To Do Today

1. **Hand back assignments**
2. **FCQs**
3. **Rob Lindsey**

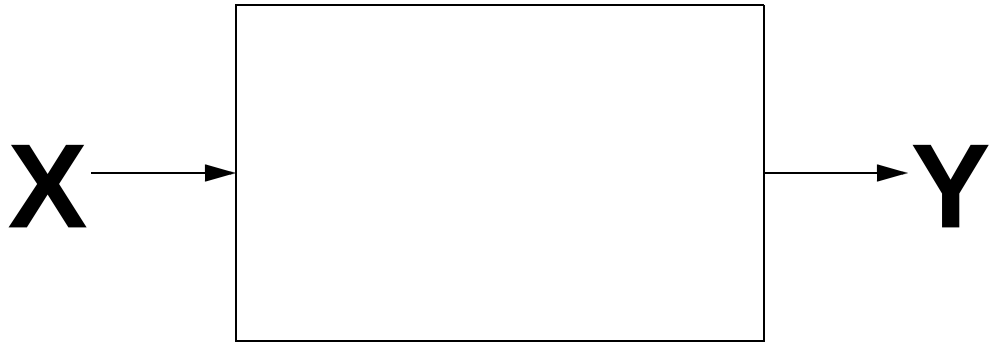
   Cool research project involving Gaussian Processes

4. **Gaussian Processes**

Gaussian Processes

For Regression

(And Classification)



Problems that GPs Solve

1. Overfitting

Large number of free parameters in models relative to number of training examples

In practice, GPs useful for problems with limited data

2. Uncertainty in prediction

Want not only to predict, but to estimate uncertainty in prediction

3. Offers a way of specifying prior knowledge in *function* space, not *parameter* space

Prior is specified over function (solution) space

Function space is more intuitive, useful

e.g., smoothness constraints on functions, trends over time, etc.

How Do We Deal With Many Parameters, Little Data?

1. Regularization

e.g., smoothing, L1 regularization, drop out in neural nets, large K for K-NN

2. Bayesian approach

specify probability of the data given weights, $P(D|W)$

specify weight priors given hyperparameter α, $P(W|\alpha)$

find posterior over weights given data, $P(W|\alpha,D)$

predict using either MAP weights or Bayesian model averaging:

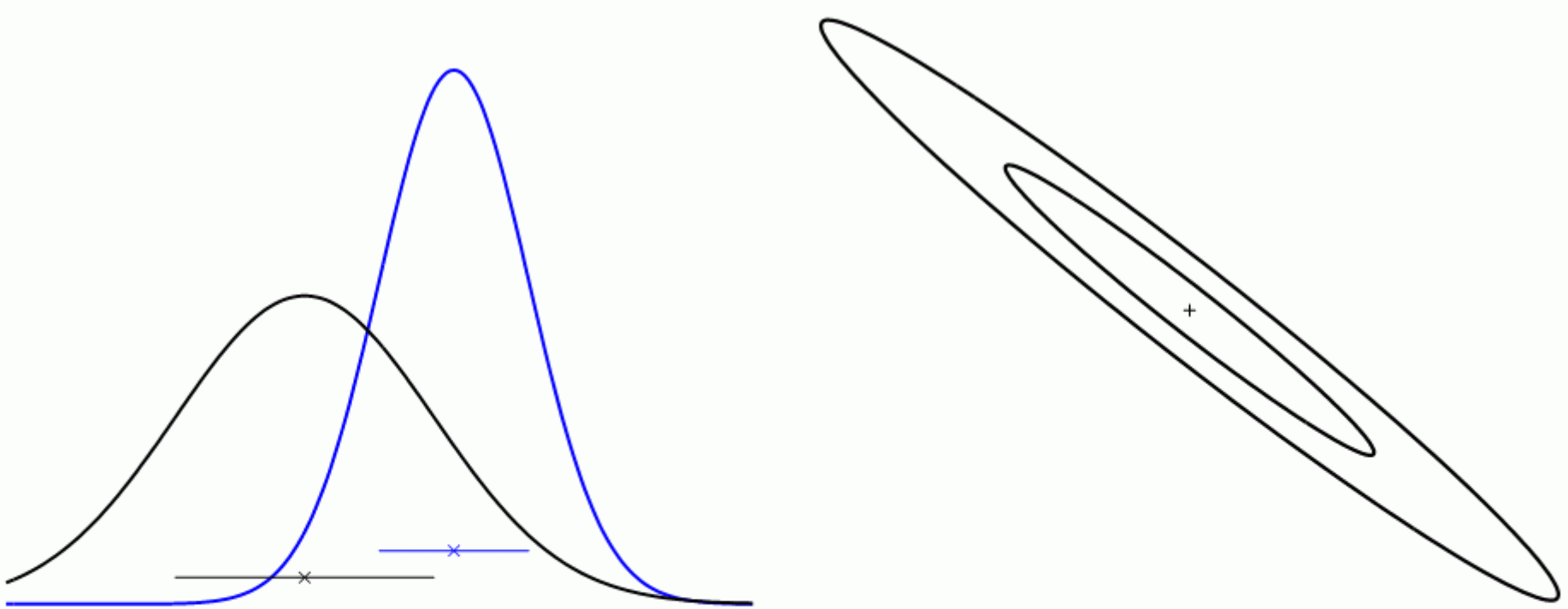
$P(y|x) = \Sigma_w \, p(y|x,w) \, P(w|\alpha,D)$

