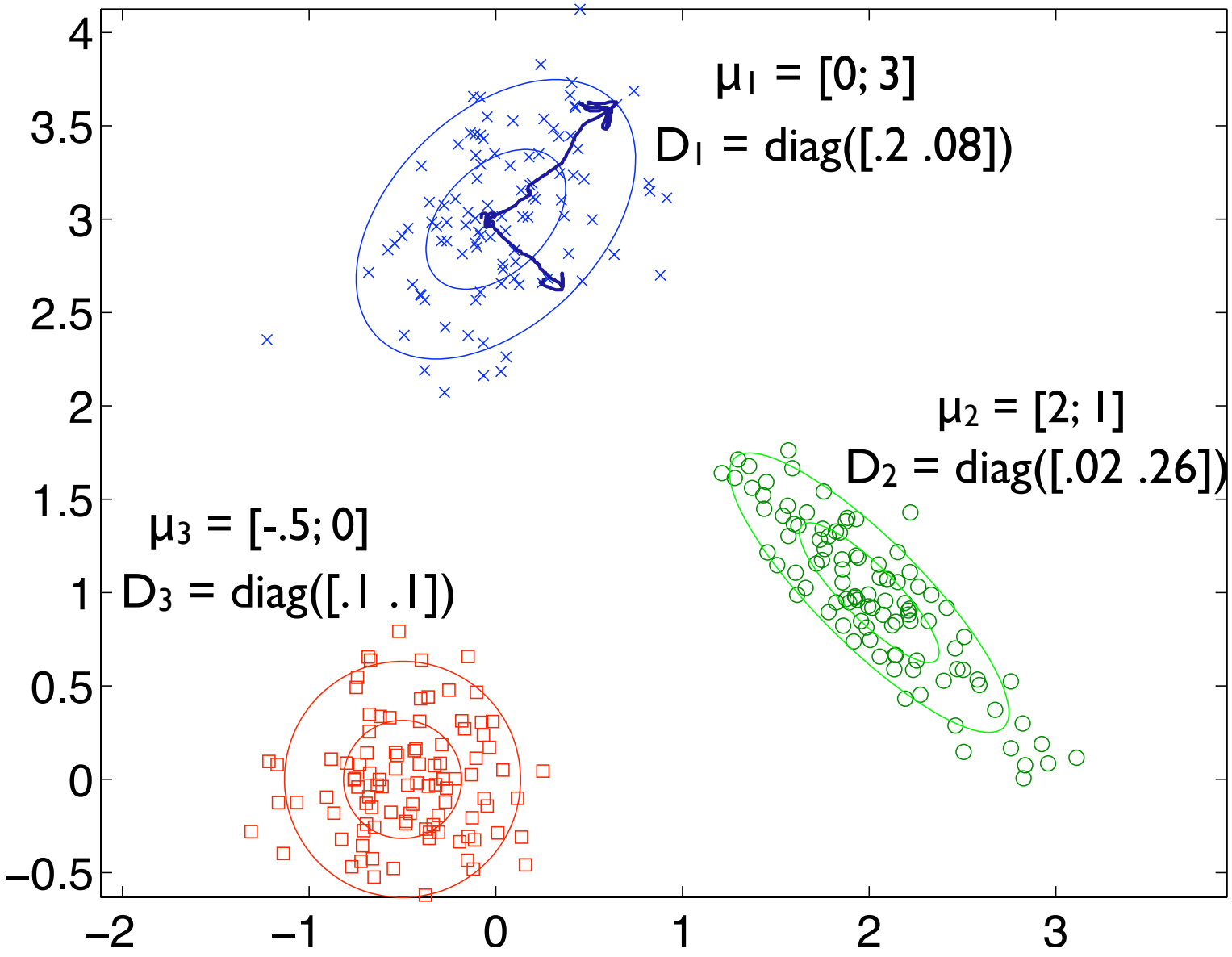
Review

- Multivariate Gaussian
  - $N(X \mid \mu, \Sigma) =$

$$(1/\sqrt{|2\pi\, \Sigma|})\, \exp\{-0.5\, (x-\mu)^T\, \Sigma^{-1}\, (x-\mu)\}$$

Multivariate Gaussians

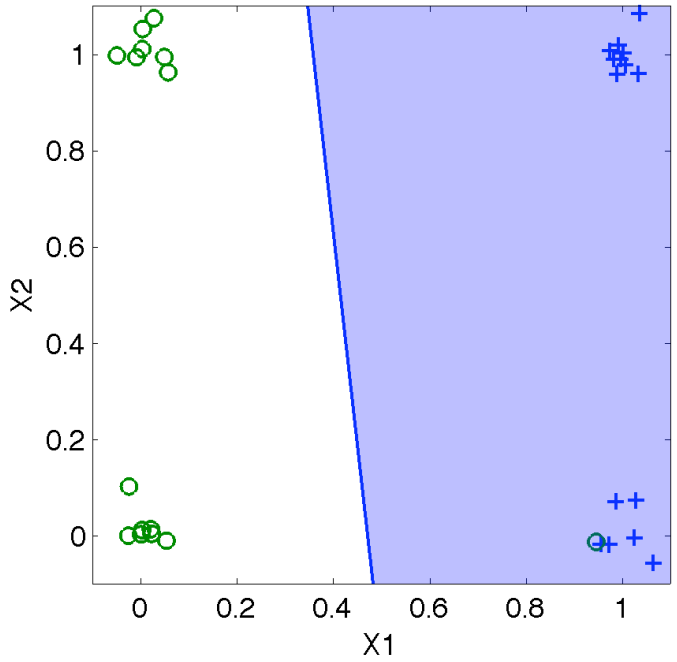


$$\Sigma_i = U D_i U^T$$

$$U = \begin{pmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ \sqrt{2}/2 & -\sqrt{2}/2 \end{pmatrix}$$

note $U^T U = I$

Review



- Naïve Bayes, Gaussian NB, Fisher model:
  - all lead to ***linear discriminants***
  - $w_0 + \sum_j w_j X_j \geq 0$ or $w_0 + w^T X \geq 0$
  - formulas for w_0, w depend on which model

Fisher linear discriminant

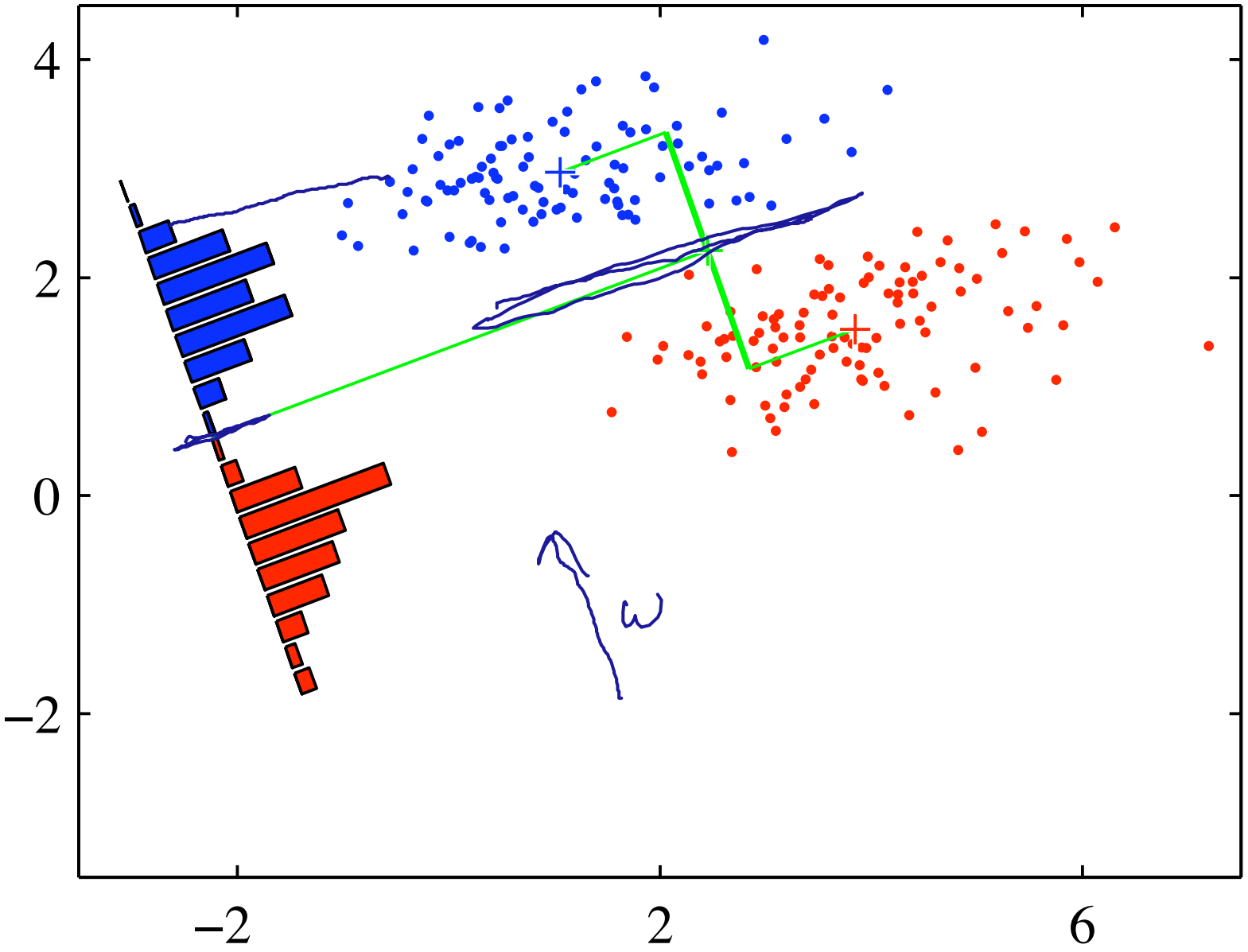


figure from book

Fisher w/ bad Σ (spherical)

