Lecture 19. Bayesian classification

COMP90051 Statistical Machine Learning

Lecturer: Feng Liu

This lecture

- Bayesian ideas in discrete settings
  * Beta-Binomial conjugacy
  * Uniqueness up to proportionality
  * Sunrise example
  * Common conjugate pairs
- Bayesian logistic regression
  * Non-conjugacy
  * Pointer: Laplace approximation
- Rejection Sampling
  * Monte Carlo sampling
  * A stochastic method of posterior approximation

How to apply Bayesian view to discrete data?

- First off consider models which *generate* the input
  - cf. *discriminative* models, which *condition* on the input
  - I.e., $p(y \mid \boldsymbol{x})$ vs $p(\boldsymbol{x}, y)$, Logistic Regression vs Naïve Bayes
- For simplicity, start with most basic setting
  - n coin tosses, of which k were heads
  - only have $\boldsymbol{x}$ (sequence of outcomes), but no 'classes' $\boldsymbol{y}$
- Methods apply to generative models over discrete data
  - e.g., topic models, generative classifiers (Naïve Bayes, mixture of multinomials)

Discrete Conjugate prior: Beta-Binomial

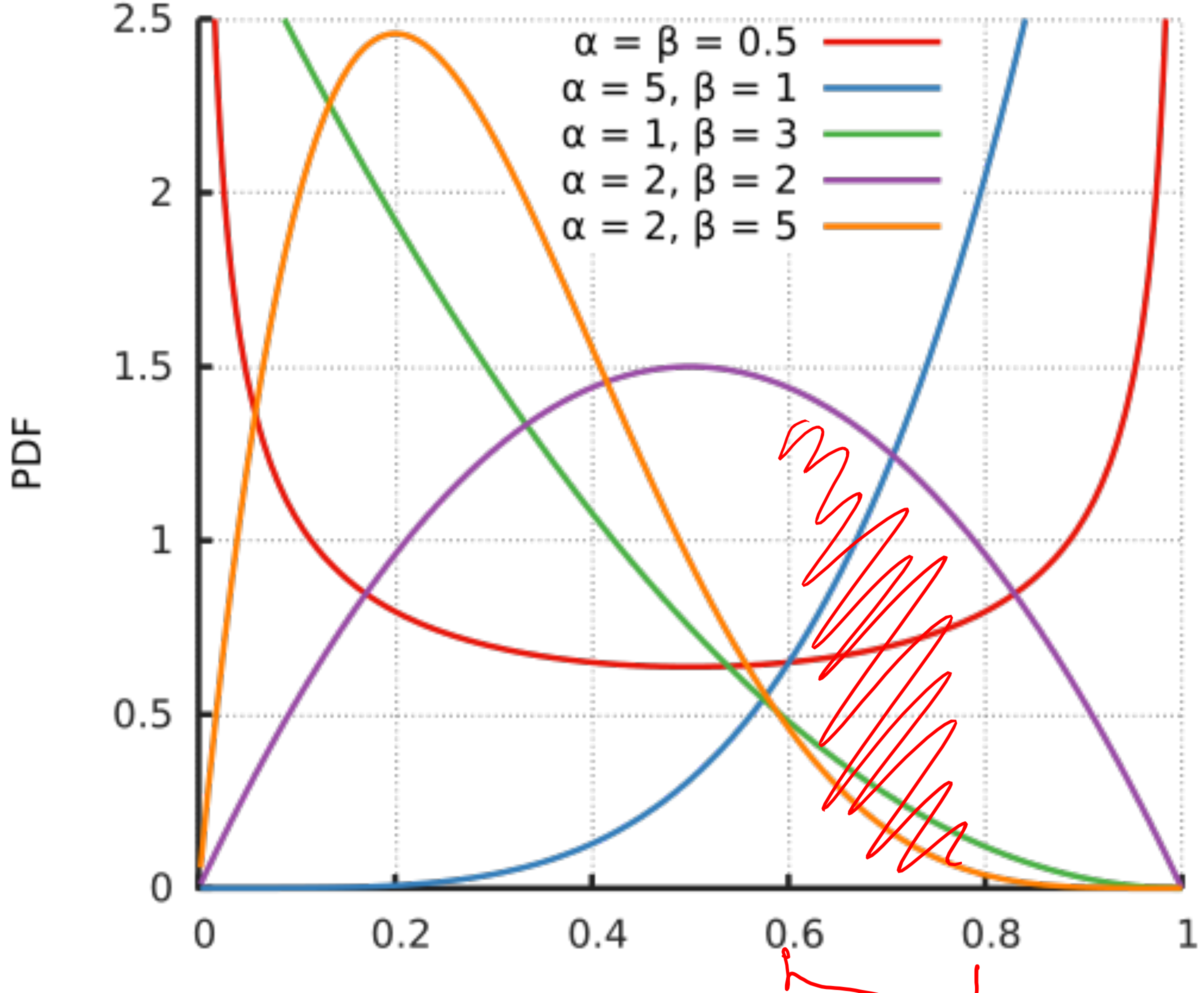
- Conjugate priors also exist for discrete spaces
- Consider *n* coin tosses, of which *k* were heads
  - let p(head) = *q* from a single toss (*Bernoulli dist*)
  - Inference question is the coin biased, i.e., is *q ≈ 0.5*
- Several draws, use *Binomial dist*
  - and its conjugate prior, *Beta dist*

