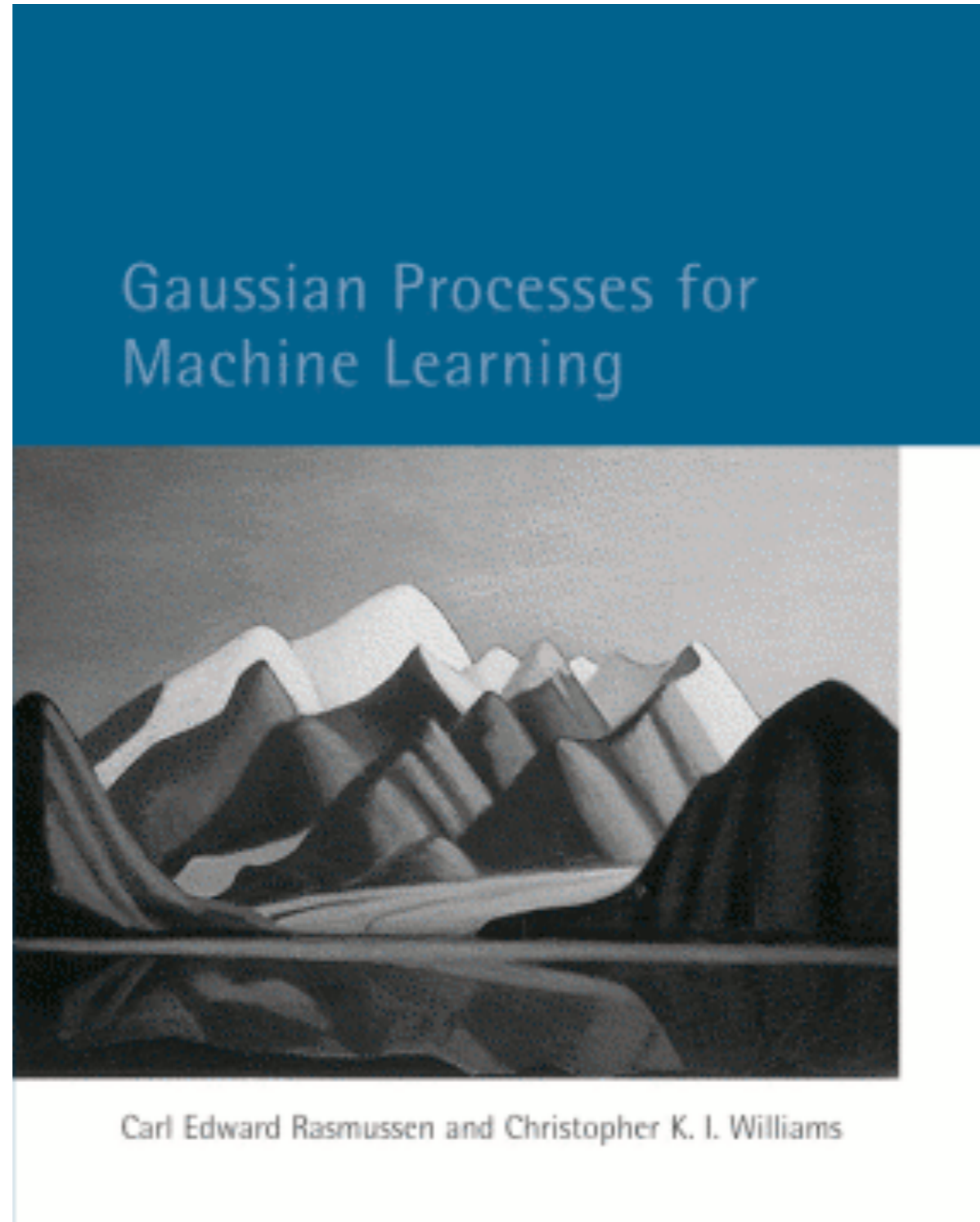


Gaussian Processes



Carl Edward Rasmussen and Christopher