Kernel Regression & Gaussian Processes

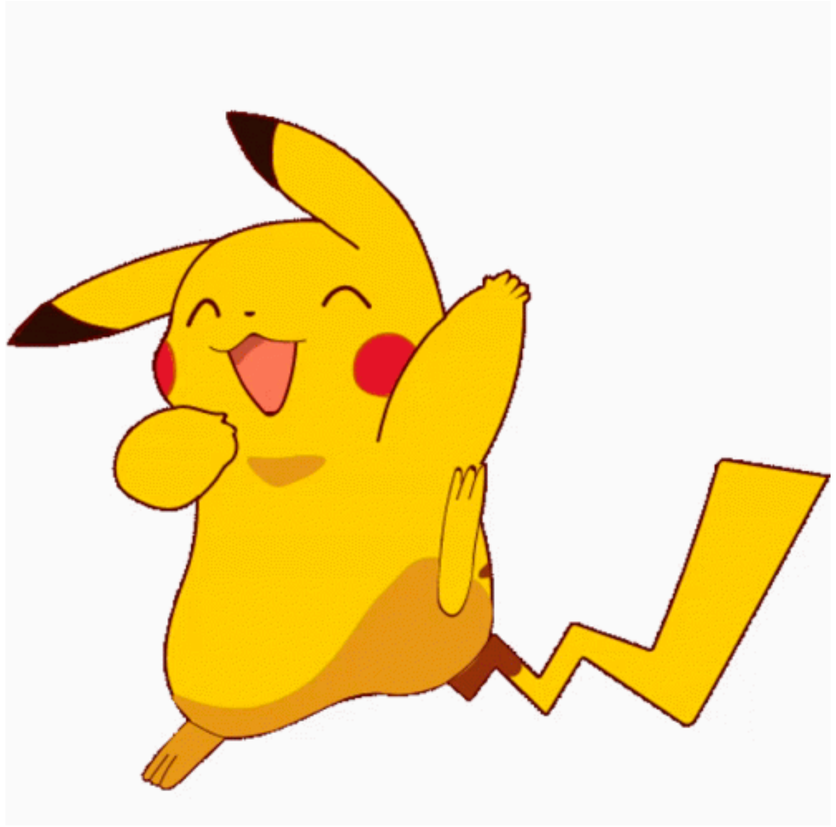
Beyond CS 181 - Lecture 1

Anna Trella

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Welcome to Beyond CS 181

Thank you for making the trek over here on a Friday!



HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Logistics

- **Completely optional**
- **No math, no readings, no assignments**
- **Each section is standalone, high-level exploration**

Motivation

Motivation

- Bridge between foundations (what you learn in class) and research
- Practical applications of theory
- Providing awareness of opportunities (classes or research groups) at Harvard

Motivation

- Bridge between foundations (what you learn in class) and research
- Practical applications of theory
- Providing awareness of opportunities (classes or research groups) at Harvard
- I think these topics are super fun and hope I can share it with y'all!

Me
ML Knowledge
You

Kernel Regression & Gaussian Processes

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Linear Regression

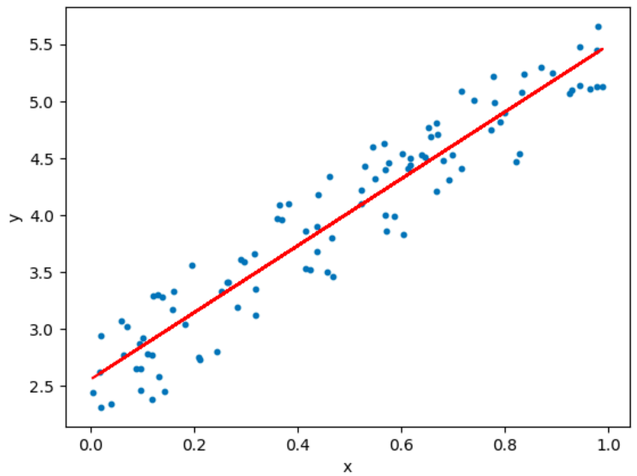
Setup:

$$y_i = x_i^T w + \epsilon, \quad \epsilon \sim \mathcal{N}(0,1)$$

Objective: Find

$$w^* = \arg\max p(Y \mid X, w)$$

$$= (X^T X)^{-1} X^T Y$$



HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Linear Regression

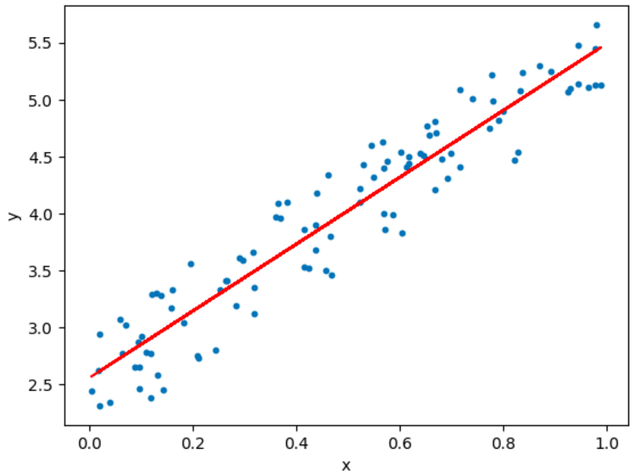
Setup:

$$y_i = x_i^T w + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_n^2)$$

Objective: Find

$$w^* = \arg\max p(Y|X, w)$$

$$= (X^T X)^{-1} X^T Y$$



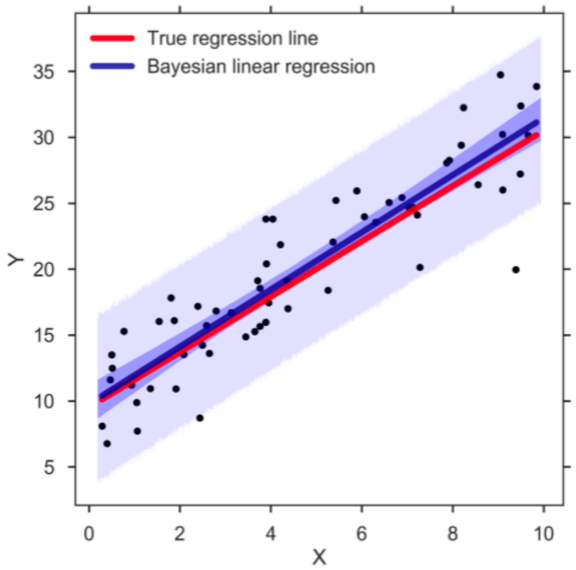
Bayesian Linear Regression

Setup:

$$y_i = x_i^T w + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_n^2), \quad w \sim \mathcal{N}(\mu, \Sigma)$$

