MAST90125: Bayesian Statistical learning

Lecture 8: Introduction to Bayesian computation

Feng Liu and Guoqi Qian

What have we covered so far

- So far, we have learned the building blocks of Bayesian inference and analysis.
- We have learned what a prior distribution, likelihood and posterior distribution are. Further, we have developed an understanding of predictive distributions, and some principles of model checking.
- However, you may have noticed that in the examples so far, the combinations of prior/likelihood considered produced closed form posteriors from recognisable distributions.

What have we covered so far

- In many problems, closed form posteriors are not guaranteed to exist. In such cases, we need to use techniques that allow us to approximate the posterior.
- In this and the following lectures, we will focus on simulation-based techniques for approximating the posterior.
- These (usually) Monte Carlo methods can either produce independent samples, which we will see in this lecture, and you have used in the first assignment; or dependent samples, which we will study in later lectures.

How to approximate the posterior

- To see how to approximate the posterior, we need to go back to Bayes Theorem,

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})} \tag{1}$$

- Of the quantities in (1), what would you know analytically?

How to approximate the posterior

- To see how to approximate the posterior, we need to go back to Bayes Theorem,

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})} \tag{1}$$

- Of the quantities in (1), what would you know analytically?
  - $p(\theta)$ and $p(\mathbf{y}|\theta)$.
- What purpose do the quantities that you do not know analytically serve?

How to approximate the posterior

- To see how to approximate the posterior, we need to go back to Bayes Theorem,

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})} \tag{1}$$

- Of the quantities in (1), what would you know analytically?
  - $p(\theta)$ and $p(\mathbf{y}|\theta)$.
- What purpose do the quantities that you do not know analytically serve?
  - $p(\mathbf{y})$ is a normalising constant. This is why people write,

$$p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)p(\theta)$$

- Hence to approximate the posterior, we often work with an un-normalised density $q(\theta|\mathbf{y})$, which must satisfy $q(\theta|\mathbf{y}) = c(\mathbf{y})p(\mathbf{y}|\theta)p(\theta) = d(\mathbf{y})p(\theta|\mathbf{y})$, where $c(\mathbf{y}), d(\mathbf{y})$ are functions of $\mathbf{y}$ but not θ.

Direct approximation

- The first method we will look at is direct approximation.
- For this approach, assume $\theta \in (a, b)$. Next define a grid of points, $\theta_1, \ldots, \theta_N$ such that $\theta_1 = a, \theta_N = b$ and $\theta_{i+1} - \theta_i = (b - a)/(N - 1)$.
- Provided N is sufficiently large then

$$\frac{p(\theta_i|\mathbf{y})}{\sum_{j=1}^{N} p(\theta_j|\mathbf{y})} = \frac{q(\theta_i|\mathbf{y})/d(\mathbf{y})}{\sum_{j=1}^{N} q(\theta_j|\mathbf{y})/d(\mathbf{y})} = \frac{q(\theta_i|\mathbf{y})}{\sum_{j=1}^{N} q(\theta_j|\mathbf{y})}$$

should approximate $\Pr(\theta_i - \epsilon/2 \leq \theta \leq \theta_i + \epsilon/2|\mathbf{y})$ and

$$\frac{\sum_{h=1}^{i} q(\theta_h|\mathbf{y})}{\sum_{j=1}^{N} q(\theta_j|\mathbf{y})}$$

should approximate $\Pr(\theta \leq \theta_i|\mathbf{y})$

Direct approximation continued

- Having thus discretised the posterior distribution, the process of taking a random draw $\tilde{\theta}$ from the posterior consists of
  - Drawing a value x from a standard uniform, $x \sim U(0, 1)$.
  - Finding $\tilde{\theta}$ using the inverse cdf of the posterior (how to do it?).
- Now for a question. What is implied about θ from the way we have looked at the algorithm so far?

