Bayesian parameter estimation

Nuno Vasconcelos
UCSD

Maximum likelihood

- parameter estimation in three steps:
  - 1) choose a parametric model for probabilities
    to make this clear we denote the vector of parameters by Θ

$$P_X(x;\Theta)$$

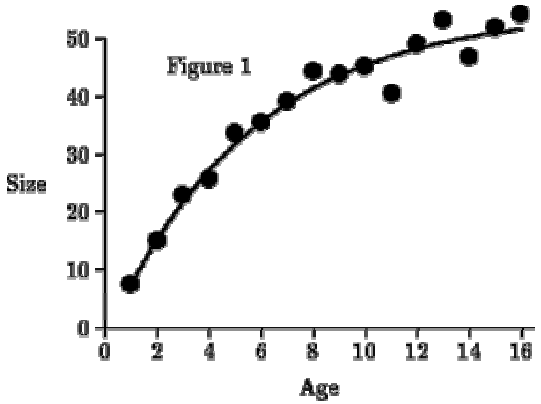
    note that this means that Θ is NOT a random variable
  - 2) assemble $\mathcal{D} = \{x_1, ..., x_n\}$ of examples drawn independently
  - 3) select the parameters that maximize the probability of the data

$$\Theta^* = \arg\max_{\Theta} P_X(D;\Theta)$$
$$= \arg\max_{\Theta} \log P_X(D;\Theta)$$

- $P_X(\mathcal{D};\Theta)$ is the likelihood of parameter Θ with respect to the data

Least squares

- there are interesting connections between ML estimation and least squares methods
- e.g. in a regression problem we have
  - two random variables X and Y
  - a dataset of examples
    $\mathcal{D} = \{(x_1,y_1), \ldots (x_n,y_n)\}$
  - a parametric model of the form



$$y = f(x;\Theta) + \varepsilon$$

  - where Θ is a parameter vector, and ε a random variable that accounts for noise
  - e.g. $\varepsilon \sim N(0,\sigma^2)$

Least squares

- assuming that the family of models is known, e.g.

$$f(x;\Theta) = \sum_{i=0}^{K} \theta_i x^i$$

  - this is really just a problem of parameter estimation
  - where the data is distributed as

$$P_{Z|X}(D \mid x;\Theta) = G\left(z,\, f(x;\Theta), \sigma^2\right)$$

  - note that X is always known, and the mean is a function of x and Θ
  - in the homework, you will show that

$$\Theta^* = \left[\Gamma^T \Gamma\right]^{-1} \Gamma^T y$$

Least squares

- where

$$\Gamma = \begin{bmatrix} 1 & \dots & x_1^K \\ & \vdots & \\ 1 & \dots & x_n^K \end{bmatrix}$$

- conclusion:
  - least squares estimation is really just ML estimation under the assumption of
    - Gaussian noise
    - independent sample
    - $\varepsilon \sim N(0, \sigma^2)$
- once again, probability makes the assumptions explicit

Least squares solution

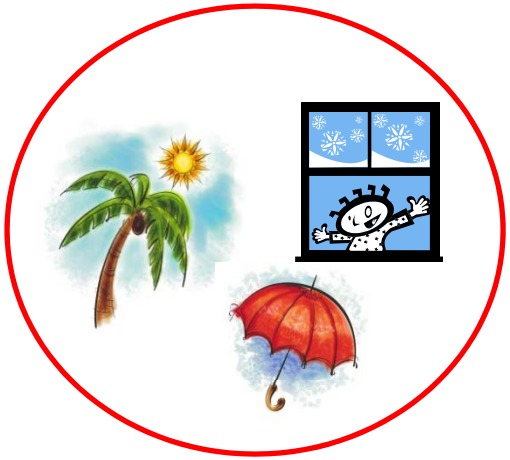
- due to the connection to parameter estimation
- we can also talk about the "quality" of the least squares solution
- in particular, we know that
  - it is unbiased
  - variance goes to zero as the number of points increases
  - it is the BLUE estimator for $f(x;\Theta)$
- under the statistical formulation we can also see how the optimal estimator changes with assumptions
- ML estimation can also lead to (homework)
  - weighted least squares
  - minimization of L_p norms
  - robust estimators

