Parametric Density Estimation:

Bayesian Estimation

C7

Bayesian Parameter Estimation

- Suppose we have some idea of the range where parameters θ should be
  - Shouldn't we formalize such prior knowledge in hopes that it will lead to better parameter estimation?

- Let θ be a random variable with prior distribution $\boldsymbol{P}(\theta)$
  - This is the key difference between ML and Bayesian parameter estimation
  - This key assumption allows us to fully exploit the information provided by the data

Bayesian Parameter Estimation

- θ is a random variable with prior $\boldsymbol{p}(\theta)$
  - Unlike MLE case, p(x|θ) is a conditional density
- The training data D allow us to convert p(θ) to a posterior probability density p(θ|D) .
  - After we observe the data D, using Bayes rule we can compute the posterior p(θ|D)
- But θ is not our final goal, our final goal is the unknown $\boldsymbol{p}(\boldsymbol{x})$
- Therefore a better thing to do is to maximize $\boldsymbol{p}(\boldsymbol{x}|\boldsymbol{D})$, this is as close as we can come to the unknown $\boldsymbol{p}(\boldsymbol{x})$!

Bayesian Estimation: Formula for $p(x|D)$

- From the definition of joint distribution:

$$p(x \mid D) = \int p(x, \theta \mid D)\, d\theta$$

- Using the definition of conditional probability:

$$p(x \mid D) = \int p(x \mid \theta, D)\, p(\theta \mid D)\, d\theta$$

- But $p(x|\theta,D)=p(x|\theta)$ since $p(x|\theta)$ is completely specified by θ

known *unknown*

$$p(x \mid D) = \int p(x \mid \theta)\, p(\theta \mid D)\, d\theta$$

- Using Bayes formula,

$$p(\theta \mid D) = \frac{p(D \mid \theta)\, p(\theta)}{\int p(D \mid \theta)\, p(\theta)\, d\theta}$$

$$p(D \mid \theta) = \prod_{k=1}^{n} p(x_k \mid \theta)$$

Bayesian Estimation vs. MLE

- So in principle $p(x|D)$ can be computed
  - In practice, it may be hard to do integration analytically, may have to resort to numerical methods

$$p(x \mid D) = \int p(x \mid \theta) \frac{\prod_{k=1}^{n} p(x_k \mid \theta) p(\theta)}{\int \prod_{k=1}^{n} p(x_k \mid \theta) p(\theta) d\theta} d\theta$$

- Contrast this with the MLE solution which requires differentiation of likelihood to get $p(x \mid \hat{\theta})$
  - Differentiation is easy and can always be done analytically

Bayesian Estimation vs. MLE

support θ receives from the data

$$p(x \mid D) = \int p(x \mid \theta) p(\theta \mid D) d\theta$$

proposed model with certain θ

- The above equation implies that if we are less certain about the exact value of **θ,** we should consider a weighted average of *p(***x**|**θ***)* over the possible values of **θ.**

- Contrast this with the MLE solution which always gives us a single model:

$$p(x \mid \hat{\theta})$$

Bayesian Estimation for Gaussian with unknown μ

- Let $p(x|\mu)$ be $N(\mu, \sigma^2)$ that is σ^2 is known, but μ is unknown and needs to be estimated, so $\theta = \mu$
- Assume a prior over μ: $p(\mu) \sim N(\mu_0, \sigma_0^2)$
- μ_0 encodes some prior knowledge about the true mean μ , while σ_0^2 measures our prior uncertainty.
- The posterior distribution is:

$$p(\mu | D) \propto p(D | \mu) p(\mu)$$

$$= \alpha' \exp\left[-\frac{1}{2} \left(\sum_{k=1}^{n} \left(\frac{x_k - \mu}{\sigma} \right)^2 + \left(\frac{\mu - \mu_0}{\sigma_0} \right)^2 \right) \right]$$

$$= \alpha'' \exp\left[-\frac{1}{2} \left[\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right) \mu^2 - 2 \left(\frac{1}{\sigma^2} \sum_{k=1}^{n} x_k + \frac{\mu_0}{\sigma_0^2} \right) \mu \right] \right]$$

Bayesian Estimation for Gaussian with unknown μ

- Where factors that do not depend on μ have been absorbed into the constants α' and α"
- $p(\mu \mid D)$ is an exponent of a quadratic function of μ i.e. it is a normal density.
- $p(\mu \mid D)$ remains normal for any number of training samples.
- If we write

$$p(\mu \mid D) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu - \mu_n}{\sigma_n} \right)^2 \right]$$

$$\alpha'' \exp\left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\mu_0}{\sigma_0^2} \right)\mu \right] \right]$$

