

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

Shie Mannor

McGill University

(with Constantine Caramanis, MIT \Rightarrow UT Austin)

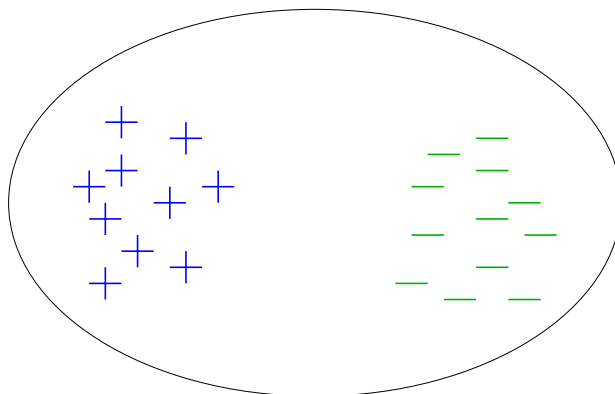
June 2006

Outline

1. Motivation
2. Nearly Log-concave functions
3. An Isoperimetric Inequality
4. Applications to Learning:
 - (a) Lower bounds on classification error
 - (b) On the size of the margin
 - (c) Regression

Part I: Motivation

Good News?

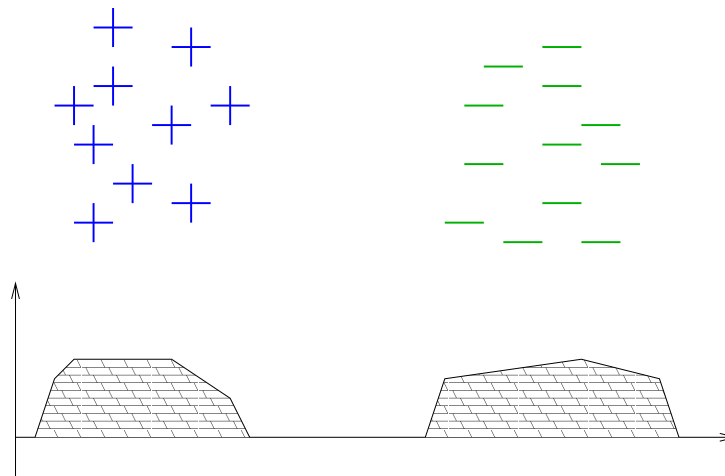


Medical applications, non-IID samples, active learning, etc.

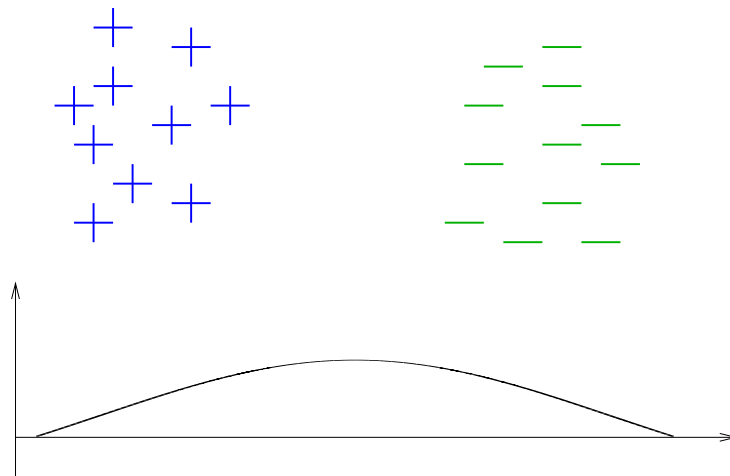
Two Main Questions:

- Is this Good News?
- How likely are we to get “Good News” if data are IID?

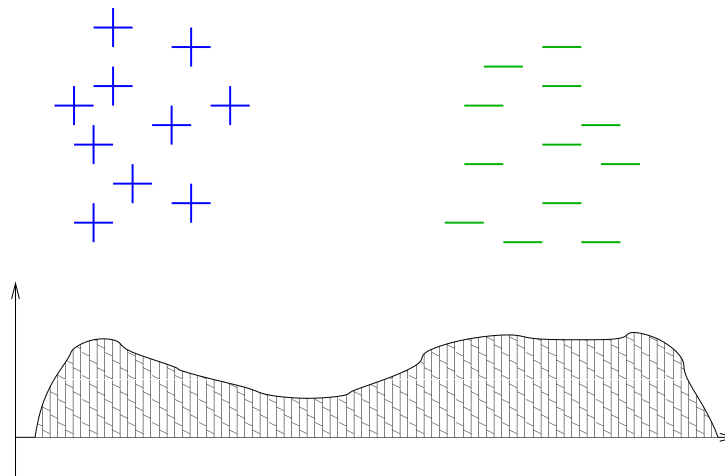
Good News??



