Learning from Positive Examples

Christos Tzamos (UW-Madison)

Based on joint work with
V Contonis (UW-Madison), **C Daskalakis** (MIT), **T Gouleakis** (MIT),
S Hanneke (TTIC), **A Kalai** (MSR), **G Kamath** (U Waterloo), **M Zampetakis** (MIT)

Typical Classification Task

Positive Examples

Hi Mike,
Do you want to come over for dinner tomorrow?

Your Amazon.com order has shipped!

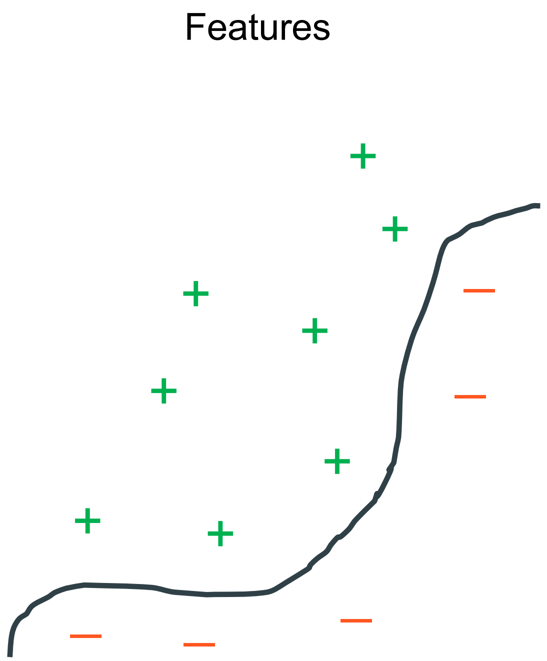
Valid Email

Dear Sir,
I am a Nigerian Prince…

Congrats! You won 1,000,000!!

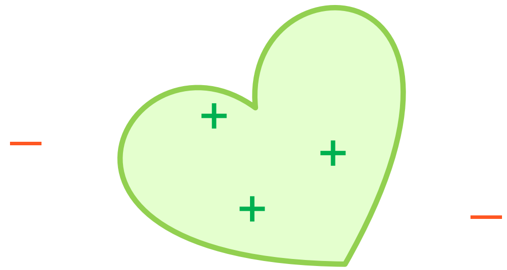
Invalid Email (Spam)

Negative Examples



Classification - Formulation

1. Unknown set **S** $\subseteq R^d$ of positive examples (target concept)
2. Points $x_1, \ldots, x_n$ in R^d are drawn from a distribution **D** (examples)
3. The examples are labeled *positive* if they are in S and *negative* otherwise.



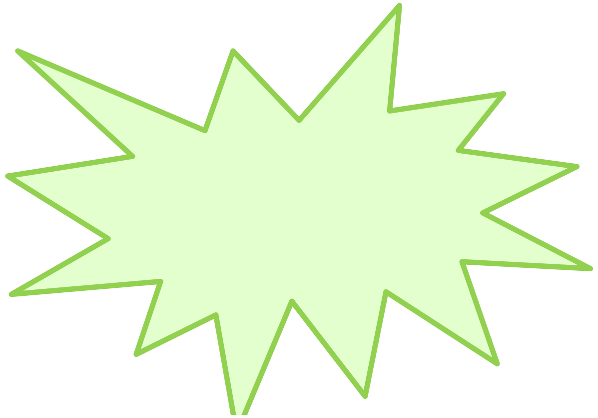
Goal: Find a set S' such that agrees with the set S on the label of a random example with high probability (> 99%)

How many examples are needed?

Complexity of Concepts

The samples needed depend on how complex the concept is.

vs



Arbitrary Distribution of Samples

Vapnik–Chervonenkis (VC) dimension

VC dimension **k** → **O(k)** samples suffice

Learning with positive examples

Learning from both positive and negative examples is well understood.
In many situations though, only positive examples are provided.

