Logistic Regression
Support Vector Machines

- Pick up assignment 1 at end of class
- Assignment 2 due now
- Assignment 3 out (due Wednesday)
- Tutorials are now 3-430pm Thursdays (FO
  330-430pm Fridays (IC
- Comments on lecture style (context, course eval ne

(*) - material

chol

group - L, re

Why do we need a valid kernel?
($K \succeq 0$)

1. Interpretation
   - valid kernels correspond to some feature space
     $(k(x,x') = \phi(x)^T \phi(x'))$
   - for Gaussian, valid kernels ensure likelihood is normalized

2. Convexity (Monday)
   - can't solve efficiently

But, non-valid kernels may still perform well

Big picture of supervised learning

	Generative model	Discriminative model	Discriminant function
MLE	$p(y_i, \bar{x}_i \mid \bar{w})$	$p(y_i \mid \bar{x}_i, \bar{w})$	$f(\bar{x}_i, \bar{w}) \to y_i$
MAP	$p(y_i, \bar{x}_i \mid \bar{w})\, p(\bar{w})$	$p(y_i \mid \bar{x}_i, \bar{w})\, p(\bar{w})$	$f(\bar{x}_i, \bar{w}) \to y_i$ "but with regularizer on $\bar{w}$"

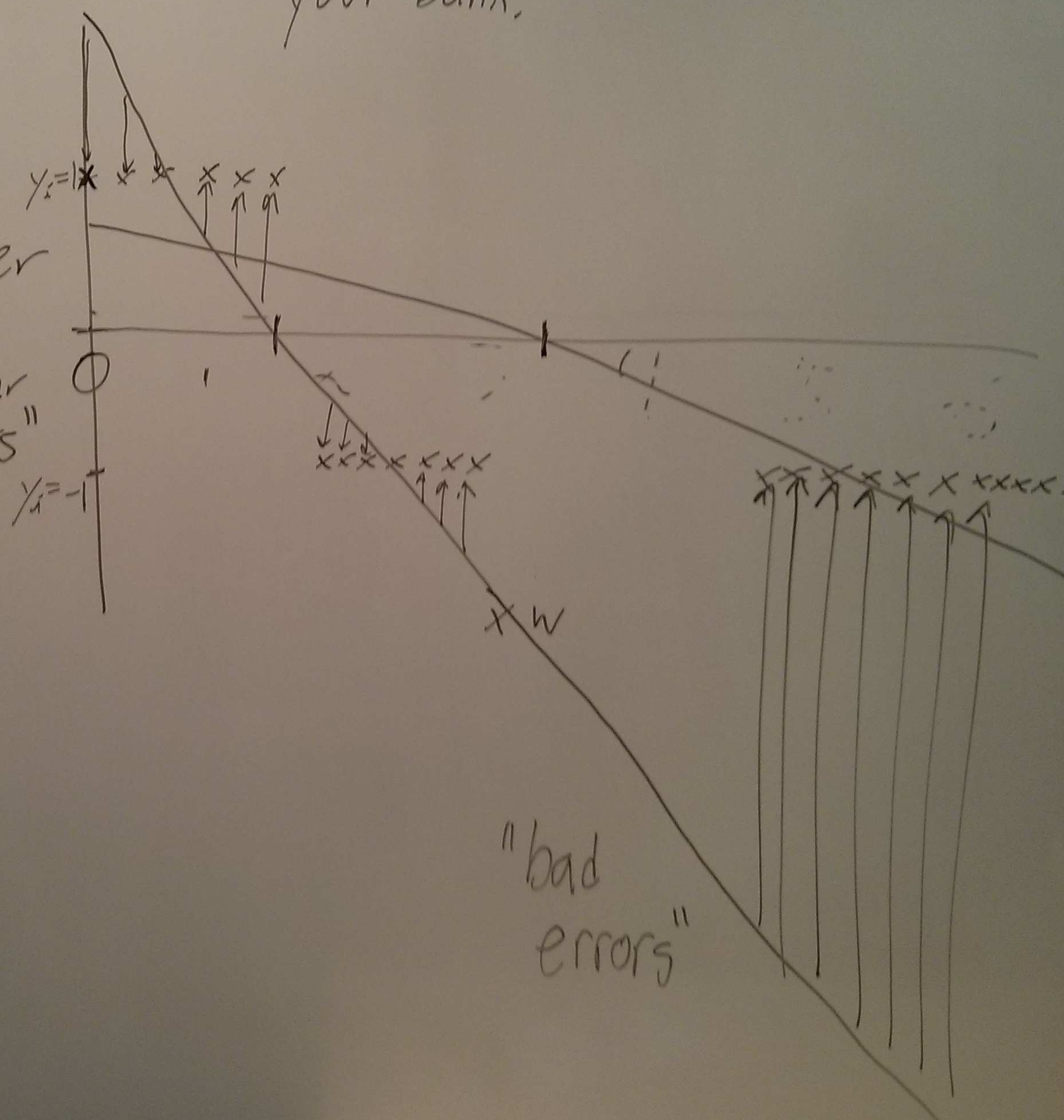
logistic regression

Support vector machines

Today: <u>linear</u> discriminative models, discriminant functions

Test time: You get an e-mail from your bank.

Training time: "number of grammar errors"



on $\bar{w}$"

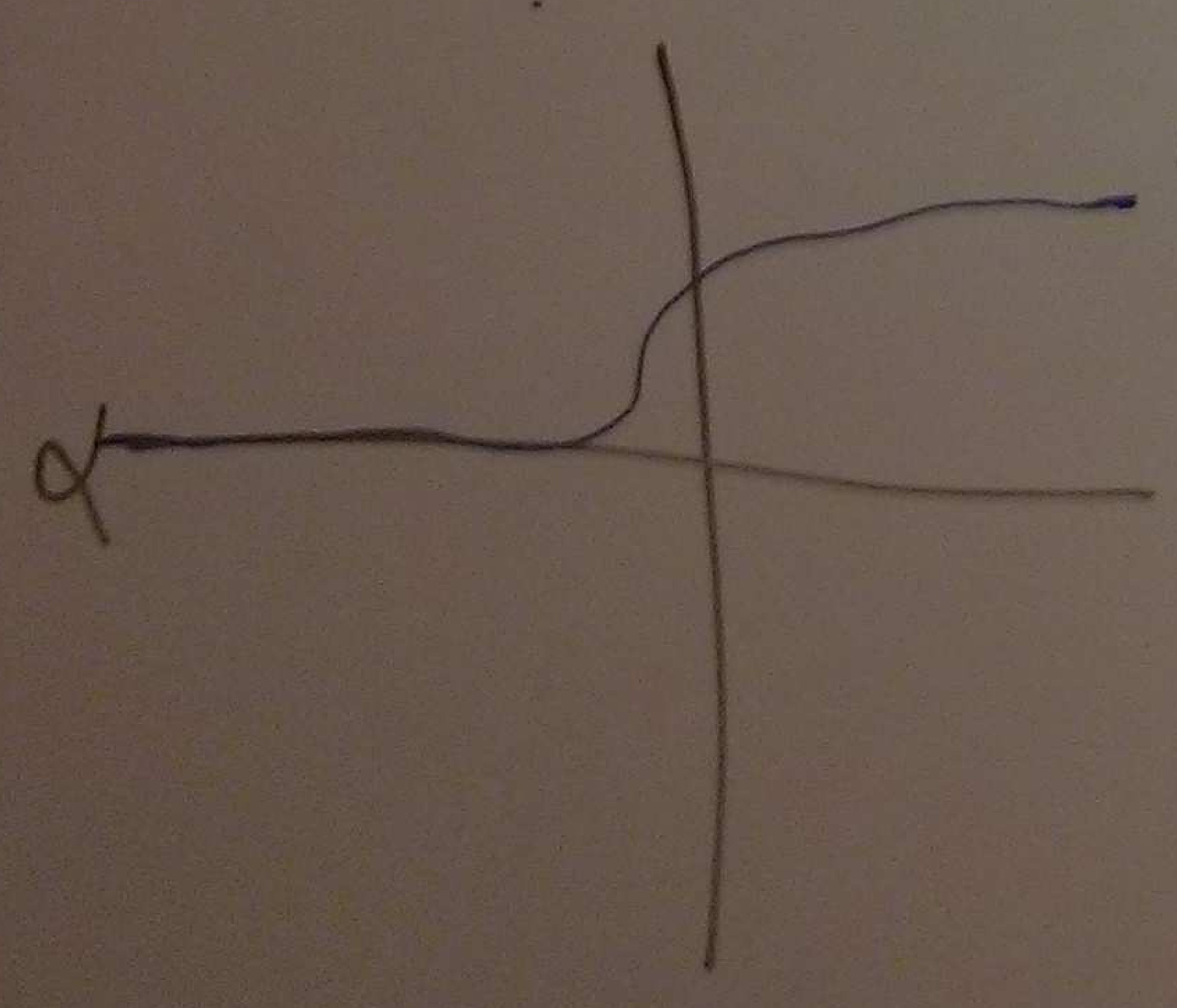
...riminative models,
...riminant functions.

