

# **An Inequality for Nearly Log-concave Distributions with Applications to Learning**

Shie Mannor

*McGill University*

(with Constantine Caramanis, MIT)

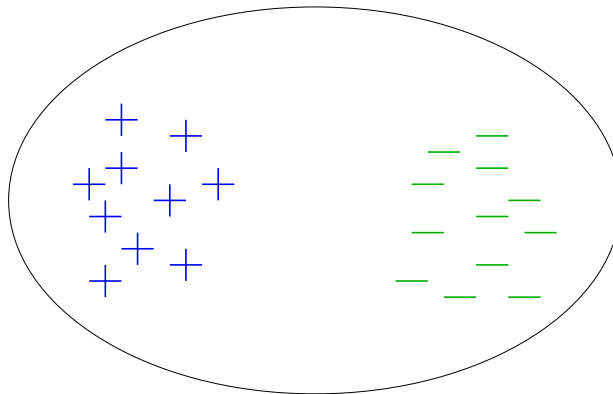
February 2006

# Outline

1. Motivation
2. Nearly Log-concave functions
3. An Isoperimetric Inequality
4. Applications to Learning: Lower Bounds
  - A General Theorem
  - Linear Classifiers
  - On the size of the margin
  - Regression

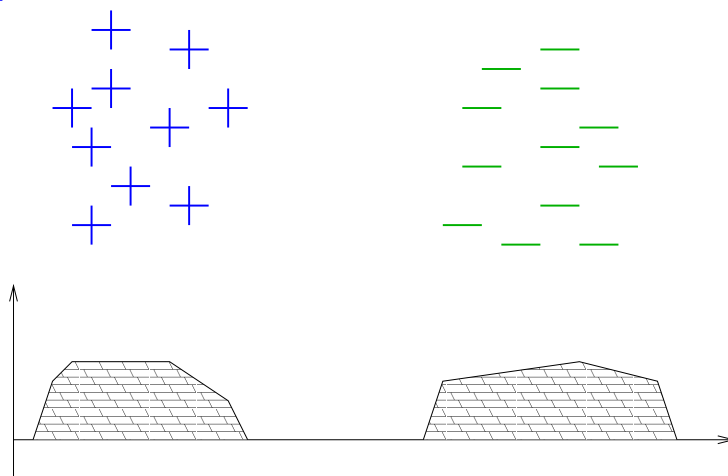
# Part I: Motivation

Good News?

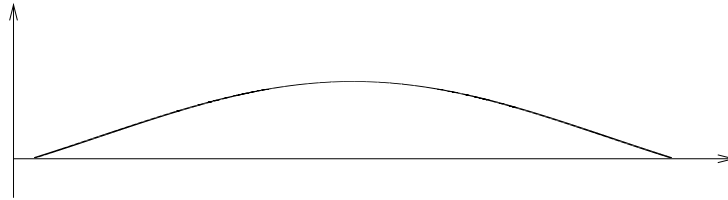
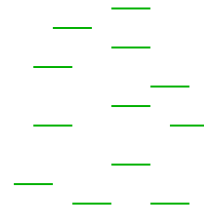
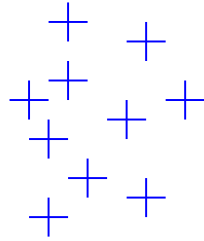


- Two Main Questions:
  - Is this Good News?
  - How Likely are we to get “Good News” ?