Objective: Find

$$w^* \sim p(Y|X, w)p(w)$$



HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Kernel Regression

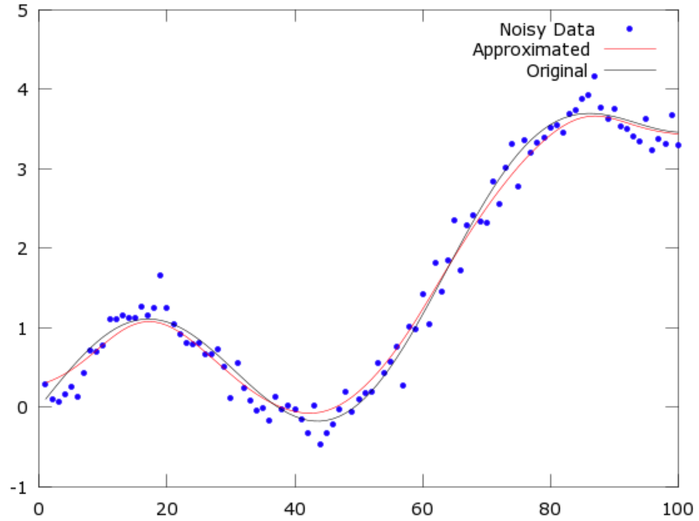
Setup:

$$y_i = \phi(x_i)^T w + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_n^2)$$

Objective: Find

$$w^* = \arg\max p(Y \mid X, w)$$

$$= (\Phi^T \Phi)^{-1} \Phi^T Y$$



Kernel Regression

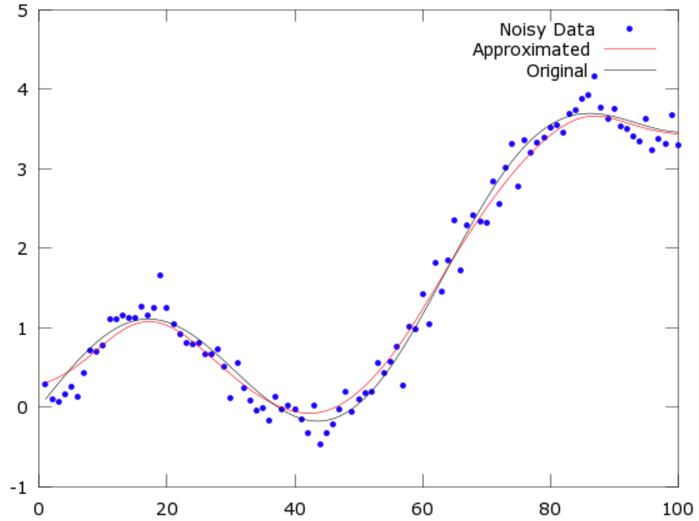
Setup:

$$y_i = \phi(x_i)^T w + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_n^2)$$

Objective: Find

$$w^* = \arg\max p(Y|X, w)$$

$$= (\Phi^T \Phi)^{-1} \Phi^T Y$$



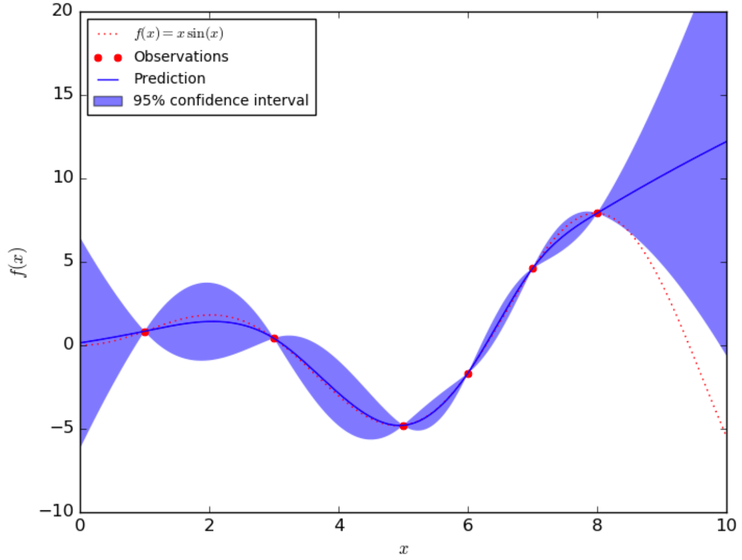
Bayesian Kernel Regression

Setup:

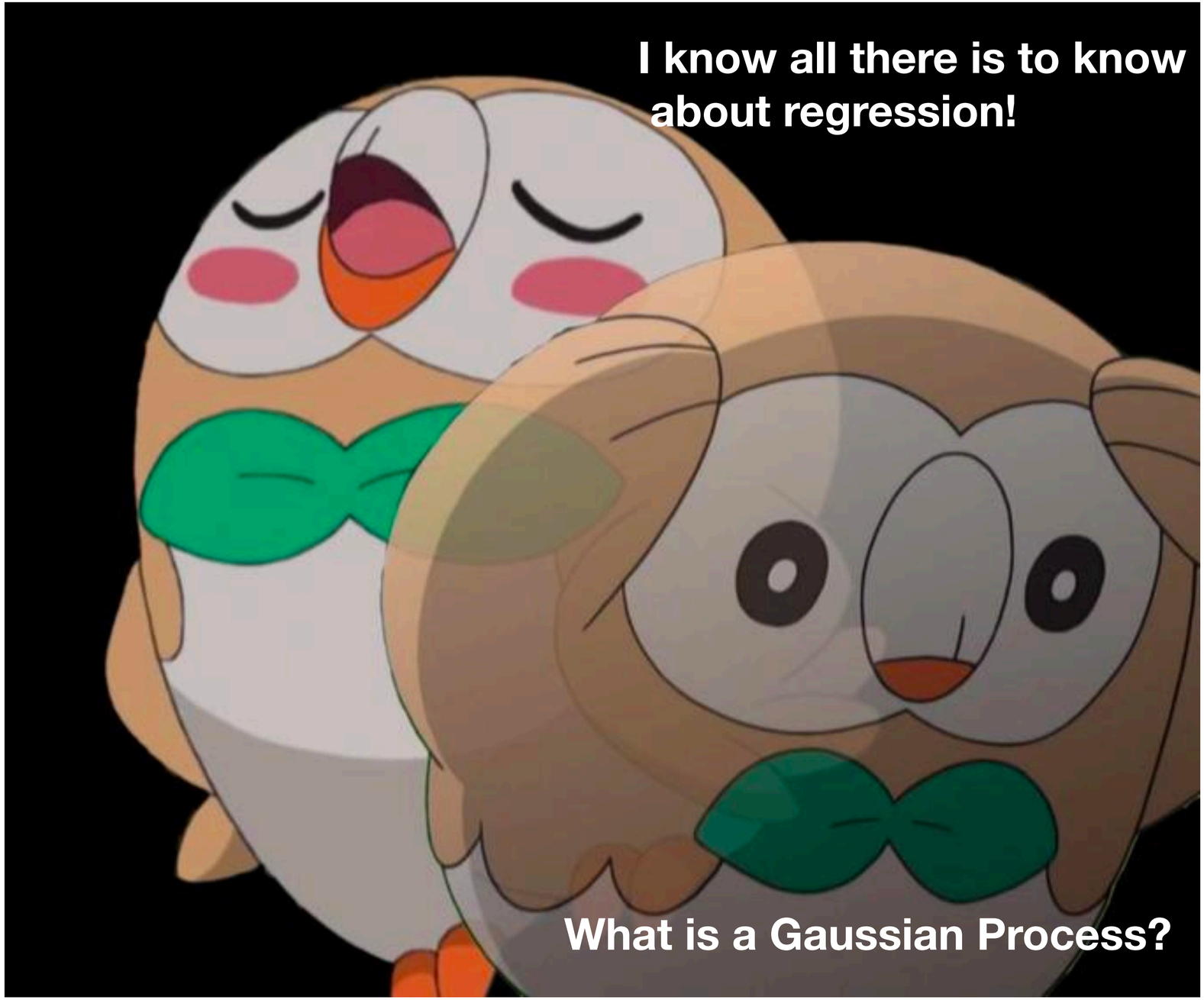
$$y_i = \phi(x_i)^T w + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_n^2), \quad w \sim \mathcal{N}(\mu, \Sigma)$$