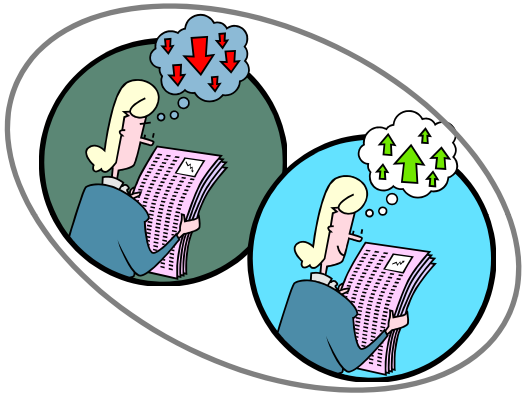
Bayesian parameter estimation

- Bayesian parameter estimation is an alternative framework for parameter estimation
  - it turns out that the division between Bayesian and ML methods is quite fundamental
- it stems from a different way of interpreting probabilities
  - frequentist vs Bayesian
- there is a long debate about which is best
  - this debate goes to the core of what probabilities mean
- to understand it, we have to distinguish two components
  - the definition of probability (this does not change)
  - the assessment of probability (this changes)
- let's start with a brief review of the part that does not change

Probability

- probability is a language to deal with processes that are non-deterministic



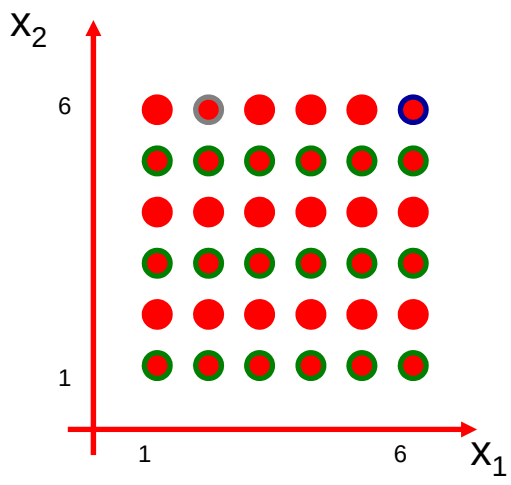


- examples:
  - if I flip a coin 100 times, how many can I expect to see heads?
  - what is the weather going to be like tomorrow?
  - are my stocks going to be up or down?
  - am I in front of a classroom or is this just a picture of it?

Sample space

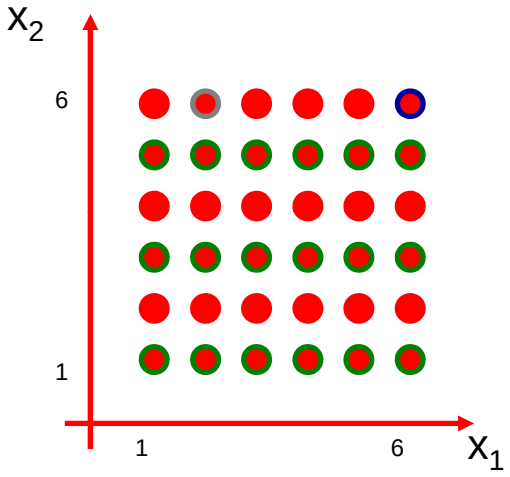
- the most important concept is that of a sample space
- our process defines a set of events
  - these are the outcomes or states of the process

- example:
  - we roll a pair of dice
  - call the value on the up face at the n^{th} toss x_n
  - note that possible events such as
    - odd number on second throw
    - two sixes
    - $x_1 = 2$ and $x_2 = 6$
  - can all be expressed as combinations of the sample space events



