An overview of Bayesian analysis

Benjamin Letham

Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA
bletham@mit.edu

May 2012

1 Introduction and Notation

This work provides a fairly rigorous yet simultaneously informal introduction to Bayesian analysis intended for those with an understanding of basic probability and interest in machine learning. We cover everything from theoretical aspects of posterior asymptotics to practical considerations in MCMC sampling. This section (Section 1) provides an introduction to the Bayesian approach and the necessary notation. In Section 2 we move to point estimation, specifically maximum likelihood and maximum *a posteriori* estimation and discuss a fascinating connection between Bayesian regression and ridge regression. Section 3 introduces exponential families and conjugate priors, which are crucial for constructing tractable Bayesian models. Section 4 discusses posterior asymptotics, and addresses the issue of what will happen if the model is wrong. Section 5 discusses hierarchical modeling and the extremely popular topic models. We finish in Section 6 with a discussion of MCMC sampling.

Most popular machine learning tools (SVM, Boosting, Decision Trees,...) are fairly agnostic to how the data were generated. In Bayesian analysis, we will work under the assumption that the data were generated from a probability distribution. Given data, our goal then becomes to determine *which* probability distribution generated the data. This is the essence of statistical inference.

Let us introduce some notation and some key definitions. We will suppose that we are given N data points $y_1, \ldots, y_N$, each of arbitrary dimension. Let $y = \{y_1, \ldots, y_N\}$ denote the full set of data. We will assume that the data were generated from a probability distribution $p(y|\theta)$ that is described by the parameters θ (not necessarily scalar). We call $p(y|\theta)$ a *likelihood function* or likelihood model for the data y, as it tells us how likely the data y are given the model specified by θ. Before seeing the data, we must specify a distribution over θ, $p(\theta)$. This distribution represents any knowledge we have about how the data are generated prior to observing them, and will play a very important role later in this section. We call this the *prior* distribution. Our end goal is the distribution $p(\theta|y)$, that is, which parameters are likely given the observed data. We call this the *posterior* distribution.

We, the modeler, specify the likelihood function and the prior using our knowledge of the system at hand. We then use these quantities, together with the data, to compute the posterior. The likelihood, prior, and posterior are all related via Bayes' rule:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta')p(\theta')d\theta'}. \tag{1}$$

Unfortunately the integral in the denominator, called the *partition function*, is often intractable. This is what makes Bayesian analysis difficult, and a big portion of what follows will essentially be avoiding that integral.

Coin flip example part 1. Suppose we have been given data from a series of N coin flips, and we are not sure if the coin is fair or not. We might assume that the data were generated by a sequence of independent draws from a Bernoulli distribution, which is parameterized by θ the probability of flipping Heads. Let $y_i = 1$ if flip i was Heads, and $y_i = 0$ otherwise. Then the likelihood model is,

$$p(y|\theta) = \theta^{\sum_{i=1}^N y_i}(1-\theta)^{(N-\sum_{i=1}^N y_i)}.$$

To determine which Bernoulli distribution generated these data, we might estimate θ as the proportion of the data which are Heads. We will see shortly that this is a principled approach.

1.1 A note on the Bayesian approach

The problem formulation we have just described has historically been a source of much controversy in statistics. There are generally two subfields of statistics: frequentist (or classical) statistics, and Bayesian statistics. Although many of the techniques overlap, there is a fundamental difference in philosophy. In the frequentist approach, θ is an unknown, but *deterministic* quantity. The goal in frequentist statistics is then to determine the range of values for θ that is supported by the data (called a confidence interval). When θ is viewed as a deterministic quantity, it is nonsensical to talk about its probability distribution. One of the greatest statisticians of our time, Fisher, wrote that Bayesian statistics "is founded upon an error, and must be wholly rejected." Another of the great frequentists, Neyman, wrote that, "the whole theory would look nicer if it were built from the start without reference to Bayesianism and priors." Nevertheless, recent advances in theory and particularly in computation have shown Bayesian statistics to be very useful for many applications.

2 Point estimates

Rather than estimate the entire distribution $p(\theta|y)$, sometimes it is sufficient to find a single 'good' value for θ. We call this a *point estimate*. For the sake of completeness, we will briefly discuss two widely used point estimates, the *maximum likelihood* (ML) estimate and the *maximum a posteriori* (MAP) estimate.

2.1 Maximum likelihood estimation

The ML estimate for θ is denoted $\hat{\theta}_{\mathrm{ML}}$ and is the value for θ under which the data are most likely:

$$\hat{\theta}_{\mathrm{ML}} \in \arg\max_{\theta} p(y|\theta). \tag{2}$$

As a practical matter, when computing the maximum likelihood estimate it is often easier to work with the *log-likelihood*, $\ell(\theta) := \log p(y|\theta)$. Because the logarithm is monotonic,

$$\hat{\theta}_{\mathrm{ML}} \in \arg\max_{\theta} \ell(\theta). \tag{3}$$

The ML estimator is very popular and has been used all the way back to Laplace. It has a number of nice properties, one of which is that it is a consistent estimator.

Definition 1. Suppose the data $y_1, \ldots, y_N$ were generated by a probability distribution $p(y|\theta_0)$. An estimator $\hat{\theta}$ is *consistent* if it converges in probability to the true value: $\hat{\theta} \xrightarrow{P} \theta_0$ as $N \to \infty$. An estimator $\hat{\theta}$ is *strongly consistent* if it converges almost surely: $\hat{\theta} \xrightarrow{a.s.} \theta_0$ as $N \to \infty$.

Under some reasonable conditions, $\hat{\theta}_{\mathrm{ML}}$ is strongly consistent. Thus, if the distribution that generated the data belongs to the family defined by our likelihood model, maximum likelihood is guaranteed to find the correct distribution, as N goes to infinity. Unfortunately, there are no small sample guarantees.

Coin flip example part 2. Returning to the coin flip example, the log-likelihood is,

$$\ell(\theta) = \left(\sum_{i=1}^{N} y_i\right) \log\theta + \left(N - \sum_{i=1}^{N} y_i\right) \log(1-\theta).$$

We can maximize this easily by differentiating and setting to zero, and doing a few lines of algebra:

$$\begin{aligned}\frac{d\ell(\theta)}{d\theta} &= \left(\sum_{i=1}^{N} y_i\right)\frac{1}{\theta} - \left(N - \sum_{i=1}^{N} y_i\right)\frac{1}{1-\theta} := 0 \\ \sum_{i=1}^{N} y_i - \hat{\theta}_{\mathrm{ML}} \sum_{i=1}^{N} y_i &= N\hat{\theta}_{\mathrm{ML}} - \hat{\theta}_{\mathrm{ML}} \sum_{i=1}^{N} y_i \\ \hat{\theta}_{\mathrm{ML}} &= \frac{\sum_{i=1}^{N} y_i}{N}. \end{aligned} \tag{4}$$

(It is easy to verify that this is indeed a maximum). In this case, the maximum likelihood estimate is exactly what we intuitively thought we should do: estimate θ as the observed proportion of Heads.

2.2 Maximum a posteriori estimation

The MAP estimate is a pointwise estimate with a Bayesian flavor. Rather than finding θ that maximizes the likelihood function, we find θ that maximizes the posterior. We don't have to worry about evaluating the partition function because it is constant with respect to θ. Again it is generally more convenient to work with the logarithm.

$$\begin{aligned}\hat{\theta}_{\mathrm{MAP}} \in \arg\max_{\theta} p(\theta|y) &= \arg\max_{\theta} \frac{p(y|\theta)p(\theta)}{\int p(y|\theta')p(\theta')d\theta'} \\ &= \arg\max_{\theta} p(y|\theta)p(\theta) = \arg\max_{\theta} \left(\log p(y|\theta) + \log p(\theta)\right). \end{aligned} \tag{5}$$

When the prior is uniform, the MAP estimate is identical to the ML estimate. For a reasonable choice of the prior (*i.e.*, one that does not assign zero probability to the true value of θ), the MAP estimate is consistent, a fact that we will discuss in more detail later. Some other properties of the MAP estimate are illustrated in the next example.

