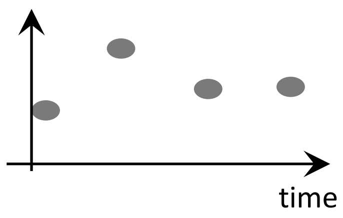
Sum of cumulative distribution differences

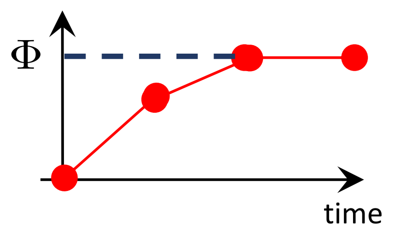


Statistical Problem Formulation - Desiderata

Measuring Drift

$$\Phi := \sum_{t=2}^{T} \text{Total-Variation}(\mathcal{D}_{t-1}, \mathcal{D}_t)$$



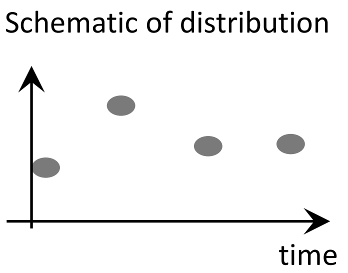


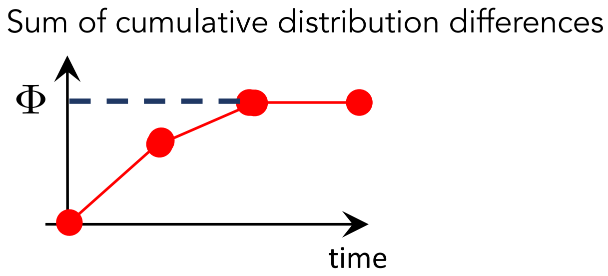
Number of Anomalies

$$\Upsilon := \sum_{t=1}^{T} \mathbf{1}(c_t \neq 0)$$

Statistical Problem Formulation - Desiderata

Measuring Drift $$\Phi := \sum_{t=2}^{T} \text{Total-Variation}(\mathcal{D}_{t-1}, \mathcal{D}_t)$$





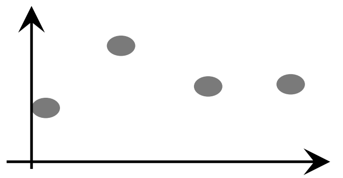
Number of Anomalies $$\Upsilon := \sum_{t=1}^{T} \mathbf{1}(c_t \neq 0)$$

The algorithm does not know these parameters.

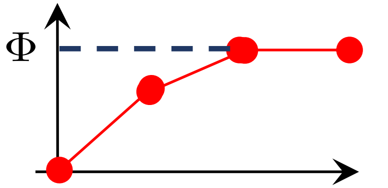
Statistical Problem Formulation - Desiderata

Measuring Drift $$\Phi := \sum_{t=2}^{T} \text{Total-Variation}(\mathcal{D}_{t-1}, \mathcal{D}_t)$$

Schematic of distribution



Sum of cumulative distribution differences



Number of Anomalies $$\Upsilon := \sum_{t=1}^{T} \mathbf{1}(c_t \neq 0)$$

The algorithm does not know these parameters.

An online algorithm $\mathcal{A}$ is said to be <u>adaptive and robust</u>, if for every $\rho < 1$ and $c < 1$,

there exists $\beta < 1$ such that $$\limsup_{T \to \infty} \sup_{\substack{\boldsymbol{\mathcal{D}} \in \mathcal{F} \text{ s.t.} \\ \Phi \leq T^\rho, \\ \Upsilon \leq T^c}} \frac{\mathbb{E}[R_T]}{T^\beta} \leq 0.$$

An algorithm has small regret, whenever the "complexity" of the problem is small.

Why is this a good benchmark ?

Regret cannot be sublinear in T, if number of corruptions is linear in T, even if there is no distribution shift

Proposition 4.1 : There is an universal constant c > 0, such that if all samples in the data stream are i.i.d., from a Gaussian distribution of unit variance and unknown mean, then $\inf_{\mathcal{A}} R_T \geq c \frac{\Upsilon}{T}(T - \Upsilon)$

Why is this a good benchmark ?