Bayesian Estimation for Gaussian with unknown μ

- then identifying the coefficients, we get

$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \qquad \frac{\mu_n}{\sigma_n^2} = \frac{n}{\sigma^2}\hat{\mu}_n + \frac{\mu_0}{\sigma_0^2}$$

where $\hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} x_k$ is the sample mean

- Solving explicitly for μ_n and σ_n^2 we obtain:

$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\right)\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0$$

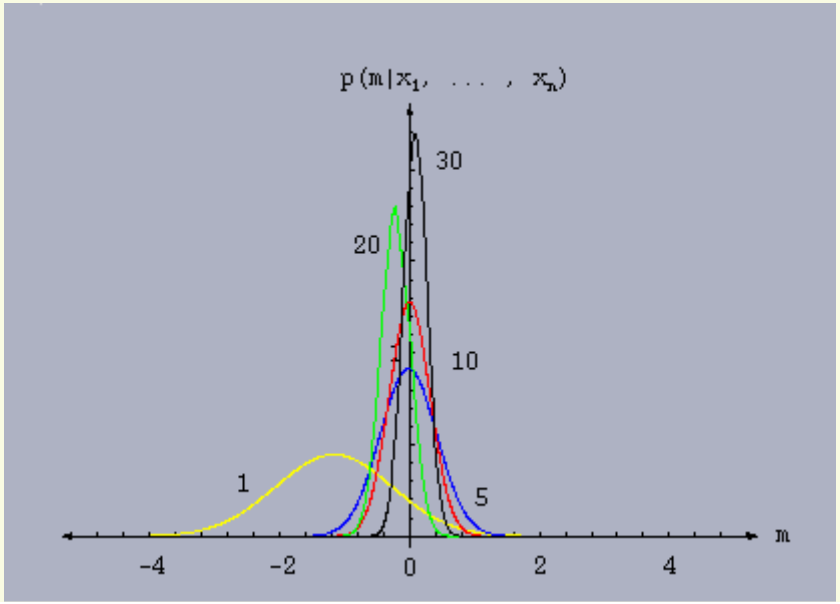
our best guess after observing n samples

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

uncertainty about the guess, decreases monotonically with n

Bayesian Estimation for Gaussian with unknown μ

- Each additional observation decreases our uncertainty about the true value of μ .
- As n increases, $p(\mu \mid D)$ becomes more and more sharply peaked, approaching a Dirac delta function as n approaches infinity. This behavior is known as Bayesian Learning.



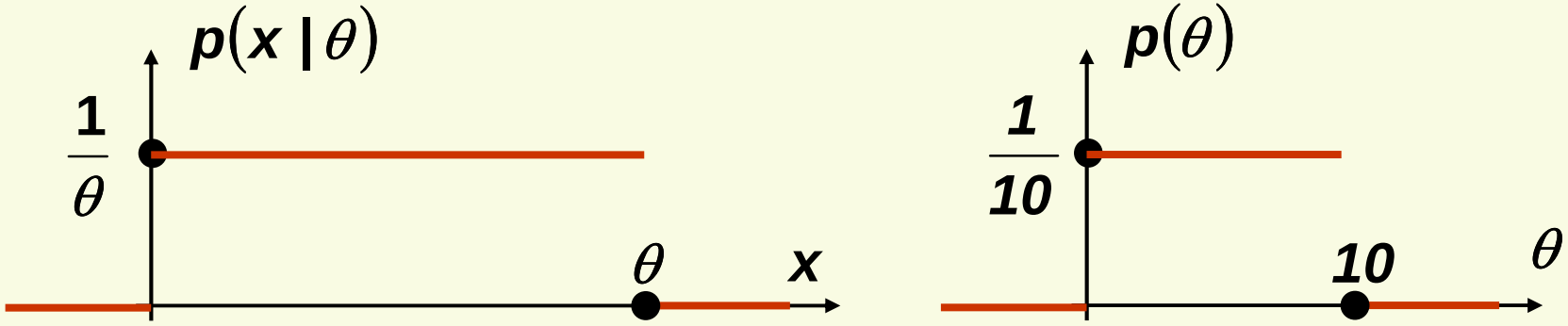
Bayesian Estimation for Gaussian with unknown μ

$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \right) \hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0$$

- In general, μ_n is a linear combination of $\hat{\mu}_n$ and μ_0 , with coefficients that are non-negative and sum to 1.
- Thus μ_n lies somewhere between $\hat{\mu}_n$ and μ_0 .
- If $\sigma_0 \neq 0$, $\mu_n \to \hat{\mu}_n$ as $n \to \infty$
- If $\sigma_0 = 0$, our a priori certainty that $\mu = \mu_0$ is so strong that no number of observations can change our opinion.
- If $\sigma_0 \approx \sigma$, a priori guess is very uncertain, and we take $\mu_n = \hat{\mu}_n$

