Anomaly Detection Via Asymmetric Risk Minimization

Aryeh Kontorovich, Danny Hendler and Eitan Menahem, Ben-Gurion University

SIMBAD 2011

Philosophy of Anomaly Detection

- How will you know that you're seeing a Alien?

Cost-Sensitive One-Class Anomaly Detection

Problem at hand:

1. We want to detect anomalies
2. During learning we see only one type of examples (positive)
3. Inherent asymmetry between classification errors
    1. False alarms are usually far less disastrous than missed anomalies
    2. we pay a fixed cost for each false alarm, but once we miss an anomaly, the “game” is over, and we pay a one-time cost *C*

mistaken call to fired dept. vs. warehouse burning down

Problem Definition

How to define this problem formally?

- this is arguably the hardest stage!

- Unlike in PAC, it's not clear what a "good" or "bad" classifier is…
  - what's to prevent the trivial learner (which label everything positive)?
- What does "probability of mistake" mean?

  There is no distribution over the negative examples!

Common Modeling Assumption: Euclidean Space

- Pros
  - Existence of inner product
  - Flexible kernels for incorporating prior knowledge
  - Efficient algorithms (SVM)
  - Good generalization bounds (margins)

- Cons
  - Euclidean structure is a strong assumption
  - Many natural settings non-Euclidean
  - Choice of kernel: artisan and partisan

What About Metric Space?

- Advantage: often much more natural
  - strings
  - images
  - audio
  - web-pages

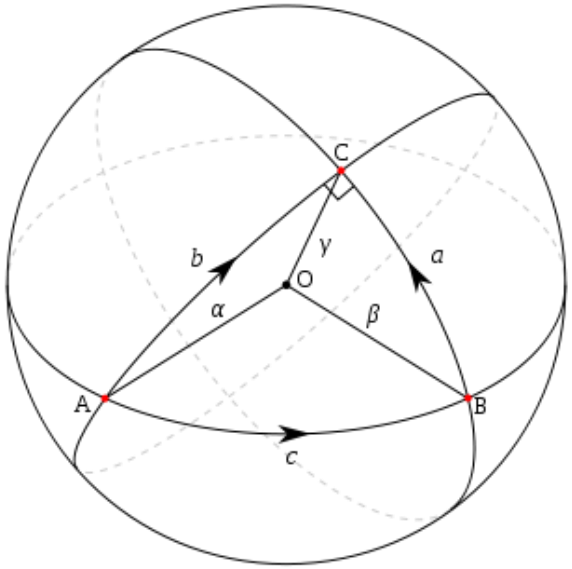
- Problem: no vector representation
  - No notion of dot-product
  - What to do?
    - Invent kernel?.. but... many natural metrics aren't Euclidean!
    - Use some NN heuristic?..
    - NN classifier has ∞ VC-dim
    - So what NN does guarantee?

Section 2

BACKGROUND

Metric Space

- (X,d) is a Metric Space if
  - X = set of points
  - d = distance function $d: x \times x \rightarrow \mathbb{R}_+$
    - Nonnegative $d(x, x') = 0 \Leftrightarrow x = x'$
    - Symmetric: $d(x, x') = d(x', x)$
    - triangle inequality: $d(x, x') \leq d(x, z) + d(z, x')$



- inner product ⇒ norm $\|x\| = \sqrt{\langle x, x \rangle}$
- norm ⇒ metric $d(x, y) = \|x - y\|$
- NOT ⇐

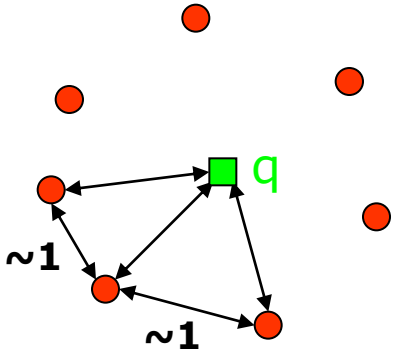
Binary Classification for Metric Data

- A powerful framework for this problem was introduced by von Luxburg & Bousquet [vLB,'04]
  - The natural hypotheses (classifiers) to consider are maximally smooth Lipschitz functions
  - Given the classifier h, the problem of evaluating h for new points in X reduces to the problem of finding a Lipschitz function consistent with h
    - Lipschitz extension problem, a classic problem in Analysis
  - The 1-NN is a special case of the Lipschitz classifier.
  - For example
    - $f(x) = min_i[y_i + 2d(x, x_i)/d(S^+, S^-)]$ over all (x_i, y_i) in S
    - Function evaluation reduces to exact Nearest Neighbor Search (NNS), assuming zero training error)
      - Strong theoretical motivation for the NNS classification heuristic

[vLB'04] : Ulrike von Luxburg, Olivier Bousquet: Distance-Based Classification with Lipschitz Functions. Journal of Machine Learning Research (JMLR) 5:669-695 (2004)

