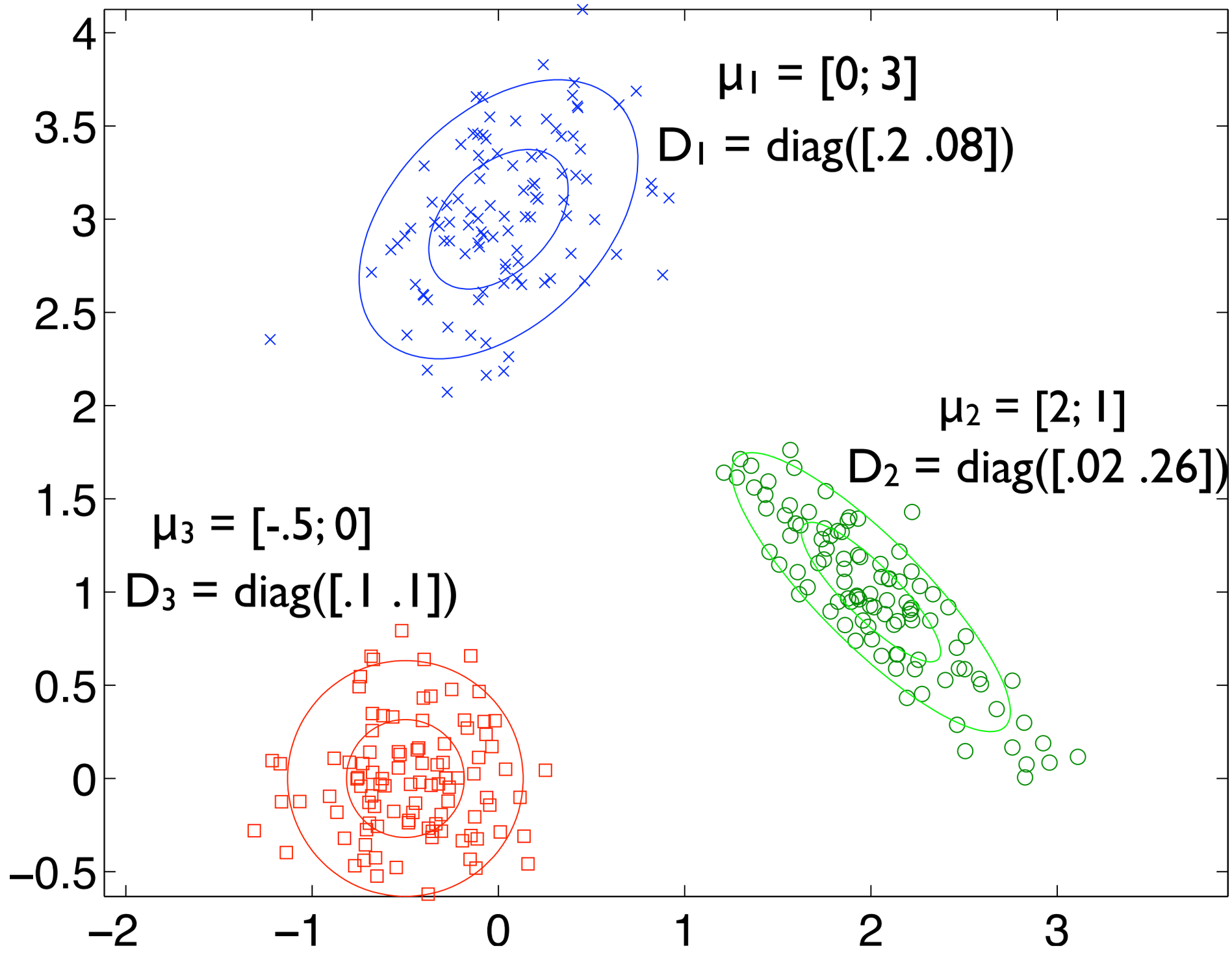
Review

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

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

Multivariate Gaussians

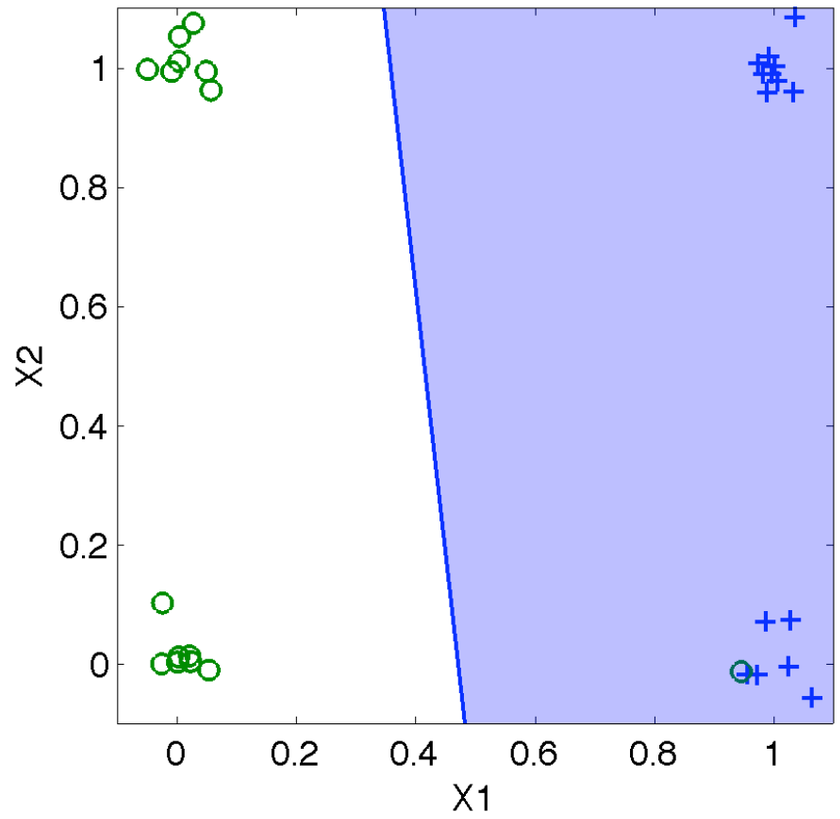


$$\Sigma_i = U D_i U^T$$