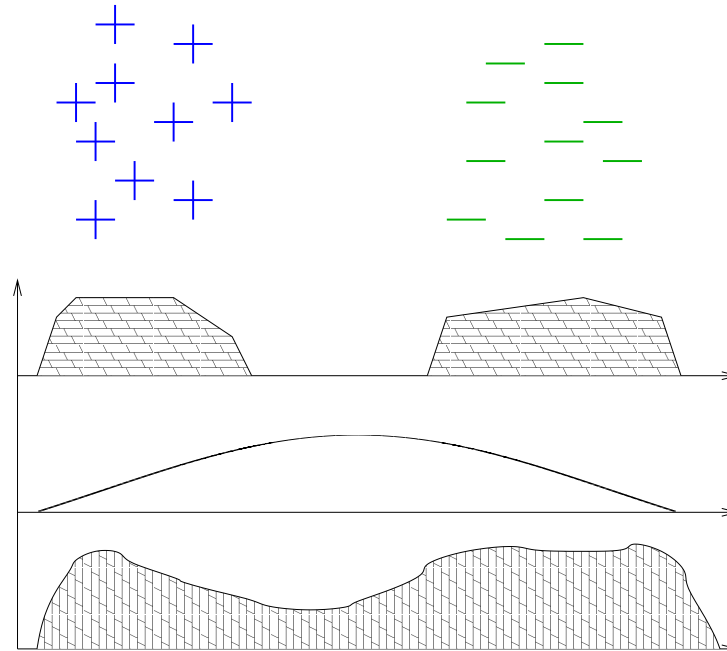
Good News???



Good News????



Good News?????



Lower Bounds in Supervised Learning (PAC)

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

$$m \geq \Omega \left(\max \left(\frac{1}{\epsilon} \text{Complexity}(\mathcal{F}), \frac{1}{\epsilon^2} \log\left(\frac{1}{\delta}\right) \right) \right)$$

samples to get an ϵ 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.

The Ideal Bounds

Given the available data x_1, \dots, x_n , a family of concepts \mathcal{F} and a family of “reasonable” distributions Ω .

Find a lower bound

$$\max_{\omega \in \Omega} \min_{F \in \mathcal{F}} \text{Err}_\omega(F) \geq \star$$

Such that:

1. Ω is “consistent” with data x_1, \dots, x_n .
2. \star depends on data, Ω and \mathcal{F} in a “simple” manner.

Scenario: sampling and testing on different but “similar” distributions.

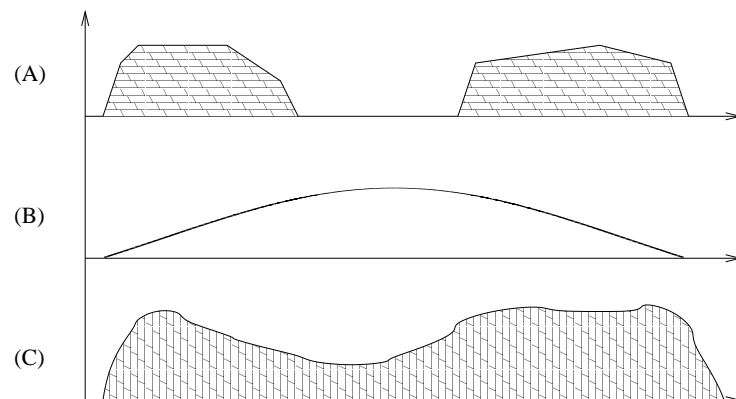
Part II: Nearly Log-Concave Functions

- Definition: $g(x)$ is β -log-concave if:

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

- 0-log-concave functions:
 - Gaussian, Uniform, Logistic, Exponential distributions.
- A much richer class: β -log-concave functions
 - Need not be continuous.
 - Mixtures of Gaussians
 - Mixtures with bounded Radon-Nikodym derivative
 - Convolutions of β_1 and β_2 log-concave functions

Nearly Log-Concave?



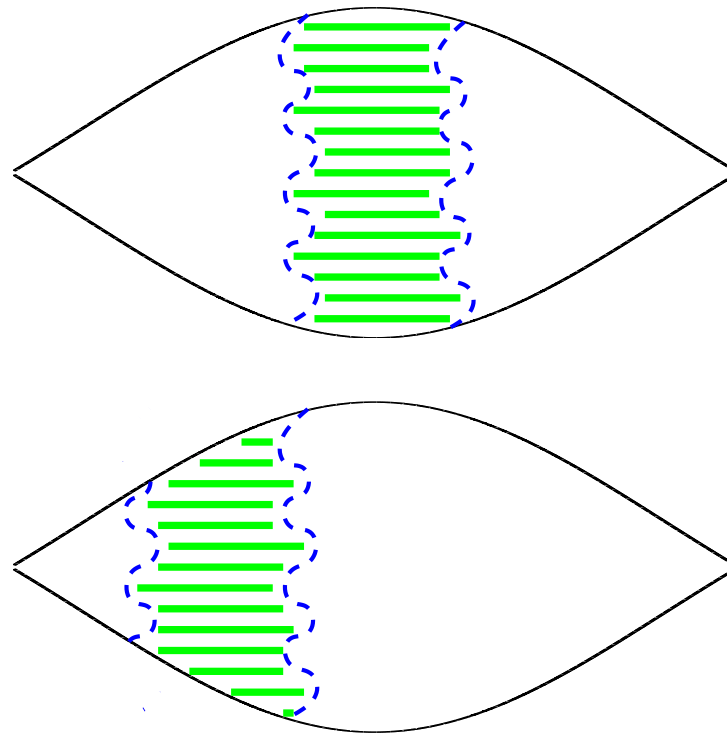
- (A) is not β -log-concave for any finite β .
- (B) is 0-log-concave (it is a Gaussian).
- (C) is β -log-concave for some finite $\beta > 0$.

Properties of Nearly Log-Concave Functions

- They:
 - Are not necessarily continuous;
 - Are not necessarily unimodal;
 - Have a convex support.

- 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 β -log-concave distribution g with induced measure μ
- 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 (open question).

