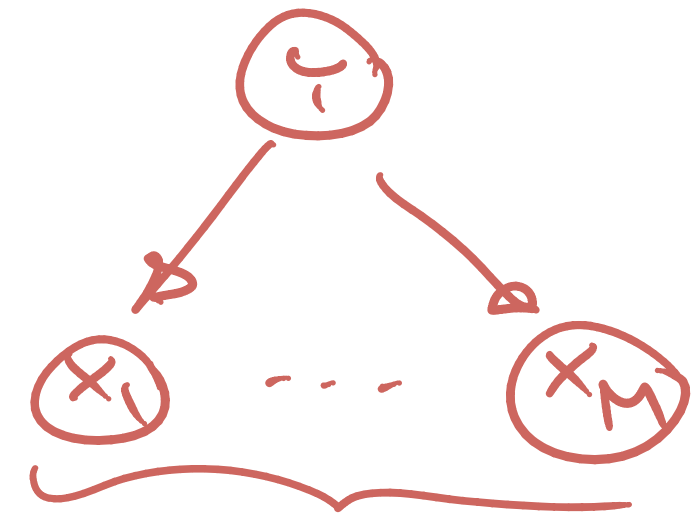
Review

- Train-test split
- Cross-validation
- Regularization and model complexity
  - L_1, L_2

Review



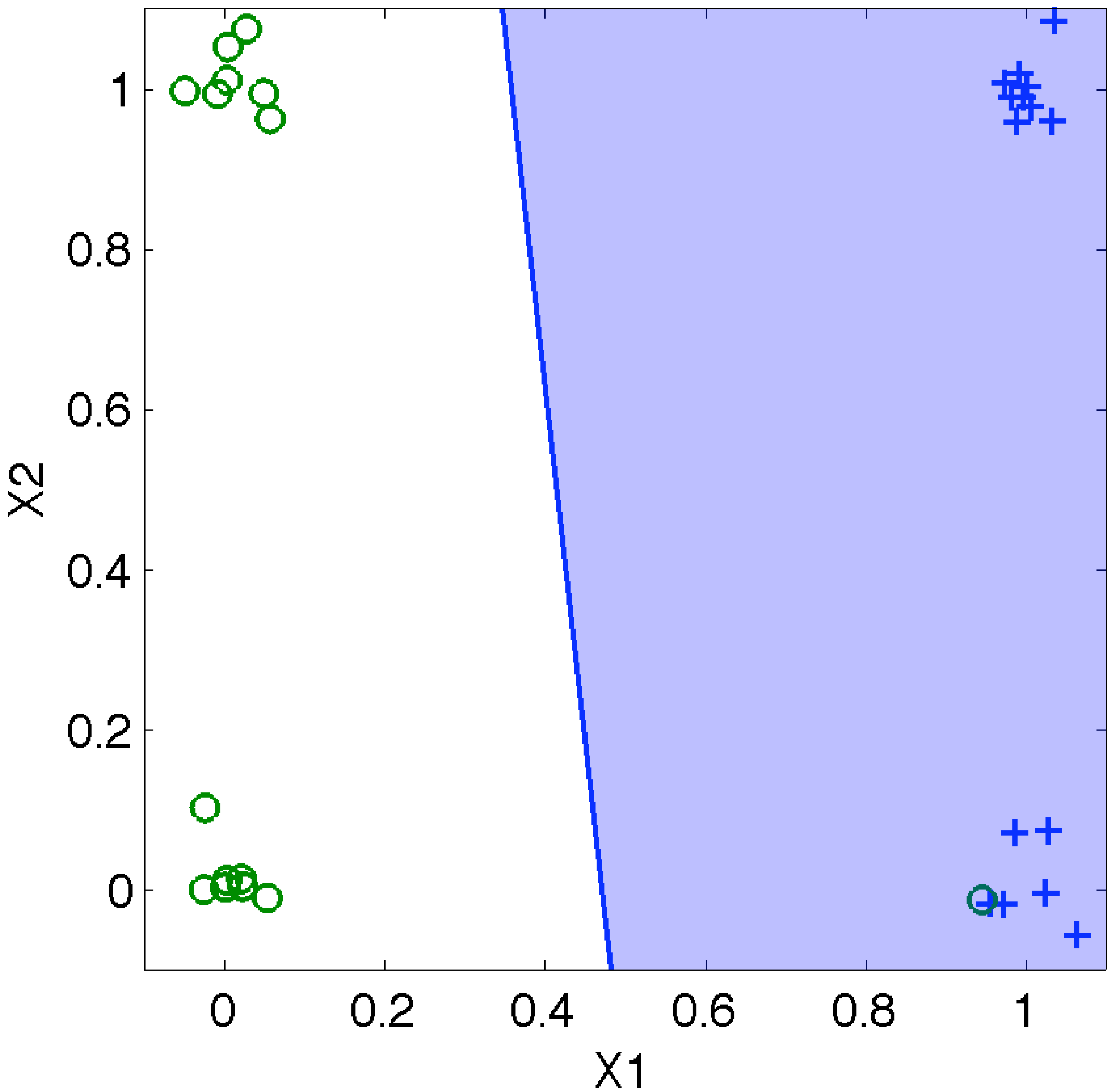
DIFFERENT FEATURES OF THE SAME EXAMPLE

- Classification w/ Naïve Bayes
- Features assumed independent given class
- Prediction: y = $\arg\max_{y} \left[p(y) \prod_{j} (x_j \mid y) \right]$
- Variations: Bag of Words, Gaussian NB

A closer look at NB

- $Y = 1$ if:

