CS 5135/6035 Learning Probabilistic Models

Course Review

Gowtham Atluri

December 02, 2018

Module 1: Probability Foundations

Topics

- Random Variables, Domain, Distribution
- Axioms, Principles
  - Conditional Probability, Bayes' Rule
  - Independence, Marginalization, etc.
- Standard Probability Distributions
  - Discrete
  - Continuous
- Multivariate Probability Distributions
- Probabilistic Reasoning
- Parameter Estimation
  - Max. Likelihood Estimation
  - Bayesian Estimation
- Properties of Estimators

Module 2: Maximum Likelihood Estimation

Topics

- General approach to MLE
  - I.I.D
  - Likelihood $\mathcal{L}(\theta|x)$, Log-Likelihood ℓ, Maximizing ℓ
  - Optimization algos: Gradient Descent/Newton Method
- Univariate Parameter Est. using MLE
- Multivariate Parameter Est. using MLE
- Logistic Regression
  - Max. Conditional Likelihood
- Latent variables
  - Mixture Models: Discrete latent vars.
  - Factor Models: Continuous latent vars.
- Expectation-Maximization
  - General Approach
  - Proof of correctness

Module 3: Bayesian Parameter Estimation

Topics

- General approach to Bayesian estimation
  - Prior, Likelihood, Posterior
  - Why/Why not Bayesian estimation?
- Priors
  - Noninformative
  - Conjugate Priors
  - Natural Conjugacy
  - Mixture of Priors
  - Jeffrey's Prior
- Posterior
  - Univariate
  - Multivariate: Nuisance Parameter, Marginal Posterior
- Summarization of Posterior
  - Point Estimation (Bayes' Risk)
  - Interval Estimation

Module 4: Bayesian Computation

Topics

- Sampling from Posterior
  - Pseudo random number generator
  - Inverse-Transform Method
  - Accept-Reject Method
- Monte Carlo Integration
  - General Approach
  - Importance Sampling
- Markov Chain Monte Carlo Methods
  - Markov Chain: Stationarity and other properties
  - Metropolis-Hastings
    - General Approach
    - Random-walk Metropolis-Hastings
    - Independent Metropolis-Hastings
  - Gibbs Sampling
    - Application: Hierarchical Models

Module 0: Course Overview and Julia

Learning **Probabilistic** Models

> "As far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality."
> — Albert Einstein

- Source of uncertainty:
  - **incomplete/noisy data**
    - not all data can be collected
  - **incomplete knowledge**
    - not all functions of a gene are known
  - **inherent randomness**
- Probability theory is a mathematical language for **representing and manupulating uncertainty**.

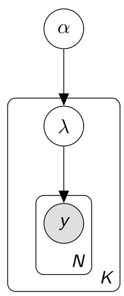
The inevitable reconciliation of **Fortuna** (goddess of chance) and **Sapientia** (wisdom incarnate). 16th century wood engraving.

Learning **Probabilistic** Models

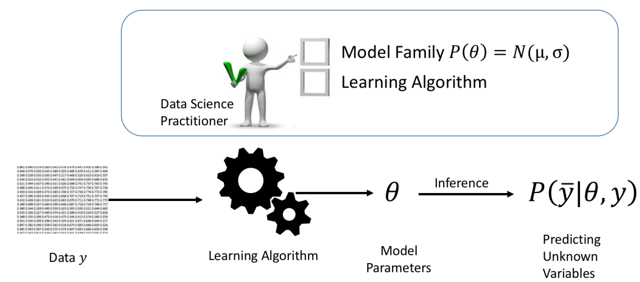
- Probability theory is a mathematical language for **representing and manupulating uncertainty**.

Advantages of probability models

- They are conceptually simple
  - Probability distributions are used to represent all uncertain unobserved quantities in a model and how they relate to the data.
- Support hierarchical construction
  - Simple probabilitic models of one or a few variables can be used to construct larger, more complex models.
- Easier to understand even complex models
  - The compositionality of probabilistic models makes it much easier to understand the models.



Learning Probabilistic Models



- Major tasks:
  - **Learning:** Given a set of samples that are known/assumed to be generated from a model, the goal is to determine the parameters of the model.
  - **Inference:** Given a set of model parameters and an observation of some variable(s), the goal is to predict states of other variables.

Module 1: Probability Foundations

Module 1: Probability Foundations

Topics

- Random Variables, Domain, Distribution
- Axioms, Principles
  - Conditional Probability, Bayes' Rule
  - Independence, Marginalization, etc.
- Standard Probability Distributions
  - Discrete
  - Continuous
- Multivariate Probability Distributions
- Probabilistic Reasoning
- Parameter Estimation
  - Max. Likelihood Estimation
  - Bayesian Estimation
- Properties of Estimators

Probability Foundations Lec 3

Given

- x is a random variable
- its domain is $\mathrm{dom}(x) = \{s1, s2, \ldots, sn\}$
  - these values/states are outcomes of a random phenomenon/experiment

A full specification of the probability values for each of the variable states, $p(x)$, is a probability distribution.

For example, in the case of a coin toss,

- $p(c = \text{heads}) = 0.5$
- $p(c = \text{tails}) = 0.5$

Kolmogorov axioms

- $0 \leq p(x = s) \leq 1$
- $\sum_x p(x) = 1$
- $p(x = s1 \cup x = s2) = P(x = s1) + p(x = s2)$

Probability Foundations Lec 3

- Joint Probability (*and*): $p(x = \text{a and } y = \text{b})$
- OR: $p(x \text{ or } y) \equiv p(x \cup y) = p(x) + p(y) - p(x \text{ and } y)$
- Marginalization: $p(x) = \sum_y p(x, y)$
- Conditional Probability: $p(x|y) = \frac{p(x,y)}{p(y)}$
- Bayes' rule: $p(x|y) = \frac{p(y|x)p(x)}{p(y)}$
- Independence:
  $x \perp\!\!\!\perp y \implies p(x, y) = p(x)p(y) \implies p(x|y) = p(x) \Leftrightarrow p(y|x) = p(y)$
  - for all states of x and y
- Conditional Independence: $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} | \mathcal{Z}$

$$p(\mathcal{X}, \mathcal{Y}|\mathcal{Z}) = p(\mathcal{X}|\mathcal{Z})p(\mathcal{Y}|\mathcal{Z}) \text{ and } p(\mathcal{X}|\mathcal{Y}, \mathcal{Z}) = p(\mathcal{X}|\mathcal{Z})$$

  - for all states of x, y, and z

Discrete Prob. Distributions Lec 4

- A random variable x is said to be discrete if it can take on only a finite number – or a countably infinite number – of possible values.
- The probability distribution of a discrete random variable is called a probability mass function (pmf).
- Cumulative distribution function $cdf(b)$ for a random variable x is $p(x \leq b) = \sum_{x=-\infty}^{b} p(x)$
- Expectation of a rand. var. $\mathbb{E}(x) = \sum_x x p(x)$
  - $\mathbb{E}(aX) = a\mathbb{E}(X)$
  - $\mathbb{E}(\sum_i a_i X_i) = \sum_i a_i \mathbb{E}(X_i)$
  - For indep. rand. vars. $\mathbb{E}(\prod_i X_i) = \sum_i \mathbb{E}(X_i)$
- Variance $\sigma^2 = \mathbb{E}[(x - \mu)^2]$

Standard Discrete Prob. Distr. Lec 4

- Bernoulli Distribution
- Binomial Distribution
- Categorical Distribution
- Multinomial Distribution
- Geometric Distribution
- Negative Binomial Distribution
- Poisson Distribution

Questions:

- What scenarios are these distributions suited for?
- What is the domain?
- What do the parameters mean?
- What is the prob. that $x = a$ or $x \leq a$ or $x \geq a$?

Continuous Prob. Distributions Lec 5

- A random variable x is said to be continuous if its domain contains continuous values.
- A function $f(x)$ that models the relative frequency behavior of the continuous valued data is called probability density function (pdf).
- Things to note:
  - $p(x = a) = \int_a^a f(x)dx = 0$
  - $p(a \leq x \leq b) = p(a < x \leq b) = p(a \leq x < b) = p(a < x < b)$
  - Cumulative distribution function $cdf(b) = \int_{-\infty}^{b} f(x)dx$
- Expectation of a rand. var. $\mathbb{E}(x) = \int_{-\infty}^{\infty} x f(x)dx$
  - $\mathbb{E}(g(x)) = \int_{-\infty}^{\infty} g(x)f(x)dx$
- Variance $\sigma^2 = \mathbb{E}[(x - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx = \mathbb{E}(x^2) - \mu^2$

Standard Continuous Prob. Distr. Lec 5/6

- Uniform distribution
- Exponential distribution
- Gamma distribution
  - Inverse Gamma
  - Chi-squared
  - Inverse Chi-squared
- Normal/Gaussian distribution
- Beta distribution
- Weibull distribution

Questions:

- What scenarios are these distributions suited for?
- What is the domain?
- What shapes can these distributions exhibit?
  - How are they influenced by the parameters?
- What is the prob. that $a \leq x \leq b$ or $x \leq a$ or $x \geq a$?

Multivariate rand. vars. Lec 6

- Univariate vs. Multivariate rand. vars.
- Joint probability
  - Discrete $p(x = a, y = b)$
  - Continuous $p(a \leq x \leq b, c \leq y \leq d) = \int_c^d \int_a^b f(x, y)dxdy$
- Cumulative distribution function
  - Discrete $cdf(x, y) = p(x \leq a, y \leq b)$
  - Continuous $cdf(x, y) = \sum_{x=-\infty}^{a} \sum_{y=-\infty}^{b} p(x, y)$
- Marginal probability
  - $f(x) = \sum_y f(x, y) = \int_{-\infty}^{\infty} f(x, y)dy$
- Conditional probability

$$f(x|y) = \begin{cases} \frac{f(x,y)}{f(y)}, & \text{for } f(y) > 0 \\ 0, & \text{elsewhere} \end{cases}$$

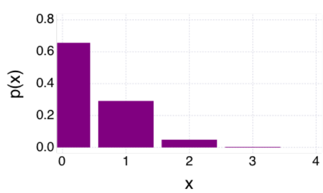
Multivariate rand. vars. Lec 6

- Independent random variables
  - Discrete: for all values of x and y, $p(x, y) = p(x)p(y)$
  - Continuous: Functional form of $f(x, y) = f(x)f(y)$
- Expectation:
  - Discrete $\mathbb{E}[g(x, y)] = \sum_x \sum_y g(x, y)p(x, y)$
  - Continuous $\mathbb{E}[g(x, y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)f(x, y)dxdy$
  - When x and y are indep., $\mathbb{E}(xy) = \mathbb{E}(x)\mathbb{E}(y)$
- Covariance is a property of the joint probability distribution
- Covariance captures joint variability of two random variables
  - $cov(x, y) = \mathbb{E}[(x - \mu_x)(y - \mu_y)]$
  - where $\mu_x = \mathbb{E}(x)$ and $\mu_y = \mathbb{E}(y)$
  - $cov(x, y) = \mathbb{E}(xy) - \mu_x\mu_y$
  - When x and y are indep., $cov(x, y) = 0$, as $\mathbb{E}(xy) = \mathbb{E}(x)\mathbb{E}(y)$

Probabilistic Inference vs. Parameter Estimation Lec 7

- Probabilistic Inference involves computation of probabilities for events, given a model family and choices for the parameters
- Parameter Estimation involves estimation of parameters given a parametric model and observed data drawn from it

Problem: 10% of a large lot of apples are damaged. If four apples are randomly sampled from the lot, find the probability that at least one apple in the sample of four is defective. $p(x \geq 1)$?