Proof method

Result strengthens previous results by Kannan and Lovász

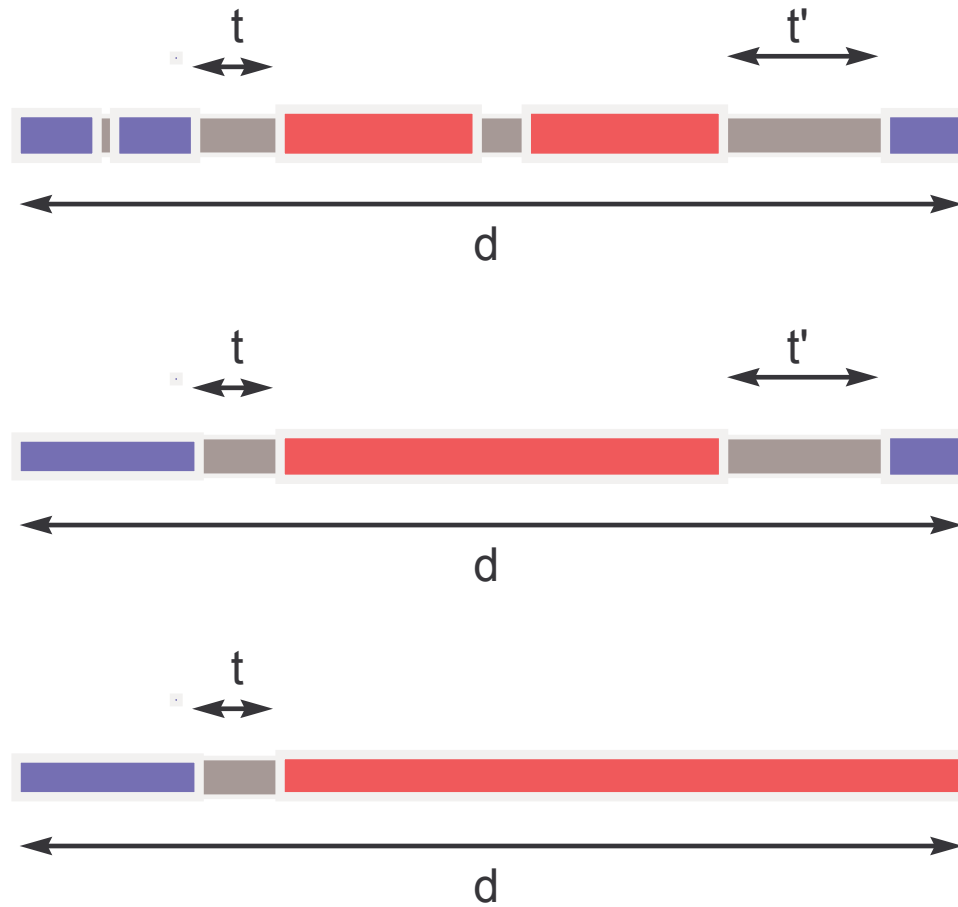
Proof by induction on the dimension

One dimension - elementary proof

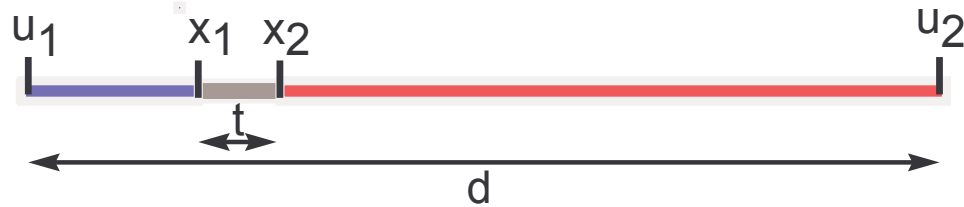
Induction step - assume result is violated in n dimensions, show it is violated in $n - 1$

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

Proof: One Dimension I



Proof: One Dimension II



Basic fact: If $g(x)$ is β -log-concave, $\forall x \in [x_1, x_2]$:

$$g(x) \geq e^{-\beta} g(y) \quad \forall y \in [u_1, x_1] \quad \text{or} \quad g(x) \geq e^{-\beta} g(y) \quad \forall y \in [x_2, u_2]$$

Let: $x_{\max} = \text{maximizer of } g \text{ on } [u_1, u_2]$

$x_{\min} = \text{minimizer of } g \text{ on } [x_1, x_2]$.

Conclusion:

$x_{\max} > x_{\min} \Rightarrow g(x) \geq e^{-\beta} g(y)$ for $x \in [x_1, x_2]$ and $y \in [u_1, x_1]$.

$x_{\max} < x_{\min} \Rightarrow g(x) \geq e^{-\beta} g(y)$ for $x \in [x_1, x_2]$ and $y \in [x_2, u_2]$.

The Reduction Step I

We assume $n > 1$ dimensions

Background: Löwner-John ellipsoid of K is the minimum volume ellipsoid containing K .

Key property: Löwner-John ellipsoid shrunk by n is contained in K .

A set is called ϵ -flat if its Löwner-John ellipsoid's smallest axis is smaller than ϵ .

We show that if the result fails for some decomposition it fails for a decomposition of an ϵ -flat set.