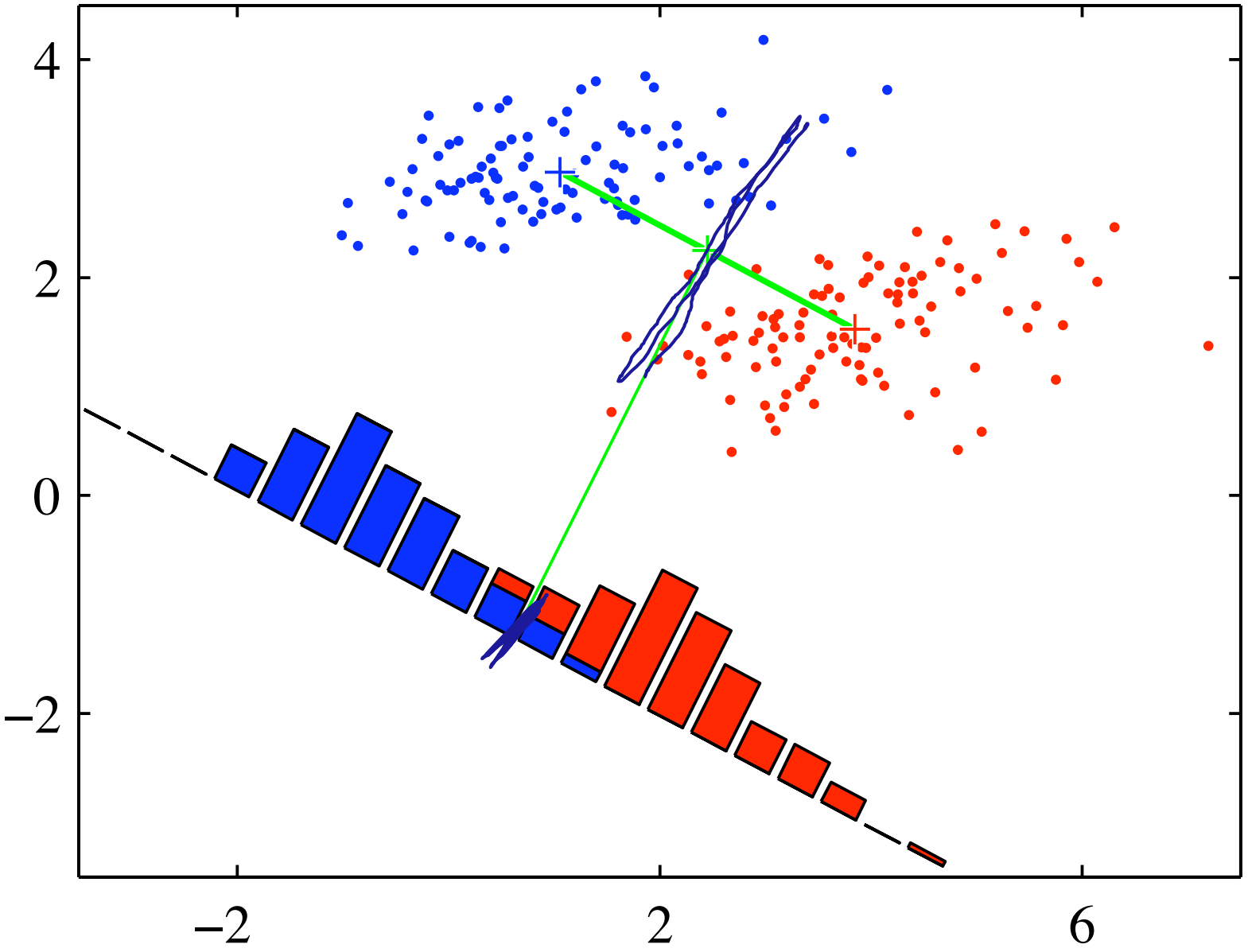


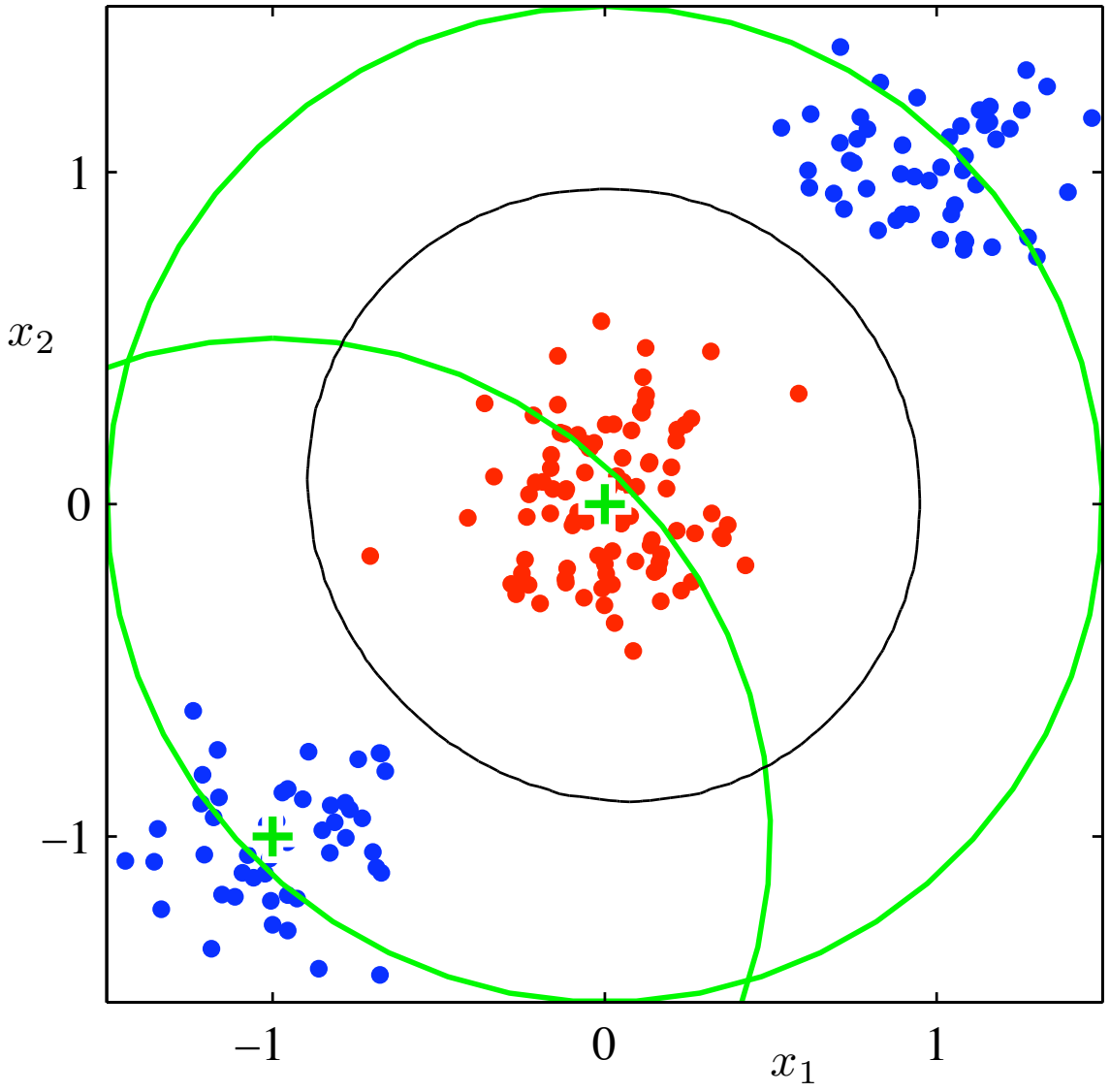
figure from book

Features

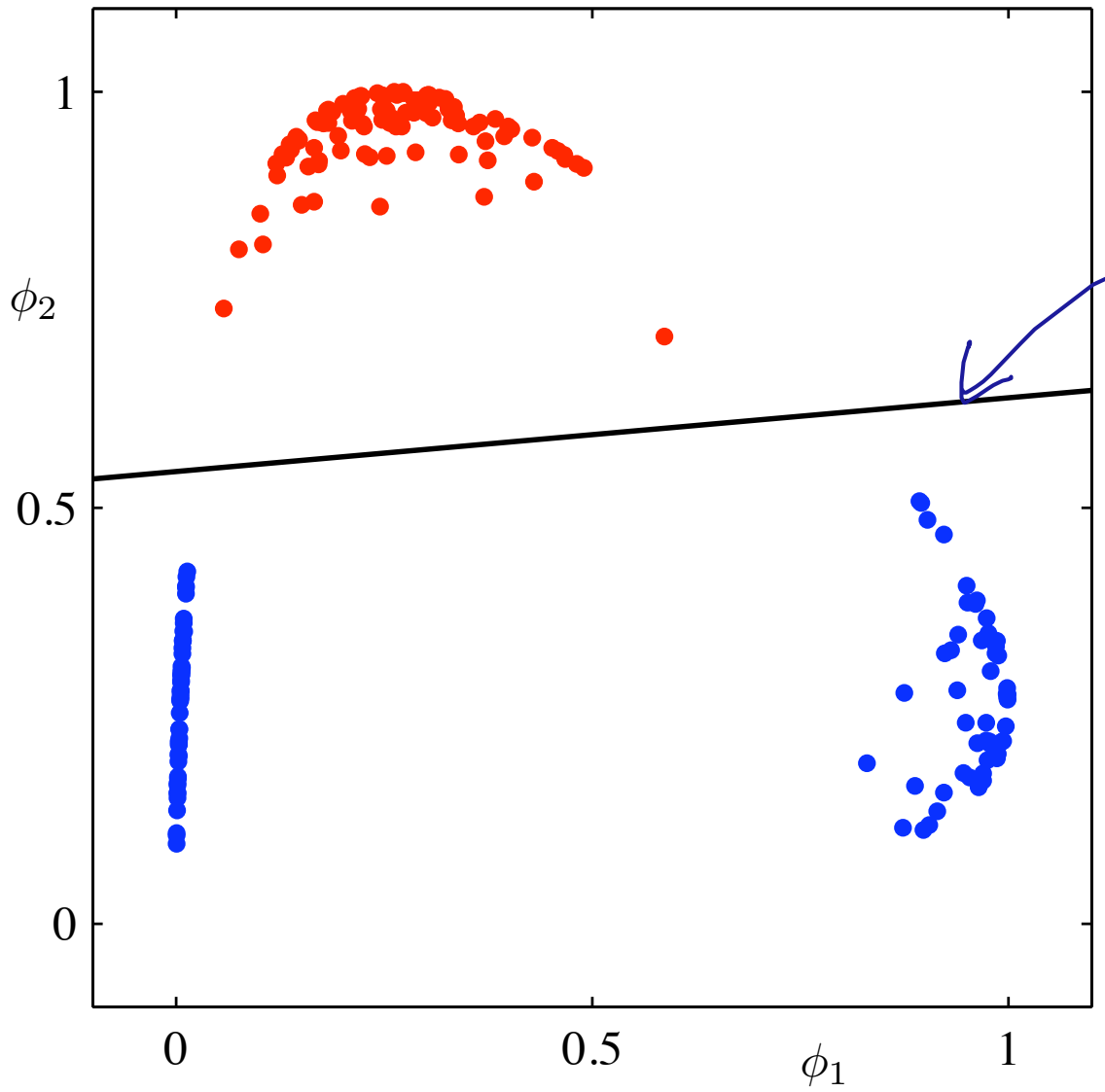
- Can generalize to use features of X:
  - $w_0 + \sum_j w_j \phi_j(\vec{X}) \geq 0$
  - $\phi_j(X)$ are ***features***
- Why might we want to do so?