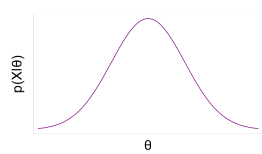
Problem: 20 apples were inspected and 3 apples were found to be damaged. What is the value of the parameter θ for the Binomial distribution?

Approaches for parameter estimation Lec 7

Maximum Likelihood Estimation (MLE)

- Parameters are assumed to be fixed but unknown
- ML solution seeks the solution that best explains the dataset X

$$\hat{\theta}_{MLE} = argmax_{\theta} p(X|\theta)$$

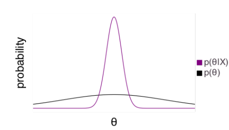


Bayesian Parameter Estimation

- Parameters are assumed to be random variables
- Prior knowledge on θ: $p(\theta)$
- Bayesian methods estimate the posterior density $p(\theta|X)$

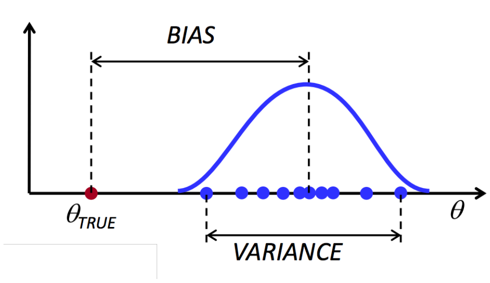
$$p(\theta|X) \propto p(X|\theta)p(\theta)$$

$$\hat{\theta}_{MAP} = argmax_{\theta} p(\theta|X)$$



Properties of Estimators Lec 7

- **Consistency**: Does the estimator converge to true value when the number of samples goes to infinity
- **Bias**: How close is the estimate to the true value (on average)?
- **Variance**: How much does it change for different datasets?



Module 2: Maximum Likelihood Estimation

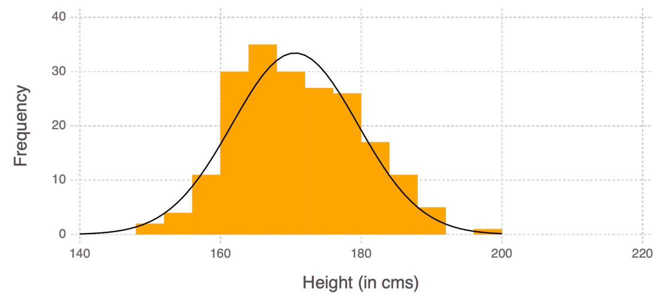
Module 2: Maximum Likelihood Estimation

Topics

- General approach to MLE
  - I.I.D
  - Likelihood $\mathcal{L}(\theta|x)$, Log-Likelihood ℓ, Maximizing ℓ
  - Optimization algos: Gradient Descent/Newton Method
- Univariate Parameter Est. using MLE
- Multivariate Parameter Est. using MLE
- Logistic Regression
  - Max. Conditional Likelihood
- Latent variables
  - Mixture Models: Discrete latent vars.
  - Factor Models: Continuous latent vars.
- Expectation-Maximization
  - General Approach
  - Proof of correctness

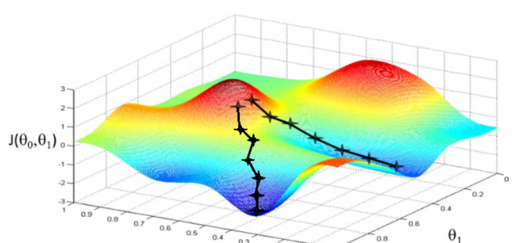
Parameter Estimation using MLE Lec 7

- Fitting Univariate distributions $p(x)$
  - E.g., Height of 200 subjects
  - $$p(x|\mu,\sigma^2) = \mathcal{N}(x|\mu,\sigma^2) \equiv \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/2\sigma^2}$$



Maximum Likelihood Estimation Lec 7

- I.I.D assumption
- Likelihood
$$p(x_1,\ldots,x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) = L(\theta|x)$$
- Log-likelihood $\ell(\theta) = \log L(\theta|x)$
- Maximization of ℓ
  - Alternatively Minimization of $-\ell(\theta)$ using a Gradient descent approach



Gradient Descent: a general algorithm Lec 8

Step 1: Pick initial value $\boldsymbol{w}_1$

Step 2: $maxIter = 10000$

Step 3: **for** $i = 2 : maxIter$

Step 4: $\quad \boldsymbol{w}_i \leftarrow \boldsymbol{w}_{i-1} - \lambda \nabla E|_{\boldsymbol{w}_{i-1}}$

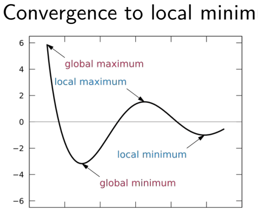
Step 5: $\quad$ **if** $|\boldsymbol{w}_i - \boldsymbol{w}_{i-1}| < \epsilon$ terminate; **end**

Step 6: **end for**

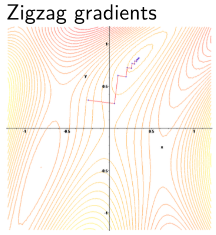
Gradient Descent: limitation

- Can converge to a local minimum
  - can result in a different value in different runs
- Tends to be slow when it is close to the minimum
- In poorly conditioned convex problems, 'zigzags' when gradients point nearly orthogonally to the shortest direction

Convergence to local minimum



Zigzag gradients



MLE for Gamma distribution Lec 8

Probability density function of Gamma distribution is
$$\frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$$
where $\Gamma(\alpha)$ is the gamma function and (α,β) are parameters that take positive values.

Likelihood function
$$L(\theta|x) = \frac{1}{\Gamma(\alpha)^n \beta^{n\alpha}} (\prod_i x_i^{\alpha-1}) e^{-\sum_i x_i/\beta}$$

Log-Likelihood function
$$\ell(\theta) = -n\log\Gamma(\alpha) - n\alpha\log\beta + (\alpha-1)\sum_i \log x_i - \frac{\sum_i x_i}{\beta}$$

Negative Log-Likelihood function
$$-\ell(\theta) = n\log\Gamma(\alpha) + n\alpha\log\beta - (\alpha-1)\sum_i \log x_i + \frac{\sum_i x_i}{\beta}$$

MLE for Gamma distribution Lec 8

Negative Log-Likelihood function

$$-\ell(\theta) = n \log \Gamma(\alpha) + n\alpha \log \beta - (\alpha - 1) \sum_i \log x_i + \frac{\sum_i x_i}{\beta}$$

Computing partial derivatives:

$$\frac{\partial \ell}{\partial \alpha} = n \frac{\partial}{\partial \alpha} \log \Gamma(\alpha) + n \log \beta - \sum_i \log x_i$$

$$\frac{\partial \ell}{\partial \beta} = n \frac{\alpha}{\beta} - \frac{\sum_i x_i}{\beta^2}$$

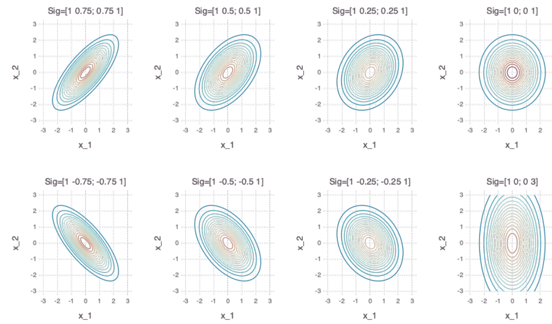
Gradient Descent update rules:

$$\alpha \leftarrow \alpha - \gamma \frac{\partial \ell}{\partial \alpha} \qquad \beta \leftarrow \beta - \gamma \frac{\partial \ell}{\partial \beta}$$

where γ is the learning rate.

Multivariate Gaussian Lec 8

- Geometric interpretatio of the covariance matrix



- Properties
  - Product of Gaussians is a Gaussian
  - Linear transformation of a Gaussian is a Gaussian
  - Partitioned Gaussian

Learning a MV Gaussian using Maximum Likelihood Lec 9

- **Scenario:** Height (in cm.) and weight (in kg.) of 200 individuals are collected. Assuming they follow a MV Gaussian distribution, estimate the parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ the MV Gaussian.

Row	Weight	Height
1	77.4	182.6
2	58.5	161.3
3	63.1	161.2
4	68.6	177.7
5	59.3	157.8
6	76.7	170.4

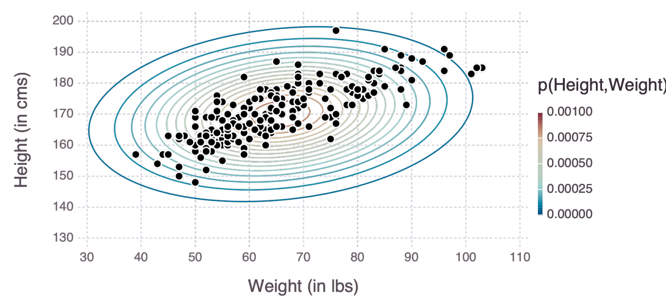
$$\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \equiv \sum_{i=1}^{n} \log p(\boldsymbol{x}_i | \boldsymbol{\mu} \boldsymbol{\Sigma}) = -\frac{1}{2} \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) - \frac{n}{2} \log det(2\pi\boldsymbol{\Sigma})$$

- Compute $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ using Gradient-descent

Learning a MV Gaussian using Maximum Likelihood Lec 9

- Fitting Multivariate distributions $p(\boldsymbol{x})$ or $p([x_1, x_2, \dots, x_d])$
  - E.g., Height and Weight of 200 subjects

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \equiv \frac{1}{\sqrt{det(2\pi\boldsymbol{\Sigma})}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}$$



Logistic Regression: Example Lec 9

- Widely used to model outcome of a categorical *dependent* variable, given the state of continuous *independent* variables
- Petal length of flowers from two different plant species are collected.