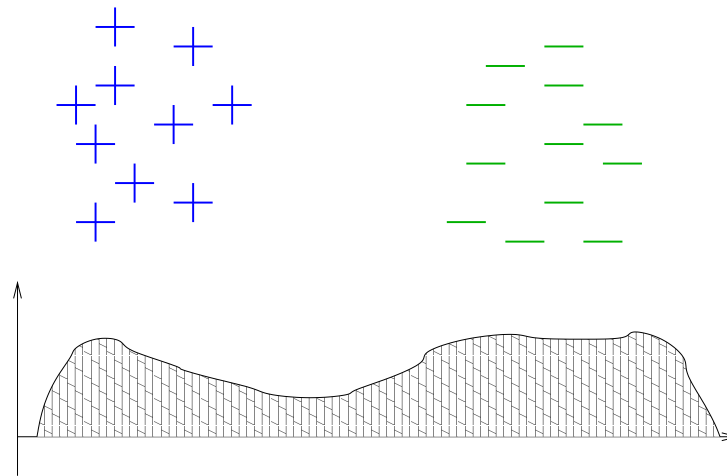
# Good News??



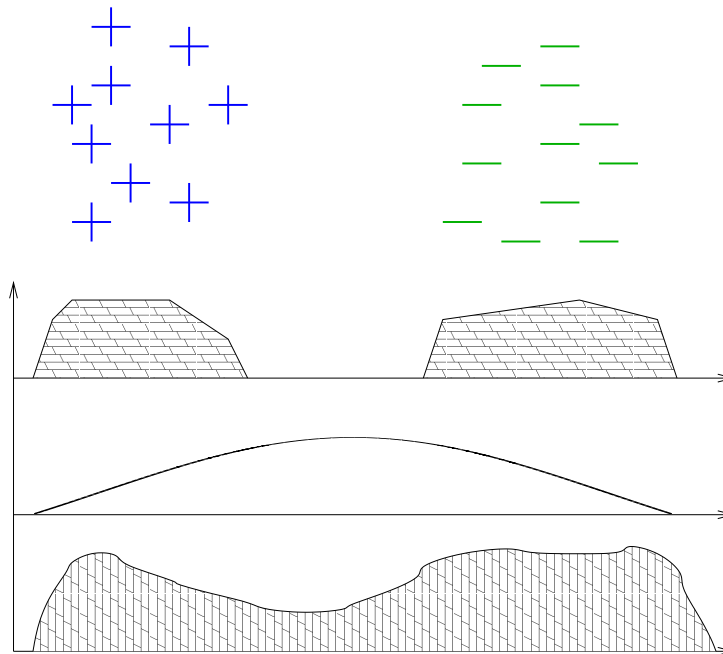
# Good News???



# Good News????



# Good News?????



# Lower Bounds in Supervised Learning (PAC)

Classical lower bounds: Given a function class  $\mathcal{F}$  one needs at least

$$m \geq \Omega \left( \frac{1}{\epsilon} \max \left( \text{Complexity}(\mathcal{F}), \log\left(\frac{1}{\delta}\right) \right) \right)$$

samples to get an  $\epsilon$  optimal solution w.p. at least  $1 - \delta$ .

Lower bounds are based on the following lemma: Given a coin with bias  $1/2 + \epsilon/2$  or  $1/2 - \epsilon/2$ , one needs  $1/\epsilon^2 \log(1/\delta)$  to decide correctly w.p. at least  $1 - \delta$ .

Pathological worst-case distribution.

Bounds are **a-priori**.

# Lower Bounds in Supervised Learning

We want bounds that are:

1. Data dependent
2. Distribution dependent (restricted class)
3. Tight
4. Computable

Our focus: lower bound on the generalization error not sample complexity.

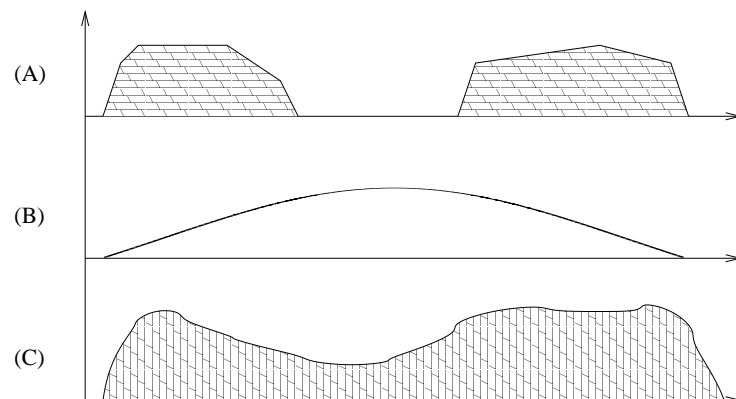
## Part II: Nearly Log-Concave Functions