Although weight prior corresponds to function prior, often difficult to determine what that function prior looks like

3. Gaussian processes

Bayesian approach using a Gaussian process prior over functions rather than over weights

Gaussian Distributions: A Reminder



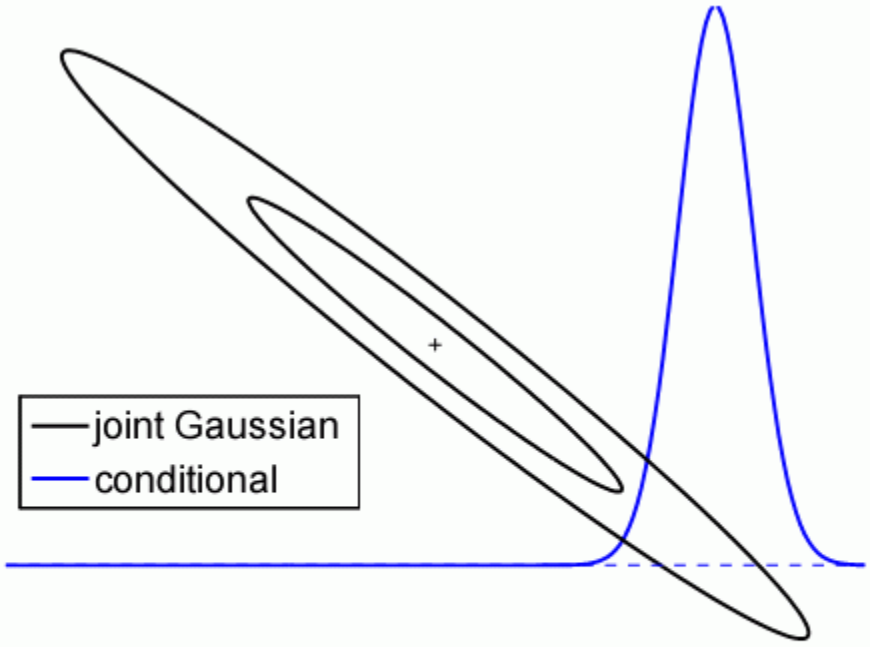
The Gaussian distribution is given by

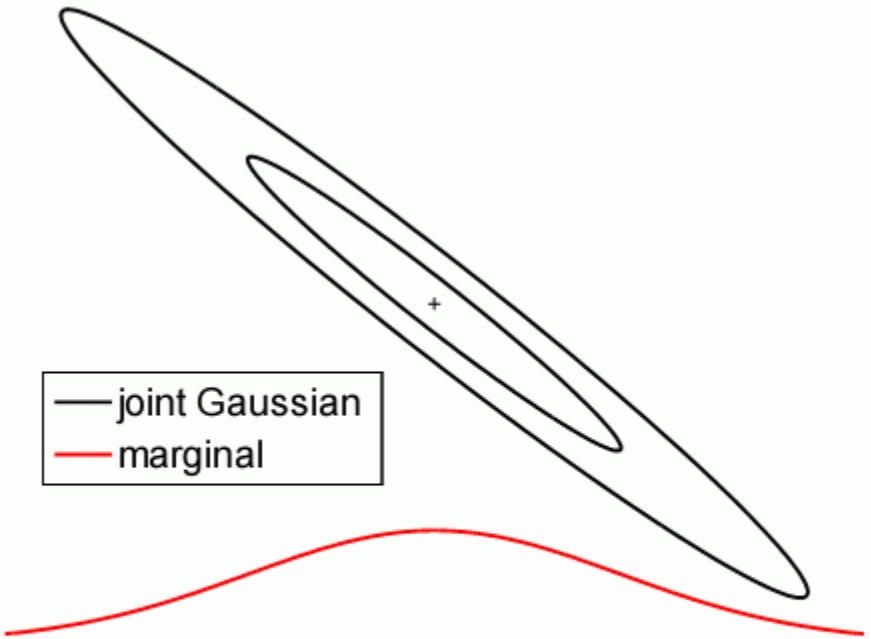
$$p(\mathbf{x}|\mu, \Sigma) = \mathcal{N}(\mu, \Sigma) = (2\pi)^{-D/2}|\Sigma|^{-1/2} \exp\left(-\tfrac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right)$$

where μ is the mean vector and Σ the covariance matrix.

Slides stolen from Karl Rasmussen NIPS 2006 tutorial

Conditionals and Marginals of Gaussians





Both the conditionals and the marginals of a joint Gaussian are again Gaussian.

Stochastic Processes