Linear discriminants



Geometry of a discriminant

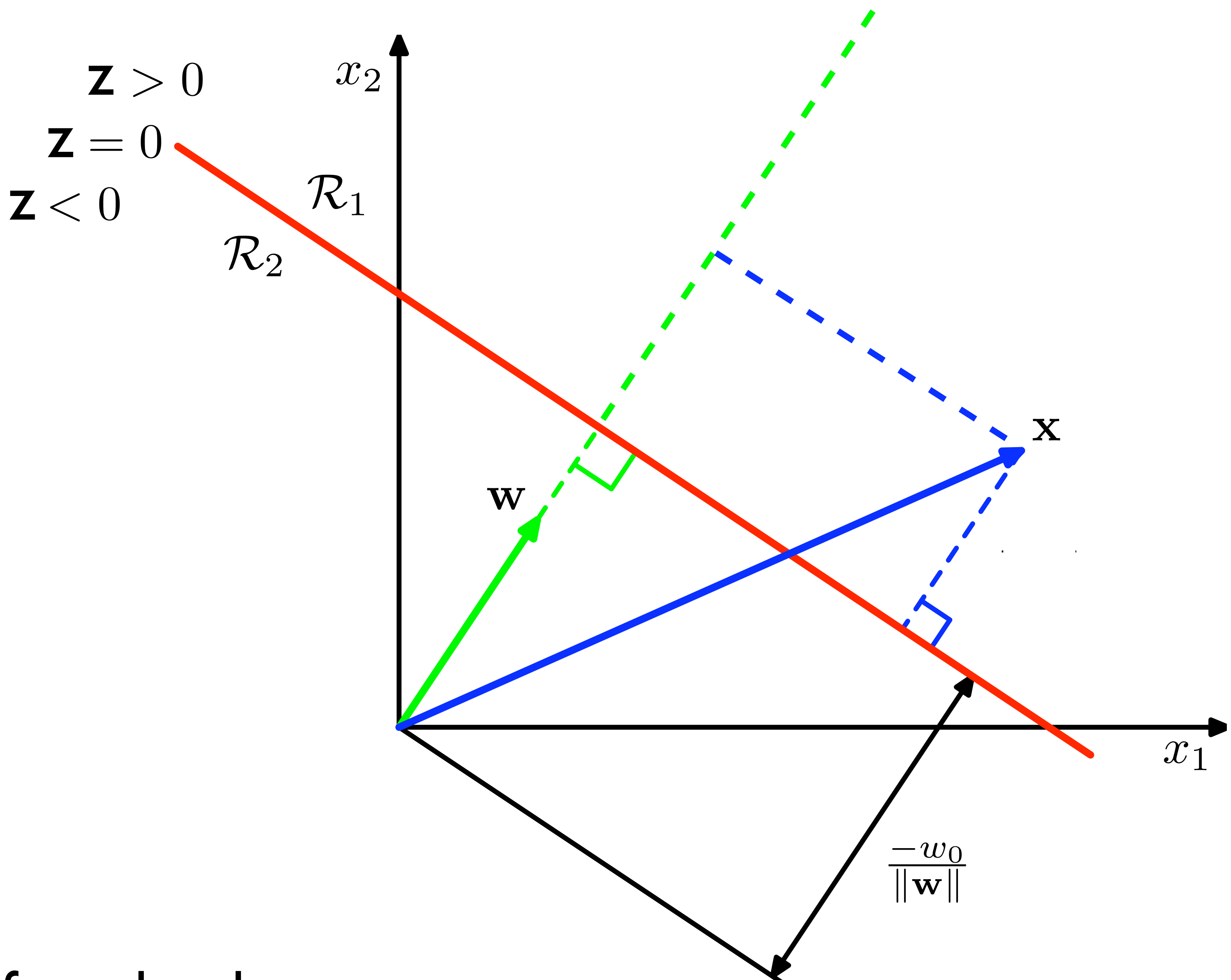
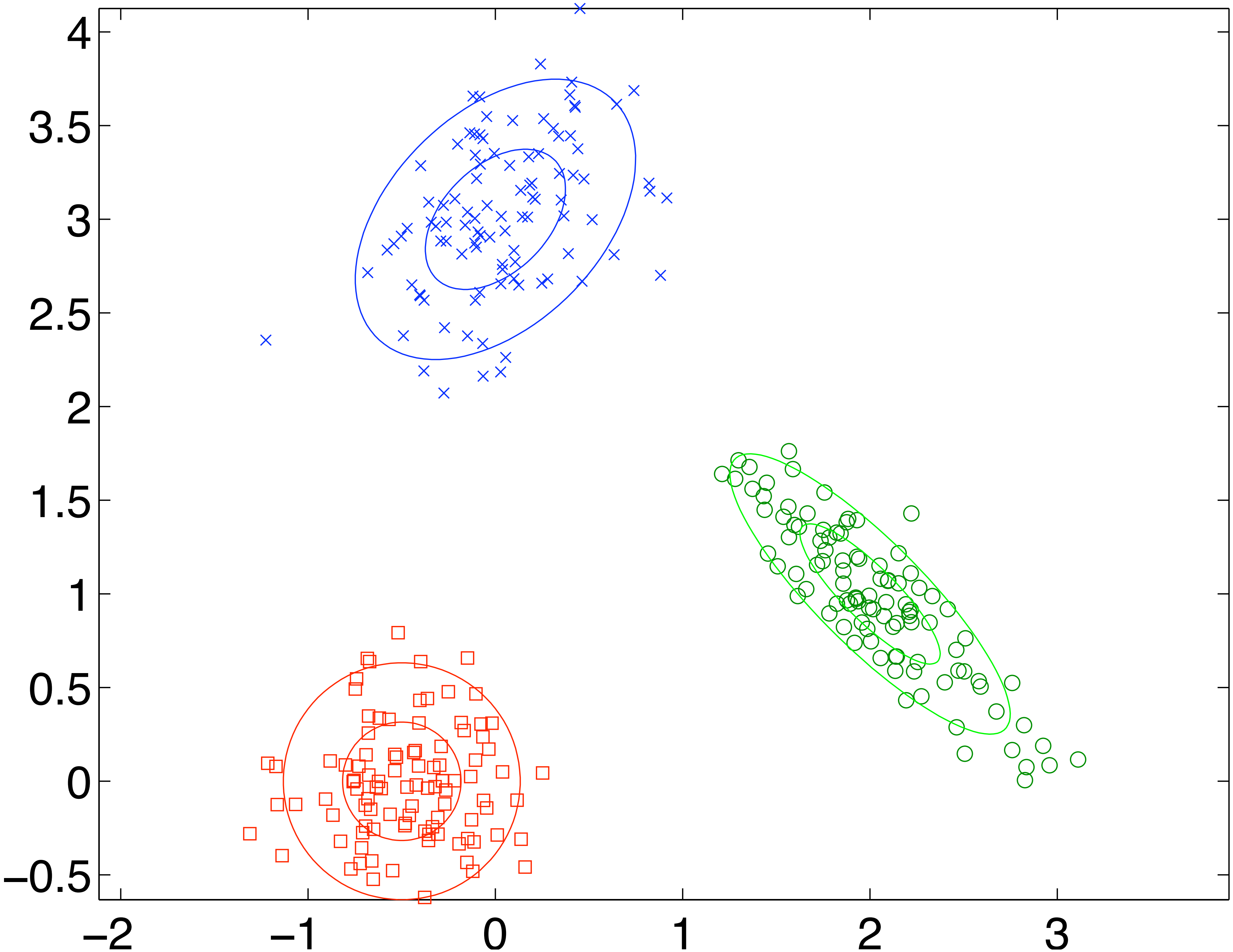


figure adapted from book

Continuous vars

- For categorical X, NB gave us a linear discriminant
- What about continuous X?
  - e.g., Gaussian NB
- Will turn out the same, but we'll work it out for a generalization

Multivariate Gaussians



A generalization of Gaussian Naïve Bayes

Generalizing GNB

- $P(X \mid Y) = N(X \mid \mu_Y, \Sigma)$
  - if Σ =
- Pick $Y=1$ if
  - $P(Y=1)\, P(X \mid \mu_1, \Sigma) \geq P(Y=0)\, P(X \mid \mu_0, \Sigma)$

