Lec 07:

① Finish up regression

② Classification

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

Regularization: Ridge: $\|\Phi w - y\|^2 + \lambda \|w\|_2^2 \rightarrow \text{argmin}$

LASSO: $\|\Phi w - y\|^2 + \lambda \|w\|_1$

OPT:

Ridge

$$\min \|\Phi w - y\|^2 \quad \text{s.t.} \quad \|w\|_2 \leq c_1$$

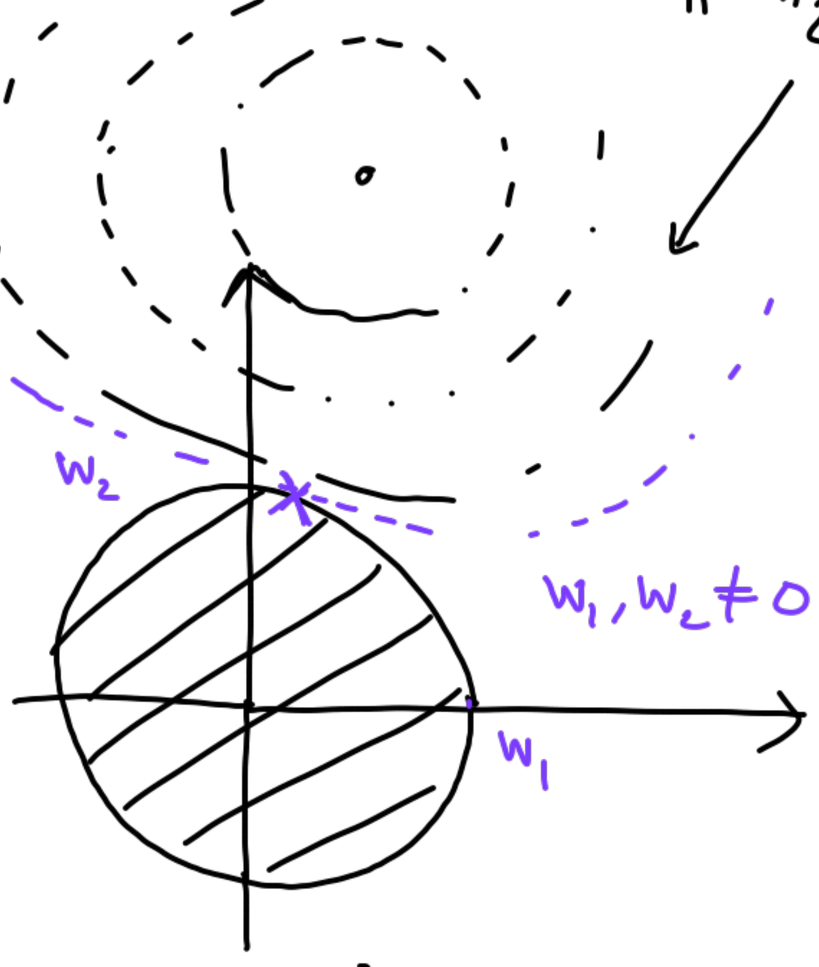
LASSO:

$$\min \|\Phi w - y\|^2 \quad \text{s.t.} \quad \|w\|_1 \leq c_2$$

Constraints : $\|w\|_2 \leq c_1$, $\|w\|_1 \leq c_2$

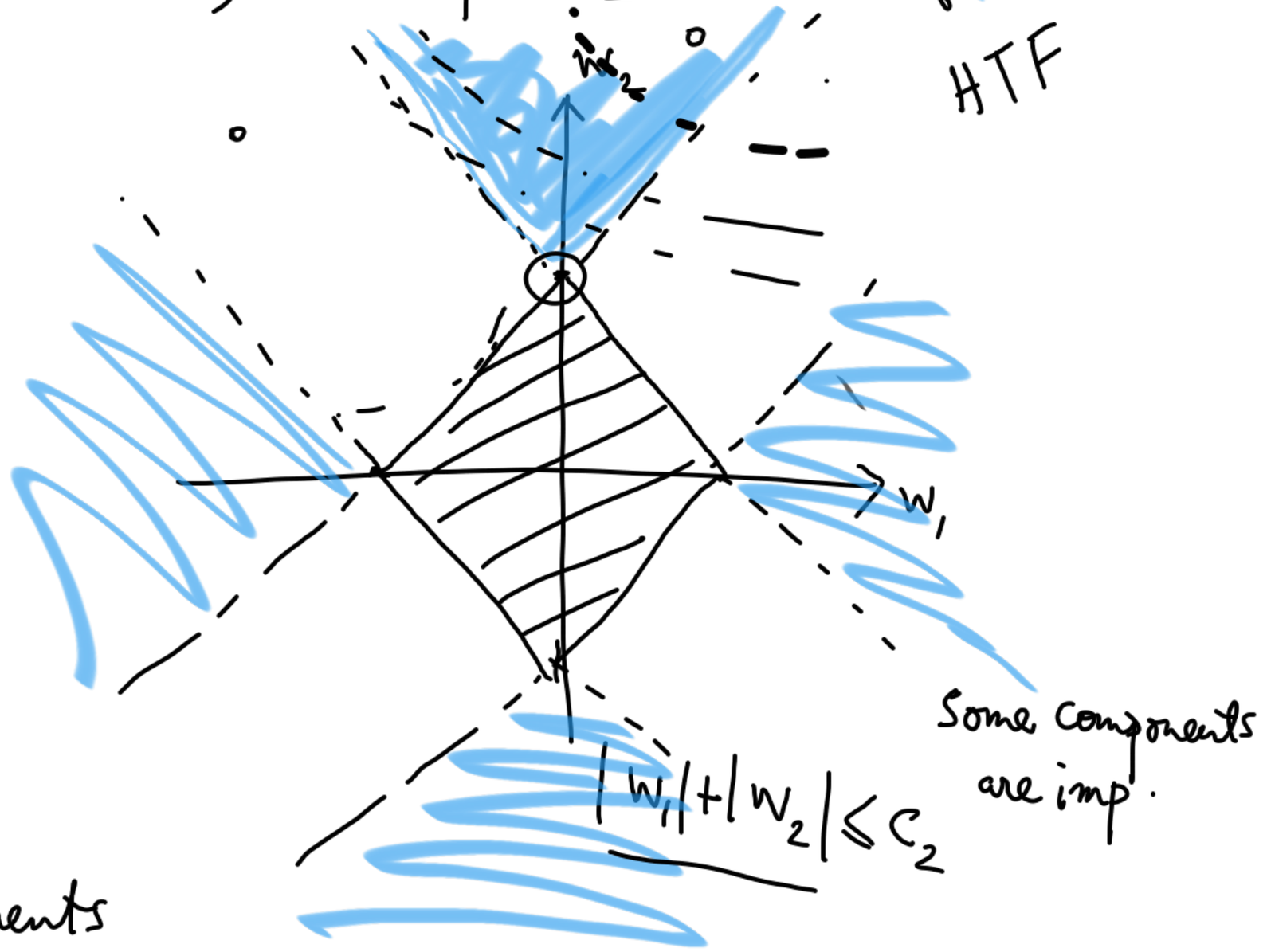
Elements of Stat learning

HTF



$w_1^2 + w_2^2 \leq c_1^2$

Applications: all components are imp



Some components are imp.

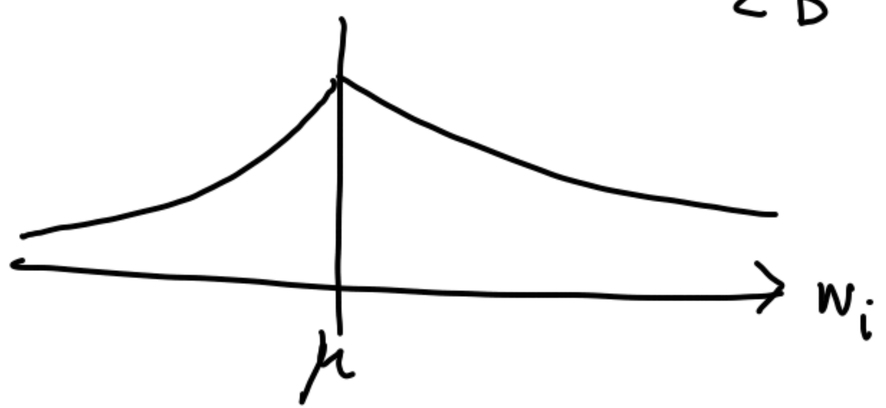
MAP : prior on w

Gaussion prior $\longrightarrow$ Ridge

Laplace prior $\longrightarrow$ LASSO

HW

$$\text{Laplace}\left(w_i \mid \mu, b\right) = \frac{1}{2b} \exp\left\{-\frac{|w_i - \mu|}{b}\right\}$$