The Reduction Step II

Suppose theorem fails by $\delta > 0$ for some K :

$$(1 + \delta)\mu(B) \leq e^{-\beta \frac{d(K_1, K_2)}{\text{diam}(K)}} \min\{\mu(K_1), \mu(K_2)\} \quad \star$$

Then there exist $K' \subseteq K$, $K'_i \subseteq K$, and $B' \subseteq B$ such that:

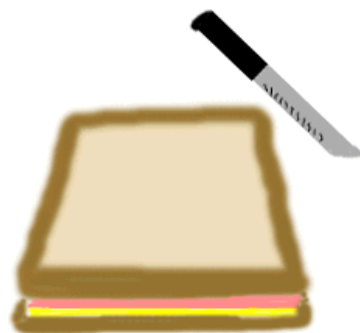
1. $d(K'_1, K'_2) \geq d(K_1, K_2)$,
2. $\text{diam}(K') \leq \text{diam}(K)$, and
3. \star holds for the prime sets.

Suppose we have K , K_1 , K_2 , B , and δ as above. We show how to construct an ϵ -flat decomposition.

The Reduction Step III

Recall the “Ham-Sandwich” Theorem: Given n non-degenerate finite Borel measures on \mathbb{R}^n , μ_1, \dots, μ_n there exists a hyperplane H such that:

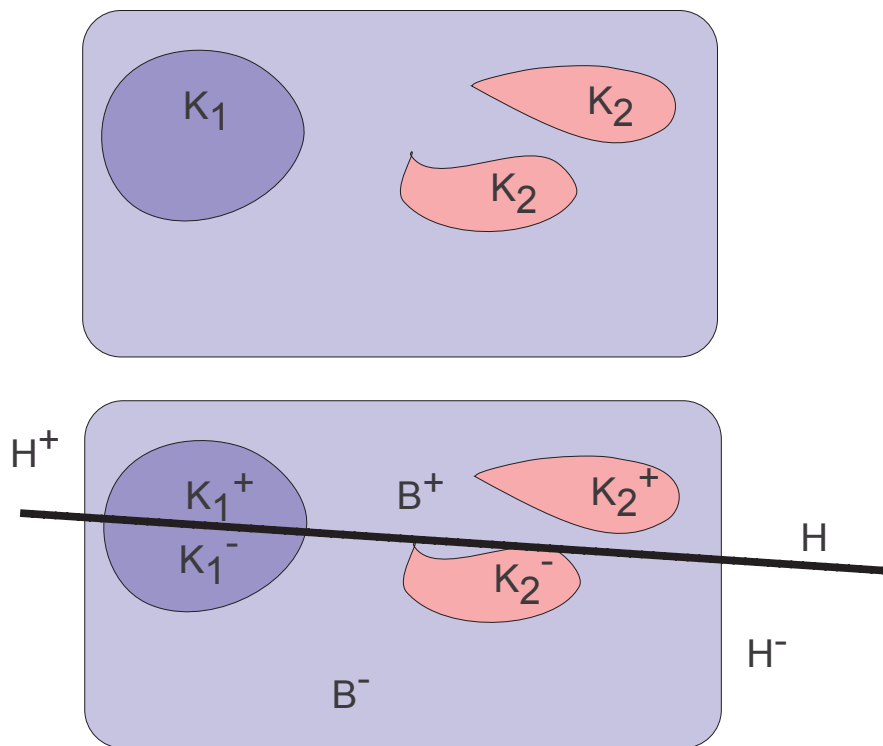
$$\mu_k(H^+) = \mu_k(H^-) = \mu_k(\mathbb{R}^n)/2 \quad k = 1, 2, \dots, n$$



We will assume $n \geq 2$ and consider the bisecting hyperplane satisfying the Ham-Sandwich Theorem for K_1 and K_2 .

The Reduction Step IV

The bisection step:



$\mu(K_i^+) = \mu(K_i^-)$, distance between sets \nearrow , diameter \searrow .

The Reduction Step V

Suppose theorem holds for both:

$$(1 + \delta)\mu(B^\pm) \geq e^{-\beta} \frac{d(K_1^\pm, K_2^\pm)}{\text{diam}(K^\pm)} \min\{\mu(K_1^\pm), \mu(K_2^\pm)\}$$

So we must have

$$\begin{aligned} (1 + \delta)\mu(B) &= (1 + \delta)(\mu(B^+) + \mu(B^-)) \\ &\geq e^{-\beta} \frac{d(K_1, K_2)}{\text{diam}(K)} \\ &\quad \left(\min\{\mu(K_1^+), \mu(K_2^+)\} + \min\{\mu(K_1^-), \mu(K_2^-)\} \right) \\ &= e^{-\beta} \frac{d(K_1, K_2)}{\text{diam}(K)} \min\{\mu(K_1), \mu(K_2)\} \end{aligned}$$

Contradicting \star .

The Reduction Step VI

Conclusion:

So we have that for at least one of the new partitions

$$(1 + \delta)\mu(\hat{B}) \leq e^{-\beta} \frac{d(K_1, K_2)}{\text{diam}(K)} \min\{\mu(\hat{K}_1), \mu(\hat{K}_2)\} \quad \hat{\star}$$

This is true for every $\delta > 0$.

The Reduction Step VII

Continue bisecting and always pick the partition with

$$(1 + \delta)\mu(\hat{B}^{(j)}) \leq e^{-\beta} \frac{t}{\text{diam}} \min\{\mu(\hat{K}_1^{(j)}), \mu(\hat{K}_2^{(j)})\} \quad \hat{\star}$$

where $K \supseteq K^{(1)} \supseteq K^{(2)} \dots \supseteq K^{(j)}$.

We claim that eventually the smallest axis of the Löwner-John ellipsoid of $K^{(j)}$ is smaller than ϵ .

