MLE/MAP [in]variance
Support Vector Machines
Convex Functions

Today:- Late assignment 2
- Marked assignment 1 (end of class)

Wed:- Assigment 3 due
- Preliminary course evaluation.

"Standardize" Features

- When do MAP, we often 'standardize columns':

$$X = \begin{bmatrix} & \cdots & \end{bmatrix}$$

In regression we sometimes do this to 'Y', so we don't need y-intercept.

Subtract mean, divide by variance } mean 0, variance 1

roughly puts weights w_i on same scale

"Bias" vs. "bias"

- Sometimes, you'll see logistic regression written using $p(y_i|\bar{x}_i,\bar{w},b) = \frac{1}{1+\exp(-y_i(\bar{w}^T\bar{x}_i + b))}$

  instead of $p(y_i|\bar{x}_i,w) = \frac{1}{1+\exp(-y_i\bar{w}^T\bar{x}_i)}$

  "bias" variable

- "Bias" variable 'b' takes into account that one class may be more likely before we see the features

  (c.f., $p(y_i=1)$ in naive Bayes)

- Equivalent to adding a column of ones to X
- Statisticians say: "do not regularize the bias"

mean 0, variance 1

Linear classifiers

Generative

Naive Bayes
LDA

Discriminative

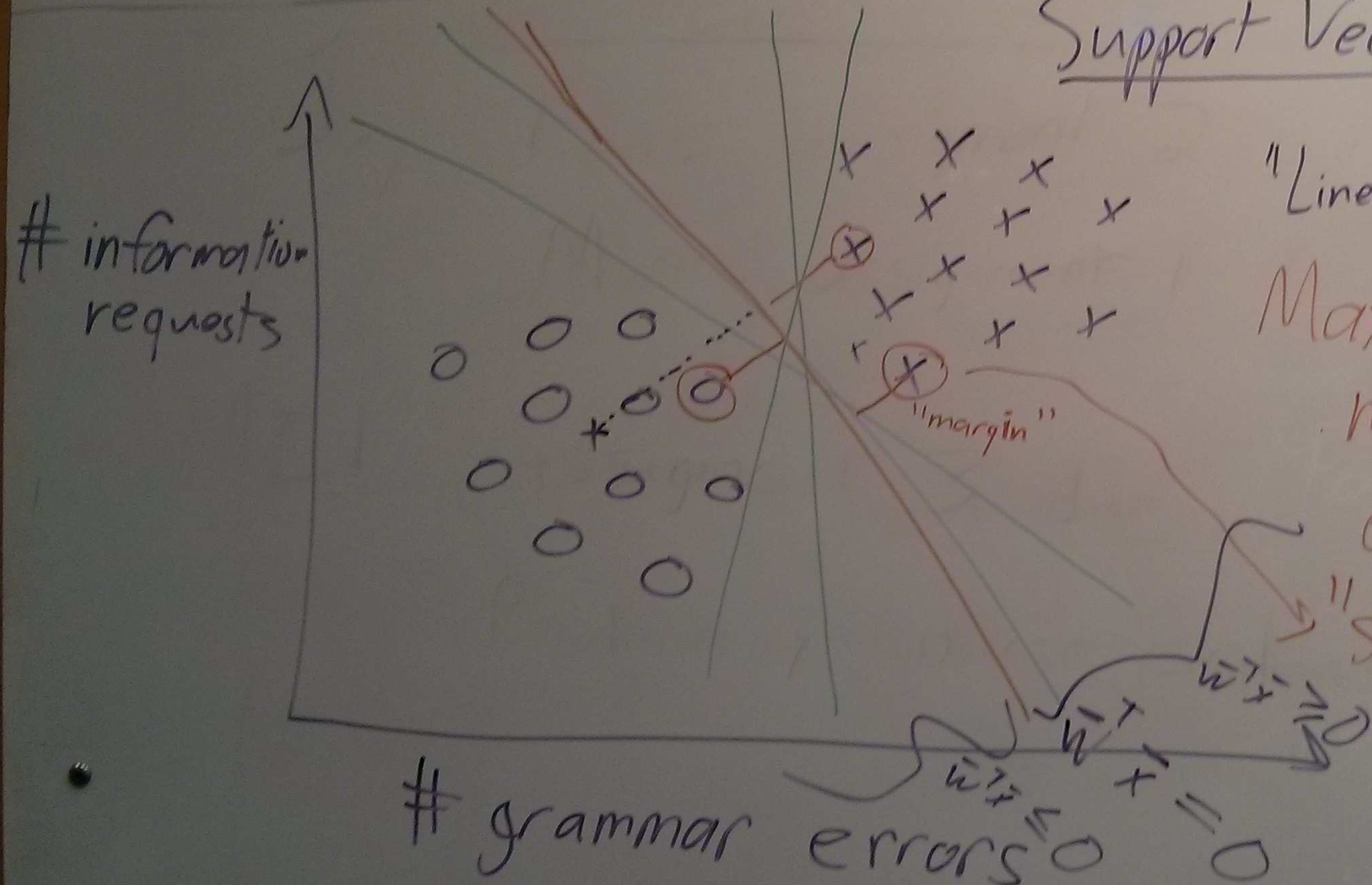
Logistic Regression

Discriminant Function

Support Vector Machines

MLE/MAP [in]variance
Support Vector Machines
Convex Functions

Support Vector Machines



"Linearly separable"

"Maximum "margin"":
maximize distance
of line to nearest points

"Support vector"

"margin" is proportional to $\frac{1}{||\bar{w}||}$

(see Wikipedia or Andrew Ng's course notes)

Maximizing $\frac{1}{||\bar{w}||}$ $\Longleftrightarrow$ minimizing $||\bar{w}||^2$

SVM:

$$\min_{w} \frac{1}{2}||\bar{w}||^2$$

$$\text{s.t. } y_i \bar{w}^T \bar{x}_i - 1 \geq 0$$

$=$ → "support vectors"

$\bar{w}^T \bar{x}_i \geq 1$, $y_i = 1$

$\bar{w}^T x_i \leq -1$, $y_i = -1$

oints

ξ

"Soft-margin" SVM

$$\underset{\bar{w}, \bar{v}}{\arg\min} \; C \sum_{i=1}^{N} v_i + \frac{1}{2} \|w\|^2$$