Coin flip example part 3. We again return to the coin flip example. Suppose we model θ using a Beta prior (we will see later why this is a good idea): $\theta \sim \mathrm{Beta}(\alpha, \beta)$. Recall the Beta distribution is,

$$\mathrm{Beta}(\theta; \alpha, \beta) = \frac{1}{B(\alpha, \beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1},$$

where $B(\alpha, \beta)$ is the beta function, and is constant with respect to θ. The quantities α and β are parameters of the prior which we are free to set according to our prior belief about θ. The MAP estimate for θ is then,

$$\hat{\theta}_{\mathrm{MAP}} \in \arg\max_{\theta} \left(\sum_{i=1}^{N} y_i\right)\log\theta + \left(N - \sum_{i=1}^{N} y_i\right)\log(1-\theta) + (\alpha-1)\log\theta + (\beta-1)\log(1-\theta) - \log B(\alpha, \beta).$$

Differentiating and setting to zero, we obtain that,

$$\hat{\theta}_{\mathrm{MAP}} = \frac{\sum_{i=1}^{N} y_i + \alpha - 1}{N + \beta - 1 + \alpha - 1}. \tag{6}$$

This is a very nice result which illustrates some interesting properties of the MAP estimate. In particular, comparing the MAP estimate in Equation 6 to the ML estimate in Equation 4, we see that the MAP estimate is equivalent to the ML estimate of a data set with $\alpha - 1$ additional Heads and $\beta - 1$ additional Tails. When we specify, for example, a prior of $\alpha = 7$ and $\beta = 3$, it is literally as if we had begun the coin tossing experiment with 6 Heads and 2 Tails on the record. This can be very useful in reducing the variance of the estimate for small samples. For example, suppose the data contain only one coin flip, a Heads. The ML estimate will be $\hat{\theta}_{ML} = 1$, which predicts that we will never flip tails! However, we, the modeler, know that the coin is probably fair, and can assign $\alpha = \beta = 2$ (or some other number with $\alpha = \beta$), and we get

$\hat{\theta}_{MAP} = 2/3$. The MAP estimate has smaller variance for small samples, and for large samples it is easy to see in this example that the effect of the prior goes to zero:

$$\lim_{N\to\infty} \hat{\theta}_{\text{MAP}} = \lim_{N\to\infty} \hat{\theta}_{\text{ML}} = \theta_{\text{true}}.$$

A useful interpretation of the MAP estimate is as a regularized version of the ML estimate.

Example 1. (Rare Events) The MAP estimate is also useful when dealing with rare events. Suppose we are the administrator for a network and we wish to estimate the probability that a unit of traffic is a network intrusion. Perhaps we monitor the traffic for a day, and there are no network intrusions. The ML estimate tells us that the probability of a network intrusion is zero. The MAP estimate would allow us to incorporate our prior knowledge that there is some probability of network intrusion, we just haven't seen one yet.

2.3 Point estimation and probabilistic linear regression

We will now apply point estimation to a slightly more interesting problem, linear regression, and on the way will discover some very elegant connections between agnostic machine learning algorithms and our new probabilistic approach. Suppose now that we have N data points $(x_1, y_1), \ldots, (x_N, y_N)$, where $x_i \in \mathbb{R}^d$ are the independent variables and $y_i \in \mathbb{R}$ are the dependent variables. We will use a likelihood model under which y_i depends linearly on x_i, and the y_i's are all independent. Specifically, we model,

$$y_i \sim \theta^T x_i + \epsilon,$$

where $\theta \in \mathbb{R}^d$ are the parameters we are interested in, and ϵ represents noise. We will assume that the noise is distributed normally with zero mean and some known variance σ^2: $\epsilon \sim \mathcal{N}(0, \sigma^2)$. This, together with our assumption of independence, allows us to express the likelihood function for the observations $y = \{y_1, \ldots, y_N\}$:

$$\begin{aligned} p(y|x,\theta) &= \prod_{i=1}^{N} p(y_i|x_i,\theta) = \prod_{i=1}^{N} \mathcal{N}(y_i; \theta^T x_i, \sigma^2) \\ &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \theta^T x_i)^2\right) \end{aligned}$$

As usual, it is more convenient to work with the log-likelihood:

$$\ell(\theta) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \theta^T x_i)^2$$

The ML estimate is,

$$\begin{aligned} \hat{\theta}_{\text{ML}} \in \arg\max_\theta \ell(\theta) &= \arg\max_\theta -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \theta^T x_i)^2 \\ &= \arg\min_\theta \sum_{i=1}^{N}(y_i - \theta^T x_i)^2. \end{aligned} \tag{7}$$

The ML estimate is equivalent to ordinary least squares! Now let us try the MAP estimate. We saw in our coin toss example that the MAP estimate acts as a regularizer to the ML estimate. For probabilistic regression, we will use a multivariate normal prior (we will see later why this is a good idea) with mean 0 and introducing a new parameter λ that will contribute to the variance:

$$\theta \sim \mathcal{N}(0, I\sigma^2/\lambda) = \frac{1}{(2\pi)^{d/2}|I\sigma^2/\lambda|^{1/2}} \exp\left(-\frac{1}{2}\theta^T \left(I\sigma^2/\lambda\right)^{-1}\theta\right).$$

Now,

$$\begin{aligned}
\hat{\theta}_{\text{MAP}} &\in \arg\max_\theta \log p(y|x,\theta) + \log p(\theta) \\
&= \arg\max_\theta -\frac{1}{2\sigma^2}\sum_{i=1}^N (y_i - \theta^T x_i)^2 - \frac{d}{2}\log 2\pi - \frac{1}{2}\log\frac{\sigma^2}{\lambda} - \frac{1}{2}\theta^T(I\lambda/\sigma^2)\theta \\
&= \arg\max_\theta -\frac{1}{2\sigma^2}\sum_{i=1}^N (y_i - \theta^T x_i)^2 - \frac{1}{2\sigma^2}\lambda\theta^T\theta \\
&= \arg\min_\theta \sum_{i=1}^N (y_i - \theta^T x_i)^2 + \lambda||\theta||_2^2. \qquad (8)
\end{aligned}$$

We see that the MAP estimate corresponds exactly to ℓ_2-regularized linear regression, and that the ℓ_2 regularization can be interpreted as a Gaussian prior. Increasing λ corresponds to increasing our certainty that θ should be close to zero.

3 Conjugate priors

Although point estimates can be useful in many circumstances (and are used in many circumstances), our true goal in Bayesian analysis is the full posterior, $p(\theta|y)$. We saw earlier that this can be obtained in principle from the prior and the likelihood using Bayes' rule, but that there is an integral in the denominator which often makes this intractable. One approach to circumventing the integral is to use conjugate priors.

The appropriate likelihood function (Binomial, Gaussian, Poisson, Bernoulli,...) is typically clear from the data. However, there is a great deal of flexibility when choosing the prior distribution. The key notion of conjugate priors is that if we choose the 'right' prior for a particular likelihood function, then we can compute the posterior without worrying about the integrating. We will formalize the notion of conjugate priors and then see why they are useful.

Definition 2. Let $\mathcal{F}$ be a family of likelihood functions and $\mathcal{P}$ a family of prior distributions. $\mathcal{P}$ is a *conjugate prior* to $\mathcal{F}$ if for any likelihood function $f \in \mathcal{F}$ and for any prior distribution $p \in \mathcal{P}$, the corresponding posterior distribution p^* satisfies $p^* \in \mathcal{P}$.

It is easy to find the posterior when using conjugate priors because we know it must belong to the same family of distributions as the prior.

Coin flip example part 4. We return yet again to the coin flip example. In our previous example, we were very wise to use a Beta prior for θ because the Beta distribution is the conjugate prior to the Bernoulli distribution. Let us see what happens when we compute the posterior:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_0^1 p(y|\theta')p(\theta')d\theta'} = \frac{\theta^{\sum_{i=1}^N y_i+\alpha-1}(1-\theta)^{N-\sum_{i=1}^N y_i+\beta-1}}{\int_0^1 \theta'^{\sum_{i=1}^N y_i+\alpha-1}(1-\theta')^{N-\sum_{i=1}^N y_i+\beta-1}d\theta'}.$$

We can recognize the Beta integral in the denominator,

$$p(\theta|y) == \frac{1}{B(\sum_{i=1}^N y_i+\alpha, N-\sum_{i=1}^N y_i+\beta)}\theta^{\sum_{i=1}^N y_i+\alpha-1}(1-\theta)^{N-\sum_{i=1}^N y_i+\beta-1},$$

and we recognize this as being a Beta distribution:

$$p(\theta|y) \sim \text{Beta}\left(\sum_{i=1}^N y_i+\alpha, N-\sum_{i=1}^N y_i+\beta\right). \qquad (9)$$

As with the MAP estimate, we can see the interplay between the data y_i and the prior parameters α and β in forming the posterior. As before, the exact choice of α and β does not matter asymptotically, the data overwhelm the prior.