Row	PetalLength	Species
1	1.6	setosa
2	1.4	setosa
3	1.3	setosa
4	5.2	virginica
5	5.0	virginica
6	5.2	virginica

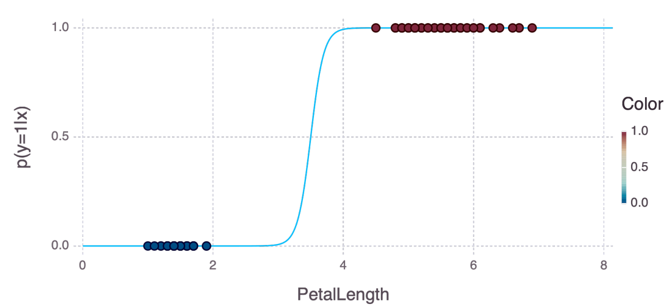
- Dependent variable
  - Species
- Independent variable
  - PetalLength

- Determine the probabilities:

$$p(y = setosa | x = 1.5) = ? \qquad p(y = virginica | x = 1.5) = ?$$

Parameter Estimation for LR (using MLE) Lec 9

- Fitting $p(y|x)$ or $p(y|\boldsymbol{x})$
  - E.g., Predicting species from petal length. $p(y = 1|x) = \dfrac{1}{1 + e^{\beta_0 + \beta_1 x}}$
  - p(species = virginica|PetalLength = 6)



Conditional Likelihood & Maximizing it Lec 9

$$f(y, x|\theta) = f(y|x, \theta) \times f(x|theta)$$
$$Joint = Conditional \times Marginal$$

Conditional Likelihood

Conditional Likelihood of θ given data x and y is

$$L(\theta; y|x) = p(y|x) = f(y|x; \theta)$$

Principle of maximum conditional likelihood

Given data consisting of pairs $\{(x_i, y_i) : i = 1, 2, \ldots n\}$, choose a parameter estimate $\hat{\theta}$ that maximizes the joint conditional likelihood expressed as the product

$$\prod_i f(y_i|x_i; \theta)$$

- suffices to assume y_i are independent (x_is need not be indep.)

Newton's method: a general algorithm Lec 9

Step 1: Pick initial value $\boldsymbol{w}_1$

Step 2: $maxIter = 10000$

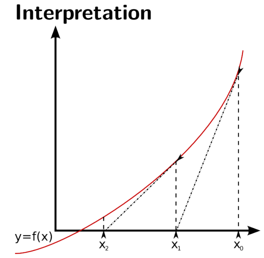
Step 3: **for** $i = 2 : maxIter$

Step 4: $\boldsymbol{w}_i \leftarrow \boldsymbol{w}_{i-1} - \frac{\nabla E(\boldsymbol{w}_{i-1})}{\nabla^2 E(\boldsymbol{w}_{i-1})}$

Step 5: **if** $|\ell_i - \ell_{i-1}| < \epsilon$ terminate; **end**

Step 6: **end for**

Interpretation



Newton's method: Advantages and Disadvantages Lec 9

Advantages

- Converges quadratically towards a stationary point.

Comparision with Gradient Descent:

$$\lambda = \frac{1}{\nabla^2 E(\boldsymbol{w}_{i-1})}$$

Disadvantages

- Does not necessarily coverge toward a minimizer
- Diverges if the starting approximation is too far
- Requires second-rder information $\nabla^2 E(\boldsymbol{w}_{i-1})$
- Not suited if $\nabla^2 E(\boldsymbol{w}_{i-1})$ is not invertible

Latent or Hidden variables Lec 10

Latent Variables

Random variables whose values are not specified in the observed data.

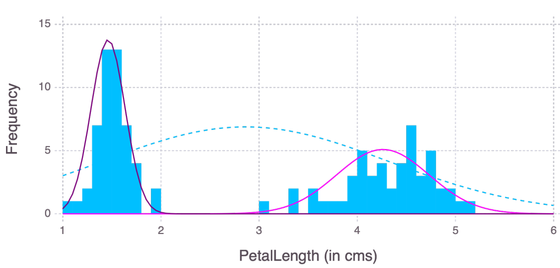
- E.g., An online survey is sent out to employees at a University to collect their height and weight. Gender is a latent variable that is not measured.

Row	Weight	Height	Gender
1	77.4	182.6	M
2	58.5	161.3	F
3	63.1	161.2	F
4	68.6	177.7	M

Observed var.	Latent Variable Continuous	Latent Variable Discrete
Continuous	Factor Analysis	Mixture Modeling
Discrete	Latent Trait Analysis	Latent Class Analysis

Mixture Models Lec 10

- Data is modelled as a mixture of several components
  - Each component has a simple parametric form (such as a Gaussian)



- Mixture Model is not 'aware' of the underlying interpretation

Mixture Models - formally Lec 10

Mixture Models

A distribution f is a **mixture** of k component distributions $f_1, f_2, \ldots, f_k$ if

$$f(x) = \sum_{i=1}^{k} \pi_i f_i(x)$$

where π_i are the **mixing weights**, $\pi_i > 0, \sum_i \pi_i = 1$

- In principle, f_is can be arbitrary distributions
- In practice, we prefer **parametric mixture models**
  - All distributions belong to the same parametric family, with different parameters
- Gaussian mixture model is a popular mixture model

Motivation for Expectation Maximization (EM) Lec 10

- To estimate parameters, maximize ℓ for Mixture Models
- To compute posterior prob. $p(M|x_i)$, we need μ_M and μ_F

$$p(M|x_i) = \frac{\pi_M \mathcal{N}(x_i; \mu_M, \sigma^2)}{\pi_M \mathcal{N}(x_i; \mu_M, \sigma^2) + \pi_F \mathcal{N}(x_i; \mu_F, \sigma^2)}$$

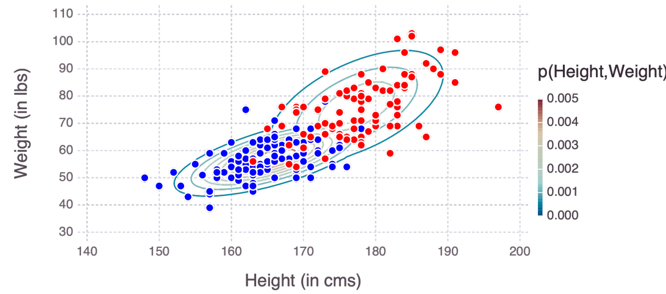
- To compute μ_M and μ_F, we need $p(M|x_i)$ and $p(F|x_i)$

$$\mu_M = \frac{\sum_{i=1}^n p(M|x_i) x_i}{\sum_{i=1}^n p(M|x_i)} \qquad \mu_F = \frac{\sum_{i=1}^n p(F|x_i) x_i}{\sum_{i=1}^n p(F|x_i)}$$

- Strategy: We will fix one and solve for the other, iteratively.
- **EM Algorithm**
  - **E Step:**, we fix parameters μ_M and μ_F, and compute the posterior distribution $p(M|x_i)$ and $p(F|x_i)$
  - **M Step:**, we fix posteriori distribution $p(M|x_i)$ and $p(F|x_i)$ and optimize for μ_M and μ_F
  - Repeat the two steps until the values converge

Mixture of Bivariate Gaussians Lec 11

- Height and Weight of 200 subjects



A general EM algorithm Lec 11

- Given a joint distribution $p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$
- **Step 1:** Choose an initial setting for parameters $\boldsymbol{\theta}^{old}$.
- **Step 2:** E Step: Evaluate $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old})$.
- **Step 3:** M Step: Evaluate $\boldsymbol{\theta}^{new}$ given by

$$\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$$

where

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{\boldsymbol{z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old}) \log p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$$

- **Step 4:** Check for convergence of either the log-likelihood or the parameter values. If convergence criteria is not met, then

$$\theta^{old} \leftarrow \theta^{new}$$

and return to step2.

Correctness of EM algorithm Lec 11

- Our goal is to maximize

$$p(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{\boldsymbol{z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$$

- We introduce a distribution $q(\boldsymbol{Z})$ defined over the latent variables
- **Claim:** For any choice of $q(\boldsymbol{Z})$, the following decomposition holds

$$\log p(\boldsymbol{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q||p)$$

where we define

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\boldsymbol{z}} q(\boldsymbol{Z}) \log \left\{ \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \right\}$$

$$KL(q||p) = -\sum_{\boldsymbol{z}} q(\boldsymbol{Z}) \log \left\{ \frac{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})}{q(\boldsymbol{Z})} \right\}$$

Note that $\mathcal{L}(q, \boldsymbol{\theta})$ is a functional of the distribution $q(\boldsymbol{Z})$, and a function of parameters $\boldsymbol{\theta}$.

Verify the claim using $\log p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) = \log p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}) + \log p(\boldsymbol{X}|\boldsymbol{\theta})$

Correctness of EM algorithm Lec 11

- For any choice of $q(\boldsymbol{Z})$, the following decomposition holds

$$\log p(\boldsymbol{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q||p)$$

where we define

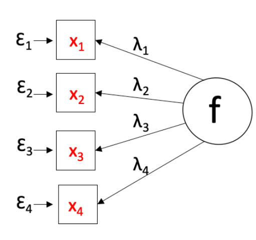
$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\boldsymbol{z}} q(\boldsymbol{Z}) \log \left\{ \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \right\}$$

$$KL(q||p) = -\sum_{\boldsymbol{z}} q(\boldsymbol{Z}) \log \left\{ \frac{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})}{q(\boldsymbol{Z})} \right\}$$

- As KL$(p||q) \geq 0$, $\mathcal{L}(q, \boldsymbol{\theta}) \leq \log p(\boldsymbol{X}|\boldsymbol{\theta})$.
  - $\mathcal{L}(q, \boldsymbol{\theta})$ is the lower bound on $\log p(\boldsymbol{X}|\boldsymbol{\theta})$.
- In E-Step: Lowed bound $\mathcal{L}(q, \boldsymbol{\theta})$ is maximized w.r.t. $q(\boldsymbol{Z})$, fixing $\boldsymbol{\theta}^{old}$
- In M-Step: $\mathcal{L}(q, \boldsymbol{\theta})$ is maximized w.r.t. $\boldsymbol{\theta}$ to give some new value $\boldsymbol{\theta}^{new}$

Factor analysis model Lec 12

$$x_i = \lambda_i f + \epsilon_i$$



- x_i are the observed variables
  - e.g., x_1, x_2, and x_3 are exam scores obtained by a student in math, English and history.
- f is the underlying commmon factor
  - e.g., student's intelligence
- λ_i are the factor loadings
  - e.g., how much is the contribution of intelligence to exam score
- ϵ_i are unique factors or residuals or random noise terms
  - e.g., how much result differs from student's general ability
- Multiple factors

$$x_i = \lambda_{i1} f_1 + \lambda_{i2} f_2 + \ldots + \lambda_{ik} f_k + \epsilon_i$$

