Bayesian Inference

Susanne Still
University of Hawaii at Manoa

This lecture mostly follows Chapter 2 of
David MacKay's PhD Thesis, Caltech (1991)
available at http://www.inference.phy.cam.ac.uk/mackay/PhD.html

Task

- Given: N measurements.
- Goal: develop and compare models to account for the measured data.
- This is a central task in science!!!

Example: Curve fitting

- Find interpolant through data.
- Criteria: explain data and predict interpolate well to new data (predict/generalize well).
- Solution: 1) Assume a parameterized model is true, find the best fit parameters.
  2) Compare different models.
- Questions: Best fit in what sense? How do we compare models?

Two levels of inference

1. Fix model (=function class). Fit parameters.
   Find the most probable parameters, given the data!
   Do this for each model.

2. Rank different models by the evidence we have for the model from the data.

1. Function fitting

- Given a function class F with parameterized functions f(x,w), where w are the parameters, find the function f(x,w*) with the most likely parameters w*, given the data D.
- "Most likely"? That's a probabilistic notion.

Finding the best parameters

- Data: $D = \{x_i, y_i\}$
- Model class M. Parameterized function: f(x,w)
- Probability of the parameters, given the data (and given that we assume the model is true): p(w|D,M) <- This is hard to compute!
- Probability (Likelihood) of the data, given the model and the parameters: p(D|w,M) <- Easy to compute

- Use Bayes' rule to calculate the posterior p(w|D,M) probability of the parameters (after we have seen the data) from the likelihood p(D|w,M) of the data given the parameters:

  p(w|D,M) = p(D|w,M)p(w|M)/p(D|M)

- p(w|M): prior prob. of parameters; p(D|M): Evidence for the model M.

  Posterior = Likelihood * Prior / Evidence

- Shorthand notation (M dropped, normalizing constant p(D) ignored):
  p(w|D) ~ p(D|w)p(w)

Example: Fitting function to noisy data

- Assume gaussian noise: y = f(x,w) + noise

$$p(y_i|x_i,w) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - f(x_i,w))^2}$$

- Likelihood:

$$p(\{y_i\}|\{x_i\},w) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - f(x_i,w))^2}$$

$$= \frac{1}{Z_D} e^{-\beta E_D}$$

with: $E_D = \sum_{i=1}^{N} (y_i - f(x_i,w))^2$ (mean squared error)

- Need a regularizer to control smoothness.

$$P(f|R,\alpha) = \frac{1}{Z_f(\alpha)} e^{-\alpha E_f(f|R)}$$

Example: Cubic spline interpolation.

$$E_f(f|R) = \int dx (f''(x))^2$$

- Can be expressed as prior on weights:

$$P(w|M) = \frac{1}{Z_w(\alpha)} e^{-\alpha E_w}$$

- Altogether: Posterior ~ Likelihood * Prior

$$P(w|D) = \frac{1}{Z_M(\alpha,\beta)} e^{-(\alpha E_w + \beta E_D)}$$

- Maximizing the (log of the) posterior is the same as minimizing

$$E_M = (\alpha E_w + \beta E_D)$$

- Under gaussian noise model, minimizing MSE finds the maximum likelihood parameters. (No regularization)

Recall: Two levels of inference

1. Fix model (=function class). Fit parameters.
   Find the most probable parameters, given the data!
   Do this for each model.

2. Rank different models by the evidence we have for the model from the data.
   Bayesian inference embodies "Occam's Razor"!

