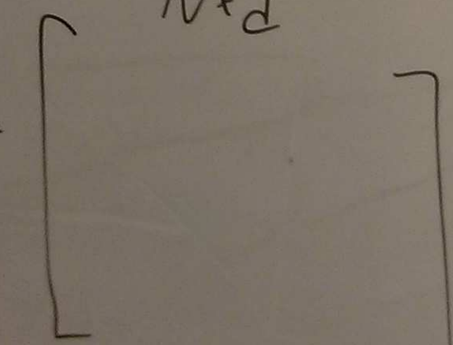
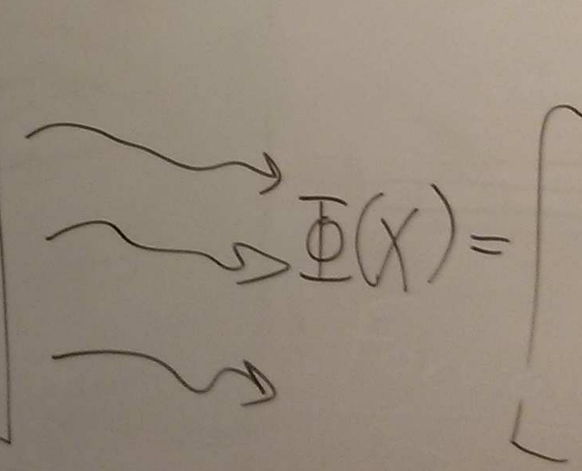
AA

Valid kernels
L_1 - Regularization
Logistic Regression

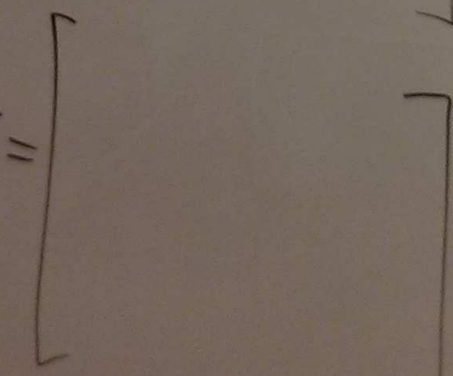
Ass 2 due Wed
Marked Ass 1 due now.

Review of kernel "trick"

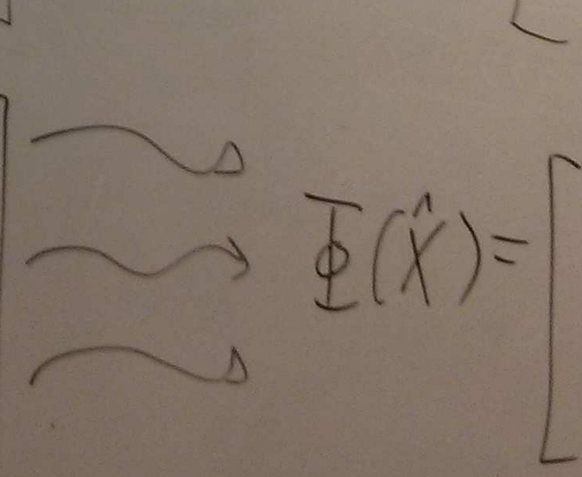
(training) $X =$ 

$\Phi(X) =$ 

$N \times d^{100}$

(testing) $\hat{X} =$ 

$T \times d$

$\Phi(\hat{X}) =$ 

$T \times d^{100}$

led

now.

$N \times d^{100}$

$T \times d^{100}$

$$\hat{Y} = \underbrace{\Phi(\hat{X})}_{T \times d^{100}} w_{MAP}$$

$$= \underbrace{\Phi(\hat{X})}_{T \times d^{100}} \Big(\underbrace{\Phi(X)^T \Phi(X)}_{d^{100} \times d^{100}} + \lambda I \Big)^{-1} \underbrace{\Phi(X)^T}_{d^{100} \times N} \underbrace{Y}_{N \times 1}$$

$$= \underbrace{\Phi(\hat{X}) \Phi(X)^T}_{T \times N} \Big(\underbrace{\Phi(X) \Phi(X)^T}_{N \times N} + \lambda I \Big)^{-1} Y$$

(*)

$$= \underbrace{k(\hat{X}, X)}_{T \times N} \Big(\underbrace{k(X, X)}_{N \times N} + \lambda I \Big)^{-1} Y$$