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

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

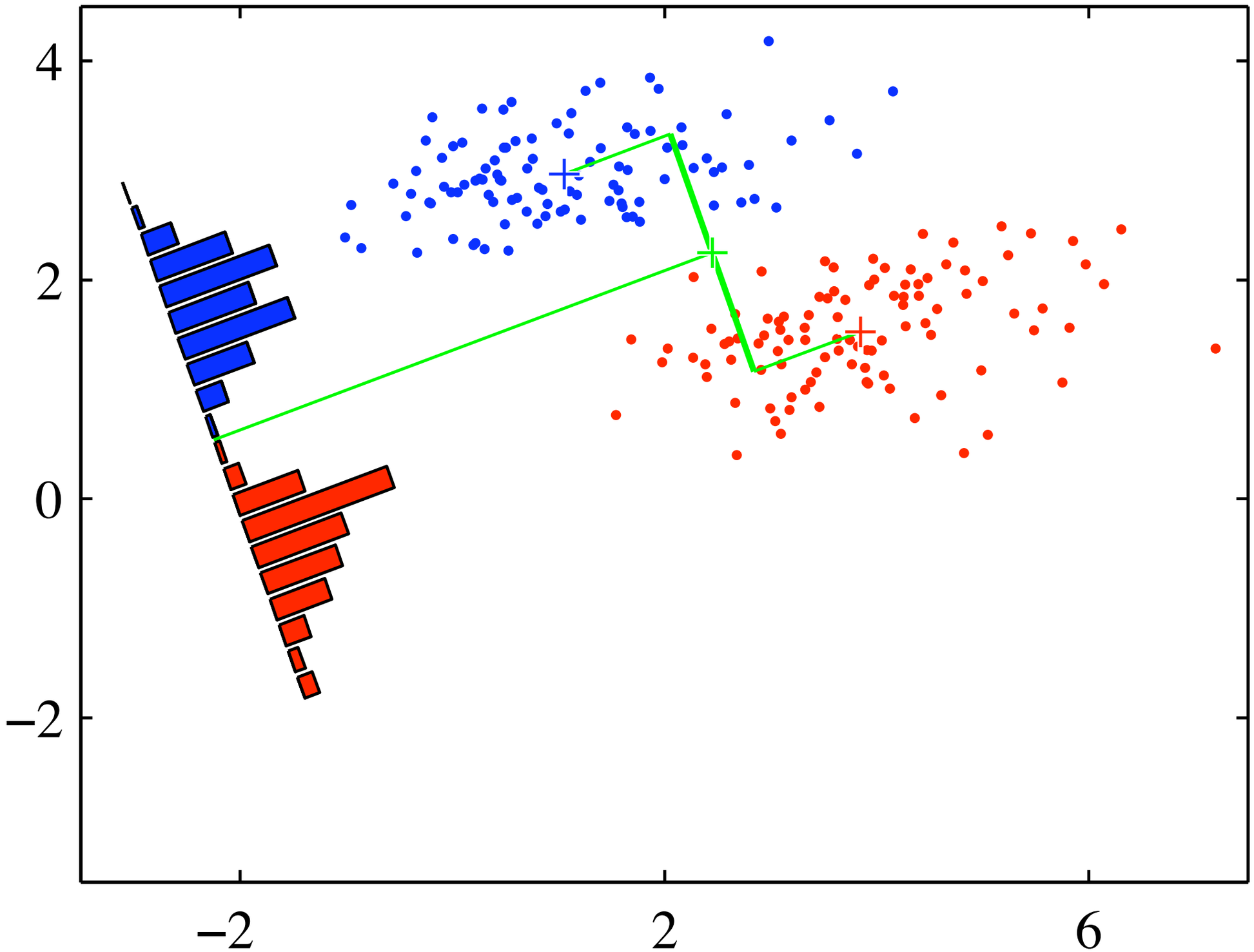


figure from book

Fisher w/ bad Σ

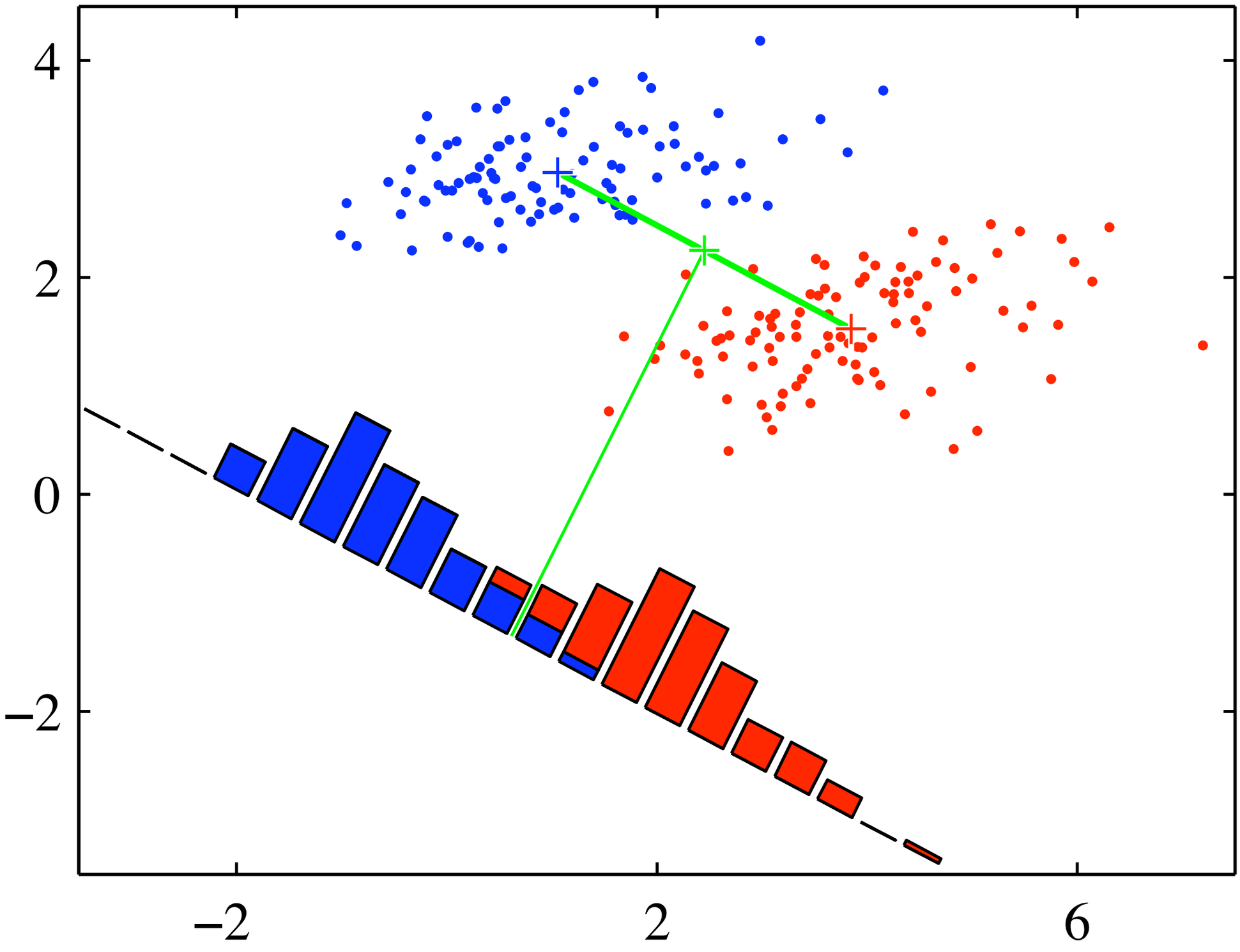
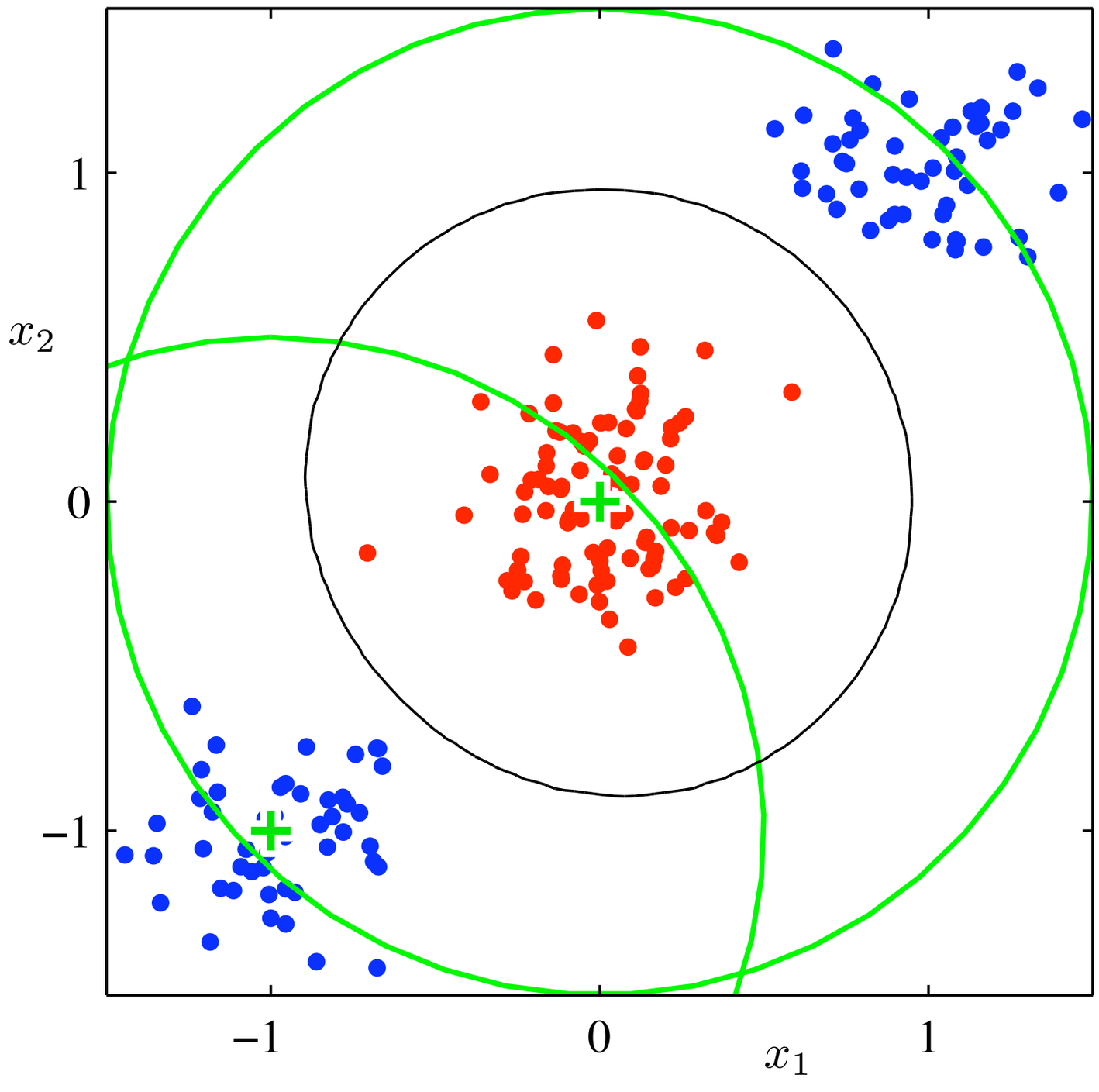