Objective: Find

$$w^* \sim p(Y|X, w)p(w)$$

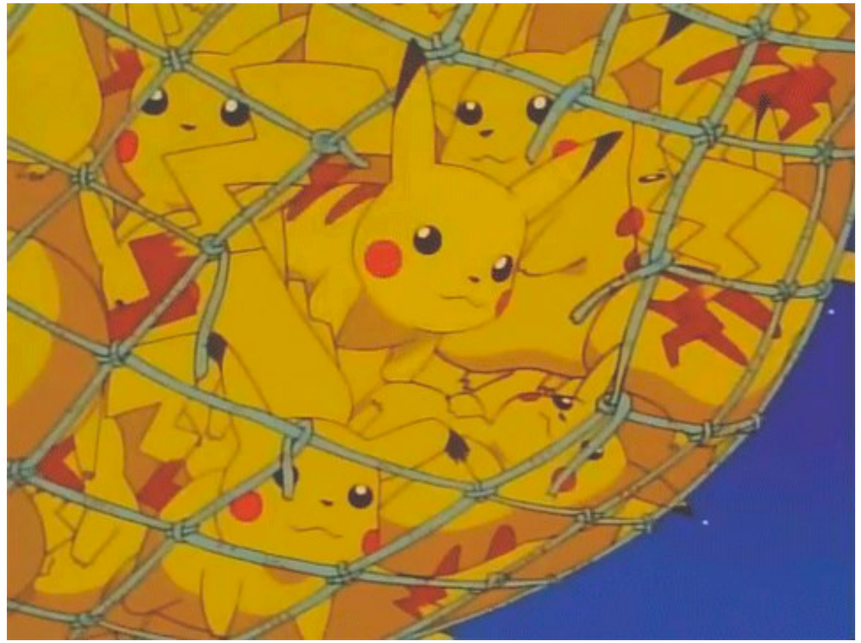


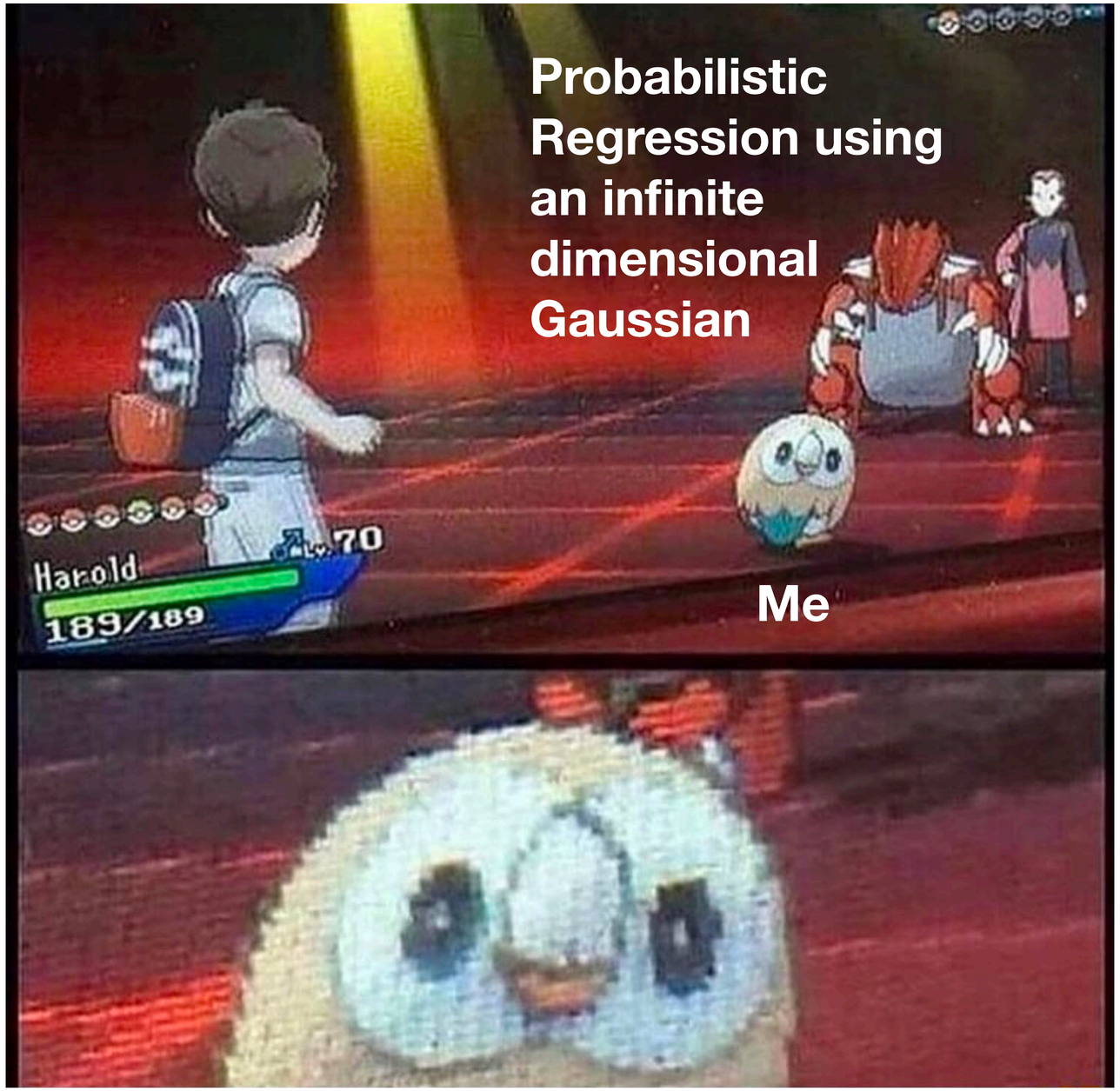
HARVARD
John A. Paulson
School of Engineering
and Applied Sciences