Sample space

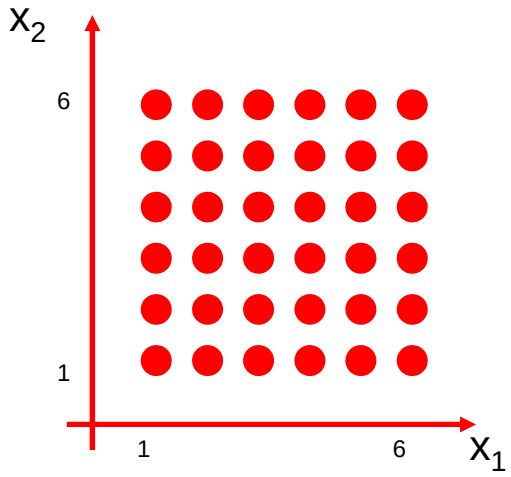
- is the list of possible events that satisfies the following properties:
  - finest grain: all possible distinguishable events are listed separately
  - mutually exclusive: if one event happens the other does not (if $x_1 = 5$ it cannot be anything else)
  - collectively exhaustive: any possible outcome can be expressed as unions of sample space events



- mutually exclusive property simplifies the calculation of the probability of complex events
- collectively exhaustive means that there is no possible outcome to which we cannot assign a probability

Probability measure

- probability of an event:
  - number expressing the chance that the event will be the outcome of the process
- probability measure: satisfies three axioms
  - $P(A) \geq 0$ for any event A
  - *P(universal event) = 1*
  - *if* $A \cap B = \varnothing$, *then* $P(A+B) = P(A) + P(B)$
- all of this
  - has to do with the definition of probability
  - is the same under Bayes and frequentist views
- what changes is how probabilities are assessed



Frequentist view

- under the frequentist view probabilities are relative frequencies
  - I throw my dice *n* times
  - in *m* of those the sum is 5
  - I say that

$$P(sum = 5) = \frac{m}{n}$$

- this is intimately connected with the ML method
  - it is the ML estimate for the probability of a Bernoulli process with states (“5”, “everything else”)
- makes sense when we have a lot of observations
  - no bias; decreasing variance; converges to true probability

Problems

- many instances where we do not have a large number of observations
- consider the problem of crossing a street
- this is a decision problem with two states
  - Y = 0: “I am going to get hurt”
  - Y = 1: “I will make it safely”

- optimal decision computable by Bayes decision rule
  - collect some measurements that are informative
  - e.g. (X = {size, distance, speed} of incoming cars)
  - collect examples under both states and estimate all probabilities
- somehow this does not sound like a great idea!

Problems

- under the frequentist view
  - you need to repeat an experiment a large number of times
  - to estimate any probabilities
- yet, people are very good at
  - estimating probabilities
  - for problems in which it is impossible to set up such experiments
- for example:
  - will I die if I join the army?
  - will Democrats or Republicans win the next election?
  - is there a God?
  - will I graduate in two years?
- to the point where they make life-changing decisions based on these probability estimates (enlisting in the army, etc.)

Subjective probability

- this motivates an alternative definition of probabilities
  - note that this has to do more with how probabilities are assessed than with the probability definition itself
  - we still have a sample space, a probability measure, etc
  - however the probabilities are not equated to relative counts
- this is usually referred to as subjective probability
- probabilities are degrees of belief on the outcomes of the experiment
  - they are individual (vary from person to person)
  - they are not ratios of experimental outcomes
- e.g.
  - for very religious person P(god exists) ~ 1
  - for casual churchgoer P(god exists) ~ 0.8 (e.g. accepts evolution, etc.)
  - for non-religious P(god exists) ~ 0

Problems

- in practice, why do we care about this?
- under the notion of subjective probability, the entire ML framework makes little sense
  - there is a magic number that is estimated from the world and determines our beliefs
  - to evaluate my estimates I have to run experiments over and over again and measure quantities like bias and variance
  - this is not how people behave, when we make estimates we attach a degree of confidence to them, without further experiments
  - there is only one model (the ML model) for the probability of the data, no multiple explanations
  - there is no way to specify that some models are, a priori, better than others