If not, $\text{Ball}_{\epsilon/n} \subseteq K^{(j)}$. \Rightarrow

$\mu(K^{(j)}) > \eta > 0$. \Rightarrow

But $\mu(K_1^{(j)}) \rightarrow 0$ and $\mu(K_2^{(j)}) \rightarrow 0$. \Rightarrow

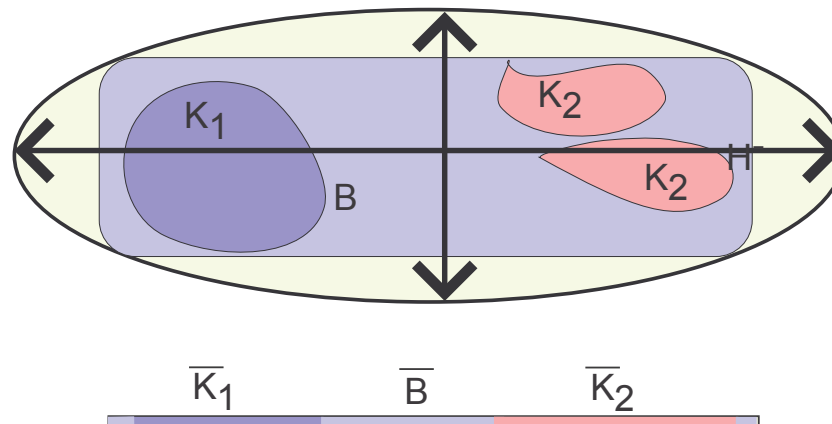
$\hat{\star}$ violated for $B^{(j)}$, $K_1^{(j)}$, and $K_2^{(j)}$ for j large.

The Induction Step I

Suppose we have $K = K_1 \cup B \cup K_2$ in \mathbb{R}^{n+1} .

The smallest axis of the Löwner-John ellipsoid of K is smaller than ϵ .

Project on the remaining axes. Obtain $\bar{K} = \bar{K}_1 \cup \bar{B} \cup \bar{K}_2$, and also \bar{g} and $\bar{\mu}$.



The Induction Step II

Projection maintains β -log concavity.

Choose ϵ small enough (smaller than $d(K_1, K_2)/2$).

Induction follows by continuity.

Part IV: Applications in Machine Learning

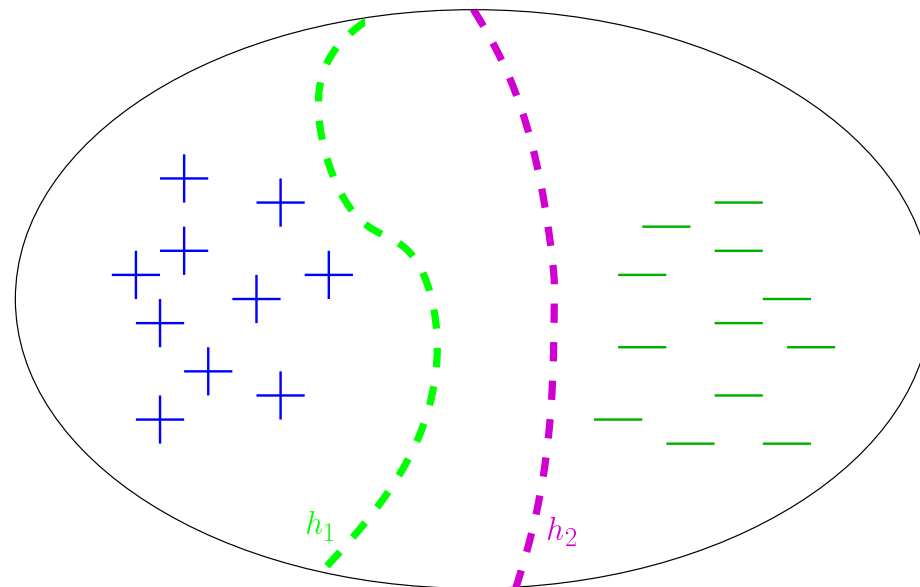
Mini-agenda

- a. A Lower Bound for Classification
- b. On the size of the margin
- c. Regression

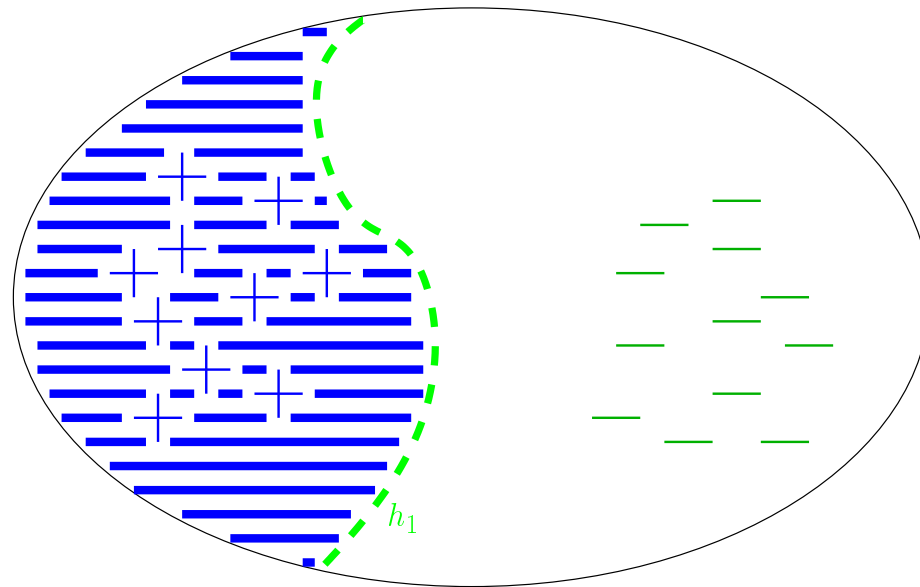
Part IV.a: Classification Error: Lower Bounds