Generalization of multivariate distribution to infinitely many variables

collection of RVs Y(x) indexed by x

Often, x is time index

Today, x is input vector: x_i is the i'th input vector; $Y(x_i)$ is the output, an RV

joint probability distribution can be specified over any subset of x, e.g., $(Y(x_1), Y(x_2), \ldots Y(x_k))$

Dirichlet process: infinite dimensional Dirichlet distribution

Infinite dimensional Dirichlet distribution

x: possible word in topic models, possible mixture component in mixture models

joint pdf on any subset of x is a Dirichlet distribution

Gaussian process: infinite dimensional Gaussian distribution

Stochastic process in which $P(Y(x_1), Y(x_2), \ldots Y(x_k))$ is multivariate Gaussian

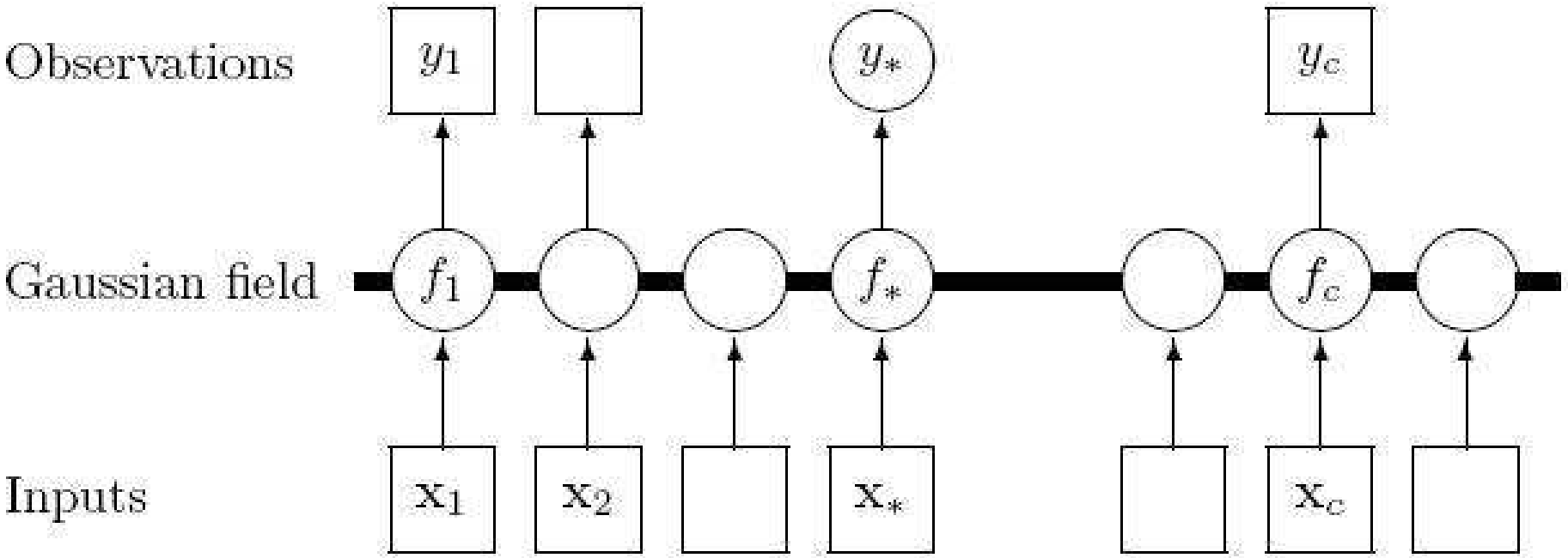
Gaussian *distribution* is for a finite set of variables, specified by a mean vector and covariance matrix

