Introduction to PHY657

Spring 2026

Learning goals

- Learn how to apply statistical methods to find the best solution to a specific problem
  - Check whether a specific model describes an observed data set reasonably well (hypothesis testing, classical statistics)
  - Develop predictions on some specific outcomes on the basis of a set of observations (linear regression in ML)

Learning goals

- Learn how to apply statistical methods to find the best solution to a specific problem
  - Check whether a specific model describes an observed data set reasonably well (hypothesis testing, classical statistics)
  - Develop predictions on some specific outcomes on the basis of a set of observations (linear regression in ML)
- This is NOT a technical course (how to become better at python, although you should be able to improve your skills through the programming labs that you will work on); the main goal is to **deepen your knowledge of the toolkits available** to you to solve a specific problem and to reflect on the assumptions and methodology used. **What is the impact on the model assumption in shaping your answer?**

Working in teams

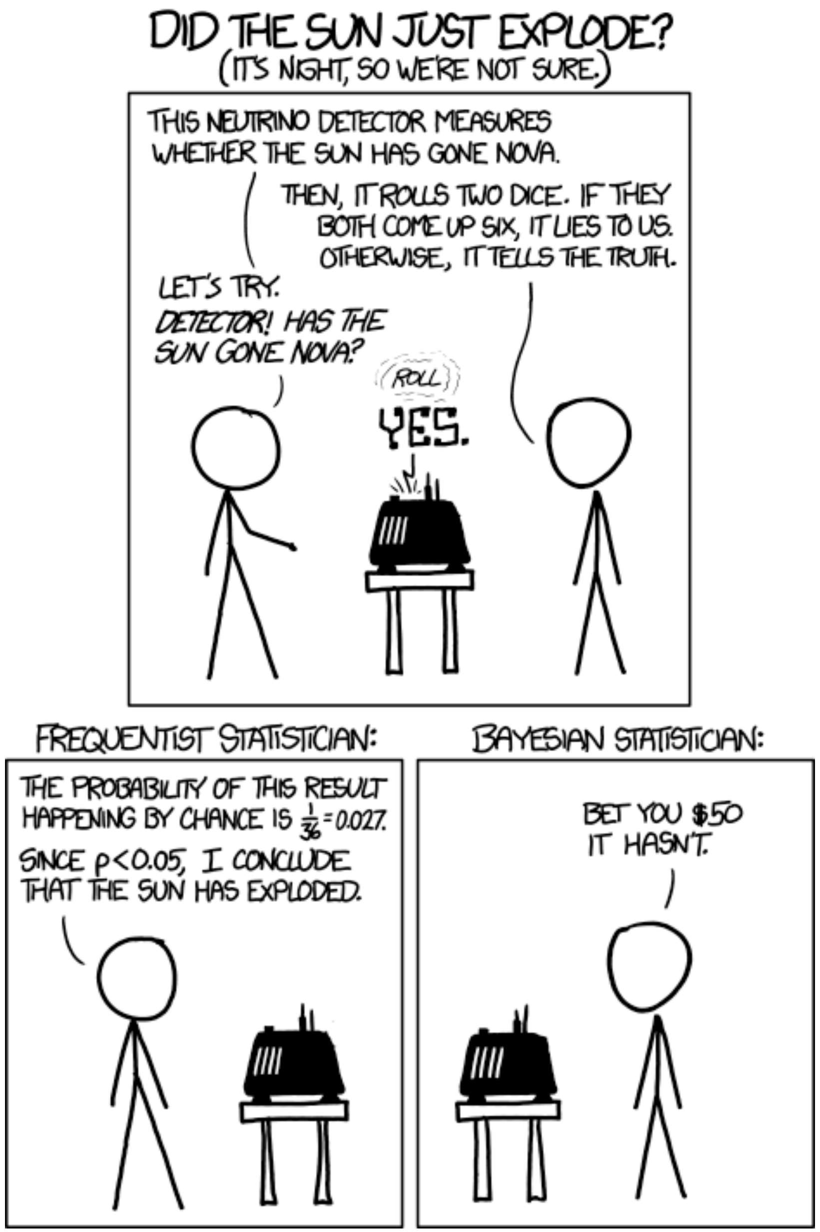
- Balance between individual responsibilities and collaborative effort, good practice for real life
- What I would like you to do:
  - Discuss project before starting coding, try to choose a common approach
  - Collaborate in refining and debug code
  - Agree on a Jupyter notebook that will be submitted on behalf of the team
- During the regular session you should limit the discussion to your partner, although if there is a problem that benefits from a discussion that involves the whole class, we will have mini-breaks