Likelihood	**Conjugate prior**	**Prior hyperparameters**	**Posterior hyperparameters**
Bernoulli	Beta	α, β	$\alpha + \sum_{i=1}^{N} y_i$, $\beta + N - \sum_{i=1}^{N} y_i$
Binomial	Beta	α, β	$\alpha + \sum_{i=1}^{N} y_i$, $\beta + \sum_{i=1}^{N} N_i - \sum_{i=1}^{N} y_i$
Poisson	Gamma	α, β	$\alpha + \sum_{i=1}^{N} y_i$, $\beta + N$
Geometric	Beta	α, β	$\alpha + N$, $\beta + \sum_{i=1}^{N} y_i$
Uniform on $[0, \theta]$	Pareto	x_m, k	$\max\{\max y_i, x_m\}$, $k + n$
Exponential	Gamma	α, β	$\alpha + N$, $\beta + \sum_{i=1}^{N} y_i$
Normal, unknown mean, known variance σ^2	Normal	μ_0, σ_0^2	$\left(\frac{\mu_0}{\sigma_0^2} + \frac{1}{\sigma^2}\sum_{i=1}^{N} y_i\right) / \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right), \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}$

Figure 1: A summary of some useful conjugate priors.

Knowing that the Beta distribution is the conjugate prior to the Bernoulli, we could have saved a few steps in the above by recognizing that,

$$p(\theta|y) \propto p(y|\theta)p(\theta) = \theta^{\sum_{i=1}^{N} y_i + \alpha - 1}(1-\theta)^{N - \sum_{i=1}^{N} y_i + \beta - 1}, \tag{10}$$

and then realizing that by the definition of a conjugate prior, the posterior must be a Beta distribution. There is exactly one Beta distribution that satisfies Equation 10, and that is Equation 9.

The parameters of the prior distribution (α and β in the case of the Beta prior) are called *prior hyperparameters*. We choose them to best represent our beliefs about the distribution of θ. The parameters of the posterior distribution are often called *posterior hyperparameters*. Anytime a likelihood model is used together with its conjugate prior, we know the posterior is from the same family of the prior, and moreover we have an explicit formula for the posterior hyperparameters. A table summarizing some of the useful conjugate prior relationships is in Figure 1. There are many more conjugate prior relationships that are not shown in Figure 1 but that can be found in reference books on Bayesian statistics[1].

3.1 Exponential families and conjugate priors

There is an important connection between exponential families (not to be confused with the exponential distribution) and conjugate priors.

Definition 3. The family of distributions $\mathcal{F}$ is an *exponential family* if every member of $\mathcal{F}$ has the form:

$$p(y_i|\theta) = f(y_i)g(\theta)\exp\left(\phi(\theta)^T u(y_i)\right), \tag{11}$$

for some $f(\cdot), g(\cdot), \phi(\cdot)$, and $u(\cdot)$.

Essentially all of the distributions that we typically work with (normal, exponential, Poisson, beta, gamma, binomial, Bernoulli,...) are exponential families. The next theorem tells us when we can expect to have a conjugate prior.

Theorem 1. *If the likelihood model is an exponential family, then there exists a conjugate prior.*

[1] *Bayesian Data Analysis* by Gelman, Carlin, Stern, and Rubin is an excellent choice, and is the source for some of the material in these notes.

Proof. Consider the likelihood of our iid data $y = \{y_1, \ldots, y_N\}$:

$$p(y|\theta) = \prod_{i=1}^{N} p(y|\theta) = \left[\prod_{i=1}^{N} f(y_i)\right] g(\theta)^N \exp\left(\phi(\theta)^T \sum_{i=1}^{N} u(y_i)\right)$$

Take the prior distribution to be:

$$p(\theta) = \frac{g(\theta)^\eta \exp\left(\phi(\theta)^T \nu\right)}{\int g(\theta')^\eta \exp\left(\phi(\theta')^T \nu\right) d\theta'}, \tag{12}$$

where η and ν are prior hyperparameters. Then, the posterior will be,

$$\begin{aligned} p(\theta|y) \propto p(y|\theta)p(\theta) &= \left[\prod_{i=1}^{N} f(y_i)\right] g(\theta)^N \exp\left(\phi(\theta)^T \sum_{i=1}^{N} u(y_i)\right) \frac{g(\theta)^\eta \exp\left(\phi(\theta)^T \nu\right)}{\int g(\theta')^\eta \exp\left(\phi(\theta')^T \nu\right) d\theta'} \\ &\propto g(\theta)^{\eta+N} \exp\left(\phi(\theta)^T \left(\nu + \sum_{i=1}^{N} u(y_i)\right)\right), \end{aligned}$$

which is in the same family as the prior, with the posterior hyperparameters being $\eta + N$ and $\nu + \sum_{i=1}^{N} u(y_i)$. □

Although this proof yields the form of the conjugate prior, we may not always be able to compute the partition function. In these cases, the result is for practical purposes existential. It turns out that in general, the converse to Theorem 1 is true, and exponential families are the only distributions with (non-trivial) conjugate priors.

Coin flip example part 5. Returning again to the coin flip example, let us first verify that the Bernoulli distribution is an exponential family:

$$\begin{aligned} p(y_i|\theta) &= \theta^{y_i}(1-\theta)^{1-y_i} \\ &= \exp\left(y_i \log\theta + (1-y_i)\log(1-\theta)\right) \\ &= \exp\left(y_i \log\theta - y_i \log(1-\theta) + \log(1-\theta)\right) \\ &= (1-\theta)\exp\left(y_i \log\frac{\theta}{1-\theta}\right), \end{aligned}$$

and we see that the Bernoulli distribution is an exponential family according to Equation 11 with $f(y_i) = 1$, $g(\theta) = 1 - \theta$, $u(y_i) = y_i$, and $\phi(\theta) = \log\frac{\theta}{1-\theta}$. Thus, according to Equation 12, the conjugate prior is,

$$\begin{aligned} p(\theta) &\propto g(\theta)^\eta \exp\left(\phi(\theta)^T \nu\right) \\ &= (1-\theta)^\eta \exp\left(\nu \log\frac{\theta}{1-\theta}\right) \\ &= (1-\theta)^\eta \left(\frac{\theta}{1-\theta}\right)^\nu \\ &= \theta^\nu (1-\theta)^{\eta-\nu}. \end{aligned}$$

Reparameterizing with $\alpha = \nu + 1$ and $\beta = \eta - \nu + 1$ gives us the Beta distribution that we expect.

4 Posterior asymptotics

Up to this point, we have defined a likelihood model that is parameterized by θ, assigned a prior distribution to θ, and then computed the posterior $p(\theta|y)$. There are two natural questions which arise. First, what if we choose the 'wrong' likelihood model? That is, what if the data were actually generated by some distribution $q(y)$ such that $q(y) \neq p(y|\theta)$ for any θ, but we use $p(y|\theta)$ as our likelihood model? Second, what if we assign

the 'wrong' prior? We can answer both of these questions asymptotically as $N \to \infty$. First we must develop a little machinery from information theory.

A useful way to measure the dissimilarity between two probability distributions is the *Kullback-Leibler (KL) divergence*, defined for two distributions $p(y)$ and $q(y)$ as:

$$D(q(\cdot)||p(\cdot)) := \mathbb{E}_{y\sim q(y)}\left[\log \frac{q(y)}{p(y)}\right] = \int q(y) \log \frac{q(y)}{p(y)} dy.$$

This is sometimes referred to as the KL distance, however it is not a metric in the mathematical sense because in general it is asymmetric: $D(q(\cdot)||p(\cdot)) \neq D(p(\cdot)||q(\cdot))$. The following property of the KL divergence will be very important for us.

Theorem 2. *$D(q(\cdot)||p(\cdot)) \geq 0$ with equality if and only if $q(y) = p(y)\,\forall y$.*

Proof. We will rely on Jensen's inequality, which states that for any convex function f and random variable X,

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]).$$

When f is strictly convex, Jensen's inequality holds with equality if and only if X is constant. Take $y \sim q(y)$ and define the random variable $X = \frac{p(y)}{q(y)}$. Let $f(X) = -\log(X)$, a strictly convex function. Now we can apply Jensen's inequality:

$$\begin{aligned}
\mathbb{E}[f(X)] \geq & f(\mathbb{E}[X]) \\
-\int q(y) \log \frac{p(y)}{q(y)} dy \geq & -\log\left(\int q(y) \frac{p(y)}{q(y)} dy\right) \\
-\int q(y) \log \frac{p(y)}{q(y)} dy \geq & -\log\left(\int p(y) dy\right) \\
-\int q(y) \log \frac{p(y)}{q(y)} dy \geq & -\log 1 \\
\int q(y) \log \frac{q(y)}{p(y)} dy \geq & 0,
\end{aligned}$$

with equality under the same conditions required for equality in Jensen's inequality: if and only if X is constant, that is, $q(y) = p(y)\,\forall y$. □

We will use the KL divergence to find the distribution from the likelihood family that is 'closest' to the true generating distribution:

$$\theta^* = \arg\min_{\theta\in\Theta} D(q(\cdot)||p(\cdot|\theta)). \tag{13}$$

For convenience, we will suppose that the arg min in Equation 13 is unique, as this is typically the case. The results can be easily extended to the case where the arg min is not unique. The main results of this section are two theorems, the first for discrete parameter spaces and the second for continuous parameter spaces.