Computational Efficiency

- Efficient construction and evaluation of the classifier h on X
  - In arbitrary metric space, exact NNS requires $\Theta(n)$ time
  - Can we do better?
  - Gottlieb et al. [GKK'10] show that the answer is YES

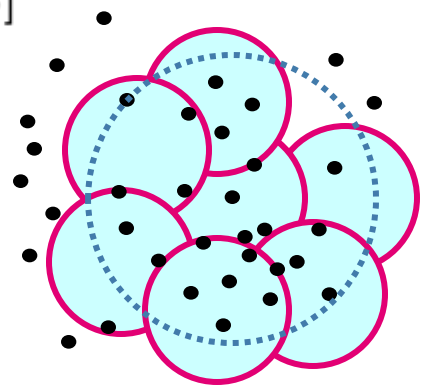


[GKK'10] : Lee-Ad Gottlieb, Leonid Kontorovich, Robert Krauthgamer: Efficient Classification for Metric Data. COLT 2010:433-440

Doubling Dimension

- Definition: Ball $B(x, r)$ = all points within distance r from x.

- The doubling constant $\lambda > 0$ (of a metric M) is the minimum value such that every ball can be covered by λ balls of half the radius
  - First used by [Ass-83], algorithmically by [Cla'97].
  - The doubling dimension is $ddim(M) = log_2\lambda(M)$ [GKL'03]
  - A metric is doubling if its doubling dimension is constant
  - Euclidean: $ddim(R^n) = O(n)$
- Cole & Gottlieb [CG'10]: (1+ε)-approximate nearest neighbor search
  - $\lambda^{O(1)} \log n + \lambda^{O(-\log\varepsilon)}$ time



Here $\lambda \geq 7$.

[Ass'83]: P. Assouad. Plongements lipschitziens dans Rn. Bull. Soc. Math. France, 111(4):429–448, 1983.
[Cla'97]: Kenneth L. Clarkson: Nearest Neighbor Queries in Metric Spaces. STOC 1997: 609-617
[CG'10]: Richard Cole, Lee-Ad : Searching dynamic point sets in spaces with bounded doubling dimension. STOC 2006: 574-583

Generalization Bound in Metric Space

- [BST99]:
  - For any f that classifies a sample of size n correctly, we have with probability at least $1-\delta$

$$P\{(x,y):\ sgn(f(x)) \neq\ y\} \leq \frac{2}{n}\left(d\ ln\left(\frac{34en}{d}\right)log_2(578n) + \ln(4/\delta)\right)$$

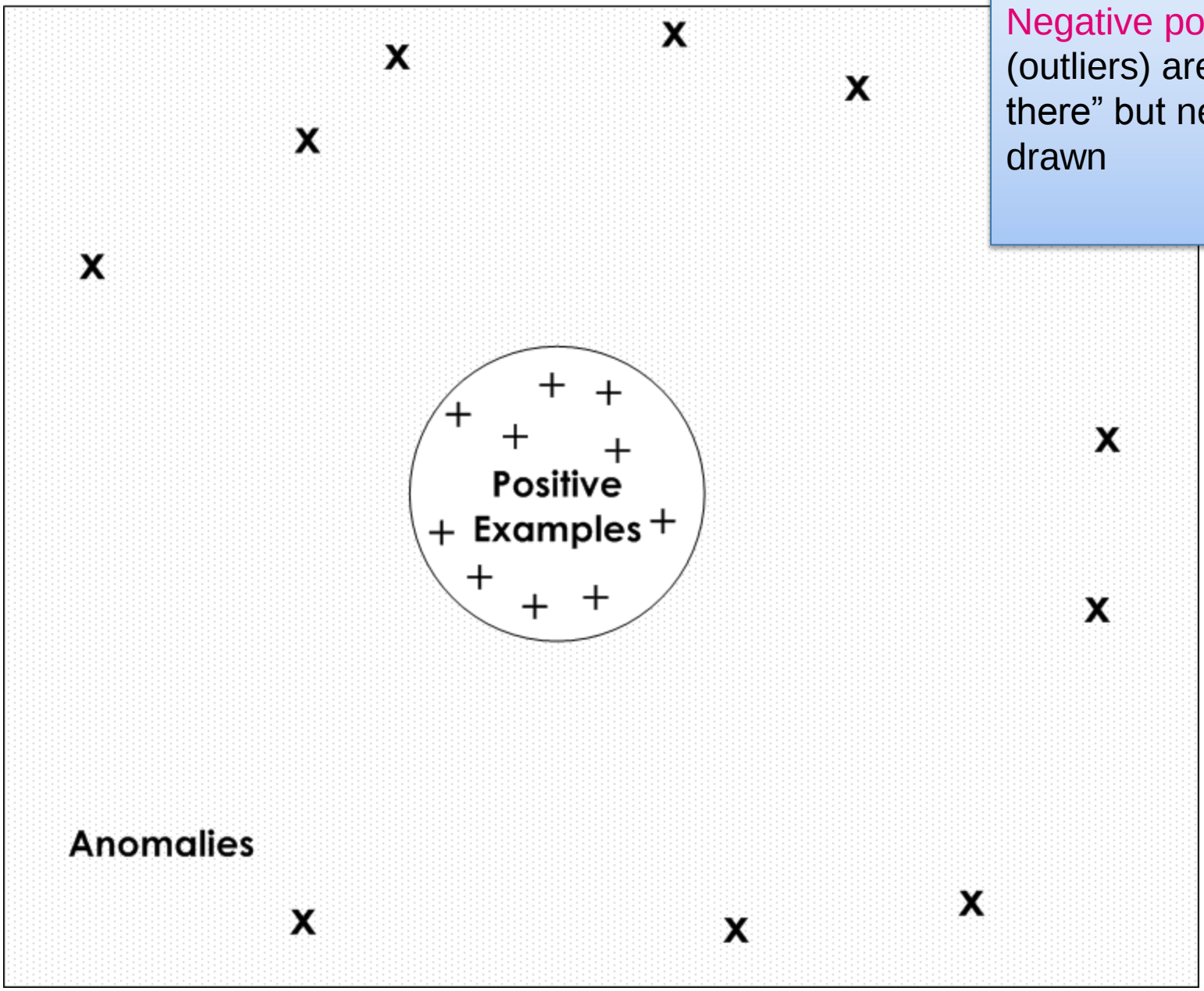
$$d \leq \lceil 8Ldiam(X)\rceil^{log\lambda+1}$$