$$p(k|n,q) = \binom{n}{k} q^k (1-q)^{n-k}$$

$$p(q) = \mathrm{Beta}(q;\alpha,\beta)$$

$$= \frac{\gamma(\alpha+\beta)}{\gamma(\alpha)\gamma(\beta)} q^{\alpha-1}(1-q)^{\beta-1}$$

Beta distribution



Sourced from https://en.wikipedia.org/wiki/Beta_distribution

Beta-Binomial conjugacy

$$p(k|n,q) = \binom{n}{k} q^k (1-q)^{n-k}$$

$$p(q) = \mathrm{Beta}(q;\alpha,\beta)$$

$$= \frac{\gamma(\alpha+\beta)}{\gamma(\alpha)\gamma(\beta)} q^{\alpha-1}(1-q)^{\beta-1}$$

Sweet! We know the normaliser for Beta

Bayesian posterior

$$p(q|k,n) \propto p(k|n,q)p(q)$$

$$\propto q^k(1-q)^{n-k} q^{\alpha-1}(1-q)^{\beta-1}$$

$$= q^{k+\alpha-1}(1-q)^{n-k+\beta-1}$$

$$\propto \mathrm{Beta}(q; k+\alpha, n-k+\beta)$$

trick: ignore constant factors (normaliser)

Uniqueness up to normalisation

- A trick we've used many times:

  When an unnormalized distribution is proportional to a recognised distribution, we say it must be that distribution

- If $f(\theta) \propto g(\theta)$ for g a distribution, $\frac{f(\theta)}{\int_{\Theta} f(\theta)d\theta} = g(\theta)$.

- Proof: $f(\theta) \propto g(\theta)$ means that $\exists C$

$$f(\theta) = C \cdot g(\theta)$$

$$\int_{\Theta} f(\theta)d\theta = C \int_{\Theta} g(\theta)d\theta = C$$

and the result follows from LHS1/LHS2 = RHS1/RHS2

Laplace's Sunrise Problem

Every morning you observe the sun rising. Based solely on this fact, what's the probability that the sun will rise tomorrow?

- Use Beta-Binomial, where *q* is the Pr(sun rises in morning)
  - posterior $$p(q|k,n) = \mathrm{Beta}(q; k+\alpha, n-k+\beta)$$
  - n = k = observer's age in days
  - let α = β = 1 (*uniform* prior)

