A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling

William M. Darling
School of Computer Science
University of Guelph

December 1, 2011

- LDA recap
- Gibbs sampling
- Markov chains
- Inverse Sampling
- Reject Sampling
- Stevens Method

Abstract

This technical report provides a tutorial on the theoretical details of probabilistic topic modeling and gives practical steps on implementing topic models such as Latent *Dirichlet* Allocation (LDA) through the Markov Chain Monte Carlo approximate inference algorithm Gibbs Sampling.

1 Introduction

Following its publication in 2003, Blei et al.'s Latent *Dirichlet* Allocation (LDA) [3] has made topic modeling – a subfield of machine learning applied to everything from computational linguistics [4] to bioinformatics [8] and political science [2] – one of the most popular and most successful paradigms for both supervised and unsupervised learning. Despite topic modeling's undisputed popularity, however, it is for many – particularly newcomers – a difficult area to break into due to its relative complexity and the common practice of leaving out implementation details in papers describing new models. While key update equations and other details on inference are often included, the intermediate steps used to arrive at these conclusions are often left out due to space constraints, and what details are given are rarely enough to enable most researchers to test the given results for themselves by implementing their own version of the described model. The purpose of this technical report is to help bridge the gap between the model definitions provided in research publications and the practical implementations that are required for performing learning in this exciting area. Ultimately, it is hoped that this tutorial will help enable the reader to build his or her own novel topic models.

This technical report will describe what topic modeling is, how various models (LDA in particular) work, and most importantly, how to implement a working system to perform learning with topic models. Topic modeling as an area will be introduced through the section on LDA, as it is the "original" topic model

and its modularity allows the basics of the model to be used in more complicated topic models.[1] Following the introduction to topic modeling through LDA, the problem of *posterior inference* will be discussed. This section will concentrate first on the theory of the stochastic approximate inference technique Gibbs Sampling and then it will discuss implementation details for building a topic model Gibbs sampler.

2 Latent *Dirichlet* Allocation

LDA is a generative probabilistic model for collections of grouped discrete data [3]. Each group is described as a random mixture over a set of latent topics where each topic is a discrete distribution over the collection's vocabulary. While LDA is applicable to any corpus of grouped discrete data, from now on I will refer to the standard NLP use case where a corpus is a collection of documents, and the data are words. The generative process for a document collection **D** under the LDA model is as follows:

1. For $k = 1...K$:
   (a) $\phi^{(k)} \sim Dirichlet(\beta)$ → k topics, each a dist over words
2. For each document $d \in \mathbf{D}$:
   (a) $\theta_d \sim Dirichlet(\alpha)$ → doc d = dist (topics)
   (b) For each word $w_i \in d$:
      i. $z_i \sim Discrete(\theta_d)$ → z_i = hidden/latent var word i comes from topic z_i
      ii. $w_i \sim Disctete(\phi^{(z_i)})$ → generate word w_i from topic ϕ^{z_i}

where K is the number of latent topics in the collection, $\phi^{(k)}$ is a discrete probability distribution over a fixed vocabulary that represents the kth topic distribution, θ_d is a document-specific distribution over the available topics, z_i is the topic index for word w_i, and α and β are hyperparameters for the symmetric Dirichlet distributions that the discrete distributions are drawn from.

The generative process described above results in the following joint distribution:

$$p(\mathbf{w}, \mathbf{z}, \theta, \phi|\alpha, \beta) = p(\phi|\beta)p(\theta|\alpha)p(\mathbf{z}|\theta)p(\mathbf{w}|\phi_z) \tag{1}$$

The unobserved (latent) variables $\mathbf{z}$, θ, and ϕ are what is of interest to us. Each θ_d is a low-dimensional representation of a document in "topic"-space, each z_i represents which topic generated the word instance w_i, and each $\phi^{(k)}$ represents a $K \times V$ matrix where $\phi_{i,j} = p(w_i|z_j)$. Therefore, one of the most interesting aspects of LDA is that it can learn, in an unsupervised manner, words that

[1] While LDA is an extension to probabilistic latent semantic analysis [12] (which in turn has ideological routes in the matrix factorization technique LSI), the topic modeling "revolution" really took off with the introduction of LDA likely due to its fully probabilistic grounding.

"environment"	"travel"	"fantasy football"
emission	travel	game
environmental	hotel	yard
air	roundtrip	defense
permit	fares	allowed
plant	special	fantasy
facility	offer	point
unit	city	passing
epa	visit	rank
water	miles	against
station	deal	team

Table 1: Three topics learned using LDA on the Enron Email Dataset.

we would associate with certain topics, and this is expressed through the topic distributions ϕ. An example of the top 10 words for 3 topics learned using LDA on the Enron email dataset[2] is shown in Figure 1 (the topic labels are added manually).

3 Inference

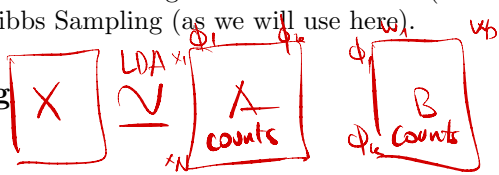
The key problem in topic modeling is *posterior inference*. This refers to reversing the defined generative process and learning the posterior distributions of the latent variables in the model given the observed data. In LDA, this amounts to solving the following equation:

$$p(\theta, \phi, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \phi, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)} \tag{2}$$

Unfortunately, this distribution is *intractable* to compute. The normalization factor in particular, $p(\mathbf{w}|\alpha, \beta)$, cannot be computed exactly. All is not lost, however, as there are a number of approximate inference techniques available that we can apply to the problem including variational inference (as used in the original LDA paper) and Gibbs Sampling (as we will use here).

3.1 Gibbs Sampling

3.1.1 Theory



Gibbs Sampling is one member of a family of algorithms from the Markov Chain Monte Carlo (MCMC) framework [9]. The MCMC algorithms aim to construct a Markov chain that has the target posterior distribution as its stationary distribution. In other words, after a number of iterations of stepping through the chain, sampling from the distribution should converge to be close to sampling

[2] http://www.cs.cmu.edu/~enron/.

from the desired posterior. Gibbs Sampling is based on sampling from conditional distributions of the variables of the posterior.

For example, to sample $\mathbf{x}$ from the joint distribution $p(\mathbf{x}) = p(x_1, ..., x_m)$, where there is no closed form solution for $p(\mathbf{x})$, but a representation for the conditional distributions is available, using Gibbs Sampling one would perform the following (from [1]):

1. Randomly initialize each x_i
2. For $t = 1, ..., T$:
   - 2.1 $x_1^{t+1} \sim p(x_1 | x_2^{(t)}, x_3^{(t)}, ..., x_m^{(t)})$
   - 2.2 $x_2^{t+1} \sim p(x_2 | x_1^{(t+1)}, x_3^{(t)}, ..., x_m^{(t)})$
   - ...
   - 2.m $x_m^{t+1} \sim p(x_m | x_1^{(t+1)}, x_2^{(t+1)}, ..., x_{m-1}^{(t+1)})$

This procedure is repeated a number of times until the samples begin to converge to what would be sampled from the true distribution. While convergence is theoretically guaranteed with Gibbs Sampling, there is no way of knowing how many iterations are required to reach the stationary distribution. Therefore, diagnosing convergence is a real problem with the Gibbs Sampling approximate inference method. However, in practice it is quite powerful and has fairly good performance. Typically, an acceptable estimation of convergence can be obtained by calculating the log-likelihood or even, in some situations, by inspection of the posteriors.

For LDA, we are interested in the latent document-topic portions θ_d, the topic-word distributions $\phi^{(z)}$, and the topic index assignments for each word z_i. While conditional distributions – and therefore an LDA Gibbs Sampling algorithm – can be derived for each of these latent variables, we note that both θ_d and $\phi^{(z)}$ can be calculated using just the topic index assignments z_i (*i.e.* $\mathbf{z}$ is a sufficient statistic for both these distributions).[3] Therefore, a simpler algorithm can be used if we integrate out the multinomial parameters and simply sample z_i. This is called a *collapsed* Gibbs sampler.

The collapsed Gibbs sampler for LDA needs to compute the probability of a topic z being assigned to a word w_i, given all other topic assignments to all other words. Somewhat more formally, we are interested in computing the following posterior up to a constant:

$$p(z_i | \mathbf{z}_{-i}, \alpha, \beta, \mathbf{w}) \tag{3}$$

where $\mathbf{z}_{-i}$ means all topic allocations *except* for z_i. To begin, the rules of conditional probability tell us that:

$$p(z_i | \mathbf{z}_{-i}, \alpha, \beta, \mathbf{w}) = \frac{p(z_i, \mathbf{z}_{-i}, \mathbf{w} | \alpha, \beta)}{p(\mathbf{z}_{-i}, \mathbf{w} | \alpha, \beta)} \propto p(z_i, \mathbf{z}_{-i}, \mathbf{w} | \alpha, \beta) = p(\mathbf{z}, \mathbf{w} | \alpha, \beta) \tag{4}$$

[3] $\theta_{d,z} = \frac{n(d,z)+\alpha}{\sum_{|Z|} n(d,z)+\alpha}$, $\phi_{z,w} = \frac{n(z,w)+\beta}{\sum_{|W|} n(z,w)+\beta}$.