[BST99] : Peter Bartlett and John Shawe-Taylor. Generalization performance of support vector machines and other pattern classifiers. In Advances in kernel methods: support vector learning, pages 43–54, Cambridge,MA, USA, 1999.MIT Press.

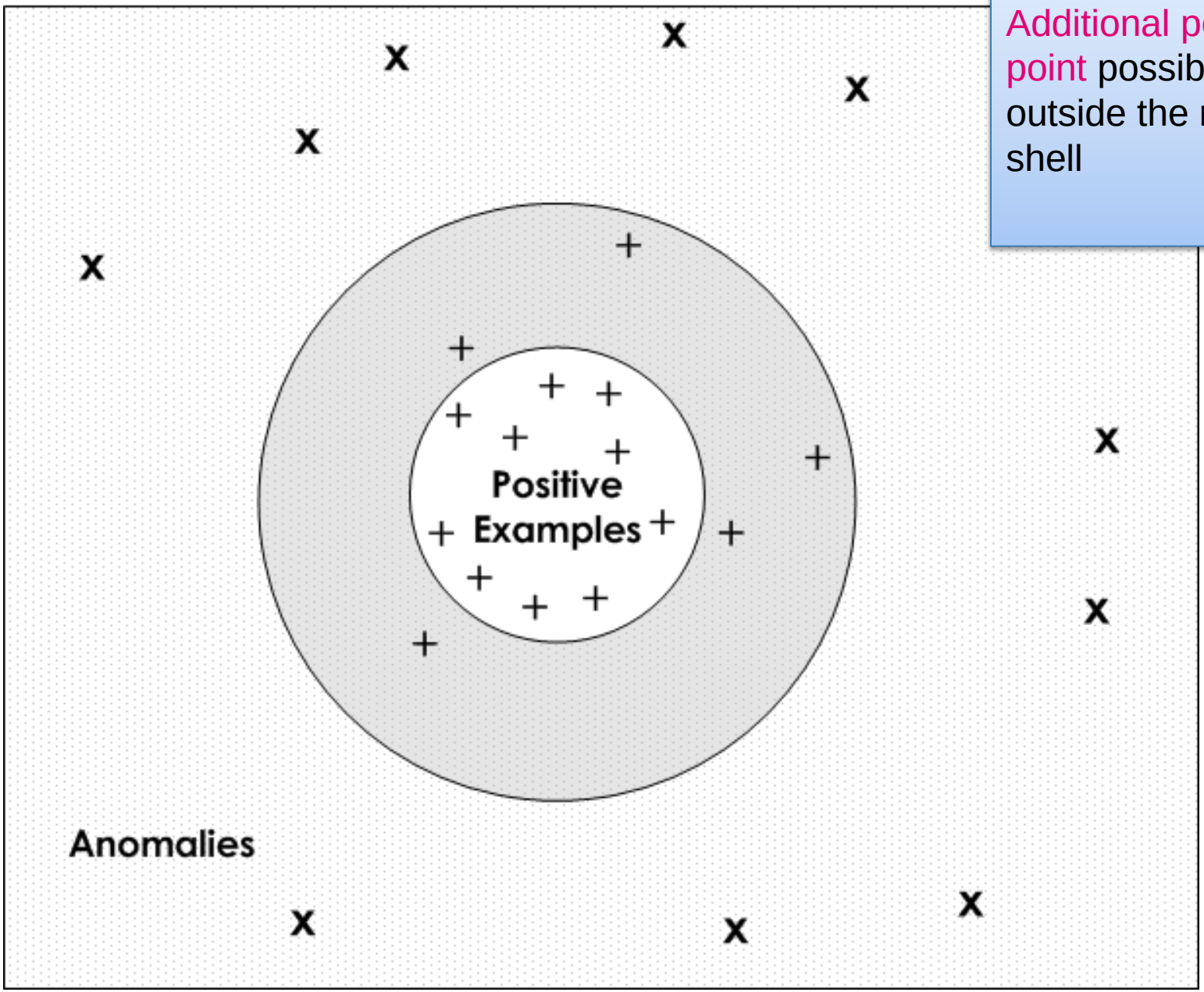
Section 3

MODEL ASSUMPTIONS