- The Definition:  $F(x)$  is  $\beta$ -log-concave if:

$$F(\lambda x + (1 - \lambda)y) \geq e^{-\beta} F(x)^\lambda F(y)^{1-\lambda}.$$

- 0-log-concave functions:
  - Gaussian, Uniform, Logistic, Exponential distributions.
- A much richer class:  $\beta$ -log-concave functions
  - Need not be continuous.
  - Mixtures of Gaussians
  - Mixtures with bounded Radon-Nikodym derivative
  - Convolutions of  $\beta_1$  and  $\beta_2$  log-concave functions

## Nearly Log-Concave?

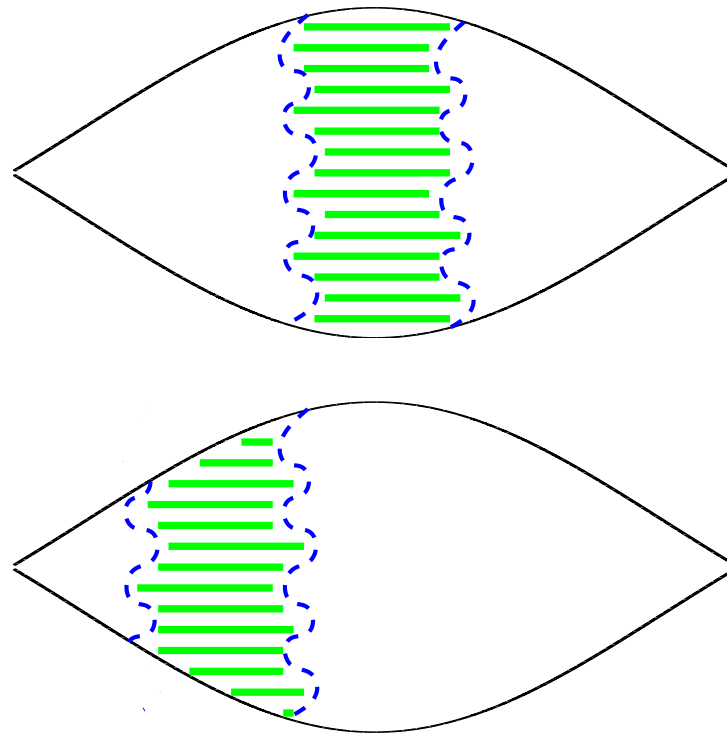


- (A) is not  $\beta$ -log-concave for any finite  $\beta$ .
- (B) is 0-log-concave (it is a Gaussian).
- (C) is  $\beta$ -log-concave for some finite  $\beta > 0$ .

# Properties of Nearly Log-Concave Functions

- They are:
  - Not necessarily continuous;
  - Not necessarily unimodal;
- However...
  - There are no big “holes” or “valleys” in the mass distribution.

## Three Way Sharing: Take the Middle Slice



## Part III: The Main Inequality: How fat is the Margin?

- For  $K$  a closed, bounded, convex set, with a **decomposition**  $K = K_1 \cup B \cup K_2$
- For any  $\beta$ -log-concave distribution  $F$  with induced measure  $\mu$
- We have:

$$\mu(B) \geq e^{-\beta \frac{d(K_1, K_2)}{\text{diam}(K)}} \min\{\mu(K_1), \mu(K_2)\}.$$

- This inequality is dimension-free (!).
- Cannot relax **any** multiplicative factor.

Inequality is tight up to a factor of 2.