We then have:

$$p(\mathbf{w}, \mathbf{z}|\alpha, \beta) = \int\int p(\mathbf{z}, \mathbf{w}, \theta, \phi|\alpha, \beta)\mathrm{d}\theta\mathrm{d}\phi \tag{5}$$

Following the LDA model defined in equation (1), we can expand the above equation to get:

$$p(\mathbf{w}, \mathbf{z}|\alpha, \beta) = \int\int p(\phi|\beta)p(\theta|\alpha)p(\mathbf{z}|\theta)p(\mathbf{w}|\phi_z)\mathrm{d}\theta\mathrm{d}\phi \tag{6}$$

Then, we group the terms that have dependent variables:

$$p(\mathbf{w}, \mathbf{z}|\alpha, \beta) = \int p(\mathbf{z}|\theta)p(\theta|\alpha)\mathrm{d}\theta \int p(\mathbf{w}|\phi_z)p(\phi|\beta)\mathrm{d}\phi \tag{7}$$

Both terms are multinomials with Dirichlet priors. Because the Dirichlet distribution is conjugate to the multinomial distribution, our work is vastly simplified; multiplying the two results in a Dirichlet distribution with an adjusted parameter. Beginning with the first term, we have:

$$\begin{aligned}
\int p(\mathbf{z}|\theta)p(\theta|\alpha)\mathrm{d}\theta &= \int \prod_i \theta_{d,z_i} \frac{1}{B(\alpha)} \prod_k \theta_{d,k}^{\alpha_k} \mathrm{d}\theta_d \\
&= \frac{1}{B(\alpha)} \int \prod_k \theta_{d,k}^{n_{d,k}+\alpha_k} \mathrm{d}\theta_d \\
&= \frac{B(n_{d,\cdot} + \alpha)}{B(\alpha)}
\end{aligned} \tag{8}$$

where $n_{d,k}$ is the number of times words in document d are assigned to topic k, a $\cdot$ indicates summing over that index, and $B(\alpha)$ is the multinomial beta function, $B(\alpha) = \frac{\prod_k \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)}$. Similarly, for the second term (calculating the likelihood of words given certain topic assignments):

$$\begin{aligned}
\int p(\mathbf{w}|\phi_z)p(\phi|\beta)\mathrm{d}\phi &= \int \prod_d \prod_i \phi_{z_{d,i},w_{d,i}} \prod_k \frac{1}{B(\beta)} \prod_w \phi_{k,w}^{\beta_w} \mathrm{d}\phi_k \\
&= \prod_k \frac{1}{B(\beta)} \int \prod_w \phi_{k,w}^{\beta_w + n_{k,w}} \mathrm{d}\phi_k \\
&= \prod_k \frac{B(n_{k,\cdot} + \beta)}{B(\beta)}
\end{aligned} \tag{9}$$

Combining equations (8) and (9), the expanded joint distribution is then:

$$p(\mathbf{w}, \mathbf{z}|\alpha, \beta) = \prod_d \frac{B(n_{d,\cdot} + \alpha)}{B(\alpha)} \prod_k \frac{B(n_{k,\cdot} + \beta)}{B(\beta)} \tag{10}$$

The Gibbs sampling equation for LDA can then be derived using the chain rule (where we leave the hyperparameters α and β out for clarity).[4] Note that the superscript $^{(-i)}$ signifies leaving the ith token out of the calculation:

$$p(z_i|\mathbf{z}^{(-i)}, \mathbf{w}) = \frac{p(\mathbf{w}, \mathbf{z})}{p(\mathbf{w}, \mathbf{z}^{(-i)})} = \frac{p(\mathbf{z})}{p(\mathbf{z}^{(-i)})} \cdot \frac{p(\mathbf{w}|\mathbf{z})}{p(\mathbf{w}^{(-i)}|\mathbf{z}^{(-i)})p(w_i)}$$

$$\propto \prod_d \frac{B(n_{d,\cdot} + \alpha)}{B(n_{d,\cdot}^{(-i)} + \alpha)} \prod_k \frac{B(n_{k,\cdot} + \beta)}{B(n_{k,\cdot}^{(-i)} + \beta)}$$

$$\propto \frac{\Gamma(n_{d,k} + \alpha_k)\Gamma(\sum_{k=1}^K n_{d,k}^{(-i)} + \alpha_k)}{\Gamma(n_{d,k}^{(-i)} + \alpha_k)\Gamma(\sum_{k=1}^K n_{d,k} + \alpha_k)} \cdot \frac{\Gamma(n_{k,w} + \beta_w)\Gamma(\sum_{w=1}^W n_{k,w}^{(-i)} + \beta_w}{\Gamma(n_{k,w}^{(-i)} + \beta_w)\Gamma(\sum_{w=1}^W n_{k,w} + \beta_w}$$

$$\propto (n_{d,k}^{(-i)} + \alpha_k) \frac{n_{k,w}^{(-i)} + \beta_w}{\sum_{w'} n_{k,w'}^{(-i)} + \beta_{w'}} \tag{11}$$

3.1.2 Implementation

Implementing an LDA collapsed Gibbs sampler is surprisingly straightforward. It involves setting up the requisite count variables, randomly initializing them, and then running a loop over the desired number of iterations where on each loop a topic is sampled for each word instance in the corpus. Following the Gibbs iterations, the counts can be used to compute the latent distributions θ_d and ϕ_k.

The only required count variables include $n_{d,k}$, the number of words assigned to topic k in document d; and $n_{k,w}$, the number of times word w is assigned to topic k. However, for simplicity and efficiency, we also keep a running count of n_k, the total number of times any word is assigned to topic k. Finally, in addition to the obvious variables such as a representation of the corpus ($\mathbf{w}$), we need an array $\mathbf{z}$ which will contain the current topic assignment for each of the N words in the corpus.

Because the Gibbs sampling procedure involves sampling from distributions conditioned on all *other* variables (in LDA this of course includes all other current topic assignments, but not the current one), before building a distribution from equation (11), we must remove the current assignment from the equation. We can do this by decrementing the counts associated with the current assignment because the topic assignments in LDA are *exchangeable* (i.e. the joint probability distribution is invariant to permutation). We then calculate the (unnormalized) probability of each topic assignment using equation (11). This discrete distribution is then sampled from and the chosen topic is set in the $\mathbf{z}$ array and the appropriate counts are then incremented. See Algorithm 1 for the full LDA Gibbs sampling procedure.

[4] For the full, nothing-left-out derivation, please see [5] and [11].

Input: words $\mathbf{w} \in$ documents $\mathbf{d}$
Output: topic assignments $\mathbf{z}$ and counts $n_{d,k}, n_{k,w}$, and n_k
begin
  randomly initialize $\mathbf{z}$ and increment counters
  foreach *iteration* **do**
    for $i = 0 \rightarrow N - 1$ **do**
      $word \leftarrow w[i]$
      $topic \leftarrow z[i]$
      $n_{d,topic}$-=1; $n_{word,topic}$-=1; n_{topic}-=1
      for $k = 0 \rightarrow K - 1$ **do**
        $p(z = k|\cdot) = (n_{d,k} + \alpha_k)\frac{n_{k,w}+\beta_w}{n_k+\beta \times W}$
      end
      $topic \leftarrow$ sample from $p(z|\cdot)$
      $z[i] \leftarrow topic$
      $n_{d,topic}$+=1; $n_{word,topic}$+=1; n_{topic}+=1
    end
  end
  return $\mathbf{z}$, $n_{d,k}, n_{k,w}, n_k$
end

Algorithm 1: LDA Gibbs Sampling

4 Extensions To LDA

While LDA – the "simplest" topic model – is useful in and of itself, a great deal of novel research surrounds extending the basic LDA model to fit a specific task or to improve the model by describing a more complex generative process that results in a better model of the real world. There are countless papers delineating such extensions and it is not my intention to go through them all here. Instead, this section will outline some of the ways that LDA can and has been extended with the goal of explaining how inference changes as a result of additions to a model and how to implement those changes in a Gibbs sampler.

4.1 LDA With a Background Distribution

One of the principal problems with LDA is that for useful results, stop-words must be removed in a pre-processing step. Without this filtering, very common words such as *the, of, to, and, a*, etc. will pervade the learned topics, hiding the statistical semantic word patterns that are of interest. While stop-word removal does a good job at solving this problem, it is an *ad hoc* measure that results in a model resting on a non-coherent theoretical basis. Further, stop-word removal is not without problems. Stop-word lists must often be domain-dependent, and there are inevitably cases where filtering results in under-coverage or over-coverage, causing the model to continue being plagued by noise, or missing patterns that may be of interest to us.