E.g. When a child learns to speak

"Mary had a little lamb"

"Twinkle twinkle little star"

"What does the fox say?"

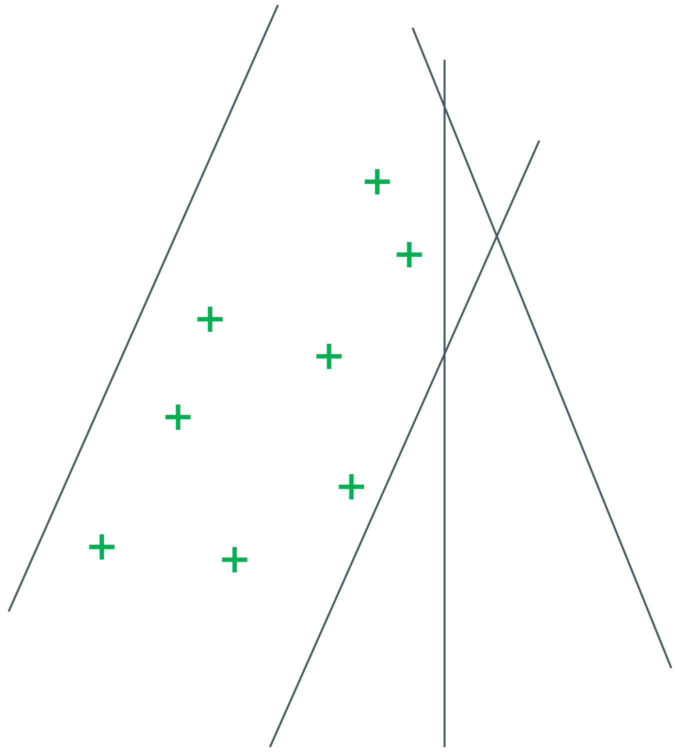
No negative examples are given

"Fox say what does"

"akjda! Fefj dooraboo"

Can we learn from positive examples?

Generally no! Need to know what examples are excluded.



Two approaches for learning

1. Assume data points are drawn from a structured distribution (e.g. Gaussian)

   "Learning Geometric Concepts from Positive Examples"

   (joint work with Contonis and Zampetakis)

2. Assume an oracle that can check the validity of examples (during training)

   "Actively Avoiding Nonsense in Generative Models"

   (joint work with Hanneke, Kalai and Kamath, **COLT 2018**)

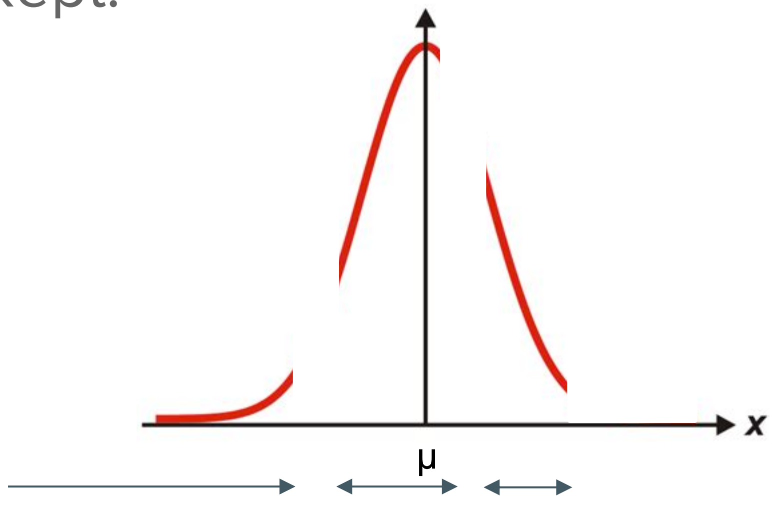
Learning from Normally Distributed Examples

Model

- Points $x_1, \ldots, x_n$ in R^d are drawn from a normal distribution **N(μ,Σ)** with unknown parameters.
- Only samples that fall into a set **S** are given.
- Assumption: at least 1% of the total samples are kept.