## How Do You Prove Such a Result?

Result is “fundamental” (strengthen results by Kannan and Lovátz)

Prove by induction on the dimension

One dimension - elementary proof

Induction step - assume result is violated in  $n$  dimensions, show it is violated in  $n - 1$

Key argument - Löwner-John polytopes make sure we can argue in terms of “flat” ellipsoids.

# Part IV: Applications in Machine Learning

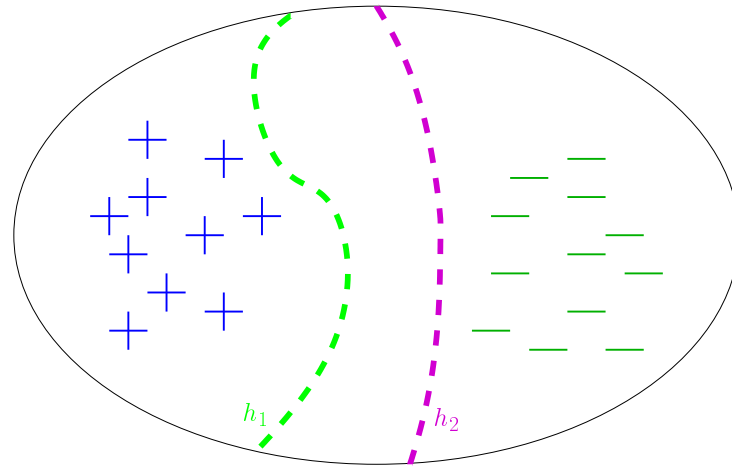
Mini- agenda

1. A Lower Bound for Classification
2. On the size of the margin
3. Regression

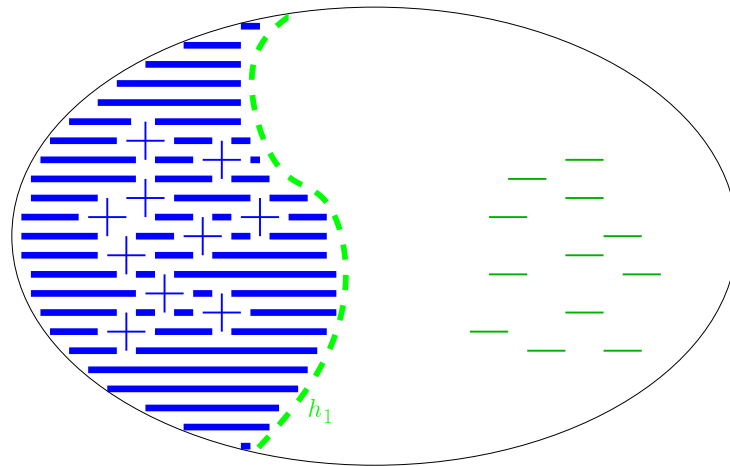
## Part IV.1: Classification Error: Lower Bounds

- The set-up:
  - Data points  $\{x_i\}$  given, with labels  $\{y_i\}$ .
  - Performance is judged against a  $\beta$ -log-concave distribution; **may be different** from the distribution that generated the data.
- Not the “classical” PAC set-up.