Assessment

- Written work, Jupyter notebooks
- Participation in in-class module-end discussion, quality of submitted work, final project (see syllabus for details)

Module 1

Introduction to frequentist and Bayesian statistics and applications to model selection



Probability theory

- Mathematical probability (Kolmogorov, 1933): any quantity that satisfies Kolmogorov axioms (defined for exclusive elementary events X_i in set Ω)
  a) $P(X_i) \geq 0$ for all i

- **Frequentist approach:** probability related to frequency of observed events
- **Bayesian approach:** probability as “degree of belief”

Frequentist probability

- The frequentist probability of an event X_i is defined as the number of times X_i occurs $N(X_i)$ in N events
- $P(X_i) = \lim_{N\to\infty} \frac{N(X_i)}{N}$
- Frequentist probability was the “gold standard” as it is **objective**.
- In principle, it can be determined to any desired accuracy and does not depend upon the observer.

Building a model of the data

- Model is the full structure of $P(\text{data}|\text{parameters})$
  - Holding parameters fixed gives the **pdf** for the data
  - Holding data fixed gives a **likelihood function** for parameters
- Model can be interpreted as a quantitative summary of the analysis: e.g. which fundamental lesson we learned in our experiment? Note that the quality of the results is tied to how convincing the story is and how tightly it connects with a model.
- Both Bayesian and Frequentist methods start with the model

Point estimator – frequentist

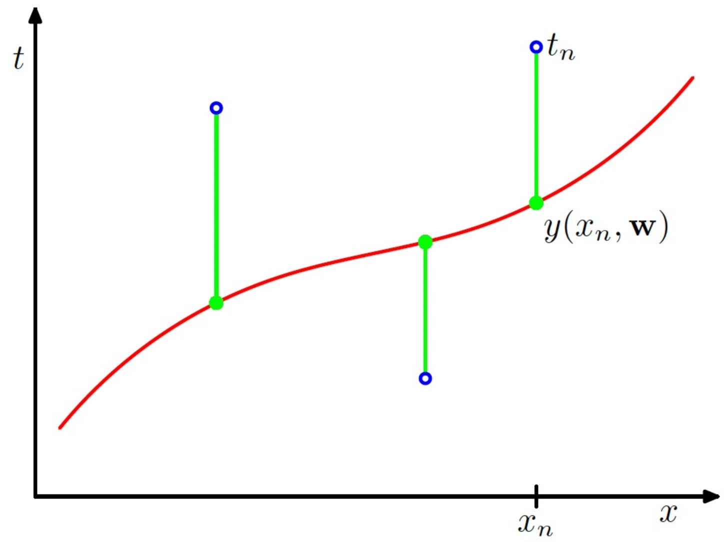
- An estimator ε_w of a set of unknown parameters μ produces an estimate $\widehat{w}$ based on a data set X
- Goal is to find the estimator which gives estimate closest to the true value
  - Estimate: $\widehat{w}$ (also commonly $\hat{\theta}$)
  - True value: w

    In general, w is a vector
- Example: the sample mean is an *estimator* for the population mean

Point estimator – frequentist

- In your programming activities you will use the error function Error function $E(w)$ given by the sum of the squares of the difference between the model $y(X, w)$ for any given w and the corresponding target values t_n so that we minimize

- $E(w) = \frac{1}{2} \sum_{n=1}^{N} \{y(X_n, w) - t_n\}^2$



Choice of the model

- In activity 1 you will experiment with different models by changing the order of the polynomial PDF
- The minimizations have unique solution because of the linear dependence of the derivatives of the error function with respect to $\{w\}$
- Check quality of the model by using a test set (predictive value):
- calculate $E(\hat{w})$ for test set and calculate $\sqrt{\frac{2E(\hat{w})}{N}}$, note: the use of training and test sets will be a dominant theme throughout the semester
- You will see the results of the overfitting