Bayesian parameter estimation

- the main difference with respect to ML is that in the Bayesian case Θ is a random variable
- basic concepts
  - training set $\mathcal{D} = \{x_1, ..., x_n\}$ of examples drawn independently
  - probability density for observations given parameter

$$P_{X|\Theta}(x \mid \theta)$$

  - prior distribution for parameter configurations

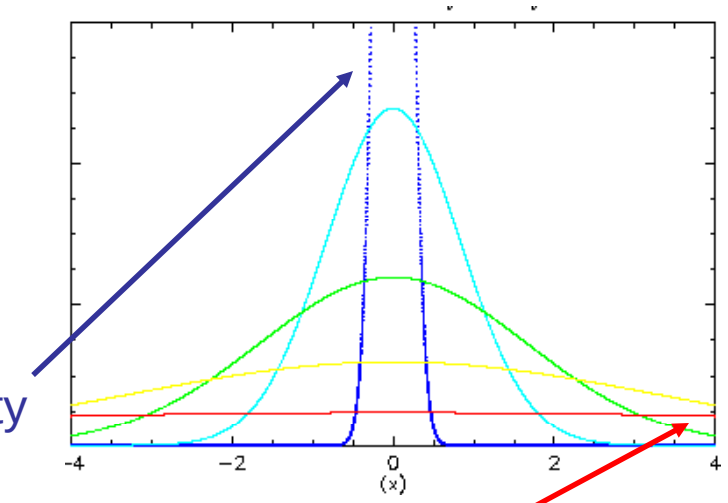
$$P_{\Theta}(\theta)$$

    that encodes prior beliefs about them
- goal: to compute the posterior distribution

$$P_{\Theta|X}(\theta \mid D)$$

Bayes vs ML

- there are a number of significant differences between Bayesian and ML estimates
- D_1:
  - ML produces a number, the best estimate
  - to measure its goodness we need to measure bias and variance
  - this can only be done with repeated experiments
  - Bayes produces a complete characterization of the parameter from the single dataset
  - in addition to the most probable estimate, we obtain a characterization of the uncertainty



Bayes vs ML

- D_2: optimal estimate
  - under ML there is one “best” estimate
  - under Bayes there is no “best” estimate
  - only a random variable that takes different values with different probabilities
  - technically speaking, it makes no sense to talk about the “best” estimate
- D_3: predictions
  - remember that we do not really care about the parameters themselves
  - they are needed only in the sense that they allow us to build models
  - that can be used to make predictions (e.g. the BDR)
  - unlike ML, Bayes uses ALL information in the training set to make predictions

Bayes vs ML

- let's consider the BDR under the "0-1" loss and an independent sample $\mathcal{D} = \{x_1, ..., x_n\}$
- ML-BDR:
  - pick i if

$$i^*(x) = \arg\max_i P_{X|Y}\left(x \mid i; \theta_i^*\right) P_Y(i)$$

$$\text{where } \theta_i^* = \arg\max_\theta P_{X|Y}\left(D \mid i, \theta\right)$$

- two steps:
  - i) find θ^*
  - ii) plug into the BDR
- all information not captured by θ^* is lost, not used at decision time

Bayes vs ML

- note that we know that information is lost
  - e.g. we can't even know how good of an estimate θ^* is
  - unless we run multiple experiments and measure bias/variance
- Bayesian BDR
  - under the Bayesian framework, everything is conditioned on the training data
  - denote $T = \{X_1, ..., X_n\}$ the set of random variables from which the training sample $\mathcal{D} = \{x_1, ..., x_n\}$ is drawn
- B-BDR:
  - pick i if

$$i^*(x) = \arg\max_i P_{X|Y,T}\left(x \mid i, D_i\right) P_Y(i)$$

- the decision is conditioned on the entire training set

Bayesian BDR

- to compute the conditional probabilities, we use the marginalization equation

$$P_{X|Y,T}\left(x \mid i, D_i\right) = \int P_{X|\Theta,Y,T}\left(x \mid \theta, i, D_i\right) P_{\Theta|Y,T}\left(\theta \mid i, D_i\right) d\theta$$

- note 1: when the parameter value is known, x no longer depends on T, e.g. $X|\Theta \sim N(\theta, \sigma^2)$
  - we can, simplify equation above into

$$P_{X|Y,T}\left(x \mid i, D_i\right) = \int P_{X|\Theta,Y}\left(x \mid \theta, i\right) P_{\Theta|Y,T}\left(\theta \mid i, D_i\right) d\theta$$