Use of $\phi_j(X)$

not linearly separable



Use of $\phi_j(X)$

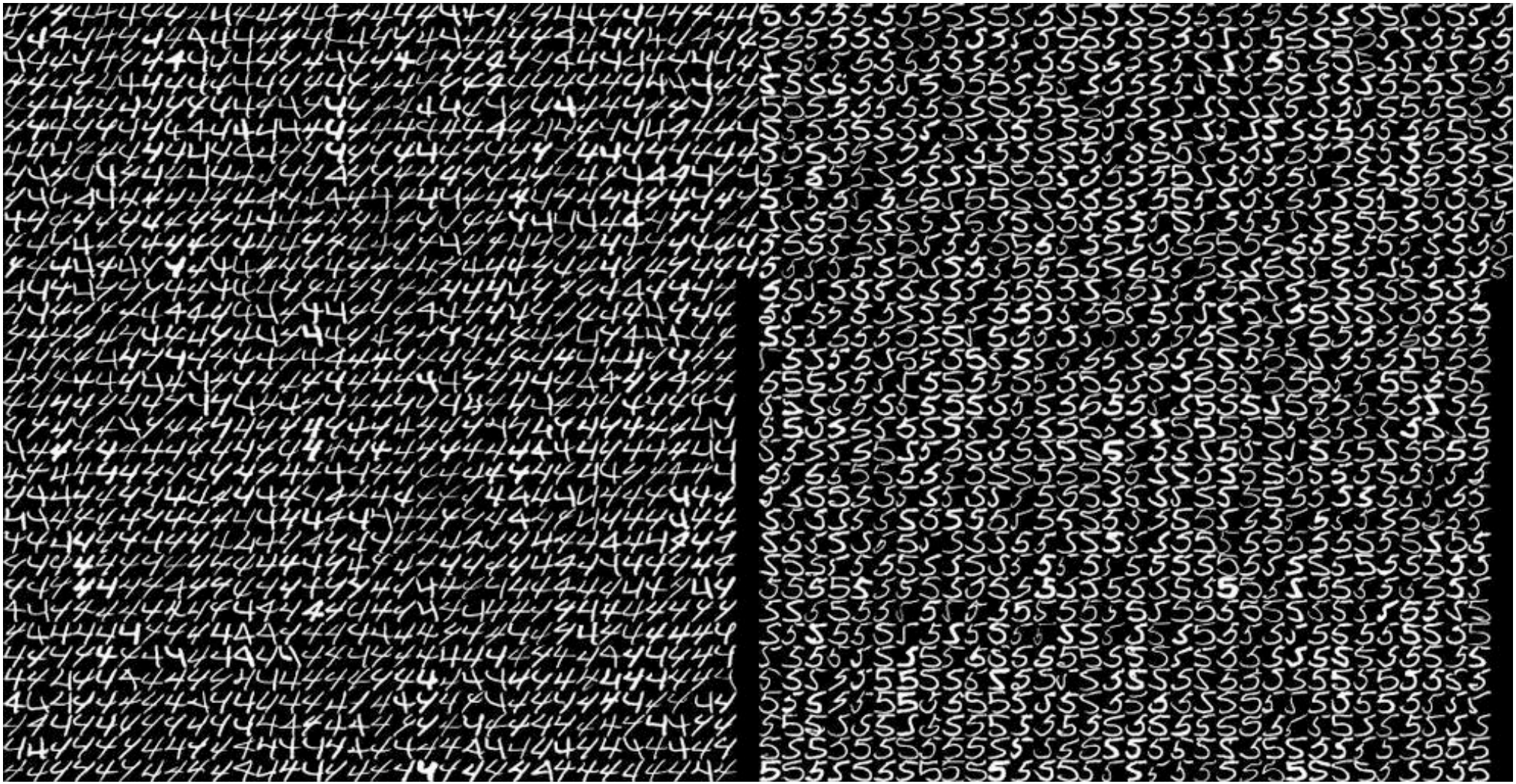


Lots of discriminants

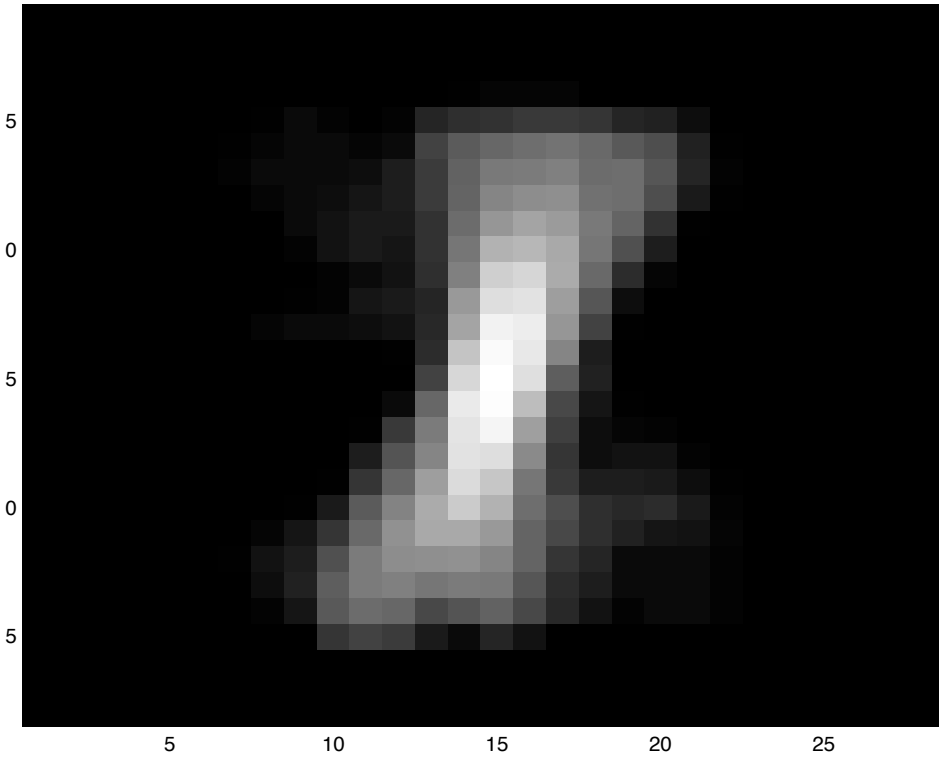
- One of most important types of classifier
- Consequently, many ways to train LDs
  - based on different assumptions about data
- We saw 3 so far
  - NB, GNB, Fisher
- Another one: coming up soon