I know all there is to know about regression!
What is a Gaussian Process?

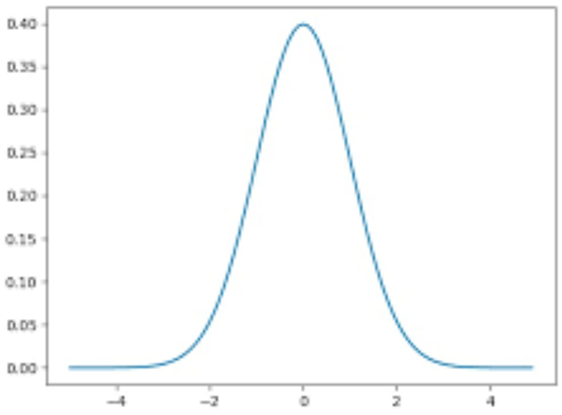
Gaussian Process Regression

We can do Bayesian kernel regression but with an infinite dimensional Gaussian!



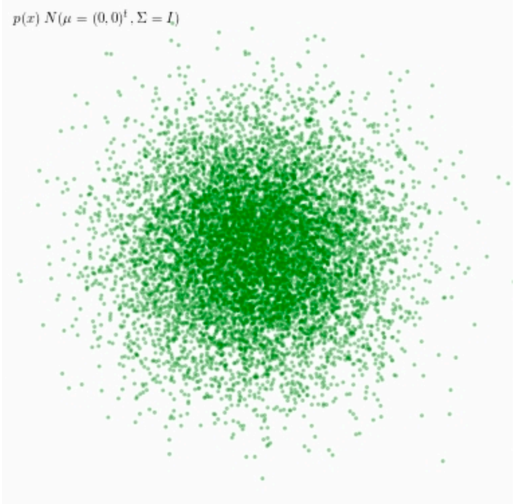

Probabilistic Regression using an infinite dimensional Gaussian
Harold
Lv. 70
189/189
Me

Gaussian Process



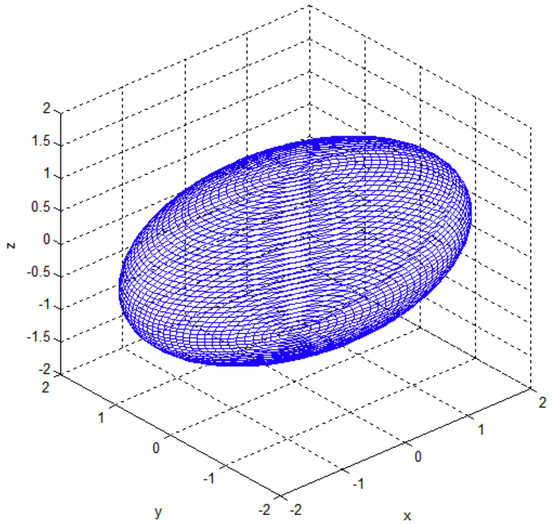
1d

$$x \sim \mathcal{N}(\mu, \sigma)$$



2d

$$[x_1, x_2] \sim \mathcal{N}([\mu_1, \mu_2], \Sigma_2)$$

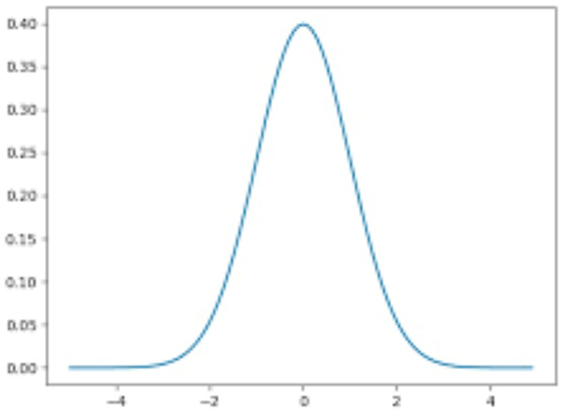


3d

$$[x_1, x_2, x_3] \sim \mathcal{N}([\mu_1, \mu_2, \mu_3], \Sigma_3)$$

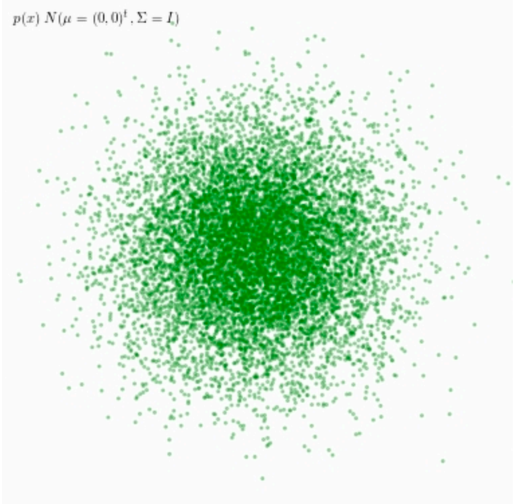
HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Gaussian Process



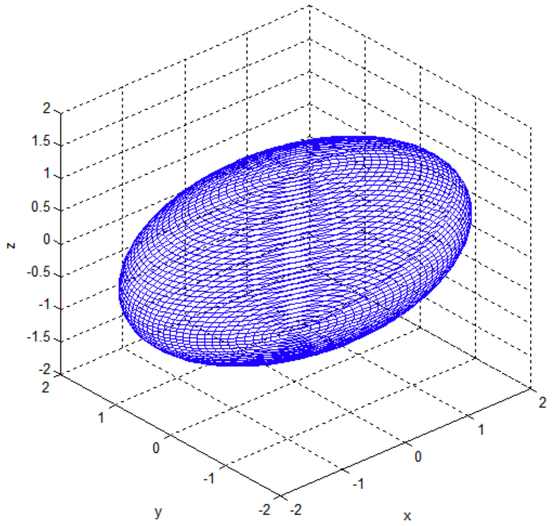
1d

$$x \sim \mathcal{N}(\mu, \sigma)$$



2d

$$[x_1, x_2] \sim \mathcal{N}([\mu_1, \mu_2], \Sigma_2)$$



3d

$$[x_1, x_2, x_3] \sim \mathcal{N}([\mu_1, \mu_2, \mu_3], \Sigma_3)$$

What does an infinite dimensional Gaussian look like?