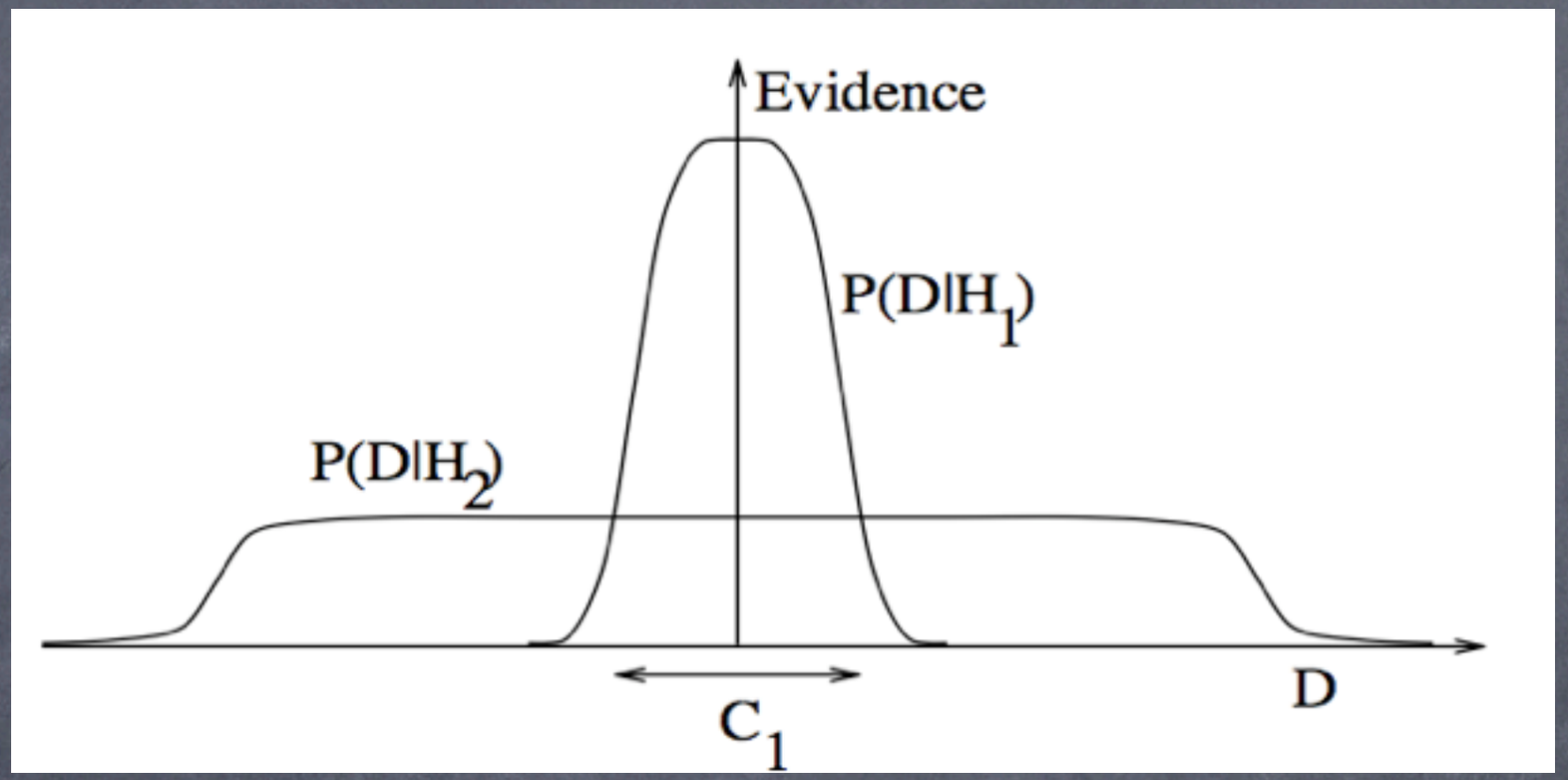
Occam's Razor

- Never use an unnecessarily complicated (or: complex) model.
- Concept goes back at least to Aristoteles
- William of Occam is often quoted for stating it.
- We will see that bayesian inference embodies "Occam's Razor"!

Bayesian Model Comparison

- Model M, Data set D.
- p(M|D) ~ p(D|M)p(M)
- Posterior probability of Model ~ Evidence*Prior
- recall: Evidence appeared as normalizing constant in bayesian parameter estimation
  p(w|D,M) = p(D|w,M)p(w|M)/p(D|M)
- If models have equal prior probability, they are ranked by evaluating the evidence.

Why are complex models penalized - Intuition



- Bayes embodies Occam's razor because a more complex model (H_2) predicts a greater variety of data sets D), thus does not predict the data in a given region (C_1) as strongly.