Fisher linear discriminant

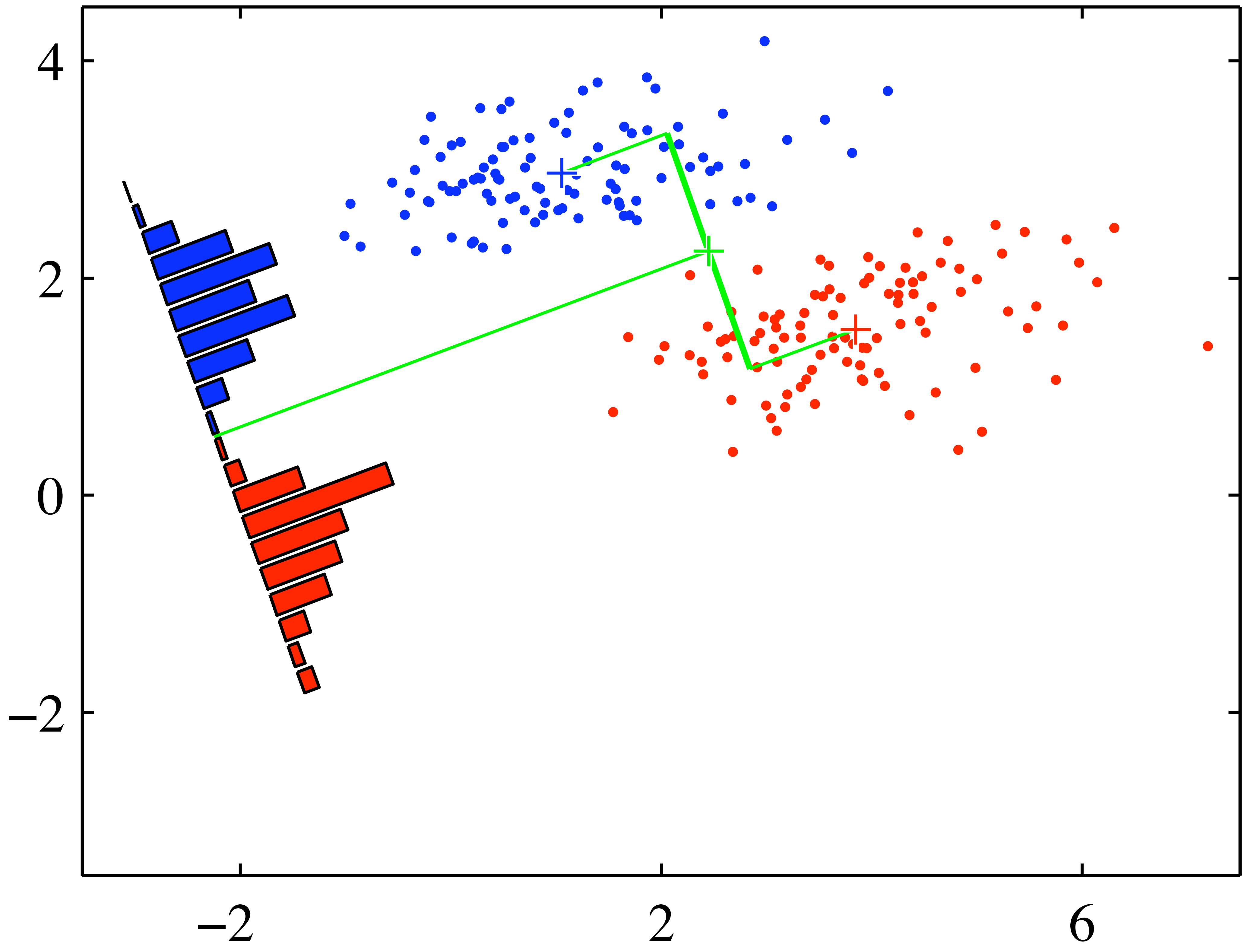


figure from book

Fisher w/ bad Σ

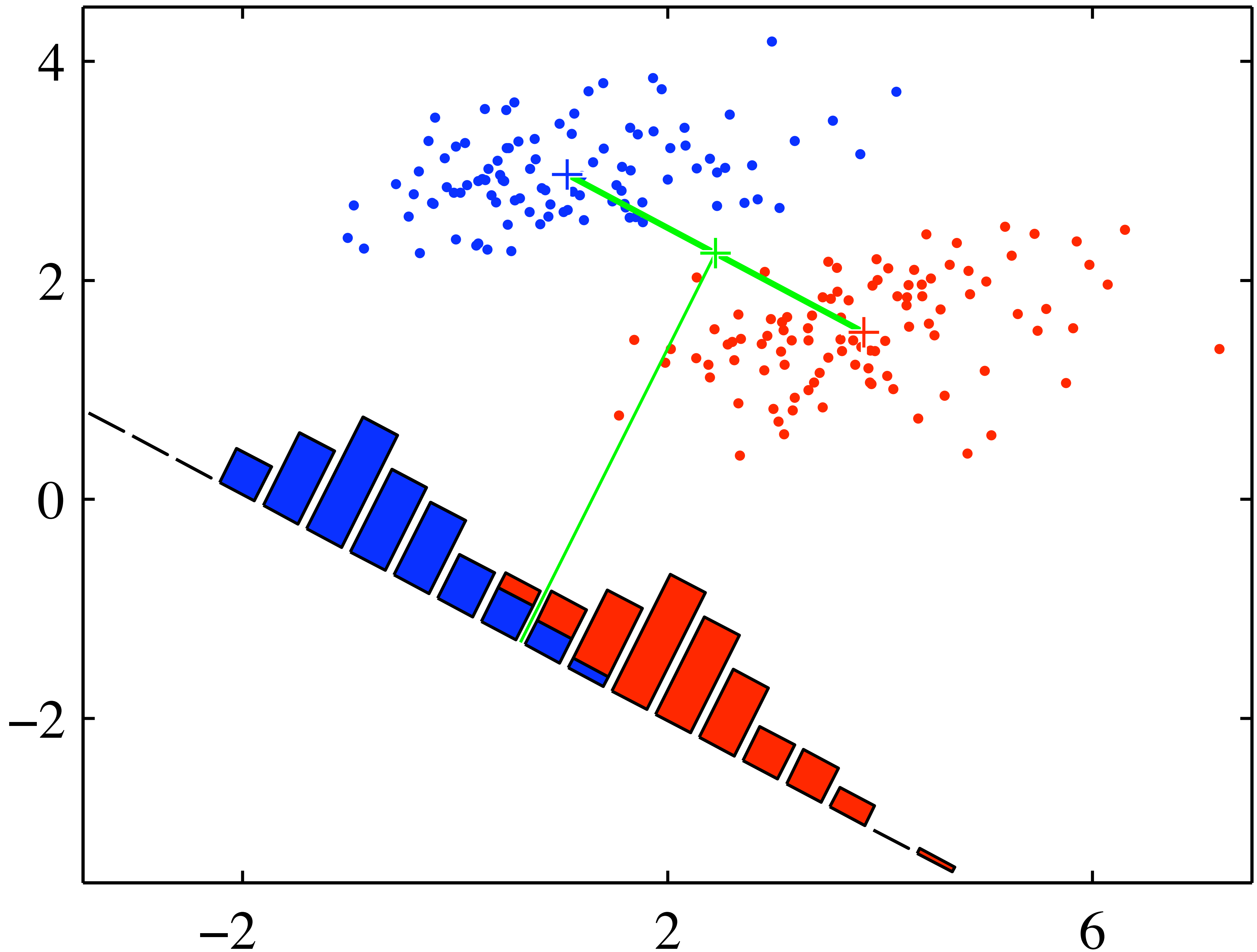


figure from book

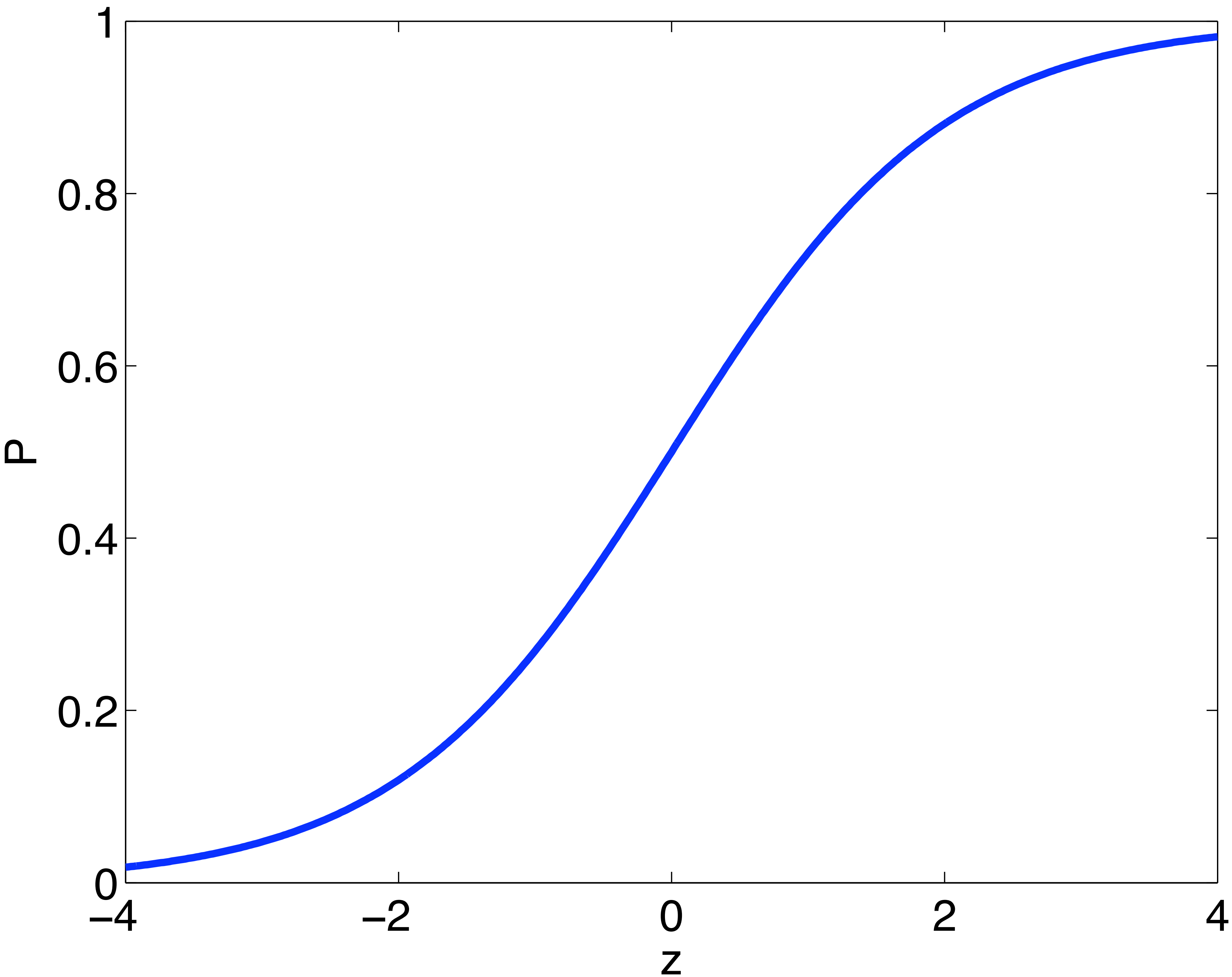
Linear discriminants

- Naïve Bayes, Gaussian NB, Fisher model
  - all lead to ***linear discriminants***
  - $w_0 + \sum_j w_j X_j \geq 0$
- One of most important types of classifier
  - esp. generalization: $w_0 + \sum_j w_j f_j(X) \geq 0$
  - $f_j(X)$ are ***features***