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Gaussian Process

What does an infinite dimensional Gaussian look like?

$$[x_1, x_2, x_3, \ldots] \sim \mathcal{N}_\infty([\mu_1, \ldots], \Sigma_\infty)$$

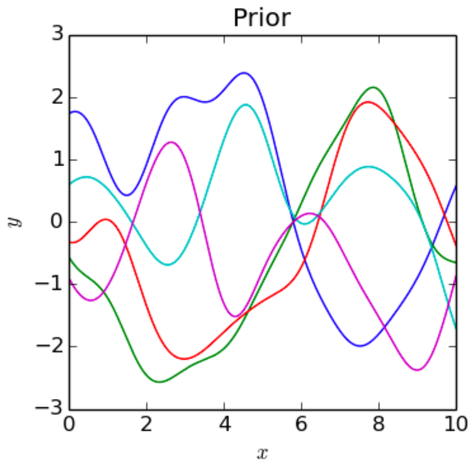
Gaussian Process

What does an infinite dimensional Gaussian look like?

$$[x_1, x_2, x_3, \ldots] \sim \mathcal{N}_\infty([\mu_1, \ldots], \Sigma_\infty)$$

We're sampling real valued functions!

$$f : \mathbb{R} \to \mathbb{R} \sim \mathcal{N}_\infty([\mu_1, \ldots], \Sigma_\infty)$$



HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

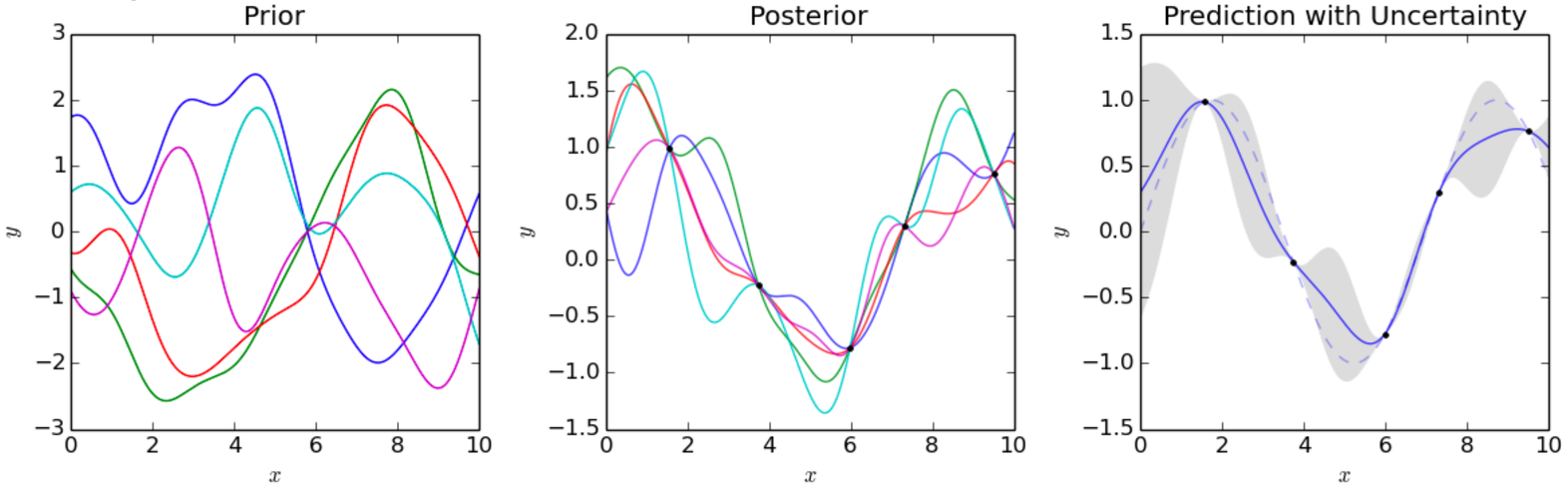
GP Process

1. Choose a prior $f_{prior} \sim \mathcal{N}(\mu, K)$

2. We sample random functions f from our prior

3. We incorporate data which rejects functions that don't agree with our data (gives us a posterior)

4. Functions from the posterior agree with the data and gives us idea of uncertainty



HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

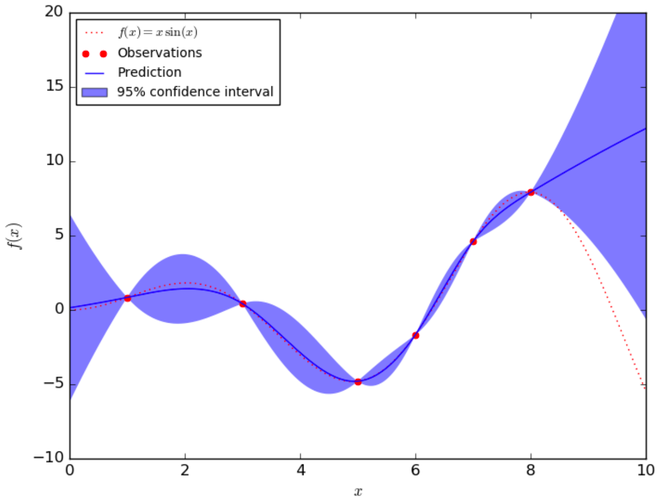
GP Process is equivalent to Bayesian Kernel Regression

Bayesian Kernel Regression

$$y_i = \phi(x_i)^T w + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_n^2), \quad w \sim \mathcal{N}(\mu, \Sigma)$$