- **Goal:** Find **μ**, **Σ**, and **S**.

- Example: When **S** is a union of 3 intervals in 1-d.



Main Structural Theorem

- Suppose the set S has low complexity
  (Gaussian Surface Area at most γ)
- Consider the moments $E[x]$, $E[x^2]$, …, $E[x^k]$ of the positive samples for $k = \Theta(\gamma^2)$

Structural Theorem [Contonis, T, Zampetakis' 2018]

For any **μ'**, **Σ'**, and a set **S'** with Gaussian Surface Area at most γ that matches all $k=\Theta(\gamma^2)$ moments,

- **S** agrees with **S'** almost everywhere and,
- The distribution **N(μ',Σ')** is almost identical to **N(μ,Σ)**

Moreover, one can identify computationally efficiently **μ'**, **Σ'**, and **S'**

Ideas behind algorithm

- The moments of the positive samples are (proportional to)
  $E[x\, 1_S(x)]$, $E[x^2\, 1_S(x)]$, …, $E[x^k\, 1_S(x)\,]$ for random x drawn from **$N(\mu,\Sigma)$**

- The function $1_S(x)$ can be written as a sum of $\sum c_k H_k(x)$ where $H_k(x)$ is the degree k Hermite polynomial.

- Hermite polynomials form an orthonormal basis similar to the Fourier Transform.

- Knowing the k first moments, we can find the top k Hermite coefficients which give a low degree approximation of the function $1_S(x)$.
- For $k = \Theta(\gamma^2)$, the approximation is very accurate.

Corollaries

- $d^{O(\gamma^2)}$ samples suffice to learn a concept with Gaussian surface area γ. Need to estimate accurately all $d^{O(\gamma^2)}$ high-dimensional moments.

- Intersection of k halfspaces: $d^{O(\log k)}$
- Degree t polynomial-threshold functions: $d^{O(t^2)}$
- Convex Sets: $d^{O(\sqrt{d})}$

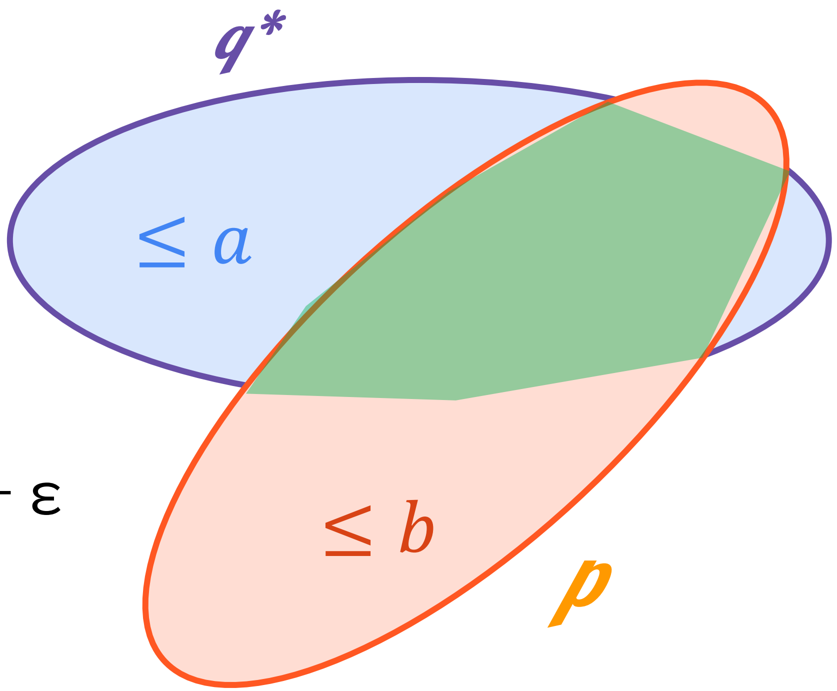
Learning with access to a Validity Oracle

Setting

Sample access to an unknown distribution p supported on an unknown set.
Can query an oracle whether an example x is in $supp(p)$.
A family $\mathcal{Q}$ of probability distributions with varying supports.