Classification

Regression: $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$

Y $y_i \in S$ ← finite set, simplest $S = \{0, 1\}$

Binary classification

Probabilistic approach:

$\hat{x} \to \hat{y}$

$$\hat{y} = \underset{y}{\operatorname{argmax}} P(Y = y \mid \hat{x}) \qquad \hat{x} = (\hat{x}_1, \dots, \hat{x}_d)^T$$

$$P(Y = y \mid \hat{x}) = \frac{P(\hat{x} \mid Y = y)\, P(Y = y)}{P(\hat{x})}$$

Naive Bayes

$x_1, x_2, \ldots, x_d \rightarrow (2^d - 1) \times (k-1)$

Assumption: $P(x \mid Y=y) = \prod_{i=1}^{d} P(x_i \mid Y=y)$

$$\underset{y}{\operatorname{argmax}}\ P(Y=y \mid x) = \underset{y}{\operatorname{argmax}} \left(\prod_{i=1}^{d} P(x_i \mid Y=y) \right) P(Y=y)$$

$(d \times 1 \times (k-1))$

$$= \underset{y}{\operatorname{argmax}} \left\{ \sum_{j=1}^{d} \log P(x_i \mid Y=y) + \log P(Y=y) \right\}$$

$= P(x_i = 1 \mid Y=y)$

$$= \frac{\#\text{ of times } x_i \text{ and } y \text{ happen together}}{\#\text{ of times } Y=y}$$

$$P(Y=y) = \frac{\#\text{ of times } Y=y}{n}.$$

Uses of NB classifier: Topic classification

Given a document → find its topic

Treat the document as a "bag of words"

given sentences → create a vocab of words

(tokenization) → index them arbitrarily

Article: $x = \{x_1, x_2, \cdots, x_N\}$ x_i is the i^{th} word

$$P(x_i \mid Y=y) = \frac{\#\text{ of times } x_i \text{ appears in the doc of } y + 1}{\sum_{w \in V} \#\text{ of times } w \text{ appears in } y + |V|}$$

for unseen words.

$$P(x \mid Y=y) = \prod_{i=1}^{N} P(x_i \mid Y=y)$$

$P(Y=y)$ = relative fraction of y in all topics.

extending x_i for a group of words → n-grams

Disadvantages of NB

True distribution of binary variable X_1 and Y

$Y=0$	$Y=1$
0.8	0.2

$P(Y)$

	$X_1=0$	$X_1=1$
$Y=0$	0.7	0.3
$Y=1$	0.3	0.7

$P(X_1|Y)$

$P(Y=0) = 0.8$

$P(Y=1) = 0.2$

NB classifier

$$\hat{y} = \underset{y}{\operatorname{argmax}}\ P(Y=y \mid X_1) = \underset{y}{\operatorname{argmax}}\ P(X_1 \mid Y=y)\, P(Y=y)$$

	$P(X_1, Y=0)$	$P(X_1, Y=1)$	$\hat{y}$
$X_1=0$	0.7×0.8	0.3×0.2	0
$X_1=1$	0.3×0.8	0.7×0.2	0