- Under these assumptions

$$p(q|k) = \mathrm{Beta}(q; k+1, 1)$$

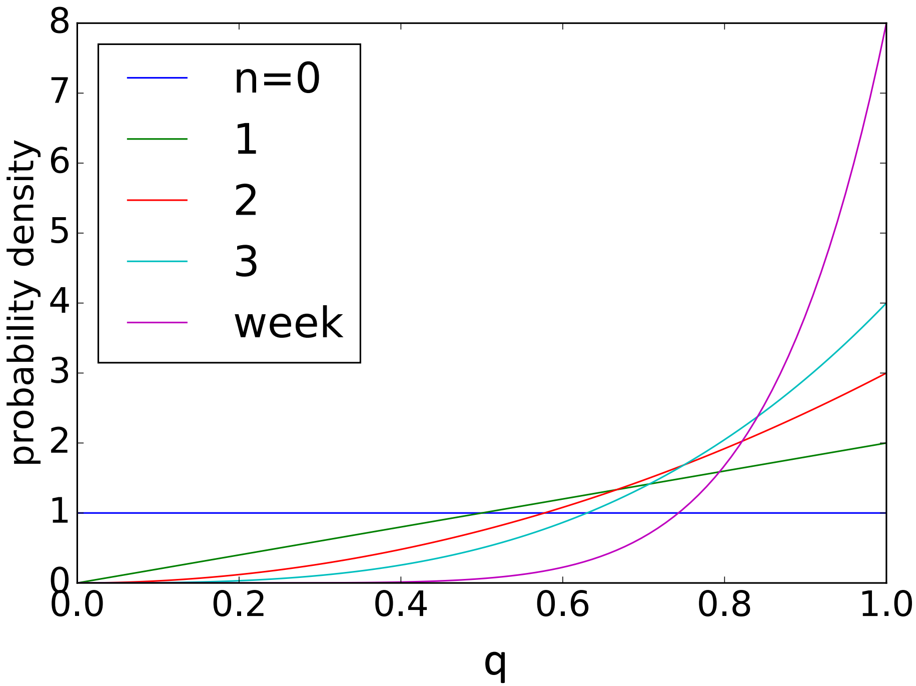
$$E_{p(q|k)}[q] = \frac{k+1}{k+2}$$

'smoothed' count of days where sun rose / did not

Sunrise Problem (cont.)

Consider human-meaningful period

Day (n, k)	k+α	n-k+β	E[q]
0	1	1	0.5
1	2	1	0.667
2	3	1	0.75
…			
365	366	1	0.997
2920 (8 years)	2921	1	0.99997



Effect of prior diminishing with data, *but never disappears completely*.

Suite of useful conjugate priors

	likelihood	conjugate prior
regression	Normal	Normal (for mean)
	Normal	Inverse Gamma (for variance) or Inverse Wishart (covariance)
classification	Binomial	Beta
	Multinomial	Dirichlet
counts	Poisson	Gamma

Mini Summary

- Bayesian ideas in discrete settings
  - Beta-Binomial conjugacy
  - Uniqueness in proportionality
  - Sunrise example
  - Conjugate pairs

