

CS145: Probability & Computing

Lecture 20: Properties of Estimates and Bayesian Parameter Estimation



Figure credits:
*Bertsekas & Tsitsiklis, **Introduction to Probability**, 2008*
*Pitman, **Probability**, 1999*

CS145: Lecture 22 Outline

- Estimator properties
- Bayesian Parameter Estimation
- Examples of Bayesian Estimators

Statistical Inference Problems

Hypothesis Testing: *How do I categorize “test data”?*

- Two (or more) mutually exclusive hypotheses: $H=0$ or $H=1$?
- The distribution of the data under each hypothesis is known:
$$p_{X|H}(x | 0), \quad p_{X|H}(x | 1) \qquad f_{X|H}(x | 0), \quad f_{X|H}(x | 1)$$
- *Goal:* Choose between hypotheses

Estimation: *How do I learn from “training data”?*

- We have n independent observations sampled from some unknown probability distribution: x_1, x_2, \dots, x_n
- We assume the distribution of our data lives in some family, but don't know the right parameter values θ
- *Goal:* Learn parameters that best “explain” the observations

Example: Bernoulli Distribution

- A *Bernoulli* or *indicator* random variable X has one parameter:

$$p_X(1) = \theta, \quad p_X(0) = 1 - \theta, \quad \mathcal{X} = \{0, 1\}$$

- The *probability mass function* for an observation x_i equals:

$$p_X(x_i; \theta) = \theta^{x_i} (1 - \theta)^{1 - x_i}$$

$$\log p_X(x_i; \theta) = x_i \log(\theta) + (1 - x_i) \log(1 - \theta)$$

- **Goal:** Estimate θ from n independent observations:

$$P(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^n p_X(x_i; \theta)$$

Example: Uniform Distribution

- A continuous *uniform distribution* between 0 and θ has the following probability density function:

$$f_X(x_i; \theta) = \frac{1}{\theta} \quad \text{if } 0 \leq x_i \leq \theta,$$

$$f_X(x_i; \theta) = 0 \quad \text{if } x_i < 0 \text{ or } x_i > \theta.$$

- **Goal:** Estimate θ from n independent observations:

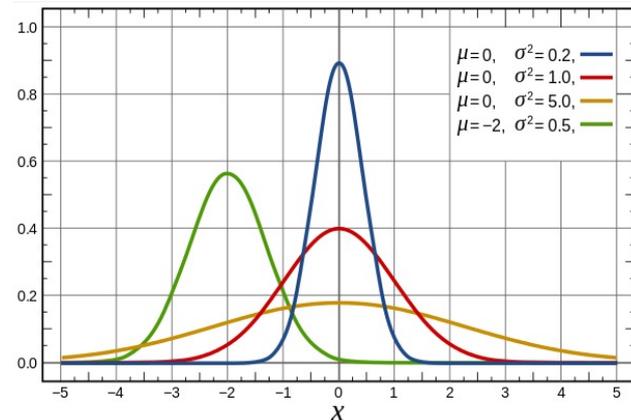
$$f(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^n f_X(x_i; \theta)$$

Example: Gaussian Distribution

- A univariate *Gaussian* distribution is parameterized by its mean and variance:

$$f_X(x_i; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2}$$

$$\theta = \{\mu, \sigma^2\}$$



- **Goal:** Estimate θ from n independent observations:

$$f(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^n f_X(x_i; \theta)$$

Maximum Likelihood (ML) Estimation

➤ Suppose I have n independent observations sampled from some unknown probability distribution: $x = \{x_1, x_2, \dots, x_n\}$

➤ Suppose I have two candidate parameter estimates where:

$$p_X(x; \theta_1) > p_X(x; \theta_2)$$

Given no other information, choose the higher likelihood model!

➤ The *maximum likelihood (ML)* parameter estimate is defined as:

$$\hat{\theta} = \arg \max_{\theta} \prod_{i=1}^n p_X(x_i; \theta) = \arg \max_{\theta} \sum_{i=1}^n \log p_X(x_i; \theta) \quad \begin{array}{l} \text{Discrete} \\ \text{Observations} \end{array}$$

$$\hat{\theta} = \arg \max_{\theta} \prod_{i=1}^n f_X(x_i; \theta) = \arg \max_{\theta} \sum_{i=1}^n \log f_X(x_i; \theta) \quad \begin{array}{l} \text{Continuous} \\ \text{Observations} \end{array}$$

Finding ML Estimates

- For many practical models, the log-likelihood is a smooth and continuous function of the parameters:

$$L(\theta) = \sum_{i=1}^n \log p_X(x_i; \theta) \qquad L(\theta) = \sum_{i=1}^n \log f_X(x_i; \theta)$$

Maximum will occur at a point where derivative equals zero!

- The *maximum likelihood (ML)* parameter estimate is defined as:

$$\hat{\theta} = \arg \max_{\theta} \prod_{i=1}^n p_X(x_i; \theta) = \arg \max_{\theta} \sum_{i=1}^n \log p_X(x_i; \theta) \quad \text{Discrete Observations}$$

$$\hat{\theta} = \arg \max_{\theta} \prod_{i=1}^n f_X(x_i; \theta) = \arg \max_{\theta} \sum_{i=1}^n \log f_X(x_i; \theta) \quad \text{Continuous Observations}$$

Example: Bernoulli Distribution

- A *Bernoulli* or *indicator* random variable X has one parameter:

$$p_X(x_i; \theta) = \theta^{x_i} (1 - \theta)^{1-x_i} \quad x_i \in \{0, 1\}$$

- The *maximum likelihood (ML)* estimate maximizes:

$$L(\theta) = \sum_{i=1}^n x_i \log(\theta) + (1 - x_i) \log(1 - \theta)$$

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{ML: Empirical fraction of successes!}$$