Evaluating the evidence

$$p(D|H_i) \quad = \quad \int p(D|\mathbf{w}, H_i) p(\mathbf{w}|H_i) d\mathbf{w}$$

- Posterior over parameters $p(\mathbf{w}|D, H_i) \sim p(D|\mathbf{w}, H_i) p(\mathbf{w}|H_i)$
- often has a strong peak around the most likely parameters w_{MP}
- Approximate Evidence by the height of the peak times it's width:

$$p(D|H_i) \quad \simeq \quad p(D|\mathbf{w}_{MP}, H_i) p(\mathbf{w}_{MP}|H_i) \Delta\mathbf{w}$$

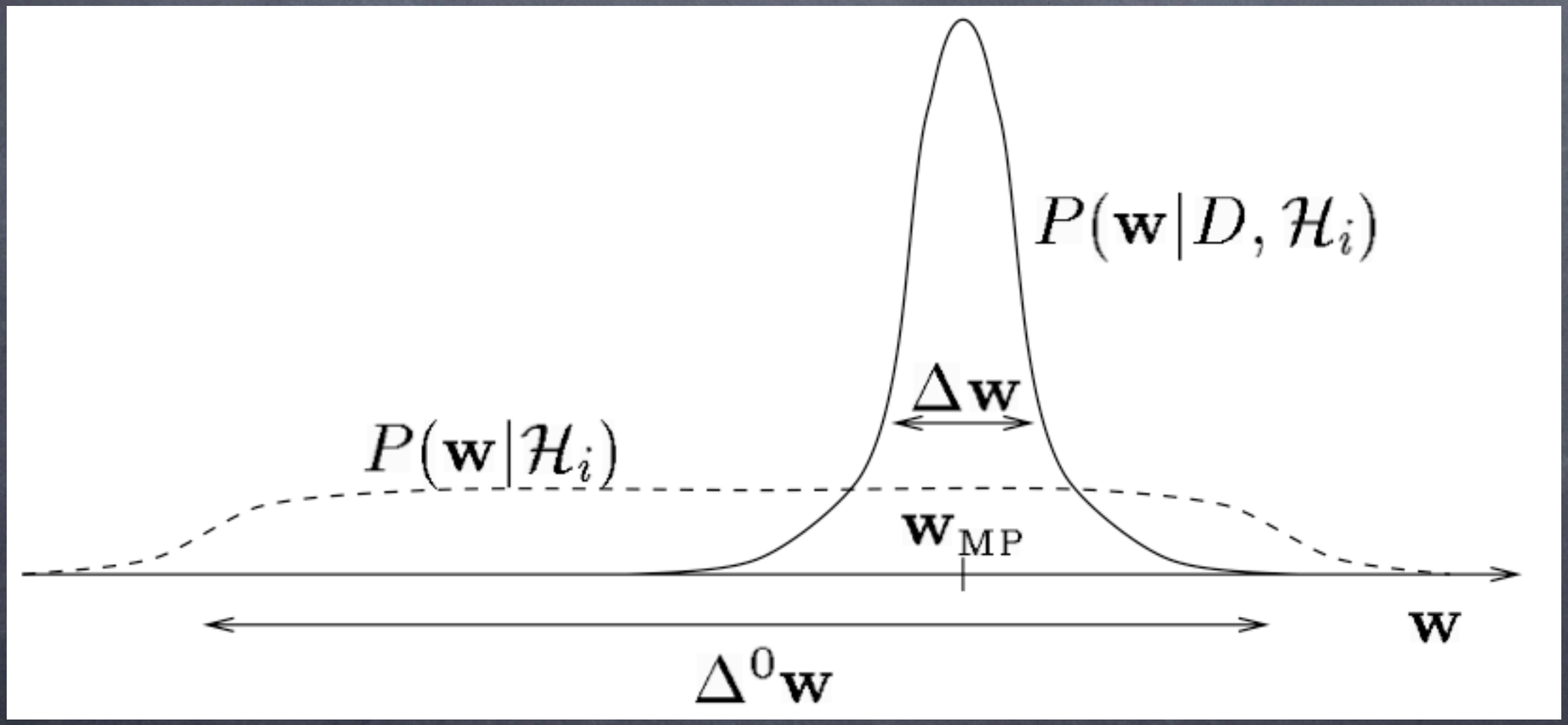
Occam Factor

$$p(D|H_i) \simeq p(D|\mathbf{w}_{MP}, H_i) p(\mathbf{w}_{MP}|H_i) \Delta\mathbf{w}$$

Evidence — Best fit Likelihood — Occam factor

- Evidence is approximately the best fit likelihood times the Occam factor.
- $\Delta\mathrm{w}$ is the posterior uncertainty in the parameters. With uniform prior: $p(\mathbf{w}_{MP}|H_i) = \frac{1}{\Delta^0\mathbf{w}}$
- Occam facor = $\frac{\Delta\mathbf{w}}{\Delta^0\mathbf{w}}$
- Ratio of posterior accessible volume of model's parameter space to the prior accessible volume.