$T \times d^{100}$

If you can compute

$K(X, X)$ and $K(X^2, X)$,

no need for $\Phi(X), \Phi(X^2)$.

$$\Phi(X)\Phi(X^T) = \begin{bmatrix} \phi(\bar{x}_1)^T\phi(\bar{x}_1) & \phi(\bar{x}_1)^T\phi(\bar{x}_2) & \cdots \\ \phi(\bar{x}_2)^T\phi(\bar{x}_1) & \ddots & \\ \vdots & & \end{bmatrix}$$

$$K(X, X) = \begin{bmatrix} k(\bar{x}_1, \bar{x}_1) & k(\bar{x}_1, \bar{x}_2) & \cdots \\ k(\bar{x}_2, \bar{x}) & \ddots & \\ \vdots & & \end{bmatrix}$$

Valid kernels

Q: When does there exist feature map ϕ, such that for kernel k we have $\phi(\bar{x}_i)^T\phi(\bar{x}_j) = k(\bar{x}_1, \bar{x}_2)$???

A: If $K(X_1, X_2) \geq 0$, for all X_1, X_2.

(may be hard to show)

Valid kernels

L_1 - Regularization Logistic Regression

Assume k_1, k_2 is valid:

- $ck, \quad c > 0$
- $k + k'$
- $k(\phi(x), \phi(x'))$
- $\exp(k)$
- $q(k)$, q polynomial with non-negative coefficients
- $f(x)\, k(x, x')\, f(x')$

Course Project Ideas:

- Take your application, design a kernel
- Scale up to large 'N'.
- Learning the kernel (MKL, etc)

Ideas:

r

kernel.

to

he

KL, etc.)

Search over features combinations

- NP-hard under most criteria (BIC)
  (CV)

Greedy methods

- forward Selection: start with empty, add best.
  (sub-modular: best approximation)
- Backward select: start w/ all vars, remove worst.
- Stagewise

L_1 - Regular

$\arg\min_{w} \frac{1}{2}||$

- regularizat
- encourages va

$\hat{X} w = [\ldots [\,] \ldots$

mb: nations

criteria (BIC)

(CV)

t with empty, add best.

be st approximation)

w all vars, remove worst.