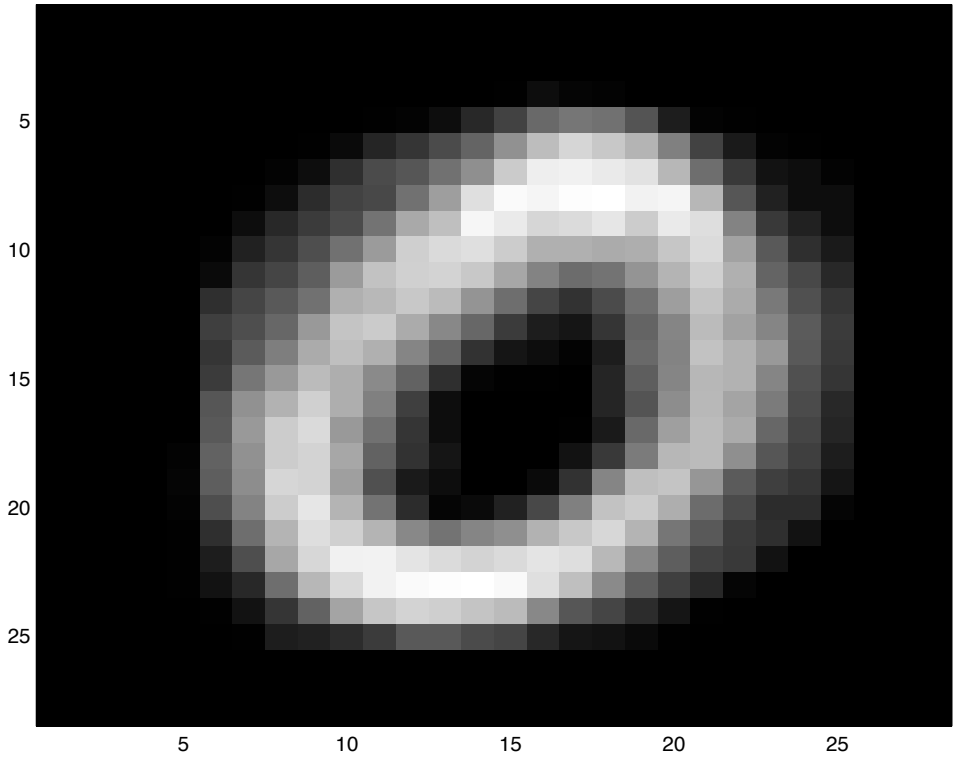
5 5 4 4

Application: USPS digits

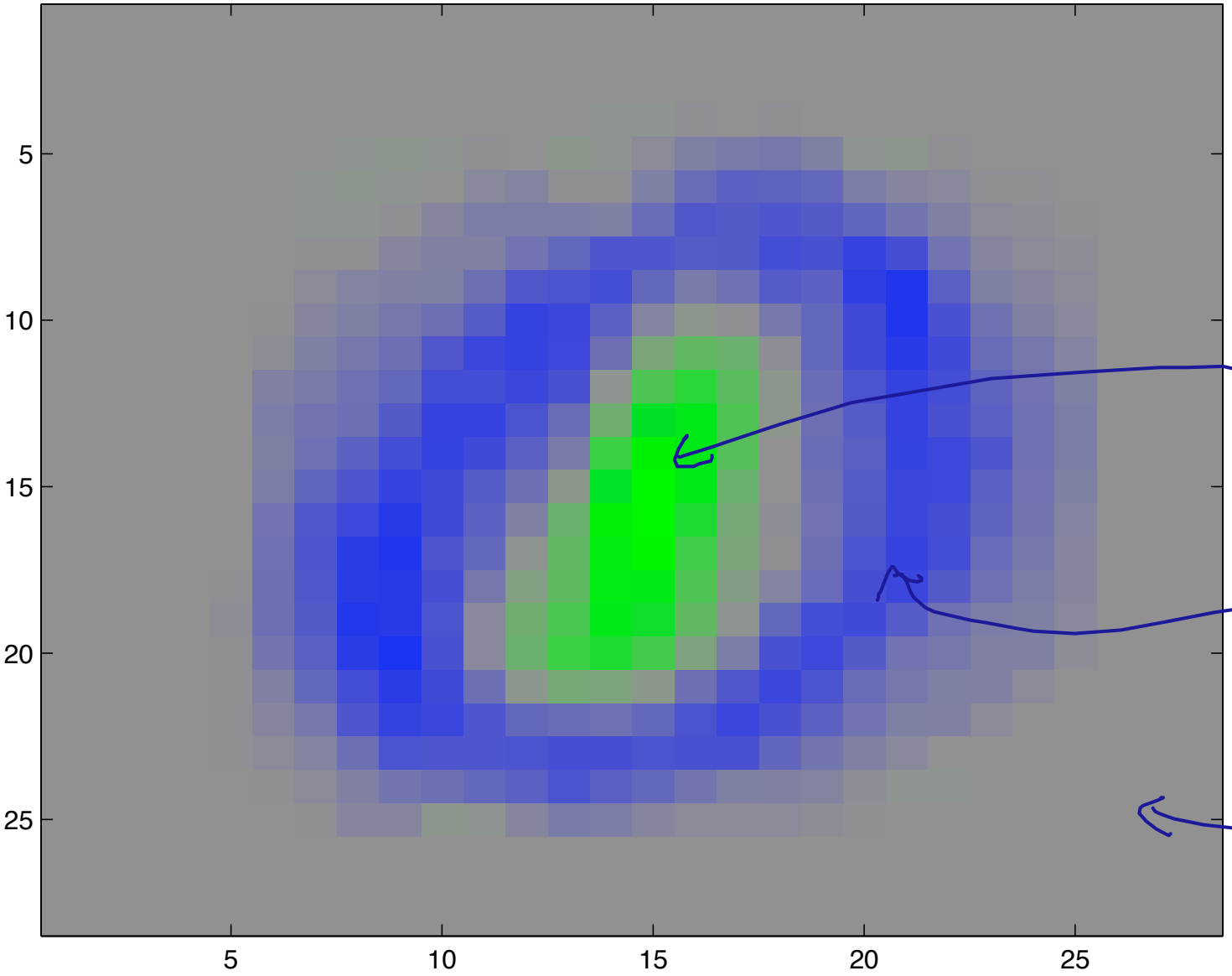


Class means for 0 and 1

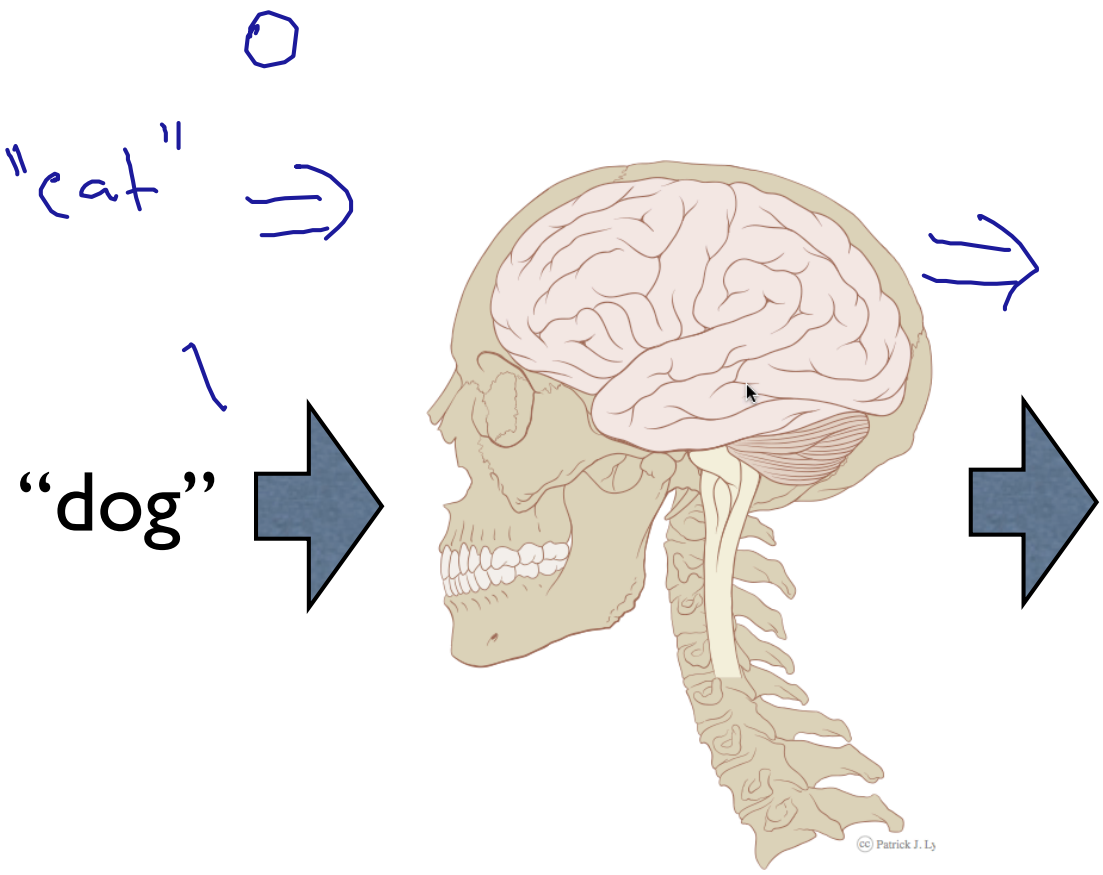


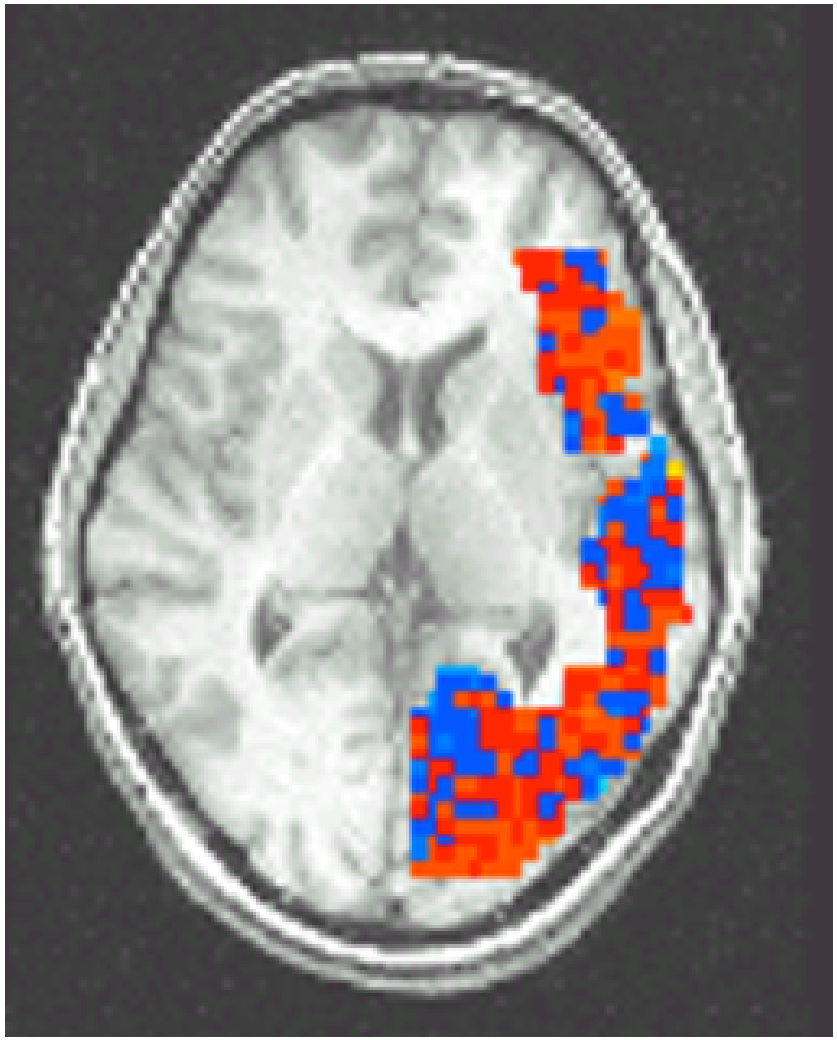


A linear discriminant



Application: fMRI classification





brain: Patrick J. Lynch, C. Carl Jaffe (CC att. 2.5)
fMRI: Mitchell et al

Class probability

- We showed:
  - log P(Y=1 | X) – log P(Y=0 | X) =

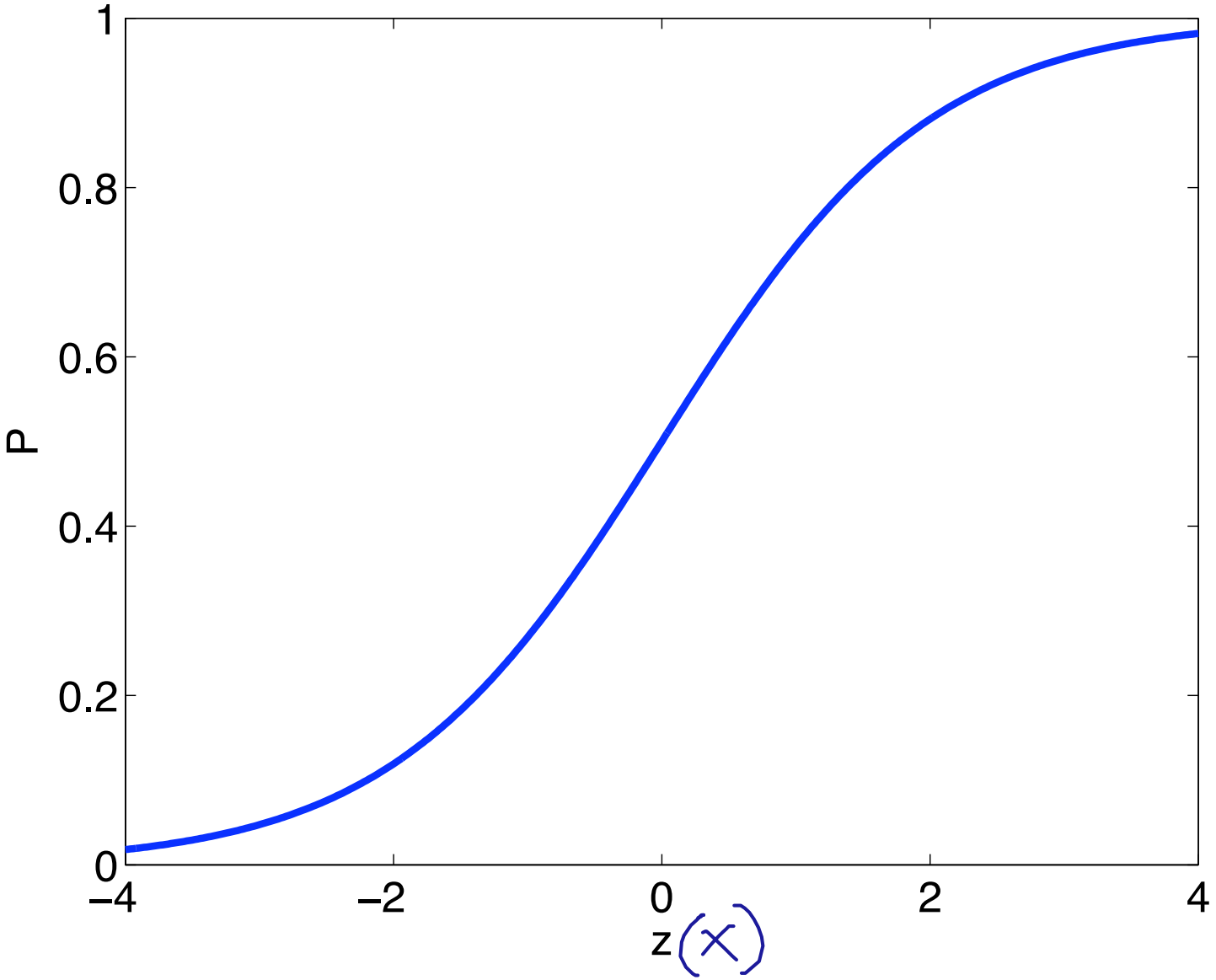
$$z(X) \equiv w_0 + \sum_j w_j X_j \quad \stackrel{?}{\geq} 0$$

- This implies
  - P(Y = 1) = $\frac{1}{1+e^{-z(X)}}$

logistic

Sigmoid: $\sigma(z) = 1/(1+\exp(-z))$

$p(y=1)$



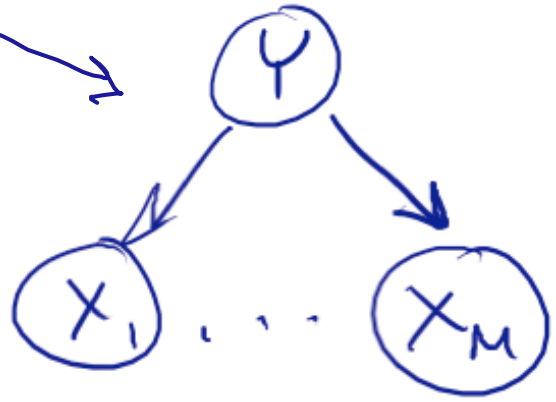
$\sigma(z)+\sigma(-z)=1$

$P(y=0|x) = \sigma(-z(x))$

NB = MLE (or MAP)

- $P(Y=1 \mid X) = \sigma(z)$
  - $z = w_0 + \sum w_j X_j$
- NB is one algorithm for finding w
- NB = maximum likelihood in this model
  - $\arg\max_w \prod_i P(Y^i, X^i \mid w)$

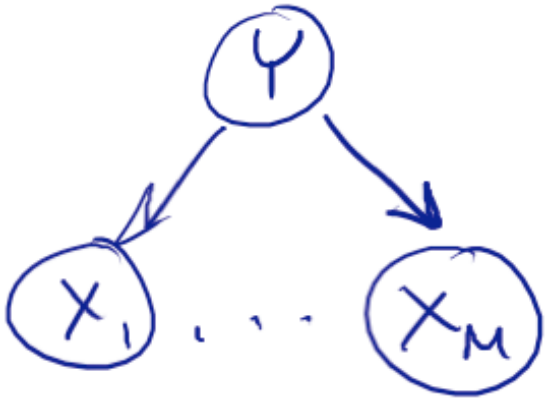
Actually, max is over all parameters of the model (w is derived from parameters, but doesn't include all of them)