Occam Factor



Noisy interpolation

- Recall: Posterior ~ Likelihood * Prior

$$P(w|D) = \frac{1}{Z_M(\alpha,\beta)} e^{-(\alpha E_w + \beta E_D)}$$

- Maximizing the (log of the) posterior is the same as minimizing $E_M = (\alpha E_w + \beta E_D)$

$$E_D = \sum_{i=1}^{N} (y_i - f(x_i, w))^2$$

- Under gaussian noise model, minimizing MSE finds the maximum likelihood parameters. (No regularization)

Noisy interpolation

- Bayesian choice of parametes α, β :
  Evaluate posterior:

$$p(\alpha, \beta|D, H) = \frac{p(D|\alpha, \beta, H)p(\alpha, \beta|H)}{p(D|H)}$$

- $p(D|\alpha, \beta, H)$ is the evidence for α, β , given by the ratio of the normalizing constants:

$$p(D|\alpha, \beta, H) = \frac{Z_M(\alpha, \beta)}{Z_W(\alpha)Z_D(\beta)}$$

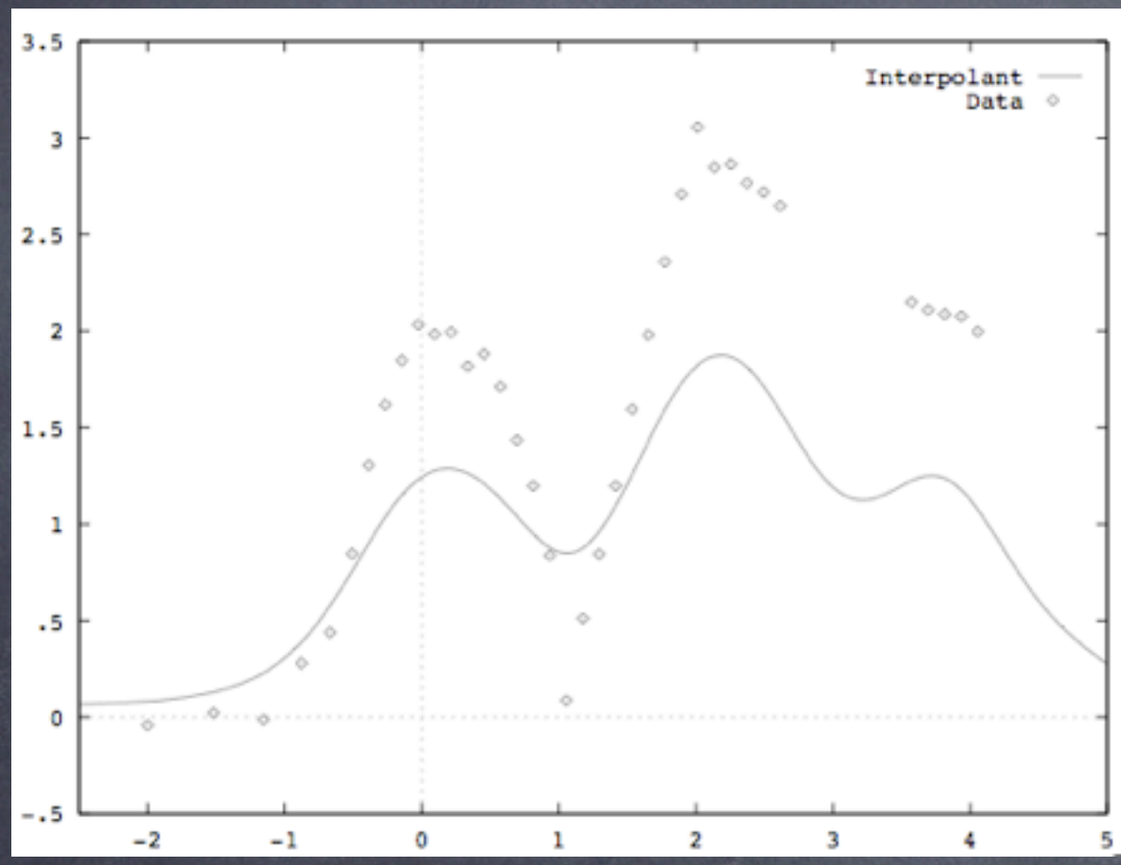
- Find the integrals Z_M, Z_D, Z_W.
  Assume that the regularizer is a quadratic functional => Z_M is gaussian.