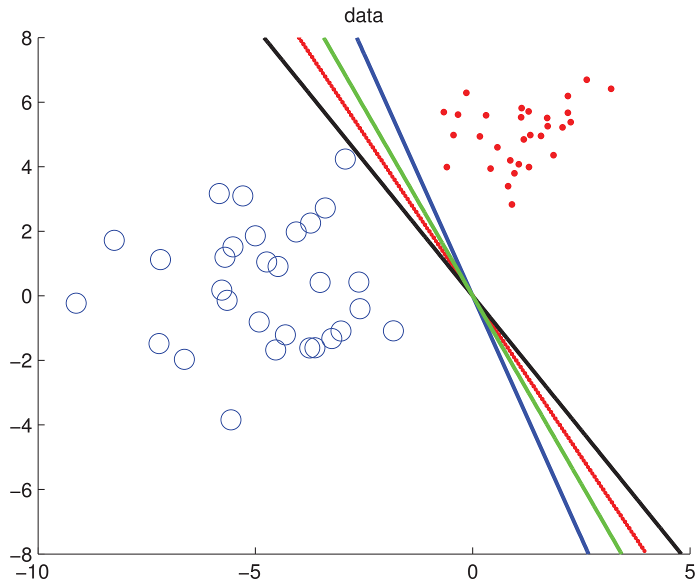
Next time: Bayesian logistic regression

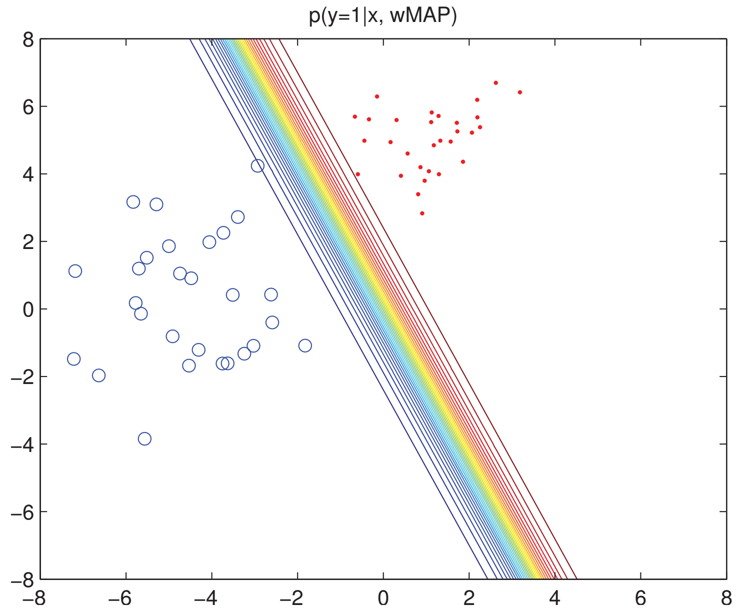
Bayesian Logistic Regression

Discriminative classifier, which conditions *on inputs. How can we do Bayesian inference in this setting?*

Now for Logistic Regression…

- Similar problems with parameter uncertainty compared to regression
  - although predictive uncertainty in-built to model outputs





Murphy Fig 8.5 & 8.6 p257-8

No conjugacy

- Can we use conjugate prior? E.g.,
  - Beta-Binomial for *generative* binary models
  - Dirichlet-Multinomial for multiclass (similar formulation)
- Model is *discriminative*, with parameters defined using logistic sigmoid*

$$p(y|q, \mathbf{x}) = q^y(1-q)^{1-y}$$
$$q = \sigma(\mathbf{x}'\mathbf{w})$$

  - need prior over ***w***, not *q*
  - no known conjugate prior (!), thus use a Gaussian prior
- Approach to inference: Monte Carlo sampling

** Or softmax for multiclass; same problems arise and similar solution*

Approximation

- No known solution for the normalising constant

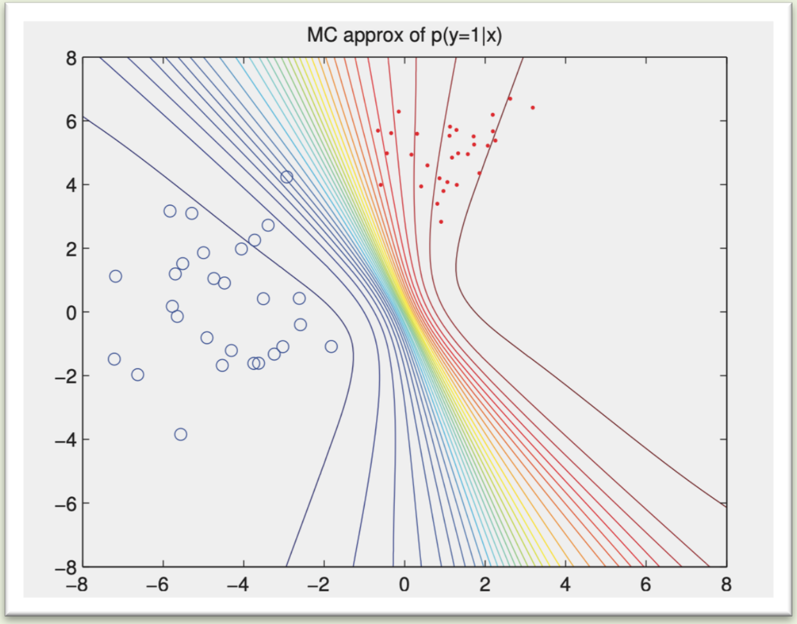
$$p(\mathbf{w}|\mathbf{X},\mathbf{y}) \propto p(\mathbf{w})p(\mathbf{y}|\mathbf{X},\mathbf{w})$$

$$= \mathrm{Normal}(\mathbf{0},\sigma^2\mathbf{I})\prod_{i=1}^{n}\sigma(\mathbf{x}_i'\mathbf{w})^{y_i}(1-\sigma(\mathbf{x}_i'\mathbf{w}))^{1-y_i}$$