Positive points are drawn iid from an unknown distribution. They are contained in a "metric shell"

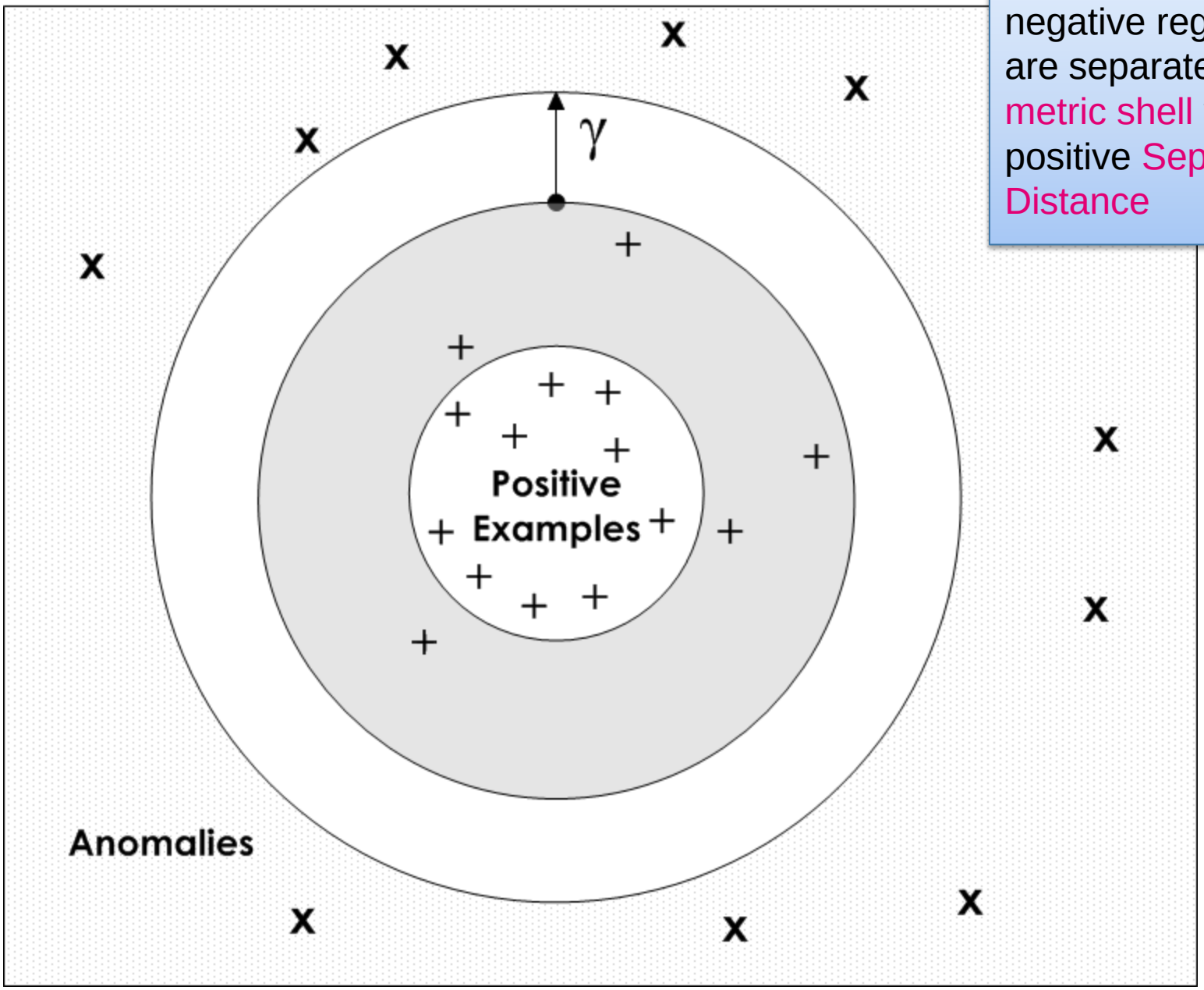


Negative points (outliers) are “out there” but never drawn



Additional positive point possibly exist outside the metric shell

The positive and negative regions are separated by a metric shell with positive Separation Distance