Expected error $= P(Y \neq \hat{y})$

$= P(Y \neq \hat{y}, X_1=0) + P(Y \neq \hat{y}, X_1=1)$

$= P(Y=1, X_1=0) + P(Y=1, X_1=1)$

$= 0.06 + 0.14 = \underline{0.2}$

X_2 identical to X_1 → copy of y

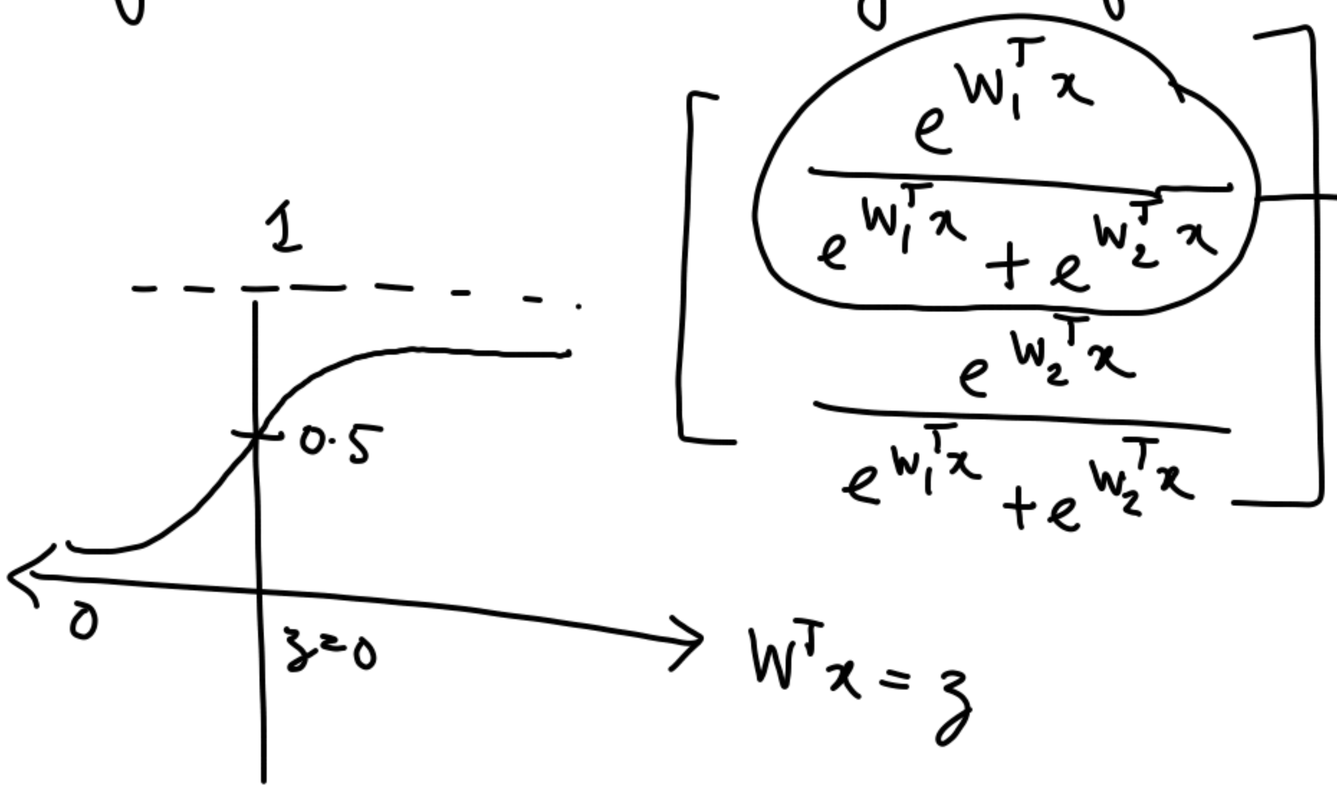
	$P(X_1, X_2, Y=0)$	$P(\ldots, Y=1)$	$\hat{y}$
$X_1=X_2=0$	$0.7^2 \times 0.8$	$0.3^2 \times 0.2$	0
$X_1=X_2=1$	$0.3^2 \times 0.8$	$(0.7)^2 \times 0.2$	1
$X_1=0, X_2=1$	✓	✓	0
$X_1=1, X_2=0$	✓	✓	0

$P(Y \neq \hat{y}) = P(Y=1, X_1=0) + P(Y=0, X_1=1)$

$= 0.3$

Logistic Regression (Classification)

$y \approx W^T x$ Binary classification $W_1^T x$ $x \in \mathbb{R}^d$



$$\frac{1}{1 + e^{(W_2 - W_1)^T x}} = \frac{1}{1 + e^{-W^T x}}$$

$-W$

$= \sigma(W^T x)$

Assumption: $P(Y=1 \mid x, W)$

$= \sigma(W^T x)$

$$\frac{P(y=1 \mid x, w)}{P(y=0 \mid x, w)} > 1 \Rightarrow \hat{y} = 1$$

ow $\Rightarrow \hat{y} = 0$

$\exp(w^T x) > 1 \Rightarrow w^T x > 0$

Linear classifier