One approach to keep stop-words out of the topic distributions is to imag-

ine all stop-words being generated by a "background" distribution [6, 7, 10]. The background distribution is the same as a topic – it is a discrete probability distribution over the corpus vocabulary – but every document draws from the background as well as the topics specific to that document. [7] and [10] use this approach to separate high-content words from less-important words to perform multi-document summarization. [6] uses a similar model for information retrieval where a word can either be generated from a background distribution, a document-specific distribution, or one of T topic distributions shared amongst all the documents. The generative process is similar to that of LDA, except that there is a multinomial variable x associated with each word that is over the three different "sources" of words. When $x = 0$, the background distribution generates the word, when $x = 1$, the document-specific distribution generates the word, and when $x = 2$, one of the topic distributions generates the word.

Here, we will describe a simpler model where only a background distribution is added to LDA. A binomial variable x is associated with each word that decides whether the word will be generated by the topic distributions or by the background. The generative process is then:

1. $\zeta \sim Dirichlet(\delta)$
2. For $k = 1...K$:
   (a) $\phi^{(k)} \sim Dirichlet(\beta)$
3. For each document $d \in \mathbf{D}$:
   (a) $\theta_d \sim Dirichlet(\alpha)$
   (b) $\lambda_d \sim Dirichlet(\gamma)$
   (c) For each word $w_i \in d$:
      i. $x_i \sim Discrete(\lambda_d)$
      ii. If $x = 0$:
         A. $w_i \sim Discrete(\zeta)$
      iii. Else:
         A. $z_i \sim Discrete(\theta_d)$
         B. $w_i \sim Disctete(\phi^{(z_i)})$

where ζ is the background distribution, and λ_d is a document-specific binomial sampled from a Dirichlet prior γ.

Developing a Gibbs sampler for this model is similar to the LDA implementation, but we have to be careful about when counts are incremented and decremented. We only adjust the background-based counts when the background was sampled as the word generator, and we only adjust the topic counts when it is the converse. We must update the x-based counts each time, however, because we sample the *route* that led to the word being generated each time. The sampler must compute the probability not only of a topic being chosen for the given document and the probability of that topic generating the given

word, it must also compute the probability that the model is in the topic-model state. This too, however, is straightforward to implement. A distribution of $T + 1$ components can be created for each word (on each iteration) where the first component corresponds to the background distribution generating the word and the other T are the probabilities for each topic having generated the word.

5 Conclusion

LDA and other topic models are an exciting development in machine learning and the surface has only been scratched on their potential in a number of diverse fields. This report has sought to aid researchers new to the field in both understanding the mathematical underpinnings of topic modeling and in implementing algorithms to make use of this new pattern recognition paradigm.

References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[2] David M. Blei and Sean Gerrish. Predicting legislative roll calls from text. In *International Conference on Machine Learning*, 2011.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[4] Jordan Boyd-Graber, David Blei, and Xiaojin Zhu. A topic model for word sense disambiguation. In *Empirical Methods in Natural Language Processing*, 2007.

[5] Bob Carpenter. Integrating out multinomial parameters in latent dirichlet allocation and naive bayes for collapsed gibbs sampling. Technical report, Lingpipe, Inc., 2010.

[6] Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. Modeling general and specific aspects of documents with a probabilistic topic model. In *NIPS*, pages 241–248, 2006.

[7] Hal Daumé, III and Daniel Marcu. Bayesian query-focused summarization. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 305–312, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[8] Georg K Gerber, Robin D Dowell, Tommi S Jaakkola, and David K Gifford. Automated discovery of functional generality of human gene expression programs. *PLoS Comput Biol*, 3(8):e148, 2007.

[9] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo In Practice*. Chapman and Hall/CRC, 1999.

[10] Aria Haghighi and Lucy Vanderwende. Exploring content models for multi-document summarization. In *NAACL '09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[11] Gregor Heinrich. Parameter estimation for text analysis. Technical report, University of Leipzig, Germany, 2008.

[12] Thomas Hofmann. Probabilistic latent semantic analysis. In *In Proc. of Uncertainty in Artificial Intelligence, UAI99*, pages 289–296, 1999.

Explaining the Gibbs Sampler

GEORGE CASELLA and EDWARD I. GEORGE*

Computer-intensive algorithms, such as the Gibbs sampler, have become increasingly popular statistical tools, both in applied and theoretical work. The properties of such algorithms, however, may sometimes not be obvious. Here we give a simple explanation of how and why the Gibbs sampler works. We analytically establish its properties in a simple case and provide insight for more complicated cases. There are also a number of examples.

KEY WORDS: Data augmentation; Markov chains; Monte Carlo methods; Resampling techniques.

1. INTRODUCTION

The continuing availability of inexpensive, high-speed computing has already reshaped many approaches to statistics. Much work has been done on algorithmic approaches (such as the EM algorithm; Dempster, Laird, and Rubin 1977), or resampling techniques (such as the bootstrap; Efron 1982). Here we focus on a different type of computer-intensive statistical method, the Gibbs sampler.

The Gibbs sampler enjoyed an initial surge of popularity starting with the paper of Geman and Geman (1984), who studied image-processing models. The roots of the method, however, can be traced back to at least Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953), with further development by Hastings (1970). More recently, Gelfand and Smith (1990) generated new interest in the Gibbs sampler by revealing its potential in a wide variety of conventional statistical problems.

The Gibbs sampler is a technique for generating random variables from a (marginal) distribution indirectly, without having to calculate the density. Although straightforward to describe, the mechanism that drives this scheme may seem mysterious. The purpose of this article is to demystify the workings of these algorithms by exploring simple cases. In such cases, it is easy to see that Gibbs sampling is based only on elementary properties of Markov chains.

Through the use of techniques like the Gibbs sampler, we are able to avoid difficult calculations, replacing them instead with a sequence of easier calculations. These methodologies have had a wide impact on practical problems, as discussed in Section 6. Although most applications of the Gibbs sampler have been in Bayesian models, it is also extremely useful in classical (likelihood) calculations [see Tanner (1991) for many examples]. Furthermore, these calculational methodologies have also had an impact on theory. By freeing statisticians from dealing with complicated calculations, the statistical aspects of a problem can become the main focus. This point is wonderfully illustrated by Smith and Gelfand (1992).

In the next section we describe and illustrate the application of the Gibbs sampler in bivariate situations. Section 3 is a detailed development of the underlying theory, given in the simple case of a 2×2 table with multinomial sampling. From this detailed development, the theory underlying general situations is more easily understood, and is also outlined. Section 4 elaborates the role of the Gibbs sampler in relating conditional and marginal distributions and illustrates some higher dimensional generalizations. Section 5 describes many of the implementation issues surrounding the Gibbs sampler, and Section 6 contains a discussion and describes many applications.

2. ILLUSTRATING THE GIBBS SAMPLER

Suppose we are given a joint density $f(x, y_1, \ldots, y_p)$, and are interested in obtaining characteristics of the marginal density

$$f(x) = \int \cdots \int f(x, y_1, \ldots, y_p)\, dy_1 \cdots dy_p, \quad (2.1)$$

such as the mean or variance. Perhaps the most natural and straightforward approach would be to calculate $f(x)$ and use it to obtain the desired characteristic. However, there are many cases where the integrations in (2.1) are extremely difficult to perform, either analytically or numerically. In such cases the Gibbs sampler provides an alternative method for obtaining $f(x)$.

Rather than compute or approximate $f(x)$ directly, the Gibbs sampler allows us effectively to generate a sample $X_1, \ldots, X_m \sim f(x)$ *without requiring* $f(x)$. By simulating a large enough sample, the mean, variance, or any other characteristic of $f(x)$ can be calculated to the desired degree of accuracy.

It is important to realize that, in effect, the end result of any calculations, although based on simulations, are the population quantities. For example, to calculate the mean of $f(x)$, we could use $(1/m)\Sigma_{i=1}^{m} X_i$, and the fact that

$$\lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} X_i = \int_{-\infty}^{\infty} xf(x)\, dx = EX. \quad (2.2)$$

*George Casella is Professor, Biometrics Unit, Cornell University, Ithaca, NY 14853. The research of this author was supported by National Science Foundation Grant DMS 89-0039. Edward I. George is Professor, Department of MSIS, The University of Texas at Austin, TX 78712. The authors thank the editors and referees, whose comments led to an improved version of this article.

Thus, by taking m large enough, any population characteristic, even the density itself, can be obtained to any degree of accuracy.

To understand the workings of the Gibbs sampler, we first explore it in the two-variable case. Starting with a pair of random variables (X, Y), the Gibbs sampler generates a sample from $f(x)$ by sampling instead from the conditional distributions $f(x \mid y)$ and $f(y \mid x)$, distributions that are often known in statistical models. This is done by generating a "Gibbs sequence" of random variables

$$Y_0', X_0', Y_1', X_1', Y_2', X_2', \ldots, Y_k', X_k'. \quad (2.3)$$

The initial value $Y_0' = y_0'$ is specified, and the rest of (2.3) is obtained iteratively by alternately generating values from

$$X_j' \sim f(x \mid Y_j' = y_j')$$
$$Y_{j+1}' \sim f(y \mid X_j' = x_j'). \quad (2.4)$$

We refer to this generation of (2.3) as Gibbs sampling. It turns out that under reasonably general conditions, the distribution of X_k' converges to $f(x)$ (the true marginal of X) as $k \to \infty$. Thus, for k large enough, the final observation in (2.3), namely $X_k' = x_k'$, is effectively a sample point from $f(x)$.