figure from book

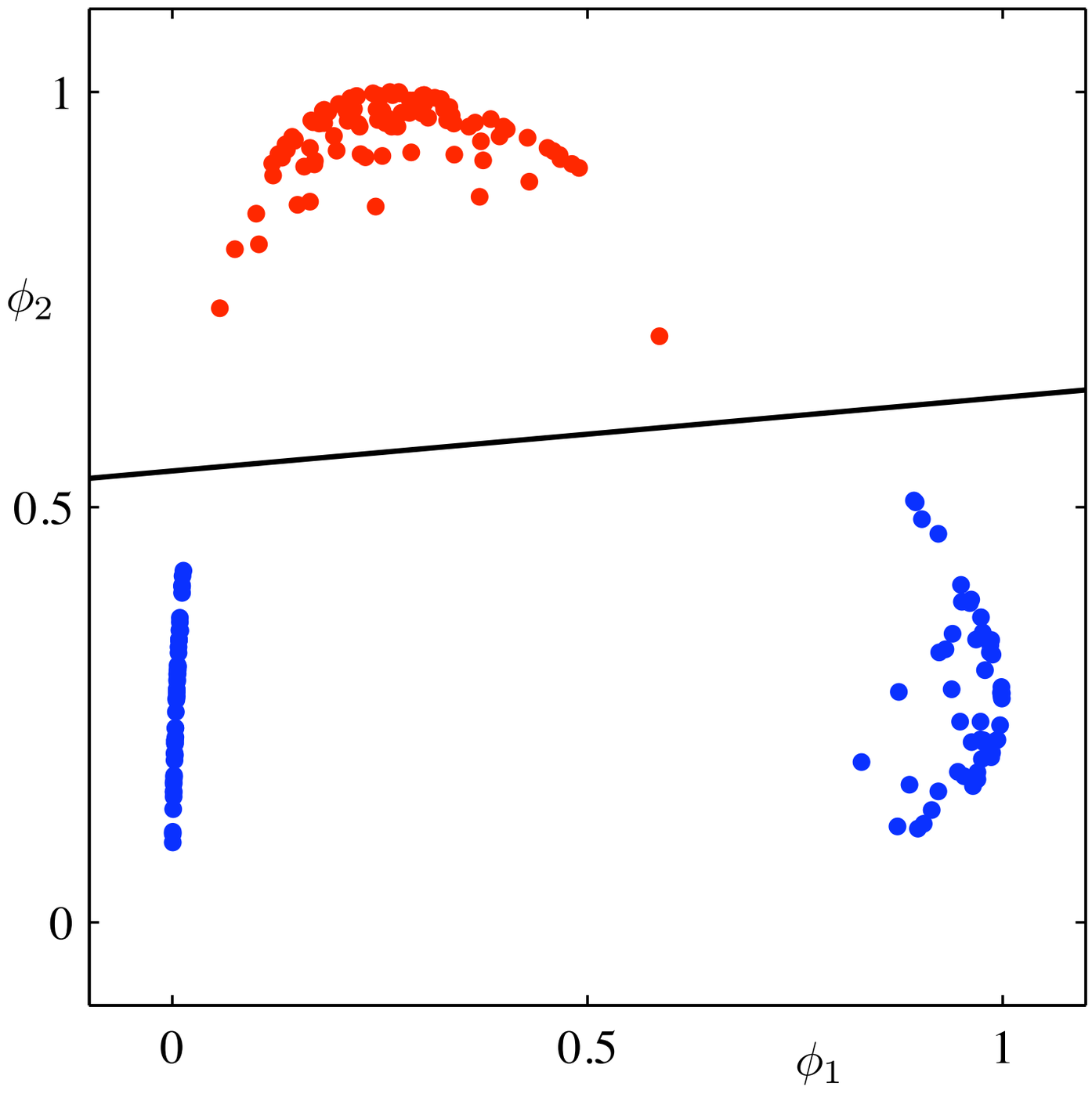
Features

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

Use of $\phi_j(X)$



Use of $\phi_j(X)$



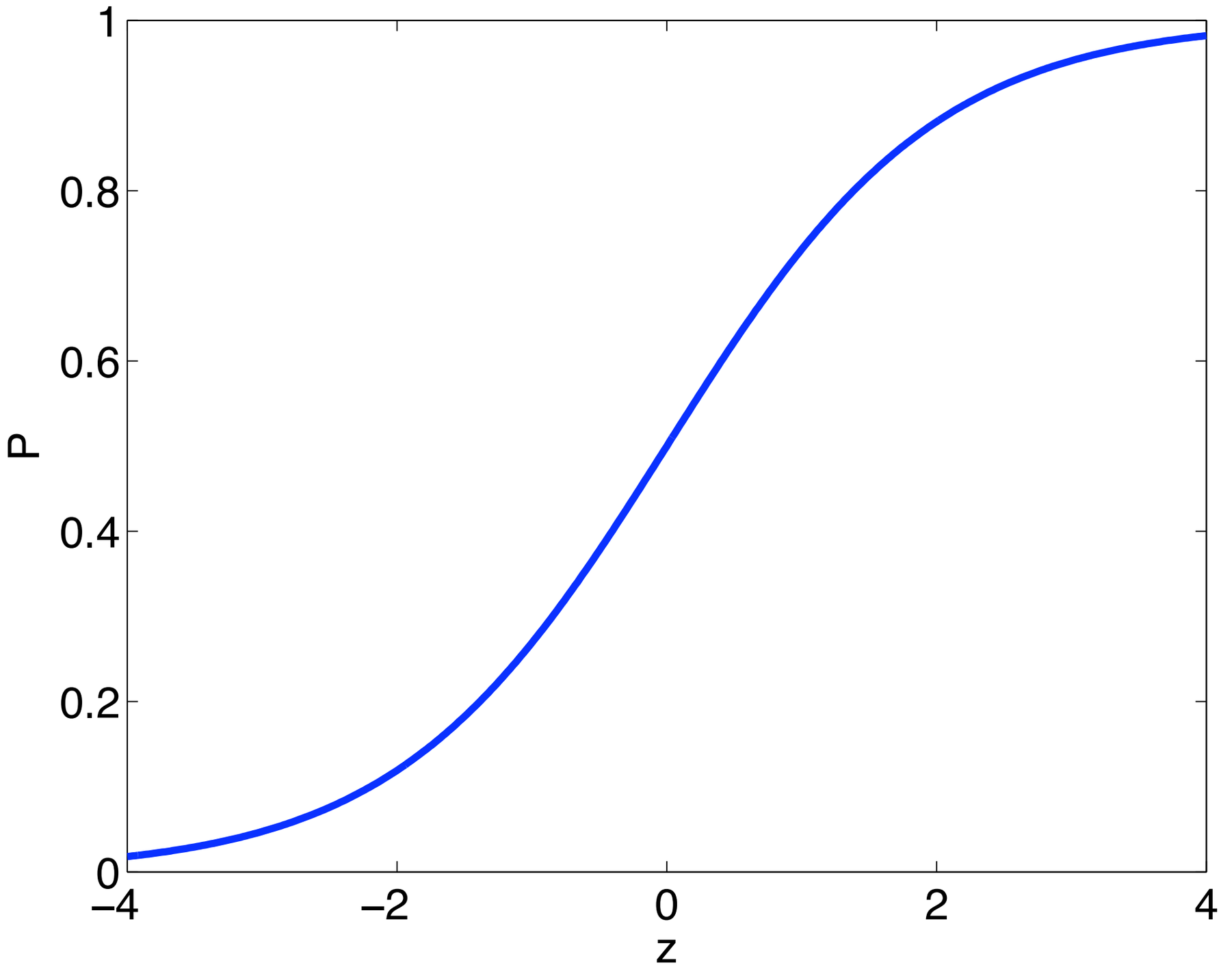
Lots of discriminants

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

Class probability

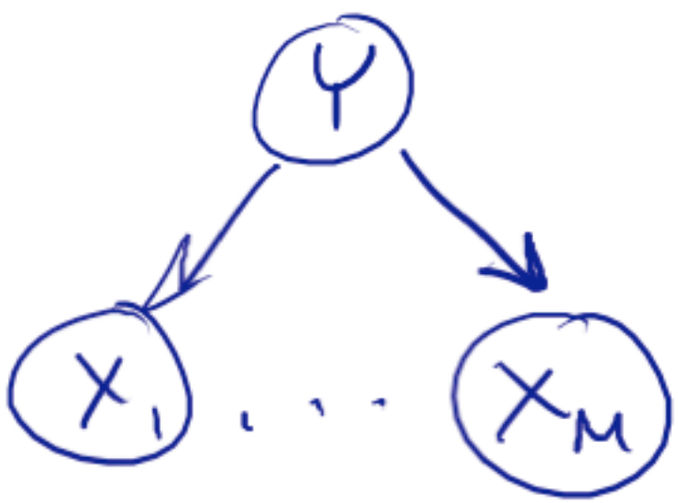
- We showed:
  - log P(Y=1 | X) – log P(Y=0 | X) =
- This implies
  - log P(Y = 1) =

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



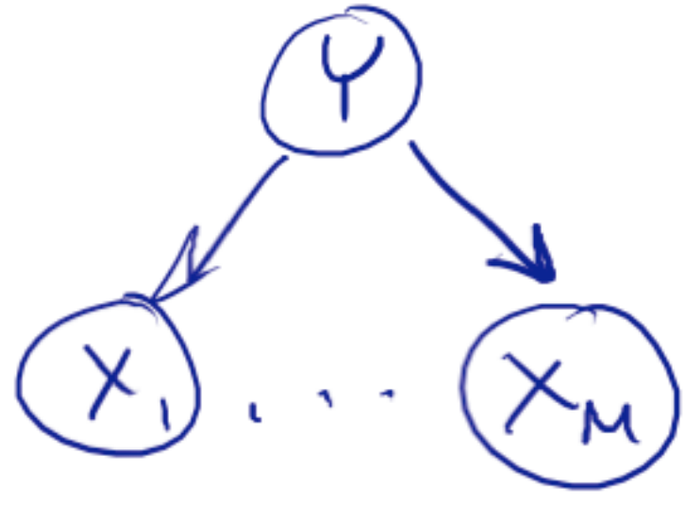
NB = MLE (or MAP)