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

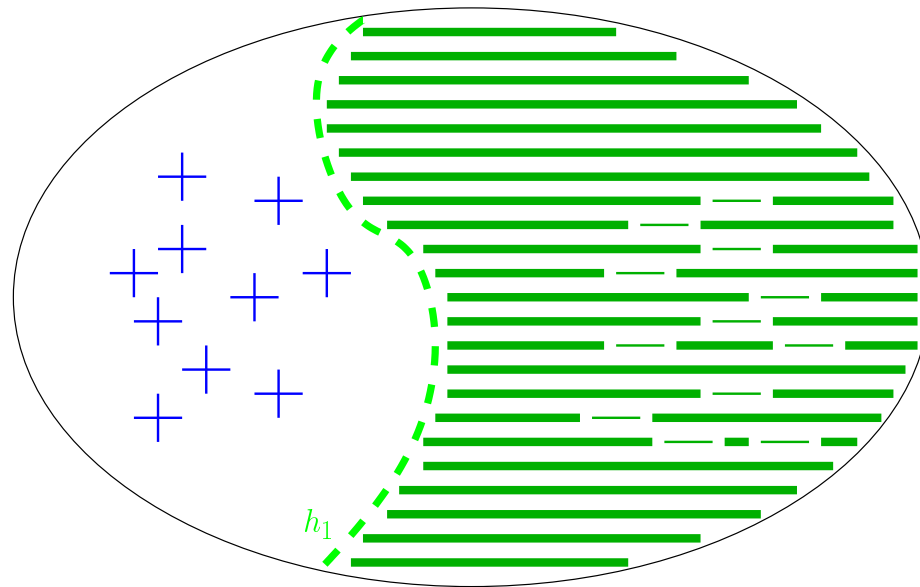
Measuring Error



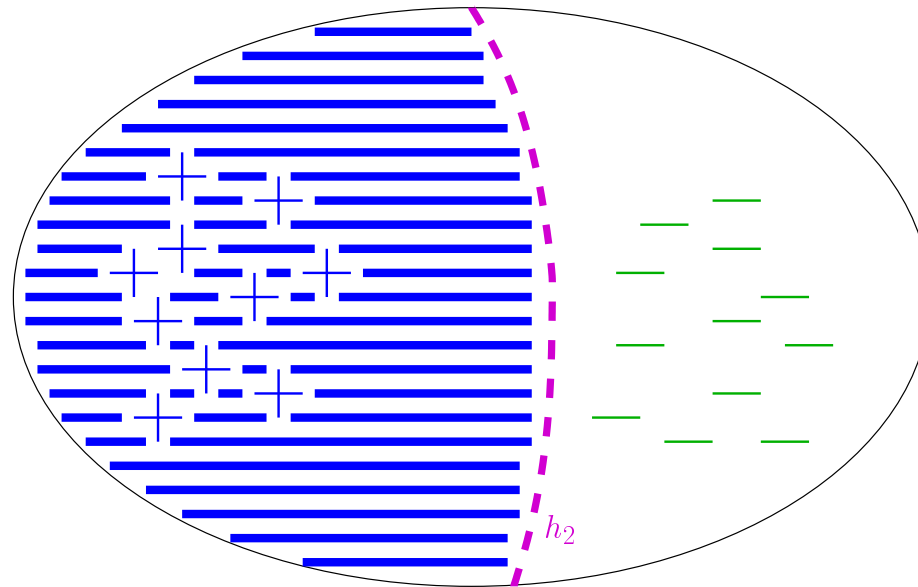
Measuring Error



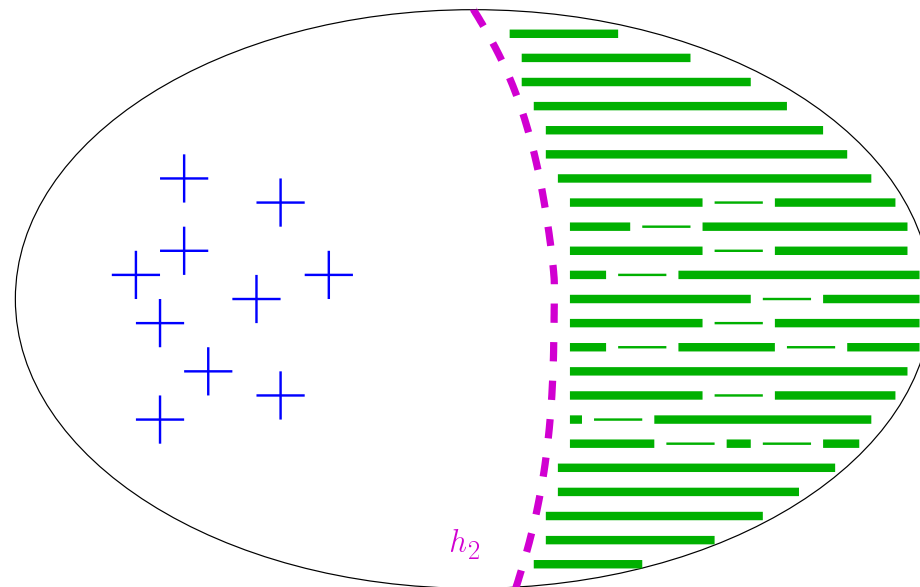
Measuring Error



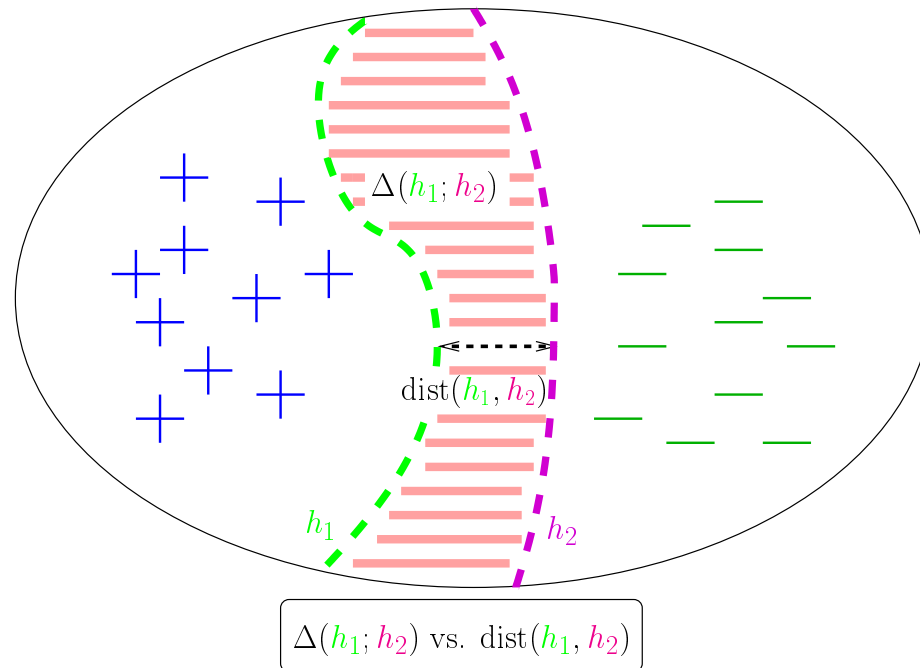
Measuring Error



Measuring Error



Measuring Error



Distance Measures

- Two distance measures: Given two classifiers, h_1, h_2 :
 - $\Delta(h_1; h_2) = \int_{h_1 \neq h_2} g(x) dx$ – measure of region $\{h_1 \neq h_2\}$.
 - $\text{dist}(h_1, h_2) = \inf_{x \in \partial h_1, z \in \partial h_2} \|x - z\|$ – separation distance.
- We cannot compute $\Delta(h_1; h_2)$ without knowledge of $g(x)$.
- 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' :
 - $\Delta(h_1; h_2) = \int_{h_1 \neq h_2} g(x) dx$ – measure of region $\{h_1 \neq h_2\}$.
 - $\text{dist}(h_1, h_2) = \inf_{x \in \partial h_1, z \in \partial h_2} \|x - z\|$ – separation distance.
- We cannot compute $\Delta(h_1; h_2)$ without knowledge of $g(x)$.
- 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 $\Delta(h_1; h_2)$.