The convergence (in distribution) of the Gibbs sequence (2.3) can be exploited in a variety of ways to obtain an approximate sample from $f(x)$. For example, Gelfand and Smith (1990) suggest generating m independent Gibbs sequences of length k, and then using the final value of X_k' from each sequence. If k is chosen large enough, this yields an approximate iid sample from $f(x)$. Methods for choosing such k, as well as alternative approaches to extracting information from the Gibbs sequence, are discussed in Section 5. For the sake of clarity and consistency, we have used only the preceding approach in all of the illustrative examples that follow.

Example 1. For the following joint distribution of X and Y,

Beta-Binomial

$$f(x, y) \propto \binom{n}{x} y^{x+\alpha-1}(1-y)^{n-x+\beta-1},$$
$$x = 0, 1, \ldots, n \quad 0 \le y \le 1, \quad (2.5)$$

suppose we are interested in calculating some characteristics of the marginal distribution $f(x)$ of X. The Gibbs sampler allows us to generate a sample from this marginal as follows. From (2.5) it follows (suppressing the overall dependence on n, α, and β) that

Cond $f(x \mid y)$ is Binomial (n, y) (2.6a)

Cond $f(y \mid x)$ is Beta $(x + \alpha, n - x + \beta)$. (2.6b)

If we now apply the iterative scheme (2.4) to the distributions (2.6), we can generate a sample $X_1, X_2, \ldots, X_m$ from $f(x)$ and use this sample to estimate any desired characteristic.

As the reader may have already noticed, Gibbs sampling is actually not needed in this example, since $f(x)$ can be obtained analytically from (2.5) as

$$f(x) = \binom{n}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(\alpha+\beta+n)},$$
$$x = 0, 1, \ldots, n, \quad (2.7)$$

the beta-binomial distribution. Here, characteristics of $f(x)$ can be directly obtained from (2.7), either analytically or by generating a sample from the marginal and not fussing with the conditional distributions. However, this simple situation is useful for illustrative purposes. Figure 1 displays histograms of two samples $x_1, \ldots, x_m$ of size $m = 500$ from the beta-binomial distribution of (2.7) with $n = 16$, $\alpha = 2$, and $\beta = 4$.

The two histograms are very similar, giving credence to the claim that the Gibbs scheme for random variable generation is indeed generating variables from the marginal distribution.

One feature brought out by Example 1 is that the Gibbs sampler is really not needed in any bivariate situation where the joint distribution $f(x, y)$ can be calculated, since $f(x) = f(x, y)/f(y \mid x)$. However, as the next example shows, Gibbs sampling may be indispensable in situations where $f(x, y)$, $f(x)$, or $f(y)$ cannot be calculated.

Example 2. Suppose X and Y have conditional distributions that are exponential distributions restricted to the interval $(0, B)$, that is,

$$f(x \mid y) \propto ye^{-yx}, \quad 0 < x < B < \infty \quad (2.8a)$$
$$f(y \mid x) \propto xe^{-xy}, \quad 0 < y < B < \infty, \quad (2.8b)$$

where B is a known positive constant. The restriction to the interval $(0, B)$ ensures that the marginal $f(x)$ exists. Although the form of this marginal is not easily calculable, by applying the Gibbs sampler to the conditionals in (2.8) any characteristic of $f(x)$ can be obtained.

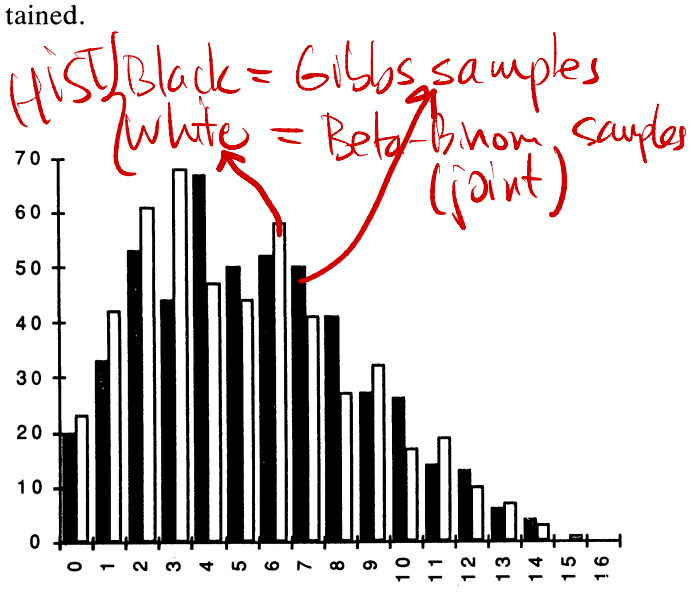


Figure 1. Comparison of Two Histograms of Samples of Size $m = 500$ From the Beta-Binomial Distribution With $n = 16$, $\alpha = 2$, and $\beta = 4$. The black histogram sample was obtained using Gibbs sampling with $k = 10$. The white histogram sample was generated directly from the beta-binomial distribution.

In Figure 2 we display a histogram of a sample of size $m = 500$ from $f(x)$ obtained by using the final values from Gibbs sequences of length $k = 15$.

In Section 4 we see that if B is not finite, then the densities in (2.8) are not a valid pair of conditional densities in the sense that there is no joint density $f(x, y)$ to which they correspond, and the Gibbs sequence fails to converge.

Gibbs sampling can be used to estimate the density itself by averaging the final conditional densities from each Gibbs sequence. From (2.3), just as the values $X'_k = x'_k$ yield a realization of $X_1, \cdots, X_m \sim f(x)$, the values $Y'_k = y'_k$ yield a realization of $Y_1, \cdots, Y_m \sim f(y)$. Moreover, the average of the conditional densities $f(x \mid Y'_k = y'_k)$ will be a close approximation to $f(x)$, and we can estimate $f(x)$ with

$$\hat{f}(x) = \frac{1}{m} \sum_{i=1}^{m} f(x \mid y_i), \tag{2.9}$$

where $y_1, \cdots, y_m$ is the sequence of realized values of final Y observations from each Gibbs sequence. The theory behind the calculation in (2.9) is that the expected value of the conditional density is

$$E[f(x \mid Y)] = \int f(x \mid y) f(y)\, dy = f(x), \tag{2.10}$$

a calculation mimicked by (2.9), since $y_1, \cdots, y_m$ approximate a sample from $f(y)$. For the densities in (2.8), this estimate of $f(x)$ is shown in Figure 2.

Example 1 (continued): The density estimate methodology of (2.9) can also be used in discrete distributions, which we illustrate for the beta-binomial of Example 1. Using the observations generated to construct Figure 1, we can, analogous to (2.9), estimate the marginal probabilities of X using

$$\hat{P}(X = x) = \frac{1}{m} \sum_{i=1}^{m} P(X = x \mid Y_i = y_i). \tag{2.11}$$

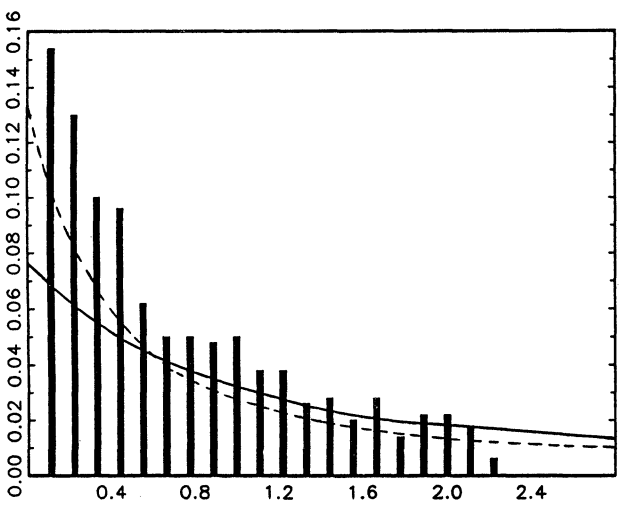


Figure 2. Histogram for x of a Sample of Size m = 500 From the Pair of Conditional Distributions in (2.8), With B = 5, Obtained Using Gibbs Sampling With k = 15 Along With an Estimate of the Marginal Density Obtained From Equation (2.9) (solid line). The dashed line is the true marginal density, as explained in Section 4.1.