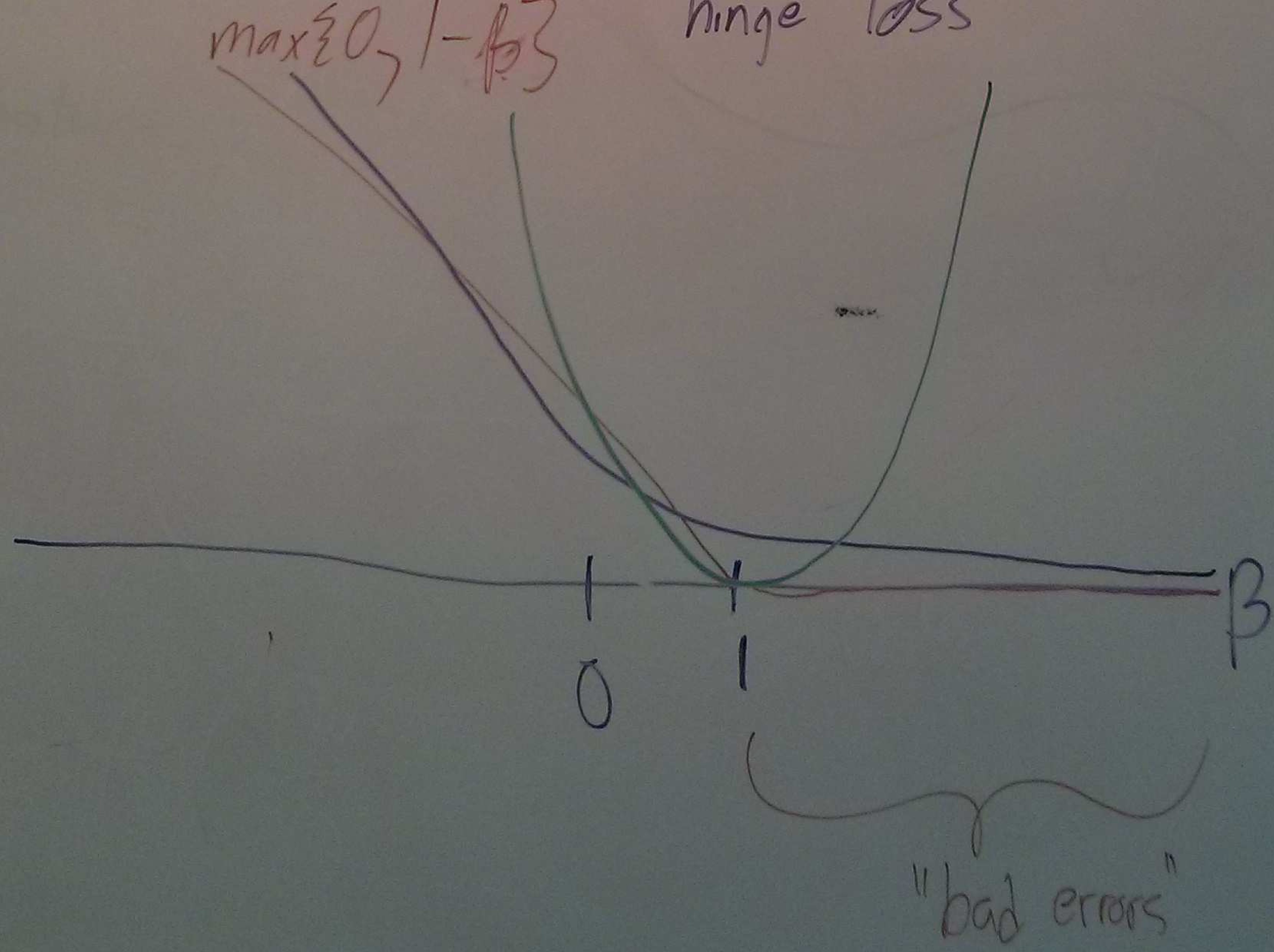
$$\text{s.t.} \quad y_i \bar{w}^T \bar{x}_i - 1 + v_i \geq 0,$$

$$v_i \geq 0$$

"for C large enough, linearly-separable is special case"

$$v_i \geq \max\{0, 1 - y_i \bar{w}^T \bar{x}_i\}$$

$$\underset{\bar{w}}{\arg\min}\; C \underbrace{\sum_{i=1}^{N} \max\{0, 1 - y_i \bar{w}^T \bar{x}_i\}}_{\text{"hinge" loss}} + \frac{1}{2}\|w\|^2$$



MLE/MAP [in]variance
Support Vector Machines
Convex Functions

Multi-Class SVM

$y_i \in \{1, 2, 3, \dots, K\}$

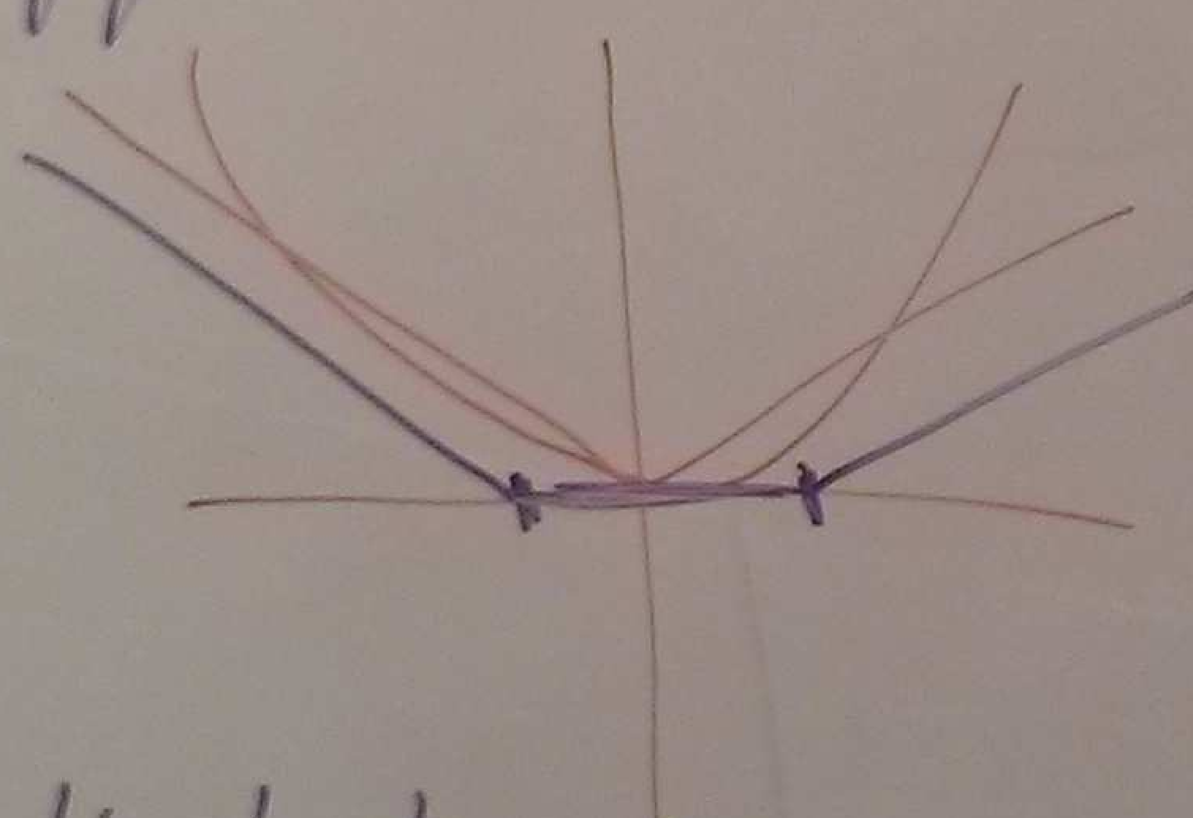
$\in \{\text{'red'}, \text{'green'}, \text{'blue'}\}$