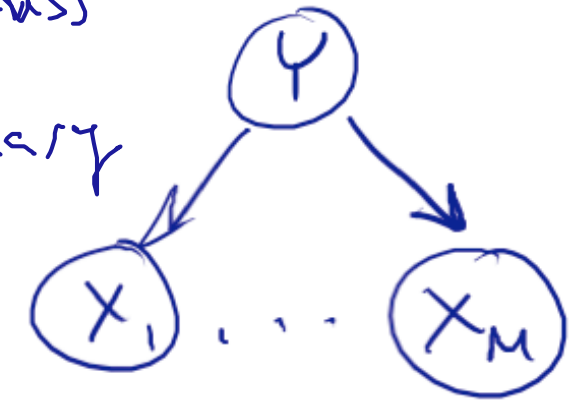
Conditional likelihood

- Another: maximum ***conditional*** likelihood
  - given data $(X^1, Y^1), \ldots, (X^N, Y^N)$
  - arg max $\arg\max_w \prod_i P(y^i \mid x^i, w)$
- ***Same*** model, ***different*** training criterion
- Cond. MLE for logistic linear discriminant: ***logistic regression***

here, the max really is just over w (no other parameters are relevant)



Discussion



- $\max_w P(X, Y \mid w)$ vs. $\max_w P(Y \mid X, w)$
- We've seen cond. MLE before:
- Why choose one?
  - MLE:
  - cond. MLE:

Generative vs discriminative

- Same trick works for any graphical model
  - if we know we're always going to be asking same query (Y given $X_1 \ldots X_M$), optimize for it
  - max $P(\text{query vars} \mid \text{evidence vars}, \theta)$ over θ
- Can improve performance, but also more risk of overfitting

$\sigma(z) = 1/(1+e^{-z})$

Logistic regression

binary $\{-1, 1\}$

- given data $(X^1, Y^1), \ldots, (X^N, Y^N)$
- $\arg\max_w \prod_i P(Y^i \mid X^i, w)$

$P(Y^i \mid X^i, w) = \begin{cases} \sigma(z(x)) & y=1 \\ 1-\sigma(z(x)) & y=0 \\ \hookrightarrow \sigma(-z(x)) & \end{cases}$