K. I. Williams

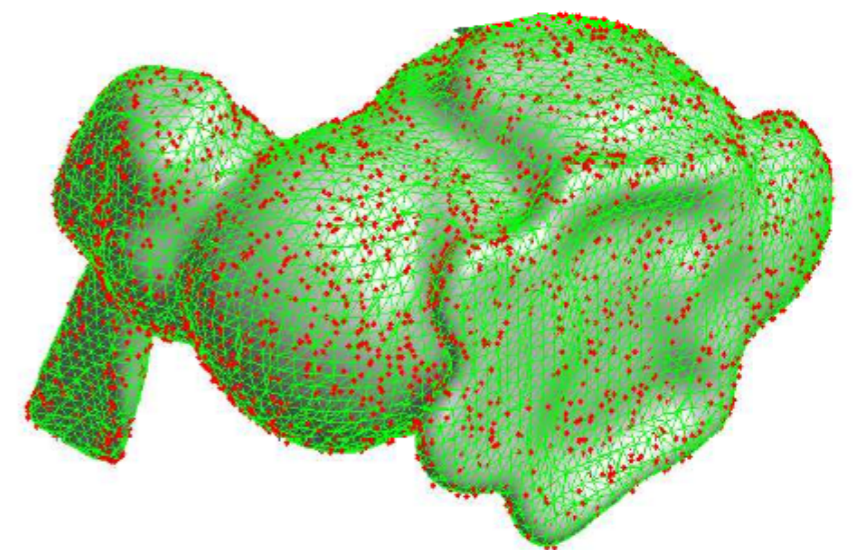
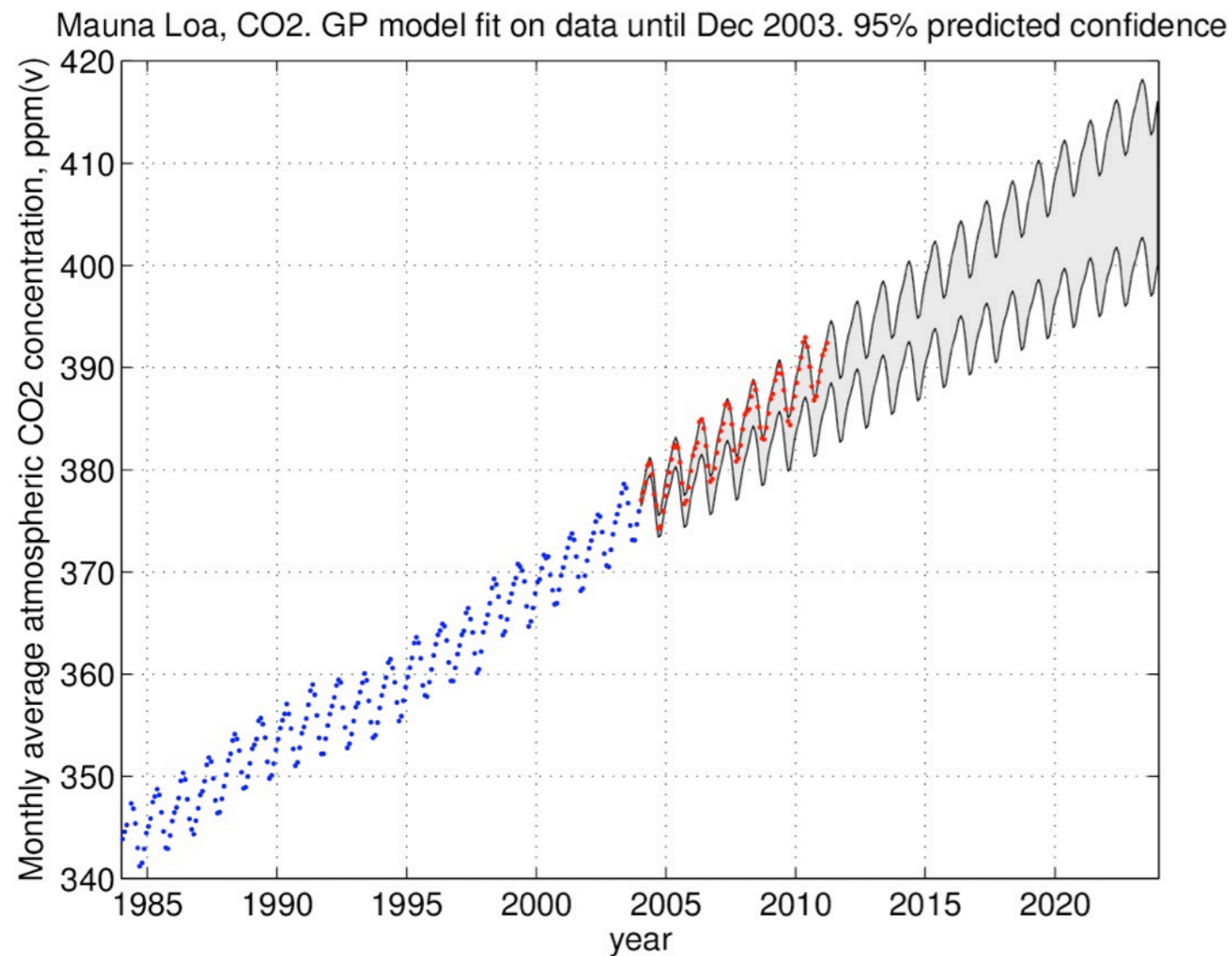