- Consequently, many ways to train LDs

Class probability

- We showed:
  - $\log P(Y=1 \mid X) - \log P(Y=0 \mid X) =$

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



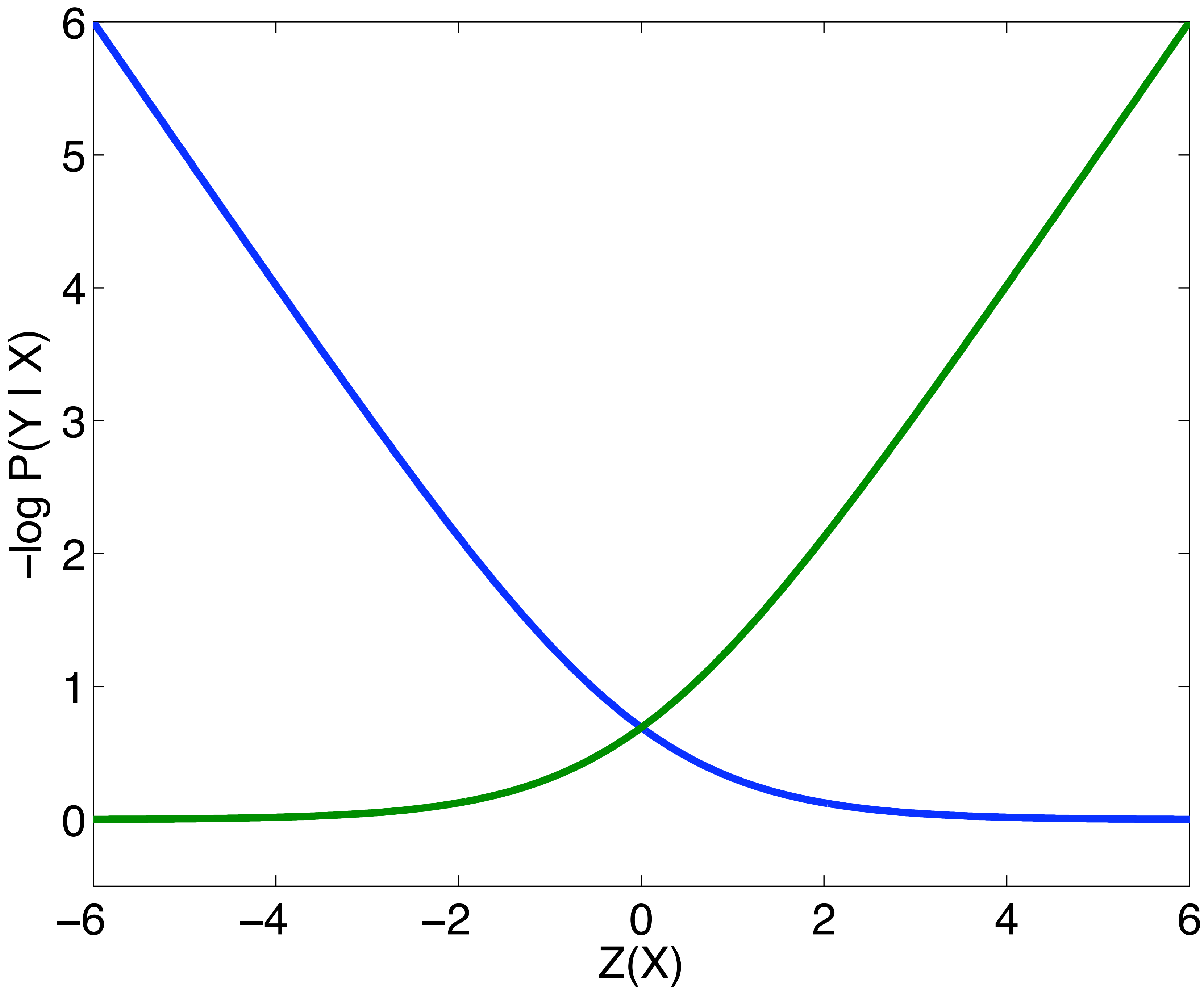
Maximum likelihood

- $P(Y=1 \mid X) = \sigma(z)$
  - $z = w_0 + \sum w_j X_j$
- NB is one algorithm for finding w
- Another: maximum (conditional) likelihood
  - given data $(X^1,Y^1), \ldots, (X^N,Y^N)$
  - arg max
- MLE for linear discriminant: ***logistic regression***