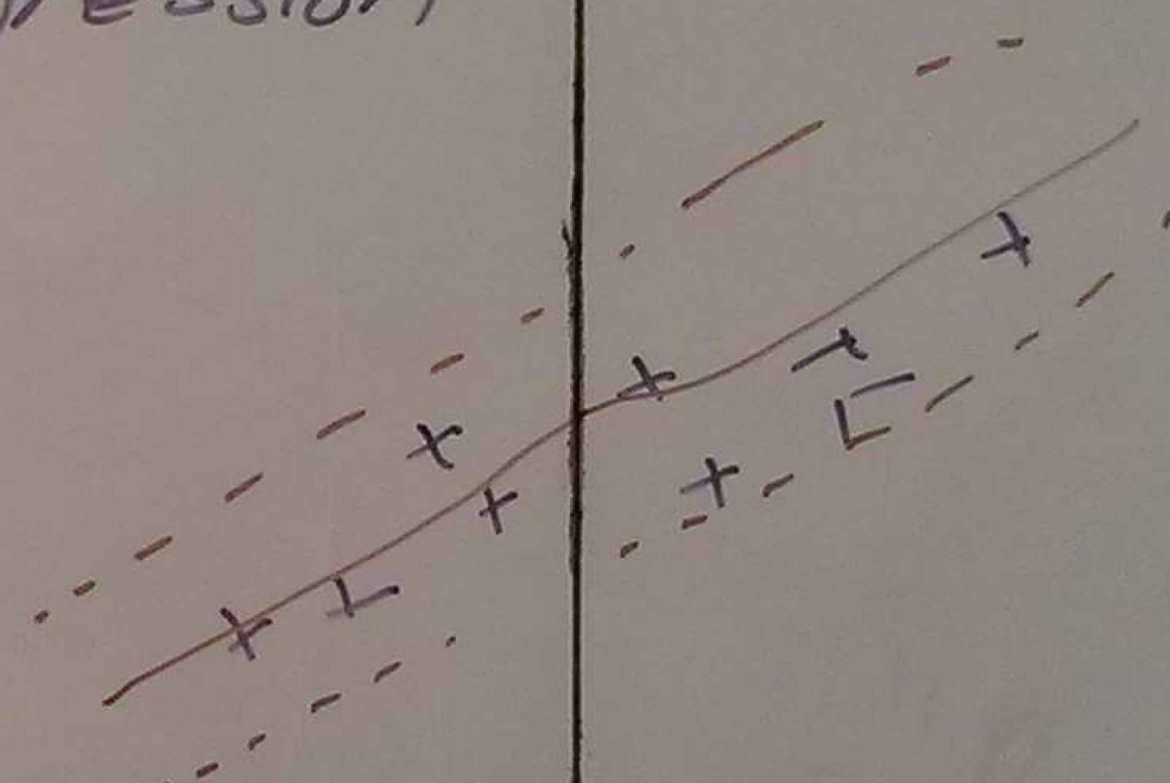
- 1 vs all: train C binary classifiers $O(dk)$
- All pairs: $\{1\}$ vs $\{2\}$
  $\{1\}$ vs $\{3\}$
  $\{2\}$ vs $\{3\}$
- ECOC
- Multi-Class: $\min_{\bar{w}_1, \bar{w}_2, \dots, \bar{w}_K} \frac{1}{2}\|\bar{w}\|^2 + C \sum_{i=1}^{N} \max_{c} \{1 - \bar{w}_{y_i}^T \bar{x}_i + w_c^T \bar{x}_i\}$

$\sum_{i=1}^{N} \sum_{c=1}^{K} \max\{0, 1 - \bar{w}_{y_i}^T \bar{x}_i + \bar{w}_c^T \bar{x}_i\}$

(*) Other types regression/classification

- Extreme-value regression
- Support vector regression





- Ordinal logistic, proportional hazards (movie ratings)
- Ranking $(a \geq b)$
- Multiple regression
- Multi-task classification

"structured prediction"

Convex Functions

Why?

- polynomial-time algorithms.
- all stationary points are global optima.

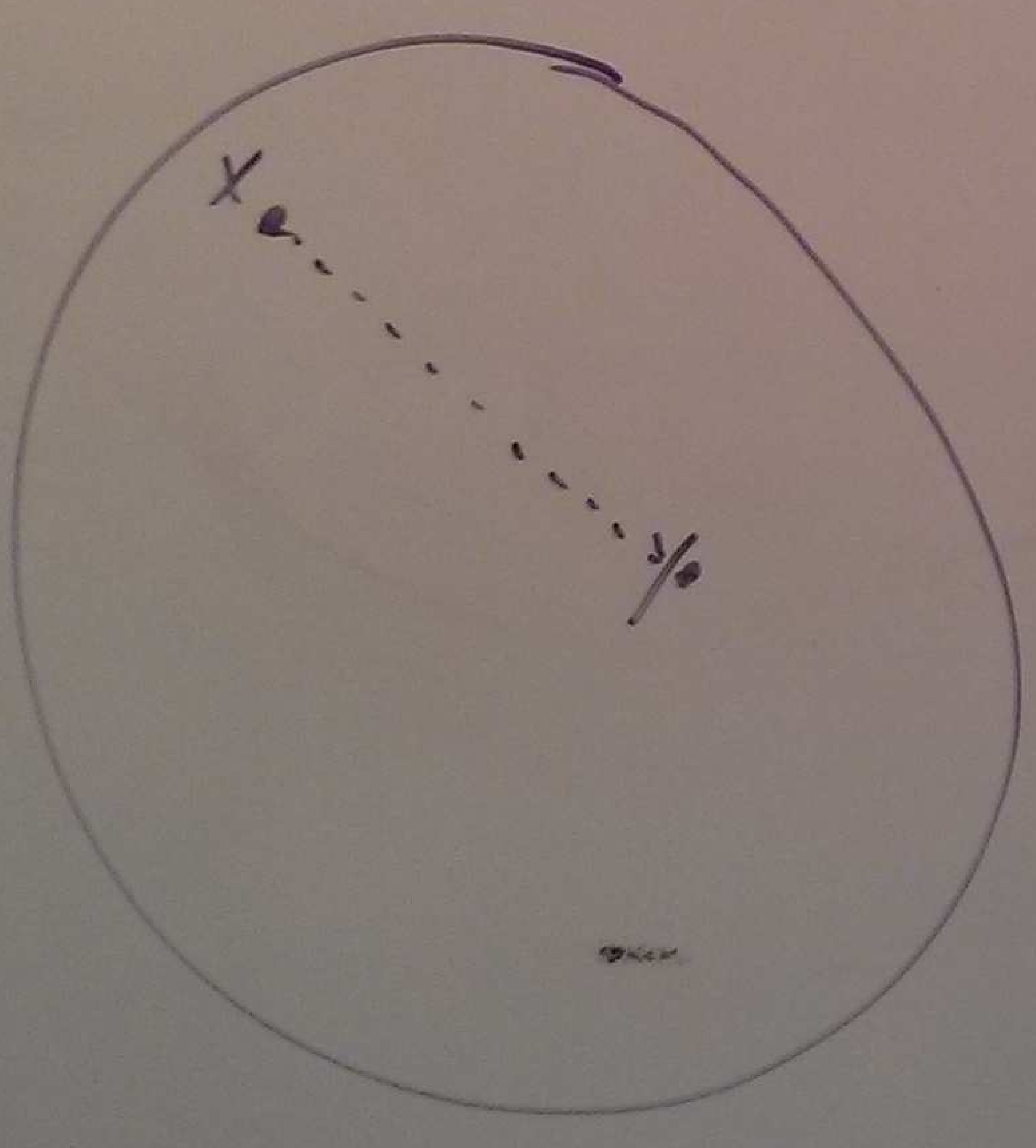
(least squares, L_2-reg, L_1-reg, Logistic, SVMs)

