Bayesian parameter estimation

Nuno Vasconcelos
UCSD

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 $\mathcal{D} = \{x_1, \ldots, x_n\}$ of examples drawn independently
  - probability density for observations given parameter

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

  - prior distribution for parameter configurations

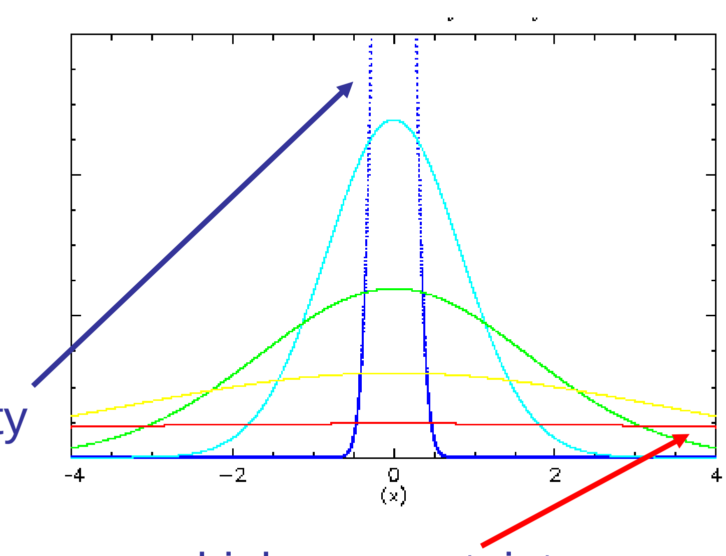
$$P_{\Theta}(\theta)$$

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

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

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



lower uncertainty

higher 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

Bayesian BDR

- this problem is avoided by Bayesian estimates
  - pick i if

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

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

- 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

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

  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

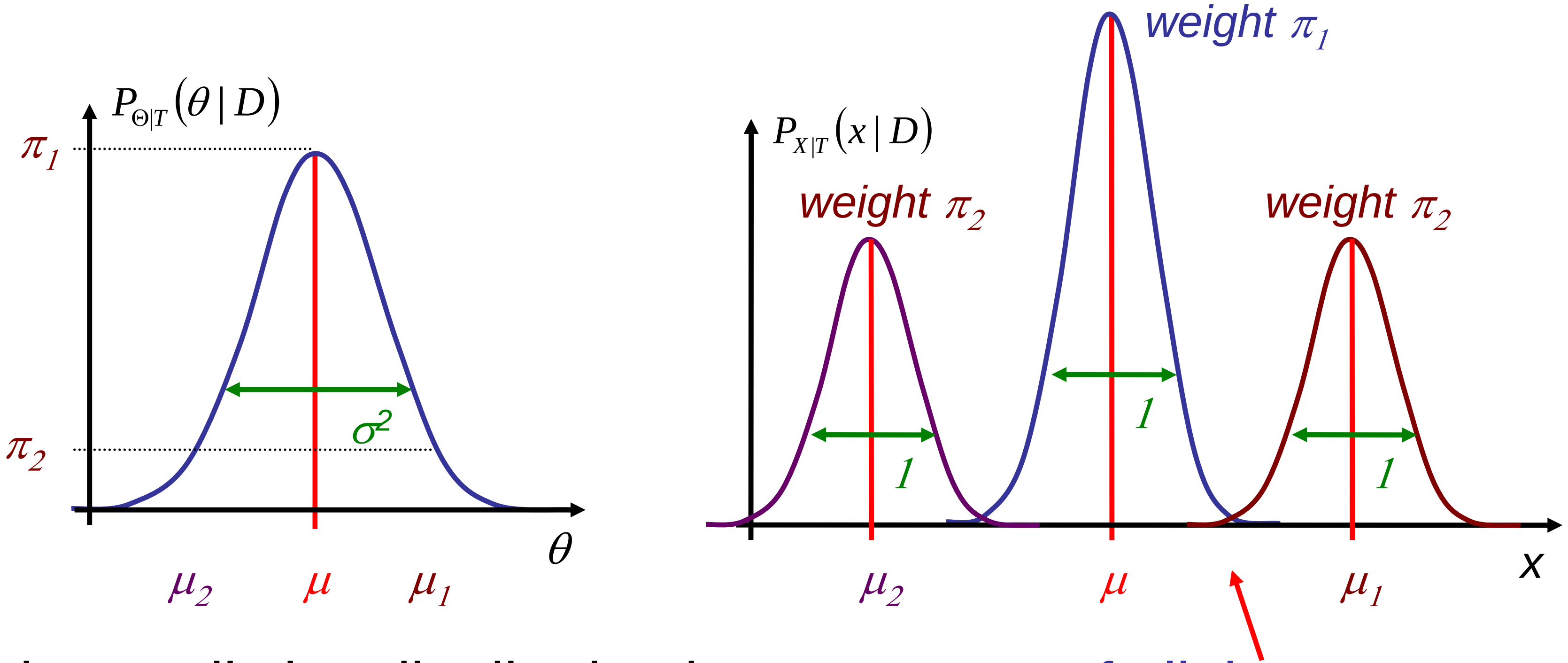
$$P_{X|T}(x|D) = E_{\Theta|T}\left[P_{X|\Theta}(x|\theta) \,|\, T = D\right]$$