Mitigation of overfitting

- Regularization adds penalty term to error function
- $E(w) = \underbrace{\frac{1}{2}\sum_{n=1}^{N}\{y(X_n, w) - t_n\}^2}_{\text{Sum of squares error}} + \underbrace{\frac{\lambda}{2}||w||^2}_{\text{Regularization term}}$
- Where $||w||^2 = w_0^2 + w_1^2 + w_2^2 + \ ... + w_M^2$ and λ controls the relative importance of the regularization
- λ is a *hyperparameter* to be tuned
- Will explore more in the exercises

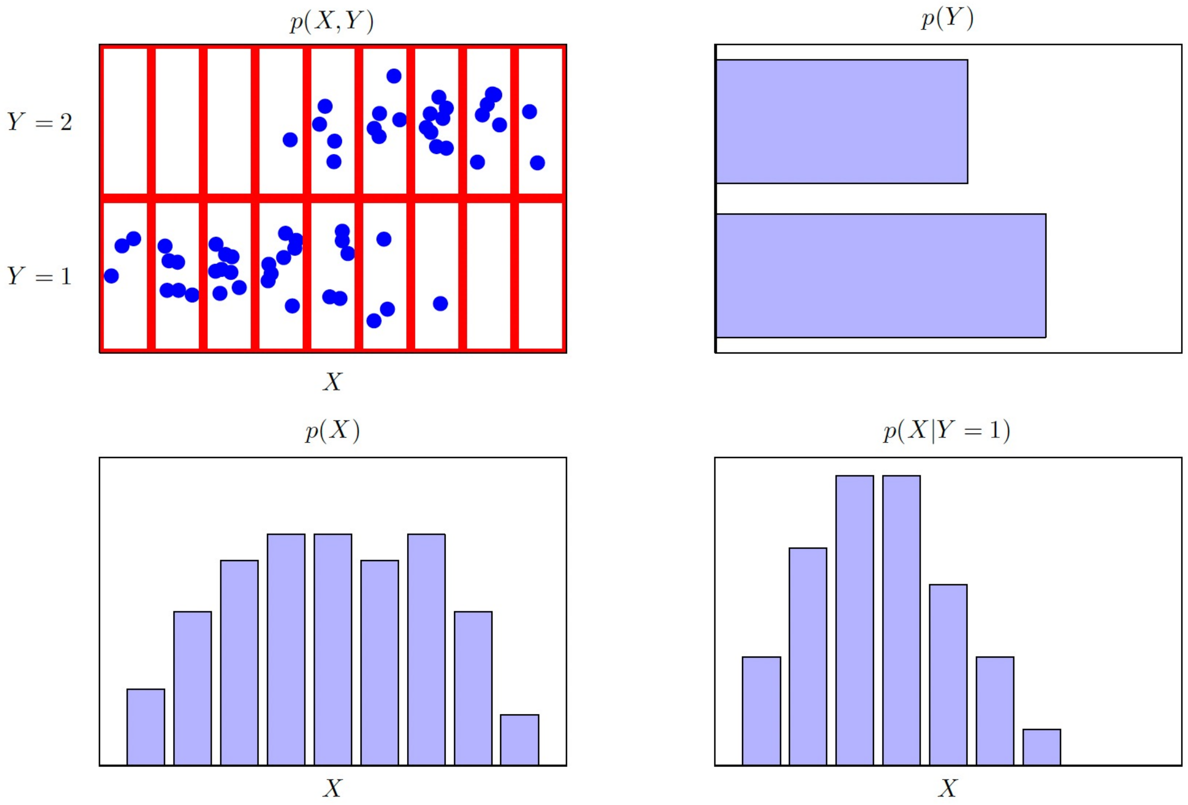
Fundamentals of Probability

- ❑ Random variable: X which can take on values of x_i
  - ❑ E.g. X is the face of a die, which can take on values x_i of {1-6}

- ❑ $p(X)$ – **Marginal probability** of random variable X

- ❑ $p(X, Y)$ – **Joint probability** of X and Y

- ❑ $p(X|Y)$ – **Conditional probability** of X *given* that Y occurred



Fundamentals of Probability

- **Sum Rule**: $p(X) = \sum_Y p(X, Y)$ - ‘marginalized over Y’