Example: Exponential Distribution

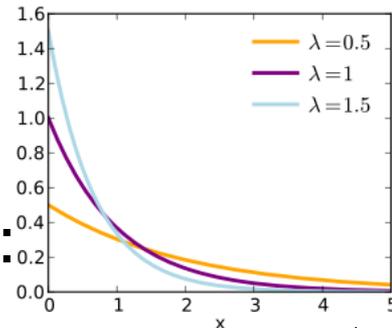
- A *geometric* random variable X has parameter:

$$f_X(x_i; \theta) = \theta e^{-\theta x_i}, \quad x_i \geq 0.$$

- The *maximum likelihood (ML)* estimate maximizes:

$$L(\theta) = \sum_{i=1}^n \log(\theta) - \theta x_i$$

$$E[X] = \frac{1}{\theta}$$



$$\hat{\theta} = \left(\frac{1}{n} \sum_{i=1}^n x_i \right)^{-1}$$

ML: Match model to empirical mean!

Example: Gaussian Distribution

- A univariate *Gaussian* distribution is:

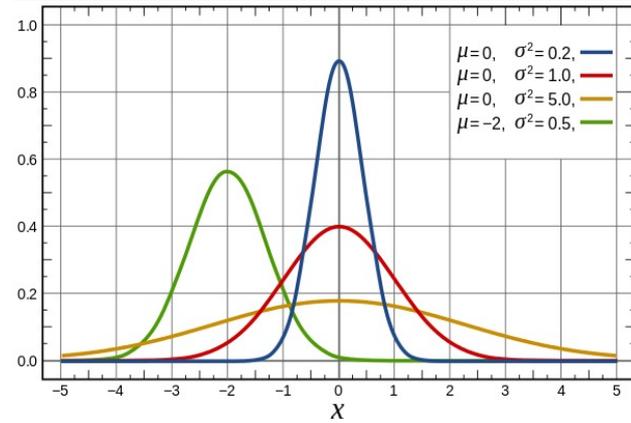
$$f_X(x_i; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2}$$

- The *ML* estimate maximizes:

$$L(\mu, \sigma) = \sum_{i=1}^n -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2) - \frac{1}{2} \left(\frac{x_i - \mu}{\sigma} \right)^2$$

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2$$



Example: Uniform Distribution

- A continuous *uniform distribution* between 0 and θ :

$$f_X(x_i; \theta) = \frac{1}{\theta} \quad \text{if } 0 \leq x_i \leq \theta,$$

$$f_X(x_i; \theta) = 0 \quad \text{if } x_i < 0 \text{ or } x_i > \theta.$$

- The *maximum likelihood (ML)* estimate maximizes:

$$L(\theta) = \prod_{i=1}^n f_X(x_i; \theta)$$

Cannot take logarithm because density can equal exactly zero.

- Optimal to choose smallest θ under which the data has positive probability:

$$\hat{\theta} = \max\{x_1, x_2, \dots, x_n\}$$

Reminder: Convergence in Probability

Convergence in Probability

Let Y_1, Y_2, \dots be a sequence of random variables (not necessarily independent), and let a be a real number. We say that the sequence Y_n **converges to a in probability**, if for every $\epsilon > 0$, we have

$$\lim_{n \rightarrow \infty} \mathbf{P}(|Y_n - a| \geq \epsilon) = 0.$$

Weak Law of Large Numbers:

X_1, X_2, \dots i.i.d.

finite mean μ and variance σ^2

$$M_n = \frac{X_1 + \dots + X_n}{n}$$

$$E[M_n] = \mu$$

➤ The “empirical mean” converges to true mean in probability:

$$\lim_{n \rightarrow \infty} P(|M_n - \mu| \geq \epsilon) = 0 \quad \text{for any } \epsilon > 0$$

Consistency of ML Estimators

$$\hat{\theta}_n = \arg \max_{\theta} \prod_{i=1}^n f_X(x_i; \theta) = \arg \max_{\theta} \sum_{i=1}^n \log f_X(x_i; \theta)$$

- An estimator is *consistent* if the sequence of estimates $\hat{\theta}_n$ converges to the true parameter in probability, for any true θ

$$\lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| \geq \epsilon) = 0 \quad \text{for any } \epsilon > 0$$

- Under “mild conditions” that are true for most distributions, the *ML parameter estimates are always consistent*

Examples: Gaussian distribution, uniform distribution, ...

- The ML estimator also satisfies a *central limit theorem*, and for large n has provably small variance (“efficiency”)

Characterization of estimations

- iid samples $\{x_1, x_2, \dots, x_n\}$ obtained from distribution \mathcal{D}
- We want to estimate a parameter θ of \mathcal{D}
- Let $\hat{\theta}_n$ be an estimate of θ obtained using n i.i.d. samples from \mathcal{D} (e.g. sample mean, ML estimators)
- $\hat{\theta}_n$ is a **Random Variable**
- What are desirable properties of the estimator $\hat{\theta}_n$?

Consistency

- An estimator is *consistent* if the sequence of estimates $\hat{\theta}_n$ converges to the true parameter in probability, for any true θ

$$\lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| \geq \epsilon) = 0 \quad \text{for any } \epsilon > 0$$

- Under “mild conditions” that are true for most distributions, the *ML parameter estimates are always consistent*
Examples: Gaussian distribution, uniform distribution, ...
- Mild conditions = log likelihood is a smooth function with unique maximum

Unbiased estimator

- An estimator is *unbiased* if its expected value corresponds to the actual correct value, that is iff

$$E_{\mathcal{D}}[\hat{\theta}] = \theta$$

- An estimator is *asymptotically unbiased* if its expected value converges to the actual correct value as $n \rightarrow \infty$, that is iff

$$\lim_{n \rightarrow \infty} E_{\mathcal{D}^n}[\hat{\theta}_n] = \theta$$

- Any unbiased estimator is also asymptotically unbiased but NOT viceversa

Example: sample average

- iid samples $\{x_1, x_2, \dots, x_n\}$ obtained from distribution \mathcal{D}
- Sample average $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n x_i$
- Is it unbiased? ✓
- Is it asymptotically unbiased? ✓
- Is it consistent? ✓