Direct approximation continued

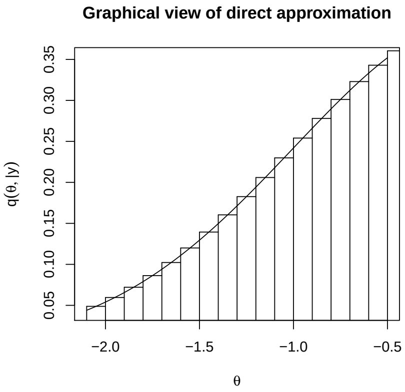
- Having thus discretised the posterior distribution, the process of taking a random draw $\tilde{\theta}$ from the posterior consists of
  - Drawing a value x from a standard uniform, $x \sim U(0, 1)$.
  - Finding $\tilde{\theta}$ using the inverse cdf of the posterior (how to do it?).
- Now for a question. What is implied about θ from the way we have looked at the algorithm so far?
  - This example is written assuming θ is univariate. While it is straight-forward to create a multi-dimensional grid for the case where θ is multi-variate, the computational cost would become prohibitive rapidly.
    - For example, if $\boldsymbol{\theta}$ is m-dimensional ($\boldsymbol{\theta} = (\theta_1 \quad \cdots \quad \theta_m)$) and direct approximation is applied so that marginally each component is considered on a grid of N points, then the number of points where evaluations are required is N^m.

Questions arising from direct approximation

- What mathematical technique is direct approximation an example of?

Questions arising from direct approximation

- What mathematical technique is direct approximation an example of?



- Direct approximation is based on a deterministic method of numerical integration. This becomes more obvious if we rewrite the discrete density as,

$$\frac{q(\theta_i|\mathbf{y})}{\sum_{j=1}^{N} q(\theta_j|\mathbf{y})} = \frac{q(\theta_i|\mathbf{y})\epsilon}{\sum_{j=1}^{N} q(\theta_j|\mathbf{y})\epsilon}.$$

- Hence in the graph to the left, each rectangle corresponds to $q(\theta_i|\mathbf{y})\epsilon$ for some θ_i, with ϵ being the rectangle width.

Example of direct approximation

- Lets say you have normally distributed data where you know the mean μ but not the variance σ^2. Further, assume that the prior distribution for $\tau = (\sigma^2)^{-1}$ is Ga(α, β).
- The joint distribution $p(\mathbf{y}, \tau)$ is thus,

$$\frac{\tau^{n/2}}{(2\pi)^{n/2}} e^{-\frac{\tau((n-1)s^2 + n(\bar{y}-\mu)^2)}{2}} \times \frac{\beta^\alpha \tau^{\alpha-1} e^{-\beta\tau}}{\Gamma(\alpha)} \propto \tau^{n/2+\alpha-1} e^{-\frac{\tau((n-1)s^2 + n(\bar{y}-\mu)^2 + 2\beta)}{2}}.$$

- While by looking at the kernel, we know that the posterior is a Gamma distribution, lets pretend you cannot sample from such a distribution.

Example of direct approximation

- Assume that $\mu = 5$, $\bar{y} = 4.88$, $n = 10$, $s^2 = 1.23$ and $\alpha = \beta = 1$
- To demonstrate direct approximation lets define a grid from $(0, 2.5)$ and let $N = 50, 200, 1000$.
- Think about how to implement the direct approximation using R program, we will show the details now.

Stochastic methods of posterior approximation

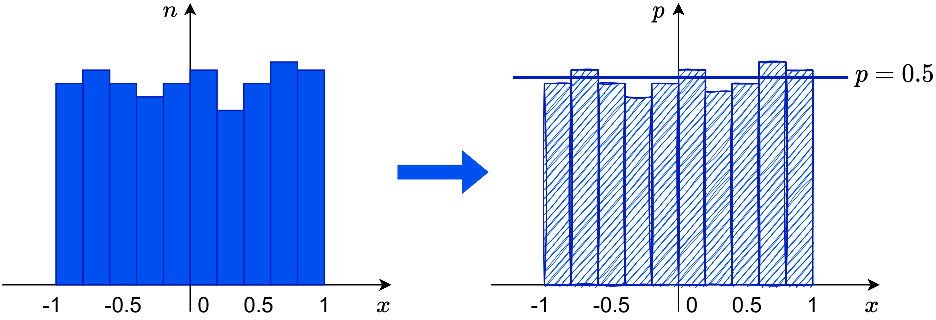
The remaining slides of this lecture will be replaced by new slides.

- While direct approximation is based on a deterministic method of numerical integration, the following methods we will study in this lecture are based on generating random numbers.
- Now, let's look at the hist graph (frequency of samples) and the probability density function.