Logistic Regression
Support Vector Machines

Logistic Regression

(parametric, linear, better handling of "bad errors"

$$p(y_i = 1 \mid \bar{x}_i, w) = \frac{1}{1 + \exp(\underbrace{-\bar{w}^T x_i}_{\alpha})}$$



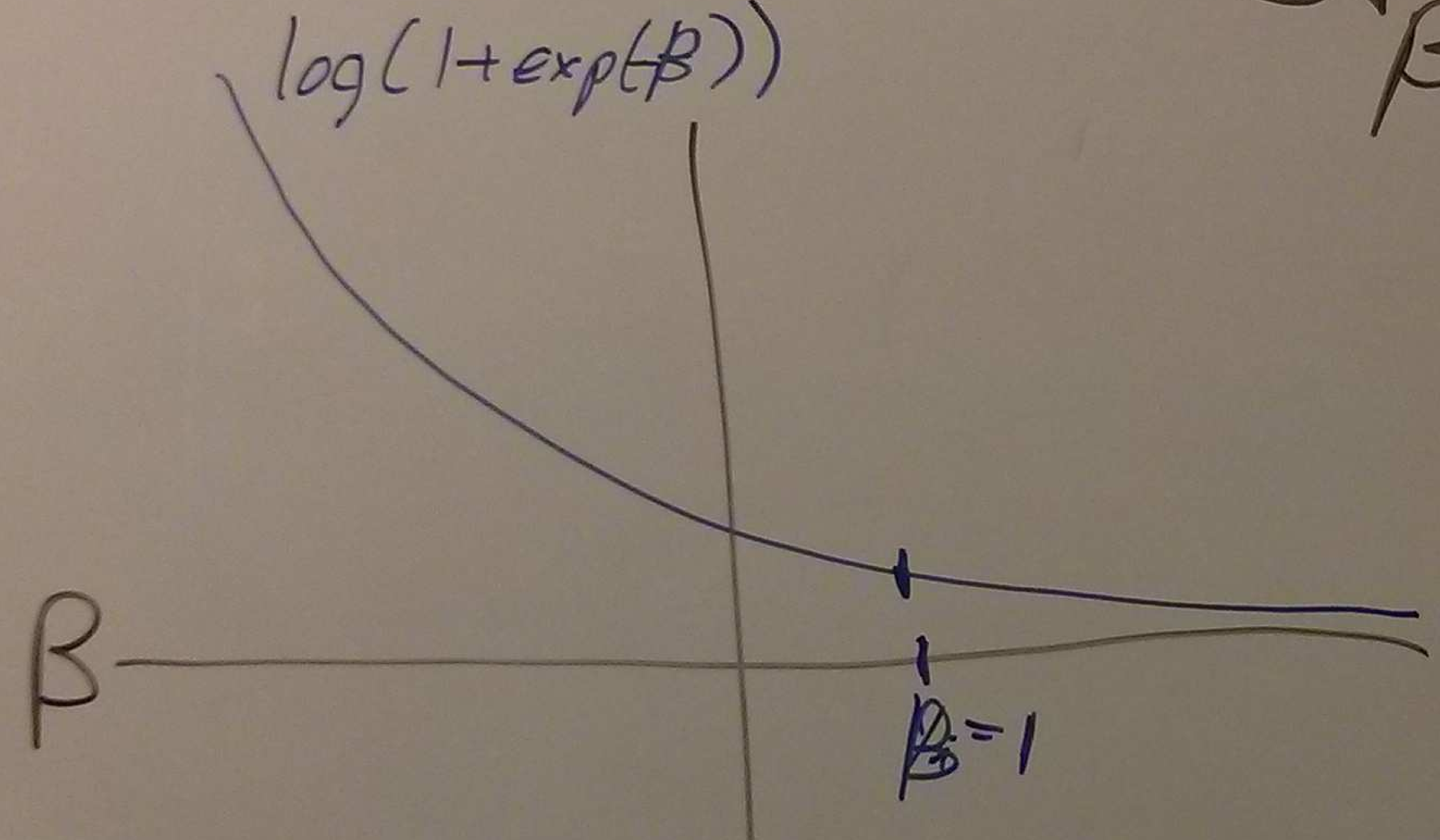
$$p(y_i = -1 \mid \bar{x}_i, \bar{w}) = \frac{1}{1 + \exp(+\bar{w}^T x_i)}$$