Theorem 3. *Let $\mathcal{F}$ be a finite family of likelihood models, with $\Theta = \{\theta : p(\cdot|\theta) \in \mathcal{F}\}$ the discrete parameter space. Let $y = (y_1, \ldots, y_N)$ be a set of independent samples from an arbitrary distribution $q(\cdot)$, and θ^* be as in Equation 13. If $p(\theta = \theta^*) > 0$, then $p(\theta = \theta^*|y) \to 1$ as $N \to \infty$.*

Proof. Consider any $\theta \neq \theta^*$:

$$\log \frac{p(\theta|y)}{p(\theta^*|y)} = \log \frac{p(\theta)p(y|\theta)}{p(\theta^*)p(y|\theta^*)} = \log \frac{p(\theta)}{p(\theta^*)} + \sum_{i=1}^{n} \log \frac{p(y_i|\theta)}{p(y_i|\theta^*)}$$

For each term in the sum,

$$\begin{aligned}\mathbb{E}\left[\log \frac{p(y_i|\theta)}{p(y_i|\theta^*)}\right] &= \mathbb{E}\left[\log \frac{p(y_i|\theta)q(y_i)}{p(y_i|\theta^*)q(y_i)}\right] \\ &= \mathbb{E}\left[\log \frac{q(y_i)}{p(y_i|\theta^*)} - \log \frac{q(y_i)}{p(y_i|\theta)}\right] \\ &= D(q(\cdot)||p(\cdot|\theta^*)) - D(q(\cdot)||p(\cdot|\theta)) \\ &< 0,\end{aligned}$$

from the definition of θ^* in Equation 13. By the strong law of large numbers, with probability 1,

$$\sum_{i=1}^{n} \log \frac{p(y_i|\theta)}{p(y_i|\theta^*)} \to n\mathbb{E}\left[\log \frac{p(y_i|\theta)}{p(y_i|\theta^*)}\right] = -\infty.$$

By supposition, $p(\theta^*) > 0$. Thus,

$$\log \frac{p(\theta)}{p(\theta^*)} + \sum_{i=1}^{n} \log \frac{p(y_i|\theta)}{p(y_i|\theta^*)} \to -\infty,$$

and,

$$\log \frac{p(\theta|y)}{p(\theta^*|y)} \to -\infty \text{ implies } p(\theta|y) \to 0.$$

This holds for every $\theta \neq \theta^*$, thus $p(\theta^*|y) \to 1$. □

This theorem tells us that the posterior eventually becomes concentrated on the value θ^* corresponding to the likelihood model that is 'closest' in the KL sense to the true generating distribution $q(\cdot)$. If $q(\cdot) = p(\cdot|\theta_0)$ for some $\theta_0 \in \Theta$, then $\theta^* = \theta_0$ is the unique minimizer of Equation 13 and the theorem tells us that the posterior will become concentrated around the true value. This is only the case if in the prior, $p(\theta^*) > 0$, which shows the importance of choosing a prior that assigns non-zero probability to every plausible value of θ.

We now present the continuous version of Theorem 3, but the proof is quite technical and is omitted.

Theorem 4. *If Θ is a compact set, θ^* is defined as in Equation 13, A is a neighborhood of θ^*, and $p(\theta^* \in A) > 0$, then $p(\theta \in A|y) \to 1$ as $N \to \infty$.*

These theorems show that asymptotically, the choice of the prior does not matter as long as it assigns non-zero probability to every $\theta \in \Theta$. They also show that if the data were generated by a member of the family of likelihood models, we will converge to the correct likelihood model. If not, then we will converge to the model that is 'closest' in the KL sense.

We give these theorems along with a word of caution. These are asymptotic results that tell us absolutely nothing about the sort of N we encounter in practical applications. For small sample sizes, poor choices of the prior or likelihood model can yield poor results and we must be cautious.

Coin flip example part 6. Suppose that the coin flip data from previous examples came from a biased coin with a 3/4 probability of Heads, but we restrict the likelihood model to only include coins with probability in the interval $[0, 1/2]$:

$$p(y|\theta) = \theta^{\sum_{i=1}^{N} y_i}(1-\theta)^{(N-\sum_{i=1}^{N} y_i)}, \quad \theta \in [0, 1/2].$$

This time we will use a uniform prior, $p(\theta) = 2, \theta \in [0, 1/2]$. The posterior distribution is then,

$$p(\theta|y) = \frac{\theta^{\sum_{i=1}^{N} y_i}(1-\theta)^{N-\sum_{i=1}^{N} y_i}}{\int_0^{1/2} \theta'^{\sum_{i=1}^{N} y_i}(1-\theta')^{N-\sum_{i=1}^{N} y_i} d\theta'}.$$

The partition function is an incomplete Beta integral, which has no closed form but can easily be solved numerically. In Figure 2, we draw samples iid from the true distribution (3/4 probability of Heads) and show how the posterior becomes increasingly concentrated around $\theta = 0.5$, the closest likelihood function to the true generating distribution.

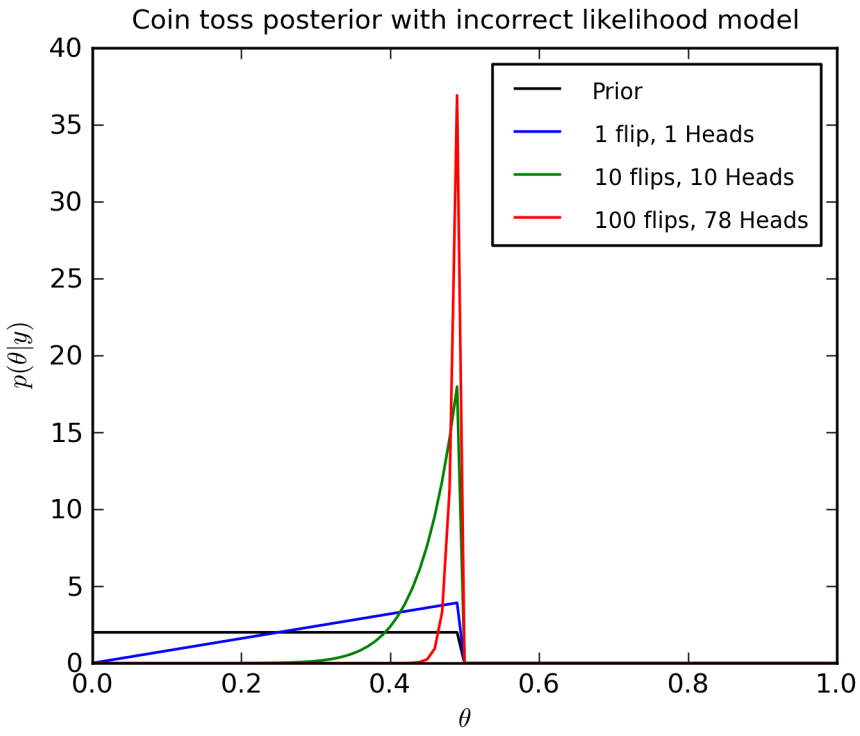


Figure 2: The posterior for a coin toss likelihood model that restricts θ to the interval $[0, 1/2]$ applied to data generated by a coin with $\theta = 3/4$.

5 Hierarchical modeling

Hierarchical models are a powerful application of Bayesian analysis. The general idea is to construct a generative model for the data under which unobserved model components are related in a hierarchical manner via conditional probability distributions. This is best explained by example, so we will develop a hierarchical model for *topic modeling*, an important information retrieval problem.

5.1 Topic models

Suppose we have been given a collection of m documents and we wish to determine how the documents are related. For example, if the documents are all news articles, the article "Patriots game canceled due to hurricane" is related to the article "New York Giants lose superbowl" because they are both about football. The article "Patriots game canceled due to hurricane" is also related to the article "Record snowfall in May" because they are both about the weather. We will now develop a hierarchical model for finding topic relationships between documents in an unsupervised setting. The method is called *Latent Dirichlet Allocation* (LDA) and it was developed by, among others, the well-known statistician Michael Jordan.