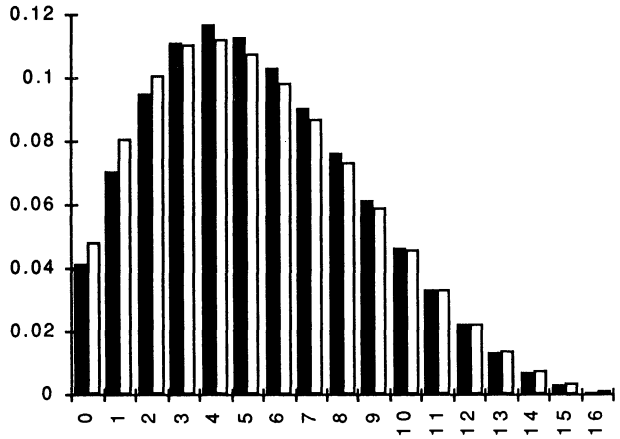


Figure 3. Comparison of Two Probability Histograms of the Beta-Binomial Distribution With n = 16, α = 2, and β = 4. The black histogram represents estimates of the marginal distribution of X using Equation (2.11), based on a sample of Size m = 500 from the pair of conditional distributions in (2.6). The Gibbs sequence had length k = 10. The white histogram represents the exact beta-binomial probabilities.

Figure 3 displays these probability estimates overlayed with the exact beta-binomial probabilities for comparison.

The density estimates (2.9) and (2.11) illustrate an important aspect of using the Gibbs sampler to evaluate characteristics of $f(x)$. The quantities $f(x \mid y_1), \cdots, f(x \mid y_m)$, calculated using the simulated values $y_1, \cdots, y_m$, carry more information about $f(x)$ than $x_1, \cdots, x_m$ alone, and will yield better estimates. For example, an estimate of the mean of $f(x)$ is $(1/m)\ \Sigma_{i=1}^{m}\ x_i$, but a better estimate is $(1/m)\ \Sigma_{i=1}^{m}\ E(X \mid y_i)$, as long as these conditional expectations are obtainable. The intuition behind this feature is the Rao-Blackwell theorem (illustrated by Gelfand and Smith 1990), and established analytically by Liu, Wong, and Kong (1991).

3. A SIMPLE CONVERGENCE PROOF

It is not immediately obvious that a random variable with distribution $f(x)$ can be produced by the Gibbs sequence of (2.3) or that the sequence even converges. That this is so relies on the Markovian nature of the iterations, which we now develop in detail for the simple case of a 2×2 table with multinomial sampling.

Suppose X and Y are each (marginally) Bernoulli random variables with joint distribution

	$X = 0$	$X = 1$
$Y = 0$	p_1	p_2
$Y = 1$	p_3	p_4

$$p_i \geq 0, p_1 + p_2 + p_3 + p_4 = 1,$$

or, in terms of the joint probability function,

$$\begin{bmatrix} f_{x,y}(0,0) & f_{x,y}(1,0) \\ f_{x,y}(0,1) & f_{x,y}(1,1) \end{bmatrix} = \begin{bmatrix} p_1 & p_2 \\ p_3 & p_4 \end{bmatrix}.$$

For this distribution, the marginal distribution of x is given by

$$f_x = [f_x(0) \quad f_x(1)] = [p_1 + p_3 \quad p_2 + p_4], \quad (3.1)$$

a Bernoulli distribution with success probability $p_2 + p_4$.

The conditional distributions of $X \mid Y = y$ and $Y \mid X = x$ are straightforward to calculate. For example the distribution of $X \mid Y = 1$ is Bernoulli with success probability $p_4/(p_3 + p_4)$. All of the conditional probabilities can be expressed in two matrices,

$$A_{y|x} = \begin{bmatrix} \frac{p_1}{p_1 + p_3} & \frac{p_3}{p_1 + p_3} \\ \frac{p_2}{p_2 + p_4} & \frac{p_4}{p_2 + p_4} \end{bmatrix}$$

and

$$A_{x|y} = \begin{bmatrix} \frac{p_1}{p_1 + p_2} & \frac{p_2}{p_1 + p_2} \\ \frac{p_3}{p_3 + p_4} & \frac{p_4}{p_3 + p_4} \end{bmatrix},$$

where $A_{y|x}$ has the conditional probabilities of Y given $X = x$, and $A_{y|x}$ has the conditional probabilities of X given $Y = y$.

The iterative sampling scheme applied to this distribution yields (2.3) as a sequence of 0's and 1's. The matrices $A_{x|y}$ and $A_{y|x}$ may be thought of as transition matrices giving the probabilities of getting to x states from y states and vice versa, that is, $P(X = x \mid Y = y)$ = probability of going from state y to state x.

If we are only interested in generating the marginal distribution of X, we are mainly concerned with the X' sequence from (2.3). To go from $X'_0 \to X'_1$ we have to go through Y'_1, so the iteration sequence is $X'_0 \to Y'_1 \to X'_1$, and $X'_0 \to X'_1$ forms a Markov chain with transition probability

$$P(X'_1 = x_1 \mid X'_0 = x_0) = \sum_y P(X'_1 = x_1 \mid Y'_1 = y) \times P(Y'_1 = y \mid X'_0 = x_0). \quad (3.2)$$

The transition probability matrix of the X' sequence, $A_{x|x}$, is given by

$$A_{x|x} = A_{y|x}A_{x|y},$$

and now we can easily calculate the probability distribution of any X'_k in the sequence. That is, the transition matrix that gives $P(X'_k = x_k \mid X'_0 = x_0)$ is $(A_{x|x})^k$. Furthermore, if we write

$$f_k = [f_k(0) \quad f_k(1)]$$

to denote the marginal probability distribution of X'_k, then for any k,

$$f_k = f_0 A^k_{x|x} = (f_0 A^{k-1}_{x|x})A_{x|x} = f_{k-1}A_{x|x}. \quad (3.3)$$

It is well known (see, for example, Hoel, Port, and Stone 1972), that as long as all the entries of $A_{x|x}$ are positive, then (3.3) implies that for any initial probability f_0, as $k \to \infty$, f_k converges to the unique distribution f that is a *stationary point* of (3.3), and satisfies

$$fA_{x|x} = f. \quad (3.4)$$

Thus, if the Gibbs sequence converges, the f that satisfies (3.4) must be the marginal distribution of X. Intuitively, there is nowhere else for this iteration to go; in the long run we will get X's in the proportion dictated by the marginal distribution. However, it is straightforward to check that (3.4) is satisfied by f_x of (3.1), that is,

$$f_x A_{x|x} = f_x A_{y|x} A_{x|y} = f_x.$$

As $k \to \infty$, the distribution of X'_k gets closer to f_x, so if we stop the iteration scheme (2.3) at a large enough value of k, we can assume that the distribution of X'_k is approximately f_x. Moreover, the larger the value of k, the better the approximation. This topic is discussed further in Section 5.

The algebra for the 2×2 case immediately works for any $n \times m$ joint distribution of X's and Y's. We can analogously define the $n \times n$ transition matrix $A_{x|x}$ whose stationary distribution will be the marginal distribution of X. If either (or both) of X and Y are continuous, then the finite dimensional arguments will not work. However, with suitable assumptions, all of the theory still goes through, so the Gibbs sampler still produces a sample from the marginal distribution of X. Equation (3.2) would now represent the conditional density of X'_1 given X'_0, and could be written

$$f_{X'_1|X'_0}(x_1 \mid x_0) = \int f_{X'_1|Y'_1}(x_1 \mid y) f_{Y'_1|X'_0}(y \mid x_0)\, dy.$$

(Sometimes it is helpful to use subscripts to denote the density.) Then, step by step, we could write the conditional densities of $X'_2|X'_0$, $X'_3|X'_0$, $X'_4|X'_0$, $\cdots$. Similar to the k-step transition matrix $(A_{x|x})^k$, we derive an "infinite transition matrix" with entries that satisfy the relationship

$$f_{X'_k|X'_0}(x \mid x_0) = \int f_{X'_k|X'_{k-1}}(x \mid t) f_{X'_{k-1}|X'_0}(t \mid x_0)\, dt, \quad (3.5)$$

which is the continuous version of (3.3). The density $f_{X'_k|X'_{k-1}}$ represents a one-step transition, and the other two densities play the role of f_k and f_{k-1}. As $k \to \infty$, it again follows that the stationary point of (3.5) is the marginal density of X, the density to which $f_{X'_k|X'_{k-1}}$ converges.