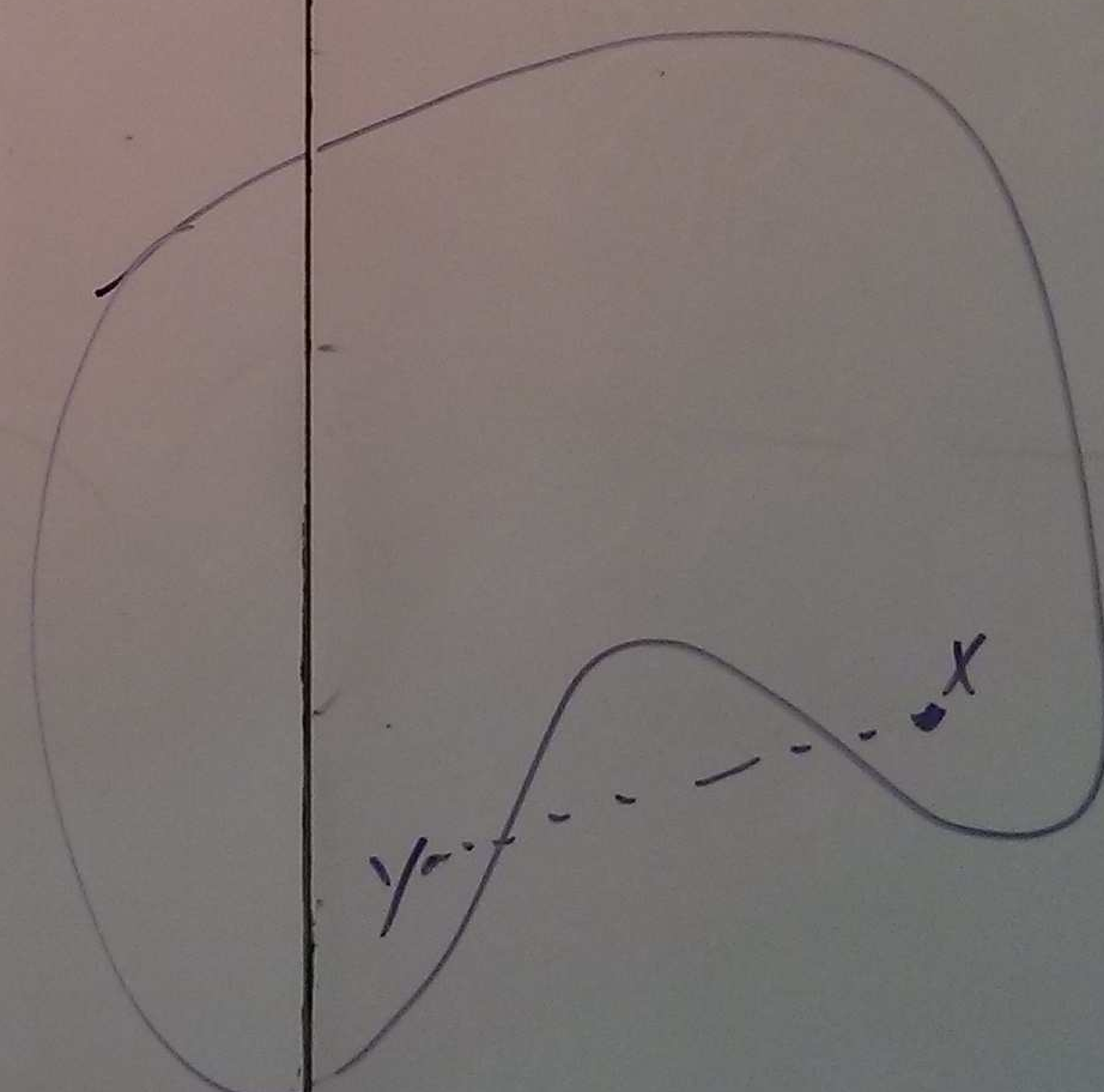
Convex Set

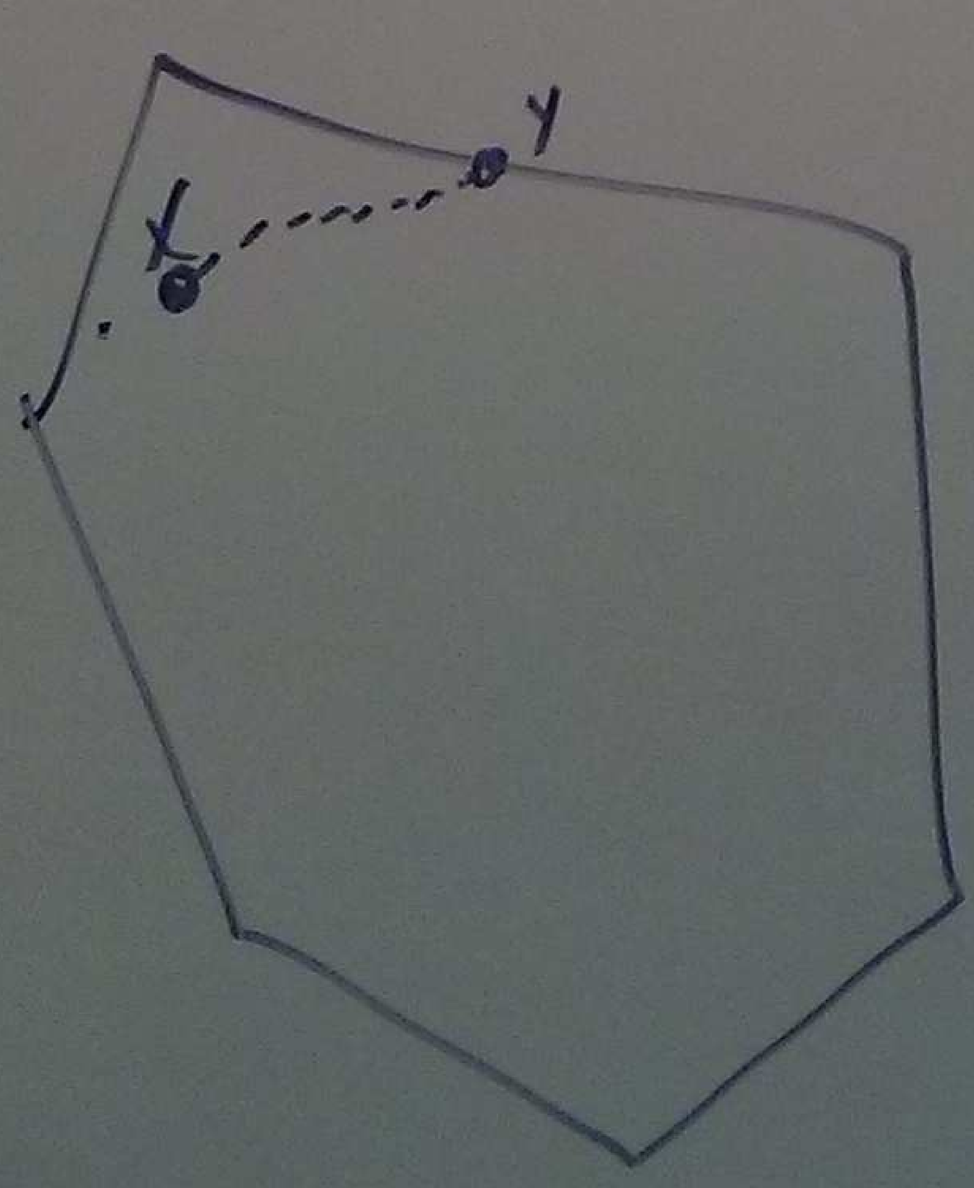
A set 'C' is convex if

$$\theta(x) + (1-\theta)y \in C,$$

$$\forall x, y \in C, \quad 0 \leq \theta \leq 1.$$



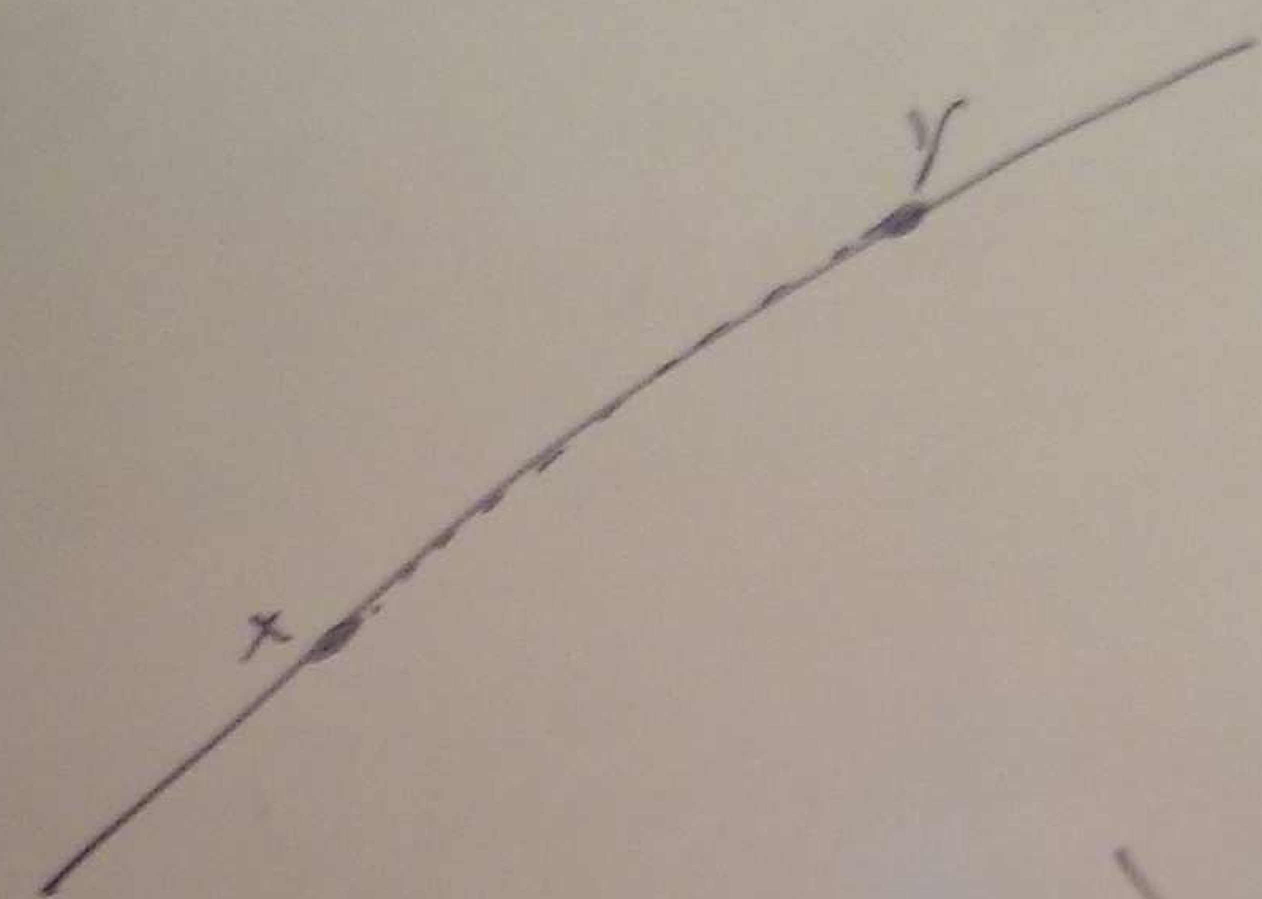