Factor analysis model — Lec 12

Formulation

$$\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$$
$$\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi})$$
$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda f} + \boldsymbol{\epsilon}$$

Parameters of this model are:

- Vector $\boldsymbol{\mu} \in \mathbb{R}^d$
- Matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{d \times k}$
  - usually $k < d$
- Diagonal matrix $\boldsymbol{\Psi} \in \mathbb{R}^{d \times d}$

- Geometric interpretation
- Identifiability problem
- Joint distribution
- Max. Likelihood Estimation
  - EM approach

Module 3: Bayesian Parameter Estimation

Module 3: Bayesian Parameter Estimation

Topics

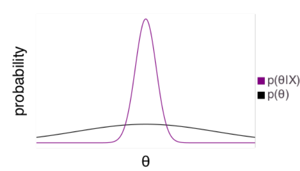
- General approach to Bayesian estimation
  - Prior, Likelihood, Posterior
  - Why/Why not Bayesian estimation?
- Priors
  - Noninformative
  - Conjugate Priors
  - Natural Conjugacy
  - Mixture of Priors
  - Jeffrey's Prior
- Posterior
  - Univariate
  - Multivariate: Nuisance Parameter, Marginal Posterior
- Summarization of Posterior
  - Point Estimation (Bayes' Risk)
  - Interval Estimation

Bayesian Parameter Estimation — Lec 13

Bayesian Estimation

- Parameters are assumed to be random variables with some known a priori distribution $p(\theta)$
- Prior distribution is either a belief or prior knowledge
- Bayesian methods seek to estimate the posterior density $p(\theta|y)$

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}$$



Terminology	Notation
Posterior	$p(\theta\|y)$
Prior	$p(\theta)$
Model	$p(y\|\theta)$
Prior predictive distribution (marginal likelihood)	$p(y)$

Bayesian estimation: Why and Why not? — Lec 13

Why do a Bayesian analysis?

- Incorporate prior belief or existing knowledge via $p(\theta)$
- Coherent with rules of probability, i.e. everything follows from specifying $p(\theta|y)$
- Captures uncertainty in the parameter estimates
- Interpretability of results, e.g. the probability the parameter is in (L, U) is 95%

Why not do a Bayesian analysis?

- Need to specify $p(\theta)$
- Computational cost of evaluating the likelihood function
- Does not guarantee coverage

Bayesian estimation: update posterior — Lec 13

- Bayes' Rule provides a formula for updating from prior beliefs to our posterior beliefs based on the data we observe, i.e.

$$p(\theta|y) = \frac{p(y|\theta)}{p(y)} p(\theta) \propto p(y|\theta)p(\theta)$$

- Suppose we gather $y_1, \ldots, y_n$ sequentially (and we assume y_i independent conditional on θ), then we have

$$\begin{aligned} p(\theta|y_1) &\propto p(y_1|\theta)p(\theta) \\ p(\theta|y_1, y_2) &\propto p(y_1, y_2|\theta)p(\theta) \\ p(\theta|y_1, y_2) &\propto p(y_2|\theta)p(y_1|\theta)p(\theta) \\ p(\theta|y_1, y_2) &\propto p(y_2|\theta)p(\theta|y_1) \end{aligned}$$

and

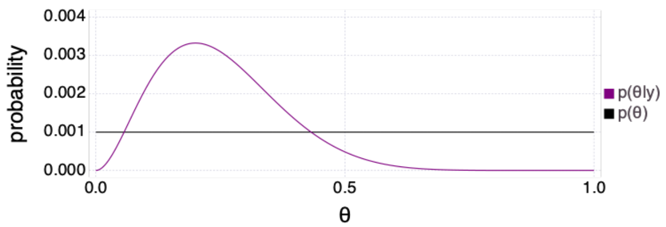
$$p(\theta|y_1, \ldots, y_i) \propto p(y_i|\theta)p(\theta|y_1, \ldots, y_{i-1})$$

So Bayesian learning is

$$p(\theta) \rightarrow p(\theta|y_1) \rightarrow p(\theta|y_1, y_2) \rightarrow \cdots \rightarrow p(\theta|y_1, \ldots, y_n).$$

Coin posterior - default prior — Lec 13

- From an experiment we have $N_H = 2$ and $N_T = 8$
- Prior distribution is $p(\theta) = 1$
- Likelihood is $\theta^{N_H}(1-\theta)^{N_T}$
- Posterior $p(\theta|y_1, \ldots, y_n) = \frac{1}{c}\theta^{N_H}(1-\theta)^{N_T} = Beta(N_H+1, N_T+1)\}$.
- We can compute the probabilities $p(\theta|y_1, \ldots, y_n)$ directly from the pdf $Beta(N_H+1, N_T+1)$



Choosing Prior — Lec 14

- How do we construct/choose prior distributions?
- Two interpretations:
  - *Population* interpretation
    - Prior distribution represents a population of possible parameter values from which θ has been drawn
  - *Knowledge* interpretation
    - We must express our knowledge about θ as if its value could be thought of as a random realization from the prior distribution.
- In many applications there is no perfectly relevant population of θ's from which the current θ has been drawn.

General guidelines:

- Prior distribution should include all possible values of θ
- Prior need not be realistically concentrated around the 'true' value.

Information about θ contained in the data will far outweigh any reasonable prior specification.

Informative Prior — Lec 14

- Let us consider using a *Beta* prior $\quad \theta \sim Beta(\alpha, \beta)$

$$Beta(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

Interpretation of *information* in the prior

- Compare this prior to the previous posterior under uniform prior
- $Beta(a, b)$ is equivalent to $a-1$ priori successes and $b-1$ prior failures.

Hyperparameters

- Parameters of the priori distribution are referred to as hyperparameters
  - These are assumed to be known
- Beta prior is indexed by two hyperparameters (a, b)
- We are essentially fixing two features of the dist. (e.g., mean and variance)

Conjugacy — Lec 14

If the posterior is of the same parametric form as the prior, then we call the prior the conjugate distribution for the likelihood distribution.

Discrete distributions

Sample Space	Sampling Dist.	Conjugate Prior	Posterior
$y \in \{0,1\}$	*Bernoulli*	*Beta*	*Beta*
$y = \mathbb{Z}_+$	*Poisson*	*Gamma*	*Gamma*
$y = \mathbb{Z}_{++}$	*Geometric*	*Gamma*	*Gamma*
$y = \mathbb{H}_K$	*Multinomial*	*Dirichlet*	*Dirichlet*

Continuous distributions

Sampling Dist.	Conjugate Prior	Posterior
$Exponential(\theta)$	$Gamma(\alpha,\beta)$	*Gamma*
$\mathcal{N}(\mu,\sigma^2)$, known σ^2	$\mathcal{N}(\mu_0,\sigma_0^2)$	*Gaussian*
$\mathcal{N}(\mu,\sigma^2)$, known μ	*InvGamma*	*InvGamma*
$\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, known $\boldsymbol{\Sigma}$	$\mathcal{N}(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0^2)$	*Gaussian*
$\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, known $\boldsymbol{\mu}$	*InvWishart*	*InvWishart*

Natural conjugate prior — Lec 15

Natural conjugate

A natural conjugate prior is a conjugate prior that has the same functional form as the likelihood.

- For example, the beta distribution is a natural conjugate prior since

$$p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1} \qquad \text{and} \qquad L(\theta) \propto \theta^{y}(1-\theta)^{n-y}.$$

- Probability distributions that belong to an exponential family have natural conjugate prior distributions.
  - This is the only class of distributions that have natural conjugate prior distributions

Exponential Family — Lec 15

- A random variable y has a distribution from an exponential family model $\mathcal{F}$ if the density of y is of the form

$$p(y|\boldsymbol{\theta}) = h(y)exp\Big(\boldsymbol{\eta}(\boldsymbol{\theta})^T\boldsymbol{T}(y) - \psi(\boldsymbol{\theta})\Big)$$

- Exponential family contains many standard distributions

Discrete	Continuous
Bernoulli	Beta
Categorical	Chi-squared
Geometric	Exponential
Poisson	Gamma
	Gaussian

Estimating parameters of a Gaussian (only unknown is μ)

- Given a training data $y = \{y_1, \ldots, y_n\}$ drawn *i.i.d* from a Gaussian $\mathcal{N}(y|\mu, \sigma^2)$ with unknown mean μ and a given variance σ^2

$$\mathcal{N}(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\Big(-\frac{1}{2\sigma^2}(y-\mu)^2\Big)$$

- Choosing a Gaussian prior over μ

$$p(\mu) = (2\pi\sigma_0^2)^{-n/2} \exp\Big[\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\Big]$$

- Our posterior over parameter μ

$$\mathcal{N}(\mu|\mu_p, \sigma_p^2) = \frac{1}{\sqrt{2\pi\sigma_p^2}} exp\Big(-\frac{1}{2\sigma_p^2}(\mu-\mu_p)^2\Big)$$

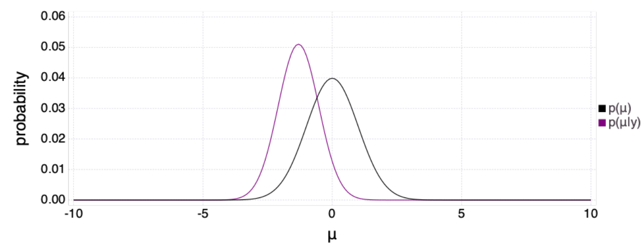
where posterior parameters are estimated by completing the square

$$\mu_p = \sigma_p^2\Big(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_i y_i}{\sigma^2}\Big); \qquad \sigma_p^2 = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}$$

Bayesian Estimation: Single-Parameter models Lec 15

Scenario: The temperatures, in Celsius, in Minneapolis during the first week of March 2018 are observed as $\{-2.5, -9.9, -12.1, -8.9, -6.0, -4.8, 2.4\}$

- Goal is to estimate μ, assuming σ^2 is known.
- Natural Conjugate Gaussian Prior $p(\mu) = \mathcal{N}(0, 1)$
- Posterior is also Gaussian $p(\mu|y) = \mathcal{N}(\mu_p, \sigma_p^2)$