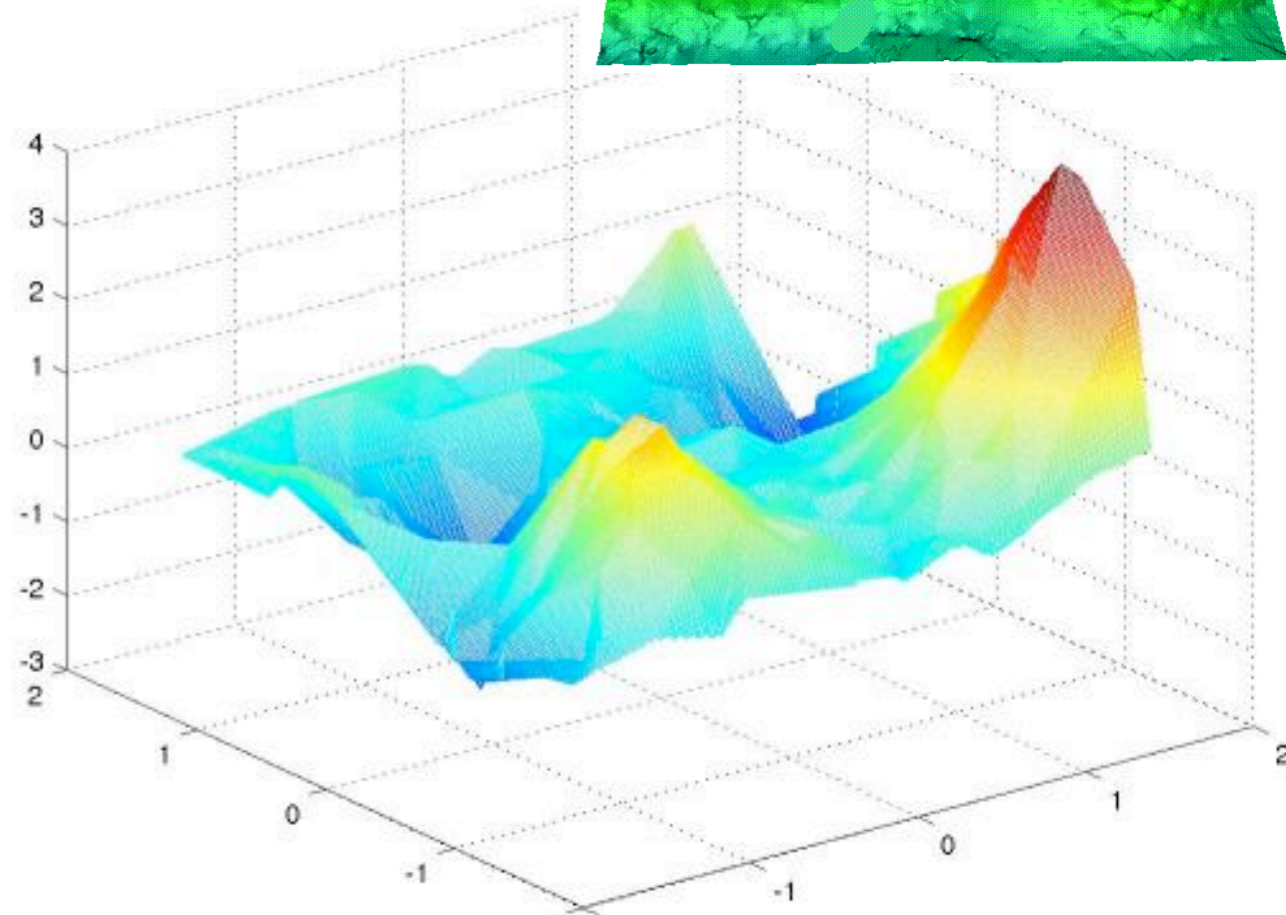
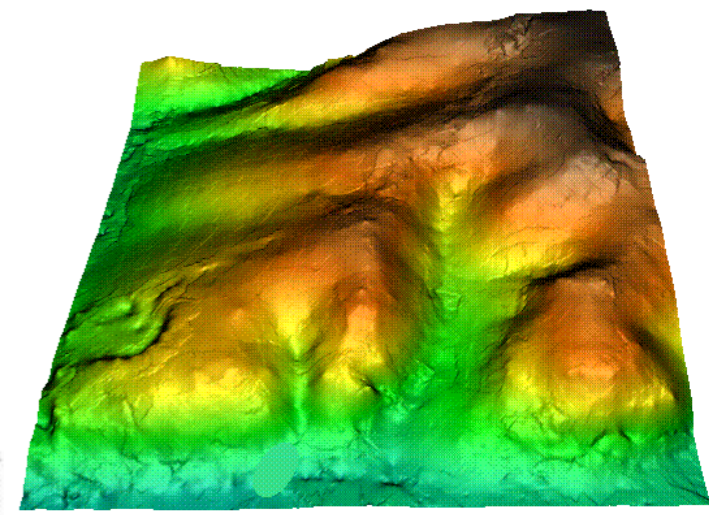
Gaussian Process