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

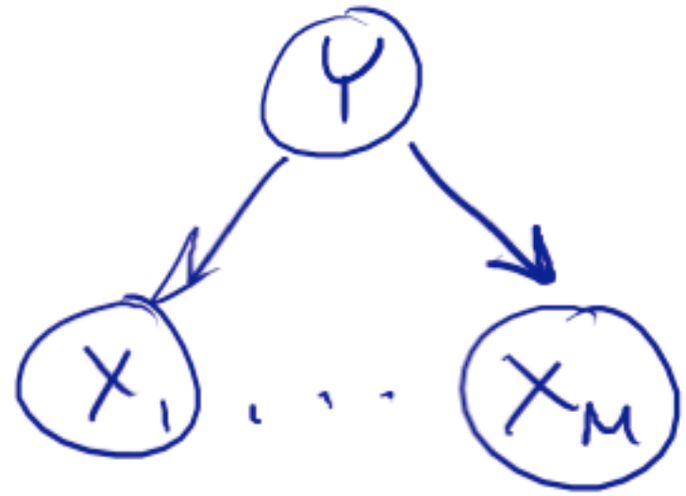


Conditional likelihood

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



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:

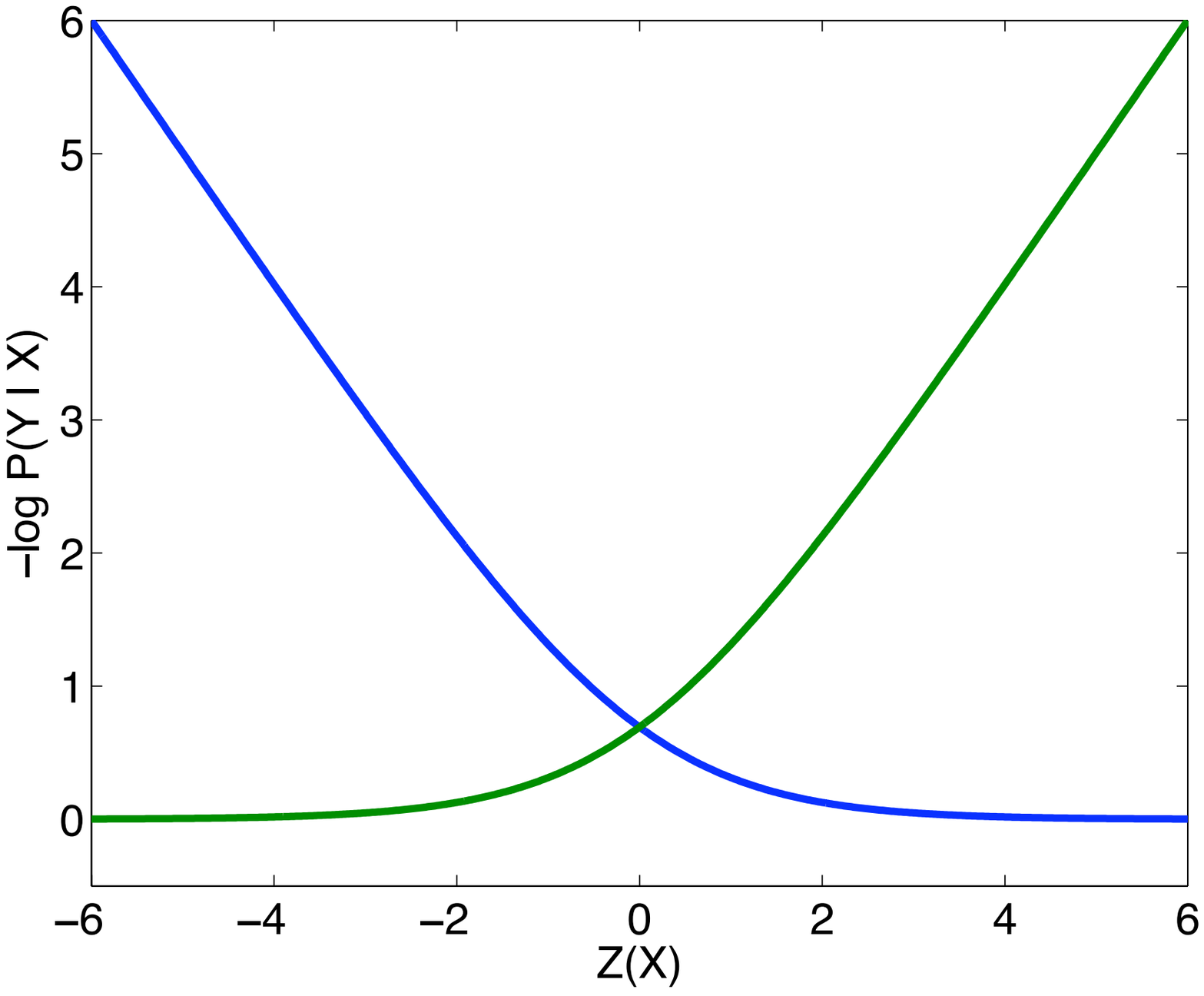
Generative vs discriminative

- 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
- Can improve performance, but also more risk of overfitting