Stochastic methods of posterior approximation

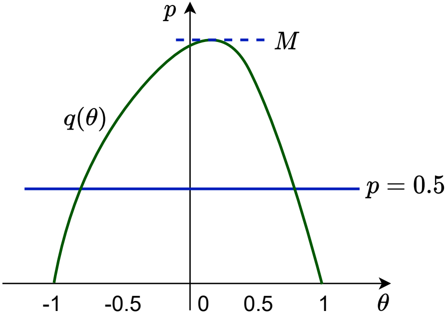
- Now, let's look at the hist graph and the probability density function.

n : Number of Samples $\quad$ p : Probability Density



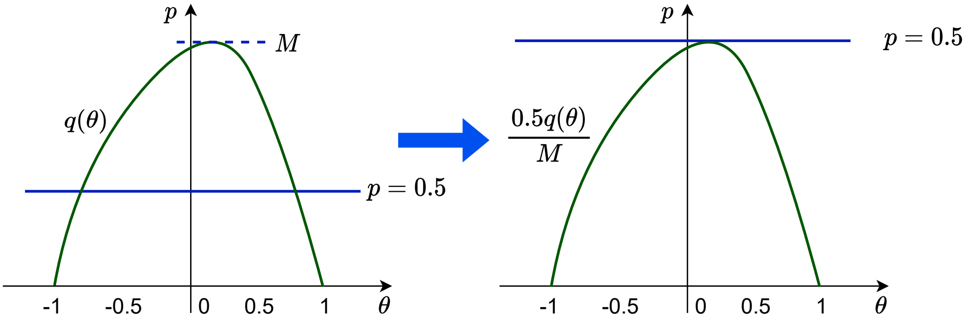
Stochastic methods of posterior approximation

- What can we do if our interested function $q(\theta)$ is like this?



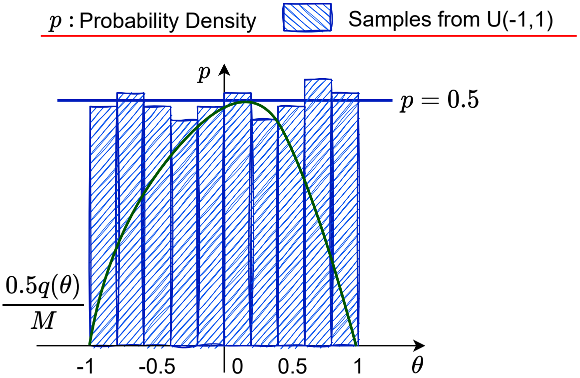
Stochastic methods of posterior approximation

- Let's scale the $q(\theta)$!



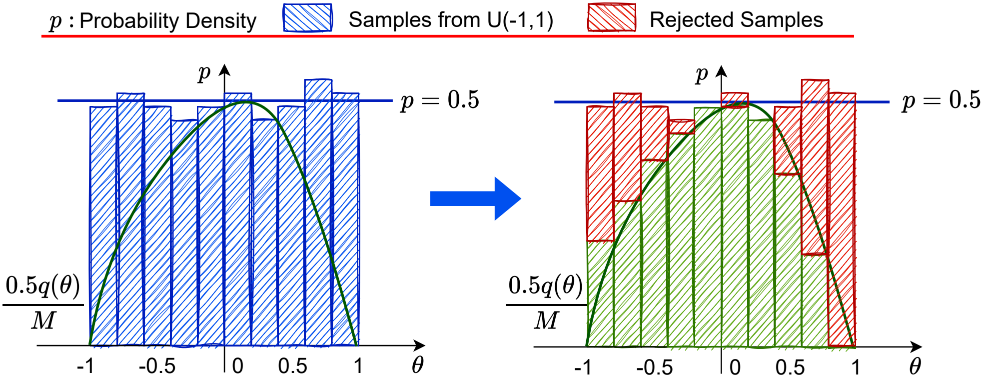
Stochastic methods of posterior approximation

- Let's show our samples back.