- Stochastic process:
 - basically, a set of random variables.
 - may be infinite.
 - usually related in some way.
- Gaussian process:
 - each variable has a Gaussian distribution
 - every finite set follows multivariate Gaussian

To date kriging has been used in a variety of disciplines, including the following:

- Environmental science^[5]
- Hydrogeology^{[6][7][8]}
- Mining^{[9][10]}
- Natural resources^{[11][12]}
- Remote sensing^[13]
- Real estate appraisal^{[14][15]}

and many others.



Gaussian Processes

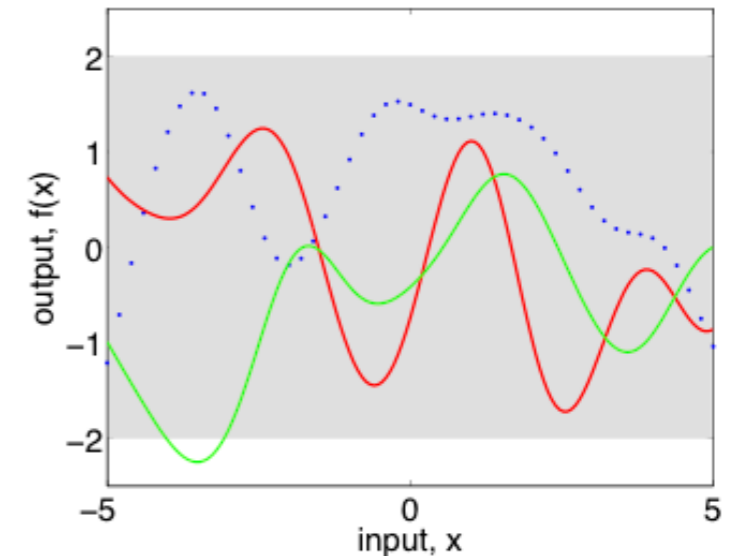
- GPs specified by mean/covariance function:
 - $m(x) = E[f(x)]$.
 - $k(x, x') = E[(f(x) - m(x))(f(x') - m(x'))^T]$.
 - $f(x) \sim GP[m(x), k(x, x')]$.
- 2 questions:
 - Why do we care?
 - How is this related to kernels?

Linear regression example

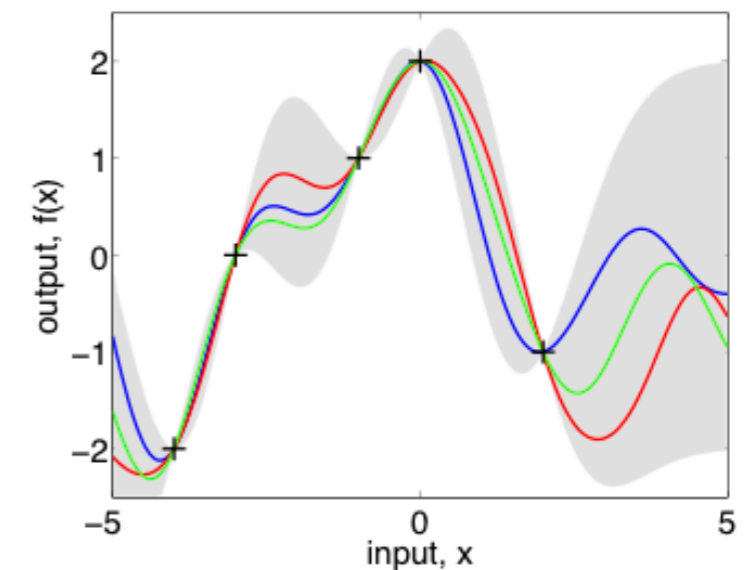
- Simple linear regression:
 - $f(x) = \phi(x)^T w$
 - $w \sim N(0, \Sigma)$
- The mean and covariance are given by
 - $E[f(x)] = \phi(x)E[w] = 0.$
 - $E[f(x)f(x')^T] = \phi(x)^T E[ww^T] \phi(x') = \phi(x)^T \Sigma \phi(x')$
 $= k(x, x').$

RBF Covariance

- Take:
 - $k(x,y) = \exp(-\|x - y\|^2/2\sigma)$.
 - prior distribution over some smooth functions
 - with efficient operations
- Use this as prior/regularizer
 - prior: $f^* \sim N(0, K(x^*, x^*))$
 - posterior: $f^* \mid x^*, x, f \sim N(K(x^*, x)K(x, x)^{-1}f, K(x^*, x^*) - K(x^*, x)K(x, x)^{-1}K(x, x^*))$




(a), prior



(b), posterior

Is this the same as using kernels?

Yes



And why exactly
are you wasting
my time?

Outline

- ~~Gaussian Process Bait and Switch~~
- Bayesian Statistics
- Marginal Likelihood





Maximum Likelihood

- Maximum likelihood:
 - $\operatorname{argmax}_w p(y \mid x, w)$
- Usually:
 - consistent (converges as $n \Rightarrow \infty$)
 - efficient (rate is fastest possible as $n \Rightarrow \infty$)
- But:
 - usually we have finite n
 - sometimes doesn't make sense
 - over-fitting

Maximum a posteriori (MAP)

- Assume a **prior $p(w)$** on random variable w .
- Bayes rule to maximize w given $\{y,x\}$:
 - $p(w | y, x) = p(y | x, w)p(w)/p(y | x)$
- Denominator does not depend on w :
 - $\operatorname{argmax}_w \log p(y | x, w) + \log p(w)$
- I.e.:
 - $\operatorname{argmax}_w \|Xw - y\|^2 + \|w\|^2/\alpha$
- Does this make the right decision?