$= \arg\min_w \sum_i -\log P(Y^i \mid X^i, w)$

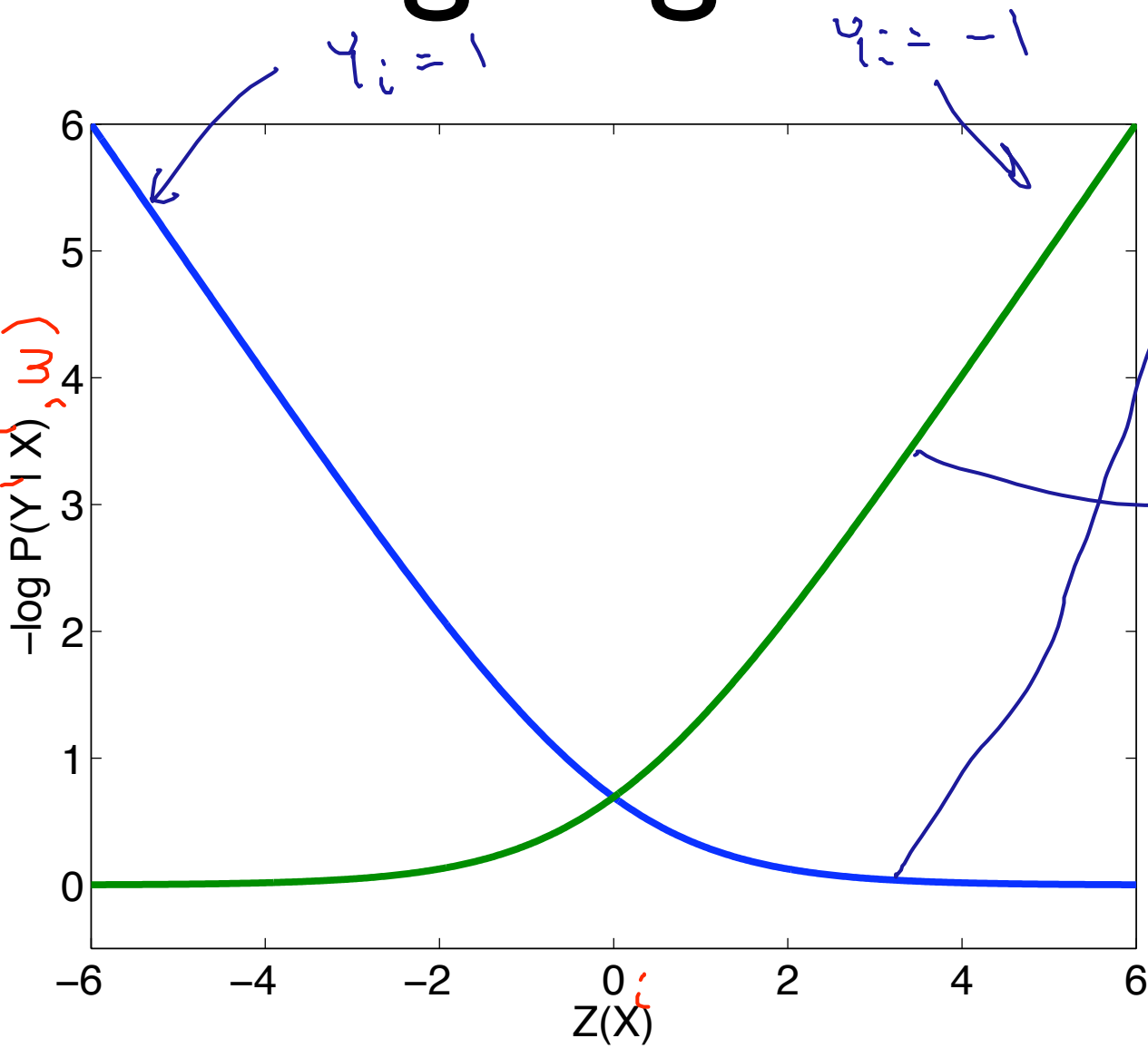
$= \arg\min_w \sum_i -\log \sigma(y^i z(x^i))$