- Resolve by *approximation*

Laplace approx.:

- assume posterior $\simeq$ Normal about mode
- can compute normalisation constant, draw samples etc.
- Tractable MAP provides parameters for this (Normal) approximate posterior



Murphy Fig 8.6 p258

How to approximate the posterior

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

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

- 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,

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

- Of the quantities in (1), what would you know analytically?
  - $p(\theta)$ and $p(y|\theta)$.
- 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,

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

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

unnormalised density
= likelihood * prior

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

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

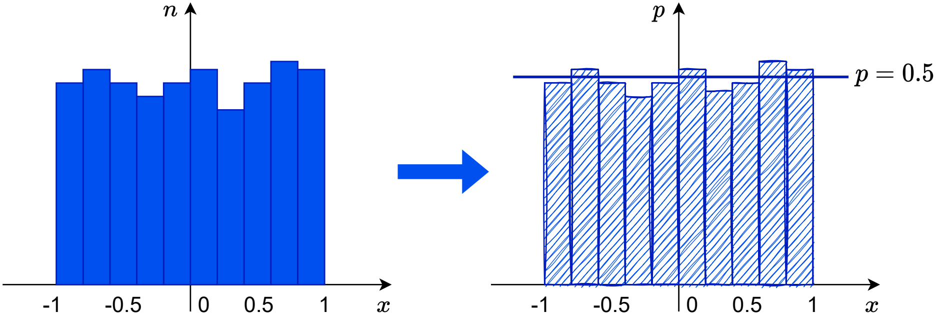
Stochastic methods of posterior approximation

- Let's first look at the hist graph (frequency of samples) and the probability density function.

Stochastic methods of posterior approximation

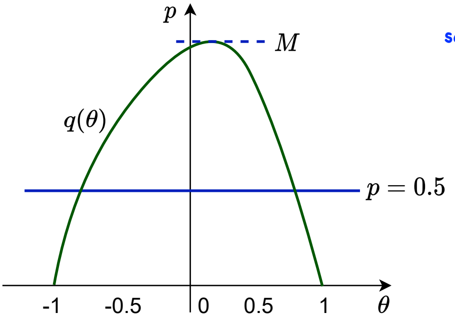
- Now, let's look at the hist graph and the probability density function.

n : Number of Samples p : Probability Density



Stochastic methods of posterior approximation

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

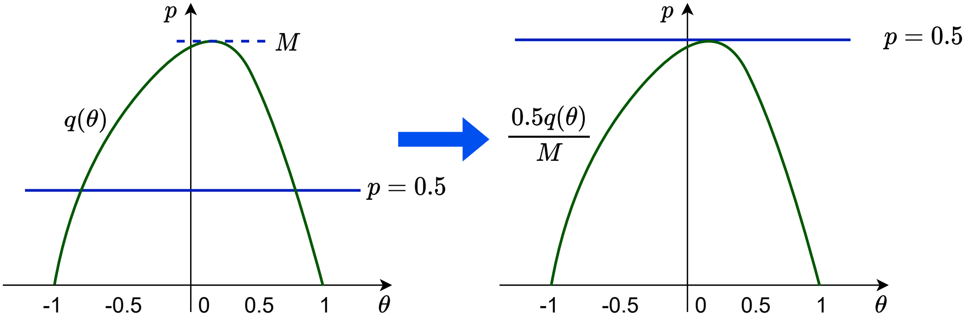


sample from the un-normalised density:

area under $q(\theta)$ > 1

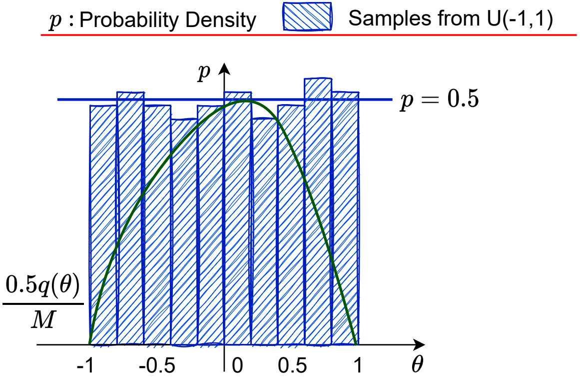
Stochastic methods of posterior approximation