4. CONDITIONALS DETERMINE MARGINALS

Gibbs sampling can be thought of as a practical implementation of the fact that knowledge of the conditional distributions is sufficient to determine a joint distribution (if it exists!). In the bivariate case, the derivation of the marginal from the conditionals is fairly straightforward. Complexities in the multivariate case, however, make these connections more obscure. We

$$A_{x|x} = A_{y|x} \cdot A_{x|y}$$

(2 steps) $A_{y|x}$: cond $(Y|x)$; $A_{x|y}$: cond $(x|y)$

$$= \begin{bmatrix} \frac{p_1}{p_1+p_3} & \frac{p_3}{p_1+p_3} \\ \frac{p_2}{p_2+p_4} & \frac{p_4}{p_2+p_4} \end{bmatrix} \begin{bmatrix} \frac{p_1}{p_1+p_2} & \frac{p_2}{p_1+p_2} \\ \frac{p_3}{p_3+p_4} & \frac{p_4}{p_3+p_4} \end{bmatrix}$$

$A_{x|x}$

$$\begin{bmatrix} \frac{p_1^2}{(p_1+p_3)(p_1+p_2)} + \frac{p_3^2}{(p_1+p_3)(p_3+p_4)} & \frac{p_1p_2}{(p_1+p_3)(p_1+p_2)} + \frac{p_3p_4}{(p_1+p_3)(p_3+p_4)} \\ \frac{p_2p_1}{(p_2+p_4)(p_1+p_2)} + \frac{p_4p_3}{(p_2+p_4)(p_3+p_4)} & \frac{p_2^2}{(p_2+p_4)(p_1+p_2)} + \frac{p_4^2}{(p_2+p_4)(p_3+p_4)} \end{bmatrix}$$

Gibbs Th ($\simeq$ Markov / convergence)

$f_x^{next} = f_x^{old} \cdot A_{x|x} \Rightarrow f_x^*$ convergent (stationary dist for x)

$f_x^* = f_x^* \cdot A_{x|x}$ verify? f_x^* = marginal of x

$$[p_1+p_3 \quad p_2+p_4] \cdot A_{x|x} \overset{?}{=} [p_1+p_3 \quad p_2+p_4]$$

(2×2)

$$\left[\frac{p_1^2}{p_1+p_2} + \frac{p_3^2}{p_3+p_4} + \frac{p_2p_1}{p_1+p_2} + \frac{p_4p_3}{p_3+p_4} \right] = p_1 + p_3$$

$$\left[p_2 \left(\frac{p_1p_2}{p_1+p_2} + \frac{p_3p_4}{p_3+p_4} \right) + p_4 \left(\frac{p_2^2}{p_1+p_2} + \frac{p_4^2}{p_3+p_4} \right) \right] = p_2 + p_4$$

begin with some illustrations in the bivariate case and then investigate higher dimensional cases.

4.1 The Bivariate Case

Suppose that, for two random variables X and Y, we know the conditional densities $f_{X|Y}(x \mid y)$ and $f_{Y|X}(y \mid x)$. We can determine the marginal density of X, $f_X(x)$, and hence the joint density of X and Y, through the following argument. By definition,

$$f_X(x) = \int f_{XY}(x, y)\, dy,$$

where $f_{XY}(x, y)$ is the (unknown) joint density. Now using the fact that $f_{XY}(x, y) = f_{X|Y}(x \mid y) f_Y(y)$, we have

$$f_X(x) = \int f_{X|Y}(x \mid y) f_Y(y)\, dy,$$

and if we similarly substitute for $f_Y(y)$, we have

$$\begin{aligned} f_X(x) &= \int f_{X|Y}(x \mid y) \int f_{Y|X}(y \mid t) f_X(t)\, dt\, dy \\ &= \int \left[\int f_{X|Y}(x \mid y) f_{Y|X}(y \mid t)\, dy \right] f_X(t)\, dt \\ &= \int h(x, t) f_X(t)\, dt, \end{aligned} \tag{4.1}$$

where $h(x, t) = [\int f_{X|Y}(x \mid y) f_{Y|X}(y \mid t)\, dy]$. Equation (4.1) defines a *fixed point integral equation* for which $f_X(x)$ is a solution. The fact that it is a unique solution is explained by Gelfand and Smith (1990).

Equation (4.1) is the limiting form of the Gibbs iteration scheme, illustrating how sampling from conditionals produces a marginal distribution. As $k \to \infty$ in (3.5),

$$f_{X_k|X_0}(x \mid x_0) \to f_X(x)$$

and

$$f_{X_k|X_{k-1}}(x \mid t) \to h(x, t), \tag{4.2}$$

and thus (4.1) is the limiting form of (3.5).

Although the joint distribution of X and Y determines all of the conditionals and marginals, it is not always the case that a set of proper conditional distributions will determine a proper marginal distribution (and hence a proper joint distribution). The next example shows this.

Example 2 (continued): Suppose that $B = \infty$ in (2.8), so that X and Y have the conditional densities

$$f(x \mid y) = ye^{-yx}, \quad 0 < x < \infty \tag{4.3a}$$

$$f(y \mid x) = xe^{-xy}, \quad 0 < y < \infty. \tag{4.3b}$$

Applying (4.1), the marginal distribution of X is the solution to

$$\begin{aligned} f_X(x) &= \left[\int ye^{-yx} te^{-ty}\, dy \right] f_X(t)\, dt \\ &= \int \left[\frac{t}{(x + t)^2} \right] f_X(t)\, dt. \end{aligned} \tag{4.4}$$

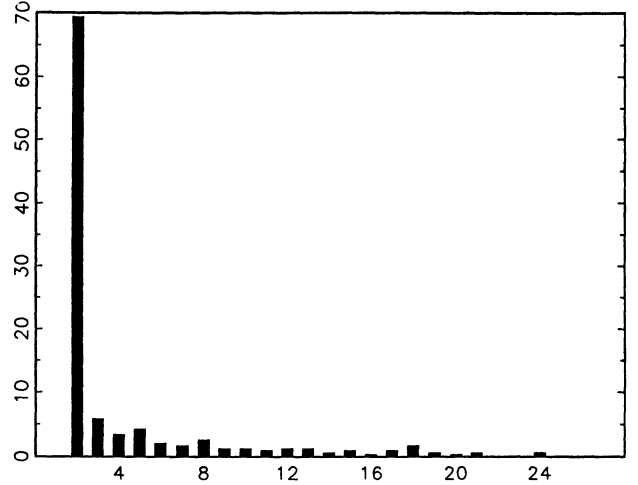


Figure 4. Histogram of a Sample of Size $m = 500$ From the Pair of Conditional Distributions in (4.3), Obtained Using Gibbs Sampling With $k = 10$.

Substituting $f_X(t) = 1/t$ into (4.4) yields

$$\frac{1}{x} = \int \left[\frac{t}{(x + t)^2} \right] \frac{1}{t}\, dt,$$

solving (4.4). Although this is a solution, $1/x$ is not a density function. When the Gibbs sampler is applied to the conditional densities in (4.3), convergence breaks down. It does not give an approximation to $1/x$, in fact, we do not get a sample of random variables from a marginal distribution. A histogram of such random variables is given in Figure 4, which vaguely mimics a graph of $f(x) = 1/x$.

It was pointed out by Trevor Sweeting (personal communication) that Equation (4.1) can be solved using the truncated exponential densities in (2.8). Evaluating the constant in the conditional densities gives $f(x|y) = ye^{-yx}/(1 - e^{-By})$, $0 < x < B$, with a similar expression for $f(y|x)$. Substituting these functions into (4.1) yields the solution $f(x) \propto (1 - e^{-Bx})/x$. This density (properly normalized) is the dashed line in Figure 2.

The Gibbs sampler fails when $B = \infty$ above because $\int f_X(x)dx = \infty$, and there is no convergence as described in (4.2). In a sense, we can say that a sufficient condition for the convergence in (4.2) to occur is that $f_X(x)$ is a proper density, that is $\int f_X(x)dx < \infty$. One way to guarantee this is to restrict the conditional densities to lie in a compact interval, as was done in (2.8). General convergence conditions needed for the Gibbs sampler (and other algorithms) are explored in detail by Schervish and Carlin (1990), and rates of convergence are also discussed by Roberts and Polson (1990).

4.2 More Than Two Variables

As the number of variables in a problem increase, the relationship between conditionals, marginals, and joint distributions becomes more complex. For example, the relationship conditional $\times$ marginal = joint does not hold for all of the conditionals and marginals. This means that there are many ways to set up a fixed-point equation like (4.1), and it is possible to use different sets of conditional distributions to calculate the

marginal of interest. Such methodologies are part of the general techniques of *substitution sampling* (see Gelfand and Smith 1990, for an explanation). Here we merely illustrate two versions of this technique.

In the case of two variables, all substitution sampling algorithms are the same. The three variable case, however, is sufficiently complex to illustrate the differences between algorithms, yet sufficiently simple to allow us to write things out in detail. Generalizing to cases of more than three variables is reasonably straightforward.

Suppose we would like to calculate the marginal distribution $f_X(x)$ in a problem with random variables X, Y, and Z. A fixed-point integral equation like (4.1) can be derived if we consider the pair (Y, Z) as a single random variable. We have

$$f_X(x) = \int \left[\iint f_{X|YZ}(x \mid y, z) f_{YZ|X}(y, z \mid t)\, dy\, dz \right] f_X(t)\, dt, \quad (4.5)$$

analogous to (4.1). Cycling between $f_{X|YZ}$ and $f_{YZ|X}$ would again result in a sequence of random variables converging in distribution to $f_X(x)$. This is the idea behind the Data Augmentation Algorithm of Tanner and Wong (1987). By sampling iteratively from $f_{X|YZ}$ and $f_{YZ|X}$, they show how to obtain successively better approximations to $f_X(x)$.