Distribution parameters: $\mu(x_i) = E[\, Y(x_i) \,]$

$C(x_i, x_j) = E[\, (Y(x_i) - \mu(x_i)) \, (Y(x_j) - \mu(x_j)) \,]$

Gaussian *process* is defined for an infinite set of variables, specified by a mean *function* and covariance *function*

Graphical Model Depiction of Gaussian Process



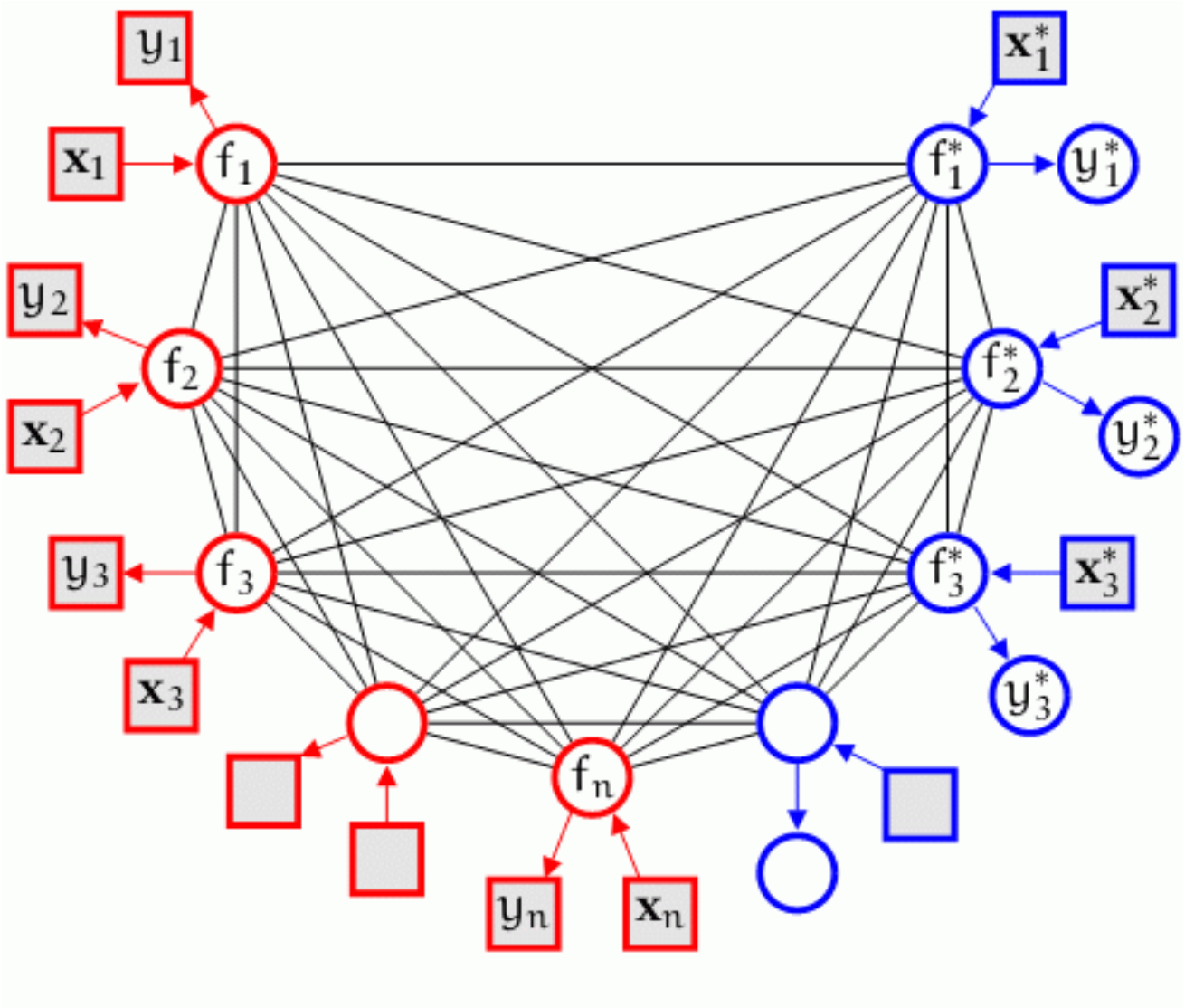
rectangle: observation; round node: free variables