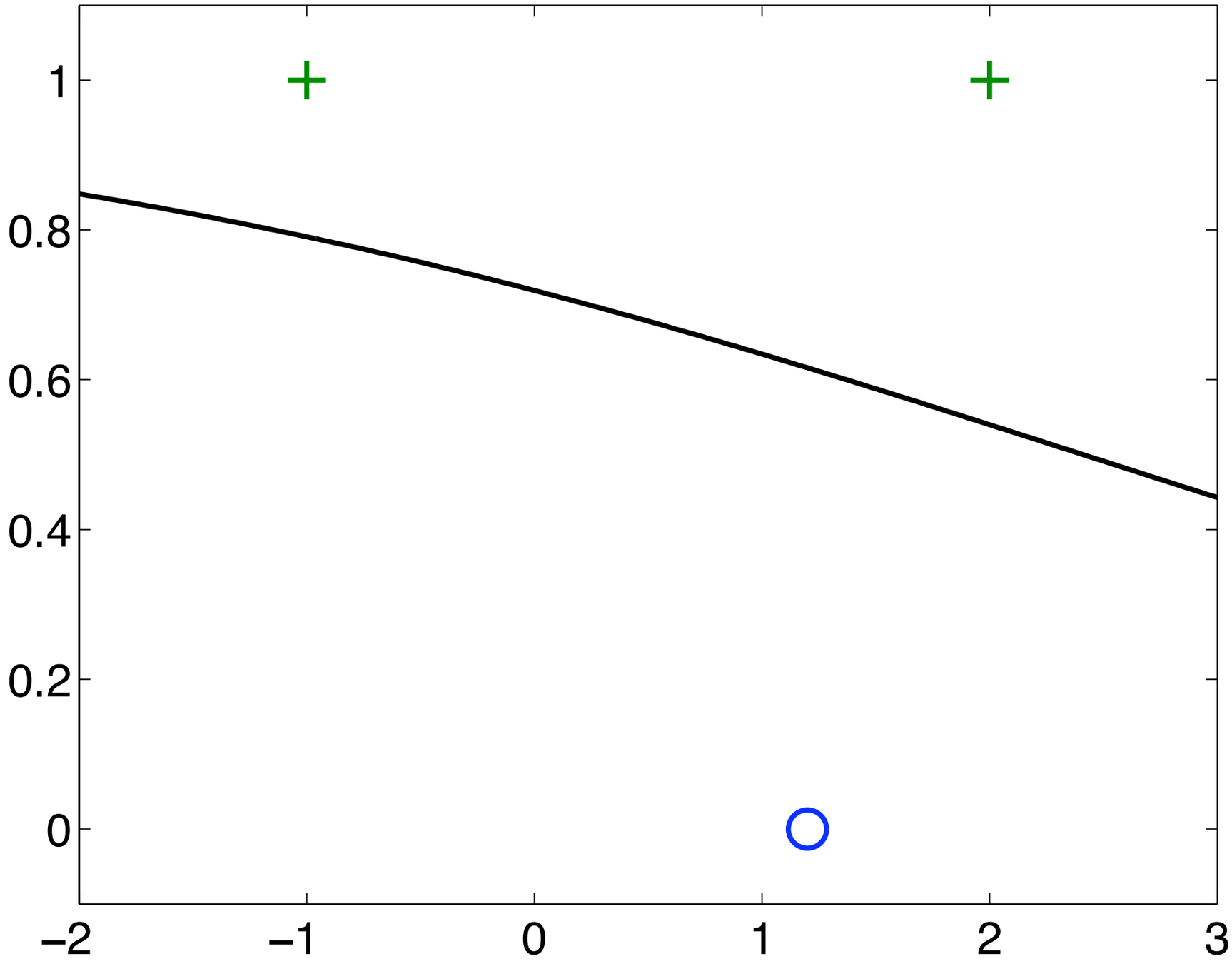
Logistic regression

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

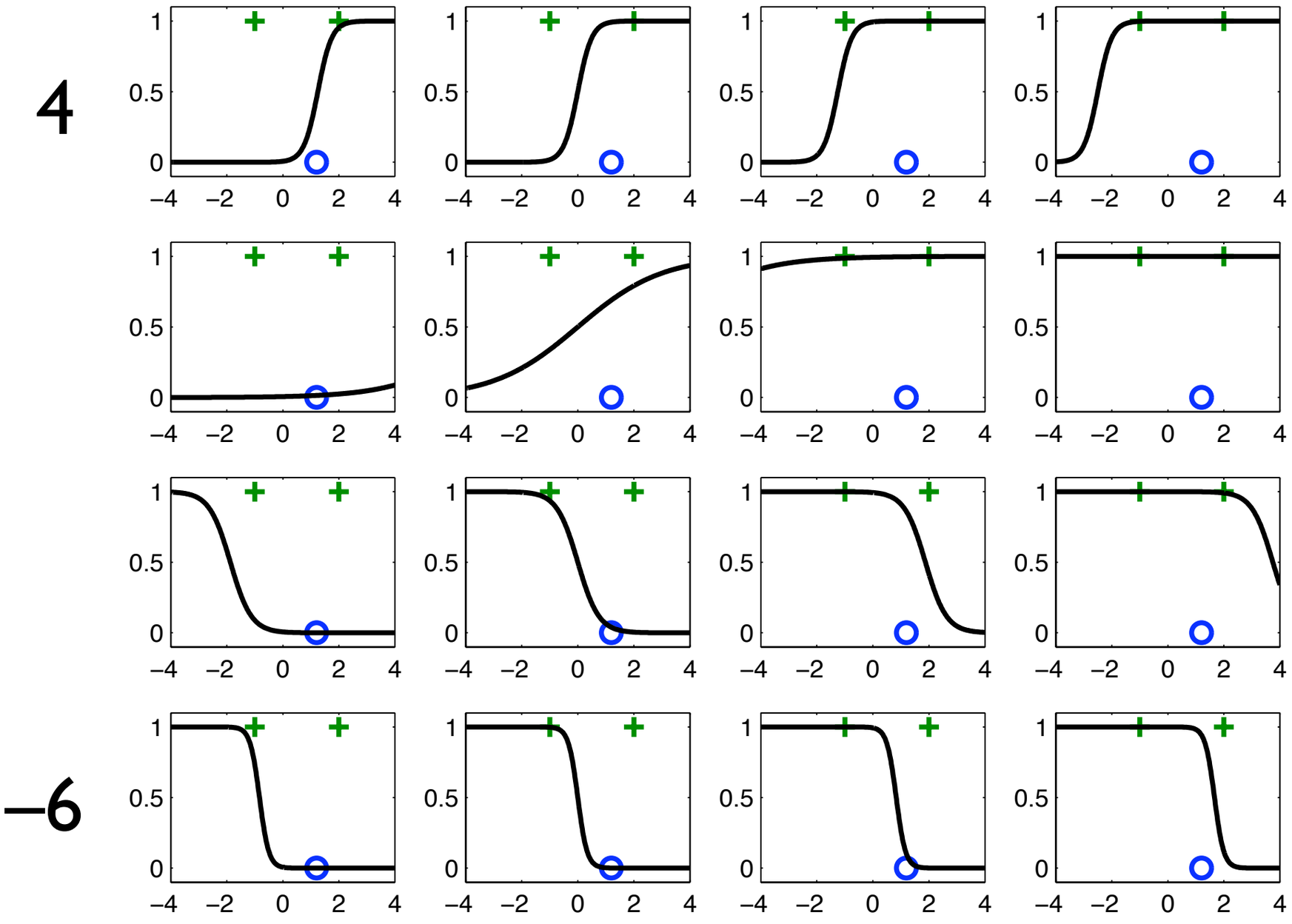
Neg. log likelihood



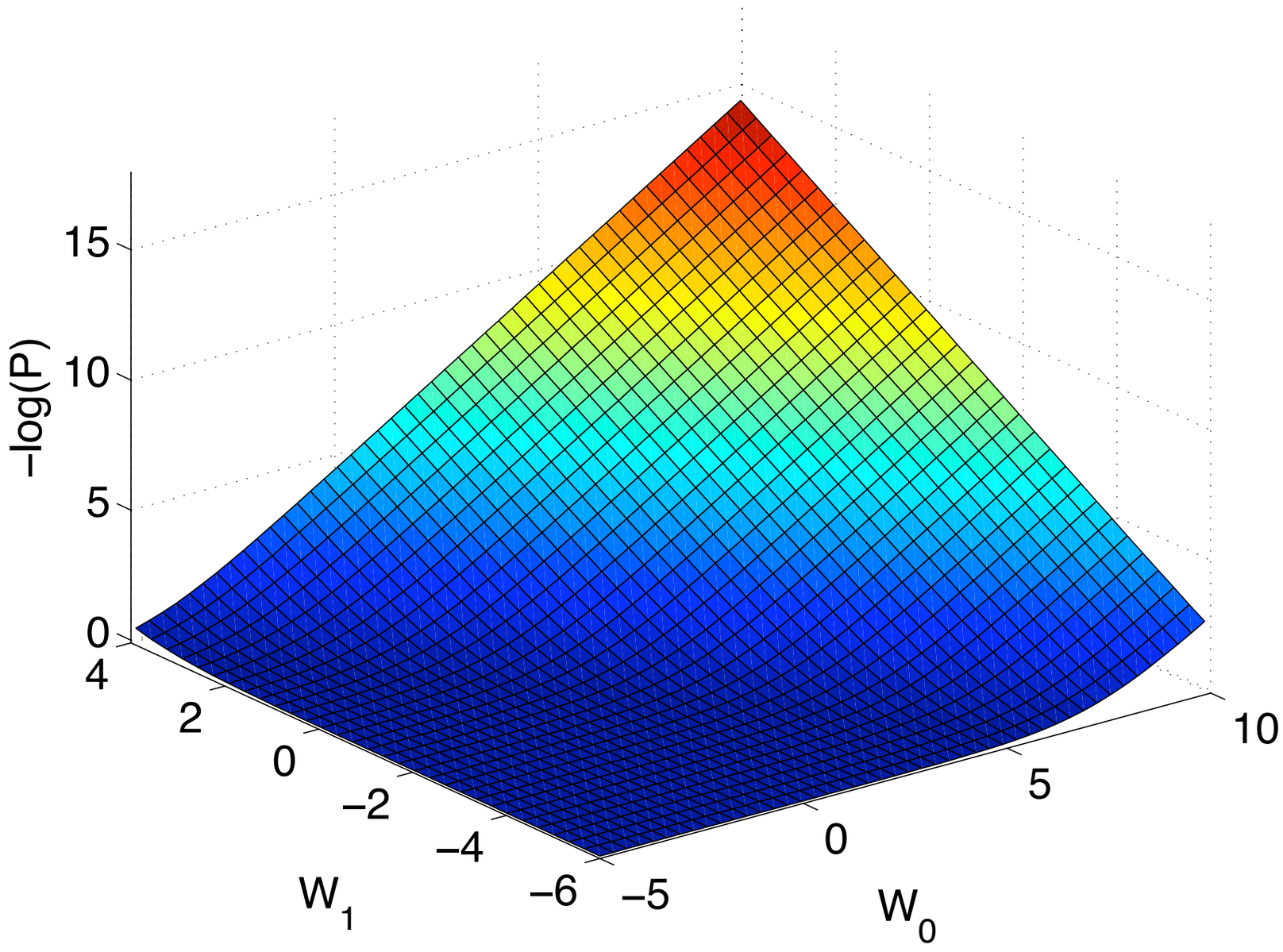
Example



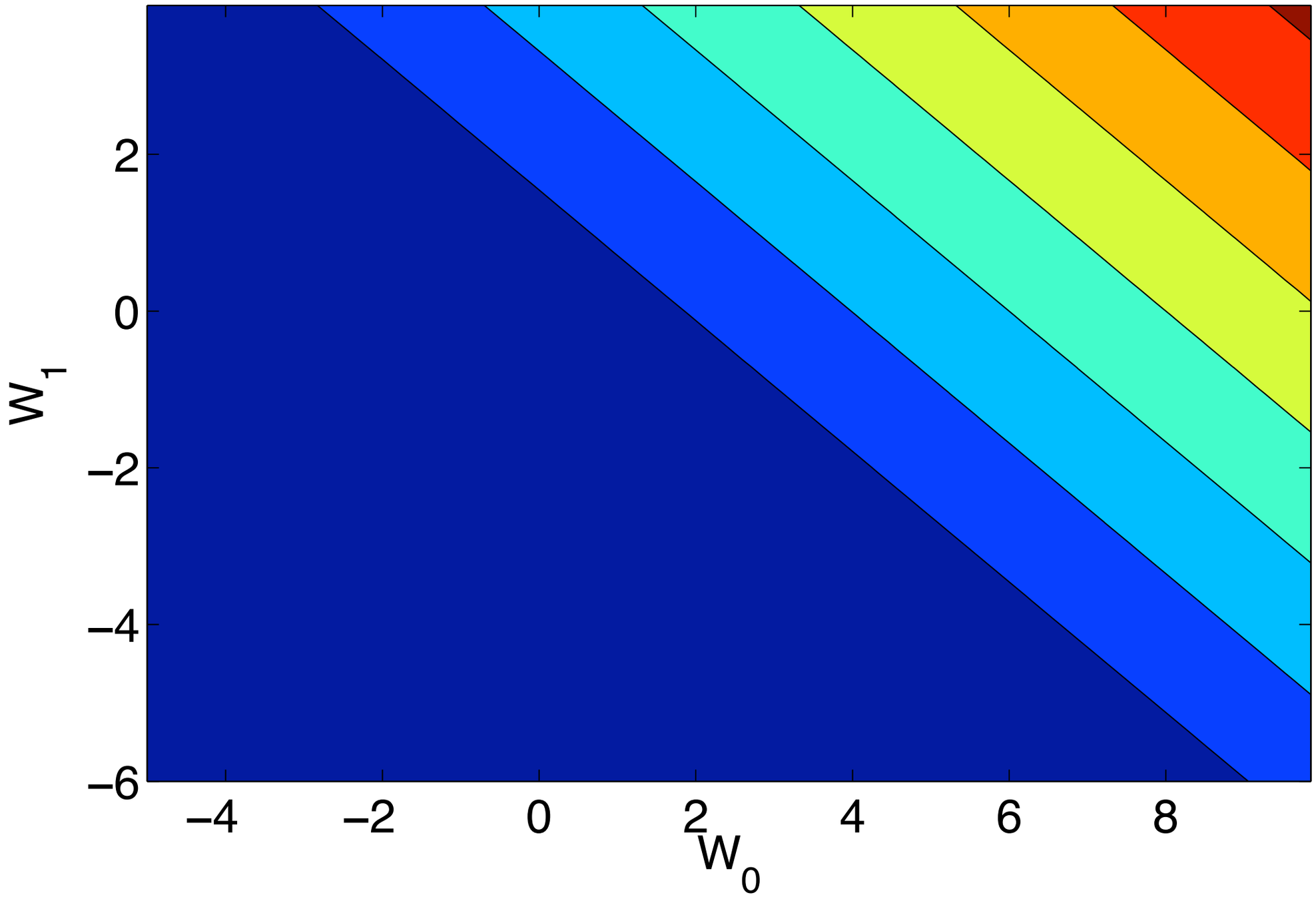
Weight space



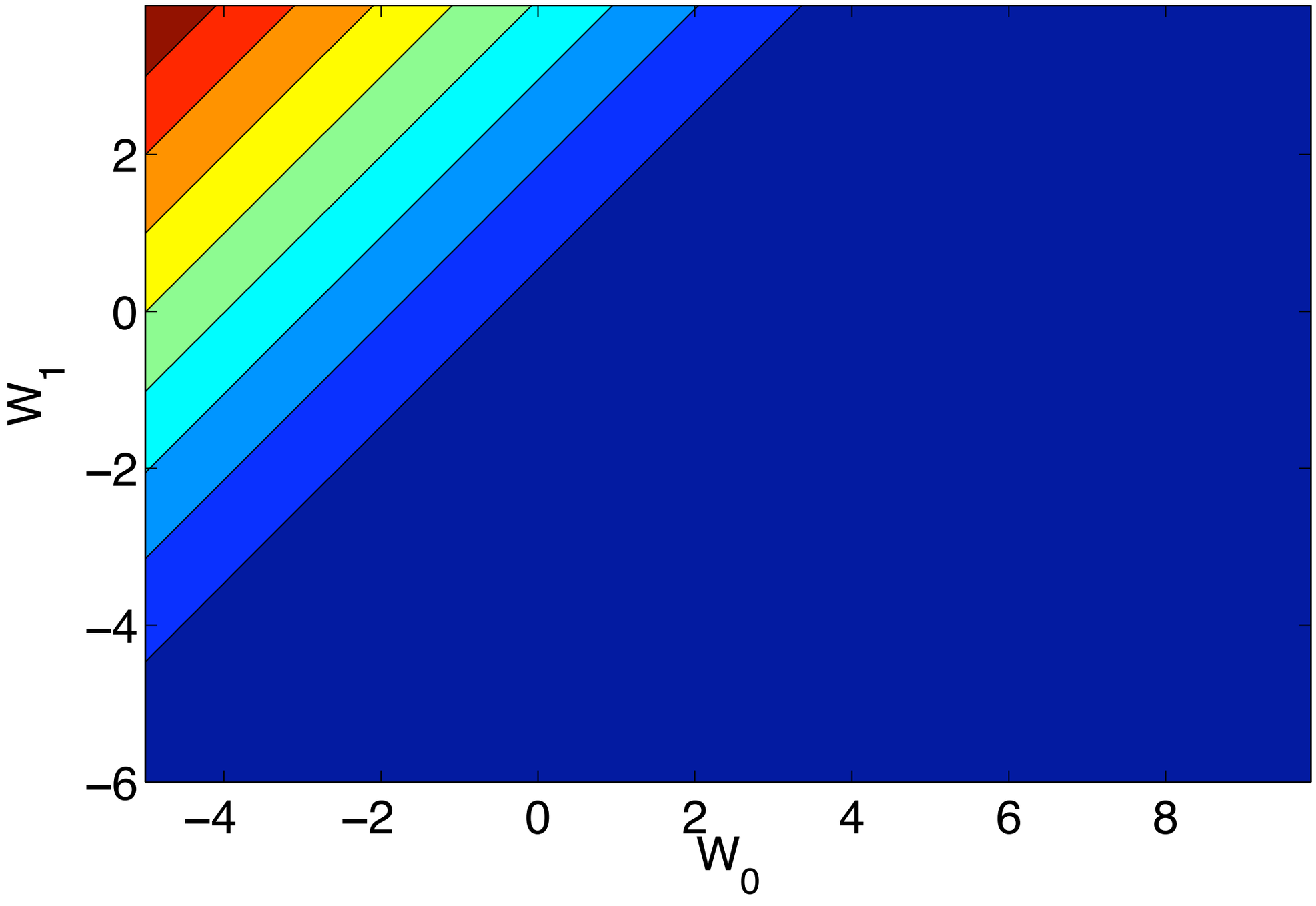