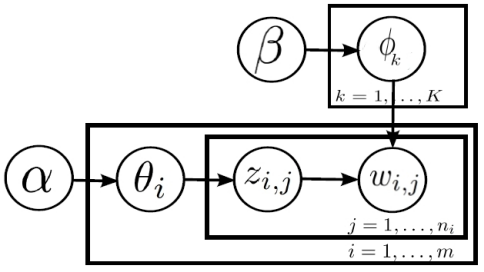
5.1.1 LDA formulation

The model has several components. The data are m documents, with document i consisting of n_i words. Each word in the document will be associated with one of K topics. We let $z_{i,j}$ denote the topic of word j in document i. We model $z_{i,j} \sim \text{Multinomial}(\theta_i)$, where $\theta_i \in \mathbb{R}^K$ describes the topic mixture of document i. For each topic, we define a multinomial distribution over all possible words. For example, given the topic is "Sports", the probability of having the word "football" might be high; if the topic were "Weather", the probability of having the word "football" might be lower. Other words, like "the" will have a high probability regardless of the topic. If words are chosen from a set of W possible words, then we let $\phi_k \in \mathbb{R}^W$ be the multinomial parameter over words for topic k. Word j of document i, denoted $w_{i,j}$, will be generated by the distribution over words corresponding to the topic $z_{i,j}$: $w_{i,j} \sim \text{Multinomial}(\phi_{z_{i,j}})$. Finally, we give prior distributions for the parameters θ_i and ϕ_k. The multinomial distribution is a generalization of the binomial distribution, and its conjugate prior is a generalization of the beta distribution: the Dirichlet distribution. Thus we model the data with the following generative model:

1. For document $i = 1, \ldots, m$, choose the document's topic distribution $\theta_i \sim \text{Dirichlet}(\alpha)$, where $\alpha \in \mathbb{R}^K$ is the prior hyperparameter.
2. For topic $k = 1, \ldots, K$, choose the topic's word distribution $\phi_k \sim \text{Dirichlet}(\beta)$, where $\beta \in \mathbb{R}^W$ is the prior hyperparameter.
3. For document $i = 1, \ldots, m$:
   For word $j = 1, \ldots, n_i$:
   Choose the topic for this word $z_{i,j} \sim \text{Multinomial}(\theta_i)$.
   Choose the word $w_{i,j} \sim \text{Multinomial}(\phi_{z_{i,j}})$.

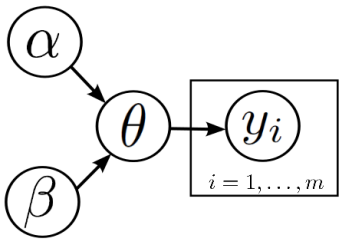
5.2 Graphical representation

Hierarchical models are illustrated with a node for every variable and arcs between nodes to indicate the dependence between variables. We give the graphical representation for LDA and then describe it:



Graphs representing hierarchical models must be acyclic. For any node x, we define Parents(x) as the set of all nodes with arcs to x. The hierarchical model consists of, for every node x, the distribution $p(x|\text{Parents}(x))$. Define Descendants(x) as all nodes that can be reached from x and Non-descendants(x) as all other nodes. Because the graph is acyclic and the distribution for each node depends only on its parents, given Parents(x), x is conditionally independent from Non-descendants(x). This is a powerful fact about hierarchical models that is important for doing inference. In the graph for LDA, this means that, for example, $z_{i,j}$ is independent of α, given θ_i. In addition to the graph structure, we use plates to denote repeated, independent draws from the same distribution. The hierarchical nature of the model is clear from the graphical representation.

Coin flip example part 7. Even simple models like the coin flip model can be represented graphically. The by-now-very-familiar coin flip model is:



5.3 Inference in hierarchical models

Hierarchical models are useful and powerful because they allow us to model complex interactions between the observed variables (in this case the words in each document) through the use of latent (or hidden) variables. Despite introducing a larger number of latent variables than we have observed variables, because of their additional structure hierarchical models are generally not as prone to overfitting as, for instance, regression is.

We are interested in infering the posterior distribution for the latent variables. Let $Z = \{z_{i,j}\}_{i=1,\ldots,m,j=1,\ldots,n_i}$, $\theta = \{\theta_i\}_{i=1,\ldots,m}$, $\phi = \{\phi_k\}_{k=1,\ldots,K}$, and $W = \{w_{i,j}\}_{i=1,\ldots,m,j=1,\ldots,n_i}$. Then, by Bayes' rule, ignoring the constant denominator, we can express the posterior as:

$$p(Z, \theta, \phi|w, \alpha, \beta) \propto p(w|Z, \phi, \theta, \alpha, \beta)p(Z, \theta, \phi|\alpha, \beta) \tag{14}$$

We will look at each of these pieces and show that they have a compact analytical form.

$$\begin{aligned} p(w|Z, \phi, \theta, \alpha, \beta) &= \prod_{i=1}^{m}\prod_{j=1}^{n_i} p(w_{i,j}|Z, \phi, \theta, \alpha, \beta) \quad \text{By iid} \\ &= \prod_{i=1}^{m}\prod_{j=1}^{n_i} p(w_{i,j}|z_{i,j}, \phi) \quad \text{By conditional independence} \\ &= \prod_{i=1}^{m}\prod_{j=1}^{n_i} \text{Multinomial}(w_{i,j}; \phi_{z_{i,j}}) \quad \text{By definition.} \end{aligned}$$

Also,

$$p(Z, \theta, \phi|\alpha, \beta) = p(Z, \theta|\alpha)p(\phi|\beta)$$

by conditional independence. Again considering each term,

$$p(Z, \theta|\alpha) = p(Z|\theta, \alpha)p(\theta|\alpha),$$

where,

$$p(Z|\theta, \alpha) = p(Z|\theta) = \prod_{i=i}^{m}\prod_{j=1}^{n_i} p(z_{i,j}|\theta_i) = \prod_{i=i}^{m}\prod_{j=1}^{n_i} \text{Multinomial}(z_{i,j}; \theta_i),$$

by conditional independence, iid, and definition as before. Further,

$$p(\theta|\alpha) = \prod_{i=1}^{m} p(\theta_i|\alpha) = \prod_{i=1}^{m} \text{Multinomial}(\theta_i; \alpha),$$

and,

$$p(\phi|\beta) = \prod_{k=1}^{K} p(\phi_k|\beta) = \prod_{k=1}^{K} \text{Multinomial}(\phi_k; \beta).$$

Plugging all of these pieces back into Equation 14, we obtain,

$$p(Z, \theta, \phi|w, \alpha, \beta) \propto \prod_{i=1}^{m} \prod_{j=1}^{n_i} \text{Multinomial}(w_{i,j}; \phi_{z_{i,j}}) \prod_{i=i}^{m} \prod_{j=1}^{n_i} \text{Multinomial}(z_{i,j}; \theta_i) \times$$
$$\prod_{i=1}^{m} \text{Multinomial}(\theta_i; \alpha) \prod_{k=1}^{K} \text{Multinomial}(\phi_k; \beta). \quad (15)$$

For any given $Z, \theta, \phi, w, \alpha$, and β, we can easily evaluate Equation 15. We will see in the next section that this is sufficient to be able to simulate draws from the posterior.

Even using conjugate priors, in general it will not be possible to recover the posterior analytically for hierarchical models of any complexity. We will rely on (among a few other options) sampling methods like the Monte Carlo Markov Chains (MCMC) that we discuss in the next section. What the statistics community call Bayesian hierarchical models are in the machine learning community often treated as a special case of Bayesian graphical models (specifically, directed acyclic graphs).

6 Markov Chain Monte Carlo sampling

With more complex Bayesian models, like the hierarchical models we discuss later, even with conjugate priors we are unable to express the posterior analytically. The reason Bayesian statistics are so widely used is because of the development of computational methods for drawing samples from the posterior distribution. Even though we are unable to express the posterior analytically, with enough samples we can compute statistics of the posterior with arbitrary precision. This approach is called Monte Carlo simulation. We will describe two commonly used monte carlo methods, which both fall under the umbrella of Markov Chain Monte Carlo (MCMC) methods: the Metropolis-Hastings algorithm, and Gibbs' sampling.