- **Product Rule**: $p(X, Y) = p(Y|X)p(X)$

- Sum rule and product rule combined show symmetry $p(X, Y) = p(Y, X)$

- Importantly $p(X|Y)$ is NOT equal to $p(Y|X)$

Bayes Theorem

- Combining sum and product rules we arrive at *Bayes' theorem*

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

- Fundamental relationship between conditional probabilities

- NOT Bayesian statistics, but has a Bayesian and frequentist interpretation

Monty Hall problem

- Suppose you are playing a shell game where 3 cups are shuffled and need to guess which cup which has the ball under it.
- You choose cup A
- The shuffler then removes cup C (they are not allowed to remove the cup with the ball)
- You are then given the option to switch to cup B

What should you choose?

Monty Hall problem

- Start of the game $P(A) = P(B) = P(C) = \frac{1}{3}$

Monty Hall problem

- Start of the game $P(A) = P(B) = P(C) = \frac{1}{3}$
- Prior probabilities$P(H_A) = P(H_B) = P(H_C) = \frac{1}{3}$

Monty Hall problem

- Start of the game $P(A) = P(B) = P(C) = \frac{1}{3}$
- Prior probabilities$P(H_A) = P(H_B) = P(H_C) = \frac{1}{3}$
- Consider 'likelihood Cup C is then taken out of the game = E_C'
  - If the ball is in A:
    - $P(E_C|H_A) = \frac{1}{2}$
  - If the ball is in B:
    - $P(E_C|H_B) = 1$
  - If the ball is in C:
    - $P(E_C|H_C) = 0$

Monty Hall problem

- Bayes theorem $P(H_i|E_C) = P(E_C|H_I)P(H_i)/P(E_C)$

Monty Hall problem

- Bayes theorem $P(H_i|E_C) = P(E_C|H_I)P(H_i)/P(E_C)$
- Compute marginal probability
  - $P(E_C) = \sum P(E_C|H_I)P(H_i)$
  - $P(E_C) = \frac{1}{2} * \frac{1}{3} + 1 * \frac{1}{3} + 0 * \frac{1}{3} = \frac{1}{2}$

Monty Hall problem

- Bayes theorem $P(H_i|E_C) = P(E_C|H_I)P(H_i)/P(E_C)$
- Compute marginal probability
  - $P(E_C) = \sum P(E_C|H_I)P(H_i)$
  - $P(E_C) = \frac{1}{2} * \frac{1}{3} + 1 * \frac{1}{3} + 0 * \frac{1}{3} = \frac{1}{2}$
- Posterior for keeping cup A vs switching to cup B

$$P(H_A|E_C) = \frac{P(E_C|H_A)P(H_A)}{P(E_C)} = \frac{\frac{1}{2}*\frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}$$

$$P(H_B|E_C) = \frac{P(E_C|H_B)P(H_B)}{P(E_C)} = \frac{1*\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$$

Day 2

- Form pairs
- Begin coding exercise for Module 1:
  - Focus on activity 1-3
- Intro to curve fitting if time

Day 3

Pairs I have so far:

- Mohammed Amal & Belal Menbari
- Billy Chu & Claire O’Connor
- Prakriti Singh & Hanieh Moradipasha
- Luis & Breck
- Lauren Sdun & Abeera Ajmal
- Javad Yousefian & Areg Zaratsyan
- Luke Matzner & Jasmine

Missing 5 people?

Todays goals:

- Finish lecture on module 1
- Work on coding assignment

Bayesian Probabilities

- For observed data $\mathcal{D}$ and model parameters w we can write Bayes' theorem as $p(w|\mathcal{D}) = \frac{p(\mathcal{D}|w)p(w)}{p(\mathcal{D})}$
- $p(\mathcal{D}|w)$ *'likelihood'* - given our data, how likely is our model parameters
- $p(w)$ *'prior'* - what do we already know (or assume) about our model before observation/analysis of data
- $p(w|\mathcal{D})$ *'posterior'* – updated probability after the 'prior' has been updated with evidence from data 'likelihood'

<u>Posterior $\propto$ likelihood $\times$ prior</u>

Curve Fitting

- So far we have seen:
  - Least squares with polynomials
  - Overfitting at high order
  - Sensitivity to noise

- What does it mean to believe a model, given noisy data?