Example: modified sample average

- iid samples $\{x_1, x_2, \dots, x_n\}$ obtained from distribution \mathcal{D}
- Sample average $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n x_i + \frac{1}{n}$
- Is it unbiased? ❌
- Is it asymptotically unbiased? ✅
- Is it consistent? ✅

Example: single sample estimate

- iid samples $\{x_1, x_2, \dots, x_n\}$ obtained from distribution \mathcal{D}
- Single sample estimate $\hat{\theta} = x_1$
- Is it unbiased? 
- Is it asymptotically unbiased? 
- Is it consistent?  Unless $\text{Var} = 0$

Example: ML estimators Bernoulli

- iid samples $\{x_1, x_2, \dots, x_n\}$ obtained from a Bernoulli dist.
- Estimate probability of head $\hat{p} = \frac{1}{n} \sum_{i=1}^n x_i$
- Is it unbiased? ✓
- Is it asymptotically unbiased? ✓
- Is it consistent? ✓

Example: ML estimator uniform distribution

- iid samples $\{x_1, x_2, \dots, x_n\}$ obtained from $\mathcal{U}(0, A)$
- Estimate for A given by $Y = \max\{x_i\}$
- Is it unbiased?

Example: ML estimator uniform dist

- Sample $\{x_1, x_2, \dots, x_n\}$ obtained from $\mathcal{U}(0, A)$
- Estimate $Y = \max\{x_i\}$
- Is it unbiased? 

$$\begin{aligned} E(\max(x_i)) &= E(Y) = \int_{-\infty}^{\infty} y f_Y(y) dy = \int_0^A y n \left(\frac{1}{A}\right)^n y^{n-1} dy = \int_0^A n \left(\frac{y}{A}\right)^n dy = \\ & n \left(\frac{1}{A}\right)^n \left(\frac{y^{n+1}}{n+1}\right) \Big|_0^A = \left(\frac{n}{n+1}\right) \left(\frac{1}{A}\right)^n A^{n+1} = \left(\frac{n}{n+1}\right) A < A \end{aligned}$$

Example: ML estimator uniform dist

- Sample $\{x_1, x_2, \dots, x_n\}$ obtained from $\mathcal{U}(0, A)$
- Estimate $Y = \max\{x_i\}$
- Is it asymptotically unbiased?

$$E(\max(x_i)) = E(Y) = \int_{-\infty}^{\infty} y f_Y(y) dy = \int_0^A y n \left(\frac{1}{A}\right)^n y^{n-1} dy = \int_0^A n \left(\frac{y}{A}\right)^n dy =$$
$$n \left(\frac{1}{A}\right)^n \left(\frac{y^{n+1}}{n+1}\right) \Big|_0^A = \left(\frac{n}{n+1}\right) \left(\frac{1}{A}\right)^n A^{n+1} = \left(\frac{n}{n+1}\right) A < A$$

Example: ML estimator uniform dist

- Sample $\{x_1, x_2, \dots, x_n\}$ obtained from $\mathcal{U}(0, A)$
- Estimate $Y = \max\{x_i\}$
- Is it asymptotically unbiased? 

$$\lim_{n \rightarrow \infty} E[Y] = \lim_{n \rightarrow \infty} \frac{n}{n+1} A = A$$

Example: ML estimator uniform dist

- iid Samples $\{x_1, x_2, \dots, x_n\}$ obtained from $\mathcal{U}(0, A)$
- Estimate $Y = \max\{x_i\}$
- Is it consistent?

Example: ML estimator uniform dist

➤ i.i.d. samples $\{x_1, x_2, \dots, x_n\}$ obtained from $\mathcal{U}(0, A)$

➤ Estimate $Y = \max\{x_i\}$

➤ Is it consistent?

$$\text{Note that } P(\max_{1 \leq i \leq n} X_i \leq t) = \begin{cases} 0 & \text{if } t \leq 0 \\ \left(\frac{t}{A}\right)^n & \text{if } t \in [0, A] \\ 1 & \text{if } t \geq A \end{cases}$$



For $\epsilon > 0$,

$$P(|\max_{1 \leq i \leq n} X_i - A| > \epsilon) = P(\max_{1 \leq i \leq n} X_i \geq A + \epsilon) + P(\max_{1 \leq i \leq n} X_i \leq A - \epsilon) = \begin{cases} \left(\frac{A - \epsilon}{A}\right)^n & \text{if } \epsilon < A \\ 0 & \text{if } \epsilon \geq A \end{cases}$$

which goes to 0 as $n \rightarrow \infty$.

Relation between properties

- If an estimate is unbiased, is it also consistent?

Relation between properties

➤ If an estimate is unbiased, is it also consistent?

➤ If $Var[\hat{\theta}] \rightarrow 0$ then YES

From Weak Law of large numbers

CS145: Lecture 22 Outline

- Estimator properties
- Bayesian Parameter Estimation
- Examples of Bayesian Estimators

Statistical Inference Problems

Hypothesis Testing: *How do I categorize “test data”?*

- Two (or more) mutually exclusive hypotheses: $H=0$ or $H=1$?
- The distribution of the data under each hypothesis is known:
 $p_{X|H}(x | 0)$, $p_{X|H}(x | 1)$ $f_{X|H}(x | 0)$, $f_{X|H}(x | 1)$
- *Goal:* Choose between hypotheses

Estimation: *How do I learn from “training data”?*

- We have n independent observations sampled from some unknown probability distribution: x_1, x_2, \dots, x_n
- We assume the distribution of our data lives in some family, but don't know the right parameter values θ
- *Goal:* Learn parameters that best “explain” the observations

Maximum Likelihood (ML) Estimation

➤ Suppose I have n independent observations sampled from some unknown probability distribution: $x = \{x_1, x_2, \dots, x_n\}$

➤ Suppose I have two candidate parameter estimates where:

$$p_X(x; \theta_1) > p_X(x; \theta_2)$$

Given no other information, choose the higher likelihood model!