- note 2: once again can be done in two steps (per class)
  - i) find $P_{\Theta|T}(\theta|D_i)$
  - ii) compute $P_{X|Y,T}(x|i, D_i)$ and plug into the BDR
- no training information is lost

Bayesian BDR

- in summary
  - pick i if

$$i^*(x) = \arg\max_i P_{X|Y,T}\left(x \mid i, D_i\right) P_Y(i)$$

$$\text{where} \quad P_{X|Y,T}\left(x \mid i, D_i\right) = \int P_{X|Y,\Theta}\left(x \mid i, \theta\right) P_{\Theta|Y,T}\left(\theta \mid i, D_i\right) d\theta$$

- note:
  - as before the bottom equation is repeated for each class
  - hence, we can drop the dependence on the class
  - and consider the more general problem of estimating

$$P_{X|T}\left(x \mid D\right) = \int P_{X|\Theta}\left(x \mid \theta\right) P_{\Theta|T}\left(\theta \mid D\right) d\theta$$

The predictive distribution

- the distribution

$$P_{X|T}(x \mid D) = \int P_{X|\Theta}(x \mid \theta) P_{\Theta|T}(\theta \mid D) d\theta$$

  is known as the predictive distribution
- this follows from the fact that it allows us
  - to predict the value of x
  - given ALL the information available in the training set
- note that it can also be written as

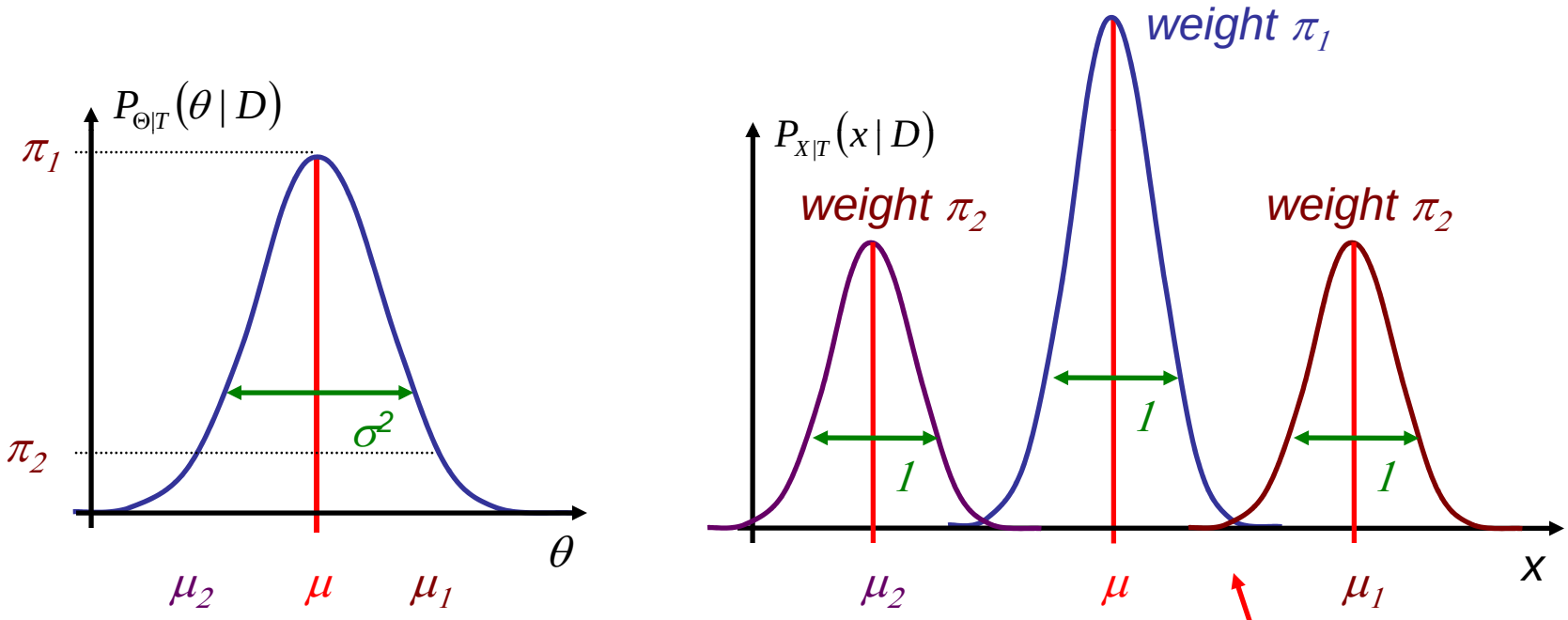
$$P_{X|T}(x \mid D) = E_{\Theta|T}\left[P_{X|\Theta}(x \mid \theta) \mid T = D\right]$$