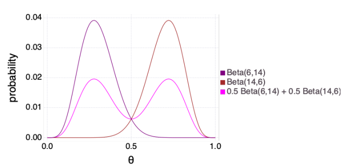
Mixture of Priors Lec 15

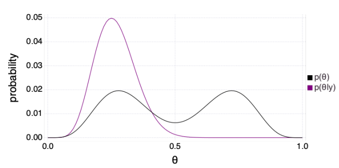
- A mixture of two prior beliefs can be used as prior density

$$p(\theta) = \pi p_1(\theta) + (1-\pi)p_2(\theta))$$

  - where $p_1(\theta) = Beta(6, 14)$ and $p_2(\theta) = Beta(14, 6)$
  - mixing probability is 0.5.
- Posterior: Mixture of priors is also a conjugate
  - Posterior is also a mixture (with updated weights)

$$p(\theta|y) = \sum_{i=1}^{k} \frac{\pi_i p_i(y)}{\sum_{j=1}^{k} \pi_j p_j(y)} p_i(\theta|y)$$





Fisher Information & Jeffreys' Prior Lec 16

- Sufficient Statistic
  - There in no information about θ left in data y, after observing summary statistic s
  - y is conditionally independent of θ, given s
- Fisher Information

$$\mathcal{I}_y(\theta) = -\mathbb{E}_{y|\theta}\Big[\frac{\partial^2}{\partial\theta^2} \log p(y|\theta)\Big]$$

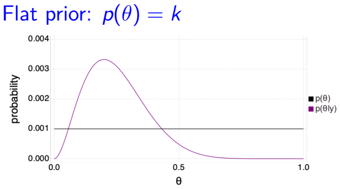
- Issue with noninformative prior
  - Posterior varies with transformations
- Jeffreys Prior
  - $p(\theta) \propto \sqrt{\mathcal{I}_y(\theta)}$
  - Posterior is invariant under transformations

Bayesian Estimation: Single-Parameter models

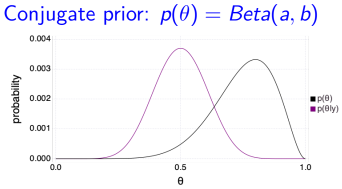
Scenario: Coin toss experiment (where 2 heads and 8 tails are observed)

- Goal is to estimate θ

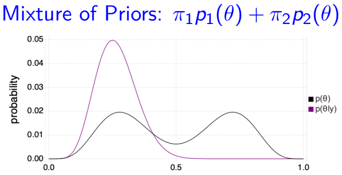
Flat prior: $p(\theta) = k$



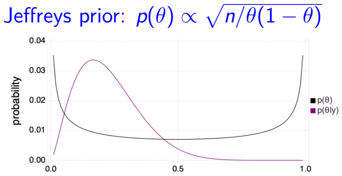
Conjugate prior: $p(\theta) = Beta(a, b)$



Mixture of Priors: $\pi_1 p_1(\theta) + \pi_2 p_2(\theta)$



Jeffreys prior: $p(\theta) \propto \sqrt{n/\theta(1-\theta)}$



Summarizing the posterior Lec 13

- Posterior distribution contains all the *current* info. about the parameter θ
- Ideally one may report the entire probability distribution $p(\theta|y)$
  - A graphical display is useful
- Bayesian estimation provides flexibility of summarizing posterior
- Two ways:
  - Point Estimate: most likely guess
    - mean
    - median
    - mode
  - Interval Estimate
    - Equal-tailed
    - One-sided
    - Highest posterior density

Bayes Risk — Lec 13

The Bayes Risk of an estimate $\hat{\theta}$ can be assessed by how much we believe we missed the true θ.

More formally, Bayes Risk is computed as the expectation of the loss function $L(\theta, \hat{\theta})$ over the posterior $p(\theta|y)$.

$$Risk = \int L(\theta, \hat{\theta})\ p(\theta|y)$$

Common estimators:

- Mean: $\hat{\theta}_{Bayes} = E[\theta|y]$ minimizes $L\Big(\theta, \hat{\theta}\Big) = \Big(\theta - \hat{\theta}\Big)^2$
- Median: $\int_{\hat{\theta}_{Bayes}}^{\infty} p(\theta|y)d\theta = \frac{1}{2}$ minimizes $L\Big(\theta, \hat{\theta}\Big) = \Big|\theta - \hat{\theta}\Big|$
- Mode: $\hat{\theta}_{Bayes} = \text{argmax}_{\theta}\ p(\theta|y)$ is obtained by minimizing $L\Big(\theta, \hat{\theta}\Big) = -\mathbb{I}\left(|\theta - \hat{\theta}| < \epsilon\right)$ as $\epsilon \to 0$, also called maximum a posterior (MAP) estimator.

Interval estimation — Lec 13

Definition

A $100(1-a)\%$ credible interval is any interval (L,U) such that

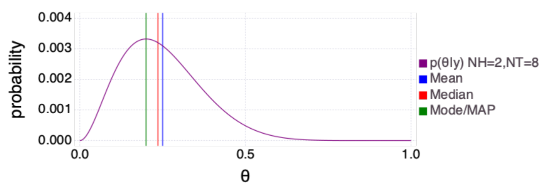
$$1 - a = \int_L^U p(\theta|y)d\theta.$$

Some typical intervals are

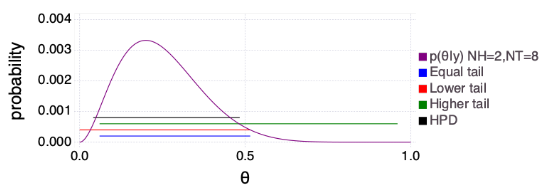
- Equal-tailed: $a/2 = \int_{-\infty}^{L} p(\theta|y)d\theta = \int_{U}^{\infty} p(\theta|y)d\theta$
- One-sided: either $L = -\infty$ or $U = \infty$
- Highest posterior density (HPD): $p(L|y) = p(U|y)$ for a uni-modal posterior which is also the shortest interval
  - one with the smallest interval width among all credible intervals

Point and Interval estimation — Lec 13

Point Estimation



Interval Estimation



Multiparameter Models — Lec 17

- Multiparameter models

$$p(y|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\Big) \qquad p(\mu, \sigma^2) \propto 1/\sigma^2$$

- Joint posterior density

$$p(\mu, \sigma^2|y) = (\sigma^2)^{-(n+2)/2} \exp\left(-\frac{1}{2\sigma^2}\left[(n-1)s^2 + n(\bar{y} - \mu)^2\right]\right)$$

$$\text{where} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \text{ is the sample variance}$$

- Nuisance parameters
- Conditional posterior density
- Marginal posterior density
  - To determine the marginal posterior for μ, we need to do marginalization

$$p(\mu|y) = \int p(\mu, \sigma^2|y)d\sigma^2 = \int p(\mu|\sigma^2, y)p(\sigma^2|y)d\sigma^2$$

Sampling Algorithm — Lec 17

$$p(\mu|y) = \int p(\mu, \sigma^2|y)d\sigma^2 = \int p(\mu|\sigma^2, y)p(\sigma^2|y)d\sigma^2$$

$$p(\mu|\sigma^2, y) = \mathcal{N}(\mu|\mu_p, \sigma_p^2); \qquad \mu_p = \sigma_p^2\Big(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_i y_i}{\sigma^2}\Big); \qquad \sigma_p^2 = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}$$

$$p(\sigma^2|y) = InvGamma(\alpha, \beta); \qquad \alpha = \frac{n-1}{2} \qquad \beta = \frac{2}{s^2(n-1)}$$

Algorithm:

- *Step 1*: Sampling σ^2
  - Compute α and β
  - Sample 1000 values of $\sigma^2|y \sim InvGamma(\alpha, \beta)$
- *Step 2*: Sampling μ
  - Assume a prior μ_0, σ_0^2
  - repeat for sample of $\sigma^2|y$
    - Compute posterior parameters μ_p, σ_p^2
    - Sample a value of μ from $\mu|\sigma^2, y \sim \mathcal{N}(\mu|\mu_p, \sigma_p^2)$

Bayesian Estimation for a MvGaussian — Lec 17

Assume we know $\mathbf{\Sigma}$, and we want to estimate $\boldsymbol{\mu} = (\mu_W, \mu_H)$,

- We first begin with a prior $p(\boldsymbol{\mu})$
  - preferably a natural conjugate prior $\mu \sim \mathcal{N}(\boldsymbol{\mu}_0, \mathbf{\Lambda}_0)$
- We write the likelihood

$$p(\mathbf{y}_1, \ldots, \mathbf{y}_k|\boldsymbol{\mu}, \mathbf{\Sigma}) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu})^{\top}\mathbf{\Sigma}^{-1}(y_i - \boldsymbol{\mu})\right)$$

- We derive the posterior (similar to the univariate Gaussian case)

$$p(\boldsymbol{\mu}|\mathbf{y}_1, \ldots, \mathbf{y}_k, , \mathbf{\Sigma}) = \mathcal{N}(\boldsymbol{\mu}_p, \mathbf{\Lambda}_p)$$

$$where\ \mathbf{\Lambda}_p^{-1} = \mathbf{\Lambda}_0^{-1} + n\mathbf{\Sigma}^{-1} \qquad \boldsymbol{\mu}_p = (\mathbf{\Lambda}_0^{-1} + n\mathbf{\Sigma}^{-1})^{-1}(\mathbf{\Lambda}_0^{-1}\boldsymbol{\mu}_o + n\mathbf{\Sigma}^{-1}\bar{y})$$

- Determining marginal posteriors is straight forward
  - $p(\mu_W|\mathbf{y}_1, \ldots, \mathbf{y}_k) \sim \mathcal{N}(\mu_1, \Sigma_{11})$ and $p(\mu_H|\mathbf{y}_1, \ldots, \mathbf{y}_k) \sim \mathcal{N}(\mu_2, \Sigma_{22})$

Module 4: Bayesian Computation

Module 4: Bayesian Computation

Topics

- Sampling from Posterior
  - Pseudo random number generator
  - Inverse-Transform Method
  - Accept-Reject Method
- Monte Carlo Integration
  - General Approach
  - Importance Sampling
- Markov Chain Monte Carlo Methods
  - Markov Chain: Stationarity and other properties
  - Metropolis-Hastings
    - General Approach
    - Random-walk Metropolis-Hastings
    - Independent Metropolis-Hastings
  - Gibbs Sampling
    - Application: Hierarchical Models

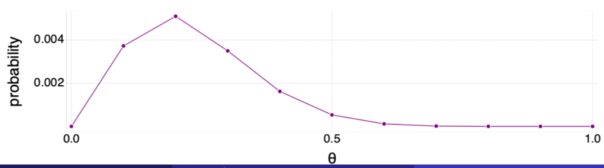
Point Estimation Lec 18

When the posterior has a standard functional form (due to conjugacy):

- we can compute a summary of the distribution analytically
  - mean of a $Beta(a, b)$ is $\frac{a}{a+b}$
- we can simulate data from the posterior and summarize
  - $\theta \sim Beta(a, b)$

When posterior does not have a standard form

- compute values of the posterior on a grid of points
- we can approximate the posterior by a discrete posterior
- High-dimensional posteriors: Computationally prohibitive



Random numbers Lec 18

- Uniform random variable is very important
  - many other random variables can be derived and transformed from it

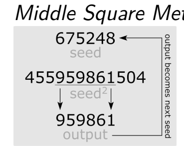
True random numbers:

- based on physical phenomenon (e.g. atmospheric noise, thermal noise, cosmic background radiation) that is known to be random
- very slow

Pseudo random numbers:

- Generated by computational algorithms
- these algorithms produce a long sequence of apparently random results
- they begin with a 'seed'
- the entire random sequence can be reproduced if 'seed' is known

von Neumann's
Middle Square Method



The inverse transform method Lec 18

For an arbitrary random variable x with density f and cdf F, define the generalized inverse of F by

$$F^{-1}(u) = \inf\{x; F(x) \geq u\}$$

If $u \sim \mathcal{U}(0, 1)$, then $F^{-1}(u)$ is distributed like x.