$$(X, Y) = (1.2, -1)$$



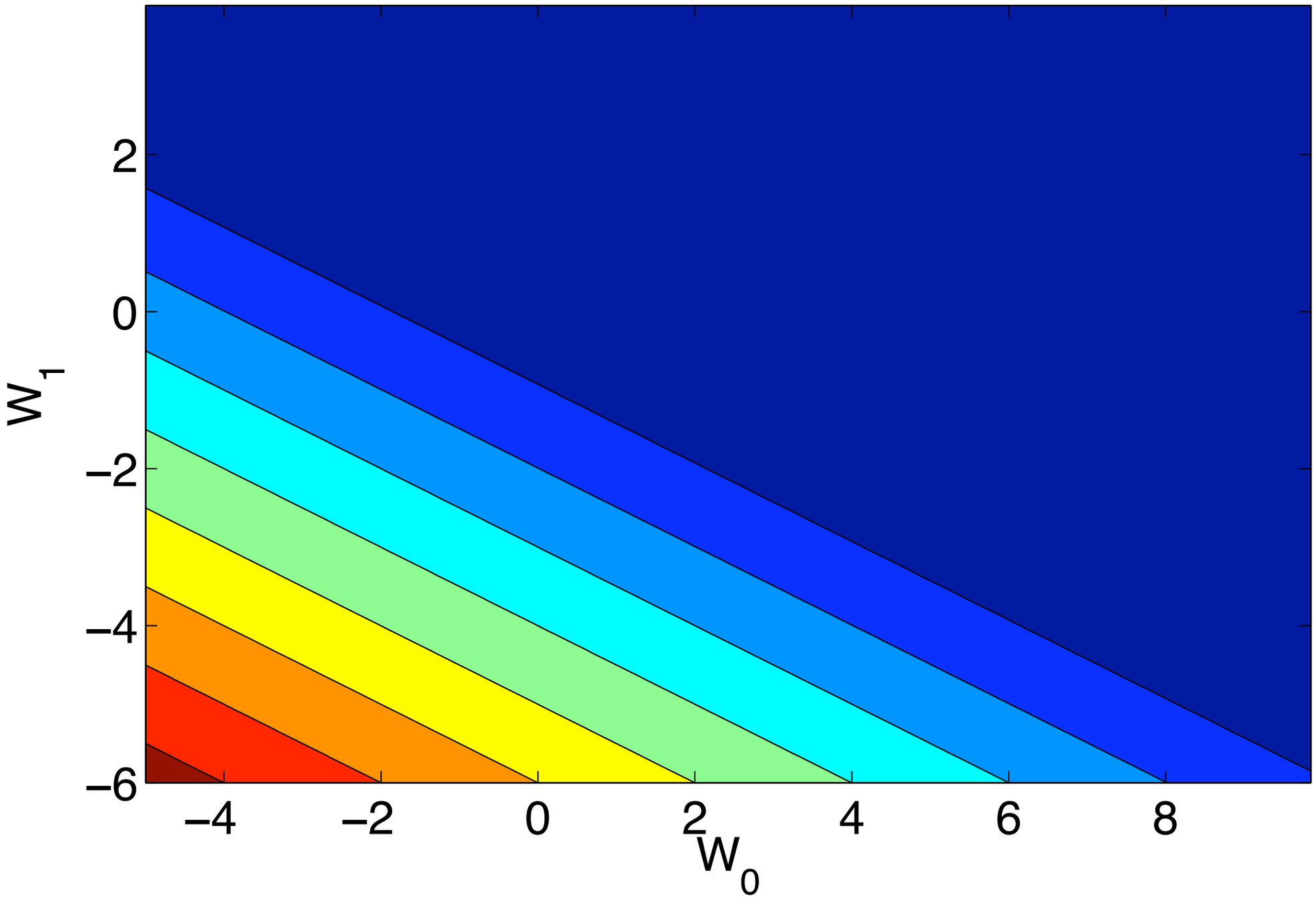
$$(X, Y) = (1.2, -1)$$



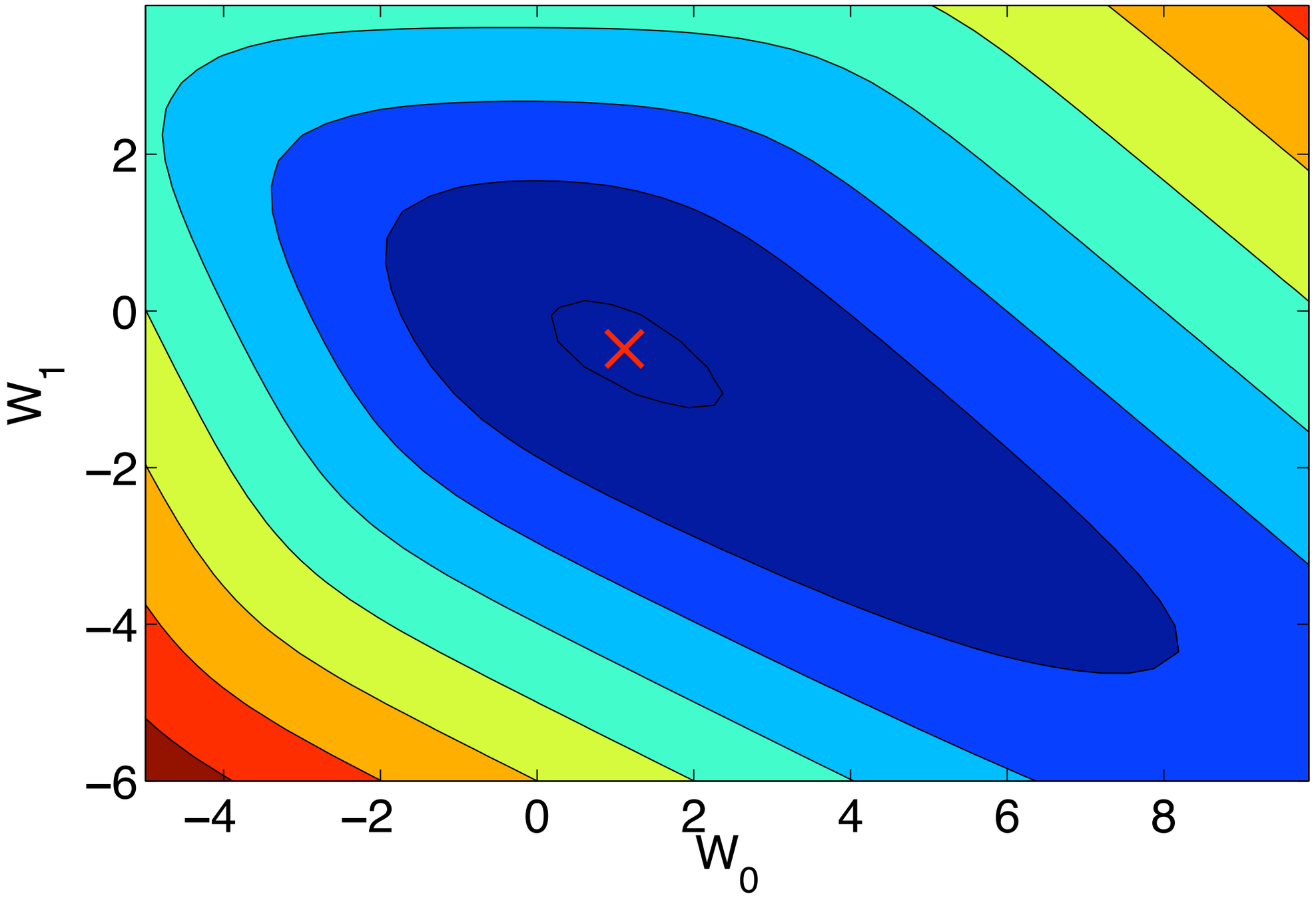
$$(X,Y) = (-1, 1)$$



(X,Y) = (2, 1)



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



Generalization: multiple classes