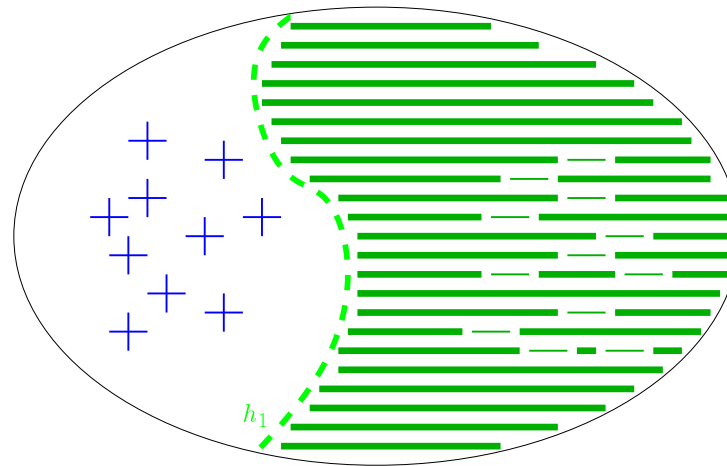
# Measuring Error



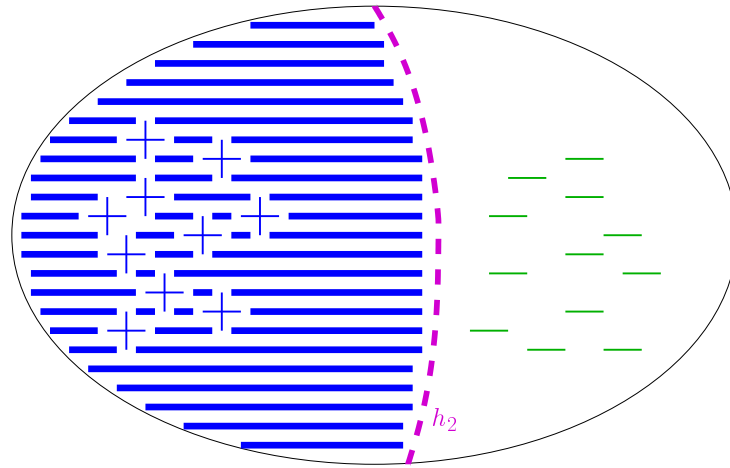
# Measuring Error



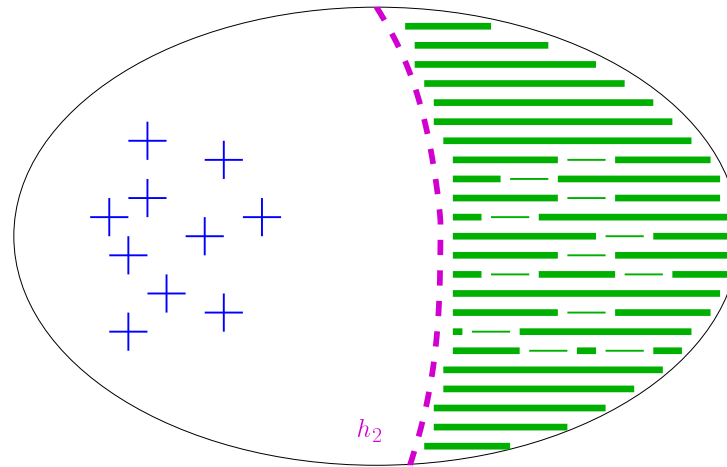
# Measuring Error



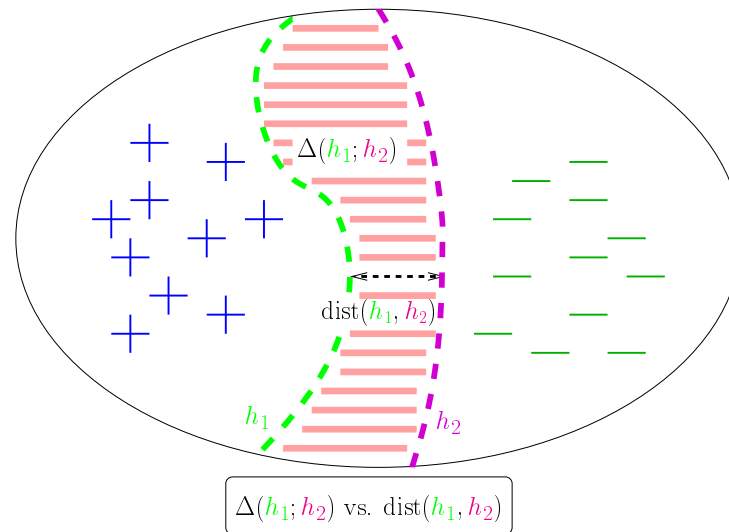
# Measuring Error



# Measuring Error



# Measuring Error



# Distance Measures

- Two distance measures: Given two classifiers,  $h, h'$ :
  - $\Delta(h_1; h_2)$  – measure of region  $\{h_1 \neq h_2\}$ .
  - $\text{dist}(h_1, h_2)$  – separation distance.
- We cannot compute  $\Delta(h_1; h_2)$  without knowledge of  $f(x)$ .
- We may be able to compute  $\text{dist}(h_1, h_2)$ , at least in principle.