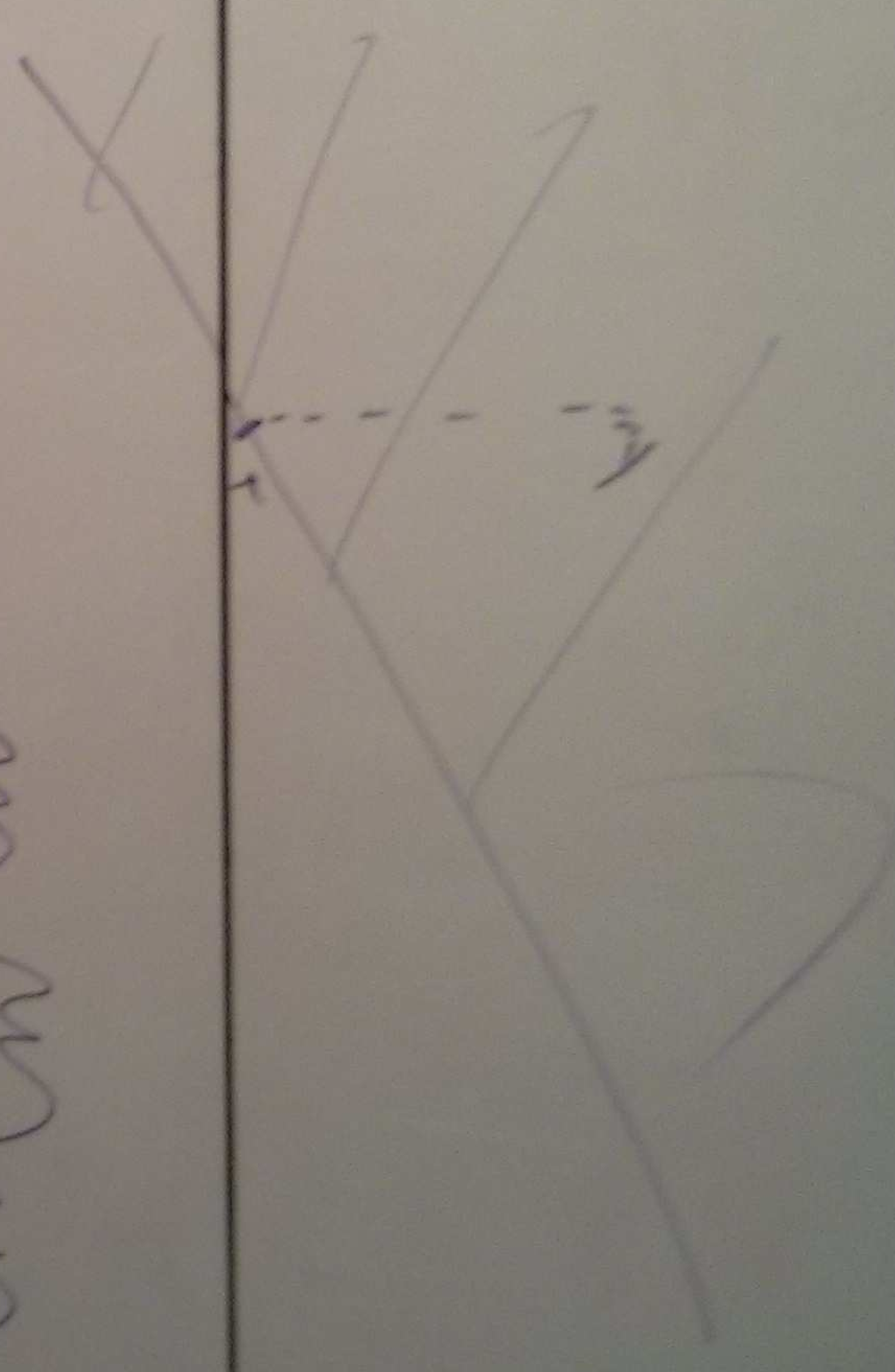


AP [in]variance

Vector Machines

Functions





Linear equality: $\{x \mid a^T x = b\}$

inequality: $\{x \mid a^T x \leq b\}$

norm-"ball" $\{x \mid \|x\| \leq r\}$