Consider simple case of five hypotheses:

w_1	w_2	w_3	w_4
$p(w_1 y, x) = 0.25$	$p(w_2 y, x) = \mathbf{0.3}$	$p(w_3 y, x) = 0.25$	$p(w_4 y, x) = 0.2$
			

According to MAP (w_2), thumbs down.

According to non-MAP, thumbs up!

$$p(w=w_2) = 0.3, p(w \neq w_2) = 0.7.$$

Thumbs down (MAP) or thumbs up (Bayesian)?

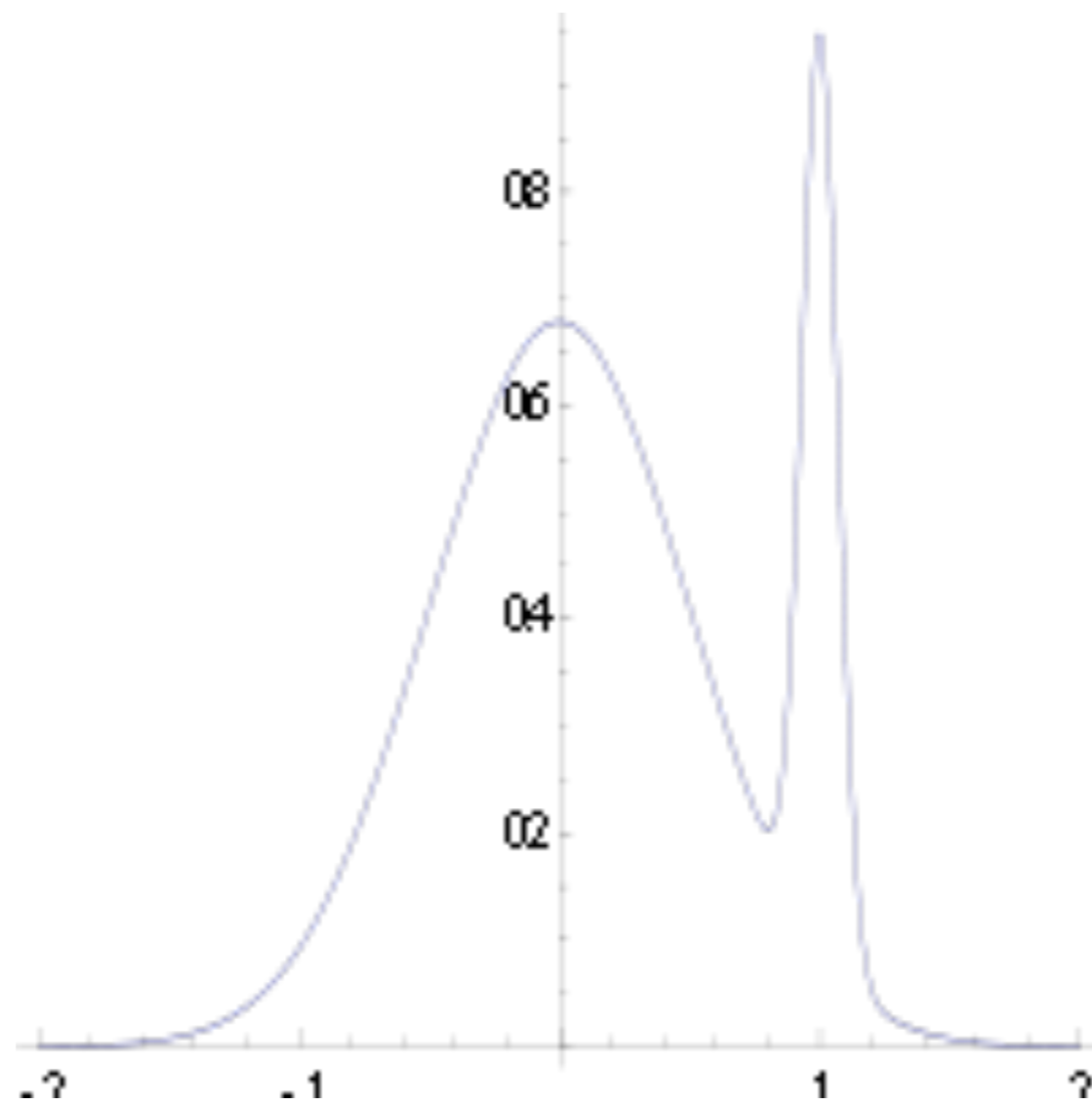


Image + annotation



MAP solution



Average of 20 samples



Error estimates

Figure 2: Example image with the boundary annotation (left) and the error estimates obtained using our method (right). Thin structures of the object are often lost in a single MAP solution (middle-left), which are recovered by averaging the samples (middle-right) leading to better error estimates.