Section 4

ANOMALY DETECTION VIA ASYMMETRIC RISK MINIMIZATION

Various Models of Uncertainty in γ

- γ is known
- We have a prior on γ
- Yet a weaker assumption on γ

Instead of Generalization Error - Asymmetric Risk

- The Risk has two components:
  - False alarm - a false report of anomaly is made by the detector
  - Missed anomaly - the detector fails to detect a real anomaly in the data.

- Risk is their weighted sum!
  - False-Alarm + Missed Anomalies* C

- The notion of Separation Distance

  Separation distance, i.e.: $d(X_+, X_-) \equiv \inf_{x \in X_+, y \in X_-} d(x, y) > \gamma$ for some separation distance $\gamma > 0$

  - a natural analogue of the Euclidean Margin

1st Case: A Known Separation Distance (γ)

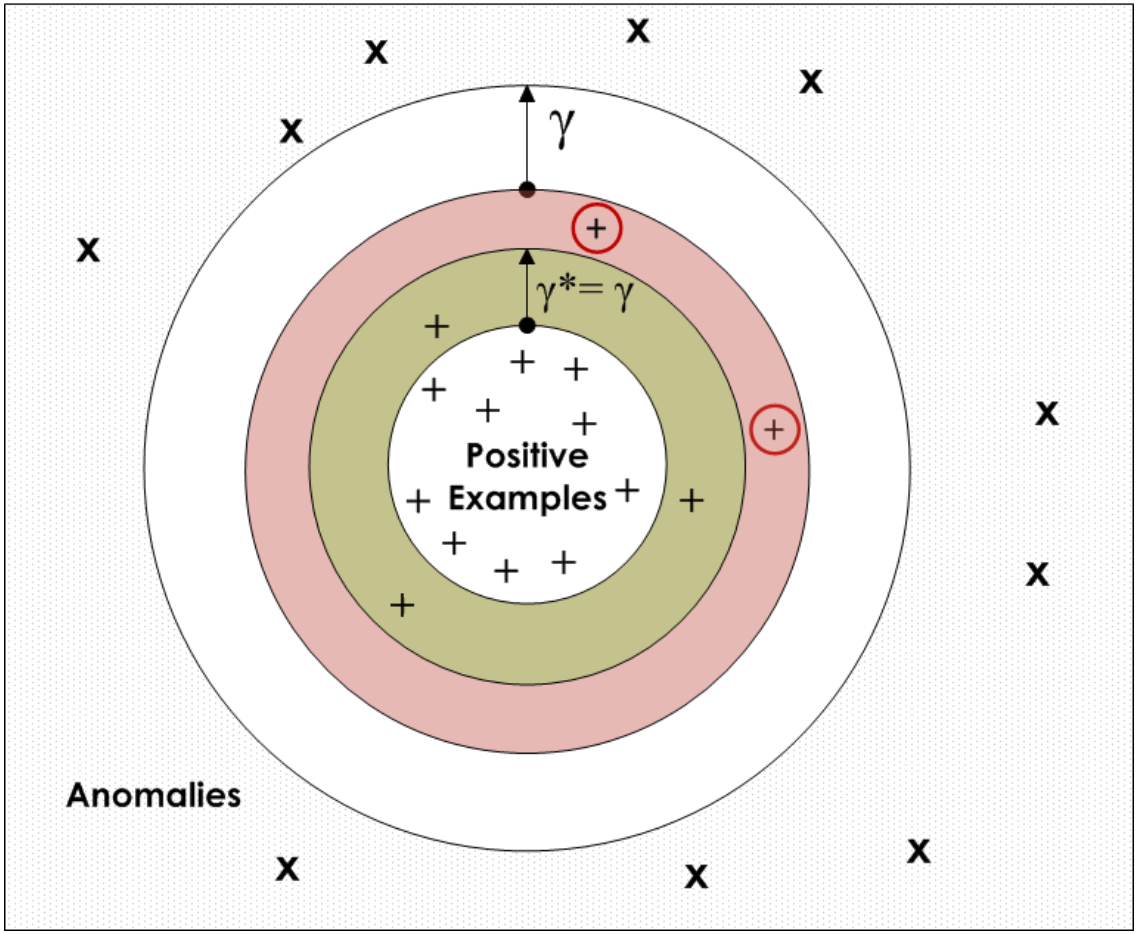
- Given a training set , $S = \{X_1, \dots, X_n\}$ drawn from χ_+ iid under the distribution P, define the proximity classifier $f_{n,\gamma}$ as :

$$f_{n,\gamma}(x) = \begin{cases} Normal, & d(x,S) \leq \gamma \\ Anomaly, & else \end{cases}$$

- In this model there are no missed anomalies

1st Case: A Known Separation Distance (γ)

- Assume that the separation distance γ is known.