➤ The *maximum likelihood (ML)* parameter estimate is defined as:

$$\hat{\theta} = \arg \max_{\theta} \prod_{i=1}^n p_X(x_i; \theta) = \arg \max_{\theta} \sum_{i=1}^n \log p_X(x_i; \theta) \quad \begin{array}{l} \textit{Discrete} \\ \textit{Observations} \end{array}$$

$$\hat{\theta} = \arg \max_{\theta} \prod_{i=1}^n f_X(x_i; \theta) = \arg \max_{\theta} \sum_{i=1}^n \log f_X(x_i; \theta) \quad \begin{array}{l} \textit{Continuous} \\ \textit{Observations} \end{array}$$

Degeneracies in ML Estimation

$$\hat{\theta}_n = \arg \max_{\theta} \prod_{i=1}^n f_X(x_i | \theta) = \arg \max_{\theta} \sum_{i=1}^n \log f_X(x_i | \theta)$$

- The theory justifying ML estimates is *asymptotic*: they have good properties as n becomes very large
- But they can have poor properties with small datasets.

Example: ML estimate of Bernoulli with no observed heads.

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n x_i = 0 \quad \text{if } x_i = 0 \text{ for all } i \quad \textit{assumes observing heads in future is impossible!}$$

- More generally, ML estimates can often give parameter estimates that are too “extreme” (too large or too small)

Bayesian Parameter Estimation

- Suppose I have n independent observations sampled from some unknown probability distribution: $x = \{x_1, x_2, \dots, x_n\}$
- We have a *likelihood model* with unknown parameters:

$$f_{X|\Theta}(x | \theta) = \prod_{i=1}^n f_{X|\Theta}(x_i | \theta)$$

- We have a *prior distribution* on parameters (possible models):

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

- *Posterior distribution* on parameters, given data, is then:

$$f_{\Theta|X}(\theta | x) = \frac{1}{f_X(x)} f_{\Theta}(\theta) \prod_{i=1}^n f_{X|\Theta}(x_i | \theta)$$

Types of priors: proper vs improper

➤ Proper prior if $\int f_{\Theta}(\theta) d\theta = 1$

➤ Improper prior if $\int f_{\Theta}(\theta) d\theta \neq 1$

Improper priors sometimes used for **uninformative priors**

Types of priors: informative vs uninformative

- **Uninformative priors** express “vague” or “general” information about a variable
 - the variable is positive
 - the value of the variable is within a limit
 - **Principle of indifference** all possible values of Θ are equally likely
- **Weakly Informative priors** express partial information about a variable to loosely constrain the value of Θ into a range – Used for **regularization**
- **Informative priors** express specific, definite information about a variable which significantly constrains the ranges of values of Θ
 - An example: a prior distribution for the temperature at noon tomorrow selected as a normal distribution with expected value equal to today's noontime temperature, with variance equal to the day-to-day variance of atmospheric temperature

Bayesian Parameter Estimation

- **Maximum a Posteriori (MAP)** parameter estimate:
Choose the parameters with largest posterior probability.

$$\hat{\theta} = \arg \max_{\theta} f_{\Theta|X}(\theta | x) = \arg \max_{\theta} f_{\Theta}(\theta) \prod_{i=1}^n f_{X|\Theta}(x_i | \theta)$$

- **Conditional Expectation** parameter estimate:
Set the parameters to the mean of the posterior distribution.

$$\hat{\theta} = E[\theta | x] = \int \theta f_{\Theta|X}(\theta | x) d\theta$$

- **Posterior distribution** on parameters, given data, is then:

$$f_{\Theta|X}(\theta | x) = \frac{1}{f_X(x)} f_{\Theta}(\theta) \prod_{i=1}^n f_{X|\Theta}(x_i | \theta)$$

Bayesian Parameter Estimation

- **Maximum a Posteriori (MAP)** parameter estimate:
Choose the parameters with largest posterior probability.

$$\hat{\theta} = \arg \max_{\theta} f_{\Theta|X}(\theta | x) = \arg \max_{\theta} f_{\Theta}(\theta) \prod_{i=1}^n f_{X|\Theta}(x_i | \theta)$$

- **Conditional Expectation** parameter estimate:
Set the parameters to the mean of the posterior distribution.

$$\hat{\theta} = E[\theta | x] = \int \theta f_{\Theta|X}(\theta | x) d\theta$$

- Both estimators pick parameters with high posterior probability
- Choice of estimator can be formalized via *decision theory* (generalization of earlier analysis of hypothesis testing)

Example: Bernoulli Distribution

- A *Bernoulli* or *indicator* random variable X has one parameter:

$$p_X(x_i | \theta) = \theta^{x_i} (1 - \theta)^{1 - x_i} \quad x_i \in \{0, 1\}$$

- Suppose we place a *uniform prior distribution*:

$$f_{\Theta}(\theta) = 1, \quad 0 \leq \theta \leq 1.$$

- The *posterior distribution* is then:

$$f_{\Theta|X}(\theta | x) = \frac{1}{f_X(x)} f_{\Theta}(\theta) \prod_{i=1}^n p_{X|\Theta}(x_i | \theta) \propto \prod_{i=1}^n p_{X|\Theta}(x_i | \theta)$$