# Distance Measures

- Two distance measures: Given two classifiers,  $h, h'$ :
  - $\Delta(h_1; h_2)$
  - $\text{dist}(h_1, h_2)$
- We cannot compute  $\Delta(h_1; h_2)$  without knowledge of  $f(x)$ .
- We may be able to compute  $\text{dist}(h_1, h_2)$ , at least in principle.
  - $\implies$  We use  $\text{dist}(h_1, h_2)$  to compute a lower bound for  $\Delta(h_1; h_2)$ .

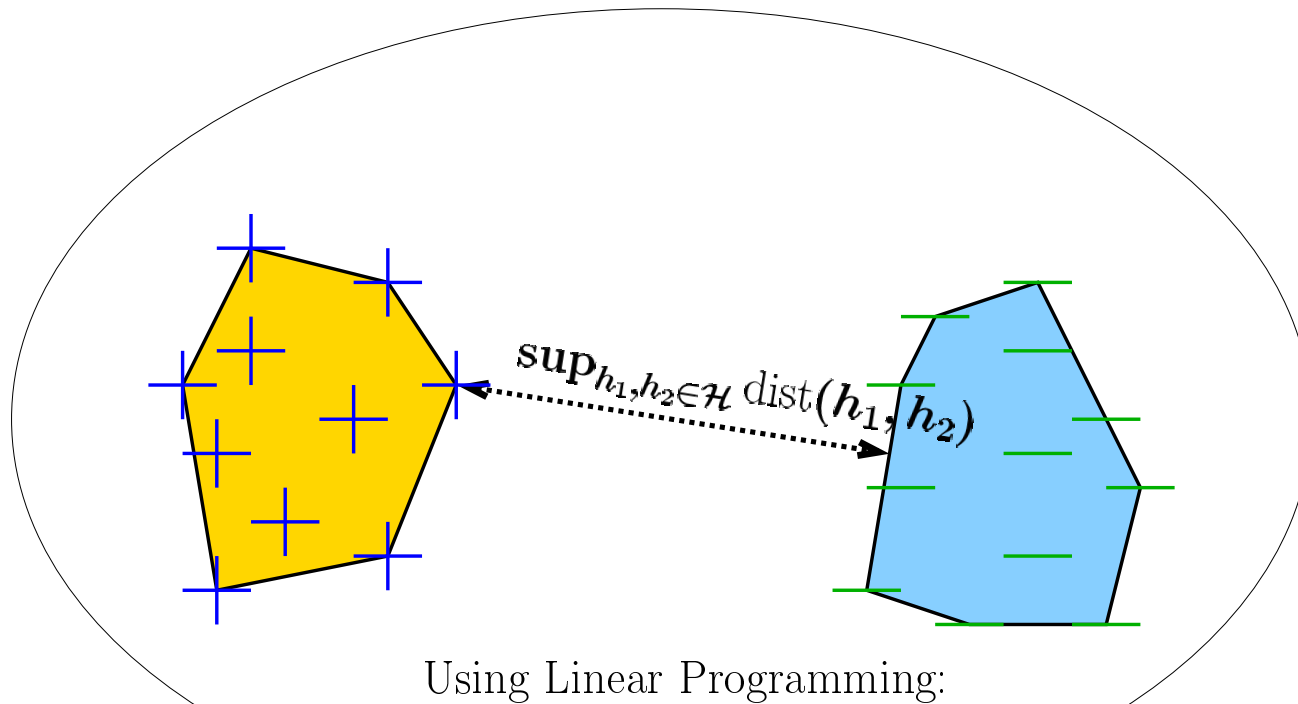
## A First Theorem: A Lower Bound on Error