  - 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

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

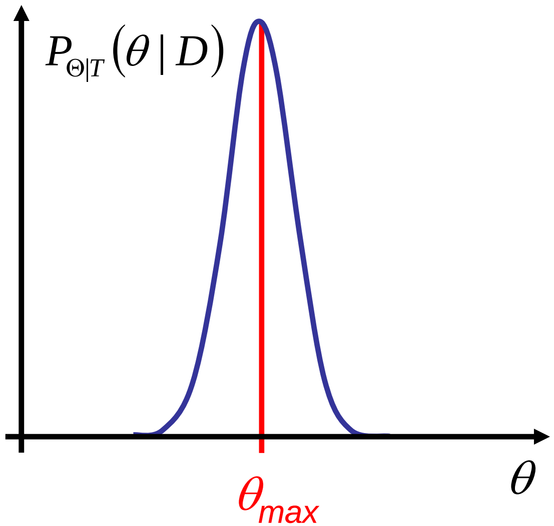


- the predictive distribution is an average of all these Gaussians

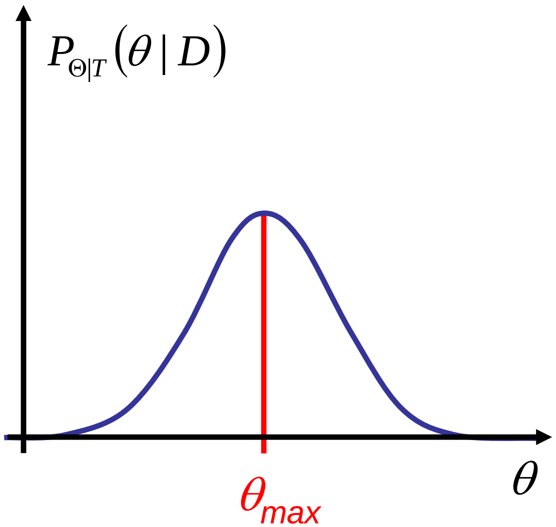
$$P_{X|T}(x|D) = \int P_{X|\Theta}(x|\theta) P_{\Theta|T}(\theta|D) d\theta$$