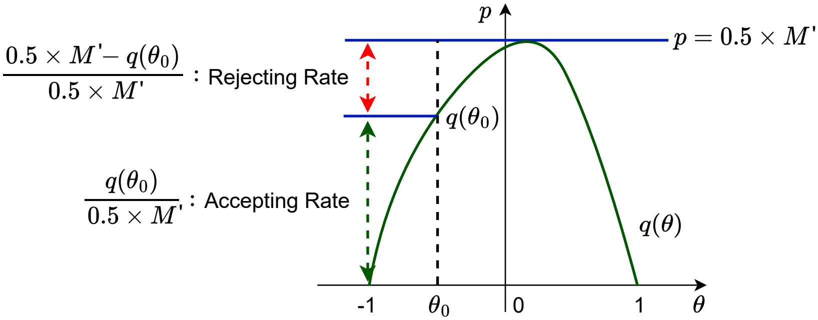
Rejection sampling

- Maybe we can reject/delete some samples.



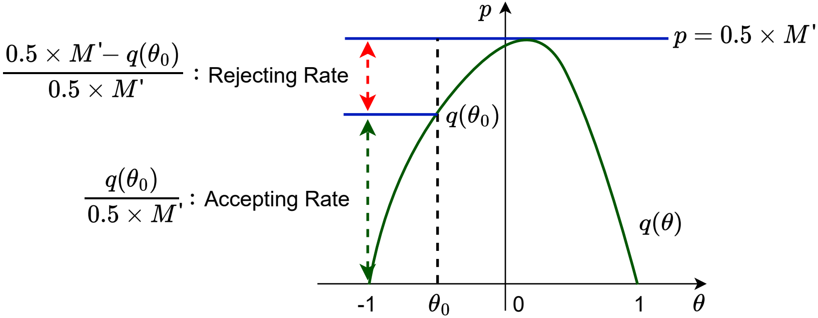
Rejection sampling

- Can we reject/delete one sample θ?



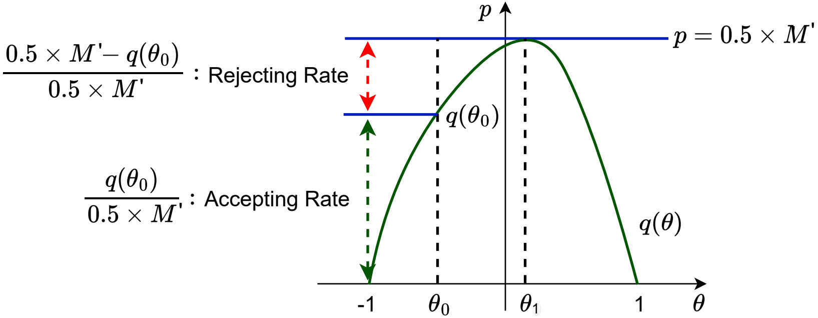
Rejection sampling

- Sure. After we sample θ_0, we can just sample a number x from U(0,1). If $x <$ the accepting rate, then we keep θ_0. Otherwise, we reject θ_0.



Rejection sampling

- It is also clear that, if we have a θ_1 such that $q(\theta_1) = 0.5 \times M$, then we will never reject θ_1, because the accepting rate of θ_1 is $1 = 100\%$.

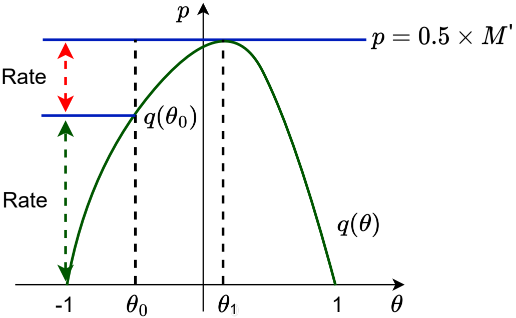


Rejection sampling

- This is the well-known Monte Carlo (MC) method!

$$\frac{0.5 \times M' - q(\theta_0)}{0.5 \times M'} : \text{Rejecting Rate}$$

$$\frac{q(\theta_0)}{0.5 \times M'} : \text{Accepting Rate}$$



Rejection sampling (more general descriptions)

- The idea behind rejection sampling is to find a density function $g(\theta)$ that completely encases the posterior $p(\theta|y)$, or in practice the un-normalised density $q(\theta|y)$, or equivalently

$$\frac{q(\theta|y)}{g(\theta)} \leq M \quad \forall \theta,$$

such that it is straight-forward to sample from $g(\theta)$. In our previous figures, $g(\theta) = 0.5$. Specifically, we sample thetas from $U(-1, 1)$.