Gaussian field: **f** is multivariate normal for any setting of **x**, **y**

solid bar: each f is connected to each other f

prediction y_* depends only on the corresponding latent f_*

adding an x_{**}, f_{**}, y_{**} does not influence the distribution



This and other slides stolen from Karl Rasmussen (NIPS 2006 tutorial)

Illustration of a Gaussian Process I

Example one dimensional Gaussian process:

$$p(f(x)) \sim \mathcal{GP}\big(m(x) = 0,\ k(x, x') = \exp(-\tfrac{1}{2}(x - x')^2)\big).$$

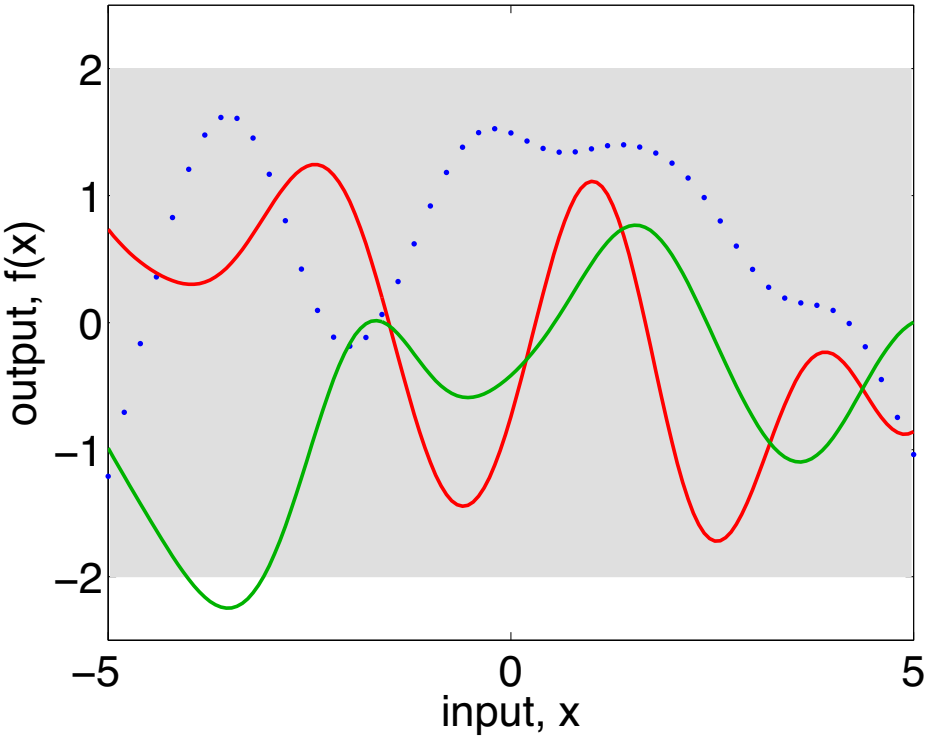
To get an indication of what this distribution over functions looks like, focus on a finite subset of function values $\mathbf{f} = (f(x_1), f(x_2), \ldots, f(x_n))^\top$, for which

$$\mathbf{f} \sim \mathcal{N}(0, \Sigma),$$

where $\Sigma_{ij} = k(x_i, x_j)$.