Example: Bernoulli Distribution

- A *Bernoulli* or *indicator* random variable X has one parameter:

$$p_X(x_i | \theta) = \theta^{x_i} (1 - \theta)^{1-x_i} \quad x_i \in \{0, 1\}$$

- The posterior distribution given n observation equals:

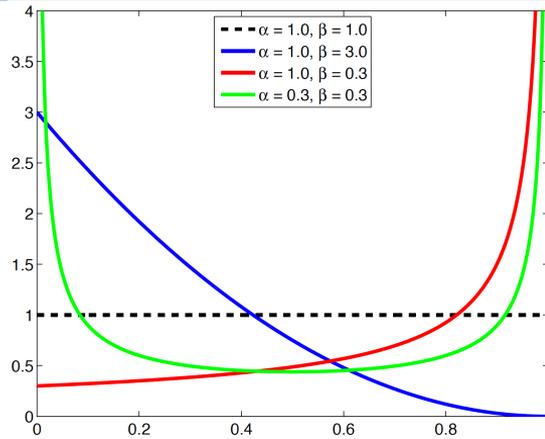
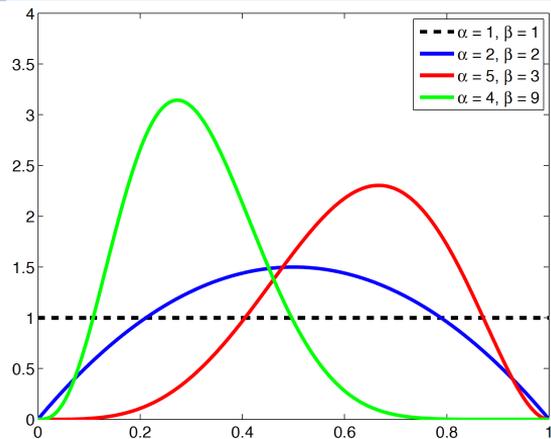
$$f_{\Theta|X}(\theta | x) \propto \theta^{n_1} (1 - \theta)^{n_0}$$
$$n_1 = \sum_{i=1}^n x_i$$
$$n_0 = n - n_1$$

This is an example of a beta distribution.

- The *posterior distribution* is then:

$$f_{\Theta|X}(\theta | x) = \frac{1}{f_X(x)} f_{\Theta}(\theta) \prod_{i=1}^n p_{X|\Theta}(x_i | \theta) \propto \prod_{i=1}^n p_{X|\Theta}(x_i | \theta)$$

The Beta Distribution



$$\text{Beta}(\theta \mid a, b) = \frac{1}{B(a, b)} \theta^{a-1} (1 - \theta)^{b-1} \quad 0 \leq \theta \leq 1$$

$$E[\theta] = \frac{a}{a + b}$$

$$\text{Var}(\theta) = \frac{ab}{(a + b)^2 (a + b + 1)}$$

$$B(a, b) \triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)}$$

Bayesian Estimators for Bernoulli

- A *Bernoulli* or *indicator* random variable X has one parameter:

$$p_X(x_i | \theta) = \theta^{x_i} (1 - \theta)^{1-x_i} \quad x_i \in \{0, 1\}$$

- The posterior distribution given n observation equals:

$$f_{\Theta|X}(\theta | x) \propto \theta^{n_1} (1 - \theta)^{n_0} \quad \begin{aligned} n_1 &= \sum_{i=1}^n x_i \\ n_0 &= n - n_1 \end{aligned}$$

$$f_{\Theta|X}(\theta | x) = \text{Beta}(\theta | n_1 + 1, n_0 + 1)$$

- This gives the following *Bayesian parameter estimates*:

$$\hat{\theta} = \frac{n_1}{n}$$

$$\hat{\theta} = \frac{n_1 + 1}{n + 2}$$

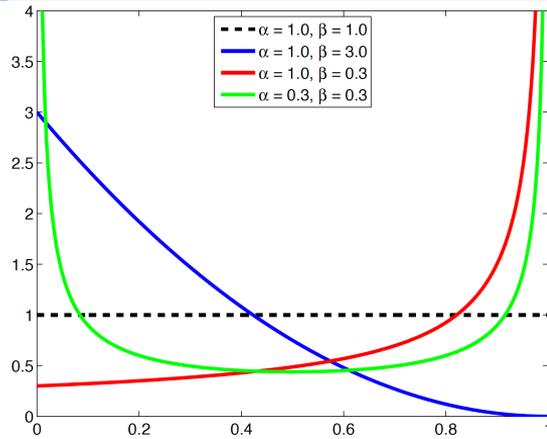
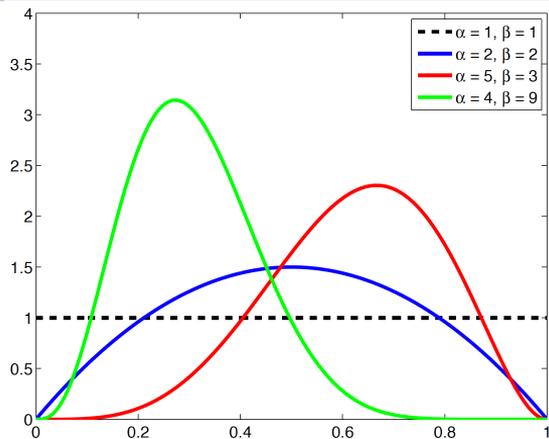
Maximum a Posteriori (MAP)

= ML with uniform prior

Conditional Expectation

= “add one” to counts

Bayesian Estimators for Bernoulli



$$f_{\Theta|X}(\theta | x) = \text{Beta}(\theta | n_1 + 1, n_0 + 1)$$

$$n_1 = \sum_{i=1}^n x_i$$

$$n_0 = n - n_1$$

➤ This gives the following *Bayesian parameter estimates*:

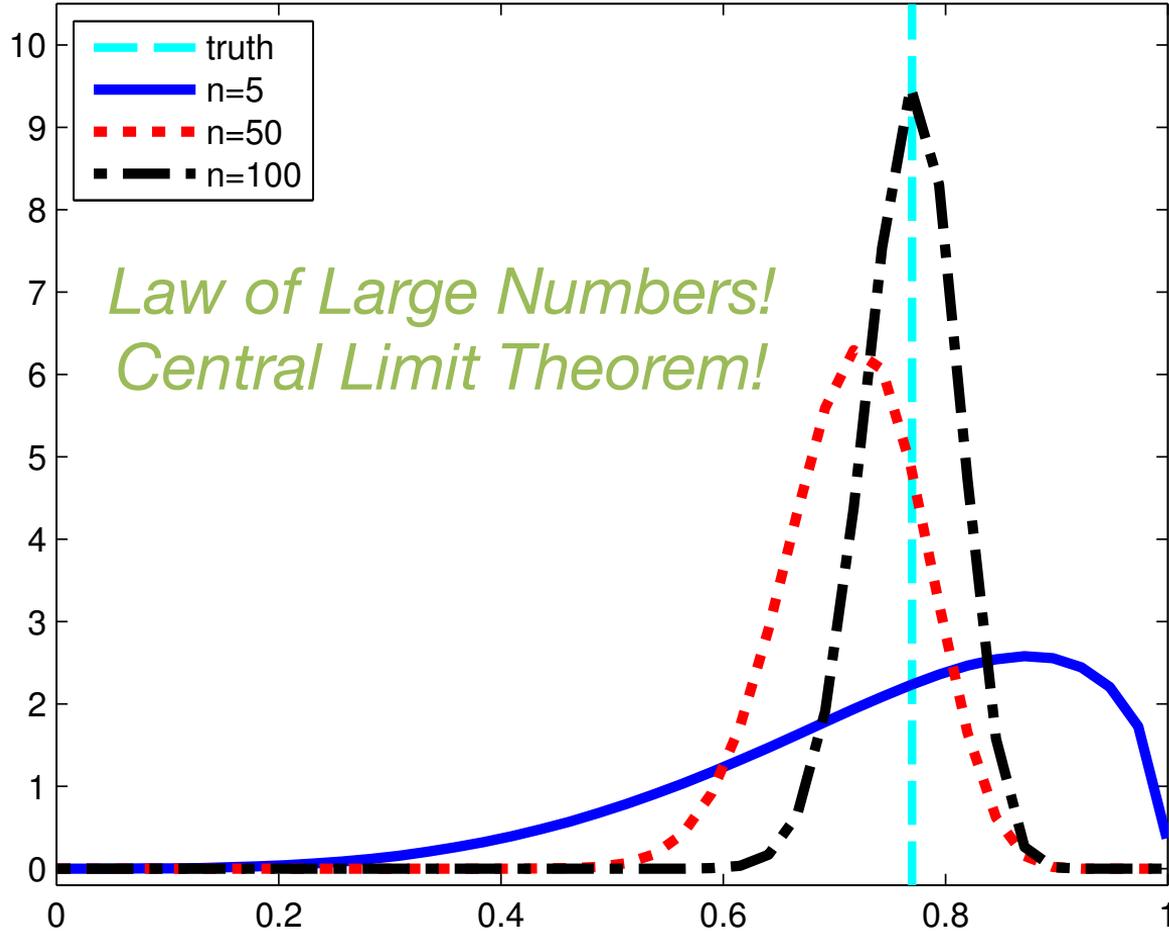
$$\hat{\theta} = \frac{n_1}{n}$$

Maximum a Posteriori (MAP)
= *ML with uniform prior*

$$\hat{\theta} = \frac{n_1 + 1}{n + 2}$$

Conditional Expectation
= “*add one*” to counts

A Sequence of Beta Posteriors



CS145: Lecture 22 Outline

- Estimator properties
- Bayesian Parameter Estimation
- Examples of Bayesian Estimators

Example: Uniform Distribution

- A continuous *uniform distribution* between 0 and θ :

$$f_X(x_i | \theta) = \frac{1}{\theta} \quad \text{if } 0 \leq x_i \leq \theta,$$

$$f_X(x_i | \theta) = 0 \quad \text{if } x_i < 0 \text{ or } x_i > \theta.$$

- Suppose we place a *uniform prior distribution*:

$$f_{\Theta}(\theta) = 1, \quad 0 \leq \theta \leq 1.$$

- The *posterior distribution*, given one observation x , is then:

$$f_{\Theta|X}(\theta | x) = \frac{f_{\Theta}(\theta) f_{X|\Theta}(x | \theta)}{\int_0^1 f_{\Theta}(\theta') f_{X|\Theta}(x | \theta') d\theta'} = \frac{1/\theta}{\int_x^1 \frac{1}{\theta'} d\theta'} = \frac{1}{\theta \cdot |\log x|}, \quad \text{if } x \leq \theta \leq 1,$$

Example: Uniform Distribution

- *Maximum a Posteriori (MAP)* estimate:
(= Maximum Likelihood with uniform prior)

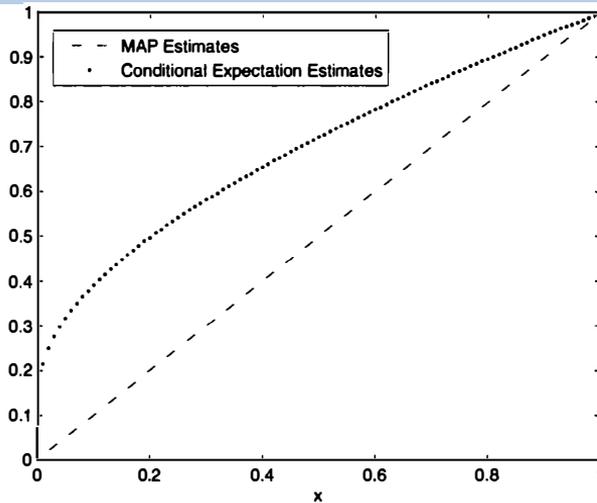
$$\hat{\theta} = x \quad \text{assumes observing larger data is impossible!}$$