Rejection sampling (more general descriptions)

- The idea behind rejection sampling is to find a density function $g(\theta)$ that completely encases the posterior $p(\theta|y)$, or in practice the un-normalised density $q(\theta|y)$, or equivalently

$$\frac{q(\theta|y)}{g(\theta)} \leq M \quad \forall\theta,$$

  such that it is straight-forward to sample from $g(\theta)$. In our previous figures, $g(\theta) = 0.5$. Specifically, we sample thetas from $U(-1, 1)$.
- The generation of draws from the posterior then proceeds as follows:
  - Sample θ^s from $g(\theta)$.
  - Sample x from a standard uniform U(0,1).
  - If $x \leq \frac{q(\theta^s|y)}{Mg(\theta^s)}$, accept θ^s, otherwise reject.

Example of rejection sampling

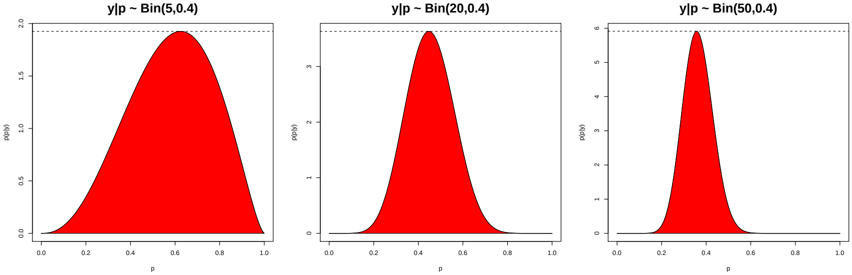
- Assume $y|p \sim \text{Bin}(n, p)$ and that the prior distribution for p is $\text{Be}(\alpha, \beta)$.
- We know that the posterior distribution $p|y$ is $\text{Be}(y+\alpha, n-y+\beta)$, but lets assume you cannot sample directly from this distribution.
- We also know that p is bounded on $[0, 1]$, so a simple choice for $g(p) = 1$, the standard uniform distribution. Then M would correspond to the maximum of the posterior, which occurs at $p_{\text{max}} = \frac{y+\alpha-1}{n+\alpha+\beta-2}$ with

$$M = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)} p_{\text{max}}^{y+\alpha-1}(1-p_{\text{max}})^{n-y+\beta-1}.$$

- Think about how to implement the MC using R program, we will show the details in the next lecture. Assume $\alpha = \beta = 0.5$, n can be either $5, 20$, or 50, and $y \sim Bin(n, 0.4)$.

Rejection sampling comments

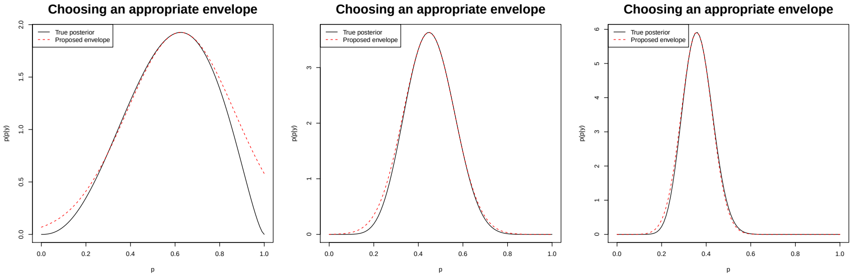
- The challenge of rejection sampling is picking $g(\theta)$ such that $q(\theta|y) \leq Mg(\theta)\ \forall\theta$ while minimising the proportion of candidate samples being rejected.



- In the case of the beta posterior example, as y, n increases, the probability of any θ^s being accepted (area in red below dashed line in figure) declines.

Rejection sampling comments

- Now, based on what you know about asymptotic theory, a normal distribution based on the posterior mode truncated at $[0,1]$ might be a better choice for $g(p)$.



- As before, and also for ease of calculation, we choose M so that $\max_p p(p|y) = M \max_p g(p)$ matched. While the choice of $g(p)$ looks better, especially for larger n, it turns out that $p(p|y)/g(p) \leq M$ does not hold $\forall p$.