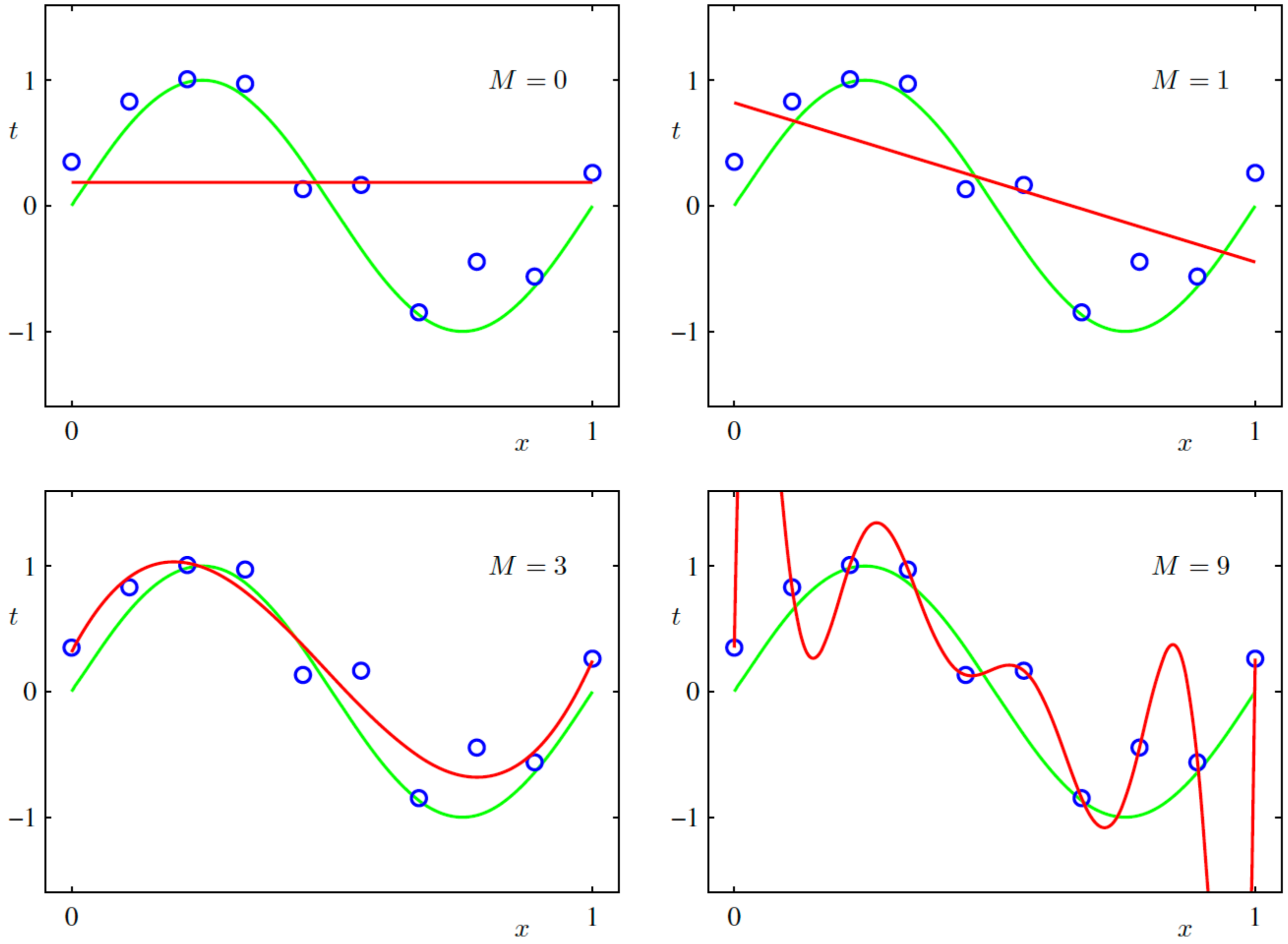


Figure 1.4 Plots of polynomials having various orders M, shown as red curves, fitted to the data set shown in Figure 1.2.