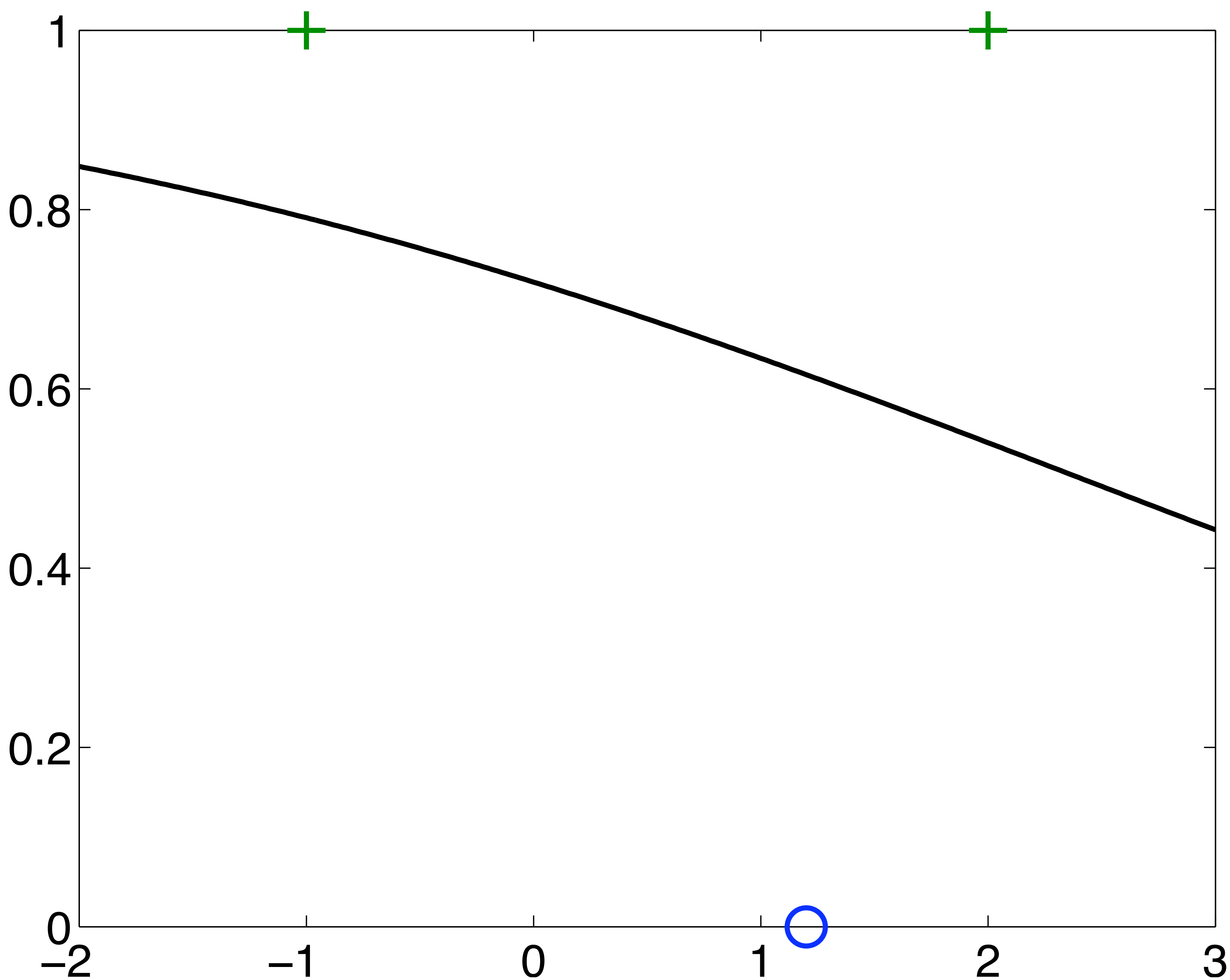
Logistic regression

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

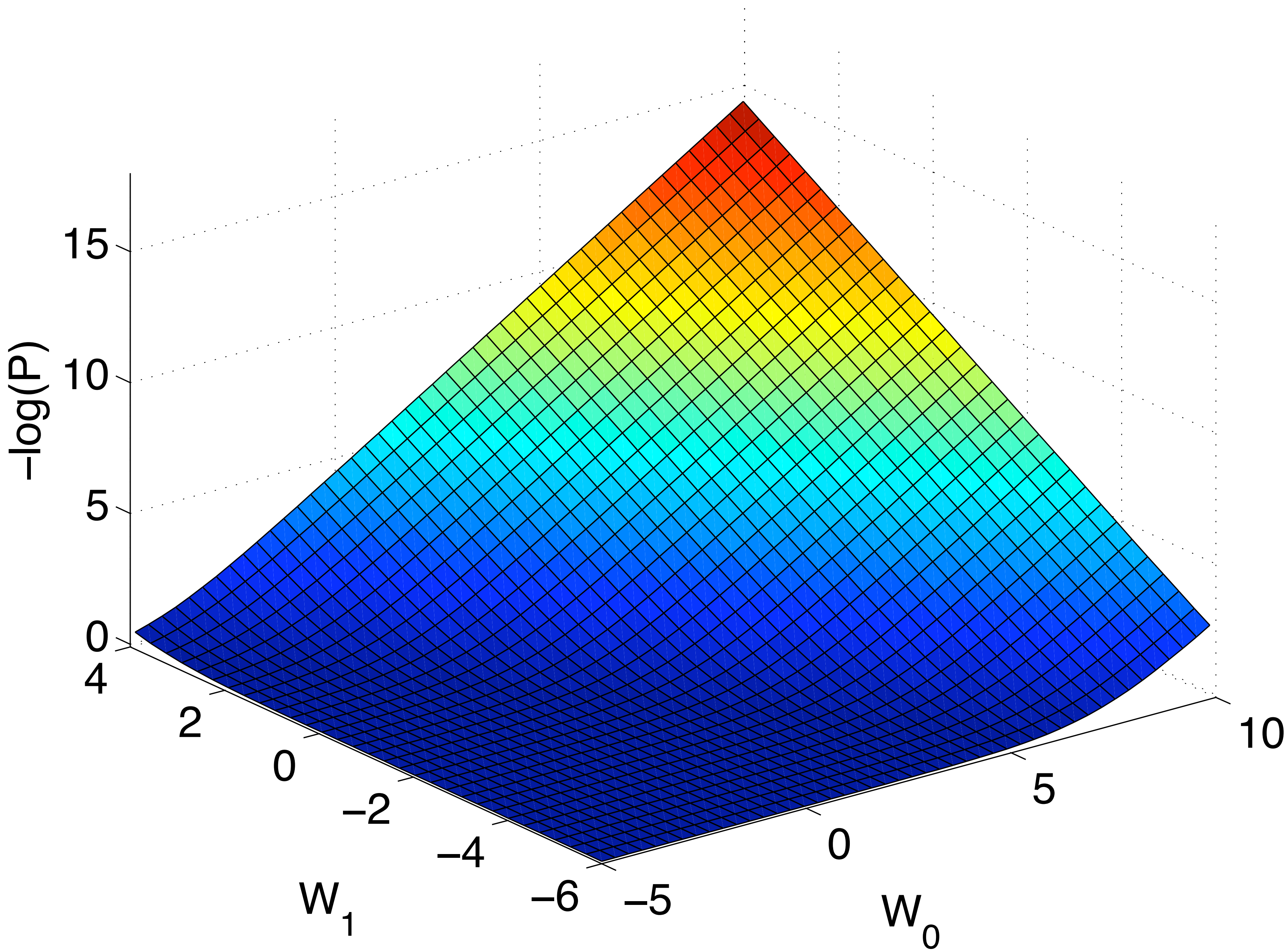
Neg. log likelihood



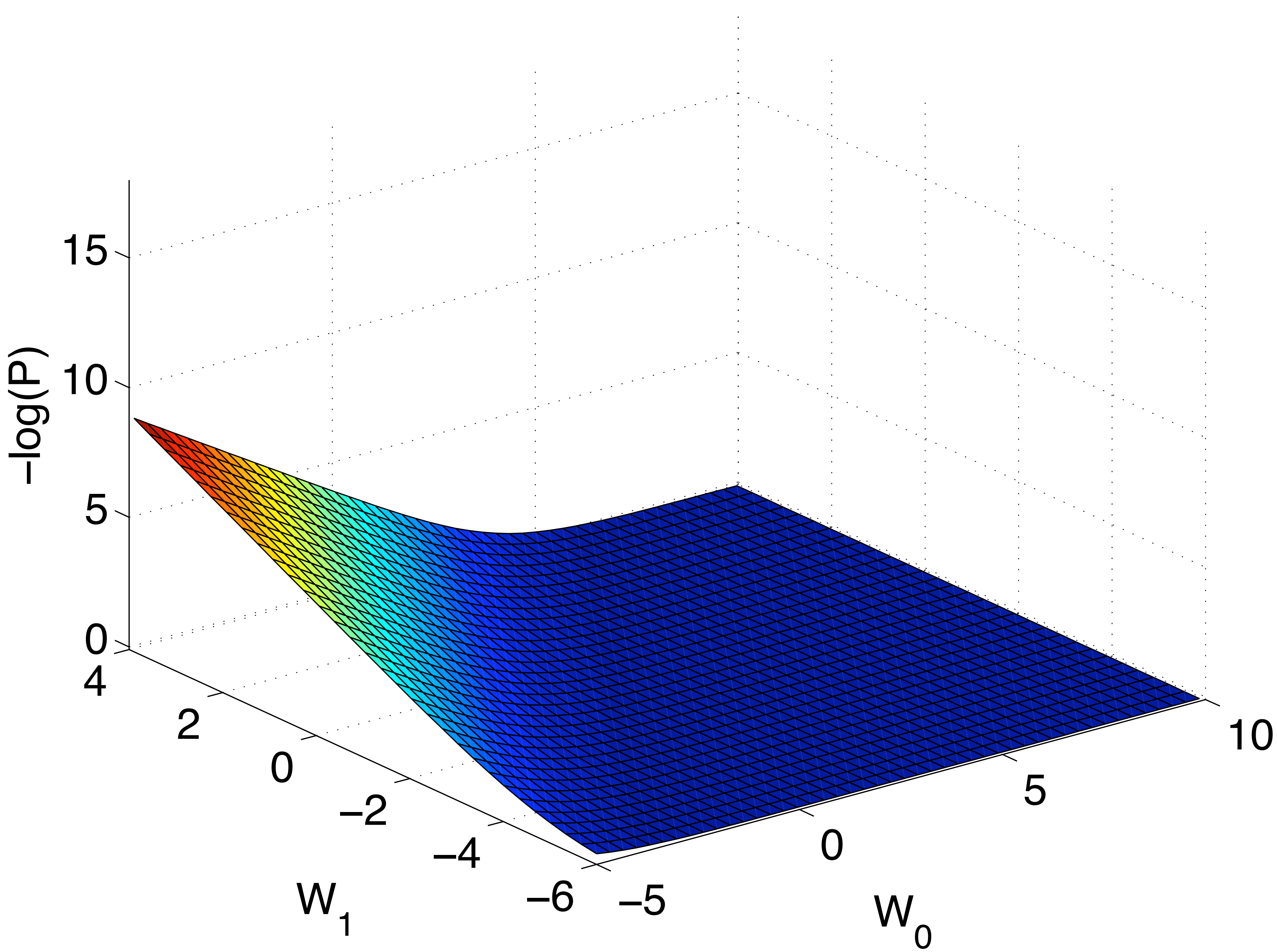
Example



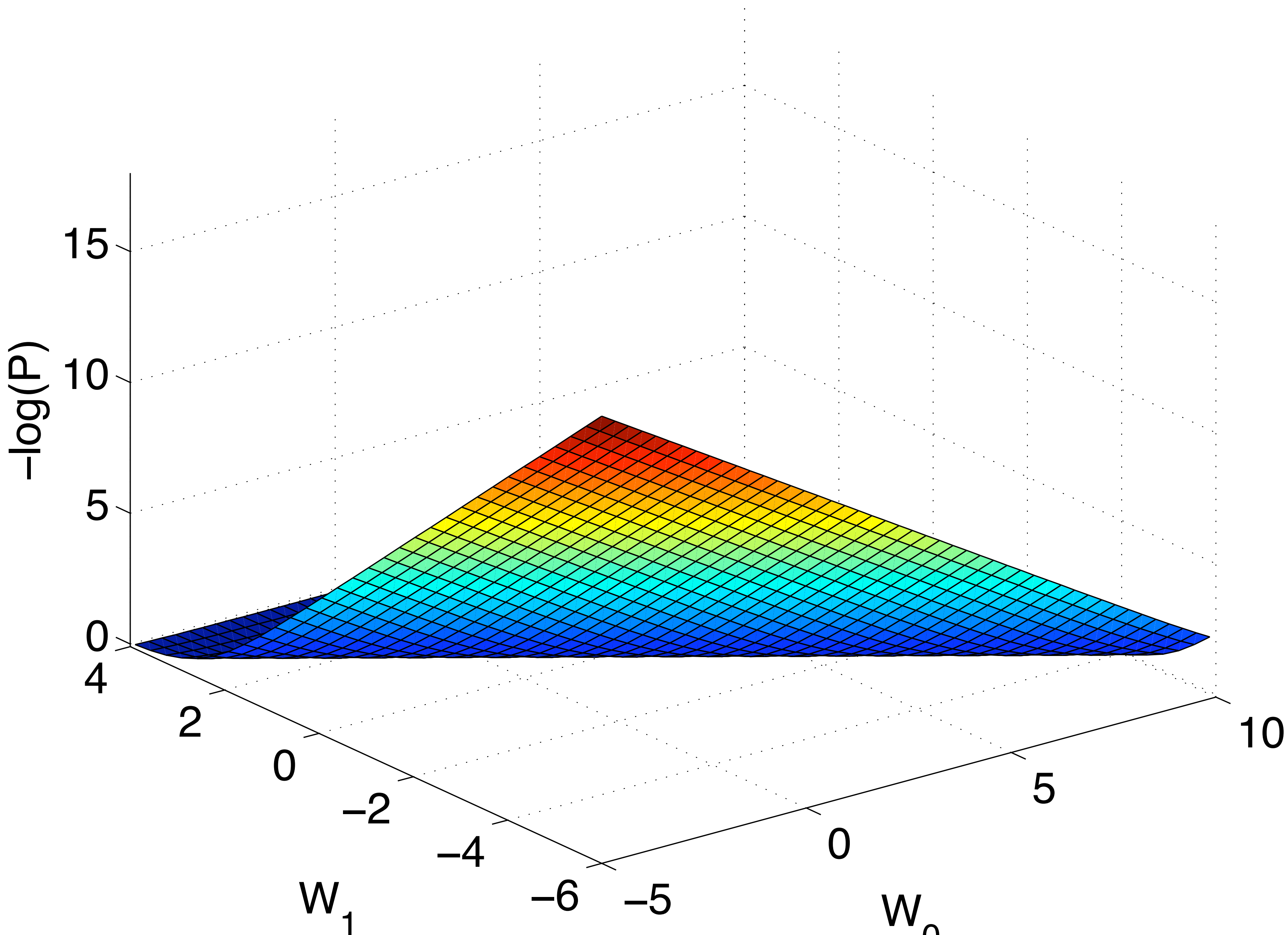
−1, 1



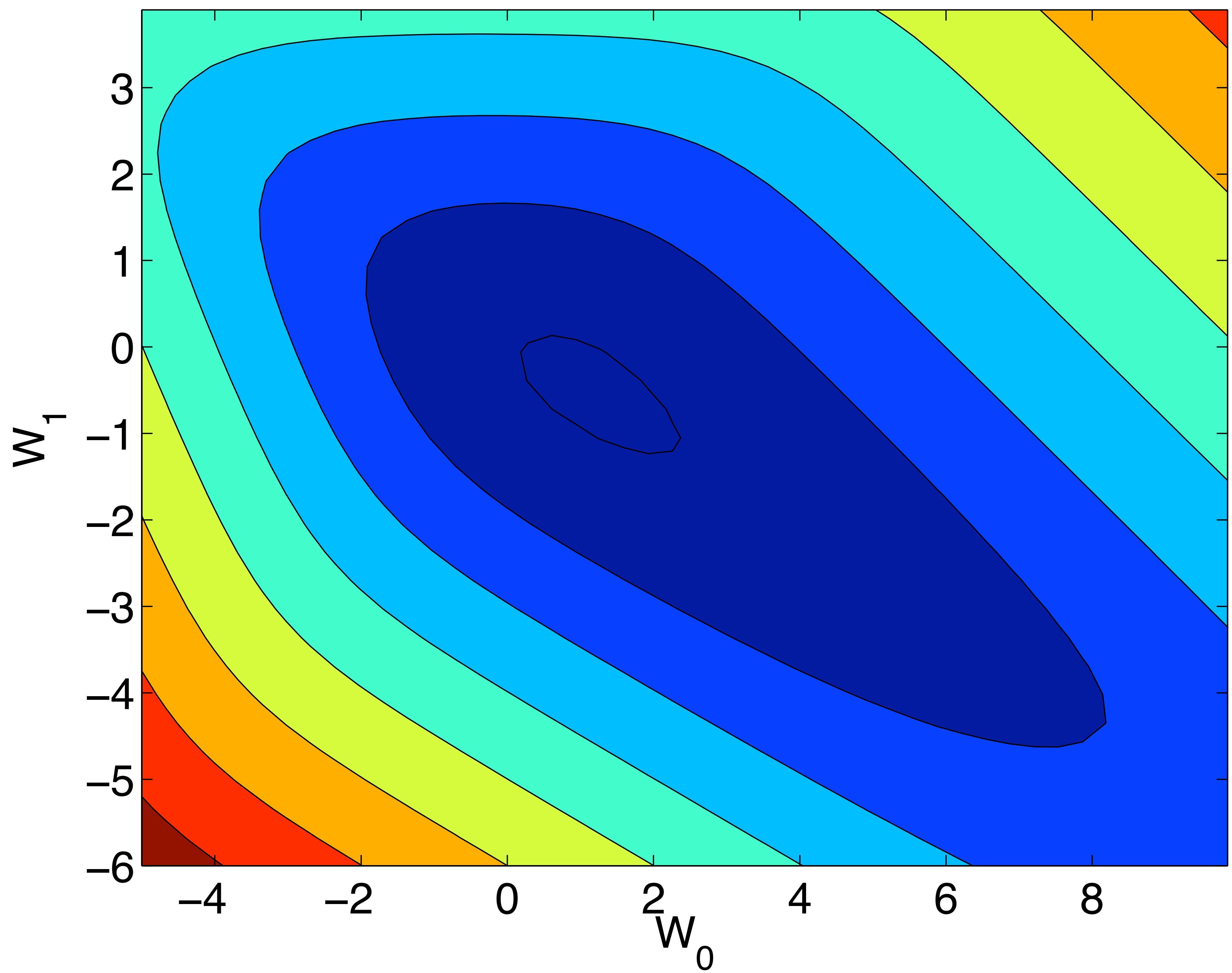
1.2, 0



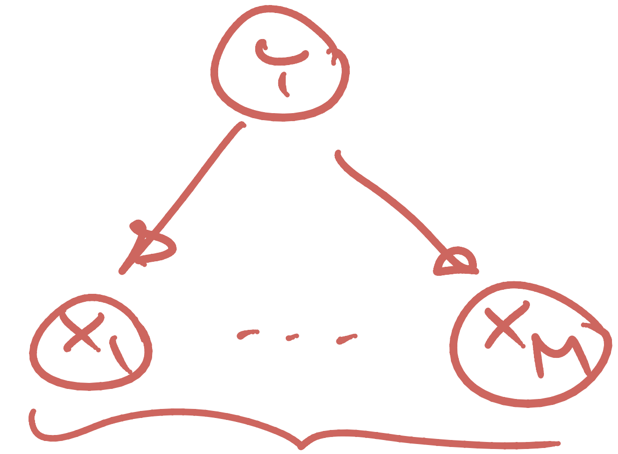