GP

$$f \sim \mathcal{N}(\mu_{GP}, K + \sigma_n^2 I)$$



HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

GP Process is equivalent to Bayesian Kernel Regression

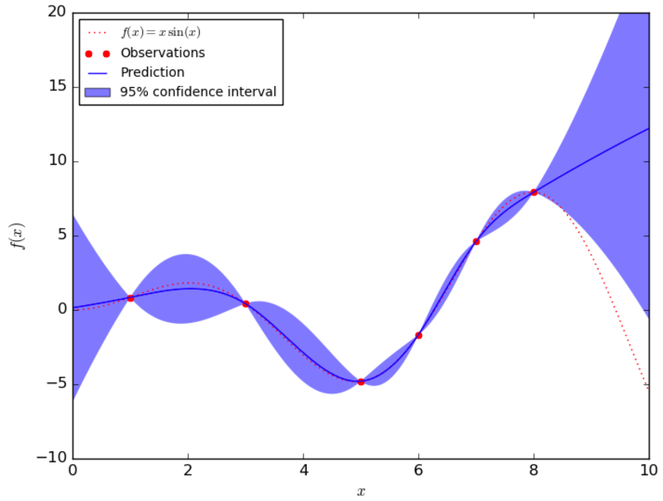
Bayesian Kernel Regression $\Longleftrightarrow$ **GP**

$$y_i = \phi(x_i)^T w + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_n^2), \quad w \sim \mathcal{N}(\mu, \Sigma)$$

$$f \sim \mathcal{N}(\mu_{GP}, K + \sigma_n^2 I)$$

Mercer's Theorem:

Every PD function $k(x, x')$ corresponds to a unique feature map $\phi(x)$ and $k(x, x') = \phi(x)^T \phi(x')$



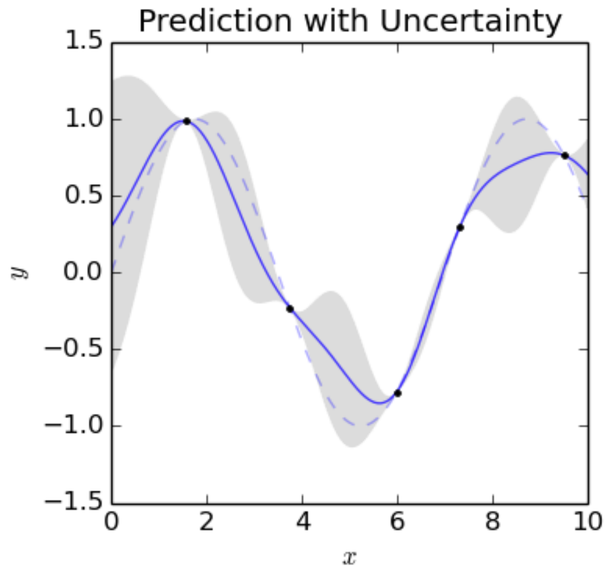
HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Why do I care about GPs?

I can do Bayesian kernel regression with infinite bases!

Outputs of my GP give me predictions AND uncertainty

$$K(x, x') = \exp\left(\frac{-||x - x'||_2^2}{\tau}\right)$$

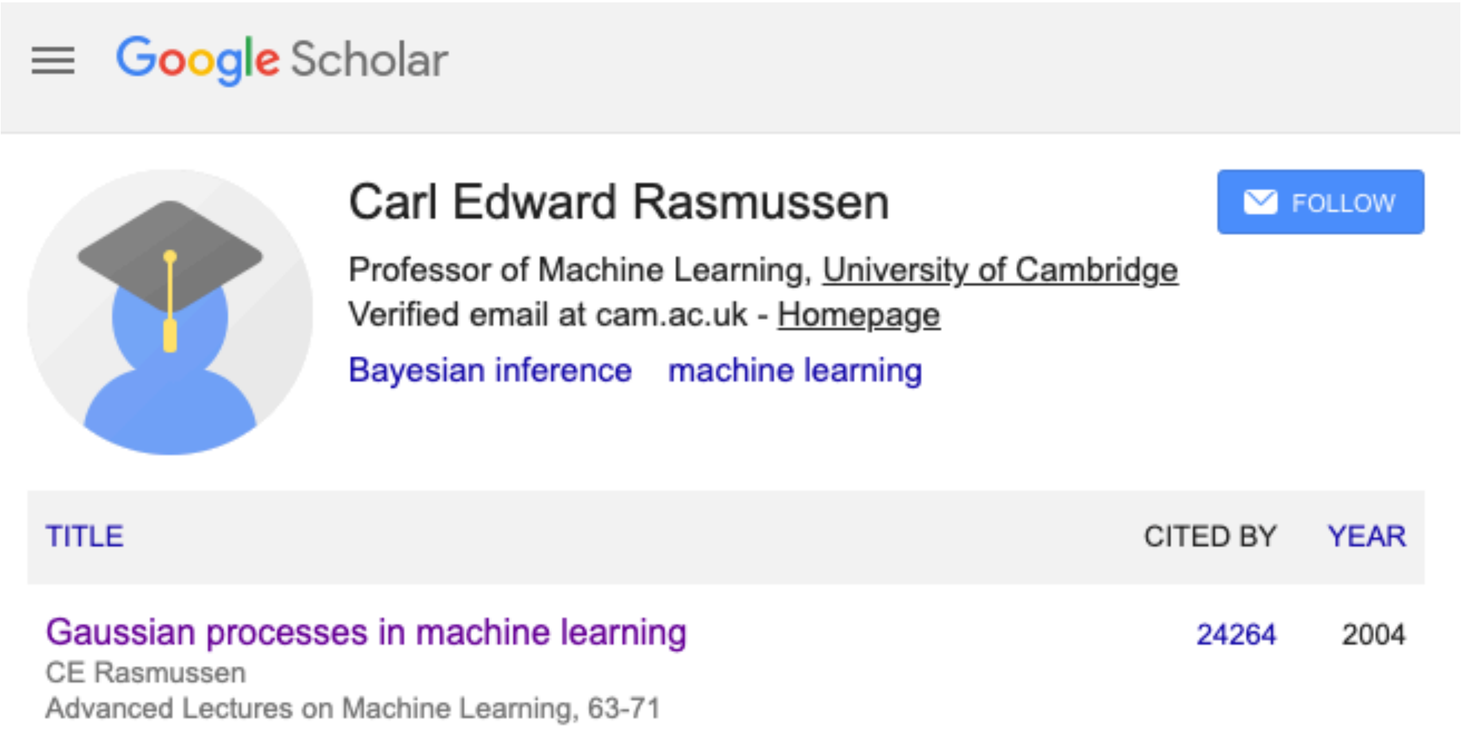


No explicit $\phi(x)$ for this kernel!

Sometimes kernel is easier to compute than basis functions.