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

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

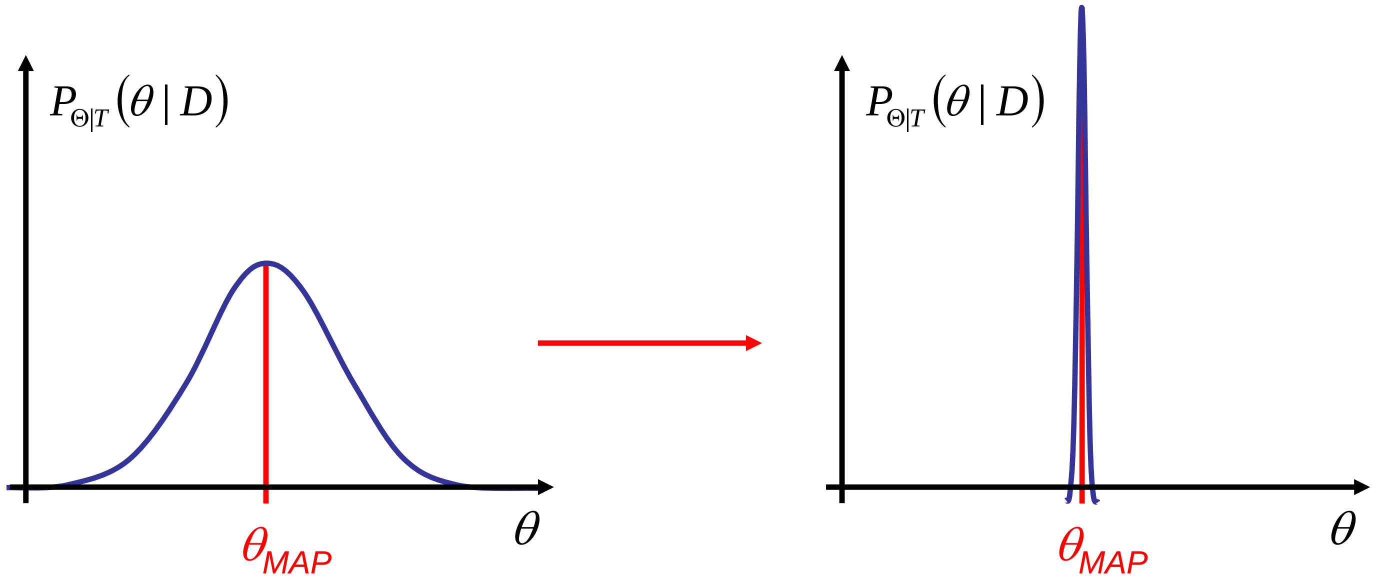
$$\theta_{MAP} = \arg\max_{\theta} P_{\Theta|T}(\theta \mid D)$$

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 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

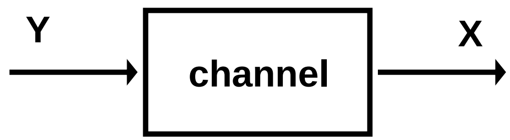
$$i^*(x) = \arg\max_i P_{X|Y}\left(x \mid i; \theta_i^{MAP}\right) P_Y(i)$$