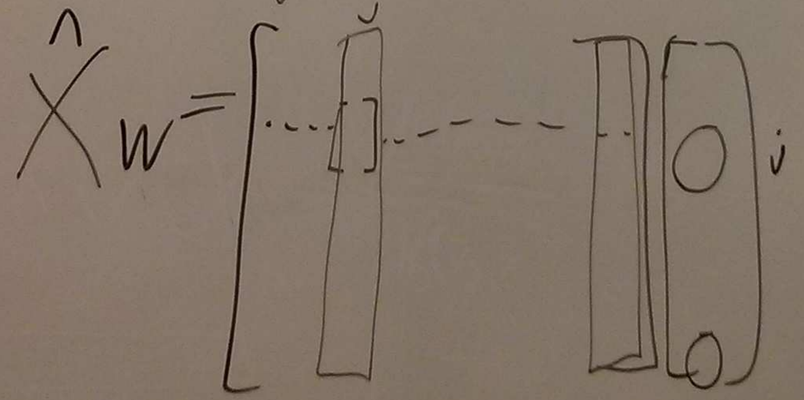
L_1 – Regularization

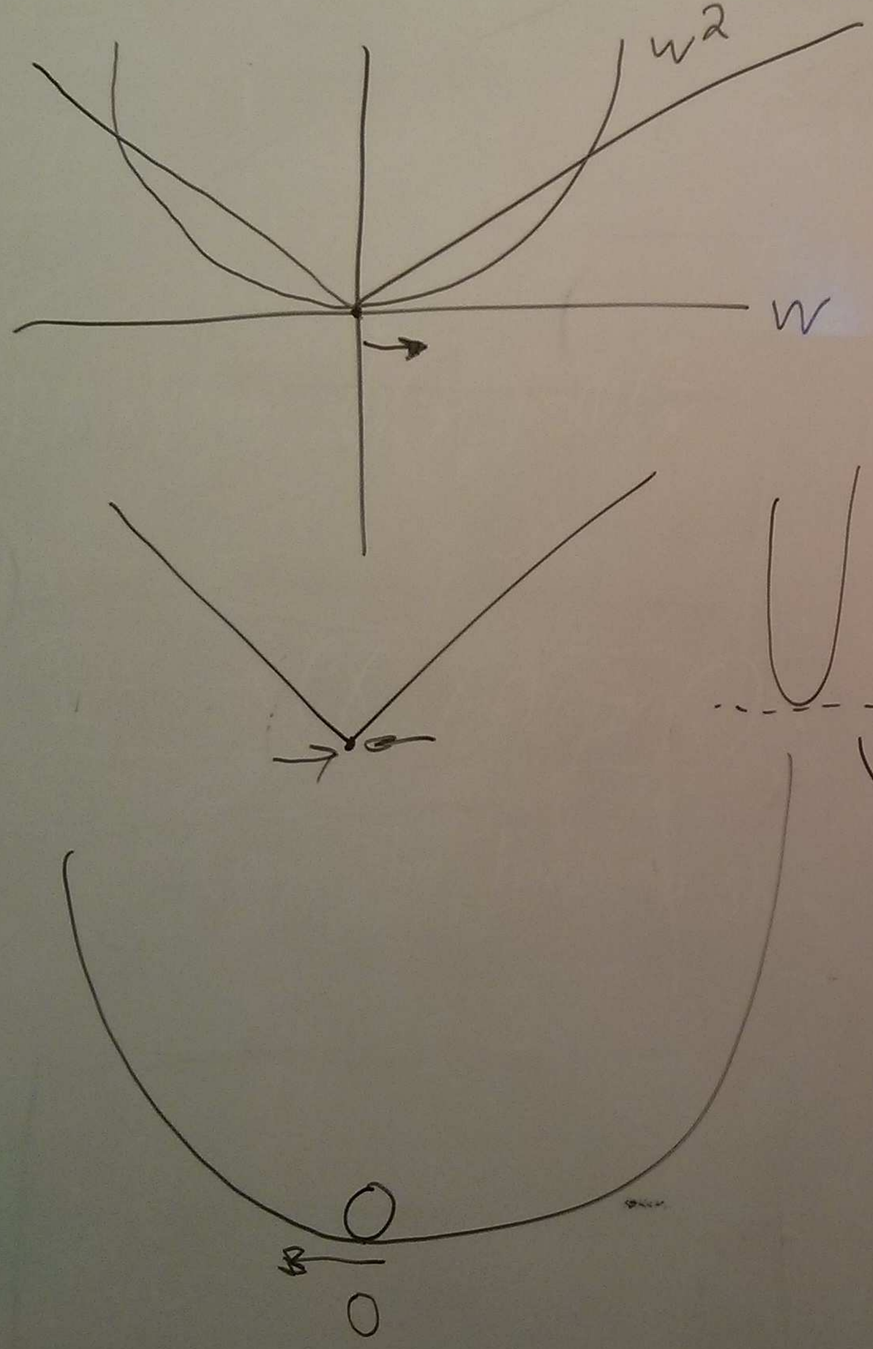
$$\underset{\bar{w}}{\operatorname{argmin}} \frac{1}{2}\|X\bar{w} - Y\|^2 + \lambda\|\bar{w}\|_1$$

$\|\bar{w}\|_1 = \sum_{i=1}^{d} |w_i|$

– regularization: protect against over-fitting (not unique)