  - since each parameter value defines a model
  - this is an expectation over all possible models
  - each model is weighted by its posterior probability, given training data

The predictive distribution

- suppose that

$$P_{X|\Theta}(x \mid \theta) \sim N(\theta, 1) \qquad \text{and} \qquad P_{\Theta|T}(\theta \mid D) \sim N(\mu, \sigma^2)$$

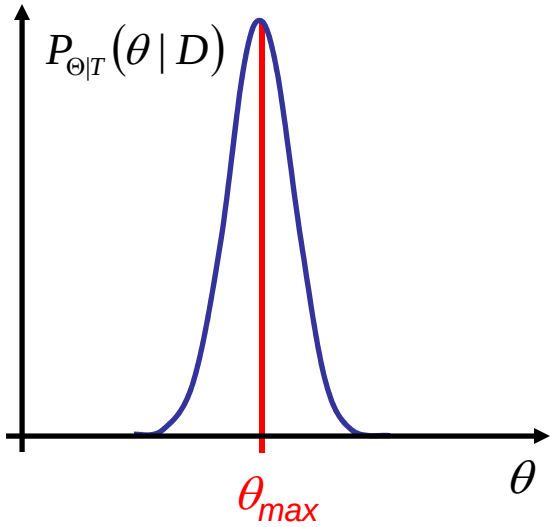


- the predictive distribution is an average of all these Gaussians

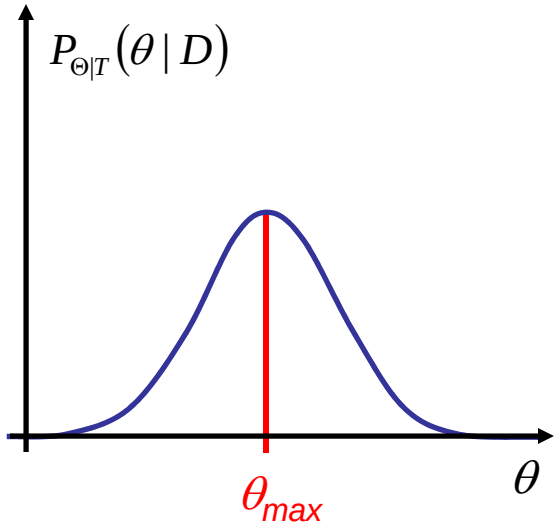
$$P_{X|T}(x \mid D) = \int P_{X|\Theta}(x \mid \theta) P_{\Theta|T}(\theta \mid D) d\theta$$

The predictive distribution

- Bayes vs ML
  - ML: pick one model
  - Bayes: average all models
- are Bayesian predictions very different than those of ML?
  - they can be, unless the prior is narrow



Bayes ~ ML



very different

The predictive distribution

- hence, ML can be seen as a special case of Bayes
  - when you are very confident about the model
  - picking one is good enough
- in coming lectures we will see that
  - if the sample is quite large, the prior tends to be narrow
  - intuitive: given a lot of training data, there is little uncertainty about what the model is
  - Bayes can make a difference when there is little data
  - we have already seen that this is the important case since the variance of ML tends to go down as the sample increases
- overall
  - Bayes regularizes the ML estimate when this is uncertain
  - converges to ML when there is a lot of certainty