Noisy interpolation

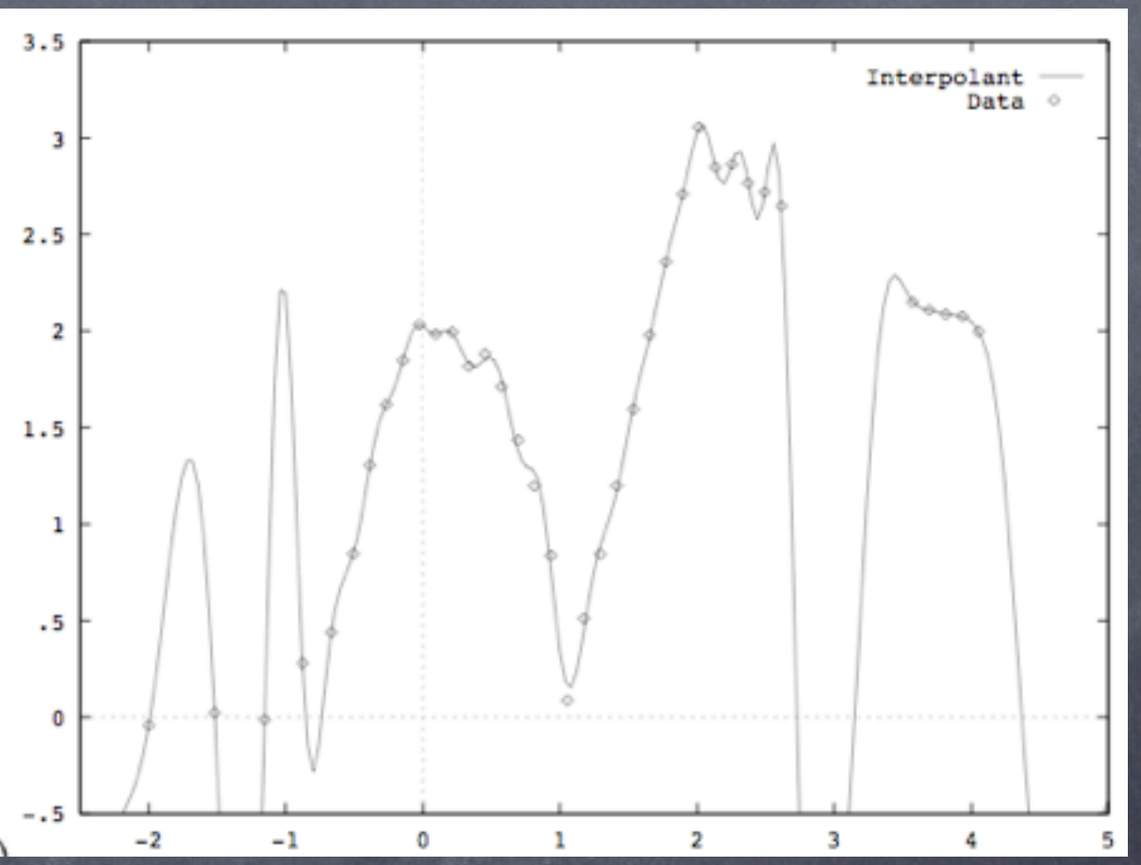
- Model comparison: $p(H|D) \sim p(D|H)p(H)$
- Evidence:

$$p(D|H) = \int d\alpha \; d\beta \; p(D|\alpha, \beta, H) p(\alpha, \beta|H)$$

Interpolation: Example



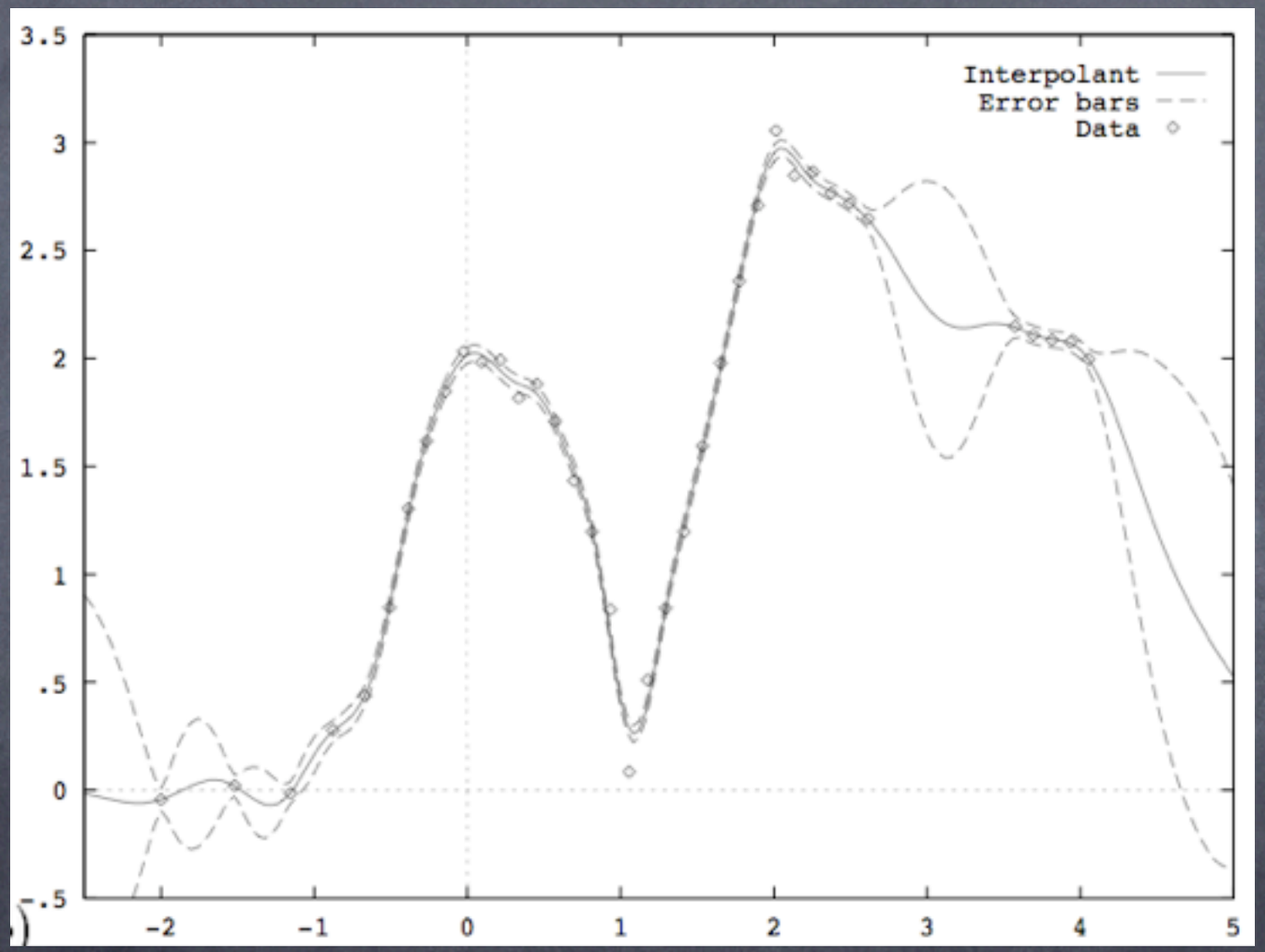
α LARGE



α SMALL

- The best interpolant depends on the regularization parameter (use flat prior on α and $E_W = \frac{1}{2}\sum_i w_i^2$).

Interpolation: Example



Most probable α with error bars

Homework

- Write a program that lets you fit a polynomial of order n to data
  - (1) without and
  - (2) with regularization ("ridge regression").