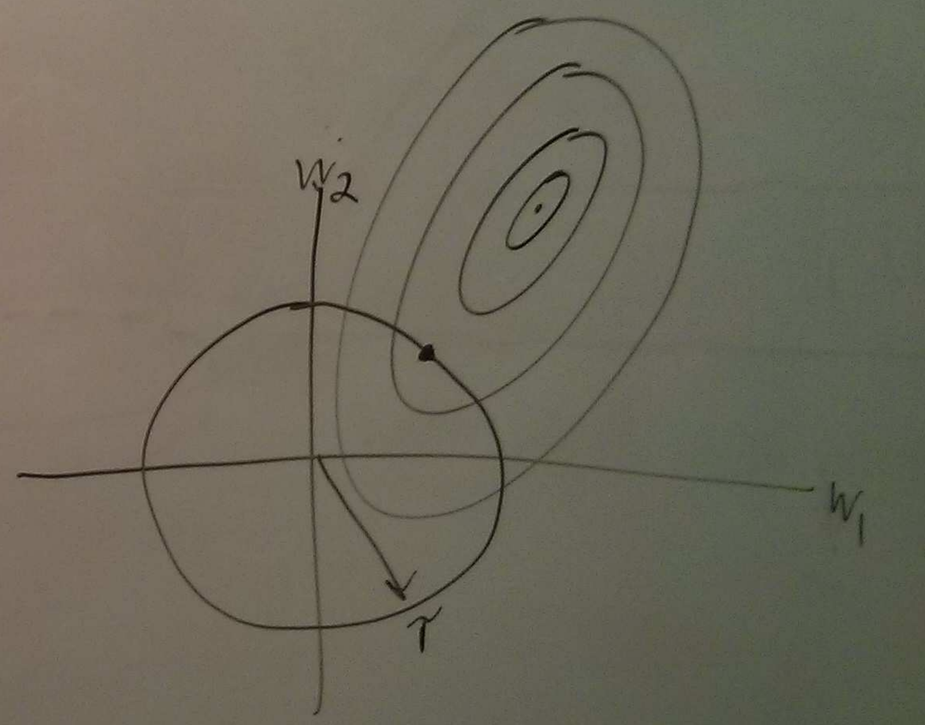
– encourages variable to be exactly 0

$\hat{X}w =$ 



$$\underset{w}{\operatorname{argmin}} \frac{1}{2}\|X\bar{w} - \bar{y}\|^2$$

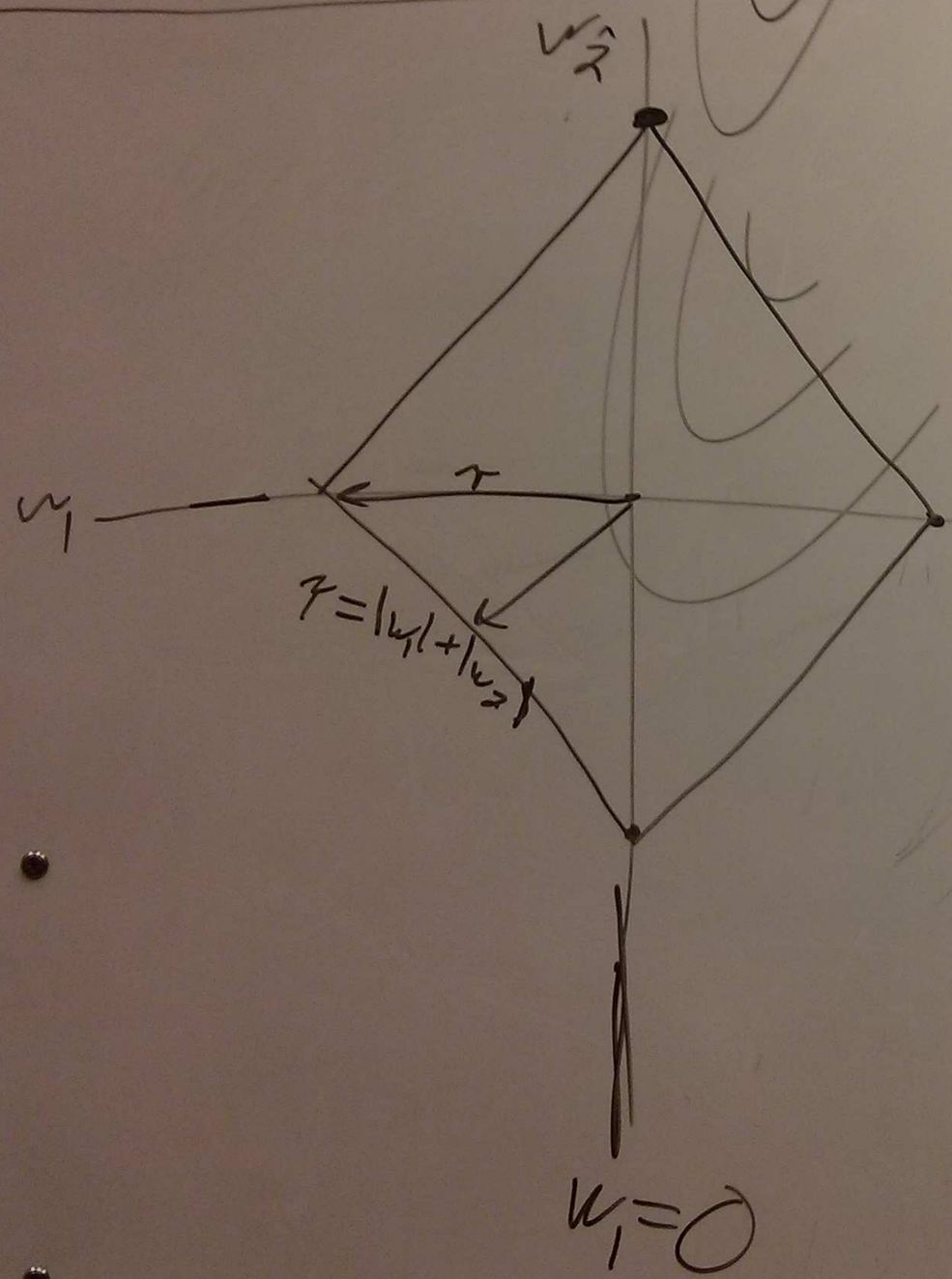
subject to $\|\bar{w}\| \leq \gamma$



Valid kernels

L_1 - Regularization

Logistic Regression



"Lasso"

$w = (0, \tau)$

(Sm

(smooth) + (non-smooth)