MAP approximation

- this sounds good, why use ML at all?
- the main problem with Bayes is that the integral

$$P_{X|T}(x \mid D) = \int P_{X|\Theta}(x \mid \theta) P_{\Theta|T}(\theta \mid D) d\theta$$

  can be quite nasty
- in practice one is frequently forced to use approximations
- one possibility is to do something similar to ML, i.e. pick only one model
- this can be made to account for the prior by
  - picking the model that has the largest posterior probability given the training data

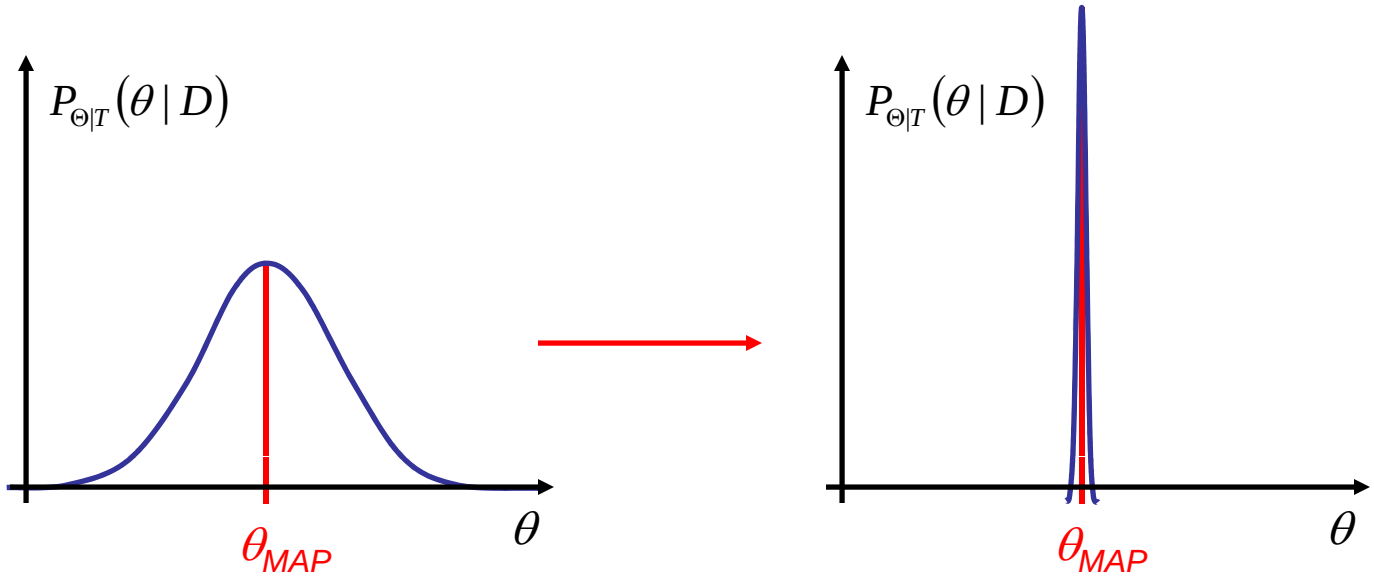
$$\theta_{MAP} = \arg\max_{\theta} P_{\Theta|T}(\theta \mid D)$$

MAP approximation

- this can usually be computed since

$$\theta_{MAP} = \arg\max_{\theta} P_{\Theta|T}(\theta \mid D)$$
$$= \arg\max_{\theta} P_{T|\Theta}(D \mid \theta) P_{\Theta}(\theta)$$

and corresponds to approximating the prior by a delta function centered at its maximum



MAP approximation

- in this case

$$P_{X|T}(x \mid D) = \int P_{X|\Theta}(x \mid \theta)\delta(\theta - \theta_{MAP})d\theta$$
$$= P_{X|\Theta}(x \mid \theta_{MAP})$$

- the BDR becomes
  - pick i if

$$i^*(x) = \arg\max_i P_{X|Y}\left(x \mid i; \theta_i^{MAP}\right)P_Y(i)$$

$$\text{where } \theta_i^{MAP} = \arg\max_\theta P_{T|Y,\Theta}(D \mid i, \theta)P_{\Theta|Y}(\theta \mid i)$$

  - when compared to the ML this has the advantage of still accounting for the prior (although only approximately)