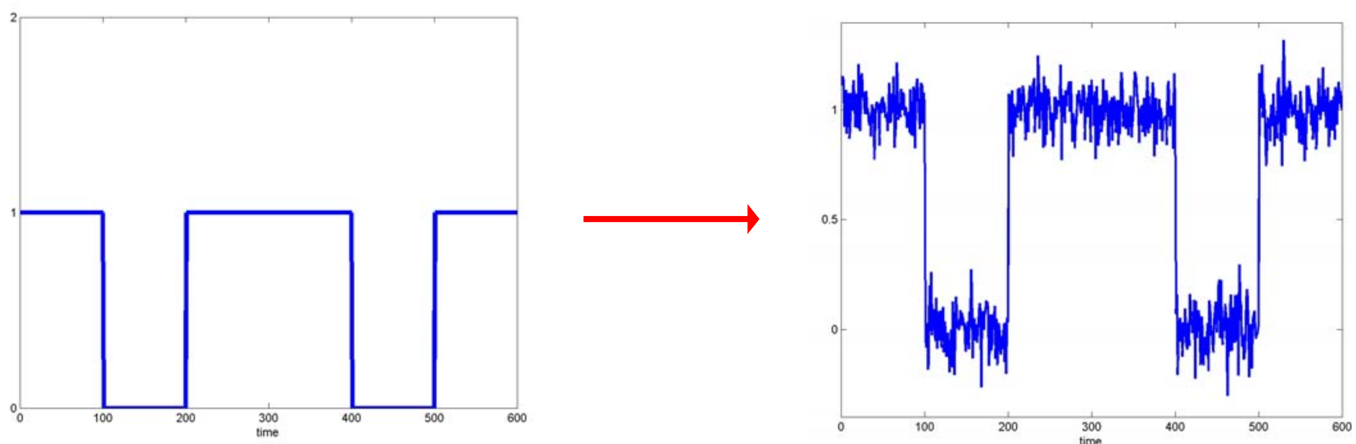
$$\text{where } \theta_i^{MAP} = \arg\max_\theta P_{T|Y,\Theta}\left(D \mid i, \theta\right) P_{\Theta|Y}\left(\theta \mid i\right)$$

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

Example

- let's consider an example of why Bayes is usefull
- example: communications
  - a bit is transmitted by a source, corrupted by noise, and received by a decoder



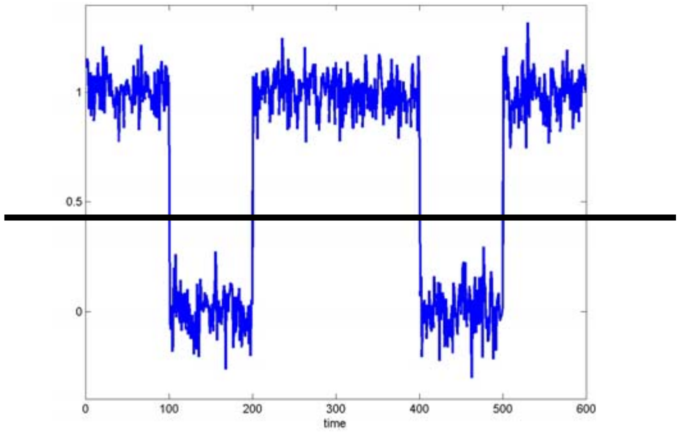


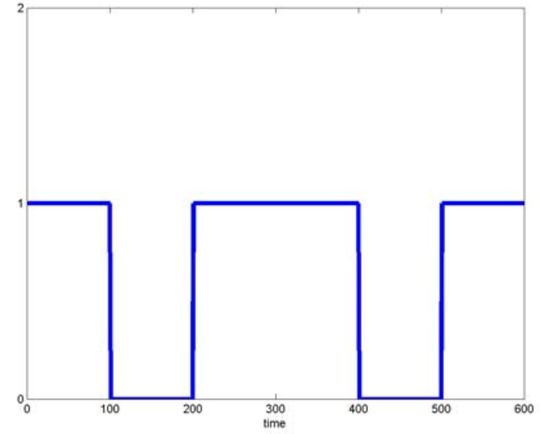
  - Q: what should the optimal decoder do to recover Y?

Example

- the optimal solution is to threshold X
  - pick T
  - decision rule

$$Y = \begin{cases} 0, & \text{if } x < \mathrm{T} \\ 1, & \text{if } x > \mathrm{T} \end{cases}$$