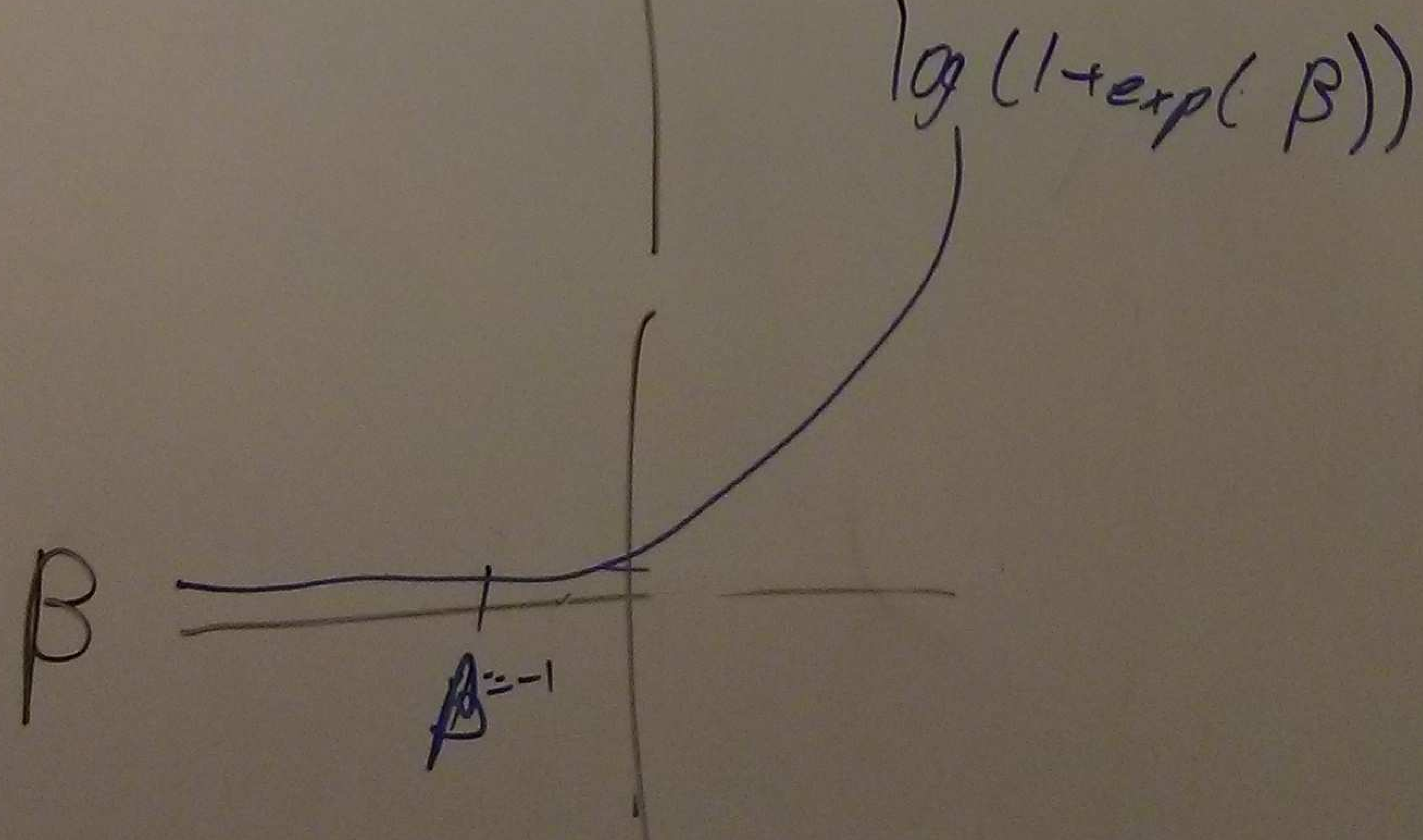
$\sum_{i=1}^{N}$

Gaussian

$$\frac{1}{2}(\bar{w}^T\bar{x}_i - y_i)^2$$

$$\sum_{i=1}^{N} -\log p(y_i \mid \bar{x}_i, \bar{w}) = \sum_{i=1}^{N} \log\left(1 + \exp(\underbrace{-y_i \bar{w}^T \bar{x}_i}_{\beta})\right)$$





MAP logistic

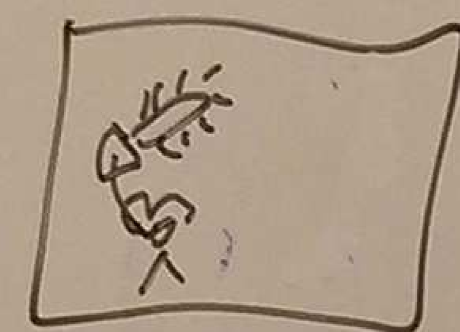
$$\overbrace{\sum \log(1+\exp(-y_i \bar{x}^T \bar{w}))}^{\text{log-likelihood}} + \overbrace{\frac{\lambda}{2}\|\bar{w}\|^2}^{\text{log-prior}}$$

$$+ \lambda\|w\|_1$$

(*) Can kernelize

Multinomial logistic regression

$\bar{x}_i \in \mathbb{R}^d$



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

"Tom", "Jerry", "Itchy", "Scratchy"

(*) Ordinal logistic regression

We'll have $\bar{w}_c$ for each class 'c'.

$$p(y_i = c \mid \bar{x}_i, \bar{w}) = \frac{\exp(\bar{w}_c^T \bar{x}_i)}{\sum_{c'=1}^{K} \exp(\bar{w}_{c'}^T x_i)}$$