Bound the False Alarm Rate for Known γ

- The false alarm rate is:

$$FA(f) = \int_{X_+} \mathbf{1}_{\{f(x)<0\}} dP(x)$$

- With probability at least $1 - \delta$, this classifiers achieves a false alarm rate that satisfies :

$$FA(f_{n,\gamma}) \leq \frac{2(D log_2(34en/D) log_2(578n) + \log_2(4/\delta))}{n}$$

where

$$D = \left\lceil \frac{8\Delta}{\gamma} \right\rceil^{ddim(X)+1}$$

[GKK] : Lee-Ad Gottlieb, Leonid Kontorovich, Robert Krauthgamer: Efficient Classification for Metric Data. COLT 2010:433-440

Bound the Risk as Function of γ

- Assuming a large n

$$Risk(\gamma) = E[FA(f_{n,\gamma})] \leq A_{n,\gamma} + B_n$$

where

$$A_{n,\gamma} = \frac{2(D_\gamma log_2(34en/D_\gamma) log_2(578n) + \log_2(4))}{n}$$

and

$$B_n = \frac{2}{nln2}$$

(*) when n is large enough, $A_{n,\gamma} < 1$

2nd Case: We Have a Prior on γ

- Although there is uncertainty regarding the separation distance γ, we might be able to model it via some distribution $G(\cdot)$ on $(0, \infty)$,
  - assumed as a prior
- Quantify the induced risk:

$$Risk(\gamma_0) = \int_{\gamma_0}^{\infty} E[FA(f_{n,\gamma})]dG(\gamma) + C\int_{0}^{\gamma_0} dG(\gamma)$$

where $G(\cdot)$ is the prior on γ

This reflects our modeling assumption that we pay a unit cost for each FA, and a large "catastrophic" cost C for any number of missed anomalies

Choosing the Optimal Separation Distance

- We define the Risk as follows:

$$Risk(\gamma_0) = \int_{\gamma_0}^{\infty} \mathrm{E}\left[FA\left(f_{n,\gamma}\right)\right] dG(\gamma) + C \int_0^{\gamma_0} dG(\gamma) \leq$$

$$\leq \int_{\gamma_0}^{\infty} (A_{n,\gamma} + B_n) dG(\gamma) + C \int_0^{\gamma_0} dG(\gamma)$$

$$=: R_n(\gamma_0)$$

- The classification rule:
  - Compute the minimizer γ^* of $R_n(\cdot)$ and use the classifier f_{n,γ^*}

- Notice that $A_{n,\gamma}$ grows inversely with γ (proportional to $1/\gamma^{ddim(X)+1}$), so that γ^* would not be arbitrary small.
- Also $R_n(\gamma_0) \to 0$ as $n \to \infty$ for any fixed γ_0

3rd Case: No Explicit Prior on γ

- We can make the weak assumption:
  - We define the maximal distance from any point in S to its nearest neighbor (isolation distance), in any discrete metric space (S, d), as follows:

$$\rho = sup_{x \in S} d(x, S \backslash \{x\})$$

assuming $\rho < \gamma$

  - We can estimate ρ empirically, as a proxy of γ

$$\hat{\rho}_n = \max_{i \in [n]} \min_{i \neq j} d(X_i, X_j)$$

Note that $\hat{\rho}_n \leq \rho$ and that $\hat{\rho}_n \rightarrow \rho$ almost surely

Estimating the False-Alarm Component

ε-Net and Unseen Mass:

- The sample S is called ε-net if every point in x has an epsilon neighbor in S.
- For $x \in S$ we define ε–ball :

$$B_\varepsilon(x) = \{y \in X : d(x,y) < \varepsilon\}$$

for $S \subset X$ we define it ε–envelope, S_ε:

$$S_\varepsilon = \bigcup_{x \in S} B_\varepsilon(x)$$

we define the ε–unseen mass as follows:

$$U_n(\varepsilon) = P(X_+ \backslash S_\varepsilon)$$

Estimating the False-Alarm Component

- Berend and Kontorovish [BK11] : the mass of all the points outside ε-net (false-alarm component) is bounded by:

$$E[U_n(\varepsilon)] \leq \frac{1}{en}(\Delta/\varepsilon)^{ddim(X)+2}$$

- For any sample $X_1, \ldots, X_n$ achieving an ε-net:

$$\hat{\rho}_n \leq \rho \leq \hat{\rho}_n + 2\epsilon$$

- If S is ε-net then choose $\hat{\gamma} := \hat{\rho}_n + 2\epsilon$

[BK11] : Daniel Berend and Aryeh Kontorovich . The missing mass problem , in preparation . 2011