Then plot the coordinates of f as a function of the corresponding x values.

draws from GP prior



GP psterior

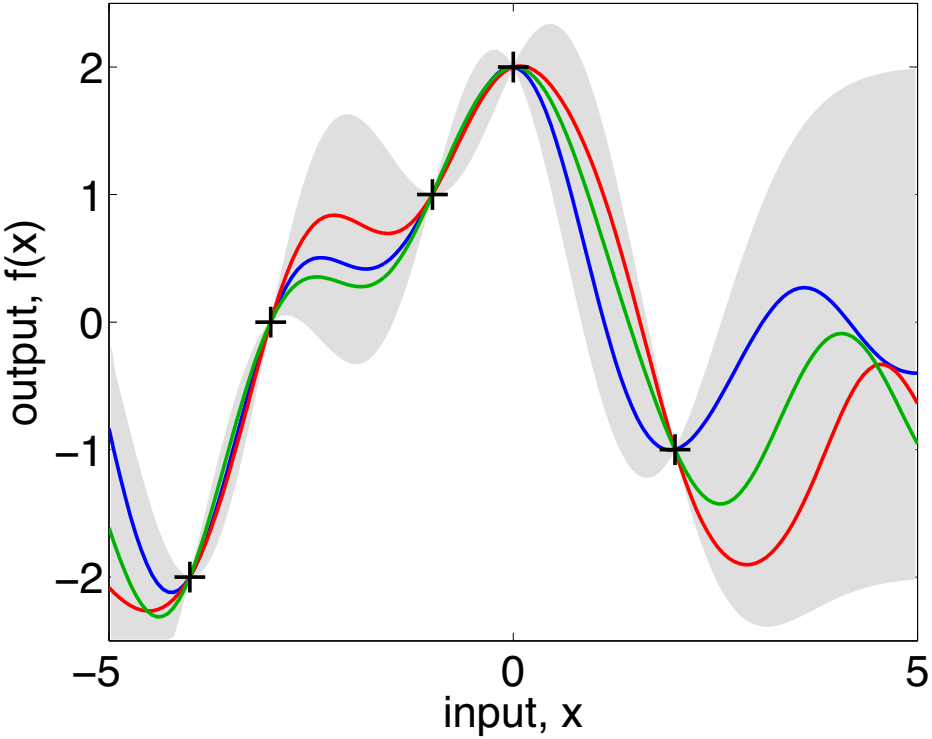


Illustration of Gaussian Process II

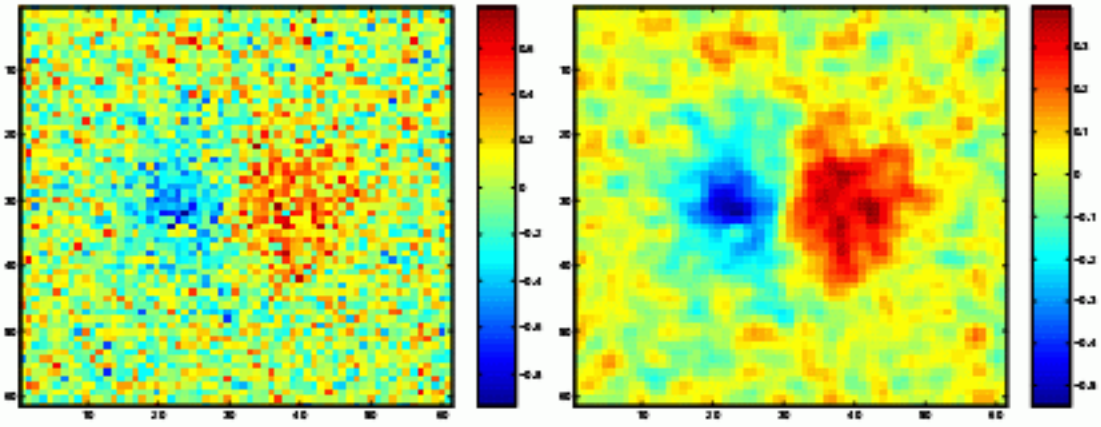
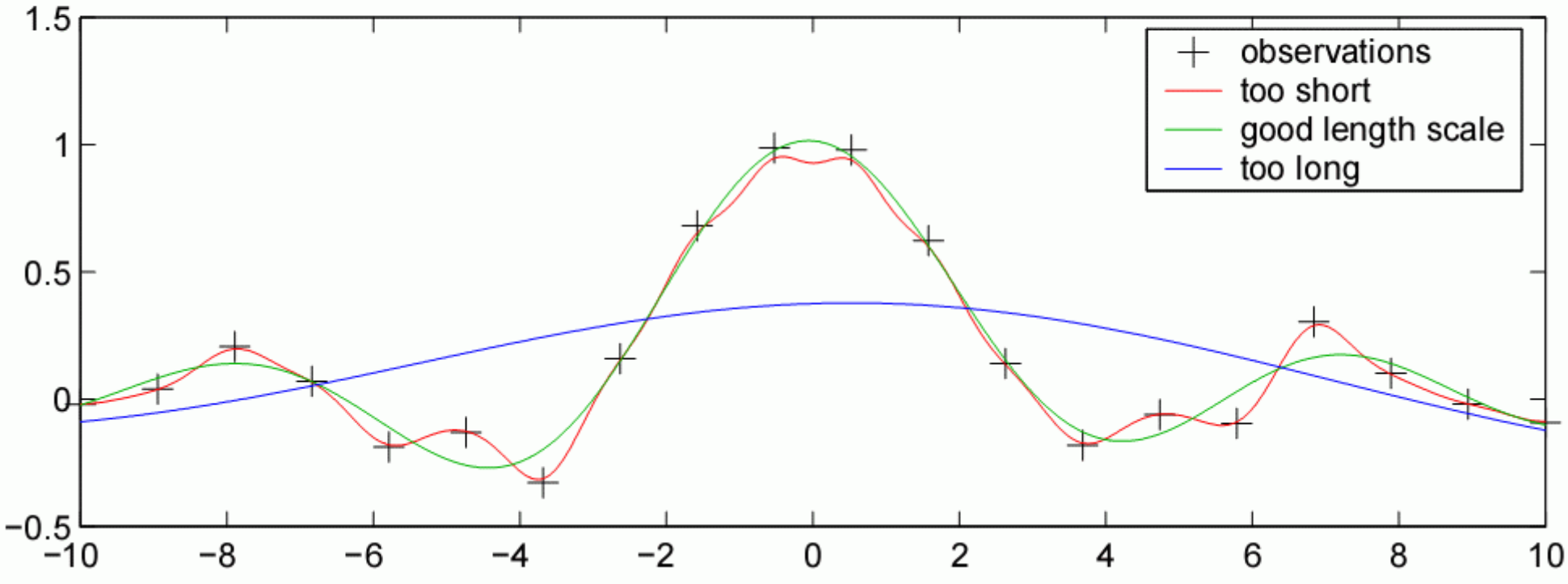


Figure 1: The first image is the realization of white noise which would be unrealistic for modelling room temperature due to discontinuity. The second image is the realization of a Gaussian field which show realistic continuous measurements.

Illustration of a Gaussian Process III

Parameterized covariance function: $k(x, x') = v^2 \exp\left(-\frac{(x - x')^2}{2\ell^2}\right) + \sigma_n^2 \delta_{xx'}$.



Quiz

How do ℓ, ν, σ_n^2 affect shape of function?

How Do We Deal With Infinite Dimensionality Of GP?

Thinking of a GP as a Gaussian distribution with an infinitely long mean vector and an infinite by infinite covariance matrix may seem impractical...