-"cone" $\{(x, r) \mid \|x\| \leq r\}$

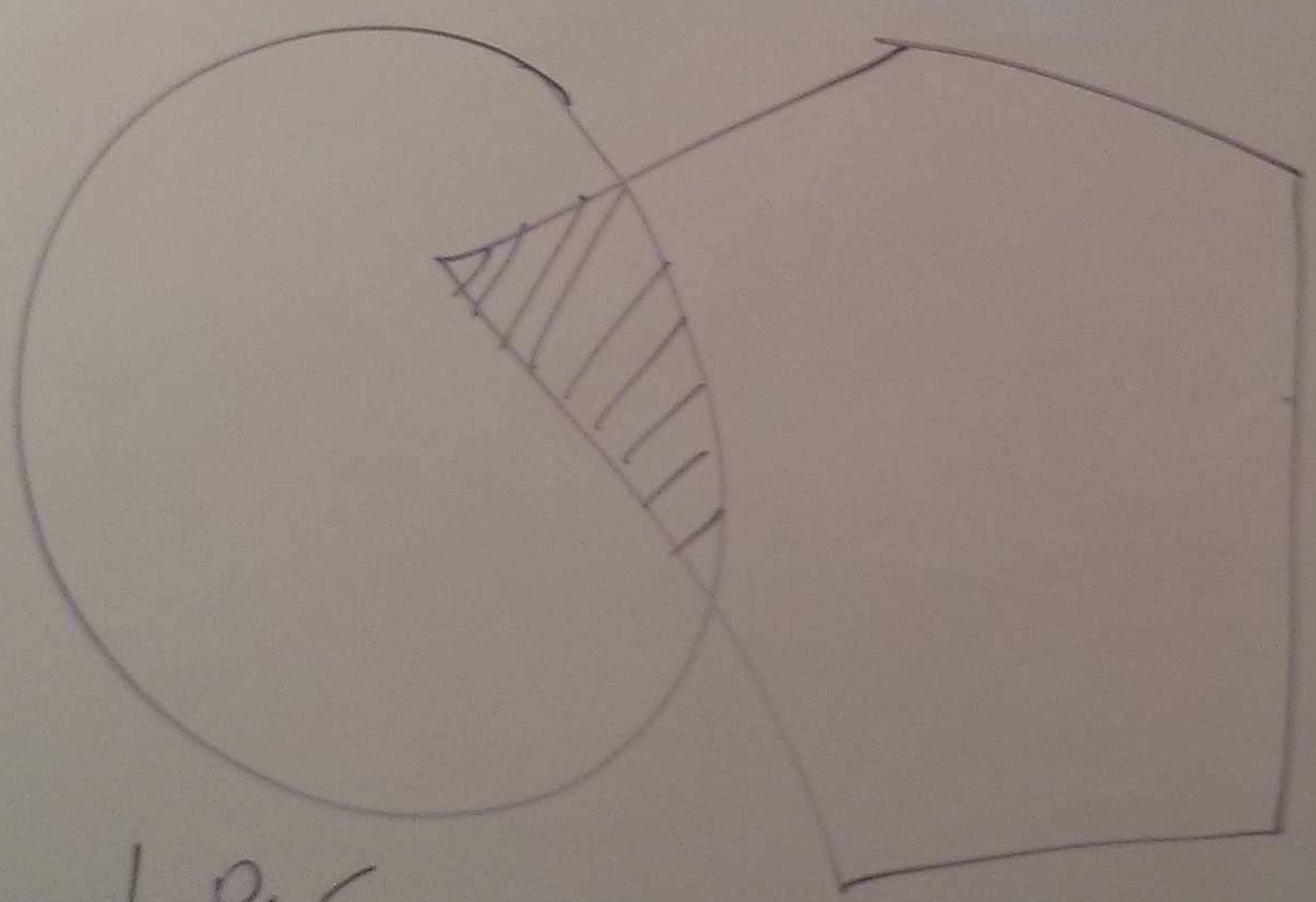
spectrahedron $\{\Sigma \mid \Sigma \succeq 0\}$

$\{\Sigma \mid \Sigma \succ 0\}$

$\{x \mid f(x) \leq 0\}$

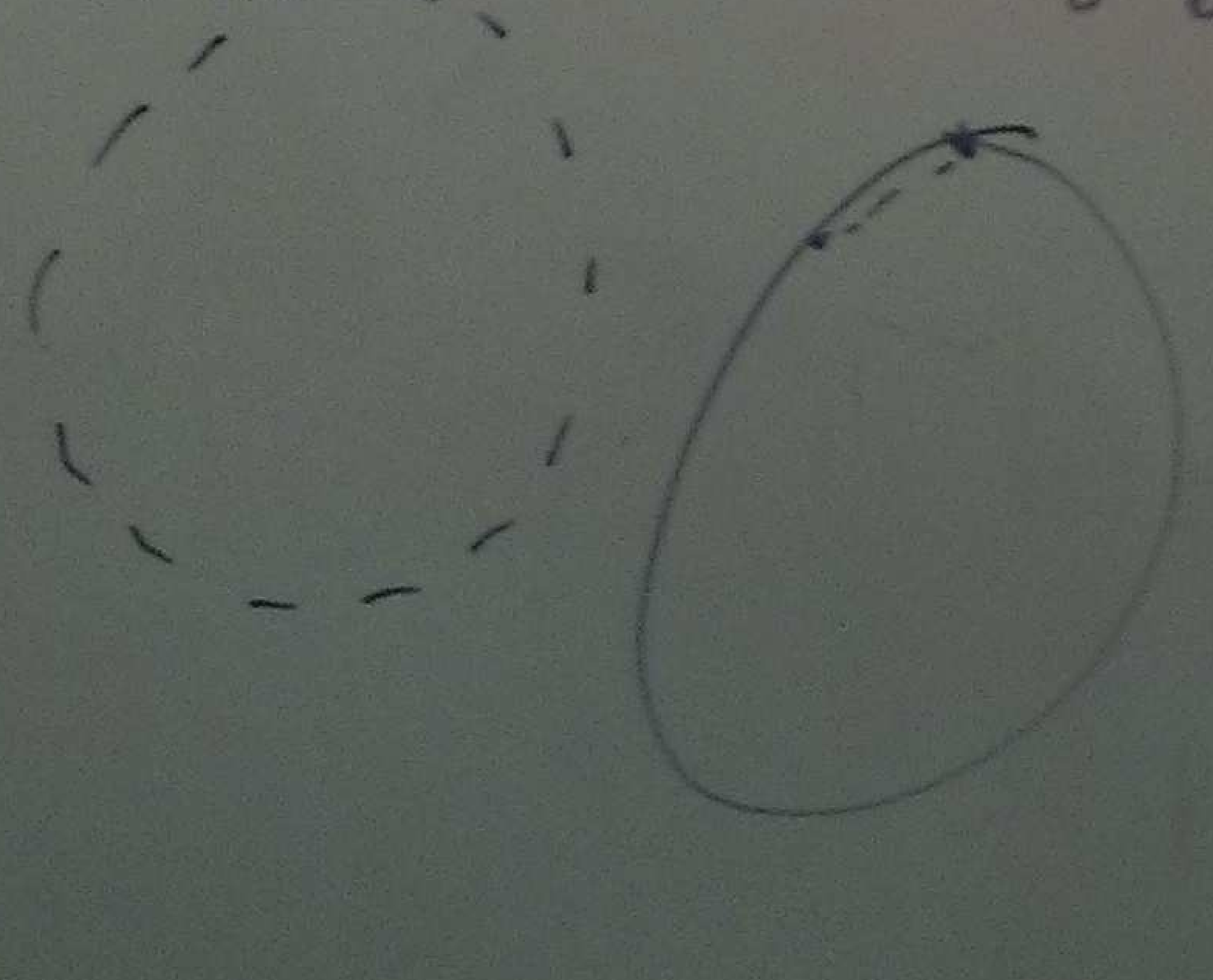
for any convex 'f'.