$\sum_i \log(1 + \exp(-y^i z(x^i)))$

$\arg\min_w \sum_i \log(1 + \exp\{-y^i z(x^i)\})$

Neg. log likelihood

for one example



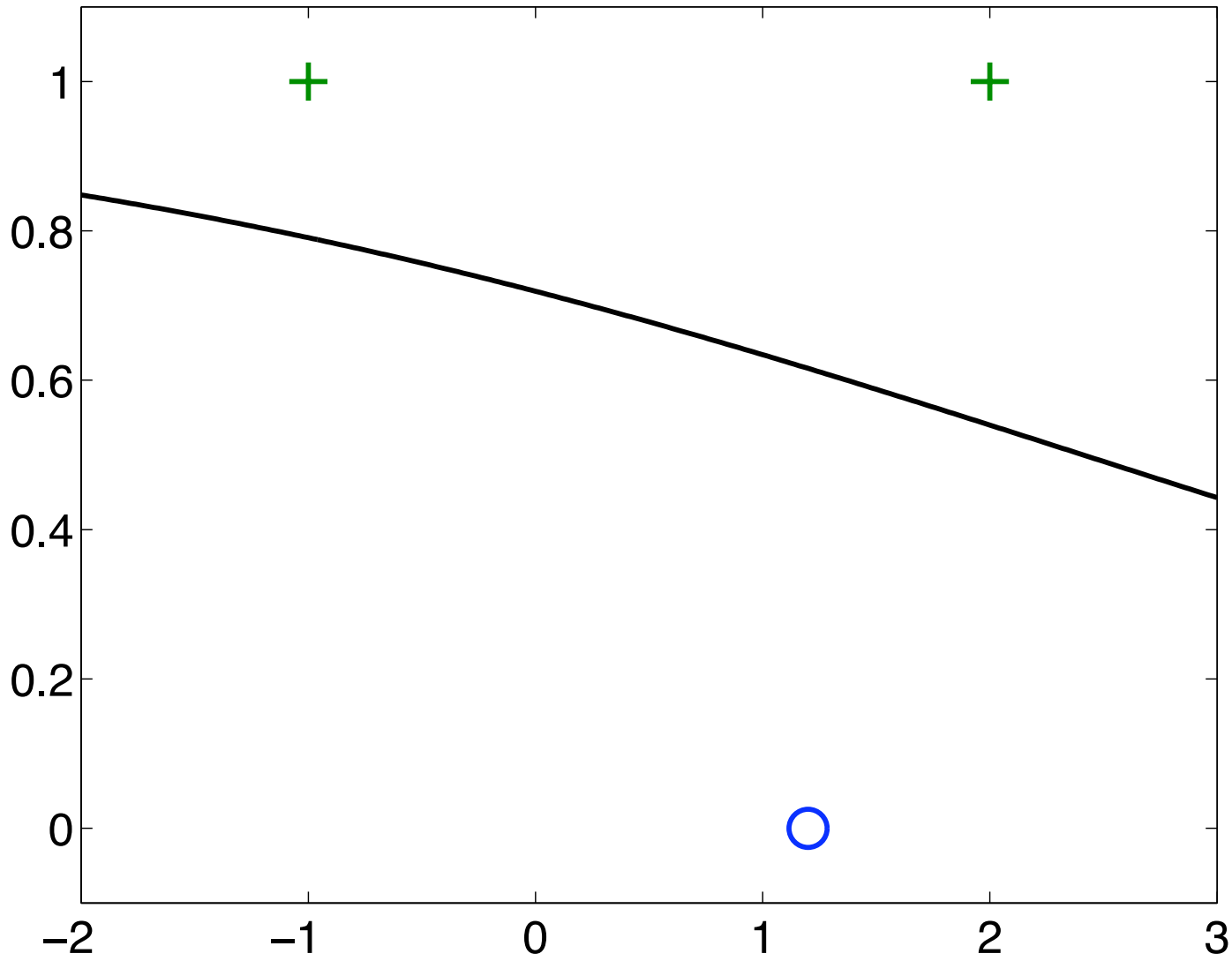
$\log(1+\exp(-z))$

$\approx \max(0, -z)$

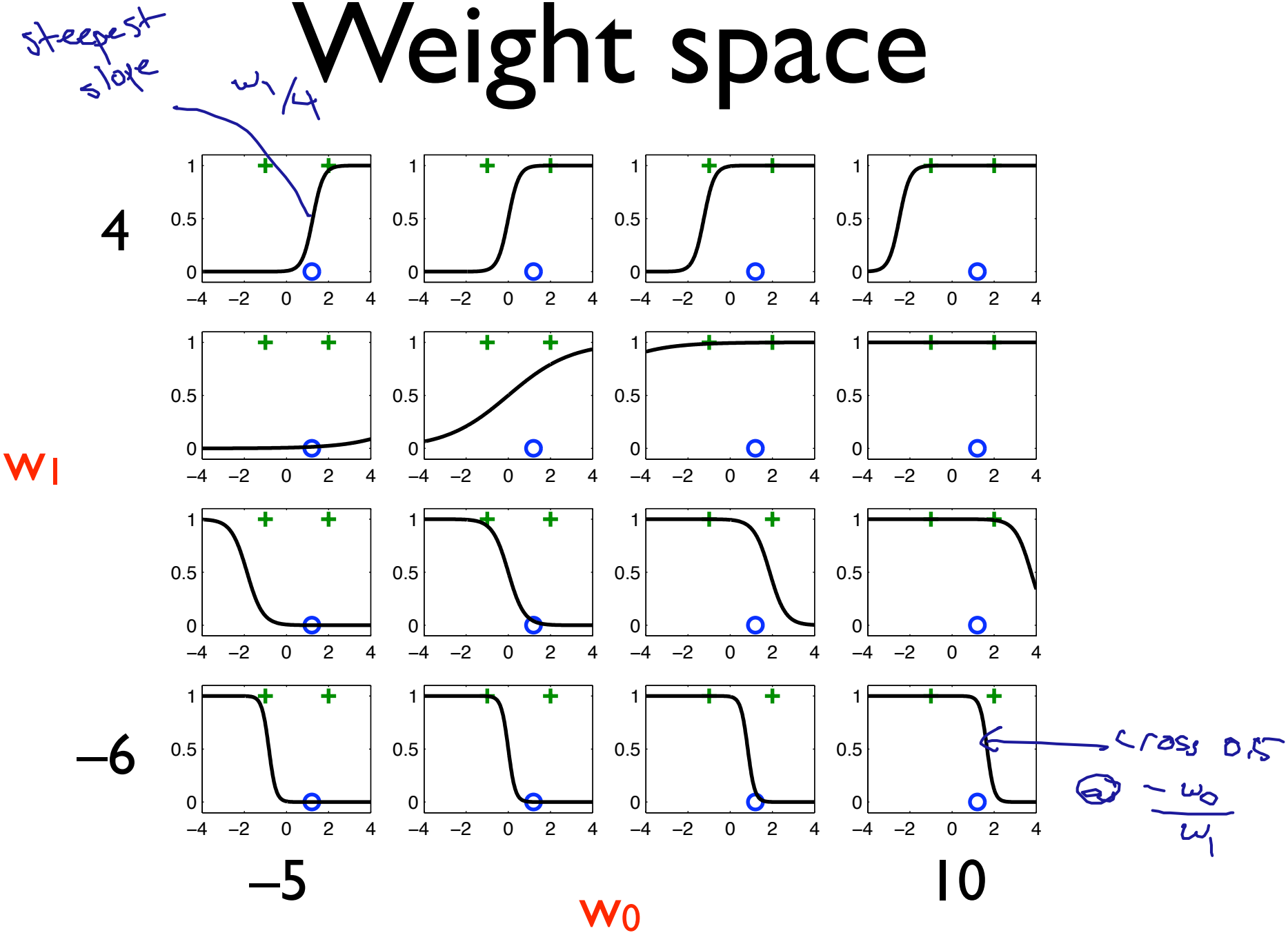
$\log(1+\exp(z))$

$\approx \max(0, z)$

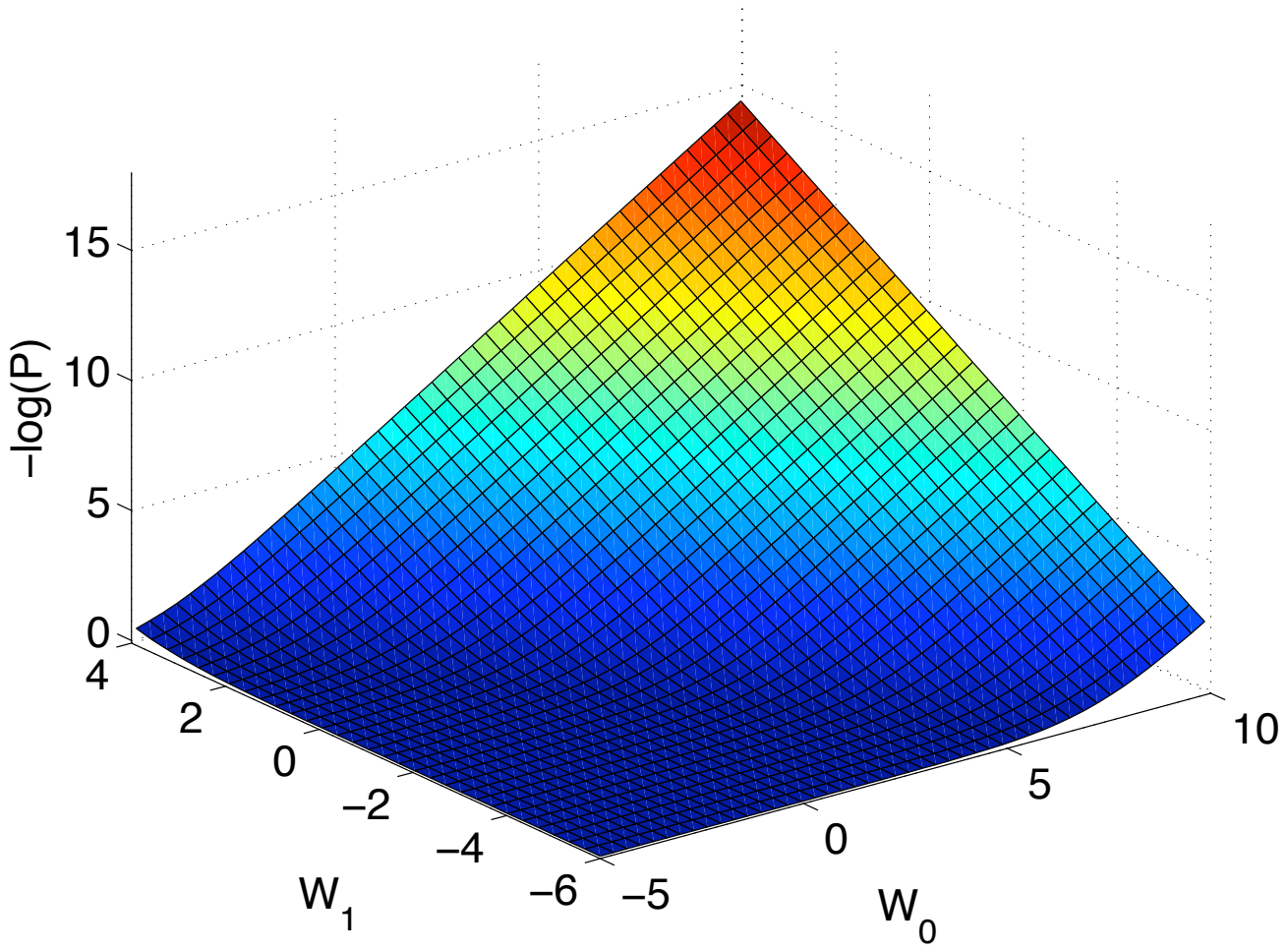
Example



Weight space



(X,Y) = (1.2, −1)