but <u>separable</u>

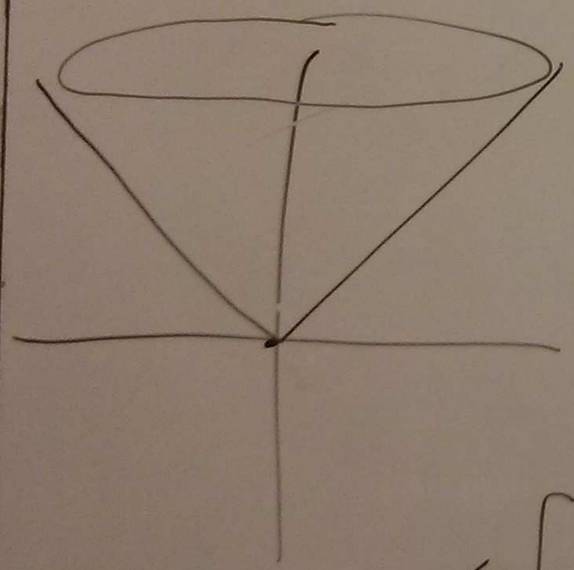
$$f(x) = \sum_{i=1}^{d} f_i(x_i)$$

Elastic net: $\frac{\lambda_2}{2}\|w\|^2 + \lambda_1\|w\|_1$

unique sparse

Group L-Regularization: $\lambda \sum_g \|w_g\|$

w_g: sub-vector



Selects "groups" of variables

[. . . . . 0 0 0 . . . . .]

w_g

N×d

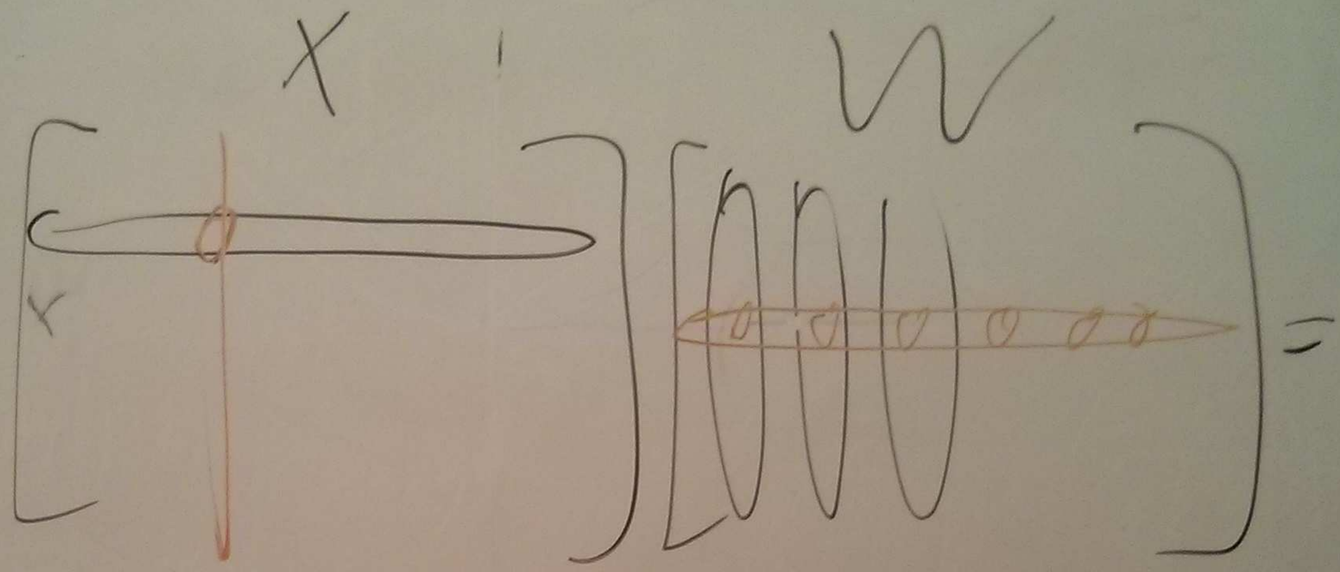
"multiple regression"