Intersection of convex sets is convex



LP: $\{x \mid Ax \le b,\ A_{eq} = b_{eq},\ LB \le x \le UB\}$

"Closed" convex set
if it includes boundary

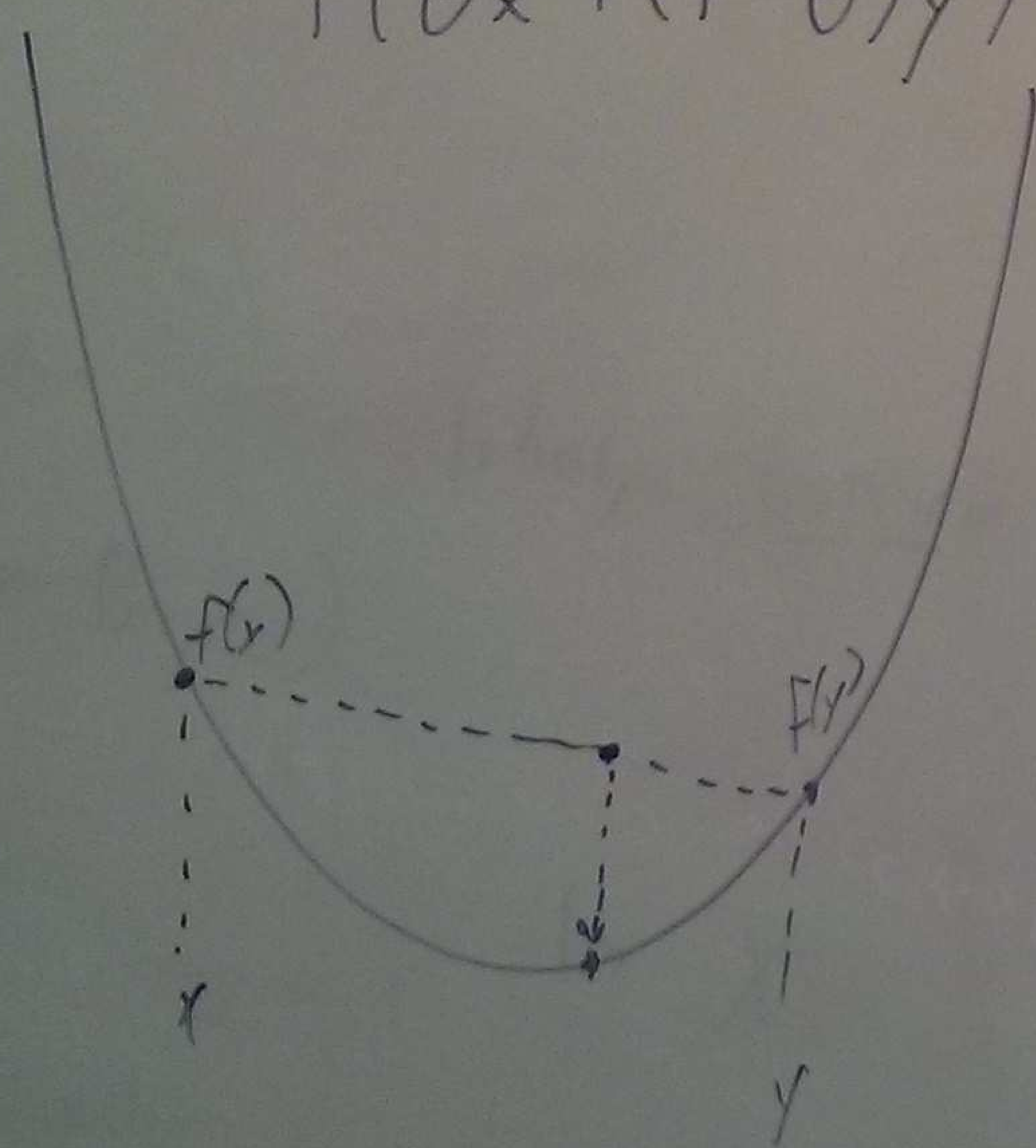


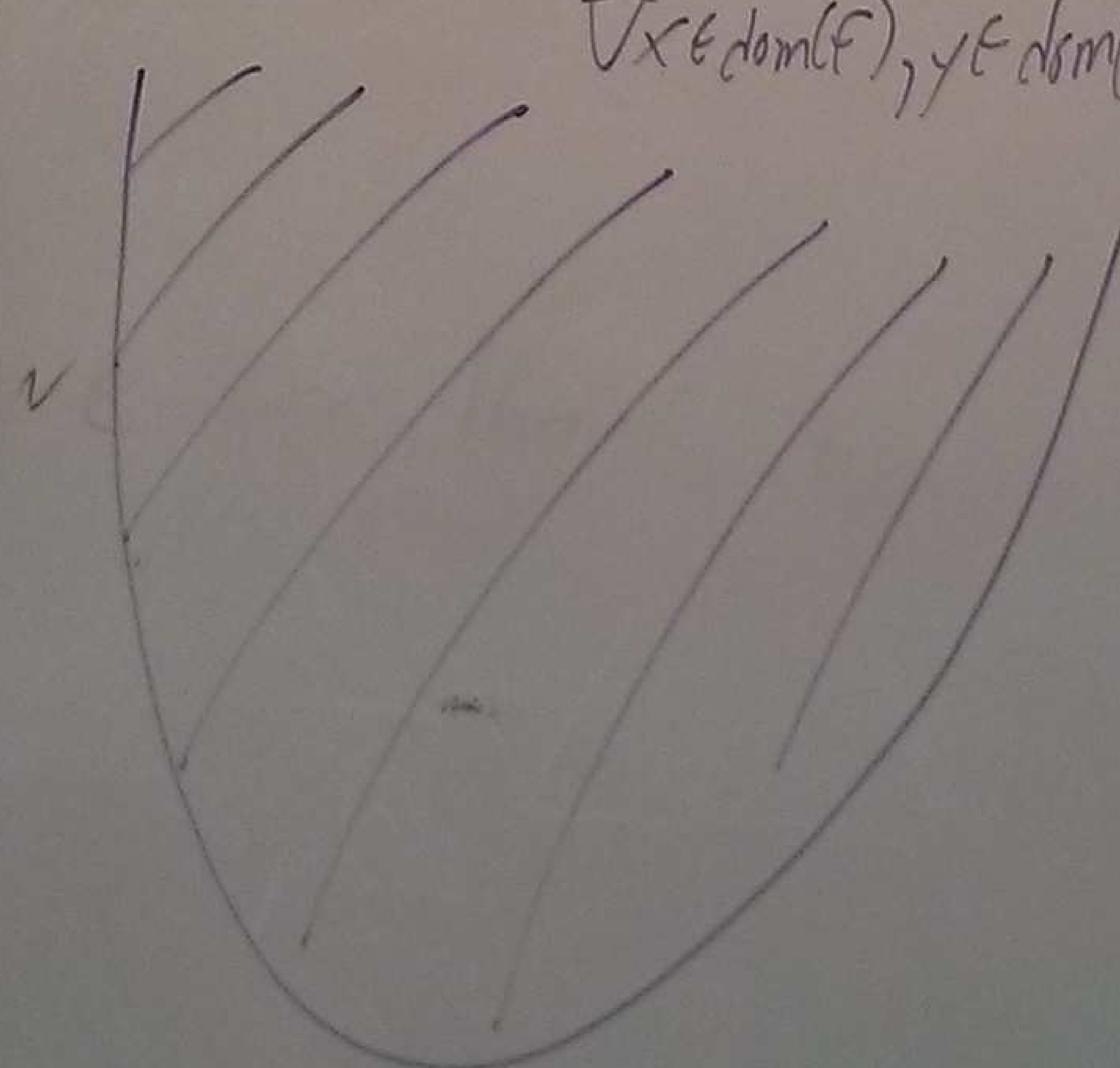
Convex Functions

A function 'f' is convex if dom(f) is a convex set, and