Assuming a q^* in $\mathcal{Q}$ exists such that

$\Pr_{x\sim q^*}[x \notin supp(p)] \leq a$ and $\Pr_{x\sim p}[x \notin supp(q^*)] \leq b$



find a q

$\Pr_{x\sim q}[x \notin supp(p)] \leq a + \varepsilon$ and $\Pr_{x\sim p}[x \notin supp(q)] \leq b + \varepsilon$

Generative Model - Neural Net

Many governments recognize the military housing of the [[Civil Liberalization and Infantry Resolution 265 National Party in Hungary]], that is sympathetic to be to the [[Punjab Resolution]] (PJS) [http://www.humah.yahoo.com/guardian.cfm/7754800786d17551963s89.htm Official economics Adjoint for the Nazism, Montgomery was swear to advance to the resources for those Socialism's rule, was starting to signing a major tripad of aid exile.]]

-- Char-RNN trained on *Wikipedia* (Karpathy)

Generative Model - Neural Net

Many governments recognize the military housing of the [[Civil Liberalization and Infantry Resolution 265 National Party in Hungary]], that is sympathetic to be to the [[Punjab Resolution]] (PJS) [http://www.humah.yahoo.com/guardian.cfm/7754800786d17551963s89.htm Official economics Adjoint for the Nazism, Montgomery was swear to advance to the resources for those Socialism's rule, was starting to signing a major tripad of aid exile.]]

-- Char-RNN trained on *Wikipedia* (Karpathy)

Generative Model - Neural Net

Many governments recognize the military housing of the [[Civil Liberalization and Infantry Resolution 265 National Party in Hungary]], that is sympathetic to be to the [[Punjab Resolution]] (PJS) [http://www.humah.yahoo.com/guardian.cfm/7754800786d17551963s89.htm Official economics Adjoint for the Nazism, Montgomery was swear to advance to the resources for those Socialism's rule, was starting to signing a major tripad of aid exile.]]

-- Char-RNN trained on *Wikipedia* (Karpathy)

Generative Model - Neural Net

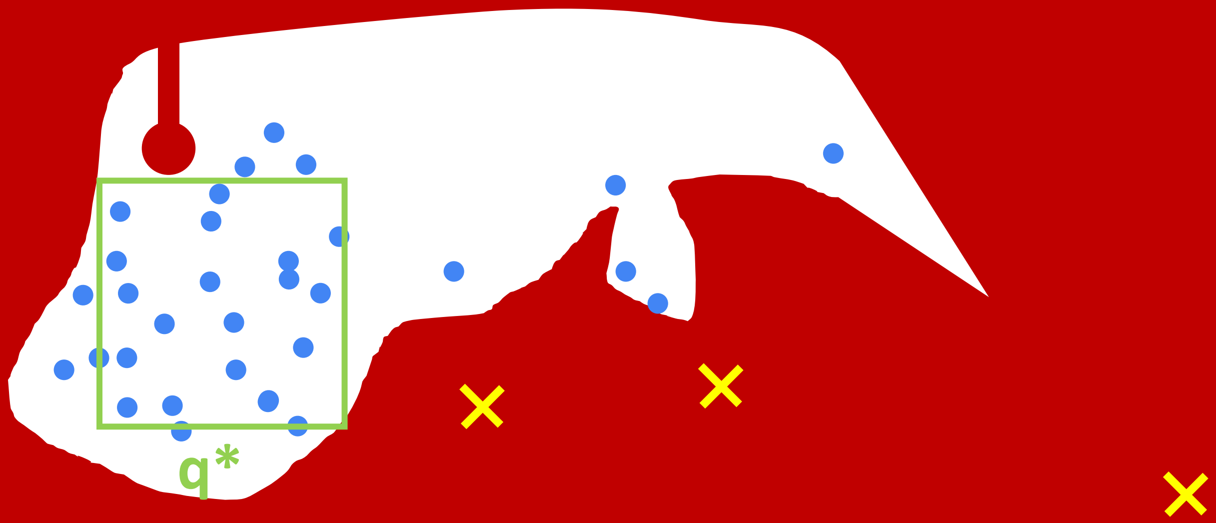
Many governments recognize the military housing of the [[Civil Liberalization and Infantry Resolution 265 National Party in Hungary]], that is sympathetic to be to the [[Punjab Resolution]] (PJS) [http://www.humah.yahoo.com/guardian.cfm/7754800786d17551963s89.htm Official economics Adjoint for the Nazism, Montgomery was swear to advance to the resources for those Socialism's rule, was starting to signing a major tripad of aid exile.]]