...luckily we are saved by the *marginalization property*:

Recall:

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{y}) d\mathbf{y}.$$

For Gaussians:

$$p(\mathbf{x}, \mathbf{y}) = \mathcal{N}\left(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix}\right) \implies p(\mathbf{x}) = \mathcal{N}(\mathbf{a}, A)$$

Unknown values of function domain don't matter.

Similar to Dirichlet process

Conjugacy Of Gaussian Processes

GP prior + Gaussian likelihood -> GP posterior

Gaussian likelihood:

$$\mathbf{y}|\mathbf{x}, f(x), M_i \sim \mathcal{N}(\mathbf{f}, \sigma^2_{\text{noise}} I)$$

(Zero mean) Gaussian process prior:

$$f(x)|M_i \sim \mathcal{GP}(m(x) \equiv 0, k(x, x'))$$

Leads to a Gaussian process posterior

$$f(x)|\mathbf{x}, \mathbf{y}, M_i \sim \mathcal{GP}(m_{\text{post}}(x) = k(x, \mathbf{x})[K(\mathbf{x}, \mathbf{x}) + \sigma^2_{\text{noise}} I]^{-1}\mathbf{y},$$
$$k_{\text{post}}(x, x') = k(x, x') - k(x, \mathbf{x})[K(\mathbf{x}, \mathbf{x}) + \sigma^2_{\text{noise}} I]^{-1} k(\mathbf{x}, x')).$$

Prediction Using Gaussian Processes

Suppose we have i = 1...N training examples

$y_i = f(x_i)$ and we want to predict function value for a new point $y_* = f(x_*)$

Assuming GP with μ = 0 and covariance function

$K+ = \begin{bmatrix} K & k_* \\ k_*^T & k_{**} \end{bmatrix}$ with $k_{**} = C(x_*, x_*)$

$k_* = C(\underline{x}, x_*)$

we can compute $P(Y_* \mid X_*, X_{1...N}, Y_{1...N})$, which is Gaussian with

$$\bar{f}_* = \mathbf{k}_*^\top (K + \sigma_n^2 I)^{-1} \mathbf{y},$$

$$\mathbb{V}[f_*] = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top (K + \sigma_n^2 I)^{-1} \mathbf{k}_*$$

Prediction can be framed as a linear combination of N kernel functions, each one centered on a training point

Like SVMs (unlike NNs), training data must be saved.

Cost of prediction is in doing matrix inversion, $O(N^3)$

What Do GPs Buy Us?

- **Smoothness hyperparameter on functions seems a good way to incorporate domain knowledge.**
- **Benefit of Bayesian methods in general is greatest with limited data.**
- **Predictions with error bars!**

STOP HERE

Intuitive Introduction to GPs: Linear Regression

Linear regression using a fixed set of *m* basis functions $\phi_i(x)$

$Y(x) = \sum_i W_i \phi_i(x)$

Weight space view

assume Gaussian prior on weights: $W \sim N(0, \Sigma_w)$ and Gaussian output noise (likelihood function)

Given these assumptions, posterior on weights is also Gaussian

Therefore, can do Bayesian thing in weight space

Function space view

What sort of functions can be generated from a fixed set of basis functions with random weights? $W \sim N(0, \Sigma_w)$

Gaussian Process! i.e., $E_w[Y(x)] = 0$,
$C(x,x') = E_w [Y(x) Y(x')] = \phi^T(x) Sw \phi(x')$

Two views are equivalent in the case of linear regression, given $\phi_i(x)$ and the weight prior

Gaussian Processes: General Case

Previous example assumed one particular covariance function

Many covariance functions are possible

Learning = search for parameters of covariance function

Any covariance function corresponds to a particular (infinite) choice of basis functions

- In general, GPs can be viewed as Bayesian linear regression with an infinite number of (unspecified) basis functions
- Like SVM kernels: kernels specify distance in (unknown and possibly infinite dimensional) space

Do not need to specify GP over entire function space, but only over the set of training points and the test point.

Covariance Function

Squared-exponential function

$$k_y(x_p, x_q) = \sigma_f^2 \exp\left(-\frac{1}{2\ell^2}(x_p - x_q)^2\right) + \sigma_n^2 \delta_{pq}$$

-