Regret cannot be sublinear in T, if number of corruptions is linear in T, even if there is no distribution shift

Proposition 4.1 : There is an universal constant c > 0, such that if all samples in the data stream are i.i.d., from a Gaussian distribution of unit variance and unknown mean, then $\inf_{\mathcal{A}} R_T \geq c\frac{\Upsilon}{T}(T-\Upsilon)$

Regret cannot be sublinear in T, if total distribution shift is linear in T, even if there are no anomalies

Proposition 4.2 : There exists a finite family of distributions $\mathcal{F}$ such that every data stream $\mathcal{D}$ from this family satisfies $\Phi(\mathcal{D}) \leq \zeta$ and incurs regret $\inf_{\mathcal{A}} \sup_{\mathcal{D}} \mathbb{E}[R_T] \geq \frac{1}{24}T^{2/3}\zeta^{1/3}$

A Simple Instantiation – Estimating the Mean

Simpler Task :

Given an unknown stream of vectors $\mu_1, \mu_2, \cdots$, let $\widetilde{X}_t \sim \mathcal{N}(\mu_t, I)$ independently, and $X_t = \widetilde{X}_t + c_t$

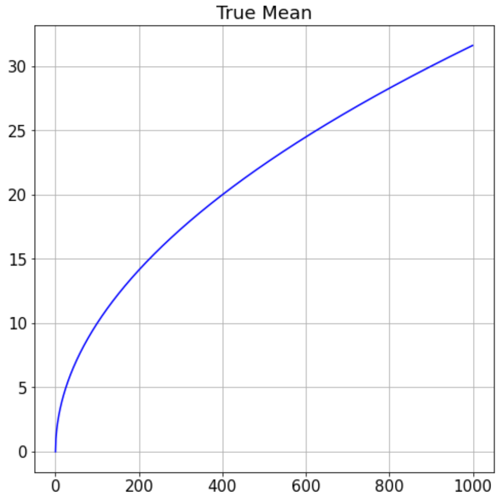
Estimate the mean $\widehat{\mu}_t$ from samples.

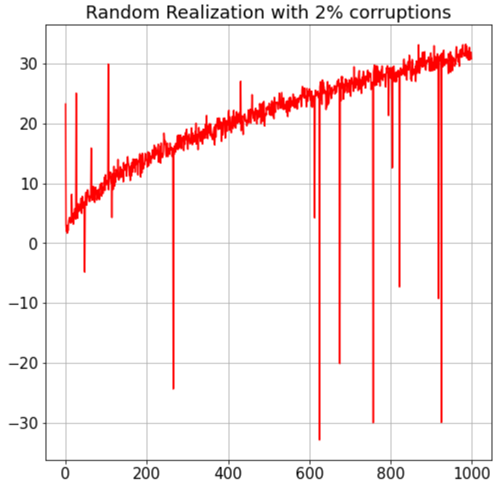
At each time t,

Input - $X_t := Z_t + \mu_t + c_t$

Output - $\widehat{\mu}_t$ an estimate of μ_t

Goal : Minimize regret $R_T := \sum_{t=1}^{T} \mathbf{1}(c_t = 0) \|\mu_t - \widehat{\mu}_t\|$