6.1 Metropolis-Hastings algorithm

The goal in MCMC is to construct a Markov Chain whose stationary distribution is the posterior $p(\theta|y)$. First we give some definitions and summarize some important properties of Markov chains. A continuous state Markov chain is a sequence of random variables $\theta^1, \theta^2, \ldots$ that satisfies the Markov property: $p(\theta^t|\theta^{t-1}, \ldots, \theta^1) = p(\theta^t|\theta^{t-1})$. The chain is *homogenous* if the probabilities do not change throughout the sequence: $p(\theta^i|\theta^{i-1}) = p(\theta^j|\theta^{j-1})$ for any i and j. We can then define the *transition kernel* to be the probability of transitioning from state θ to θ': $K(\theta, \theta') = p(\theta'|\theta)$. We define $\pi_t(\theta)$ to be the probability distribution over possible states θ at time t: $\pi_t(\theta) = p_{\theta^t}(\theta)$. Then,

$$\pi_{t+1}(\theta) = \int \pi_t(\theta') K(\theta', \theta) d\theta'.$$

Under some conditions that will be satisfied by the chains we are interested in (specifically, irreducible, aperiodic, and non-transient), the distributions $\pi_1, \pi_2, \ldots$ will converge to a unique *stationary distribution* (or equilibrium distribution, or steady-state distribution) π^* that satisfies:

$$\pi^*(\theta) = \int \pi^*(\theta') K(\theta', \theta) d\theta'.$$

A sufficient condition for π^* to be the stationary distribution is the detailed balance equation:

$$K(\theta, \theta')\pi^*(\theta) = K(\theta', \theta)\pi^*(\theta'), \text{ for all } \theta, \theta'. \tag{16}$$

Now we are ready to present the Metropolis-Hastings algorithm. In addition to the distributions we have already used (likelihood and prior), we will need a *proposal distribution* (or jumping distribution) $J_t(\theta, \theta')$ which will propose a new state θ' given the current state θ. There are many options when choosing a proposal distribution which we will discuss later.

Step 1. Choose a starting point θ^0. Set $t = 1$.

Step 2. Sample θ^* from the proposal distribution $J_t(\theta^{t-1}, \cdot)$. $\theta^{t-1} \to \theta^*$ is now the proposed move for time t.

Step 3. Compute the ratio,

$$\alpha(\theta^{t-1}, \theta^*) = \min\left\{\frac{p(\theta^*|y)J_t(\theta^*, \theta^{t-1})}{p(\theta^{t-1}|y)J_t(\theta^{t-1}, \theta^*)}, 1\right\} = \min\left\{\frac{p(y|\theta^*)p(\theta^*)J_t(\theta^*, \theta^{t-1})}{p(y|\theta^{t-1})p(\theta^{t-1})J_t(\theta^{t-1}, \theta^*)}, 1\right\}$$

The fact that we can compute ratios of posterior probabilities without having to worry about the normalization integral is the key to monte carlo methods.

Step 4. With probability $\alpha(\theta^{t-1}, \theta^*)$, accept the move $\theta^{t-1} \to \theta^*$ by setting $\theta^t = \theta^*$ and incrementing $t \leftarrow t+1$. Otherwise, discard θ^*.

Step 5. Until stationary distribution and the desired number of samples are reached, return to **Step 2.**

Because the proposal distribution and $\alpha(\theta^{t-1}, \theta^*)$ depend only on the current state, the sequence $\theta_0, \theta_1, \ldots$ forms a Markov chain. What makes the Metropolis-Hastings algorithm special is the following theorem, which shows that if we sample from the chain long enough, we will sample from the posterior.

Theorem 5. *If $J_t(\theta, \theta')$ is such that that the Markov chain $\theta^0, \theta^1, \ldots$ produced by the Metropolis-Hastings algorithm has a unique stationary distribution, then the stationary distribution is $p(\theta|y)$.*

Proof. We will give a proof for the case where $J_t(\theta, \theta') = J(\theta, \theta')\, \forall t$, so the chain is homogenous. To show that $p(\theta|y)$ is the stationary distribution, it is sufficient to show that it satisfies Equation 16, the detailed balance equation:

$$K(\theta, \theta')p(\theta|y) = K(\theta', \theta)p(\theta'|y), \text{ for all } \theta, \theta'.$$

The transition kernel is,

$$K(\theta, \theta') = J(\theta, \theta')\alpha(\theta, \theta').$$

Take any θ and θ', and without loss of generality, suppose that,

$$J(\theta, \theta')p(\theta|y) \geq J(\theta', \theta)p(\theta'|y).$$

Then,

$$\alpha(\theta, \theta') = \frac{J(\theta', \theta)p(\theta'|y)}{J(\theta, \theta')p(\theta|y)},$$

and $\alpha(\theta', \theta) = 1$. Now,

$$\begin{aligned}
K(\theta, \theta')p(\theta|y) =& J(\theta, \theta')\alpha(\theta, \theta')p(\theta|y) \\
=& J(\theta, \theta')p(\theta|y)\frac{J(\theta', \theta)p(\theta'|y)}{J(\theta, \theta')p(\theta|y)} \\
=& J(\theta', \theta)p(\theta'|y) \\
=& J(\theta', \theta)\alpha(\theta', \theta)p(\theta'|y) \\
=& K(\theta', \theta)p(\theta'|y)
\end{aligned}$$

□

We have now proven that the Metropolis-Hastings algorithm will eventually yield samples from the posterior distribution. However, there are a number of important questions to be addressed. What proposal distribution should we use? How many samples will it take for the chain to converge to the stationary distribution? How will we know when the chain has reached its stationary distribution? We will discuss these important issues after we provide some intuition into the Metropolis-Hastings algorithm and introduce the Gibbs' sampler.

6.1.1 Some intuition into the Metropolis-Hastings algorithm

To gain some intuition for the Metropolis-Hastings algorithm, consider the slightly more simple case when the proposal distribution is symmetric: $J_t(\theta, \theta') = J_t(\theta', \theta)$. Then, we accept the move $\theta \to \theta'$ with probability,

$$\alpha(\theta, \theta') = \min\left\{\frac{p(\theta'|y)}{p(\theta|y)}, 1\right\}.$$

Consider now the two cases. If $\alpha(\theta, \theta') = 1$, then the posterior value for θ' is larger than the posterior value for θ. Thus, for every accepted sample θ, we should have at least as many accepted samples θ', and so we always accept the transition $\theta \to \theta'$. On the other hand, if $\alpha(\theta, \theta') < 1$, then $\alpha(\theta, \theta')$ is the ratio of the posterior value of θ' to that of θ. For every accepted sample θ, we should have on average $\alpha(\theta, \theta')$ accepted values of θ', and so we accept the transition with probability $\alpha(\theta, \theta')$.

6.2 Gibbs' Sampler

The Gibbs' sampler is a very powerful MCMC sampling technique for the special situation when we have access to conditional distributions. Let us express $\theta \in \mathbb{R}^d$ as $\theta = [\theta_1, \ldots, \theta_d]$. Suppose that even though we are not able to sample directly from $p(\theta|y)$, we are able to sample from the conditional distribution $p(\theta_j|\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_d, y)$. It turns out there are many situations like this when working with hierarchical models, which we will discuss later. The Gibbs' sampler is the following very natural algorithm:

Step 1. Initialize $\theta^0 = [\theta_1^0, \ldots, \theta_d^0]$. Set $t = 1$.

Step 2. For $j \in \{1, \ldots, d\}$, sample θ_j^t from $p(\theta_j|\theta_1^t, \ldots, \theta_{j-1}^t, \theta_{j+1}^{t-1}, \ldots, \theta_d^{t-1}, y)$.