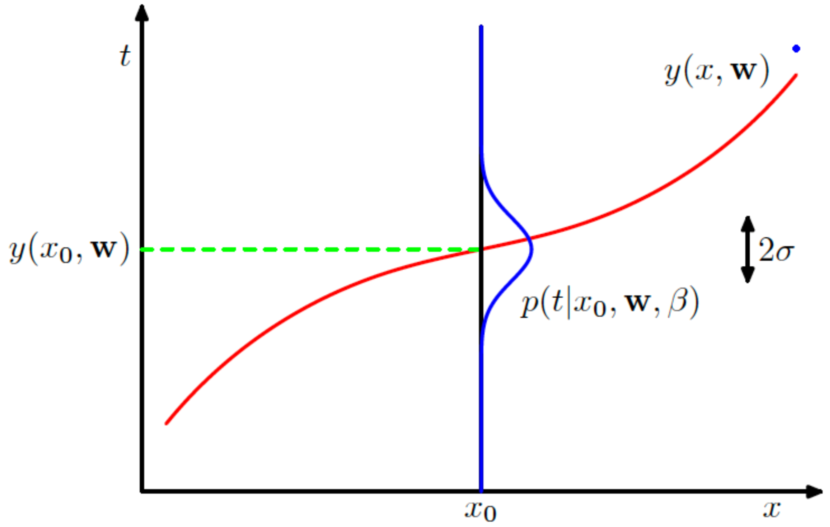
Fitting as a probability (MLE)

- Assume data are drawn from a normal distribution – get the likelihood function

$$p(\mathbf{t} \mid \mathbf{w}) = \prod_n \underbrace{\mathcal{N}\left(t_n \mid y(x_n, \mathbf{w}), \beta^{-1}\right)}$$

"Probability that for a value x_n our observed value t_n is described by our polynomial model $y(x_n, \boldsymbol{w})$ with gaussian random noise given by β"

Important: this assumption represents the noise in the data



Fitting as a probability (MLE)

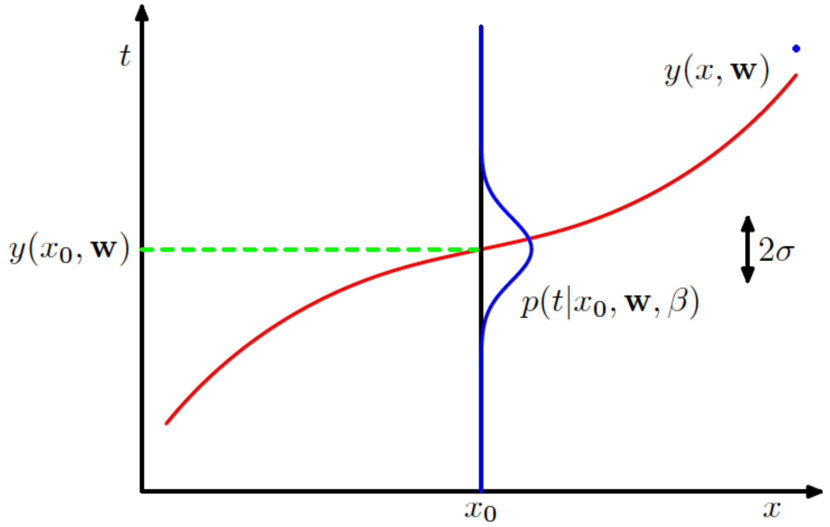
- Assume data are drawn from a normal distribution – get the likelihood function

$$p(\mathbf{t} \mid \mathbf{w}) = \prod_n \mathcal{N}(t_n \mid y(x_n, \mathbf{w}), \beta^{-1})$$

- Minimizing the negative log is equivalent to the SSE from before

$$\underbrace{-\log p(\mathbf{t} \mid \mathbf{w})}_{\text{“Maximum Likelihood”}} \propto \underbrace{\sum_n (t_n - y(x_n, \mathbf{w}))^2}_{\text{Sum of squares error “SSE”}}$$

Important: this assumption represents the noise in the data



Regularization as probability (MAP)

- We can introduce a Gaussian prior to the distribution of the parameters

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}, \alpha^{-1} I)$$

This represents the variation we expect in the weights of the model

Regularization as probability (MAP)

- We can introduce a Gaussian prior to the distribution of the parameters

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}, \alpha^{-1} I)$$

This represents the variation we expect in the weights of the model

- We can write posterior as

$$p(\mathbf{w}|\mathbf{t}) \propto p(\mathbf{t}|\mathbf{w})p(\mathbf{w})$$

- Maximizing the posterior (MAP) becomes

$$-\log p(\mathbf{t}|\mathbf{w}) - \log p(\mathbf{w}) \iff \text{SSE} + \alpha \|\mathbf{w}\|^2$$

Regularization as probability (MAP)

- Not (yet) Bayesian!