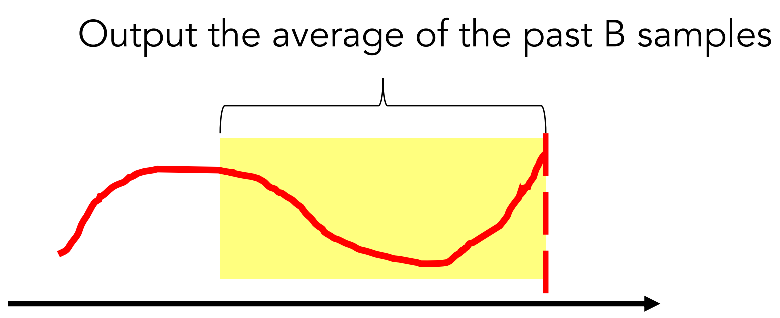


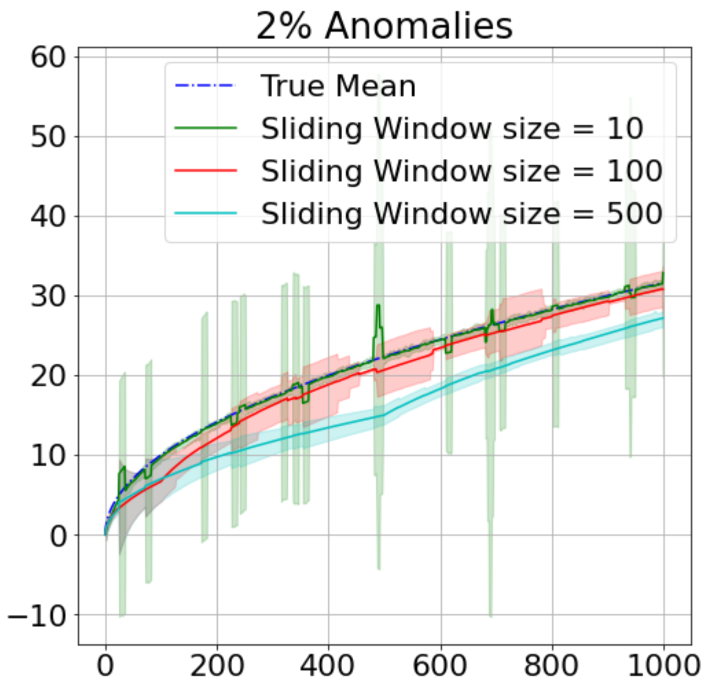


We show this to be an instantiation of our general model

Sliding Window Methods fail

$$\widehat{\mu}_t = \begin{cases} \frac{1}{B}\sum_{s=0}^{B-1} X_{t-s} & \text{if } \left\| \frac{1}{B}\sum_{s=0}^{B-1} X_{t-s} - X_t \right\| \leq \lambda, \\ X_t & \text{otherwise} \end{cases}$$





Main Dilemma

Need many past samples for the average to concentrate around the mean

Samples too far in the past may not be reflective of the current distribution

Naïve Dynamic Windows Fail

Idea – Only average those points that are not declared an anomaly at the time of arrival

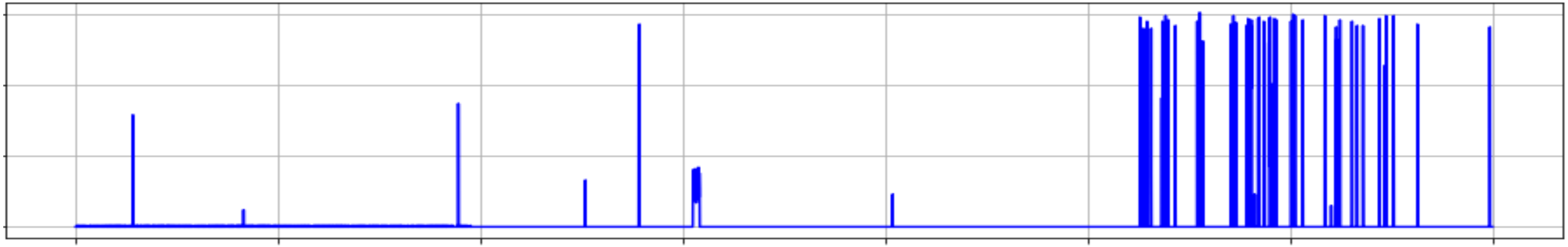
This is a popularly used paradigm in many published algorithms

Naïve Dynamic Windows Fail

Idea – Only average those points that are not declared an anomaly at the time of arrival

This is a popularly used paradigm in many published algorithms

This will clearly fail to adapt to the “new normal” in the example below.