$X = [\quad]$ $Y = [\quad]$

N×C

XW

Nuclear



Nuclear-norm regularization:

- sum of singular values
- $\Rightarrow$ low-rank matrices

Structured Sparsity: "$w_2 \neq 0$ only if $w_1 \neq 0$"

(good project)

$\Rightarrow$ publication

W

$[(W)] =$ | x_1 x_2 x_3 x_4

x_1^2 x_1x_2 - - - - - - -

$x_1x_2x_5x_6$

...ation:

...of singular values

...low-rank matrices

$...x_2 \neq 0$ only

... if $w_1 \neq 0$"

nels

eation

ession

Supervised Learning

Generative $p(y_i | \vec{x}_i)$

Naive Bayes

GDA

Discrimina tive $p(y_i | x_i)$

Least squares

[Logistic Regression]

Discriminant function

KNN

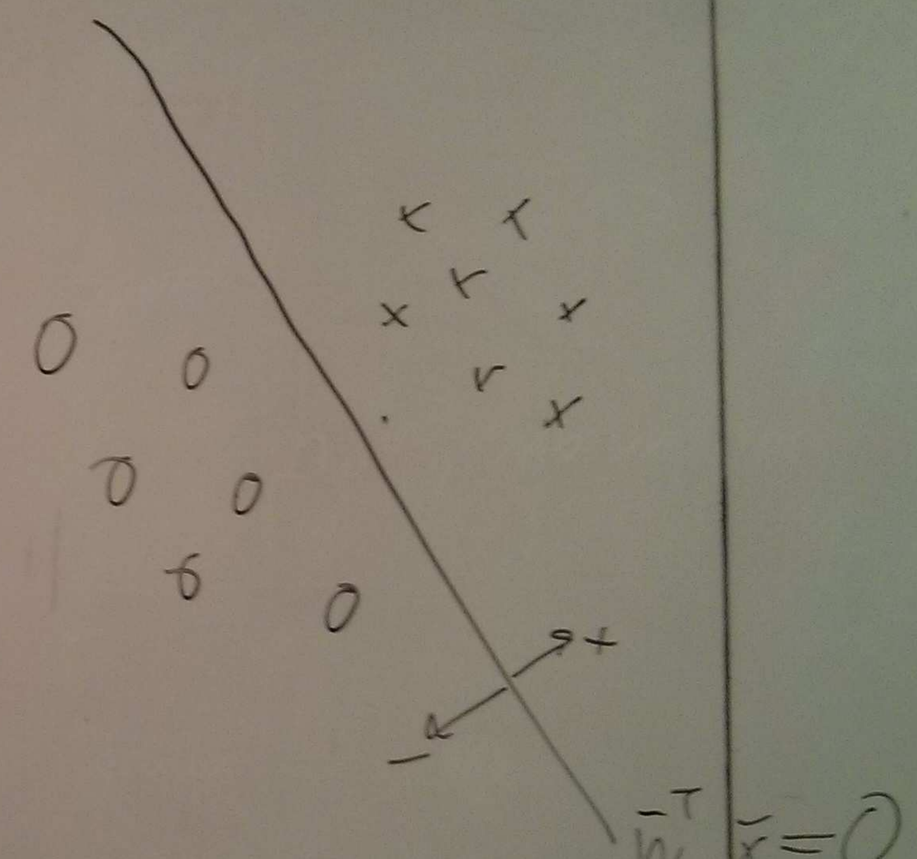
[Support vector machine]