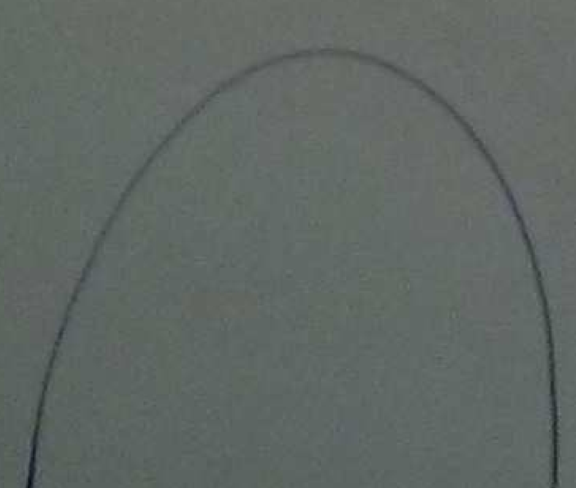
$$f(\theta x + (1-\theta) y) \le \theta f(x) + (1-\theta) f(y),$$

$\forall x \in dom(f), y \in dom(f), 0 \le \theta \le 1$



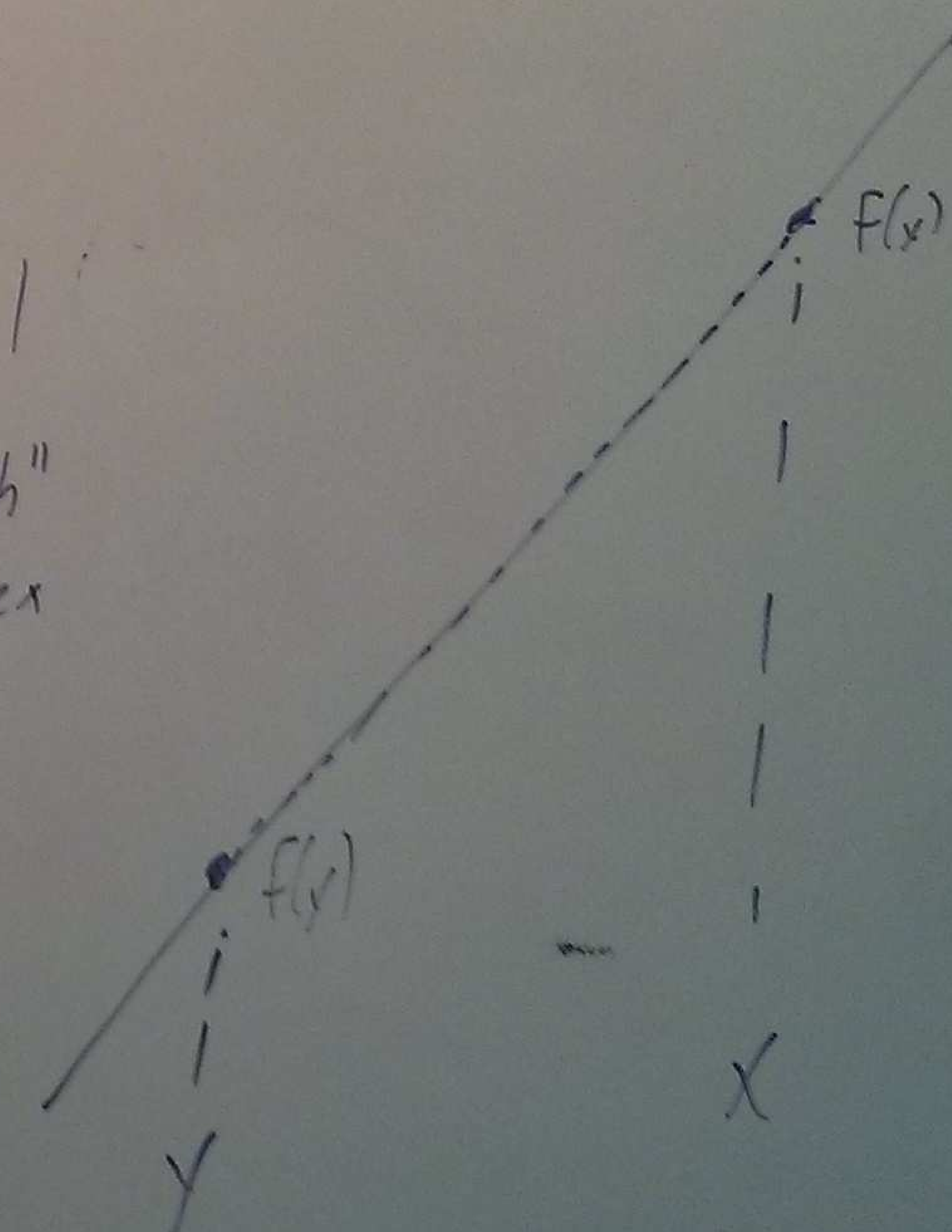


"epigraph" is convex set



"concave"

Examples



f(x)