"softmax"

"over-parameterized"

$$\sum_{i=1}^{K} p(y_i = c \mid \bar{x}_i, \bar{w}) = 1$$

MLE: set $\bar{w}_c = 0$ for 1 class 'c'.

$$\sum_{i=1}^{N} -\log p(y_i \mid \bar{w}, \bar{x}_i) = -\log\left(\frac{\exp(\bar{w}_{y_i}^T \bar{x}_i)}{\sum_{c'=1}^{K} \exp(\bar{w}_{c'}^T x_i)}\right)$$

$$= \sum_{i=1}^{N} -\bar{w}_{y_i}^T \bar{x}_i + \log\left(\sum_{c'=1}^{K} \exp(\bar{w}_{c'}^T x_i)\right)$$

$$\sum_{i=1}^{N} -I(c = y_i)\bar{x}_i + \underbrace{\frac{\exp(\bar{w}_c^T x_i)}{\sum_{c'=1}^{K} \exp(\bar{w}_{c'}^T \bar{x}_i)}}_{p(y_i = c \mid \bar{x}_i, \bar{w})} x_i$$

Log-sum-exp trick

"overflow"

$$\log\left(\sum_{i=1}^{N} \exp(a_i)\right)$$

$$= \log\left(\sum_{i=1}^{N} \underbrace{\exp(a_i - A + A)}_{\exp(a_i - A)\exp(A)}\right), \quad A$$