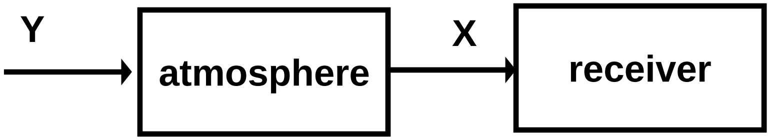


  - what is the threshold?
  - the midpoint between signal values

$$x < \frac{\mu_1 + \mu_0}{2}$$

Example

- today we consider a slight variation



- still:
  - two states:
    - Y=0 transmit signal s = $-\mu_0$
    - Y=1 transmit signal s = μ_0
  - same noise model

$$X = Y + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)$$

Example

- the BDR is still
  - pick “0” if

$$x < \frac{\mu_0 + (-\mu_0)}{2} = 0$$

  - this is optimal and everything works wonderfully
  - one day we get a phone call: the receiver is generating a lot of errors!
  - something must have changed in the rover
  - there is no way to go to Mars and check
- goal: to do as best as possible with the info that we have at X and our knowledge of the system

Example

- what we know:
  - the received signal is Gaussian, with same variance σ^2, but the means have changed
  - there is a calibration mode:
    - rover can send a test sequence
    - but it is expensive, can only send a few bits
  - if everything is normal, received means should be μ_0 and $-\mu_0$
- action:
  - ask the system to transmit a few 1s and measure X
  - compute the ML estimate of the mean of X

$$\mu = \frac{1}{n}\sum_i X_i$$

- result: the estimate is different than μ_0

Example

- we need to combine two forms of information
  - our prior is that $X \sim N(\mu_0, \sigma^2)$
  - our "data driven" estimate is that $X \sim N(\hat{\mu}, \sigma^2)$
- Q: what do we do?
  - $\mu_n = f(\hat{\mu}, \mu_0, n)$
  - for large n, $\mu_n \approx f(\hat{\mu})$
  - for small n, $\mu_n \approx f(\mu_0)$
  - intuitive combination

$$\mu_n = \alpha_n \hat{\mu} + (1 - \alpha_n)\mu_0$$

$$\alpha_n \in [0,1], \quad \alpha_n \underset{n \to \infty}{\longrightarrow} 1, \quad \alpha_n \underset{n \to 0}{\longrightarrow} 0$$

Bayesian solution

- Gaussian likelihood (observations)

$$P_{T|\mu}(D|\mu) = G(D,\mu,\sigma^2) \qquad \sigma^2 \text{is known}$$

- Gaussian prior (what we know)

$$P_\mu(\mu) = G(\mu,\mu_0,\sigma_0^2)$$

  – μ_0,σ_0^2 are known hyper-parameters

- we need to compute
  – posterior distribution for μ

$$P_{\mu|T}(\mu|D) = \frac{P_{T|\mu}(D|\mu)P_\mu(\mu)}{P_T(D)}$$

Bayesian solution

- posterior distribution

$$P_{\mu|T}(\mu \mid D) = \frac{P_{T|\mu}(D \mid \mu) P_{\mu}(\mu)}{P_T(D)}$$

- note that
  - this is a probability density
  - we can ignore constraints (terms that do not depend on μ)
  - and normalize when we are done
- we only need to work with

$$\begin{aligned} P_{\mu|T}(\mu \mid D) &\propto P_{T|\mu}(D \mid \mu) P_{\mu}(\mu) \\ &\propto \prod_i P_{X|\mu}(x_i \mid \mu) P_{\mu}(\mu) \end{aligned}$$

Bayesian solution

- plugging in the Gaussians

$$P_{\mu|T}(\mu|D) \propto \prod_i P_{X|\mu}(x_i|\mu) P_\mu(\mu)$$

$$\propto \prod_i G(x_i, \mu, \sigma^2) G(\mu, \mu_0, \sigma_0^2)$$

$$\propto \exp\left\{ - \sum_{\mathrm{i}} \frac{(x_i - \mu)^2}{2\sigma^2} - \frac{(\mu - \mu_0)^2}{2\sigma_0^2} \right\}$$

$$\propto \exp\left\{ - \sum_{\mathrm{i}} \frac{\mu^2 - 2x_i\mu + x_i^{\ 2}}{2\sigma^2} - \frac{\mu^2 - 2\mu\mu_0 + \mu_0^{\ 2}}{2\sigma_0^2} \right\}$$

$$\propto \exp\left\{ - \left(\frac{\mathrm{n}}{2\sigma^2} + \frac{1}{2\sigma_0^2} \right)\mu^2 + 2\left(\frac{\sum_i x_i}{2\sigma^2} + \frac{\mu_0}{2\sigma_0^2} \right)\mu - \left(\frac{\sum_i x_i^{\ 2}}{2\sigma^2} + \frac{\mu_0}{2\sigma_0^2} \right) \right\}$$