Bayesian Inference

- Bayesian approach considers the **full posterior**:
 - $p(w \mid y, x) = p(y \mid x, w)p(w)/p(y \mid x)$
- **Prediction by integrating over uncertainty**:
 - $p(y^* \mid x^*, y, x) = \int p(y^* \mid x^*, w)p(w \mid y, x)dw.$
 - Note: integrate instead of maximize.
- Can also add risk function:
 - not generally optimized by posterior mode (MAP)
 - **squared error minimized by posterior mean.**
 - **absolute error minimized by posterior median.**

Solving Integrals

- How do we solve these integrals?
 - numerical integration (low dim)
 - conjugate priors (Gaussian likelihood w/ GP prior)
 - subset methods (Nystrom)
 - fast linear algebra (Krylov, fast transforms, KD-trees)
 - variational methods (Laplace, mean-field, EP)
 - Monte Carlo methods (Gibbs, MH, particle)

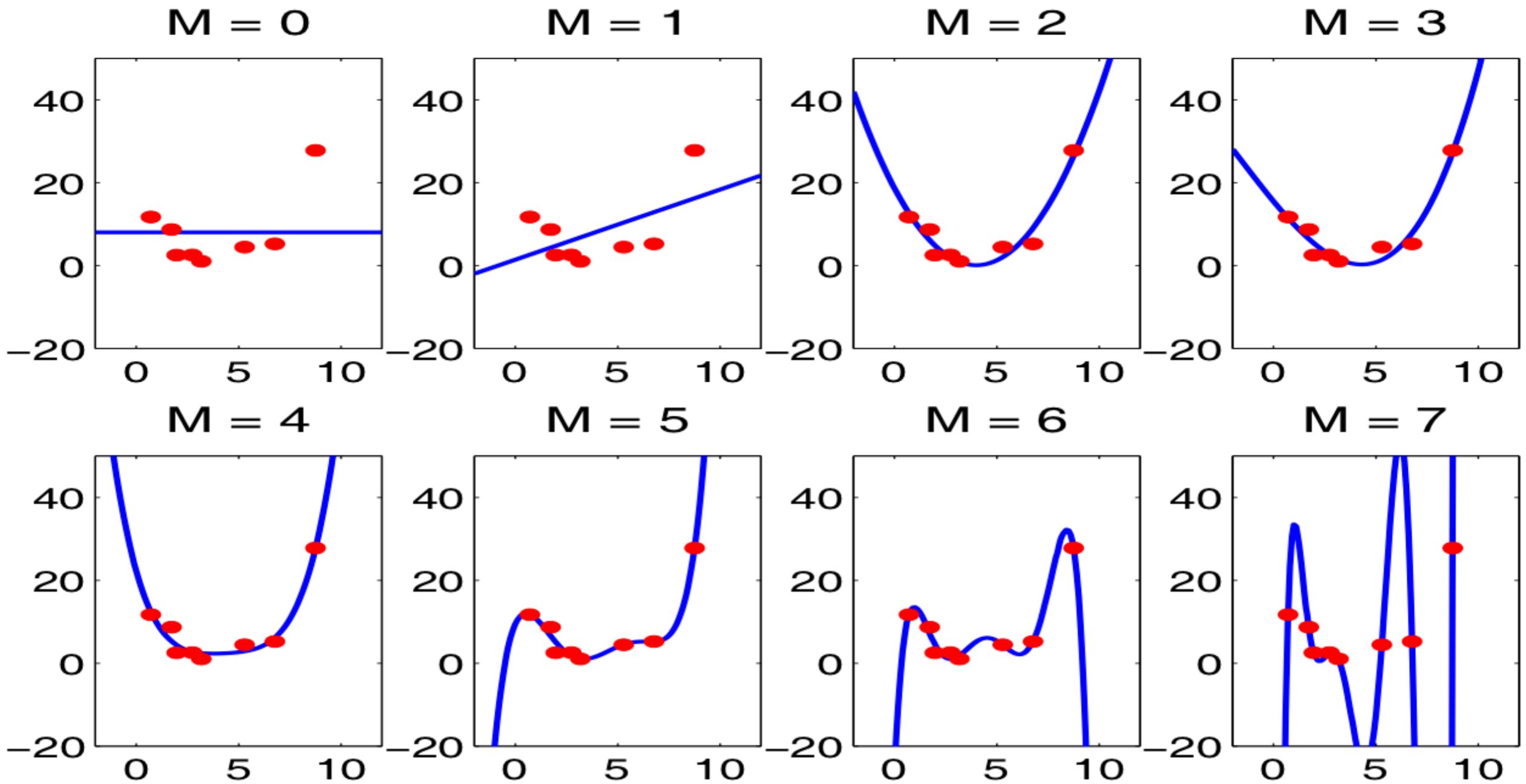
Outline

- ~~Gaussian Process Bait and Switch~~
- ~~Bayesian Statistics~~
- Marginal Likelihood