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

target: $y_i \in \{-1, 1\}$

Linear classifiers:

$$\hat{y}_i = \text{sign}(w^T \bar{x}_i)$$



"logistic regression"

"hinge loss"

"support vector machine"

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

good

good

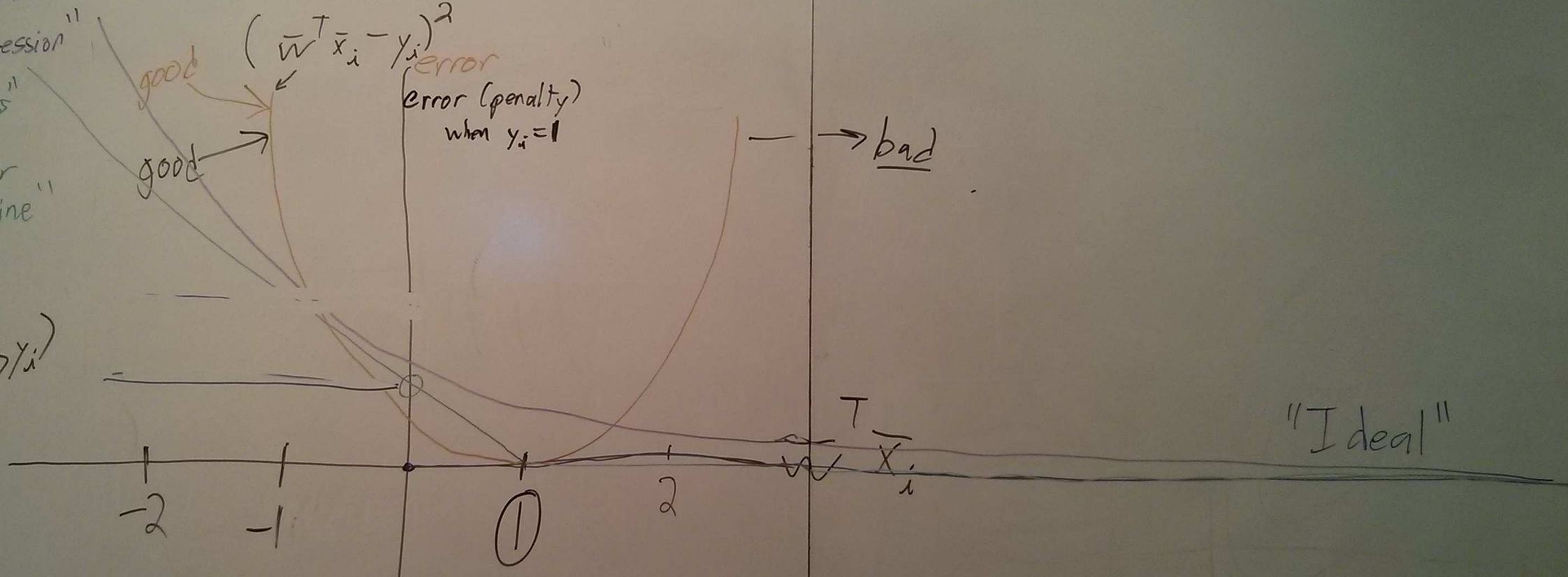
error

error (penalty) when $y_i = 1$

→ bad

$\sum_{i=1}^{N} f_i(\bar{w}^T \bar{x}_i, y_i)$

$= 0$



"Ideal"

Kernels

gularization

c Regression

log-Odds ratio

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

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

$$1 - p(y_i=1 \mid \bar{x}_i, \bar{w})$$

$$p_1 = (1 - p_1)\exp(\bar{w}^T \bar{x}_i)$$

$$p_1(1 + \exp(\bar{w}^T x_i)) = \exp(\bar{w}^T x_i)$$

$$p_1 = \frac{\exp(\bar{w}^T \bar{x}_1)}{1 + \exp(\bar{w}^T \bar{x}_i)} = \frac{1}{1 + \exp(-\bar{w}^T \bar{x}_i)}$$

"sigmoid function"

$$\therefore p_2 = \frac{1}{1 + \exp(\bar{w}^T \bar{x}_i)}$$