Bayesian solution

$$P_{\mu|T}(\mu \mid D) \propto \exp\left\{ -\left(\frac{\mathrm{n}}{2\sigma^2} + \frac{1}{2\sigma_0^2} \right)\mu^2 + 2\left(\frac{\sum_i x_i}{2\sigma^2} + \frac{\mu_0}{2\sigma_0^2} \right)\mu \right\}$$

- this is a Gaussian, we just need to put it in the standard quadratic form to know its mean and variance
- use the completing the squares trick

$$ax^2 + 2bx + c = a\left(x^2 + 2\frac{b}{a}x + \frac{c}{a} \right)$$

$$= a\left(x^2 + 2\frac{b}{a}x + \left(\frac{b}{a}\right)^2 - \left(\frac{b}{a}\right)^2 + \frac{c}{a} \right) = a\left(x + \frac{b}{a} \right)^2 + c - \frac{b^2}{a}$$

Bayesian solution

$$P_{\mu|T}(\mu \mid D) \propto \exp\left\{ -\left(\frac{\mathrm{n}}{2\sigma^2} + \frac{1}{2\sigma_0^2} \right)\mu^2 + 2\left(\frac{\sum_i x_i}{2\sigma^2} + \frac{\mu_0}{2\sigma_0^2} \right)\mu \right\}$$

- in this case

$$ax^2 + 2bx + c = a\left(x + \frac{b}{a} \right)^2 + c - \frac{b^2}{a} \propto a\left(x + \frac{b}{a} \right)^2$$

- we have

$$P_{\mu|T}(\mu \mid D) \propto \exp\left\{ -\left(\frac{\mathrm{n}}{2\sigma^2} + \frac{1}{2\sigma_0^2} \right)\left[\mu - \left(\frac{\sum_i x_i}{2\sigma^2} + \frac{\mu_0}{2\sigma_0^2} \right) \Big/ \left(\frac{\mathrm{n}}{2\sigma^2} + \frac{1}{2\sigma_0^2} \right) \right]^2 \right\}$$

Bayesian solution

- and using

$$1/\left(\frac{\mathrm{n}}{2\sigma^2}+\frac{1}{2\sigma_0^2}\right)=\frac{2\sigma^2\sigma_0^2}{\left(\sigma^2+n\sigma_0^2\right)}$$

- we have

$$P_{\mu|T}(\mu\,|\,D)\propto \exp\left\{-\left(\frac{\mathrm{n}}{2\sigma^2}+\frac{1}{2\sigma_0^2}\right)\left[\mu-\left(\frac{2\sigma^2\sigma_0^2}{\sigma^2+n\sigma_0^2}\right)\left(\frac{\sigma_0^2\sum_i x_i+\mu_0\sigma^2}{2\sigma^2\sigma_0^2}\right)\right]^2\right\}$$

$$\propto \exp\left\{-\left(\frac{2\sigma^2\sigma_0^2}{\sigma^2+n\sigma_0^2}\right)^{-1}\left[\mu-\left(\frac{\sigma_0^2\sum_i x_i+\mu_0\sigma^2}{\sigma^2+n\sigma_0^2}\right)\right]^2\right\}$$