It’s so prolific

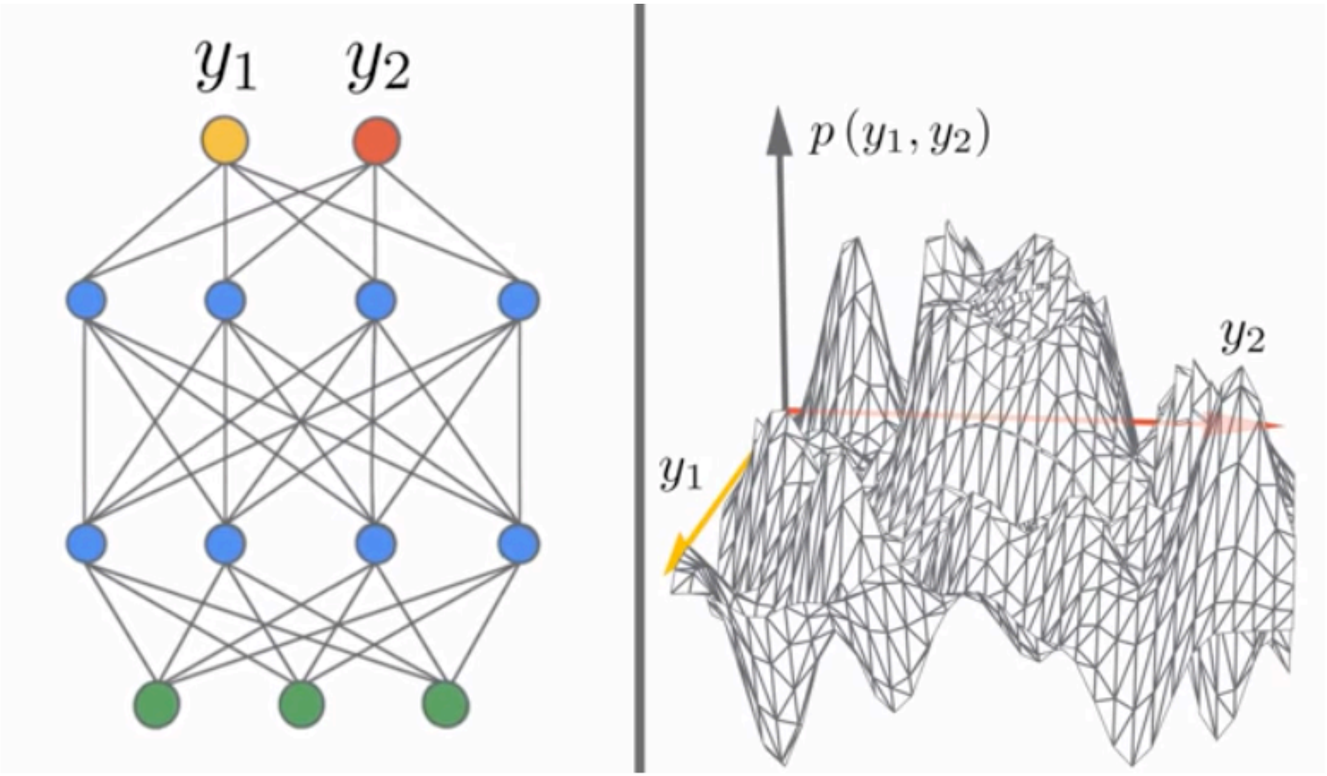
Applications in physical sciences, biological sciences, ML, Reinforcement Learning, etc.



HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

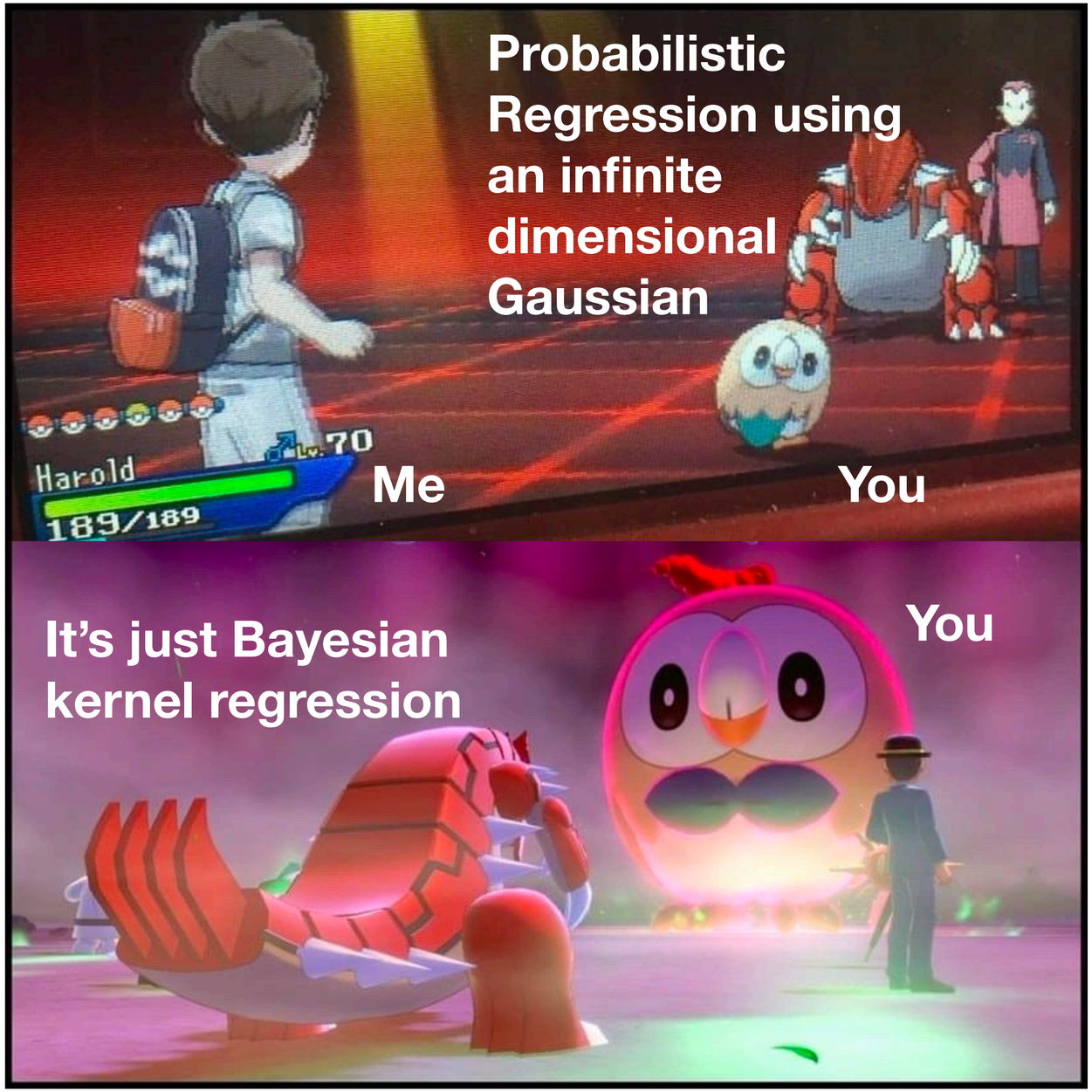
An infinitely wide Neural Network converges to a GP!