-- Char-RNN trained on *Wikipedia* (Karpathy)

NONSENSE!

NONSENSE!

NONSENSE!



NONSENSE!

NONSENSE!

NONSENSE!

Example: Rectangle Learning

Consider again the problem instance where:

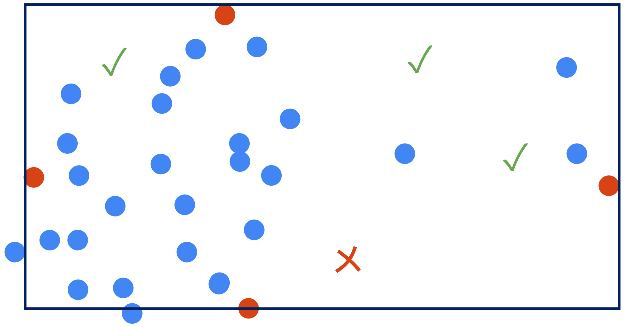
$\mathcal{Q}$ is the class of all *Uniform* distributions over rectangles [a,b]x[c,d]

Draw many samples from p

For any quadruple of points

Choose $q \in \mathcal{Q}$ specified by their bounding box

Draw many samples from q to estimate validity querying the oracle $supp(p)$



★ **Can learn using O(1/ε^2) samples from p and O(1/ε^5) queries to $supp(p)$.**

In d-dimensions, uses **O(d/ε^2)** samples and **O(1/ε^{2d+1})** queries.

Curse of dimensionality

(The previous algorithm is tight…)

Theorem: To find a *d-dimensional* box q in $\mathcal{Q}$ such that

$$\Pr_{x\sim p}[x \notin supp(q)] \leq \Pr_{x\sim p}[x \notin supp(q^*)] + \epsilon \quad \text{and} \quad \Pr_{x\sim q}[x \notin supp(p)] \leq \epsilon$$

one needs to make **exp**(d) queries to the $supp(p)$ oracle.

Lower-bound requires q in $\mathcal{Q}$ (proper learning)!!!

We show that if q is not required to be in $\mathcal{Q}$, it is possible to learn efficiently.