- One weight vector per class: $Y \in \{1,2,\ldots,C\}$
  - $P(Y=k) =$
  - $Z_k =$
- In 2-class case:

Multiclass example

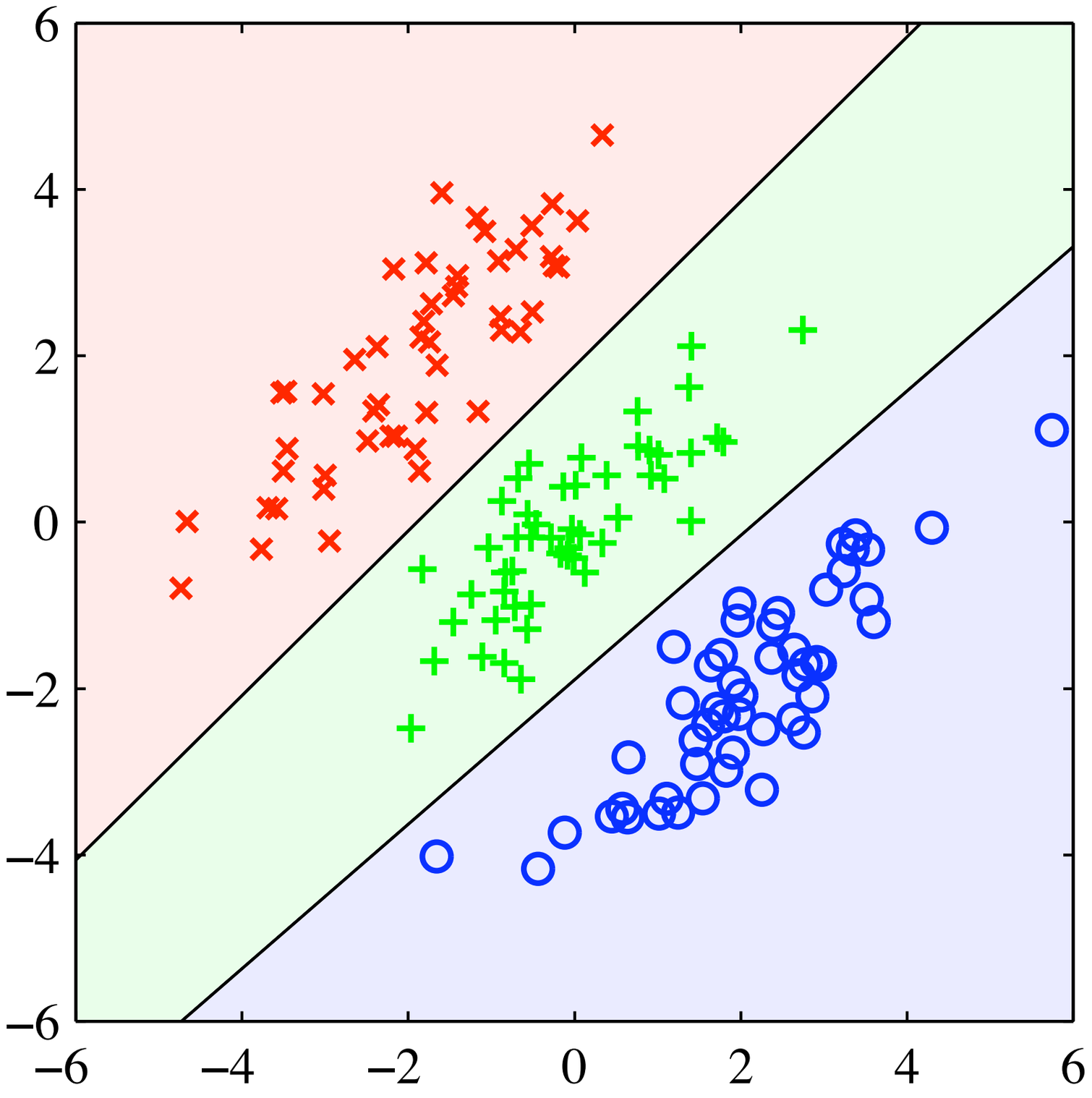


figure from book

Conditional MAP logistic regression

- P(Y | X, W) =
  - Z =

- As in linear regression, can put prior on W
  - common priors: L_2 (ridge), L_1 (sparsity)