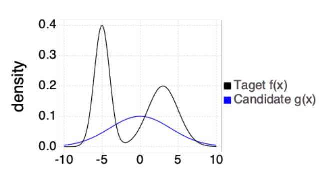
Using a uniform random number generator, we can draw samples from f

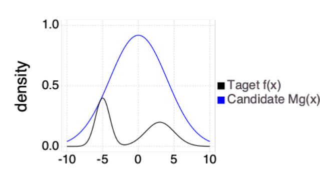
Example: Develop a procedure to draw samples for $x \sim Exp(\lambda = 1)$ with density $f(x) = \lambda e^{-\lambda x} = e^{-x}$, using a uniform random number generator?

- Approach
  1. *Determine cdf for a given density $f(x)$: $F(x) = \int_0^x e^{-t}dt = 1 - e^{-x}$*
  2. *Set $u = F(x)$: $u = 1 - e^{-x}$*
  3. *Solve for x: $x = F^{-1}(u) = -log(1 - u)$*
  4. *Draw $u \sim \mathcal{U}(0, 1)$, then compute $x = F^{-1}(u)$: $x = -\log(u)$*
- Continuous, Discrete, Mixture Representations

Accept-Reject Methods Lec 19

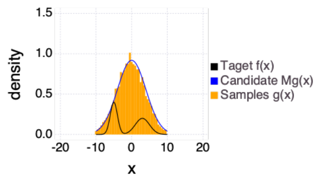
- These *Accept-Reject* methods require us to know the functional form of density f upto a multiplicative constant
  - f is known as *target density*
- We choose a simpler density g, called the *candidate density*
  - to generate random variables for which simulation is done
- Constraints:
  1. f and g have compatible supports (i.e., $g(x) > 0$, when $f(x) > 0$)
  2. There is constant M such that $f(x)/g(x) \leq M$ for all x
     - So, $Mg(x)$ *envelopes* $f(x)$

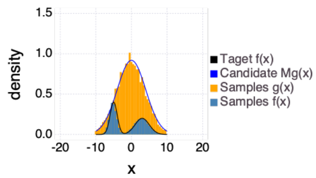




Accept-Reject Methods Lec 19

- Approach
  1. Generate $y \sim g$
  2. Independently generate $u \sim \mathcal{U}(0,1)$
  3. If $\quad u \leq \frac{1}{M}\frac{f(y)}{g(y)}$, then *accept* y as a sample
  4. else *reject* y, discard u, and start again with step 1.





Properties of Accept-Reject algorithm Lec 19

- It suffices to know $f(x)$ upto a multiplicative constant
  - The normalizing constant can be absorbed into M
  - $\frac{f(x)}{cg(x)} \leq M \implies \frac{f(x)}{g(x)} \leq M'$
- Efficiency of Accept-Reject algorithm can be measured in terms of its acceptance probability
  - $u \leq \frac{1}{M}\frac{f(y)}{g(y)}$
  - higher the acceptance probability, fewer wasted simulations from g
- If the bound $f(x) \leq Mg(x)$ is not tight (i.e., M is replaced by a larger constant)
  - the algorithm is still valid, but less efficient
- The probability of acceptance is $1/M$
  - M should be as small as possible for computational efficiency.

Integrals in Bayesian approaches Lec 20

Bayesian approaches require solving integrals in different scenarios:

1. Normalization (e.g., for determining the posterior distribution)
$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}$$
2. Marginalization (e.g., for averaging nuisance parameters)
$$p(\theta_1|y) = \int_{\theta_2 \dots \theta_k} p([\theta_1, \theta_2, \dots, \theta_k]|y)d\theta_2 \dots d\theta_k$$
3. Expectation (e.g., to obtain summary statistics of the posterior)
$$\mathbb{E}(f(\theta)) = \int f(\theta)p(\theta|y)d\theta$$

Challenges:

- Integrals in large dimensional spaces
  $p(\theta_1|y) = \int_{\theta_2 \dots \theta_k} p([\theta_1, \theta_2, \dots, \theta_k]|y)d\theta_2 \dots d\theta_k$
- Closed form solutions to integrals are not always possible

Monte Carlo Integration Lec 20

$$I(f) = \int_{\mathbf{x}^{min}}^{\mathbf{x}^{max}} f(\mathbf{x})d\mathbf{x} = \int_{\mathbf{x}^{min}}^{\mathbf{x}^{max}} g(\mathbf{x})p(\mathbf{x})d\mathbf{x} = \mathbb{E}_{p(\mathbf{x})}[g(\mathbf{x})] \approx \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i)$$

Steps:

1. Factorize $f(\mathbf{x}) = g(\mathbf{x})p(\mathbf{x})$
   - $p(\mathbf{x})$ can be interpreted as a probability density
     - $p(\mathbf{x}) \geq 0 \qquad \int p(\mathbf{x})d\mathbf{x} = 1$
2. Samples $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ are drawn i.i.d. from density $p(\mathbf{x})$
3. Compute $I(f) \approx \frac{1}{n}g(x_i)$

- Factorization of $f(\mathbf{x}) = g(\mathbf{x})p(\mathbf{x})$ is *key* for MC to work
  - We need to find $g(\mathbf{x})$ and $p(\mathbf{x})$ such that $I(f) = \mathbb{E}_{p(\mathbf{x})}[g(\mathbf{x})]$

Monte Carlo Integration Lec 20

$$I(f) = \int_{\mathbf{x}^{min}}^{\mathbf{x}^{max}} f(\mathbf{x})d\mathbf{x} \qquad \text{In MC integration } f(\mathbf{x}) = g(\mathbf{x})p(\mathbf{x})$$

Often $p(x)$ is chosen to be Uniform $\implies$ *ordinary* Monte Carlo Integration

Algorithm:

1. Initialize $x_1, \dots, x_n$ to 0s
2. **for** $i = 1, \dots, n$ times
3. Draw $x_i \sim U(0,5)$
4. **end**
5. Compute $S_n = \frac{1}{n}\sum_{i=1}^{n} \delta f(x_i)$
6. Return S_n

Monte Carlo methods: Convergence Lec 20

- *Strong Law of Large Numbers:* Let $x_1, x_2, \dots, x_n$ be i.i.d. with $\mathbb{E}[x_i] = \mu \in \mathbb{R}$, $Var(x_i) = \sigma^2 \in (0, \infty)$.
$$\text{If } \bar{x}_i = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ then } \bar{x}_i \to \mu$$
- LLN gives us the mean of the estimate S_n behavior when $n \to \infty$
- *Central Limit Theorem:*
  - Let $x_1, x_2, \dots, x_n$ be i.i.d. with $\mathbb{E}[x_i^2] < +\infty$.
  - Let σ^2 denote the variance of x_i, i.e., $\sigma^2 = E((x_i - E(x_i))^2)$ and
  - $\epsilon_n = \mathbb{E}(x) - \frac{1}{n}\sum_{i=1}^{n} x_i$.

  then $(\frac{\sqrt{n}}{\sigma}\epsilon_n)$ converges in distribution to $\mathcal{N}(0,1)$
- CLT gives us a distribution for error ϵ_n

Monte Carlo methods: Importance Sampling Lec 19

- Importance Sampling is a **MC Integration** approach
  - not a *sampling approach*
- The *idea* is to sample random numbers from a density that is close to the shape of the integrand.
  - Shape of $f(\mathbf{x})$ and $q(\mathbf{x})$ should look similar, $support(f) \subset support(q)$

$$I(f) = \int f(\mathbf{x})dx = \int \frac{f(\mathbf{x})}{q(\mathbf{x})} q(\mathbf{x})d\mathbf{x}$$

  - Choosing $q(\mathbf{x})$ requires some effort
    - $q(\mathbf{x})$ must be a probability density, i.e., $q(\mathbf{x}) \geq 0 \qquad \int p(\mathbf{x})d\mathbf{x} = 1$
- Using Monte Carlo integration on this 'factorization', we have Importance Sampling approach

Monte Carlo methods: Importance Sampling Lec 20

$$I(f) = \int f(\mathbf{x})dx = \int \frac{f(\mathbf{x})}{q(\mathbf{x})} q(\mathbf{x})dx$$

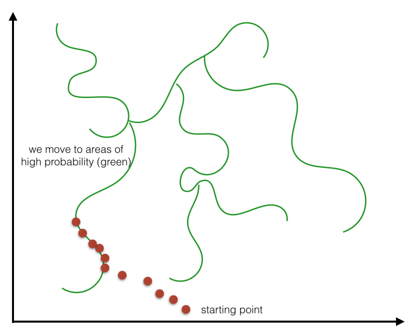
Importance Sampling Approach:

1. Initialize $\mathbf{x}_1, \ldots, \mathbf{x}_n$ to 0s
2. **for** $i = 1, \ldots, n$ times
3. Draw $\mathbf{x}_i \sim q(\mathbf{x})$
4. **end**
5. Compute $S_n = \frac{1}{n}\sum_{i=1}^{n} \frac{f(x_i)}{q(x_i)}$
6. Return S_n

- Importance Sampling
  - reduces variance of the estimate
  - by reducing the value of the term $Var[g(x)] = Var[\frac{f(x)}{q(x)}]$

Idea behind Markov Chain Monte Carlo Methods Lec 21

- Instead of sampling i.i.d., sample from a Markov Chain



- *Markov Chain*- where we go next depends on our current state
- *Monte Carlo* - Simulating data

Advantages/Disadvantes of MCMC Lec 21

Advantages:

- applicable even when we can't directly draw samples
- works for complicated distributions in high-dimensional spaces, even when we don't know where the regions of high probability are
- relatively easy to implement
- fairly reliable

Disadvantages:

- slower than simple Monte Carlo or importance sampling (i.e., requires more samples for the same level of accuracy)
- can be very difficult to assess accuracy and evaluate convergence, even empirically