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



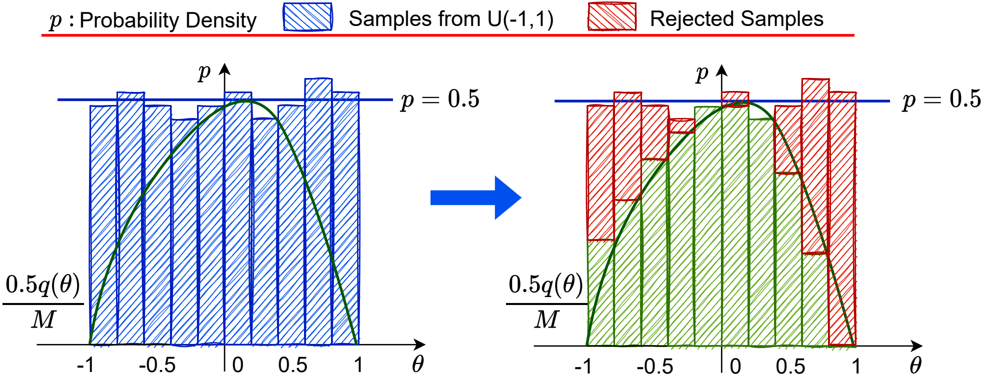
Stochastic methods of posterior approximation

- Let's show our samples back.



Rejection sampling

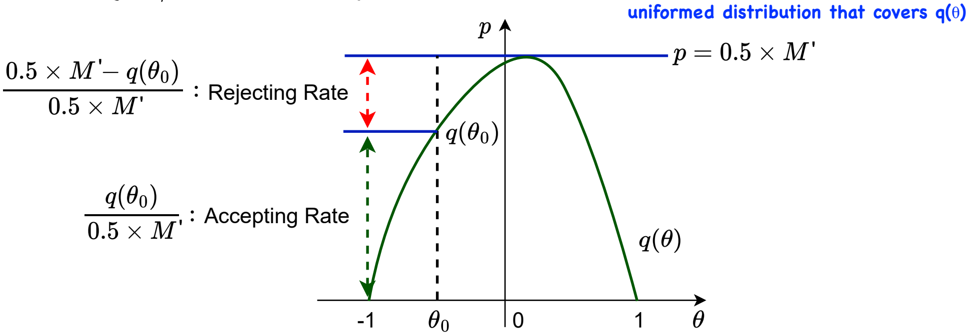
- Maybe we can reject/delete some samples.



reject some sample, as we need to sample some distribution that can cover our posterior distribution

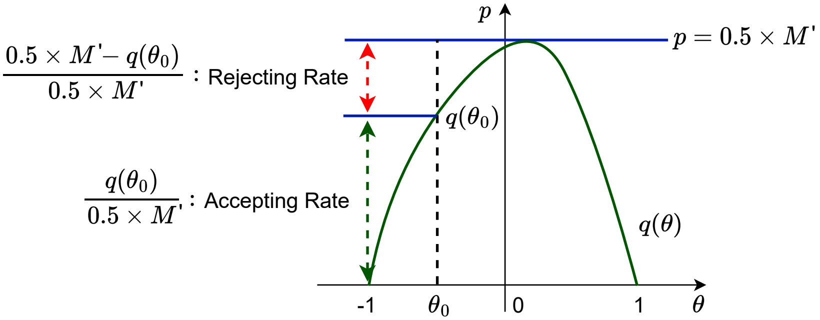
Rejection sampling

- Can we reject/delete one sample θ?