- $\max_w$ P(W=w | X, Y)

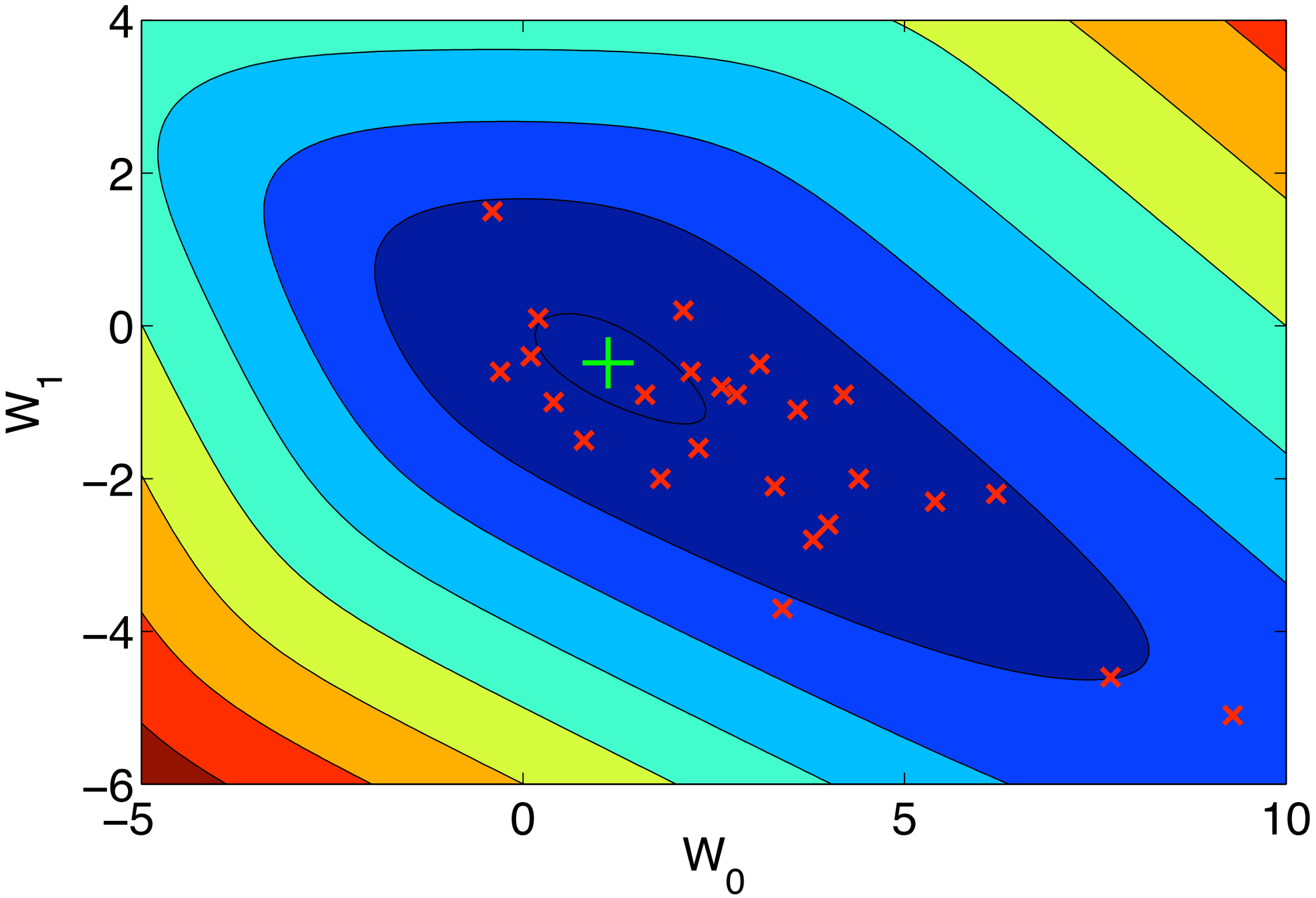
Software

- Logistic regression software is easily available: most stats packages provide it
  - e.g., `glm` function in R
  - or, http://www.cs.cmu.edu/~ggordon/IRLS-example/
- Most common algorithm: Newton's method on log-likelihood (or L_2-penalized version)
  - called "iteratively reweighted least squares"
  - for L_1, slightly harder (less software available)

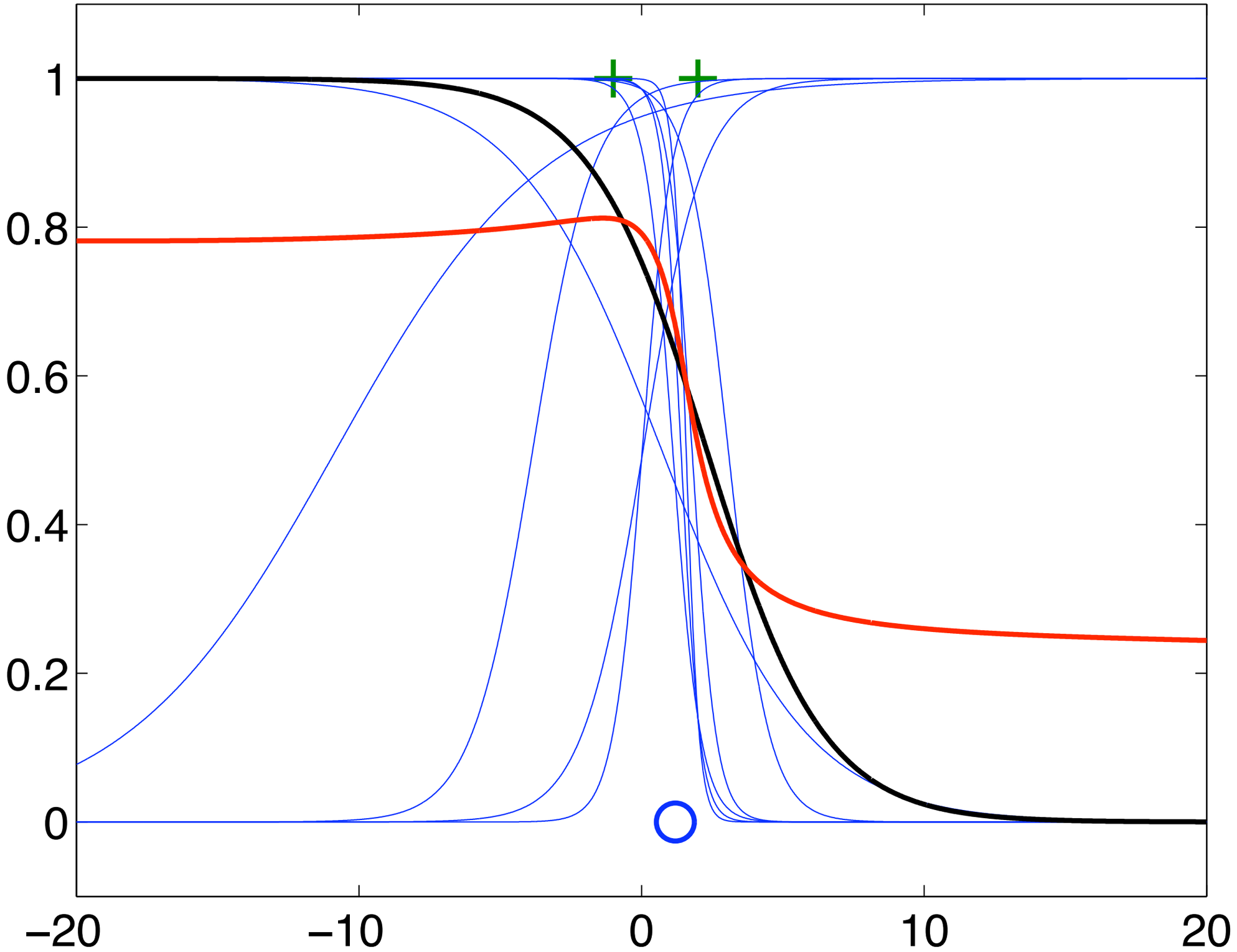
Bayesian regression

- In linear and logistic regression, we've looked at
  - conditional MLE: $\max_w P(Y \mid X, w)$
  - conditional MAP: $\max_w P(W=w \mid X, Y)$
- But of course, a true Bayesian would turn up nose at both
  - why?

Sample from posterior



Predictive distribution



Overfitting

- True Bayesian inference ***never*** leads to overfitting
  - may still lead to bad results for other reasons!
  - e.g., not enough data, bad model class, …
- Overfitting is an indicator that the MLE or MAP approximation is a bad one