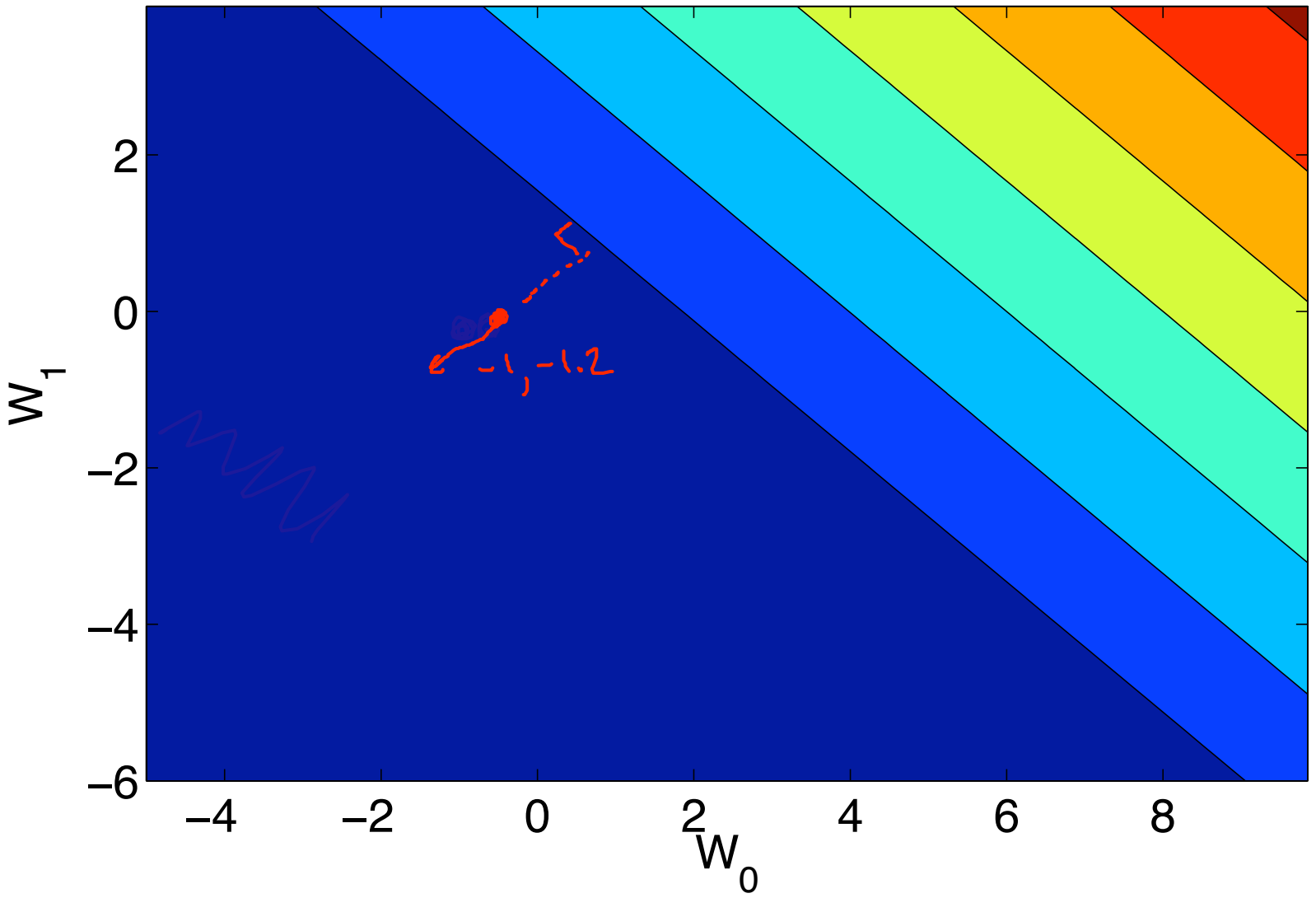


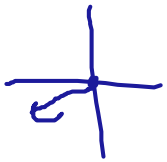
$\log \sigma(y \cdot w^T (1, x))$

$\begin{pmatrix} w_0 \\ w_1 \end{pmatrix} = w$

$(X, Y) = (1.2, -1)$

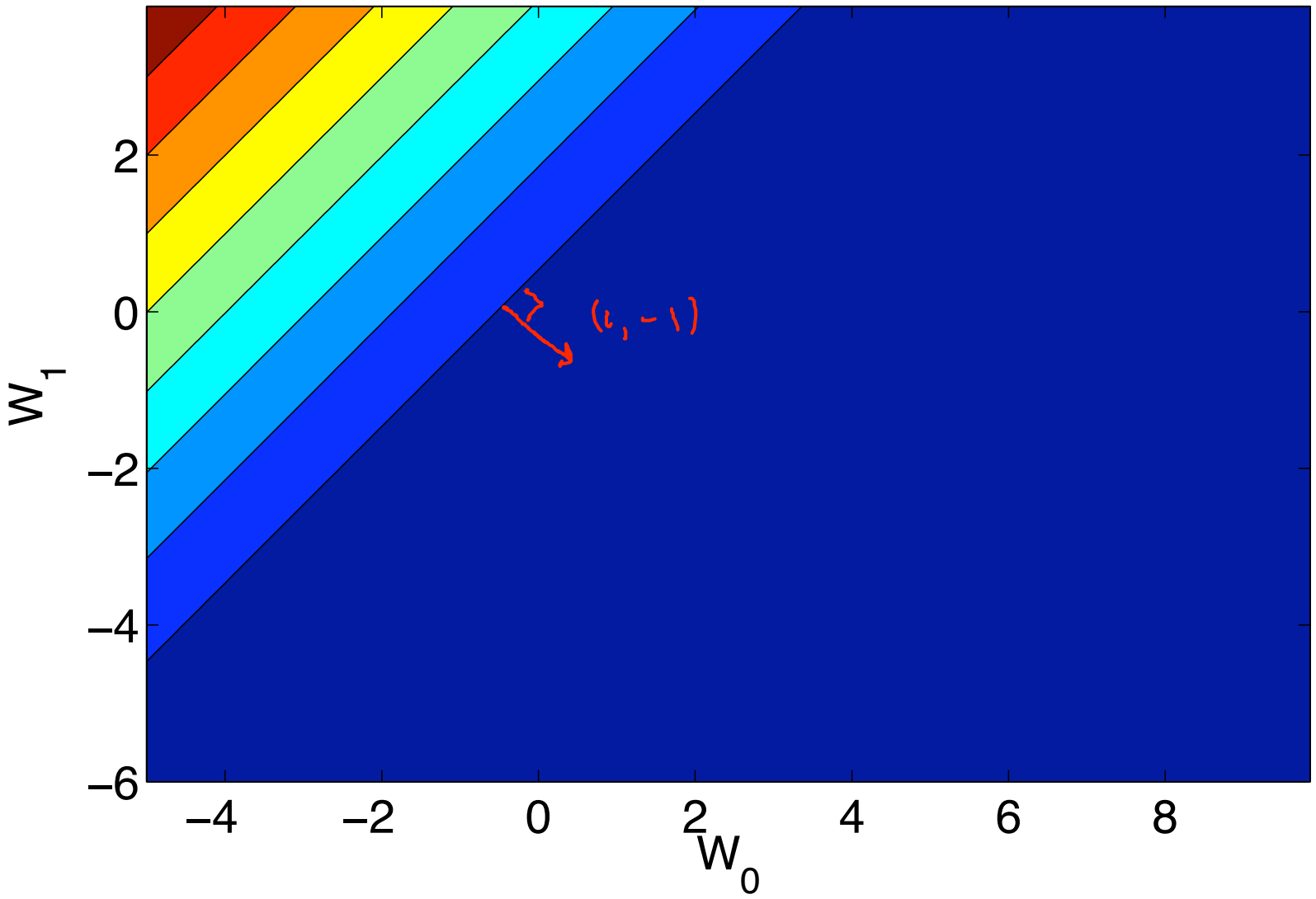
$-1 \begin{pmatrix} 1 \\ 1.2 \end{pmatrix} = \begin{pmatrix} -1 \\ -1.2 \end{pmatrix}$



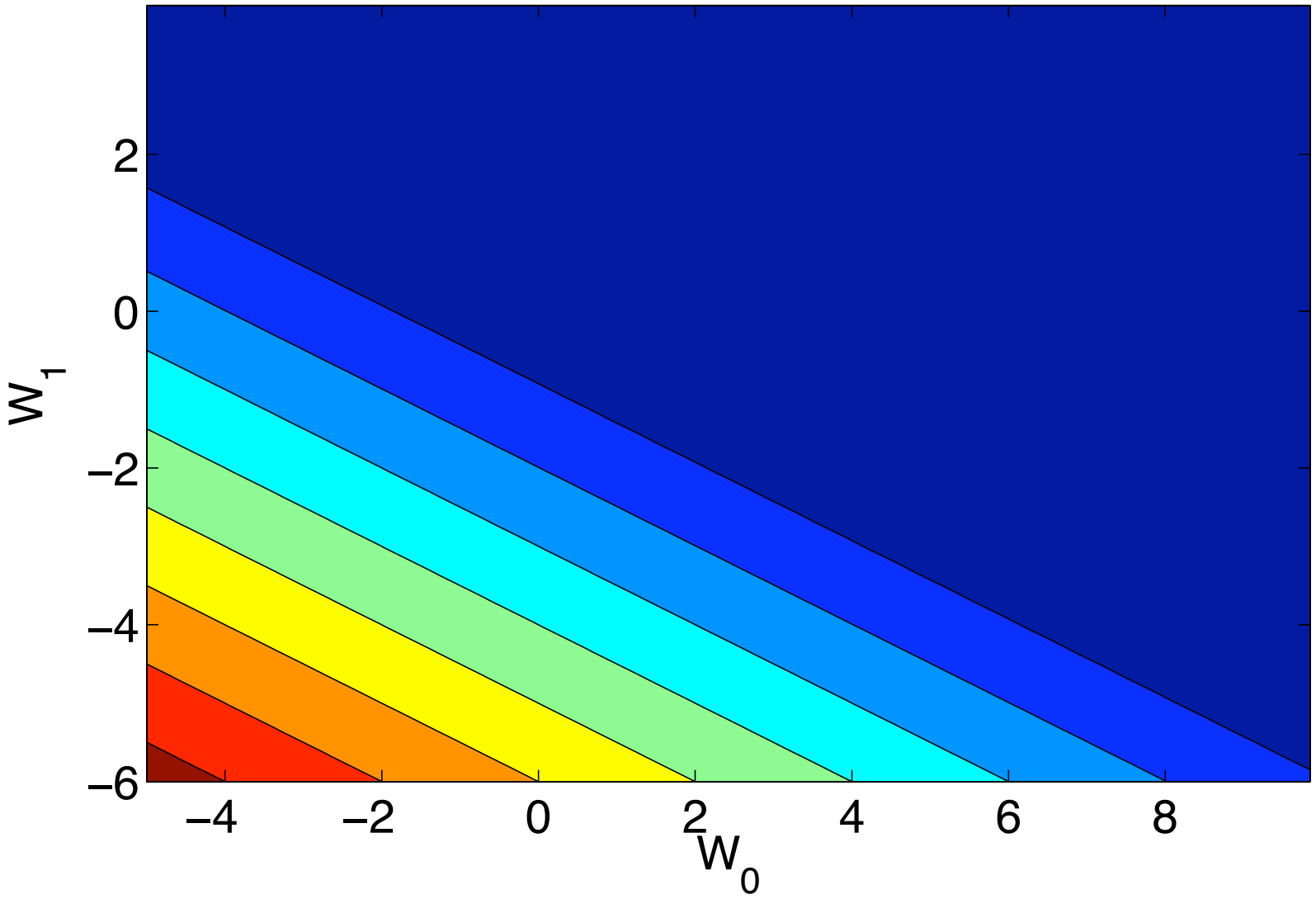


$y\begin{pmatrix}1\\x\end{pmatrix} = 1\begin{pmatrix}1\\-1\end{pmatrix} = \begin{pmatrix}1\\-1\end{pmatrix}$

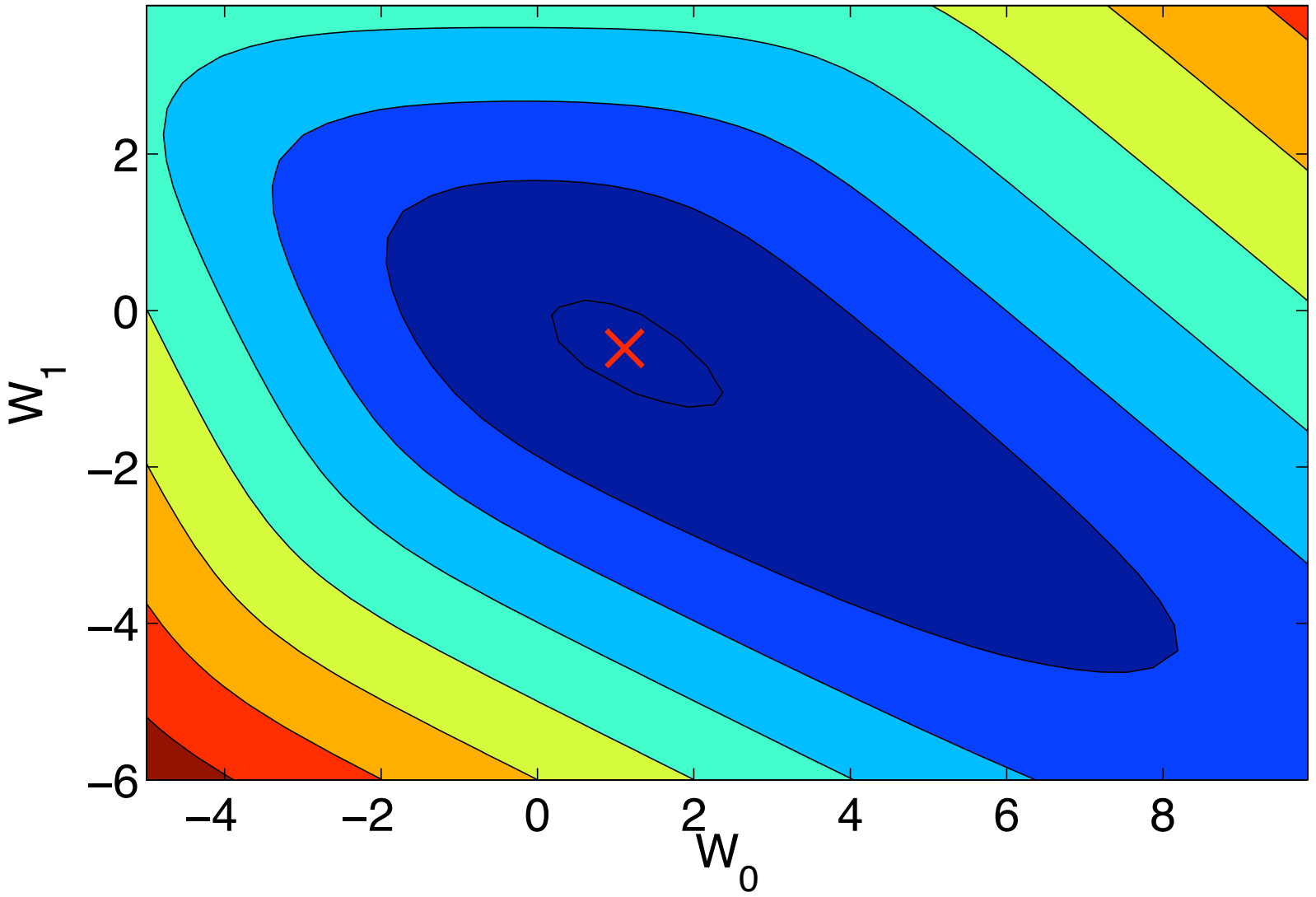
$(X, Y) = (-1, 1)$



(X,Y) = (2, 1)



$-\log(P(Y_{1..3} \mid X_{1..3}, W))$



$w_0^* = 1.12$

$w_1^* = -.45$