- If  $f(x)$  is  $\beta$ -log-concave, with measure  $\mu$ , and  $K$  is bounded,
- For every  $h \in \mathcal{H}$ , there exists  $h' \in \mathcal{H}$  such that

$$\Delta(h; h') \geq \frac{1}{2} \left[ \frac{e^{-\beta} P_0}{\text{diam}(K)} \right] \sup_{h_1, h_2 \in \mathcal{H}} \text{dist}(h_1, h_2),$$

- Note that this inequality is **dimension free**.

# Example: Linear Classifiers



Using Linear Programming:

$$\sup_{h_1, h_2 \in \mathcal{H}} \text{dist}(h_1, h_2) = d(\text{conv}(+), \text{conv}(-))$$

## Extensions

Main result still works with an unbounded set  $K$ , but finite second moment:

$$\sigma^2 = \int_K \|x - \bar{x}\|_2^2 f(x) dx < \infty.$$

Then the induced measure  $\mu$  satisfies for every partition  $K = K_1 \cup K_2 \cup B$ :

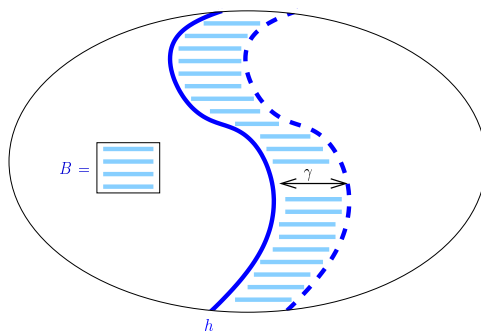
$$\mu(B) \geq e^{-\beta d(K_1, K_2)} \frac{1}{4\sqrt{2}\sigma} \min\{\mu(K_1)^{3/2}, \mu(K_2)^{3/2}\}.$$

## Part Part IV.2 : How Often Do We Get Lucky...?

- Suppose we sample from  $\beta$ -log-concave distribution.
- A classifier  $h$  is given.

$$K^-(h) = \{x : h(x) = -1\}.$$

- How likely that a sample does not land within the geometric margin?
  - That is: how big is the measure of  $B^-$ ?



We want to bound the event that

$$\left\{ \min_{i: x_i \in K^+(h)} d(x_i, K^-(h)) > \gamma \right\}$$

Let  $B = \{x \in K^+(h) : d(x, K^-(h)) < \gamma\}$ . We have

$$\mu(B) \geq \gamma C_1 \min \left\{ \mu(K^-(h)), C_2 \mu(K^+(h)) \right\}.$$

**Proposition:** For every  $\gamma > 0$  given  $N$  samples from a  $\beta$ -log-concave  $f$ :

$$\begin{aligned} & \Pr \left( \min_{\{i: x_i \in K^-(h)\}} d(x_i, K^+(h)) > \gamma \right) \\ & \leq \exp \left( -N \gamma C \min \left\{ \mu(K^+(h)), \frac{\mu(K^-(h))}{1 + \gamma C} \right\} \right), \end{aligned}$$

## The Symmetric Case

Let

$$B^{symm} = \{x \in K^-(h) : d(x, K^+(h)) < \gamma\} \\ \cup \{x \in K^+(h) : d(x, K^-(h)) < \gamma\}.$$

Let  $f$  be a  $\beta$ -log-concave distribution on  $K$  with induced measure  $\mu$ . Then

$$\mu(B^{symm}) \geq \gamma \frac{e^{-\beta}}{\text{diam}(K)} \min \left\{ \mu(K^+(h)), \mu(K^-(h)) \right\}.$$

A similar probabilistic bound follows.