- and

$$P_{\mu|T}(\mu\,|\,D)=G\left(\mu,\mu_n,\sigma_n^2\right),\qquad \mu_n=\frac{\sigma_0^2\sum_i x_i+\mu_0\sigma^2}{\sigma^2+n\sigma_0^2},\ \sigma_n^2=\left(\frac{\sigma^2\sigma_0^2}{\sigma^2+n\sigma_0^2}\right)$$

Bayesian solution

- this can be rewritten as

$$P_{\mu|T}(\mu \mid D) = G\left(\mu, \mu_n, \sigma_n^2\right)$$

$$\mu_n = \frac{\sigma_0^2 \sum_i x_i + \mu_0 \sigma^2}{\sigma^2 + n\sigma_0^2} \Rightarrow \mu_n = \underbrace{\frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2}}_{\alpha_\mathrm{n}} \mu_{ML} + \underbrace{\frac{\sigma^2}{\sigma^2 + n\sigma_0^2}}_{1-\alpha_n} \mu_0$$

$$\sigma_n^2 = \left(\frac{\sigma^2 \sigma_0^2}{\sigma^2 + n\sigma_0^2}\right) \Rightarrow \frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}$$

- we can compare with our “intuitive” solution

Bayesian solution

- we had

$$\mu_n = \alpha_n \hat{\mu} + (1 - \alpha_n)\mu_0$$

$$\alpha_n \in [0,1], \quad \alpha_n \underset{n\to\infty}{\to} 1, \quad \alpha_n \underset{n\to 0}{\to} 0$$

- the Bayesian solution is

$$\mu_n = \underbrace{\frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2}}_{\alpha_\mathrm{n}} \mu_{ML} + \underbrace{\frac{\sigma^2}{\sigma^2 + n\sigma_0^2}}_{1-\alpha_n} \mu_0$$

- note that $\alpha_n \in [0,1], \quad \alpha_n \underset{n\to\infty}{\to} 1, \quad \alpha_n \underset{n\to 0}{\to} 0$

- it is exactly the same as our heuristic

Bayesian solution

- for free, Bayes also gives us
  - the weighting constants

$$\alpha_n = \frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2}$$

  - a measure of the uncertainty of our estimate

$$\frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}$$

  - note that $1/\sigma^2$ is a measure of precision
  - this should be read as

$$P_{Bayes} = P_{ML} + P_{prior}$$

  - Bayesian precision is greater than both that of ML and prior

Observations

- 1) note that precision increases with n, variance goes to zero

$$\frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}$$

  we are guaranteed that in the limit of infinite data we have convergence to a single estimate

- 2) for large n the likelihood term dominates the prior term

$$\mu_n = \alpha_n \hat{\mu} + (1-\alpha_n)\mu_0$$

$$\alpha_n \in [0,1], \quad \alpha_n \underset{n\to\infty}{\longrightarrow} 1, \quad \alpha_n \underset{n\to 0}{\longrightarrow} 0$$

  the solution is equivalent to that of ML

- for small n, the prior dominates
- this always happens for Bayesian solutions

$$P_{\mu|T}(\mu \mid D) \propto \prod_i P_{X|\mu}(x_i \mid \mu) P_\mu(\mu)$$

Observations

– 3) for a given n

$$\alpha_n = \frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2}$$

$$\mu_n = \alpha_n \hat{\mu} + (1-\alpha_n)\mu_0$$
$$\alpha_n \in [0,1], \quad \alpha_n \underset{n\to\infty}{\longrightarrow} 1, \quad \alpha_n \underset{n\to 0}{\longrightarrow} 0$$

if $\sigma_0^2 >> \sigma^2$, i.e. we really don't know what μ is a priori
then $\mu_n = \mu_{ML}$

– on the other hand, if $\sigma_0^2 << \sigma^2$, i.e. we are very certain a priori, then $\mu_n = \mu_0$

- in summary,
  - Bayesian estimate combines the prior beliefs with the evidence provided by the data
  - in a very intuitive manner

Any questions?