Step 3. Until stationary distribution and the desired number of samples are reached, increment $t \leftarrow t + 1$ and return to **Step 2.**

In each iteration of the Gibbs' sampler, we sequentially update each component of θ^t. We could do that updating in any order, it does not have to be $1, \ldots, d$. To see that the Gibbs' sampler will eventually yield samples from the posterior $p(\theta|y)$, we will show that it is a special case of the Metropolis-Hastings algorithm. The Gibbs' sampler performs d component updates for every time step. we will treat each component update as a time step in the Metropolis-Hastings algorithm: $t = d(t' - 1) + j$, where $j = 1, \ldots, d$ represents the component we are currently updating and t' is the equivalent time step in the Gibbs' sampler. We will set the proposal distribution to ensure that θ^* differs from θ^{t-1} only on the j'th component:

$$J_t(\theta^{t-1}, \theta^*) = J_{d(t'-1)+j}(\theta^{t-1}, \theta^*) := \begin{cases} p(\theta_j^*|\theta_{i\neq j}^{t-1}, y) & \text{if } \theta_i^* = \theta_i^{t-1}\ \forall i \neq j \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

Then, for any θ^* that is proposed, the probability of accepting the move is,

$$\begin{aligned}
\alpha(\theta^{t-1}, \theta^*) =& \frac{p(\theta^*|y)/J_t(\theta^{t-1}, \theta^*)}{p(\theta^{t-1}|y)/J_t(\theta^*, \theta^{t-1})} \\
=& \frac{p(\theta^*|y)/p(\theta_j^*|\theta_{i\neq j}^{t-1}, y)}{p(\theta^{t-1}|y)/p(\theta_j^{t-1}|\theta_{i\neq j}^*, y)} \\
=& \frac{p(\theta^*|y)/p(\theta_j^*|\theta_{i\neq j}^*, y)}{p(\theta^{t-1}|y)/p(\theta_j^{t-1}|\theta_{i\neq j}^{t-1}, y)} \text{ because } \theta_{i\neq j}^* = \theta_{i\neq j}^{t-1} \\
=& \frac{p(\theta_{i\neq j}^*|y)}{p(\theta_{i\neq j}^{t-1}|y)} \\
=& 1.
\end{aligned}$$

Thus every Metropolis-Hastings step is accepted, and the Metropolis-Hastings algorithm with the proposal distribution in Equation 17 is equivalent to the Gibbs' sampler. This result (along with some properties of the proposal distribution) guarantees that the Gibbs' sampler will converge to the posterior $p(\theta|y)$.

6.3 Practical considerations

Now that we have seen the general idea of MCMC algorithms and some theory behind them, let us dive into the details.

6.3.1 Proposal distributions

To guarantee existence of a stationary distribution, all that is required (with rare exceptions) is for the proposal distribution $J_t(\cdot, \cdot)$ to be such that there is a positive probability of eventually reaching any state from any other state. A typical proposal distribution is a random walk $J(\theta, \theta') = \mathcal{N}(\theta, \sigma^2)$ for some σ^2. There are several important features of proposal distributions that work well in practice. First, we must be able to sample from it efficiently. Second, we must be able to compute the ratio $\alpha(\theta, \theta')$ in an efficient way. Third, the jumps should not be too large or we will reject them frequently and the chain will not move quickly. Fourth, the jumps should not be too small or it will take a long time to explore the whole space. The balance between 'too small' and 'too large' is the subject of hundreds of papers on 'adaptive MCMC', but there is really no good way to know which proposal distribution to use. In practice, we often try several proposal distributions to see which is most appropriate, for example, by adjusting σ^2 in the above proposal distribution.

6.3.2 On reaching the stationary distribution

Unfortunately, it is impossible to know how many iterations it will take to reach the stationary distribution, or even to be certain when we have arrived. This is probably the largest flaw in MCMC sampling. There are a large number of heuristics which are used to assess convergence, which generally involve looking at how θ^t varies with time. In general, the initial samples depend strongly on the starting point and are thrown away. This is referred to as *burn-in*, and often involves discarding the first half of all samples. To reduce the effect of autocorrelations between samples, we typically only store a small fraction of them, for example we store only 1 out of every 1000 accepted samples. This process is called *thinning*. Finally, we can assess convergence by comparing the distribution of the first half of stored samples to the distribution of the second half of stored samples. If the chain has reached its stationary distribution, these should be the same. These (and similar) approaches have been shown to be successful in practice. Unfortunately, there is no way to be entirely certain that we are truly drawing samples from the posterior and we must be very cautious.

Coin flip example part 8. Let us return to our coin flip example a final time. We will draw samples from the posterior using the Metropolis-Hastings algorithm. Our model has a scalar parameter $\theta \in [0, 1]$. Our proposal distribution $J_t(\theta^{t-1}, \theta^*)$ will be uniform on an interval of size r around θ^{t-1}:

$$J_t(\theta^{t-1}, \theta^*; r) = \begin{cases} \frac{1}{r} & \text{if } \theta^* \in [\theta^{t-1} \ominus \frac{r}{2}, \theta^{t-1} \oplus \frac{r}{2}] \\ 0 & \text{otherwise} \end{cases}$$

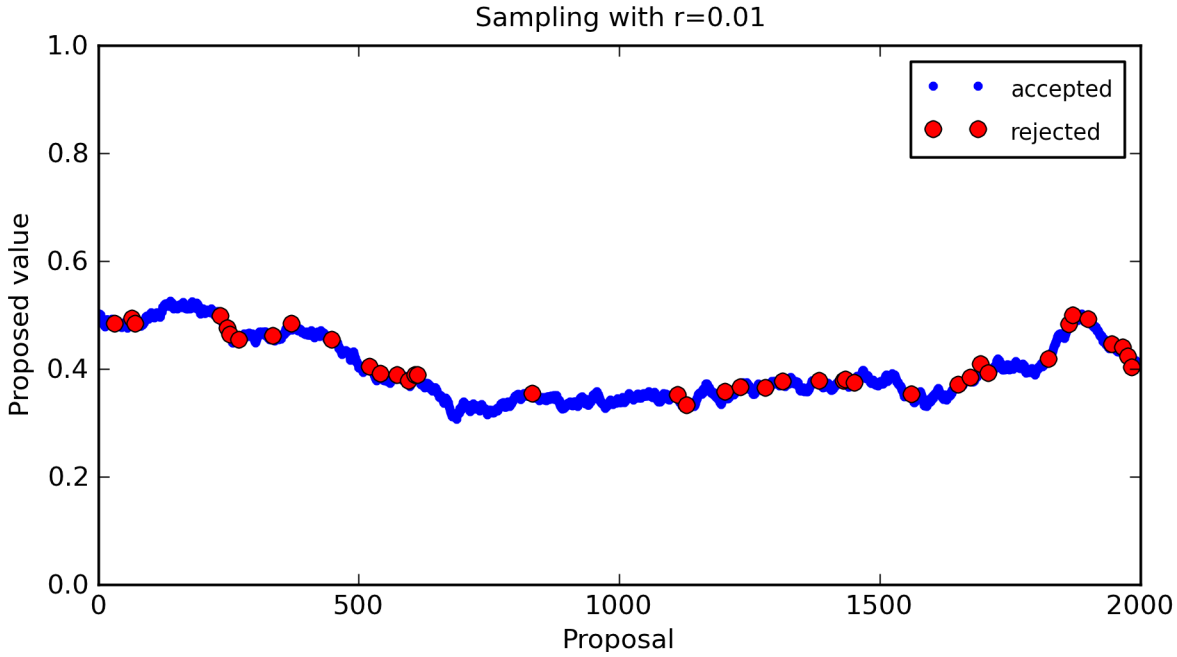


Figure 3: The proposed MCMC samples for the coin toss example with $r = 0.01$.

where $\oplus$ and $\ominus$ represent addition and subtraction on $[0, 1]$, *e.g.*, $0.7 \oplus 0.5 = 0.2$. Notice that this proposal distribution is symmetric: $J_t(\theta^{t-1}, \theta^*; r) = J_t(\theta^*, \theta^{t-1}; r)$. We accept the proposed θ^* with probability,

$$\begin{aligned}
\alpha(\theta^{t-1}, \theta^*) &= \min\left\{\frac{p(y|\theta^*)p(\theta^*)J_t(\theta^*, \theta^{t-1}; r)}{p(y|\theta^{t-1})p(\theta^{t-1})J_t(\theta^{t-1}, \theta^*; r)}, 1\right\} \\
&= \min\left\{\frac{p(y|\theta^*)p(\theta^*)}{p(y|\theta^{t-1})p(\theta^{t-1})}, 1\right\} \\
&= \min\left\{\frac{(\theta^*)^{\sum_{i=1}^N y_i+\alpha-1}(1-\theta^*)^{N-\sum_{i=1}^N y_i+\beta-1}}{(\theta^{t-1})^{\sum_{i=1}^N y_i+\alpha-1}(1-\theta^{t-1})^{N-\sum_{i=1}^N y_i+\beta-1}}, 1\right\},
\end{aligned}$$

which we can easily compute. This formula and a uniform random number generator for the proposal distribution are all that is required to implement the Metropolis-Hastings algorithm. Consider the specific case of $N = 25$, $\sum_{i=1}^N y_i = 6$, and $\alpha = \beta = 5$. Figures 3, 4, and 5 show the proposals θ^* for chains with $r = 0.01$, $r = 0.1$, and $r = 1$ respectively, with the colors indicating whether each proposed θ^* was accepted or not. The chain in Figure 3 shows that with $r = 0.01$, the step sizes are too small and after 2000 proposals we have not reached the stationary distribution. On the other hand, the chain in Figure 5 shows that with $r = 1$ the steps are too large and we reject most of the proposals. This leads to a small number of accepted samples after 2000 proposals. The chain with $r = 0.1$, in Figure 4 is the happy medium, where we rapidly reach the stationary distribution, and accept most of the samples. To compare this to the analytic distribution that we obtained in Equation 9, we run a chain with $r = 0.1$ until we have collected 25,000 accepted samples. We then discard the initial 200 samples (burn-in) and keep one out of every 100 samples from what remains (thinning). A normalized histogram of the resulting samples is compared to the analytical posterior in Figure 6. The running time to generate the MCMC samples in Figure 6 was less than a second, and they are a reasonable approximation to the posterior.