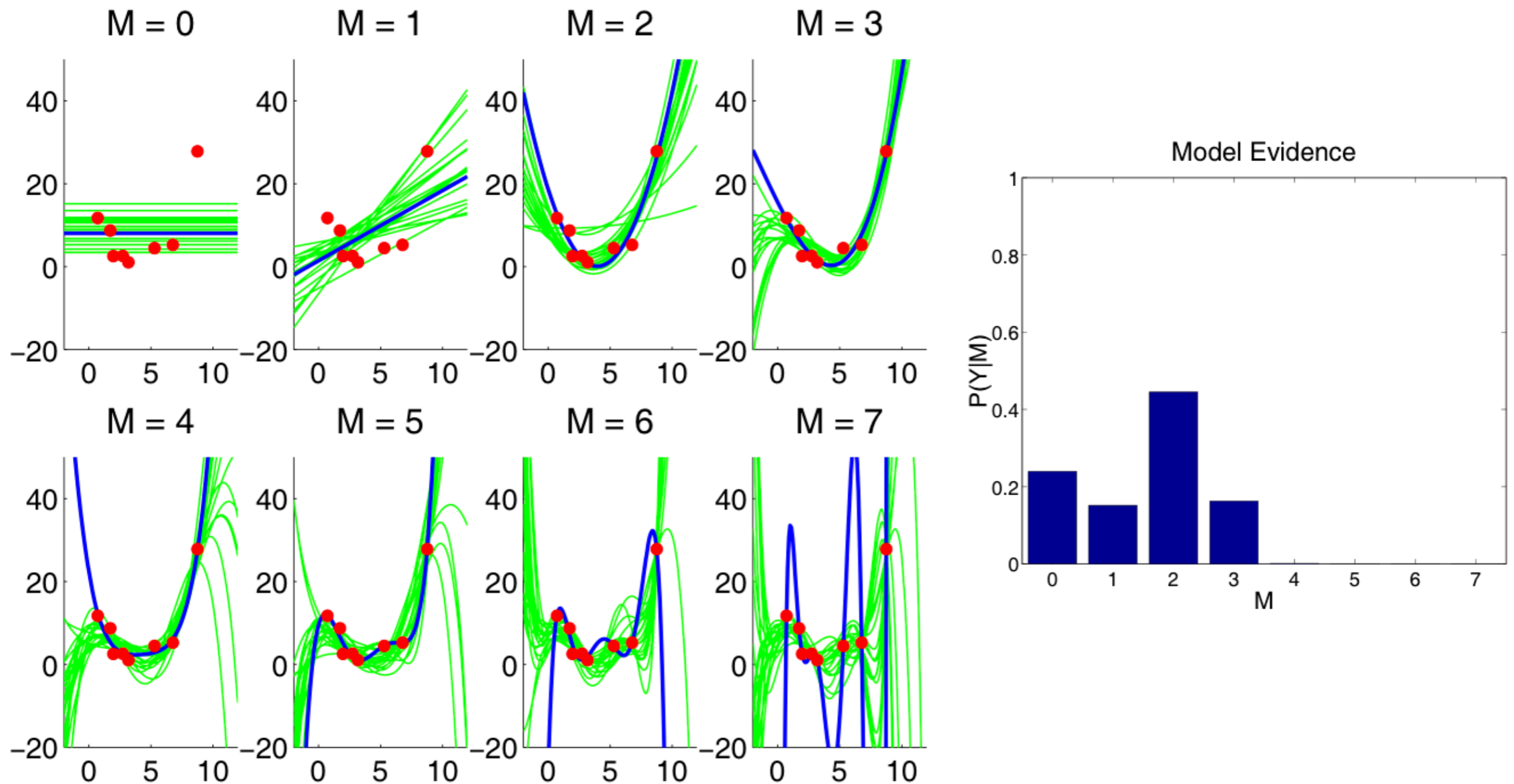
Marginal Likelihood

- **Marginal likelihood** is the denominator:
 - $p(y | x) = \int p(y | x, w)p(w)dw.$
- *Likelihood of data given hypothesis class:*
 - $p(y | x, H_0) = \int p(y | x, w)p_0(w)dw$
 - $p(y | x, H_1) = \int p(y | x, w)p_1(w)dw$
 - Called 'evidence' instead of 'likelihood'.
 - H_1 can include H_0
- Alternative to cross-validation. (???)

Example: Polynomial Regression



Example: Polynomial Regression



- Idea: *favours simplest model that explains data.*
- But note: doesn't say whether any of your models makes sense.

Bayesian Model Selection

- Which hypothesis class should we use?
- Bayesian model selection idea 1:
 - Choose H_i to maximize marginal likelihood.
- Bayesian model averaging:
 - Integrate over H_i , weighted by posterior (harder)
- Bayesian model selection idea 2:
 - Optimize parameters of H
 - Type II Maximum Likelihood (or Type II MAP)

Type II Maximum Likelihood

- Maximum likelihood:
 - $\operatorname{argmax}_w p(y \mid x, w)$
- MAP:
 - $\operatorname{argmax}_w p(y \mid x, w)p(w)$
- Type II maximum likelihood:
 - $\operatorname{argmax}_\alpha p(y \mid x, \alpha) = \int p(y \mid x, w)p(w \mid \alpha)dw$
- Type II MAP:
 - $\operatorname{argmax}_\alpha p(y \mid x, \alpha)p(\alpha)$.