## So What's The Catch?

People often consider the gap as:

$$\text{gap}(x_1, \dots, x_N; h) = \min_{i, j: h(x_i) \neq h(x_j)} d(x_i, x_j)$$

This **cannot** be bounded in a dimension free matter.

Conclusion: Large gap can only occur if

1. The distribution is not  $\beta$ -log-concave, or
2. Dimensions are added artificially.

## And What About the Margin?

If  $h$  is Lipschitz, the margin between  $x_1, \dots, x_n$  and  $h$  is proportional to:

$$\text{margin}(x_1, \dots, x_N; h) \propto \min \left\{ d((x_1, \dots, x_N \cap K^-(h)), K^+(h)), \right. \\ \left. ((x_1, \dots, x_N \cap K^+(h)), K^-(h)) \right\}.$$

A similar bound holds - the probability of a large margin decreases exponentially to 0.

Can also bound  $\Pr\{\sup_{h \in \mathcal{H}} \text{margin}(x_1, \dots, x_N; h) > \gamma\}$  using covering numbers.

## Part IV.3: Regression Tubes

Suppose we have a problem of the form

$$Y = k(X) + \text{Noise},$$

where  $k$  is unknown and  $x$  is sampled according to some pdf.

We let the tube be defined as:

$$T_{\epsilon_0, \epsilon_1}^k = \{(x, y) : \epsilon_0 \leq \|k(x) - y\| \leq \epsilon_1\}.$$

Basic question: How “fat” is the tube around  $k$ ?

Cost of converting  $\epsilon_0$ -sensitive error to  $\epsilon_1$ -sensitive error.

We will look at more general noise models

# Independent Additive Noise

If:

1.  $Y = k(X) + N$
2.  $N$  is independent of  $x$  with support  $K_Y$
3.  $N$  is  $\beta$ -log-concave.

Then

$$\mu(T_{\epsilon_0, \epsilon_1}^k) \geq (\epsilon_1 - \epsilon_0) \cdot \frac{e^{-\beta}}{\text{diam}(K_Y)} \min \left\{ \mu(T_{0, \epsilon_0}^k), \mu(T_{\epsilon_1, \text{diam}(K)}^k) \right\}.$$

Bound still holds if we replace  $k$  with  $k'$ .

Nearly linear error differential on the boundary.

## Joint Distribution

If the joint distribution is  $\beta$ -log-concave and  $k$  Lipschitz continuous:

$$\mu(T_{\epsilon_0, \epsilon_1}^k) \geq (\epsilon_1 - \epsilon_0) \frac{e^{-\beta}}{L \operatorname{diam}(K)} \min \left\{ \mu(T_{0, \epsilon_0}^k), \mu(T_{\epsilon_1, \operatorname{diam}(K)}^k) \right\}.$$

The linear case:

$$Y = a^\top X + N$$

if  $X$  is  $\beta_1$ -log-concave and  $N$  is  $\beta_2$ -log-concave we obtain the result with  $\beta := \beta_1 + \beta_2$ .

## Other Goodies

Results hold for finite moments as well.

Additional results hold under different assumptions on noise.

For example, if  $X$  is  $\beta$ -log-concave and  $Y|X$  is  $\beta'$ -log-concave the main theorem still holds with  $\beta + \beta'$ .

General conclusion: the boundary of the tube must carry a lot of weight.

Roughly linear in the differential

# Overview

- A new look at lower bounds - not PAC at all
- A weak structural assumption leading to using distances instead of measures.
- Main Point: For  $\beta$ -log-concave distributions, good separation means the no-man's-land must carry a lot of weight.
- If not  $\beta$ -log-concave, problem is “easy” to start with (?)

# Applications to learning

Consider two scenarios

- Data are generated by **unknown** distribution; performance judged by a (different)  $\beta$ -log-concave distribution:

A large margin is Bad News.

- Data are generated by a  $\beta$ -log-concave distribution:

A large margin/gap is unlikely.