GPs provide a closed form way to evaluate Bayesian Neural Networks



HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

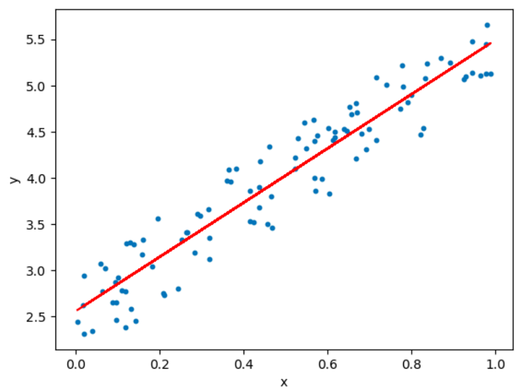
Link here


Probabilistic Regression using an infinite dimensional Gaussian
Harold
Lv. 70
189/189
Me
You
It's just Bayesian kernel regression
You

Summary and Takeaways

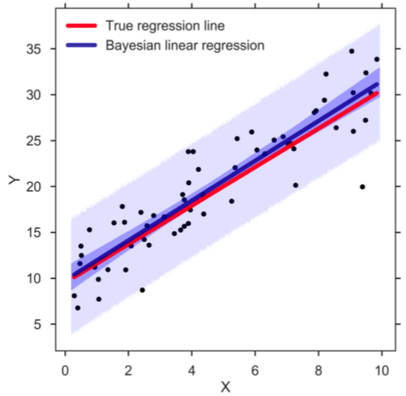
Linear Regression

Closed Form MLE, Predictions



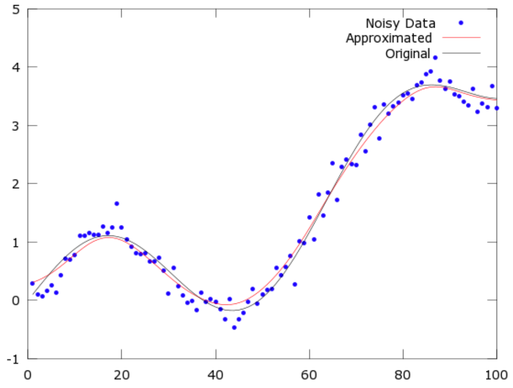
Bayesian Linear Regression

Closed Form Posterior Distribution, Predictions and Uncertainty



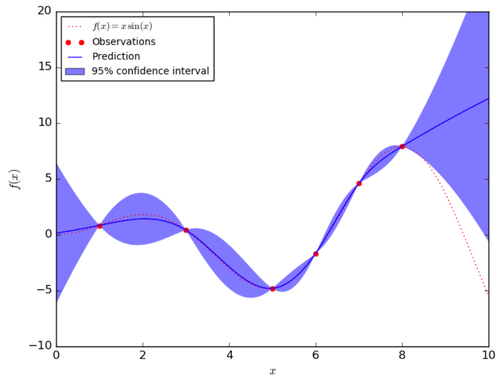
Kernel Regression

Nonlinear, Closed Form MLE, Predictions



Bayesian Kernel Regression

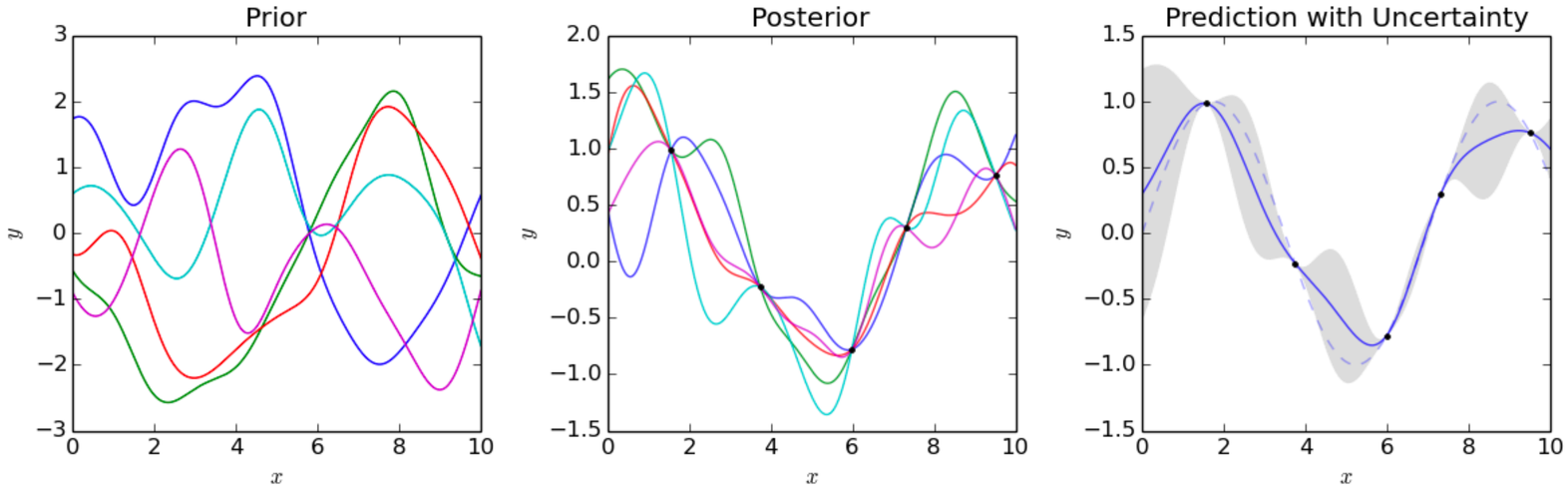
Nonlinear, Closed Form Posterior Distribution, Predictions and Uncertainty



HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

GP is a view of doing Bayesian Kernel Regression

- We assume everything is Gaussian
- We define a prior (kernel)
- We draw prior function samples from the Gaussian
- We use Gaussian marginalization to “delete” functions that don’t agree with our data



HARVARD
John A. Paulson
School of Engineering
and Applied Sciences