Markov Chain Lec 21

- A *Markov Chain* is a *sequence* of random variables $x_1, x_2, \ldots, x_n$ such that, given the present state, future and past states are independent

$$p(x_{n+1}|x_1, x_2, \ldots x_n) = p(x_{n+1}|x_n)$$

Defining a Markov chain:

- **State space** of the Markov Chain: the set from which x_i take values
- **Initial distribution (π_0):** the distribution of x_0
- **Transition probability distribution** or **Markov kernel** $K(x_n, x_{n+1})$: conditional distribution $p(x_{n+1}|x_n)$
  - *Time-homogeneous chain* when $p(x_{n+1}|x_n)$ does not depend on n

Markov Chain Lec 21

- Stationary Distribution
  - Probability distr. remains unchanged $\pi = \pi K$
- Irreducibility
  - every state reachable from every other state
- Reversibility (detailed balance eqns)
  - $p(x_0, x_1, \ldots, x_{n-1}, x_n) = p(x_n, x_{n-1}, \ldots, x_1, x_0)$
- Recurrent states/chain
  - a state is guaranteed to be revisited in finite time
- Periodicity
  - revisiting a state at regular intervals?
- Ergodicity, Convergence, Ergodic Theorem
  - a state is ergodic if it is recurrent and a-periodic
  - an ergodic Markov Chain converges to stationary distribution

Metropolis-Hastings Algorithm — Lec 21

Algorithm:

1. Initialize $x_0 \sim q$
2. **for** iteration $i = 1, 2, \ldots$ **do**
3. Propose: $x_{cand} \sim q(x_i|x_{i-1})$
4. Acceptance Prob.:

$$\alpha(x_{cand}|x_{i-1}) = min\{1, \frac{q(x_{i-1}|x_{cand})f(x_{cand})}{q(x_{cand}|x_{i-1})f(x_{i-1})}\}$$

5. $u \sim Uniform(0, 1)$
6. **if** $u < \alpha$ **then**
7. Accept the proposal $x_i \leftarrow x_{cand}$
8. **else**
9. Reject the proposal $x_i \leftarrow x_{i-1}$
10. **end if**
11. **end for**

Random Walk Metropolis-Hastings — Lec 22

- From our example, proposal distr. $q(x_{cand}|x) = \mathcal{N}(x, 0.1)$; $x_{cand} \sim \mathcal{N}(x, 0.1)$
  - Alternatively $x_{cand} = x + \epsilon$; $\epsilon \sim \mathcal{N}(0, 0.1)$
- More generally, $x_{cand} = x_{i-1} + \epsilon$
  - ϵ is a *random perturbation* with a distribution independent of current state
  - E.g., $x_{cand} = x_{i=1} + \epsilon_t$, where $\epsilon_t \sim Uniform(-\delta, \delta)$
  - E.g., $x_{cand} = x_{i=1} + \epsilon_t$, where $\epsilon_t \sim Normal(0, \tau^2)$
- In the context of the general Metropolis-Hastings algorithm
  - $q(x|y) = q(y - x)$
- Markov chain associated with q is a *radom walk*, when it is symmetric around 0, i.e, $q(-t) = q(t)$
  - due to acceptance step in M-H,M-H samples are *not* a random walk

Random Walk Metropolis-Hastings — Lec 22

- Acceptance probability

$$\alpha(x_{cond}|x_{i-1}) = min\{1, \frac{q(x_{i-1}|x_{cand})f(x_{cand})}{q(x_{cand}|x_{i-1})f(x_{i-1})}\} = min\{1, \frac{f(x_{cand})}{f(x_{i-1})}\}$$

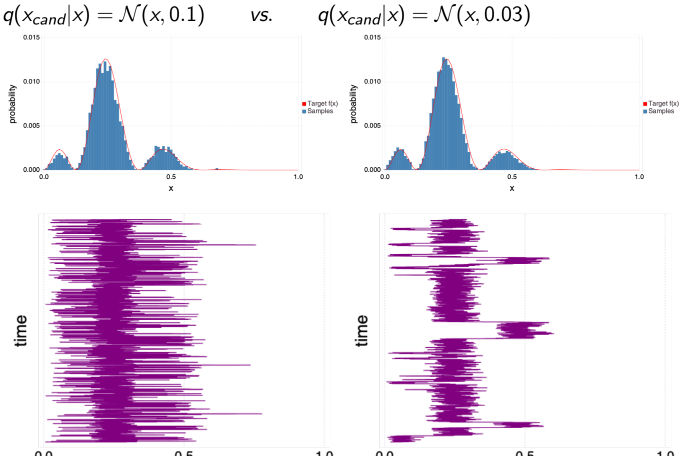
- 'Uphill' proposals are always accepted
  - when $f(x_{cand}) > f(x_{i-1})$, $\alpha = 1$
- 'Downhill' proposals are accepted with probability equal to the relative 'heights' of the target at the proposed and current values.
  - When $f(x_{cand}) < f(x_{i-1})$, $\alpha = \frac{f(x_{cand})}{f(x_{i-1})}$
- The above simplification of α is not unique to random-walk M-H
  - If $q(x_{i-1}|x_{cand}) = q(x_{cand}|x_{i-1})$, $\alpha = min\{1, \frac{f(x_{cand})}{f(x_{i-1})}\}$

Choosing $q(x, y)$ — Lec 22

- The induced Markov chain should be irreducible, with short mixing time, to allow full coverage of the state-space
  - Support of q should include support of f ($support(f) \subset support(q)$)
- Typically $q(x|y)$ is selected from a family of distributions
  - that requires specification of location and scale parameters
  - E.g., Normal, Uniform, Cauchy, Laplace, Student's T-distribution
- A $q(x|y)$ with a small 'scale' will limit the step size of the Markov Chain

Wide vs. Narrow Proposal density — Lec 21

$q(x_{cand}|x) = \mathcal{N}(x, 0.1)$ *vs.* $q(x_{cand}|x) = \mathcal{N}(x, 0.03)$



Independent Metropolis-Hastings — Lec 22

- Choosing $q(x|y)$ that is independent of the current state y - $q(x|y) = q(x)$

Algorithm:

1. Initialize $x_0 \sim q$
2. **for** iteration $i = 1, 2, \ldots$ **do**
3. Propose: $x_{cand} \sim q(x_i)$
4. Acceptance Prob.:

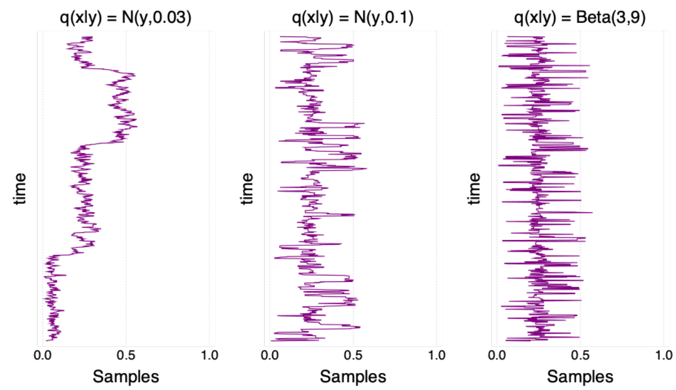
$$\alpha(x_{cond}|x_{i-1}) = min\{1, \frac{q(x_{i-1})f(x_{cand})}{q(x_{cand})f(x_{i-1})}\}$$

5. $u \sim Uniform(0, 1)$
6. **if** $u < \alpha$ **then**
7. Accept the proposal $x_i \leftarrow x_{cand}$
8. **else**
9. Reject the proposal $x_i \leftarrow x_{i-1}$
10. **end if**
11. **end for**

Independent MH vs. Accept-Reject Method Lec 22

- Independent Metropolis-Hastings
  - appears to be a straightforward generalization of Accept-reject method
- Repeated occurrences
  - no repeated occurrences in Accept-Reject Method
  - repeated occurrences possible in Independent Metropolis-Hastings
    - Step **9**: Reject the proposal $x_i \leftarrow x_{i-1}$
- Samples are
  - i.i.d in Accept-Reject Method
  - Not i.i.d in Independent Metropolis-Hastings
- Determining upper bound M using $f(x)/g(x) \leq M$
  - required in Accept-Reject Method
  - not required in Independent Metropolis-Hastings

Independent Metropolis-Hastings Lec 22



Choosing proposal density Lec 22

- The spread of the of the proposal density affects
  1. acceptance rate
  2. region of the sample space covered by the chain
- When the chain converged and density is sampled around the mode
  - If spread is extremely large, next sample will be far from current value
    - low probability of being accepted
  - If spread is too small, it will take too long to traverse support of target density
    - low probability regions will be undersampled
- Proposal density needs to be tuned appropriately

Integration using MCMC Lec 22

- While the examples we considered involve ‘sampling’
  - MCMC methods are suited for integration as well
- **Ergodic Theorem**: For a finite irreducible chain with stationary distribution π

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} h(x_t) = \mathbb{E}_\pi(h(x))$$

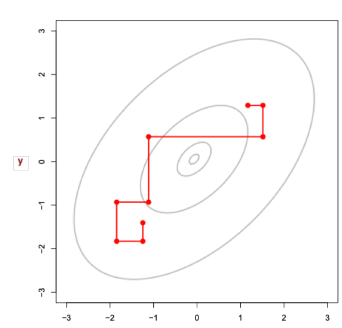
  - This expectation is the same as the integral $\int h(x)\pi(x)dx$
- **Approach:**
  - Draw n samples from $\pi(x)$ using Metropolis-Hastings
  - Compute the values for $h(x)$ using these samples
  - Compute the average of the $h(x)$ values

Gibbs sampling Lec 22

- Gibbs sampling allows us to generate samples from joint target density functions
  - Useful for sampling from a joint posterior $p(\theta_1, \theta_2, \ldots, \theta_d|y)$
- Gibbs sampling simplifies a complex high-dimensional problem
  - by breaking it down into simple, low-dimensional problems
- To draw samples from $f(x, y)$, Gibbs sampler draws from $f(x|y)$ and $f(y|x)$
  - Draw $x_{t+1} \sim f(x|y_t)$
  - Draw $y_{t+1} \sim f(y|x_t)$
  - Samples $x_0, y_0, x_1, y_1, \ldots, x_n, y_n$
- Assumes we can generate samples from $f(x|y)$ and $f(y|x)$

Gibbs sampling in 2D Lec 22

- To draw samples from $f(x, y)$
  - Draw $x_{t+1} \sim f(x|y_t)$
  - Draw $y_{t+1} \sim f(y|x_t)$
- Each step is parallel to one of the parameter axis
  - as only one component value is changed



Gibbs sampling — Lec 22

Algorithm:

1. Initialize $x^{(0)} \sim q(x)$
2. **for** iteration $i = 1, 2, \ldots$ **do**
3. $\quad x_1^{(i)} \sim p(x_1 | x_2 = x_2^{(i-1)}, x_3 = x_3^{(i-1)}, \ldots, x_d = x_d^{(i-1)})$
4. $\quad x_2^{(i)} \sim p(x_2 | x_1 = x_1^{(i-1)}, x_3 = x_3^{(i-1)}, \ldots, x_d = x_d^{(i-1)})$

$\vdots$

5. $\quad x_d^{(i)} \sim p(x_d | x_2 = x_2^{(i-1)}, x_3 = x_3^{(i-1)}, \ldots, x_{d-1} = x_{d-1}^{(i-1)})$
6. **end for**

- GS assumes that we can draw samples from the full conditionals

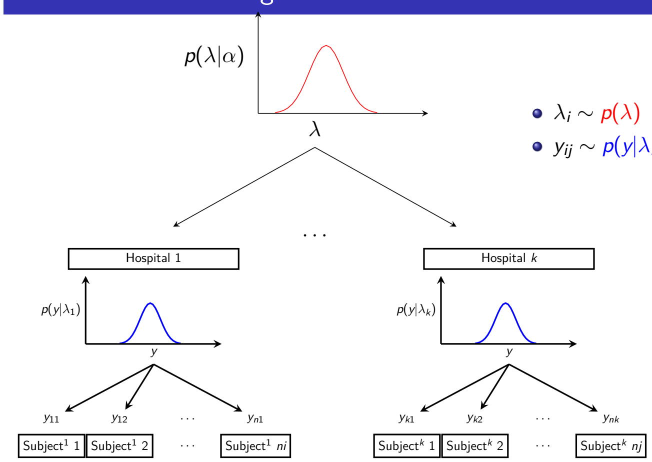
Gibbs Sampling - a special case of MH — Lec 22

- Let x_i be the i^{th} variable and x_{-i} be all variables except x_i
- Let $p(x_1, \ldots, x_d)$ be the target distribution we want to simulate
- Let $Q(x_i', x_{-i} | x_i, x_{-i}) = \frac{1}{k} p(x_i' | x_{-i})$
  - because at each step, we are drawing $x_i' \sim p(x_i' | x_{-i})$
- Let $\alpha(x_i', x_{-i} | x_i, x_{-i}) = min(1, \rho)$, where

$$\rho = \frac{q(x_{i-1}|x_{cand}) f(x_{cand})}{q(x_{cand}|x_{i-1}) f(x_{i-1})} = \frac{Q(x_i, x_{-i}|x_i', x_{-i})}{Q(x_i', x_{-i}|x_i, x_{-i})} \frac{p(x_i', x_{-i})}{p(x_i, x_{-i})}$$
$$= \frac{p(x_i', x_{-i})}{p(x_i, x_{-i})} \frac{p(x_i|x_{-i})}{p(x_i'|x_{-i})} = \frac{p(x_i'|x_{-i}) p(x_{-i})}{p(x_i|x_{-i}) p(x_{-i})} \frac{p(x_i|x_{-i})}{p(x_i'|x_{-i})}$$
$$= 1$$

- Hence, acceptance probability $\alpha = 1$

Hierarchical Modeling — Lec 23



- $\lambda_i \sim p(\lambda)$
- $y_{ij} \sim p(y|\lambda_i)$

Traditional vs. Hierarchical Modeling — Lec 23

At each hospital i

- $y_{ij} \sim p(y|\lambda_i)$

Estimate $\lambda_1, \ldots, \lambda_k$, separately

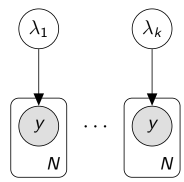
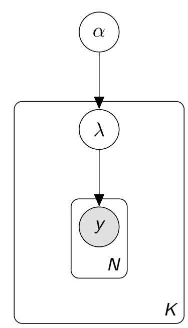


Plate-diag. interpretation:

- Nodes are random vars.
- Arrows show dependency
- Shaded nodes are obs. var.
- Plates for multiple samples

- $\lambda_i \sim p(\lambda|\alpha)$
- $y_{ij} \sim p(y|\lambda_i)$

Estimate $\lambda_1, \ldots, \lambda_k$, α



Individual vs. Combined estimation of λ_i's — Lec 23

- Individual estimates λ_i can be highly variable
  - Particularly due to hospitals with a small number of cancer patients
  - There may not be enough samples to accurately estimate *survival rates*
- As individual estimate are poor, it may seem desirable to combine the individual estimates λ_is
  - Treat λ_is as data points and estimate parameter α of the distribution $p(\lambda)$
- Since individual estimates λ_i are already noisy, estimating the parameters of the $p(\lambda)$ is ineffective
- In hierarchical modeling λ_i's and α are estimated simultaneously
  - Overcomes the above limitations with individual modeling

Traditional vs. Hierarchical Modeling — Lec 23

Traditional Model

At each hospital i

- $y_{ij} \sim p(y|\lambda_i)$

Estimate λ_i's

Bayesian setup:

- Likelihood: $p(y_{ij}|\lambda_i)$
- Prior: $p(\lambda_i|\tau)$
- Posterior $p(\lambda_i|y_{ij})$

Prior is on $\lambda_1, \ldots, \lambda_k$

Hierarchical Model

- $\lambda_i \sim p(\lambda|\alpha)$
- $y_{ij} \sim p(y|\lambda_i)$

Estimate λ_i's, α

Bayesian setup:

- Likelihood: $\prod_{ij} p(y_{ij}|\lambda_i) p(\lambda_i|\alpha)$
- Prior: $p(\alpha|\phi)$
- Posterior $p(\lambda_1, \ldots, \lambda_k, \alpha|y)$

Prior is only on α, not for $\lambda_1, \ldots, \lambda_k$

Normal Hierarchical Model — Lec 23

We assume y_{ij} and λ_i follow Gaussian distribution

- λ_i is the mean for hospital i
- variance is σ^2 and is the same for all hospitals

General Version

- $y_{ij} \sim p(y|\lambda_i)$
- $\lambda_i \sim p(\lambda|\alpha)$
- Prior: $p(\alpha|\phi)$
- Likelihood: $\prod_{ij} p(y_{ij}|\lambda_i)p(\lambda_i|\alpha)$

Specific Version: Using Normal distr.

- $y_{ij} \sim \mathcal{N}(\lambda_i, \sigma^2)$
  - where $i = 1, \ldots, k,\ j = 1, \ldots, n_i,\ n = \sum_{i=1}^{k} n_i$
- $\lambda_i \sim \mathcal{N}(\mu, \tau^2)$
- (flat) Prior: $p(\mu, \sigma^2, \tau^2) = p(\mu)p(\sigma^2)p(\tau^2) \propto \frac{1}{\sigma^2\tau^2}$

Normal Hierarchical Model — Lec 23

- Generative Model:
  - $y_{ij} \sim \mathcal{N}(\lambda_i, \sigma^2)$
    - where $i = 1, \ldots, k,\ j = 1, \ldots, n_i,\ n = \sum_{i=1}^{k} n_i$
  - $\lambda_i \sim \mathcal{N}(\mu, \tau^2)$
- Non-Inf. Prior: $p(\mu, \sigma^2, \tau^2) = p(\mu)p(\sigma^2)p(\tau^2) \propto \frac{1}{\sigma^2\tau^2}$

$$
\begin{aligned}
\text{Posterior } p(\lambda_1, \ldots, \lambda_k, \alpha|y) &\propto p(y|\lambda)p(\lambda|\alpha)p(\alpha) \\
&\propto \prod_{ij} p(y_{ij}|\lambda_i)p(\lambda_i|\alpha)p(\alpha) \\
&\propto \prod_{ij} p(y_{ij}|\lambda_i, \sigma^2)p(\lambda_i|\mu, \tau^2)p(\sigma^2, \mu, \tau^2) \\
&\propto \prod_{ij} \mathcal{N}(y_{ij}|\lambda_i, \sigma^2)\mathcal{N}(\lambda_i|\mu, \tau^2)\frac{1}{\sigma^2\tau^2}
\end{aligned}
$$

Gibbs Sampling — Lec 23

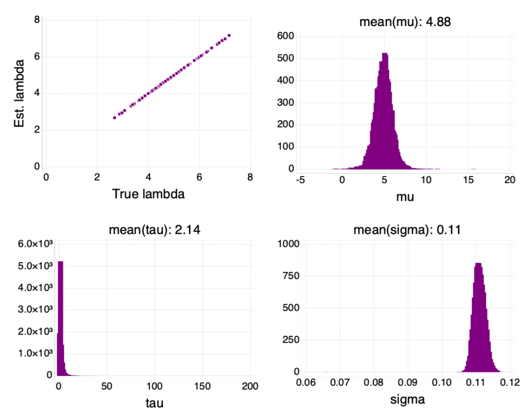
$$p(\lambda_1, \ldots, \lambda_k, \sigma^2, \mu, \tau^2|y) \propto \prod_{ij} \mathcal{N}(y_{ij}|\lambda_i, \sigma^2)\mathcal{N}(\lambda_i|\mu, \tau^2)\frac{1}{\sigma^2\tau^2}$$

1. Initialize $\lambda_1^{(1)}, \ldots, \lambda_k^{(1)}, \sigma^{2(1)}, \mu^{(1)}, \tau^{2(1)}$
2. **for** run = 2:n
3. **for** $i = 1, \ldots, k$ $\quad \lambda_i^{(run)} \sim p(\lambda_i|, \ldots)$ **end**
4. $\sigma^{2(run)} \sim p(\sigma^2| \ldots)$
5. $\mu^{(run)} \sim p(\mu| \ldots)$
6. $\tau^{2(run)} \sim p(\tau^2| \ldots)$
7. **end**

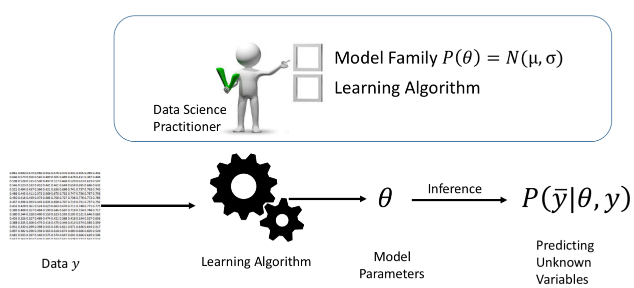
These full conditionals can be written by retaining only the terms in the posterior that has the parameter of interest

Gibbs Sampling — Lec 23

Results: True vs. Estimated parameters of Normal Hierarchical Model



Learning Probabilistic Models



- Major tasks:
  - **Learning:** Given a set of samples that are known/assumed to be generated from a model, the goal is to determine the parameters of the model.
  - **Inference:** Given a set of model parameters and an observation of some variable(s), the goal is to predict states of other variables.