Main Result

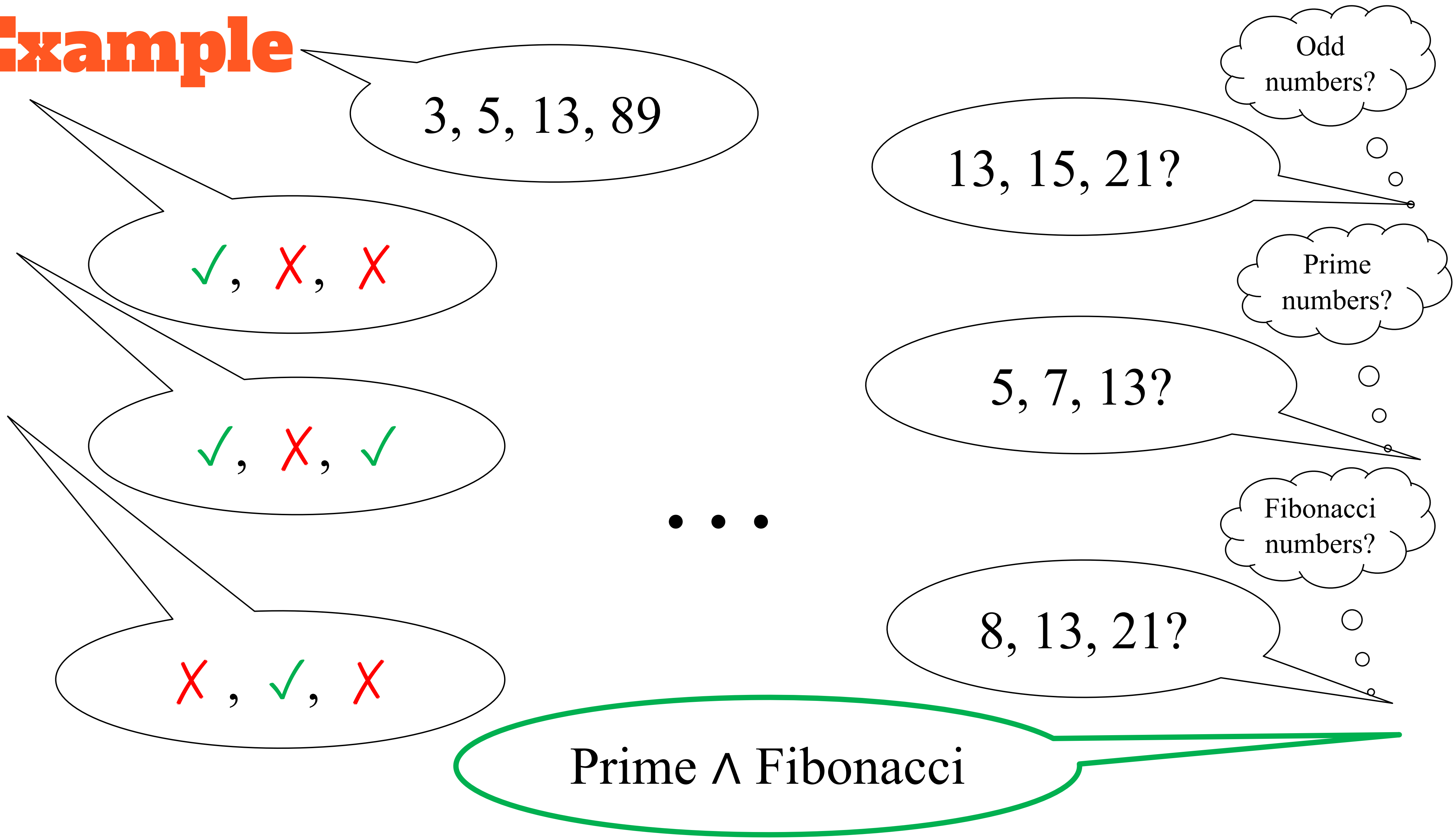
Theorem [Hanneke, Kalai, Kamath, T, COLT'18]:

For any class of distributions $\mathcal{Q}$, one can find a q such that

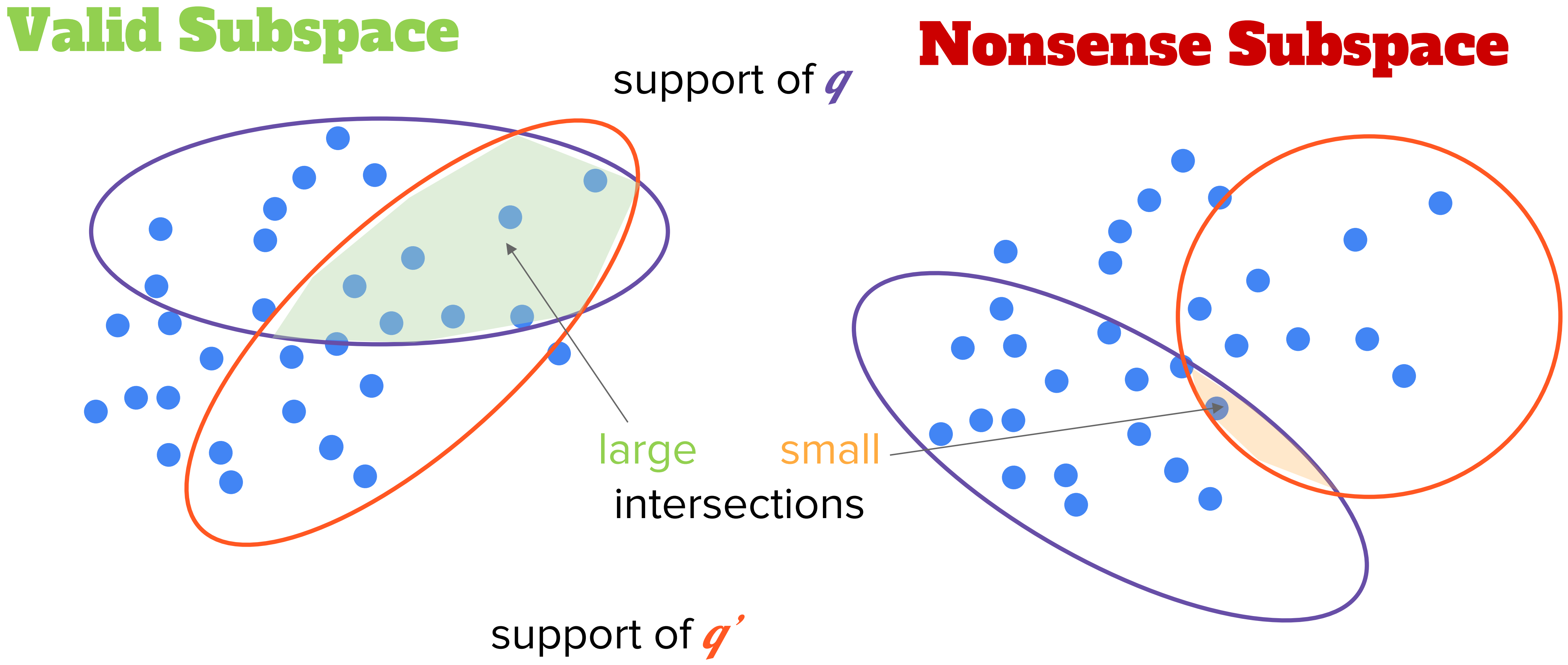
$\Pr_{x\sim p}[x \notin supp(q)] \leq \Pr_{x\sim p}[x \notin supp(q^*)] + \epsilon$ and $\Pr_{x\sim q}[x \notin supp(p)] \leq \epsilon$

using only **poly**(VC-dim($\mathcal{Q}$), ϵ^{-1}) samples from p and queries to $supp(p)$.

Example


3, 5, 13, 89
✓, ✗, ✗
✓, ✗, ✓
✗, ✓, ✗
Odd numbers?
13, 15, 21?
Prime numbers?
5, 7, 13?
Fibonacci numbers?
8, 13, 21?
. . .
Prime ∧ Fibonacci

Why does this work?



Summary

Learning from positive examples

- Not possible without assumptions
- Proposed a framework for learning when samples are normally distributed
- Alternatively, possible to learn if one can query an oracle for validity

Further work

- Learning the Gaussian parameters requires only $O(d^2)$ samples for any concept class with validity oracle [Daskalakis, Gouleakis, **T**, Zampetakis, FOCS'2018]

Thank You!