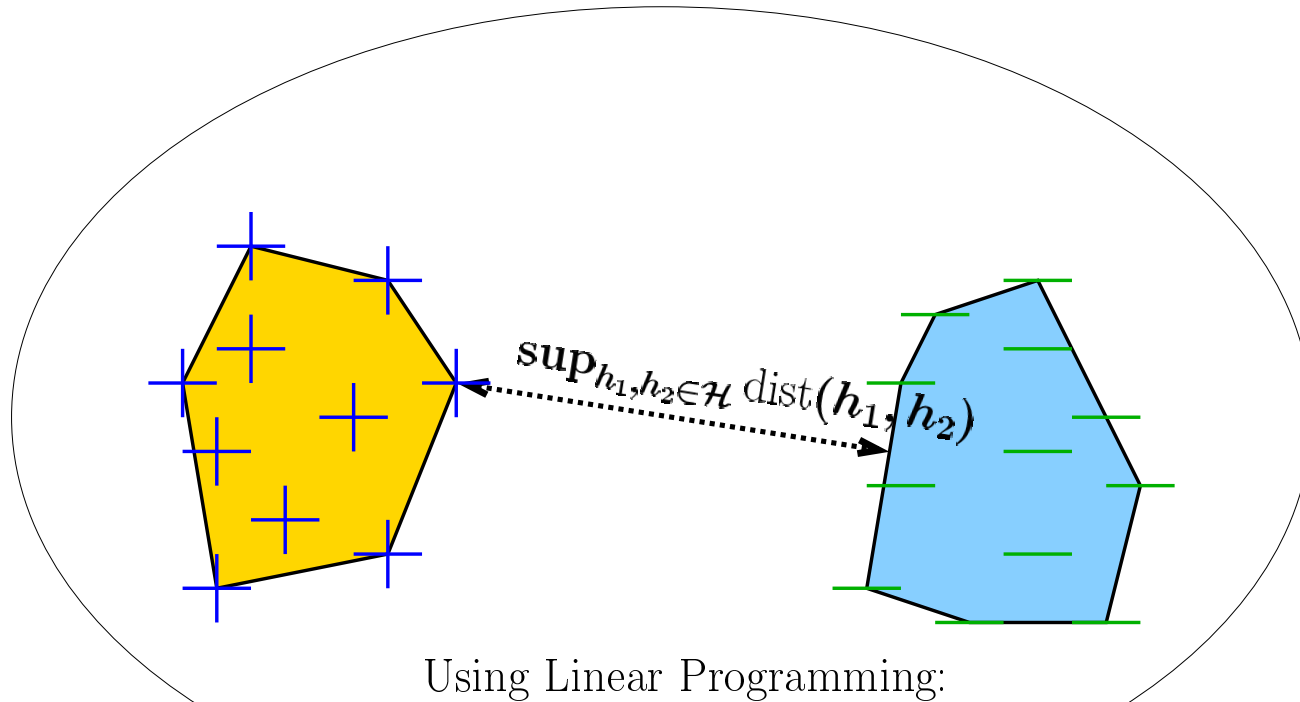
A First Theorem: A Lower Bound on Error

- If $g(x)$ is β -log-concave, with measure μ , and K is bounded,
- For every $h \in \mathcal{H}$ and $\epsilon > 0$ there exists $h' \in \mathcal{H}$ (where \mathcal{H} agree with h on the data) such that

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

- 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:

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

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

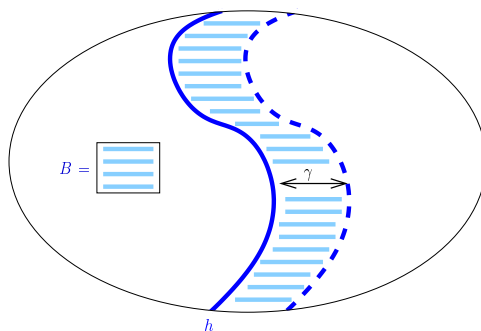
$$\mu(B) \geq e^{-\beta d(K_1, K_2)} \frac{1}{4\sqrt{2}\sigma} \min\{\mu(K_1)^{3/2}, \mu(K_2)^{3/2}\}.$$

Part IV.b : How Often Do We Get Lucky...?

- Suppose we sample from β -log-concave distribution (IID)
- 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 β -log-concave g :

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

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 g be a β -log-concave distribution on K with induced measure μ . 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.

What Is 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 β -log-concave, or
2. Dimensions are added artificially.

And What About the Margin?

If $h = \text{sign}(f(x))$ for a Lipschitz f , the margin between x_1, \dots, x_n and f is proportional to:

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

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

Can also bound $\Pr\{\sup_{f \in \mathcal{F}} \text{margin}(x_1, \dots, x_N; f) > \gamma\}$ 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.

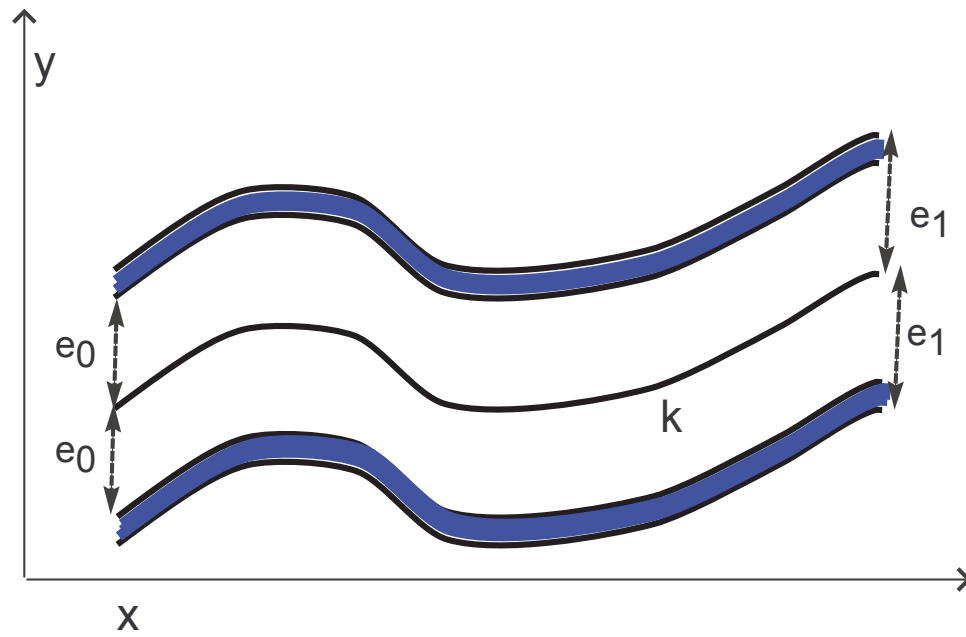
We let the tube be defined as:

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

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

Cost of converting ϵ_0 -sensitive error to ϵ_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 β -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 β -log-concave and k Lipschitz continuous:

$$\mu(T_{\epsilon_0, \epsilon_1}^k) \geq (\epsilon_1 - \epsilon_0) \frac{e^{-\beta}}{\sqrt{L^2 + 1} \text{diam}(K)} \min \left\{ \mu(T_{0, \epsilon_0}^k), \mu(T_{\epsilon_1, \text{diam}(K)}^k) \right\}.$$

The linear case:

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

if X is β_1 -log-concave and N is β_2 -log-concave we obtain the result with $\beta := \beta_1 + \beta_2$.

Other Goodies

Results hold for finite second moments as well.

Additional results hold under different assumptions on noise.

For example, if X is β -log-concave and $Y|X$ is β' -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 on the generating distribution leads to using distances instead of measures.
- Main Point: For β -log-concave distributions, good separation means the no-man's-land must carry a lot of weight.
- If not β -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) β -log-concave distribution:

A large margin is Bad News.

- Data are generated by a β -log-concave distribution:

A large margin/gap is unlikely.

Questions and Comments