2, 1



Likelihood



Discussion



DIFFERENT FEATURES OF THE SAME EXAMPLE

- Two ways to train linear discriminants: naïve Bayes and logistic regression
  - based on same graphical model
  - max P(X,Y) vs max P(Y | X)
  - max likelihood vs max ***conditional*** likelihood
- NB lets us predict Y from X or X from Y; logistic regression can only predict Y from X
  - ***generative*** vs ***discriminative***

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

Generalization: multiple classes

- One weight vector per class: for Y=k
  - P(Y=k) =
  - Z_k =
- In 2-class case:

Multiclass example

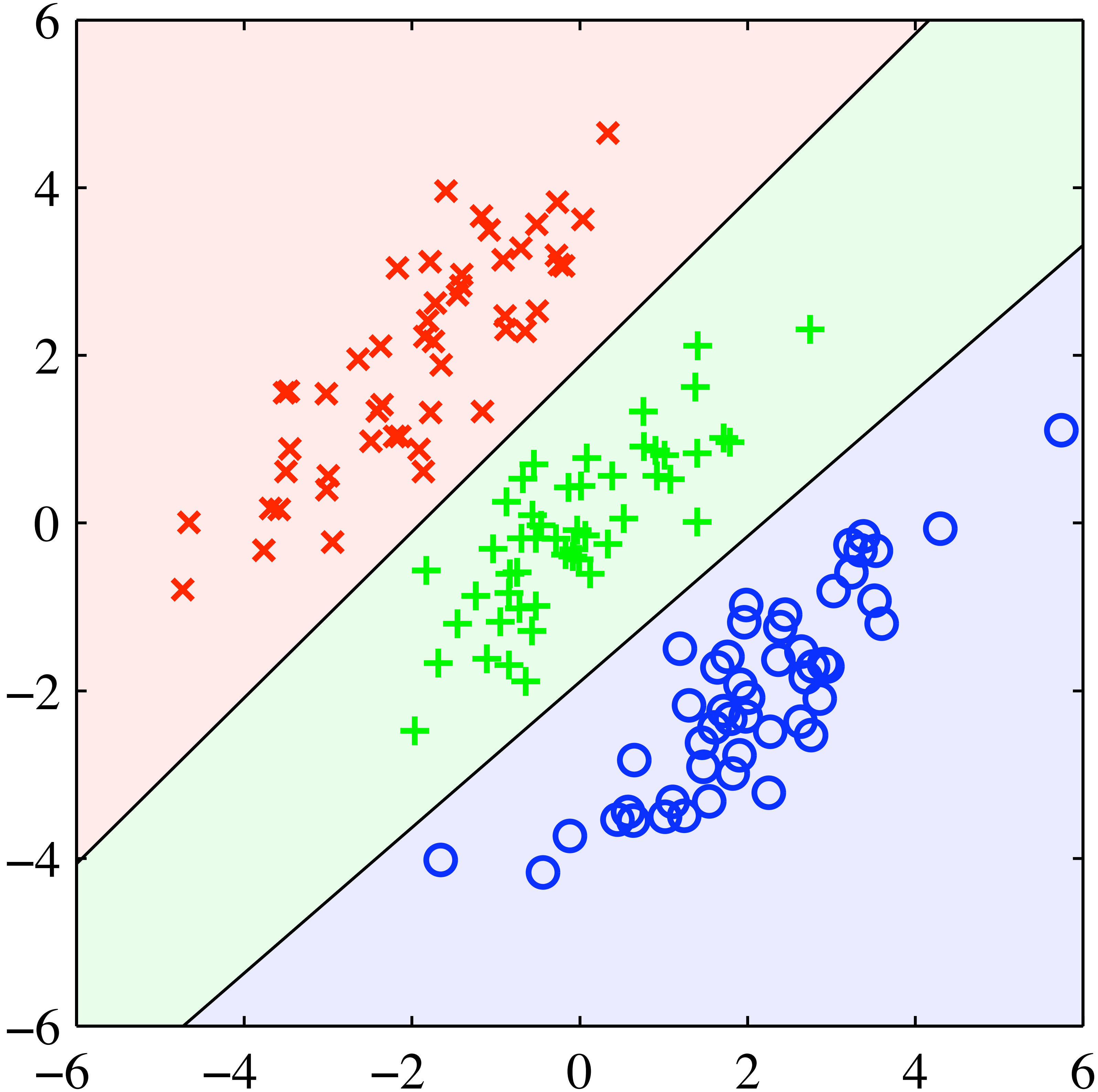


figure from book

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)