In contrast, the Gibbs sampler would sample iteratively from $f_{X|YZ}$, $f_{Y|XZ}$, and $f_{Z|XY}$. That is, the jth iteration would be

$$\begin{aligned} X'_j &\sim f(x \mid Y'_j = y'_j, Z'_j = z'_j) \\ Y'_{j+1} &\sim f(y \mid X'_j = x'_j, Z'_j = z'_j) \\ Z'_{j+1} &\sim f(z \mid X'_j = x'_j, Y'_{j+1} = y'_{j+1}). \end{aligned} \quad (4.6)$$

The iteration scheme of (4.6) produces a Gibbs sequence

$$Y'_0, Z'_0, X'_0, Y'_1, Z'_1, X'_1, Y'_2, Z'_2, X'_2, \ldots, \quad (4.7)$$

with the property that, for large k, $X'_k = x'_k$ is effectively a sample point from $f(x)$. Although it is not immediately evident, the iteration in (4.6) will also solve the fixed-point equation (4.5). In fact, a defining characteristic of the Gibbs sampler is that it always uses the full set of univariate conditionals to define the iteration. Besag (1974) established that this set is sufficient to determine the joint (and any marginal) distribution, and hence can be used to solve (4.5).

As an example of a three-variable Gibbs problem, we look at a generalization of the distribution examined in Example 1.

Example 3. In the distribution (2.5), we now let n be the realization of a Poisson random variable with mean λ, yielding the joint distribution

$$f(x, y, n) \propto \binom{n}{x} y^{x+\alpha-1}(1-y)^{nx+\beta-1} e^{-\lambda} \frac{\lambda^n}{n!},$$

$$x = 0, 1, \ldots, n,\ 0 \le y \le 1,\ n = 1, 2, \ldots. \quad (4.8)$$

Again, suppose we are interested in the marginal distribution of X. Unlike Example 1, here we cannot calculate the marginal distribution of X in closed form.

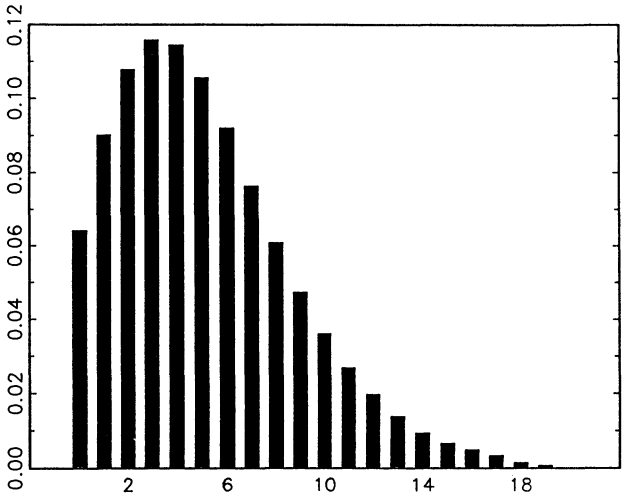


Figure 5. Estimates of Probabilities of the Marginal Distribution of X Using Equation (2.11), Based on a Sample of Size m = 500 From the Three Conditional Distributions in (4.9) With $\lambda = 16$, $\alpha = 2$, and $\beta = 4$. The Gibbs sequences had length k = 10.

However, from (4.8) it is reasonably straightforward to calculate the three conditional densities. Suppressing dependence on λ, α, and β,

$$\begin{aligned} &f(x \mid y, n) \text{ is binomial } (n, y) \\ &f(y \mid x, n) \text{ is beta } (x + \alpha, n - x + \beta) \\ &f(n \mid x, y) \propto e^{-(1-y)\lambda} \frac{[(1-y)\lambda]^{n-x}}{(n-x)!}, \\ &\qquad n = x, x+1, \cdots. \end{aligned} \quad (4.9)$$

If we now apply the iterative scheme (4.6) to the distributions in (4.9), we can generate a sequence $X_1, X_2, \ldots, X_m$ from $f(x)$ and use this sequence to estimate the desired characteristic. The density estimate of $P(X = x)$, using Equation (2.11) can also be constructed. This is done and is given in Figure 5. This figure can be compared to Figure 3, but here there is a longer right tail from the Poisson variability.

The model (4.9) can have practical applications. For example, conditional on n and y, let x represent the number of successful hatchings from n insect eggs, where each egg has success probability y. Both n and y fluctuate across insects, which is modeled in their respective distributions, and the resulting marginal distribution of X is a typical number of successful hatchings among all insects.

5. EXTRACTING INFORMATION FROM GIBBS SEQUENCE

Some of the more important issues in Gibbs sampling surround the implementation and comparison of the various approaches to extracting information from the Gibbs sequence in (2.3). These issues are currently a topic of much debate and research.

5.1 Detecting Convergence

As illustrated in Section 3, the Gibbs sampler generates a Markov chain of random variables which converge to the distribution of interest $f(x)$. Many of the

popular approaches to extracting information from the Gibbs sequence exploit this property by selecting some large value for k, and then treating any X_j' for $j \geq k$ as a sample from $f(x)$. The problem then becomes that of choosing the appropriate value of k.

A general strategy for choosing such k is to monitor the convergence of some aspect of the Gibbs sequence. For example, Gelfand and Smith (1990) and Gelfand, Hills, Racine-Poor, and Smith (1990) suggest monitoring density estimates from m independent Gibbs sequences, and choosing k to be the first point at which these densities appear to be the same under a "felt-tip pen test." Tanner (1991) suggests monitoring a sequence of weights that measure the discrepancy between the sampled and the desired distribution. Geweke (in press) suggests monitoring based on time series considerations. Unfortunately, such monitoring approaches are not foolproof, illustrated by Gelman and Rubin (1991). An alternative may be to choose k based on theoretical considerations, as in Raftery and Banfield (1990). M.T. Wells (personal communication) has suggested a connection between selecting k and the cooling parameter in simulated annealing.

5.2 Approaches to Sampling the Gibbs Sequence

A natural alternative to sampling the kth or final value from many independent repetitions of the Gibbs sequence, as we did in Section 2, is to generate one long Gibbs sequence and then extract every rth observation (see Geyer, in press). For r large enough, this would also yield an approximate iid sample from $f(x)$. An advantage of this approach is that it lessens the dependence on initial values. A potential disadvantage is that the Gibbs sequence may stay in a small subset of the sample space for a long time (see Gelman and Rubin 1991).

For large, computationally expensive problems, a less wasteful approach to exploiting the Gibbs sequence is to use all realizations of X_j' for $j \leq k$, as in George and McCulloch (1991). Although the resulting data will be dependent, it will still be the case that the empirical distribution of X_j' converges to $f(x)$. Note that from this point of view one can see that the "efficiency of the Gibbs sampler" is determined by the rate of this convergence. Intuitively, this convergence rate will be fastest when X_j' moves rapidly through the sample space, a characteristic that may be thought of as mixing. Variations on these and other approaches to exploiting the Gibbs sequence have been suggested by Gelman and Rubin (1991), Geyer (in press), Muller (1991), Ritter and Tanner (1990), and Tierney (1991).

6. DISCUSSION

Both the Gibbs sampler and the Data Augmentation Algorithm have found widespread use in practical problems and can be used by either the Bayesian or classical statistician. For the Bayesian, the Gibbs sampler is mainly used to generate posterior distributions, whereas for the classical statistician a major use is for calculation of the likelihood function and characteristics of likelihood estimators.

Although the theory behind Gibbs sampling is taken from Markov chain theory, there is also a connection to "incomplete data" theory, such as that which forms the basis of the EM algorithm. Indeed, both Gibbs sampling and the EM algorithm seem to share common underlying structure. The recent book by Tanner (1991) provides explanations of all these algorithms and gives many illustrative examples.

The usefulness of the Gibbs sampler increases greatly as the dimension of a problem increases. This is because the Gibbs sampler allows us to avoid calculating integrals like (2.1), which can be prohibitively difficult in high dimensions. Moreover, calculations of the high dimensional integral can be replaced by a series of one-dimensional random variable generations, as in (4.6). Such generations can in many cases be accomplished efficiently (see Devroye 1986; Gilks and Wild 1992; Ripley 1987).

The ultimate value of the Gibbs sampler lies in its practical potential. Now that the groundwork has been laid in the pioneering papers of Geman and Geman (1984), Tanner and Wong (1987), and Gelfand and Smith (1990), research using the Gibbs sampler is exploding. A partial (and incomplete) list includes applications to generalized linear models [Dellaportas and Smith (1990), who implement the Gilks and Wild methodology, and Zeger and Rizaul Karim (1991)]; to mixture models (Diebolt and Robert 1990; Robert 1990; to evaluating computing algorithms (Eddy and Schervish 1990); to general normal data models (Gelfand, Hill, and Lee 1992); to DNA sequence modeling (Churchill and Casella 1991; Geyer and Thompson, in press); to applications in HIV modeling (Lange, Carlin, and Gelfand 1990); to outlier problems (Verdinelli and Wasserman 1990); to logistic regression (Albert and Chib 1991); to supermarket scanner data modeling (Blattberg and George 1991); to constrained parameter estimation (Gelfand et al. 1992); and to capture–recapture modeling (George and Robert 1991).