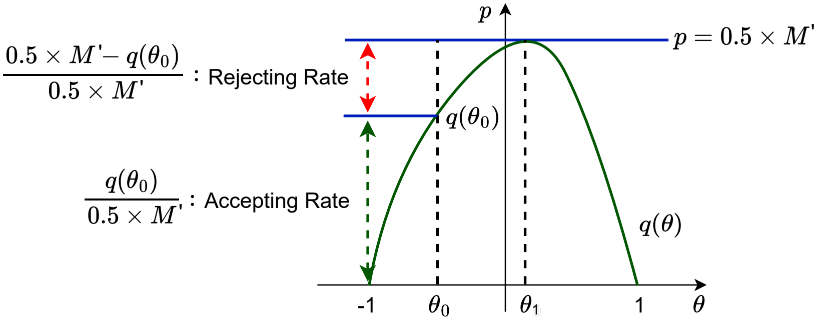
Rejection sampling

- 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.



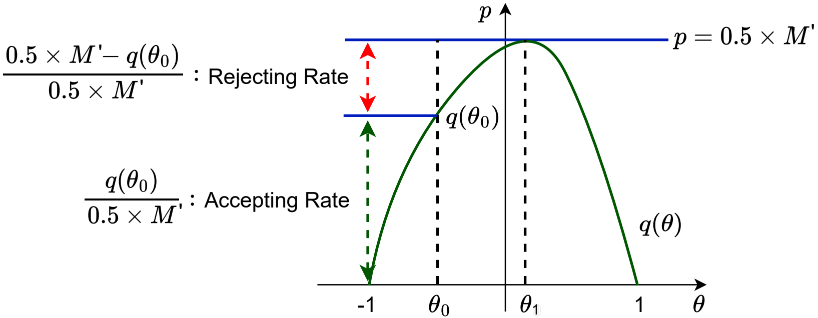
Rejection sampling

- 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!



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).

g: uniform distribution between 0 and 1, then $g(\theta) = 1$,
and M would be the max value of $q(\theta)$

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 $x \leq \frac{q(\theta^s|y)}{M' g(\theta^s)}$, accept θ^s, otherwise reject.

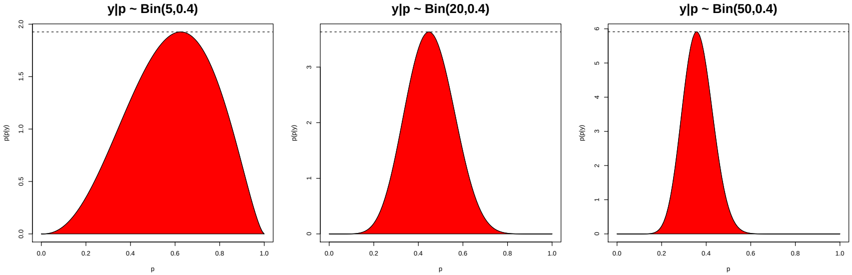
Example of rejection sampling

- 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

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

Rejection sampling comments

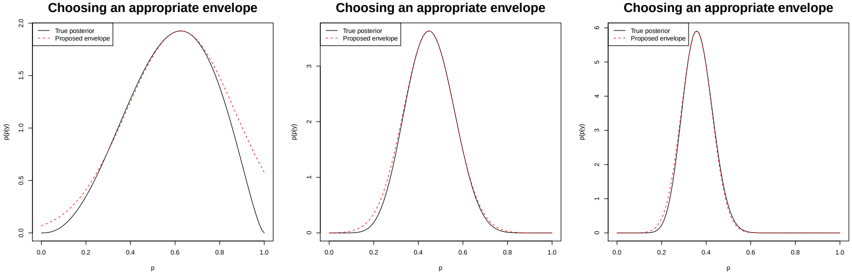
- The challenge of rejection sampling is picking $g(\theta)$ such that $q(\theta|y) \leq Mg(\theta)\ \forall\theta$ while minimising the proportion of candidate samples being rejected.



- 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 $\forall p$.

Mini Summary

- Bayesian ideas in discrete settings
  - Beta-Binomial conjugacy
  - Conjugate pairs; Uniqueness in proportionality
- Bayesian classification (logistic regression)
  - Non-conjugacy necessitates approximation
- Rejection sampling
  - Monte Carlo sampling: A classic method to approximate posterior

Next time: probabilistic graphical models