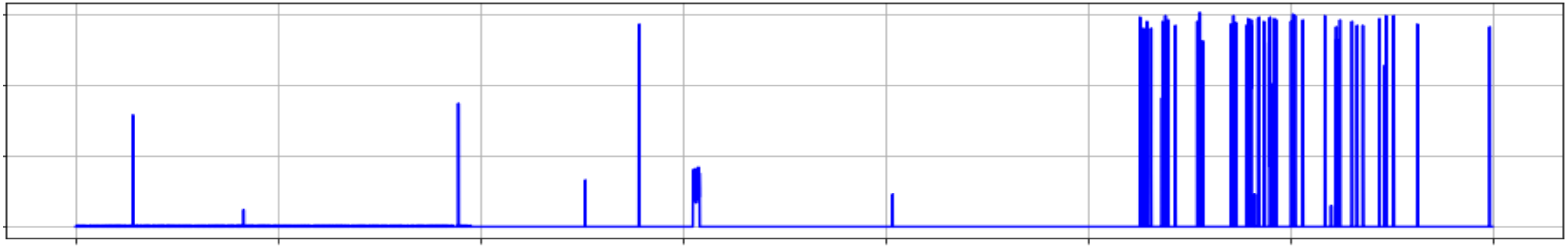


Naïve Dynamic Windows Fail

Idea – Only average those points that are not declared an anomaly at the time of arrival

This is a popularly used paradigm in many published algorithms

This will clearly fail to adapt to the "new normal" in the example below.



Thus, it is important to not discard a sample even if it looks anomalous.

FITNESS Achieves the Desiderata

Our Proposal – Estimate the mean from the largest set of recent samples that are relevant.

FITNESS Achieves the Desiderata

Our Proposal – Estimate the mean from the largest set of recent samples that are relevant.

Algorithm 1: `FITNESS :GAUSSIAN`

Input: $\sigma \geq 0$, Slack parameter $\delta \in (0, 1)$, Time horizon T, C as given in Definition 11

1 **for** *each time* $t \geq 1$ **do**

2   Receive Input $X_t \in \mathbb{R}^d$

3     $j \leftarrow 1$

4   **while** $\left\| \frac{1}{j} \sum_{s=0}^{j-1} X_{t-s} - X_t \right\| \leq C_1 \left(1 + \frac{2}{\sqrt{j}}\right) \sqrt{d\sigma} \log\left(\frac{T^2}{\delta}\right)$ **do**

5         $j \leftarrow j + 1$

6   **return** $\widehat{\mu}_t := \frac{1}{j} \sum_{s=0}^{j-1} X_{t-s}$

Key trick is to introduce j on the RHS of the condition in line 4.

As more samples from the past are averaged, we want the concentration to be higher.

FITNESS Achieves the Desiderata

$J^*(t)$ is the first time instant while scanning backwards from t, when μ_t significantly deviates from the average of the means in the time-window [J*(t), t].

Definition 17. *For every* $t \in \{1, 2, \cdots, T\}$ *that is non-anomalous (i.e.,* $c_t = 0$*), define* $J^*(t)$ *as*

$$J^*(t) := \inf \left\{ j \in \{1, 2, \cdots, t\}, s.t. \left\| \mu_t - \frac{1}{j} \sum_{s=0}^{j-1} (\mu_{t-s} + c_{t-s}) \right\| > C \sqrt{\frac{d\sigma}{j}} \log\left(\frac{T^2}{\delta}\right) \right\},$$

where inf of an empty set is defined as $J^*(t) := t + 1$.

Theorem 18. *If Algorithm 1 is run with slack parameter* $\delta \in (0, 1)$*, then with probability at-least* $1 - \delta$*, the following regret bound holds*

$$R_T \leq \sum_{t=1}^{T} 2C \sqrt{\frac{d\sigma}{J^*(t) - 1}} \log\left(\frac{T^2}{\delta}\right).$$

This result implies that the FITNESS is both adaptive and robust.

Shortcomings and Future Work

1. Computational Complexity is not added as a desiderata

   - FITNESS takes O(t) time per sample. Ideally need O(1) computation time per sample

Shortcomings and Future Work

1. Computational Complexity is not added as a desiderata

    - FITNESS takes O(t) time per sample. Ideally need O(1) computation time per sample

2. We only have provable robustness and adaptivity in the Gaussian case

    Practical Anomaly Detection are typically in heavy-tailed and time-series settings

Thank You

More details in the paper

FITNESS (<u>Fi</u>ne <u>T</u>une on <u>Ne</u>w and <u>S</u>imilar <u>S</u>amples) to detect anomalies in streams with drifts and outliers