Estimating the Missed Anomaly Components

- What do we do about the missed anomaly?
  - We can't give a non-trivial bound $P(\hat{\gamma} > \gamma)$ since we don't know how close ρ is to γ
  - Instead, use the following heuristic:
  - Corresponding roughly to the assumption $\Pr[\rho + t\Delta > \gamma] \approx t$

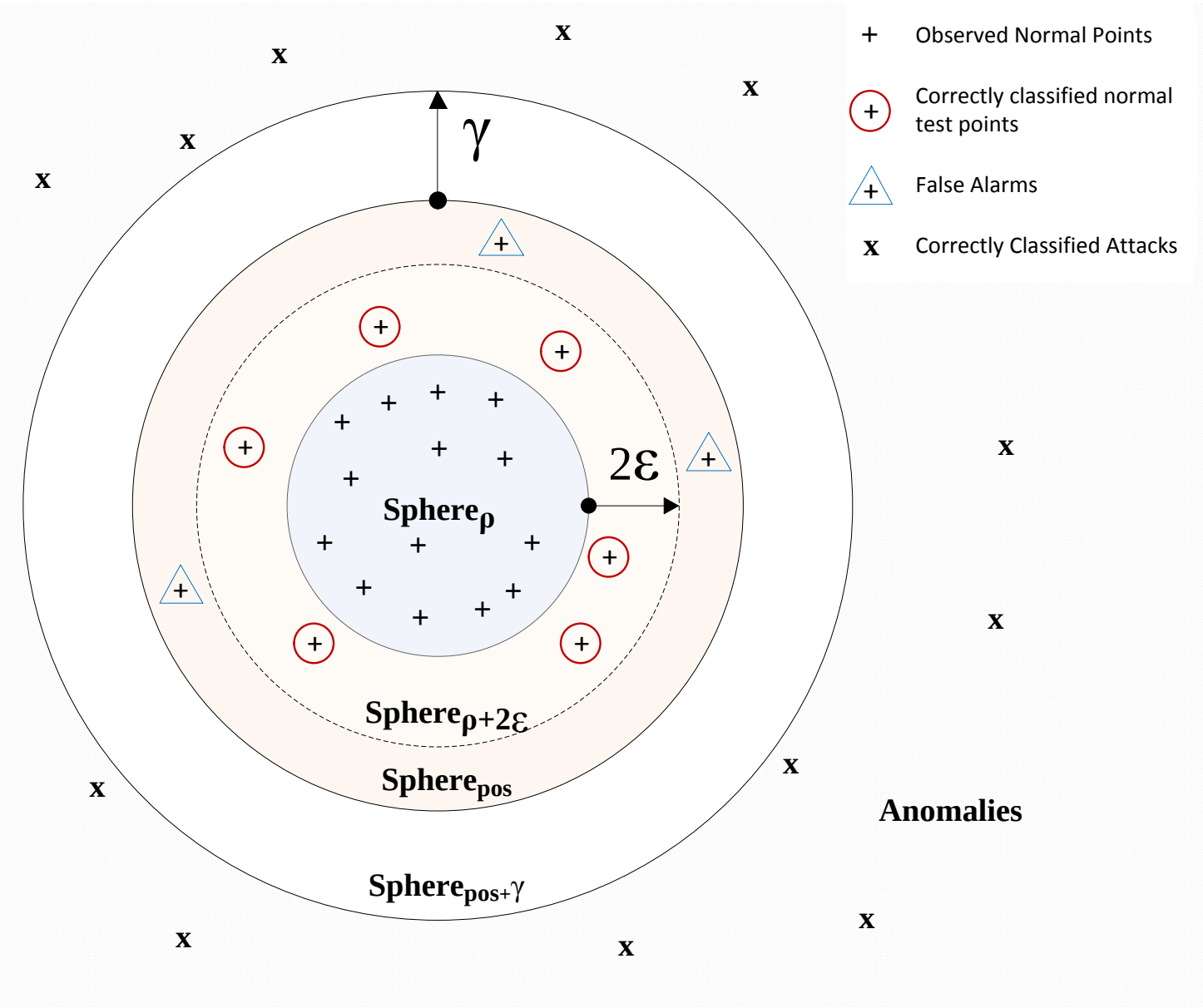
$$Missed\_Anomalies = \frac{2C\epsilon}{\Delta}$$

- Combining the two risk components:

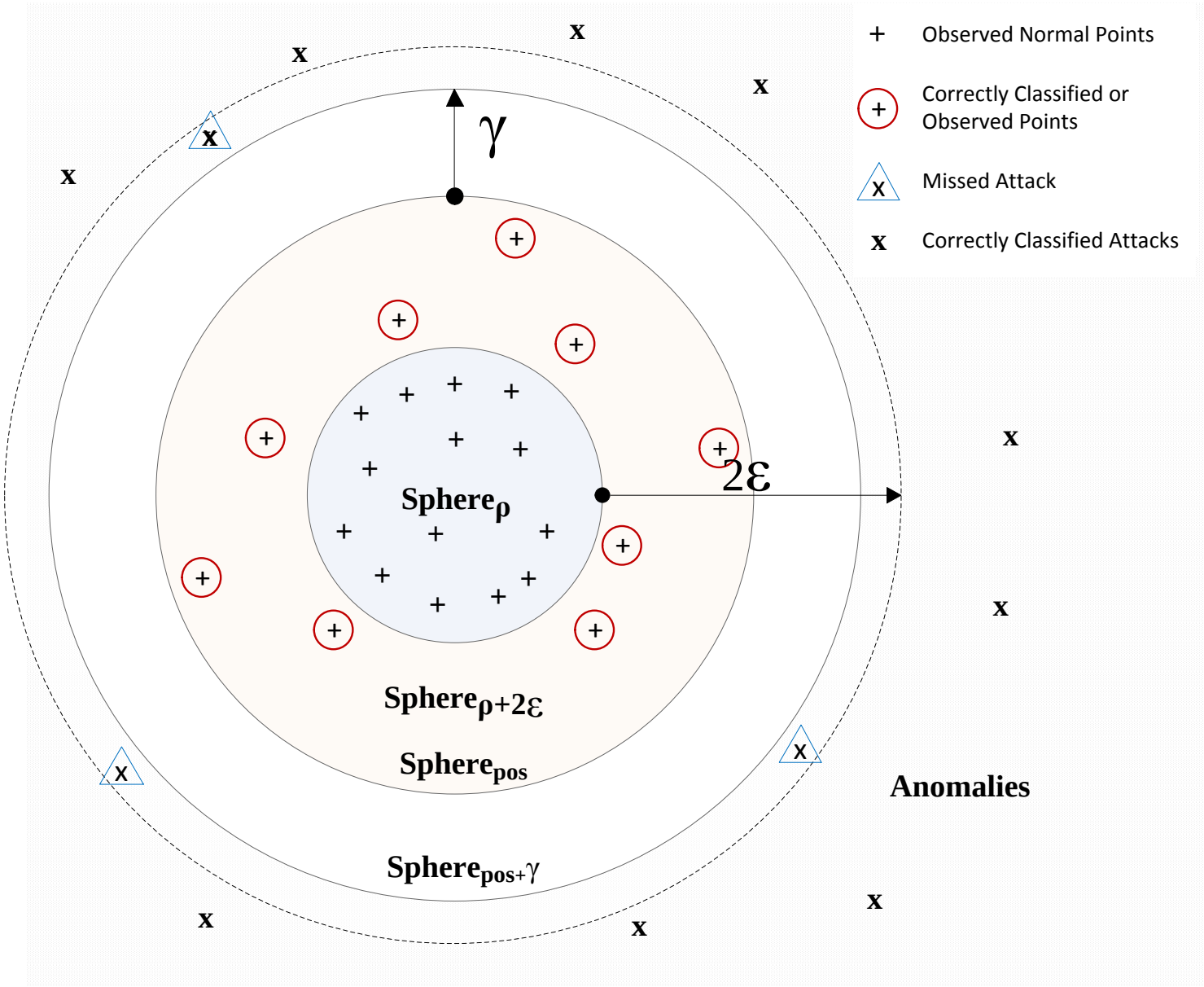
$$R_n(\varepsilon) := \frac{1}{en}(\Delta/\varepsilon)^{ddim(X)+2} + \frac{2C\epsilon}{\Delta}$$

which is minimized at $\varepsilon_n = \Delta^{ddim(X)+3}/2Cen$

False Alarms Are Possible if $\hat{\gamma} < \gamma$



Missed Anomalies Are Possible if $\hat{\gamma} > \gamma$



Section 5

EMPIRICAL EXPERIMENT

Classification Results and The Incurred Classification Cost

- The participating classifiers are the proposed cost-sensitive-classifier, denoted as "AAD", the Peer-Group-Analysis classifier, denoted as "PGA" and the Global-Density-Estimation, denoted as "GDE"

Dataset	Classifier	% Classification Error	% False Alarms	% Missed Attacks	Incurred Cost
2D-Single-Cluster	*AAD*	0.44	0.00	0.01	24,000.08
	GDE	16.03	0.00	0.91	273,000.1
	PGA	1.24	0.01	0.03	57,000.24
9D-Sphere	*AAD*	0.24	0.00	0.00	0.13
	GDE	28.45	0.29	0.00	15.65
	PGA	1.11	0.01	0.07	21,000.54
BGU ARP	*AAD*	0.18	0.00	0.00	0.14
	GDE	59.10	0.61	0.00	45.57
	PGA	4.55	0.01	1.00	300,000.9


Thank You!