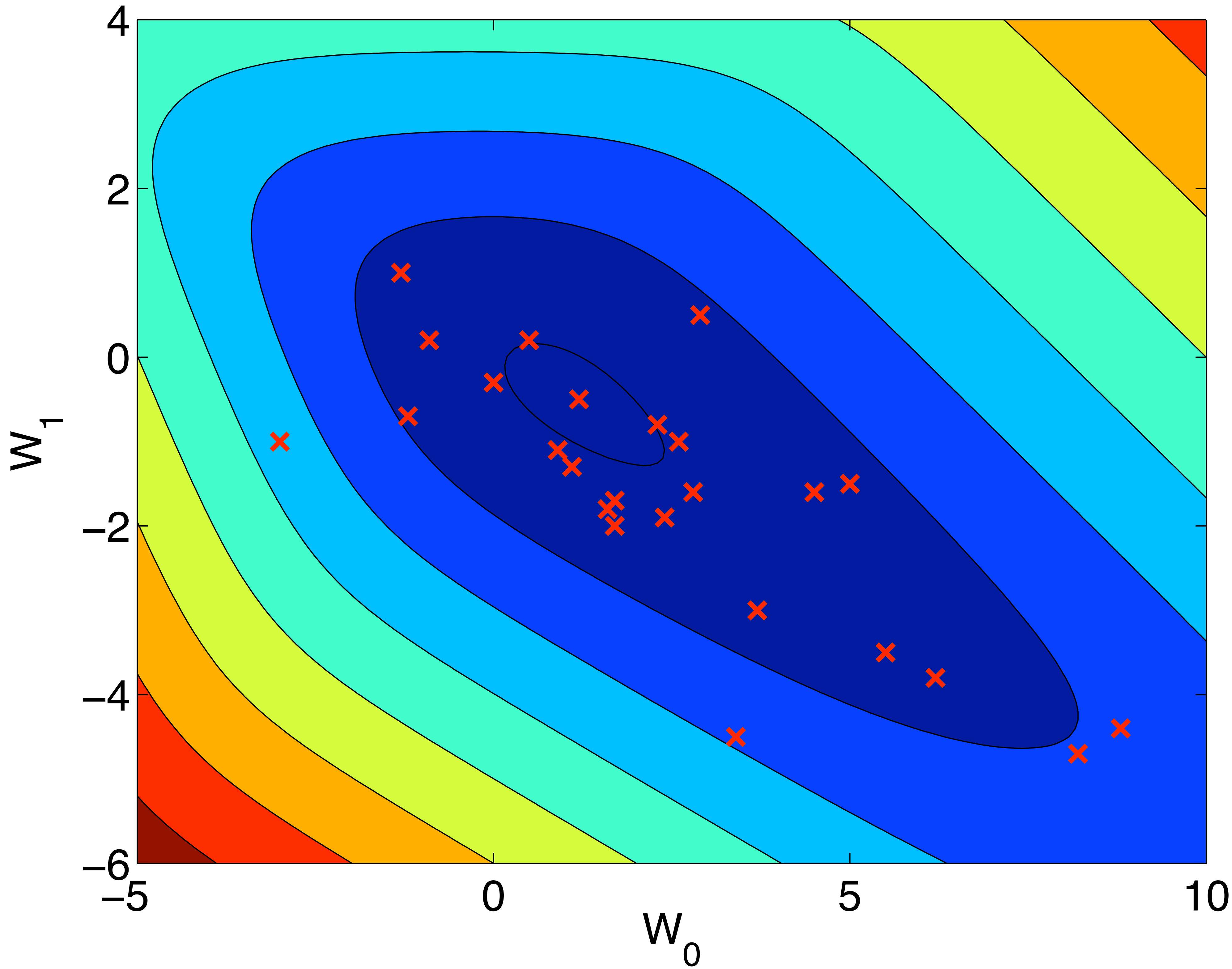
Software

- 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)

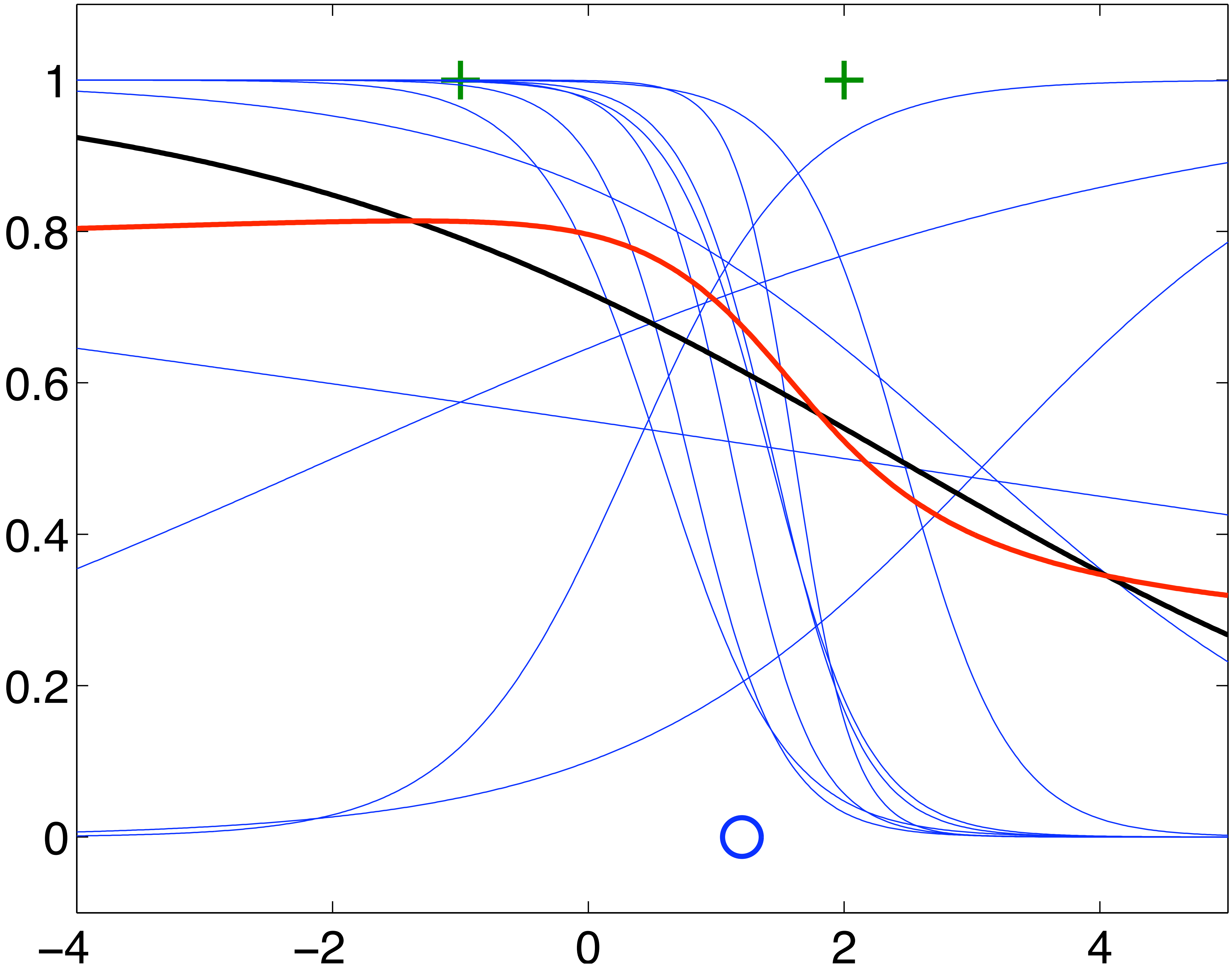
Bayesian regression

- In linear and logistic regression, we've looked at
  - MLE: $\max_w P(Y \mid X, w)$
  - 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