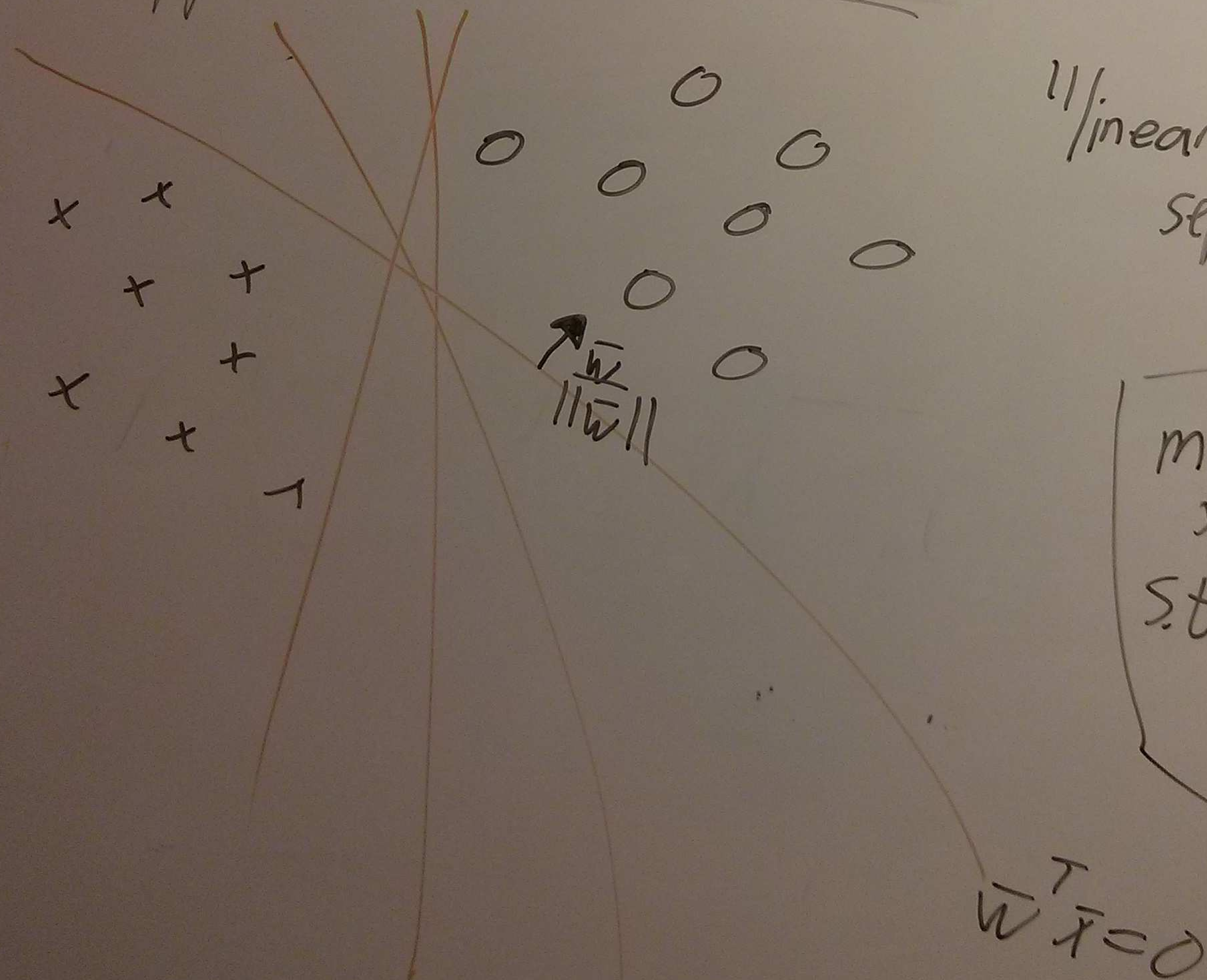
$$= \log\left(\exp(A) \sum_{i=1}^{N} \exp(a_i - A)\right)$$

$$= A + \log\left(\sum_{i=1}^{N} \exp(a_i - A)\right)$$

Logistic Regression

Support Vector Machines

Support Vector Machines



"linearly separable"

$$\min_{x} \; 0^T x$$
$$\text{s.t. } \bar{w}^T x_i \geq 1, \quad y_i = 1$$
$$\bar{w}^T x_i \leq -1, \quad y_i = -1$$