Bayesian Estimation: Example for U[0,θ]

- Let X be U[**0,**θ]. Recall $p(x|\theta)$=1/θ inside [**0,**θ], else 0

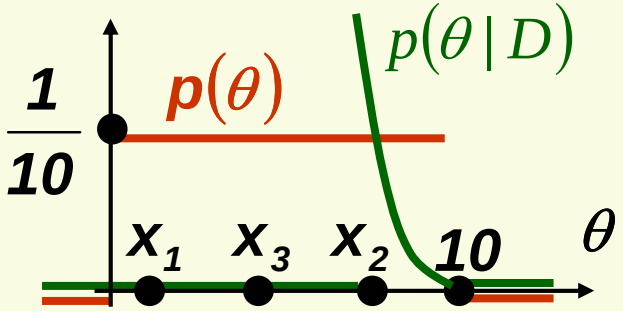


- Suppose we assume a U[0,10] prior on θ
  - good prior to use if we just know the range of θ but don't know anything else

Bayesian Estimation: Example for U[0,θ]

- We need to compute $p(x \mid D) = \int p(x \mid \theta) p(\theta \mid D) d\theta$
  - using $p(\theta \mid D) = \frac{p(D \mid \theta) p(\theta)}{\int p(D \mid \theta) p(\theta) d\theta}$ and $p(D \mid \theta) = \prod_{k=1}^{n} p(x_k \mid \theta)$
- When computing MLE of θ, we had

$$p(D \mid \theta) = \begin{cases} \frac{1}{\theta^n} & \text{for } \theta \geq \max\{x_1, \ldots, x_n\} \\ 0 & \text{otherwise} \end{cases}$$



- Thus

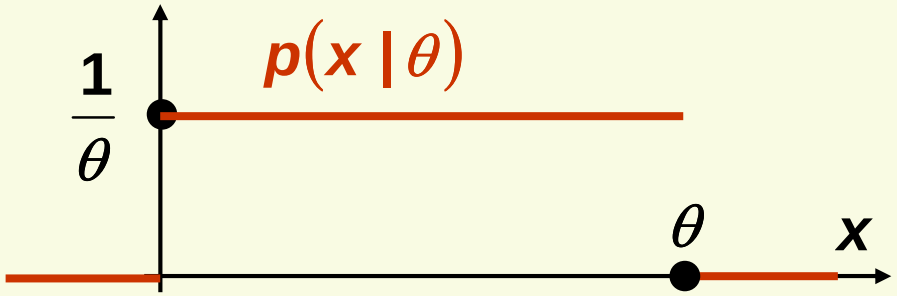
$$p(\theta \mid D) = \begin{cases} c \frac{1}{\theta^n} & \text{for } \max\{x_1, \ldots, x_n\} \leq \theta \leq 10 \\ 0 & \text{otherwise} \end{cases}$$

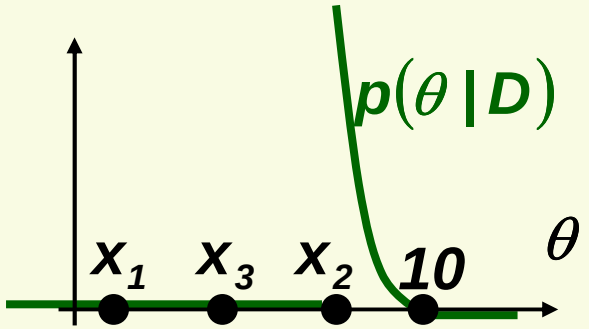
- where **c** is the normalizing constant, i.e. $c = \frac{1}{\int_{\max\{x_1, \ldots, x_n\}}^{10} \frac{d\theta}{\theta^n}}$

Bayesian Estimation: Example for U[0,θ]

- We need to compute $p(x \mid D) = \int p(x \mid \theta) p(\theta \mid D) d\theta$

$$p(\theta \mid D) = \begin{cases} c \dfrac{1}{\theta^n} & \text{for } \max\{x_1, \ldots, x_n\} \le \theta \le 10 \\ 0 & \text{otherwise} \end{cases}$$





- We have 2 cases:

1. case $x < \max\{x_1, x_2, \ldots, x_n\}$

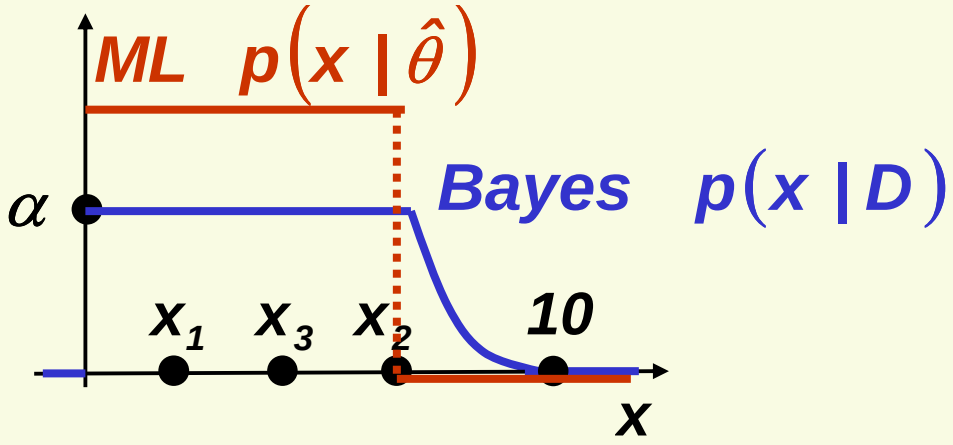
constant independent of x

$$p(x \mid D) = \int_{\max\{x_1, \ldots x_n\}}^{10} c \frac{1}{\theta^{n+1}} d\theta = \boxed{\alpha}$$

2. case $x > \max\{x_1, x_2, \ldots, x_n\}$

$$p(x \mid D) = \int_x^{10} c \frac{1}{\theta^{n+1}} d\theta = \frac{c}{-n\theta^n}\Big|_x^{10} = \boxed{\frac{c}{nx^n}} - \frac{c}{n10^n}$$

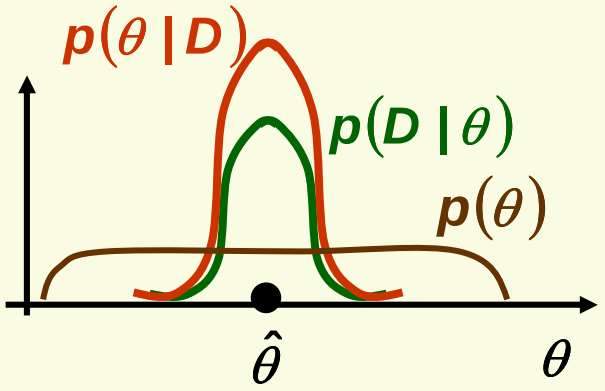
Bayesian Estimation: Example for U[0,θ]



- Note that even after x >max {x_1, x_2,…, x_n}, Bayes density is not zero, which makes sense
- curious fact: Bayes density is not uniform, i.e. does not have the functional form that we have assumed!

ML vs. Bayesian Estimation with Broad Prior

- Suppose $p(\theta)$ is flat and broad (close to uniform prior)
- $p(\theta|D)$ tends to sharpen if there is a lot of data



- Thus $p(D|\theta) \propto p(\theta|D)p(\theta)$ will have the same sharp peak as $p(\theta|D)$
- But by definition, peak of $p(D|\theta)$ is the ML estimate $\hat{\theta}$
- The integral is dominated by the peak:

$$p(x \mid D) = \int p(x \mid \theta)p(\theta \mid D)d\theta \approx p\left(x \mid \hat{\theta}\right)\int p(\theta \mid D)d\theta = p\left(x \mid \hat{\theta}\right)$$

- Thus as n goes to infinity, Bayesian estimate will approach the density corresponding to the MLE!

ML vs. Bayesian Estimation

- **Number of training data**
  - The two methods are equivalent assuming infinite number of training data (and prior distributions that do not exclude the true solution).
  - For small training data sets, they give different results in most cases.
- **Computational complexity**
  - ML uses differential calculus or gradient search for maximizing the likelihood.
  - Bayesian estimation requires complex multidimensional integration techniques.

ML vs. Bayesian Estimation

- Solution complexity
  - Easier to interpret ML solutions (i.e., must be of the assumed parametric form).
  - A Bayesian estimation solution might not be of the parametric form assumed. Hard to interpret, returns weighted average of models.
- Prior distribution
  - If the prior distribution $p(\theta)$ is uniform, Bayesian estimation solutions are equivalent to ML solutions.

ML vs. Bayesian Estimation

- Broad or asymmetric $p(\theta/D)$
  - In this case, the two methods will give different solutions.
  - Bayesian methods will explicitly exploit such information.
- General comments
  - There are strong theoretical and methodological arguments supporting Bayesian estimation.
  - In practice, ML estimation is simpler and can lead to comparable performance.