Type II ML for GPs

- Type II ML for Gaussian processes:
 - $\operatorname{argmin}_{\alpha} y^T (K(\alpha) + \sigma^2 I) y + \log \det(K(\alpha) + \sigma^2 I)$
- Parameters α could be strength of prior:
 - $-\log p(w) = \|w_i\|^2 / \alpha$
- Use one α_i for each $w_i \Rightarrow$ variable selection
 - automatic relevance determination (ARD).
- Use one α_i for each $x_i \Rightarrow$ example selection
 - relevance vector machine (RVM).

ARD Prior

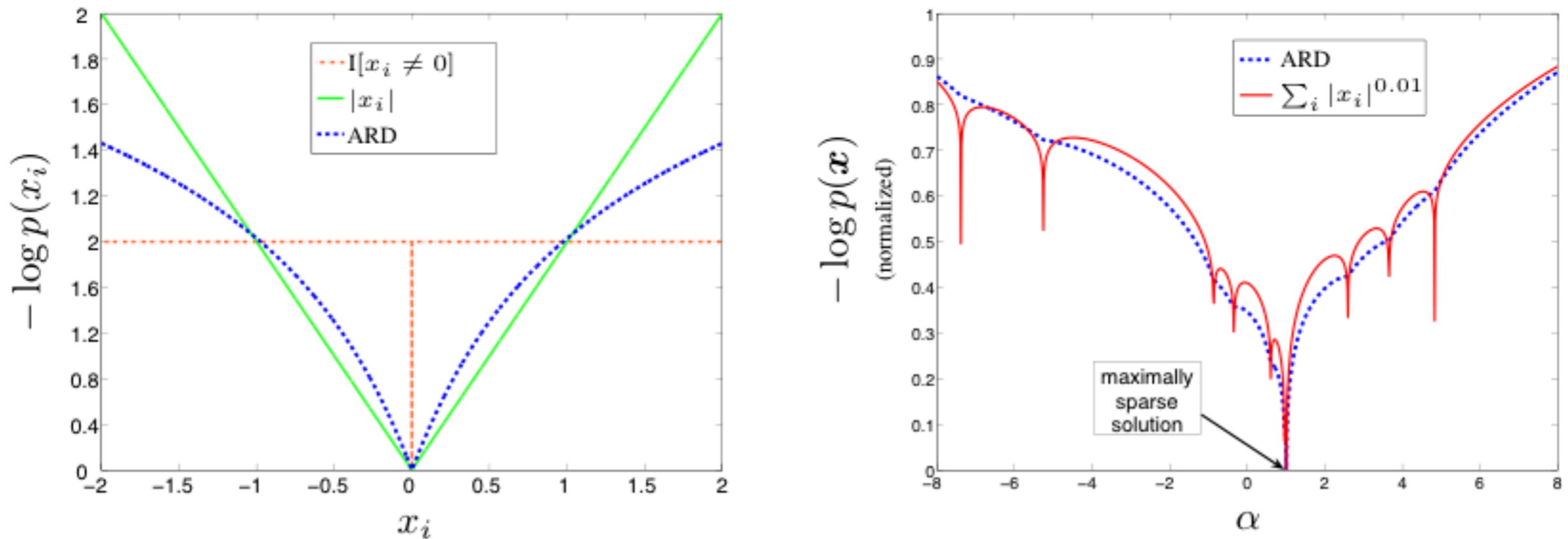


Figure 1: *Left*: 1D example of the implicit ARD prior. The ℓ_1 and ℓ_0 norms are included for comparison. *Right*: Plot of the ARD prior across the feasible region as parameterized by α . A factorial prior given by $-\log p(\mathbf{x}) \propto \sum_i |x_i|^{0.01} \approx \|\mathbf{x}\|_0$ is included for comparison. Both approximations to the ℓ_0 norm retain the correct global minimum, but only ARD smooths out local minima.

- Sparser solutions than L1-regularization.
- Fewer local minima than Lp-regularization ($p < 1$)

Sparseness of RVM

Total Number of Classification Errors and Average Number of Retained Kernel Basis Functions (in Parentheses, Rounded to the Nearest Integer) for Various Classifiers on Six of the Seven Benchmark Data Sets Described in Table 1

	Crabs RBF	Iris RBF	FGlass RBF	AML/ALL linear	Colon linear	Yeast linear
SVM	2 (53)	5 (124)	62 (365)	3 (31)	75 (32)	7 (155)
RVM	0 (4)	10 (37)	61 (74)	5 (3)	84 (6)	12 (49)

Conclusion

- MAP estimation:
 - it's easy to make work
 - but sometimes it does weird stuff
- Bayesian:
 - it's hard to make work
 - but sometimes it makes more sense