MAP vs ML

- ML-BDR
  - pick i if

$$i^*(x) = \arg\max_i P_{X|Y}\left(x \mid i; \theta_i^*\right) P_Y(i)$$

$$\text{where } \theta_i^* = \arg\max_\theta P_{X|Y}\left(D \mid i, \theta\right)$$

- Bayes MAP-BDR
  - pick i if

$$i^*(x) = \arg\max_i P_{X|Y}\left(x \mid i; \theta_i^{MAP}\right) P_Y(i)$$

$$\text{where } \theta_i^{MAP} = \arg\max_\theta P_{T|Y,\Theta}\left(D \mid i, \theta\right) P_{\Theta|Y}\left(\theta \mid i\right)$$

  - the difference is non-negligible only when the dataset is small
- there are better alternative approximations

The Laplace approximation

- this is a method for approximating any distribution $P_X(x)$
  - consists of approximating $P_X(x)$ by a Gaussian centered at its peak
- let's assume that

$$P_X(x) = \frac{1}{Z} g(x)$$

  - where g(x) is an unormalized distribution (g(x) > 0, for all x)
  - and Z the normalization constant

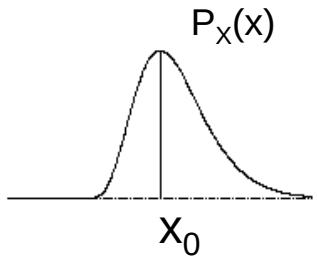
$$Z = \int g(x) dx$$

- we make a Taylor series approximation of g(x) at its maximum x_0

Laplace approximation

- the Taylor expansion is

$$\log g(x) = \log g(x_o) - c/2\,(x - x_0)^2 + \ldots$$



  - (the first-order term is zero because x_0 is a maximum)
  - with

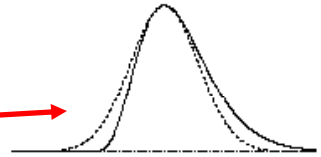
$$c = -\frac{\partial^2}{\partial x^2} \log g(x)\Big|_{x=x_0}$$

  - and we approximate g(x) by an unormalized Gaussian

$$g'(x) = g(x_o) \exp\left\{-c/2\,(x - x_0)^2\right\}$$

  - and then compute the normalization constant

$$Z = g(x_o)\sqrt{\frac{2\pi}{c}}$$



Laplace approximation

- this can obviously be extended to the multivariate case
- the approximation is

$$\log g(x) = \log g(x_o) - \tfrac{1}{2}(x - x_0)^T A (x - x_0)$$

  - with A the Hessian of g(x) at x_0

$$A_{ij} = -\frac{\partial^2}{\partial x_i \partial x_j} \log g(x) \Big|_{x=x_0}$$

  - and the normalization constant

$$Z = g(x_o) \sqrt{\frac{(2\pi)^d}{|A|}}$$

- in physics this is also called a saddle-point approximation

Laplace approximation

- note that the approximation can be made for the predictive distribution

$$P_{X|T}(x \mid D) = G(x, x^*, \mathrm{A}_{X|T})$$

- or for the parameter posterior

$$P_{\Theta|T}(\theta \mid D) = G(\theta, \theta_{MAP}, A_{\Theta|T})$$

in which case

$$P_{X|T}(x \mid D) = \int P_{X|\Theta}(x \mid \theta) G(\theta, \theta_{MAP}, A_{\Theta|T}) d\theta$$

- this is clearly superior to the MAP approximation

$$P_{X|T}(x \mid D) = \int P_{X|\Theta}(x \mid \theta) \delta(\theta - \theta_{MAP}) d\theta$$

Other methods

- there are two other main alternatives, when this is not enough
  - variational approximations
  - sampling methods (Markov Chain Monte Carlo)
- variational approximations consist of
  - bounding the intractable function
  - searching for the best bound
- sampling methods consist
  - designing a Markov chain that has the desired distribution as its equilibrium distribution
  - sample from this chain
- sampling methods
  - converge to the true distribution
  - but convergence is slow and hard to detect

Any questions?