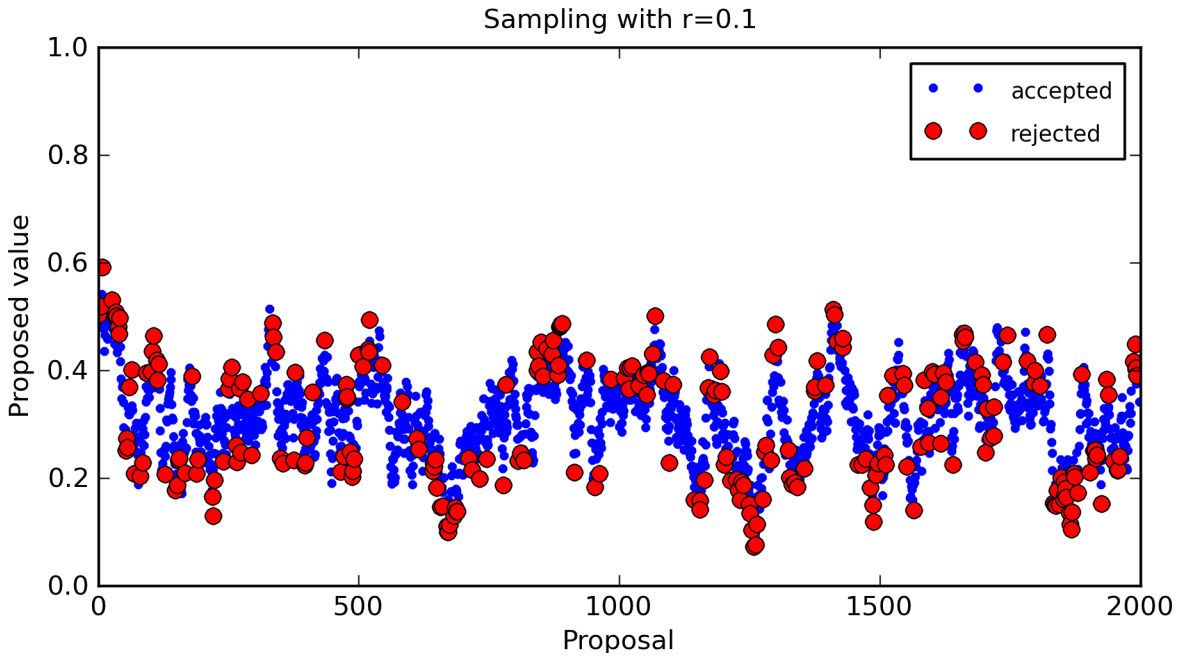


Figure 4: The proposed MCMC samples for the coin toss example with $r = 0.1$.

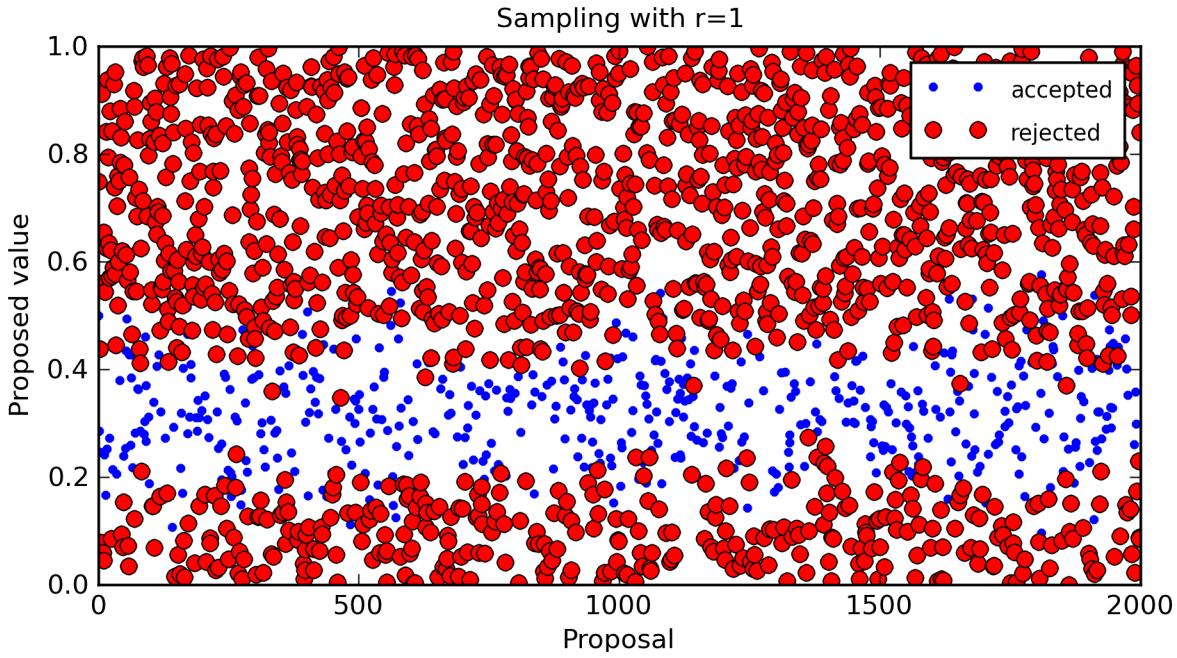


Figure 5: The proposed MCMC samples for the coin toss example with $r = 1$.

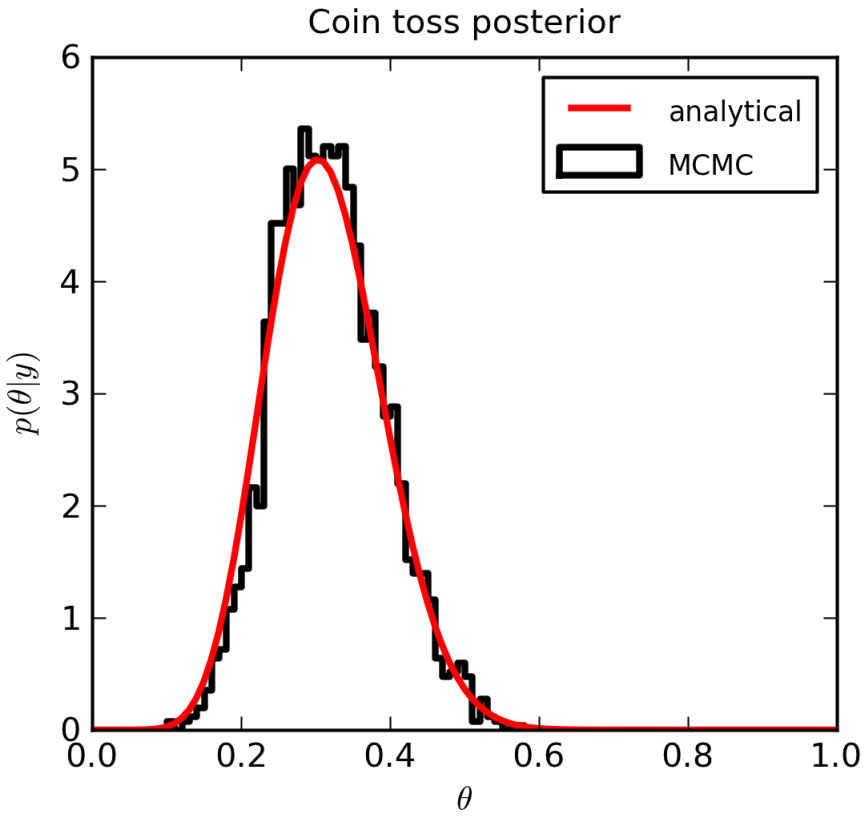


Figure 6: The MCMC and analytical posteriors for the coin toss example.