- *Conditional expectation* estimate:

$$\mathbf{E}[\Theta | X = x] = \int_x^1 \theta \frac{1}{\theta \cdot |\log x|} d\theta = \frac{1-x}{|\log x|}.$$

- The *posterior distribution*, given one observation x , is then:

$$f_{\Theta|X}(\theta | x) = \frac{f_{\Theta}(\theta) f_{X|\Theta}(x | \theta)}{\int_0^1 f_{\Theta}(\theta') f_{X|\Theta}(x | \theta') d\theta'} = \frac{1/\theta}{\int_x^1 \frac{1}{\theta'} d\theta'} = \frac{1}{\theta \cdot |\log x|}, \quad \text{if } x \leq \theta \leq 1,$$



Example: Gaussian distribution

- Gaussian distribution with fixed variance but uncertain mean:

$$f_X(x_i | \theta) = \frac{1}{\sqrt{2\pi\nu}} \exp \left\{ -\frac{(x_i - \theta)^2}{2\nu} \right\}$$

- Suppose we place a zero-mean *Gaussian prior distribution*:

$$f_\Theta(\theta) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp \left\{ -\frac{\theta^2}{2\sigma_0^2} \right\}$$

- The *Gaussian posterior distribution* given n observations:

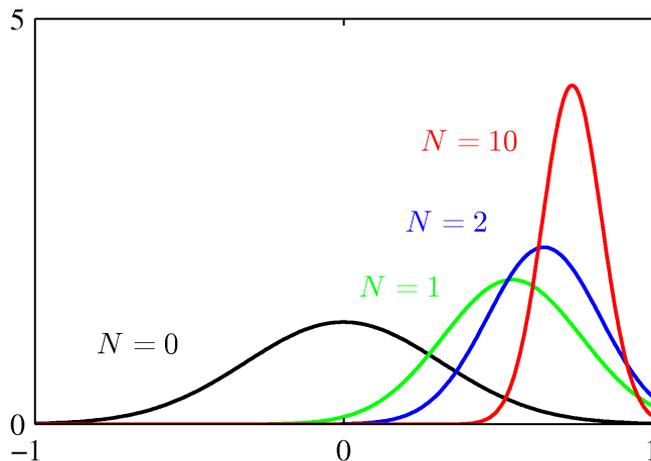
$$f_{\Theta|X}(\theta | x) = \frac{1}{f_X(x)} f_\Theta(\theta) \prod_{i=1}^n f_{X|\Theta}(x_i | \theta) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp \left\{ -\frac{(\theta - \mu_n)^2}{2\sigma_n^2} \right\}$$
$$\mu_n = \left(\frac{\sigma_0^2}{\sigma_0^2 + \nu/n} \right) \cdot \left(\frac{1}{n} \sum_{i=1}^n x_i \right) \quad \frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\nu}$$

Posterior Mean versus Empirical Mean

Optimal Estimator:

Posterior mean,
Posterior mode, &
Posterior median

$$\hat{\theta} = \mu_N = E[\theta | x]$$



Example:

Posterior given varying
amounts of data n

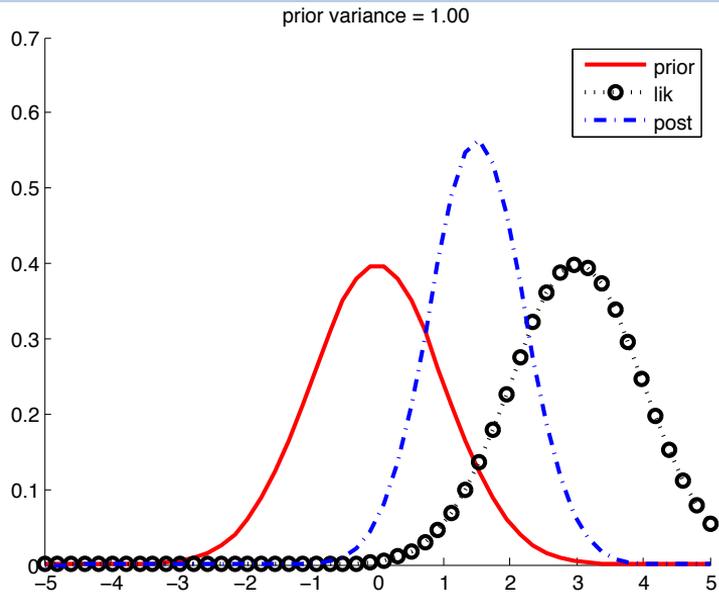
$$\theta = 0.8$$

$$\nu = 0.1$$

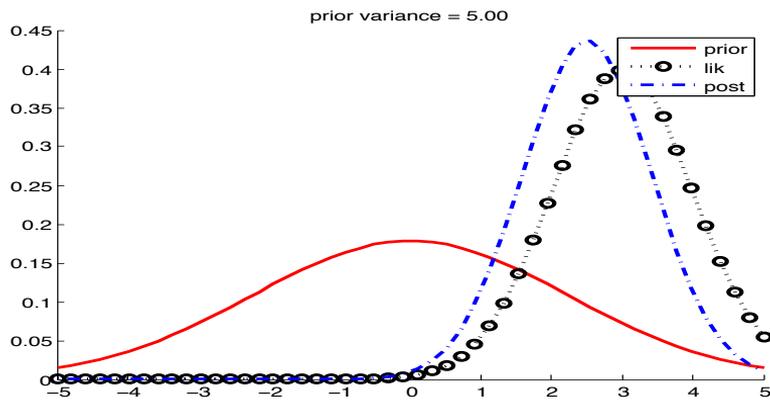
- The *posterior mean* approaches *empirical mean* (ML) as $n \rightarrow \infty$
- The posterior variance shrinks (*law of large numbers*) as $n \rightarrow \infty$