- It only gives one ‘best fit’ curve – but suggests some curves are less likely than others

Bayesian curve fitting asks a different question

- Frequentist/MAP: *“What is the best curve?”* – point estimate

- Bayesian: *“Given the data, what curves are plausible, and what should we predict?”*

We stop treating parameters as *unknown numbers* and treat them as *random variables*

Curves become probability distributions

- Statement of problem: given training data **x** and **t** for new test point x we wish to predict the value of t.

Curves become probability distributions

- Statement of problem: given training data **x** and **t** for new test point x we wish to predict the value of t.
- We can write a predictive distribution for this: $p(t|x, \mathbf{x}, \mathbf{t})$

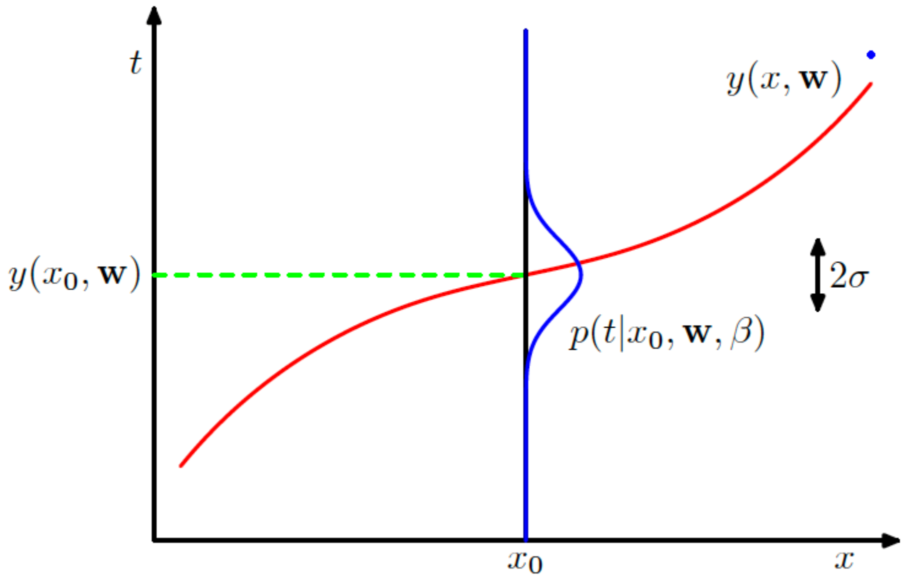
Curves become probability distributions

- Statement of problem: given training data **x** and **t** for new test point x we wish to predict the value of t.
- We can write a predictive distribution for this: $p(t|x, \mathbf{x}, \mathbf{t})$
- Use the sum and product rules to determine this distribution

$$p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w}) p(\mathbf{w}|\mathbf{x}, \mathbf{t}) d\mathbf{w}$$

Curves become probability distributions

$$p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w})p(\mathbf{w}|\mathbf{x}, \mathbf{t})d\mathbf{w}$$



Same distribution as before that we used to create the likelihood

"Probability that for a value x_n our observed value t_n is described by our polynomial model $y(x_n, \boldsymbol{w})$ with gaussian random noise given by β"

Curves become probability distributions

$$p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w})\underbrace{p(\mathbf{w}|\mathbf{x}, \mathbf{t})}_{\text{Posterior distribution} \,\propto\, \text{likelihood + prior}} d\mathbf{w}$$

Curves become probability distributions

Key difference from MLE/MAP: Model parameters have been integrated away!

$$p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w}) p(\mathbf{w}|\mathbf{x}, \mathbf{t}) d\mathbf{w}$$

Curves become probability distributions

- We are now marginalizing over *all possible parameters* for the model!

$$p(t|x, \mathbf{x}, \mathbf{t})$$