$$f_{\Theta|X}(\theta | x) = \frac{1}{f_X(x)} f_{\Theta}(\theta) \prod_{i=1}^n f_{X|\Theta}(x_i | \theta) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp \left\{ -\frac{(\theta - \mu_n)^2}{2\sigma_n^2} \right\}$$
$$\mu_n = \left(\frac{\sigma_0^2}{\sigma_0^2 + \nu/n} \right) \cdot \left(\frac{1}{n} \sum_{i=1}^n x_i \right) \quad \frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\nu}$$

Impact of Prior Variance



Example: Posteriors given same single observation, for two different priors.

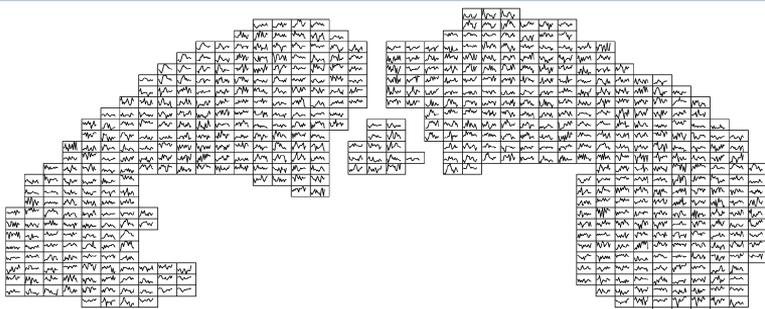


$$f_{\Theta|X}(\theta | x) = \frac{1}{f_X(x)} f_{\Theta}(\theta) \prod_{i=1}^n f_{X|\Theta}(x_i | \theta) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp \left\{ -\frac{(\theta - \mu_n)^2}{2\sigma_n^2} \right\}$$
$$\mu_n = \left(\frac{\sigma_0^2}{\sigma_0^2 + \nu/n} \right) \cdot \left(\frac{1}{n} \sum_{i=1}^n x_i \right) \quad \frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\nu}$$

Food for thought

- Are the example Bayesian ML estimates unbiased?
- Asymptotically unbiased?
- Consistent?

Brain State Classification from fMRI



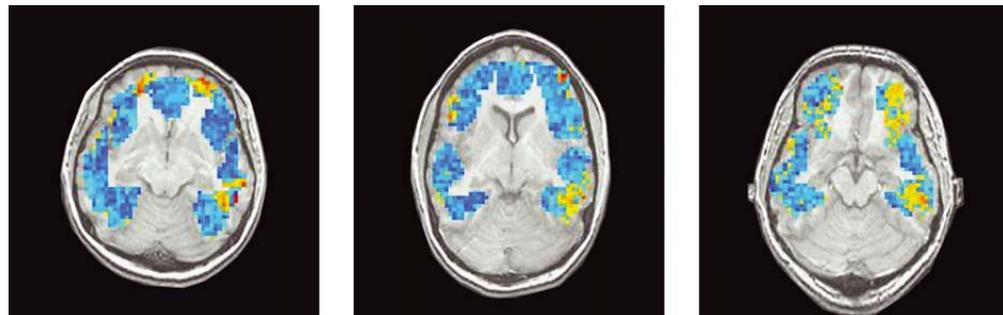
“fish,” “four-legged animals,” “trees,”
 “flowers,” “fruits,” “vegetables,” “family
 members,” “occupations,” “tools,” “kitchen
 items,” “dwellings,” & “building parts.”

$$f : \text{fMRI}(t) \rightarrow \text{WordCategory}$$



Table 1. Error rates for classifiers across all studies.

Study	Examples per class	Feature selection	Classifier					
			GNB	SVM	1NN	3NN	5NN	9NN
Picture vs. Sentence	40	Yes	0.18	0.11	0.22	0.18	0.18	0.19
	40	No	0.34	0.34	0.44	0.44	0.41	0.38
Semantic Categories	32	Yes	0.08	N/A	0.31	0.21	0.17	0.14
	32	No	0.10	N/A	0.40	0.40	0.40	0.25
Syntactic Ambiguity	10	Yes	0.25	0.28*	0.39	0.39	0.38	0.34
	10	No	0.41	0.38	0.50	0.46	0.47	0.43



single-voxel accuracy (red high, blue low)

Gaussian Naïve Bayes Classifiers, Mitchell et al., Machine Learning 2004.

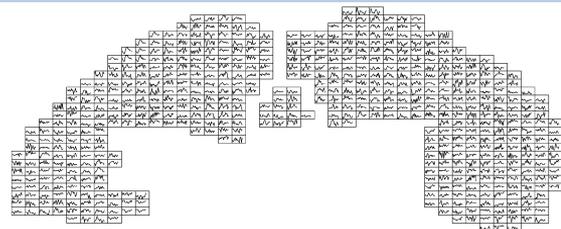
Naïve Bayes for Continuous Features

- Every feature has class-specific mean, variance:

$$\mu_{kd} = E[x_{nd} \mid y_n = k]$$

$$\nu_{kd} = \text{Var}[x_{nd} \mid y_n = k]$$

- Maximum likelihood estimates would compute the empirical mean and variance of every voxel/pixel, for every every class
- Because there are many voxels and only a limited number of scans, performance improves by placing appropriate priors on (especially the variance) parameters



N scans, D voxels, $x_{nd} \in \mathbb{R}$

$y_n = k$ if data n is class k

N digits, D pixels, $x_{nd} \in \mathbb{R}$