[*Received December 1990. Revised September 1991.*]

REFERENCES

Albert, J., and Chib, S. (1991), "Bayesian Analysis of Binary and Polychotomous Response Data," technical report, Bowling Green State University, Dept. of Mathematics and Statistics.

Besag, J. (1974), "Spatial Interaction and the Statistical Analysis of Life Systems," *Journal of the Royal Statistical Society*, Ser. B, 36, 192–236.

Blattberg, R., and George, E. I. (1991), "Shrinkage Estimation of Price and Promotional Elasticities: Seemingly Unrelated Equations," *Journal of the American Statistical Association*, 86, 304–315.

Churchill, G., and Casella, G. (1991), "Sampling Based Methods for the Estimation of DNA Sequence Accuracy," Technical Report BU-1138-M, Cornell University, Biometrics Unit.

Dellaportas, P., and Smith, A. F. M. (1990), "Bayesian Inference for Generalized Linear Models," Technical Report, University of Nottingham, Dept. of Mathematics.

Dempster, A. P., Laird, N., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm" (with

discussion), *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Diebolt, J., and Robert, C. (1990), "Estimation of Finite Mixture Distributions Through Bayesian Sampling," technical report, Université Paris VI, L.S.T.A.

Devroye, L. (1986), *Non-Uniform Random Variate Generation*, New York: Springer-Verlag.

Eddy, W. F., and Schervish, M. J. (1990), "The Asymptotic Distribution of the Running Time of Quicksort," technical report, Carnegie Mellon University, Dept. of Statistics.

Efron, B. (1982), *The Jackknife, the Bootstrap, and Other Resampling Plans*, National Science Foundation-Conference Board of the Mathematical Sciences Monograph 38, Philadelphia: SIAM.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. M. (1990), "Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling," *Journal of the American Statistical Association*, 85, 972–985.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Gelfand, A. E., Smith, A. F. M., and Lee, T-M. (1992), "Bayesian Analysis of Constrained Parameter and Truncated Data Problems Using Gibbs Sampling," *Journal of the American Statistical Association*, 87, 523–532.

Gelman, A., and Rubin, D. (1991), "An Overview and Approach to Inference From Iterative Simulation," Technical Report, University of California-Berkeley, Dept. of Statistics.

Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

George, E. I., and McCulloch, R. E. (1991), "Variable Selection via Gibbs Sampling," Technical Report, University of Chicago, Graduate School of Business.

George, E. I., and Robert, C. P. (1991), "Calculating Bayes Estimates for Capture–Recapture Models," Technical Report 94, University of Chicago, Statistics Research Center, and Cornell University, Biometrics Unit.

Geweke, J. (in press), "Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments," *Proceedings of the Fourth Valencia International Conference on Bayesian Statistics*.

Geyer, C. J. (in press), "Markov Chain Monte Carlo Maximum Likelihood," *Computer Science and Statistics: Proceedings of the 23rd Symposium on the Interface*.

Geyer, C. J., and Thompson, E. A. (in press), "Constrained Maximum Likelihood and Autologistic Models With and Application to DNA Fingerprinting Data, *Journal of the Royal Statistical Society*, Ser. B.

Gilks, W. R., and Wild, P. (1992), "Adaptive Rejection Sampling for Gibbs Sampling," *Journal of the Royal Statistical Society*, Ser. C, 41, 337–348.

Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97–109.

Hoel, P. G., Port, S. C., and Stone, C. J. (1972), *Introduction to Stochastic Processes*, New York: Houghton-Mifflin.

Lnage, N., Carlin, B. P., and Gelfand, A. E. (1990), "Hierarchical Bayes Models for the Progression of HIV Infection Using Longitudinal CD4+ Counts," technical report, Carnegie Mellon University, Dept. of Statistics.

Liu, J., Wong, W. H., and Kong, A. (1991), "Correlation Structure and Convergence Rate of the Gibbs Sampler (I): Applications to the Comparison of Estimators and Augmentation Schemes," Technical Report 299, University of Chicago, Dept. of Statistics.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), "Equations of State Calculations by Fast Computing Machines," *Journal of Chemical Physics*, 21, 1087–1091.

Muller, P. (1991), "A Generic Approach to Posterior Integration and Gibbs Sampling," Technical Report 91-09, Purdue University, Dept. of Statistics.

Raftery, A. E., and Banfield, J. D. (1990), "Stopping the Gibbs Sampler, the Use of Morphology, and Other Issues in Spatial Statistics," technical report, University of Washington, Dept. of Statistics.

Ripley, B. D. (1987), *Stochastic Simulation*, New York: John Wiley.

Ritter, C., and Tanner, M. A. (1990), "The Griddy Gibbs Sampler," technical report, University of Rochester.

Robert, C. P. (1990), "Hidden Mixtures and Bayesian Sampling," technical report, Université Paris VI, L.S.T.A.

Roberts, G. O., and Polson, N. G. (1990), "A Note on the Geometric Convergence of the Gibbs Sampler," Technical Report, University of Nottingham, Dept. of Mathematics.

Schervish, M. J., and Carlin, B. P. (1990), "On the Convergence of Successive Substitution Sampling," Technical Report 492, Carnegie Mellon University, Dept. of Statistics.

Smith, A. F. M., and Gelfand, A. E. (1992), "Bayesian Statistics Without Tears: A Sampling–Resampling Perspective, *The American Statistician*, 46, 84–88.

Tanner, M. A. (1991), *Tools for Statistical Inference*, New York: Springer-Verlag.

Tanner, M. A., and Wong, W. (1987), "The Calculation of Posterior Distributions by Data Augmentation" (with discussion), *Journal of the American Statistical Association*, 82, 528–550.

Tierney, L. (1991), "Markov Chains for Exploring Posterior Distributions," Technical Report 560, University of Minnesota, School of Statistics.

Verdinelli, I., and Wasserman, L. (1990), "Bayesian Analysis of Outlier Problems Using the Gibbs Sampler," Technical Report 469, Carnegie Mellon University, Dept. of Statistics.

Zeger, S., and Rizaul Karim, M. (1991), "Generalized Linear Models With Random Effects: A Gibbs Sampling Approach," *Journal of the American Statistical Association*, 86, 79–86.

Markov Chain example

1Mil population = agents

3 restaurants B, M, S

Fixed transition probab

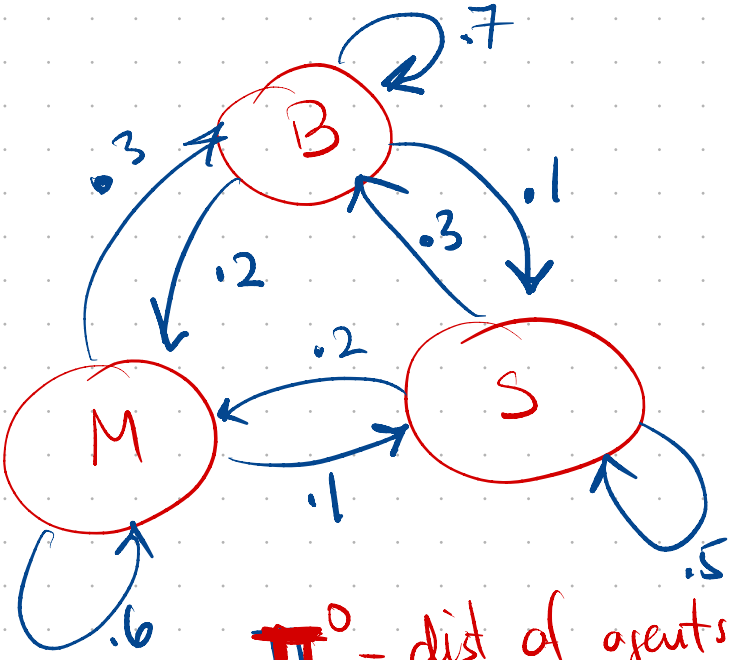
days

eat every day: 0, 1, 2, ...

B/M/S restaurant day k → B/M/S restaurant day k+1 → S states

per agent

transition



transition probab P matrix

	B	M	S
B →	.7	.2	.1
M →	.3	.6	.1
S →	.3	.2	.5

$P(\text{B next day} \mid \text{S this day})$

π^0 = dist of agents to restaurants initially (day 0)

$[\pi^0_B \quad \pi^0_M \quad \pi^0_S]$

π^1 = dist of agents in day 1 = $[\pi^1_B \quad \pi^1_M \quad \pi^1_S]$

π^1_B = fraction agents day 1 at B = $\pi^0_B \cdot P(B \to B) + \pi^0_M \cdot P(M \to B) + \pi^0_S \cdot P(S \to B)$

eat day before at B and trans (B→B) + eat day before at M + transition (B→B)

$$\pi^{\text{next day}}_{k+1} = \pi^{\text{prev day}}_{k} \cdot P$$

prev day dist · fixed transition probab.

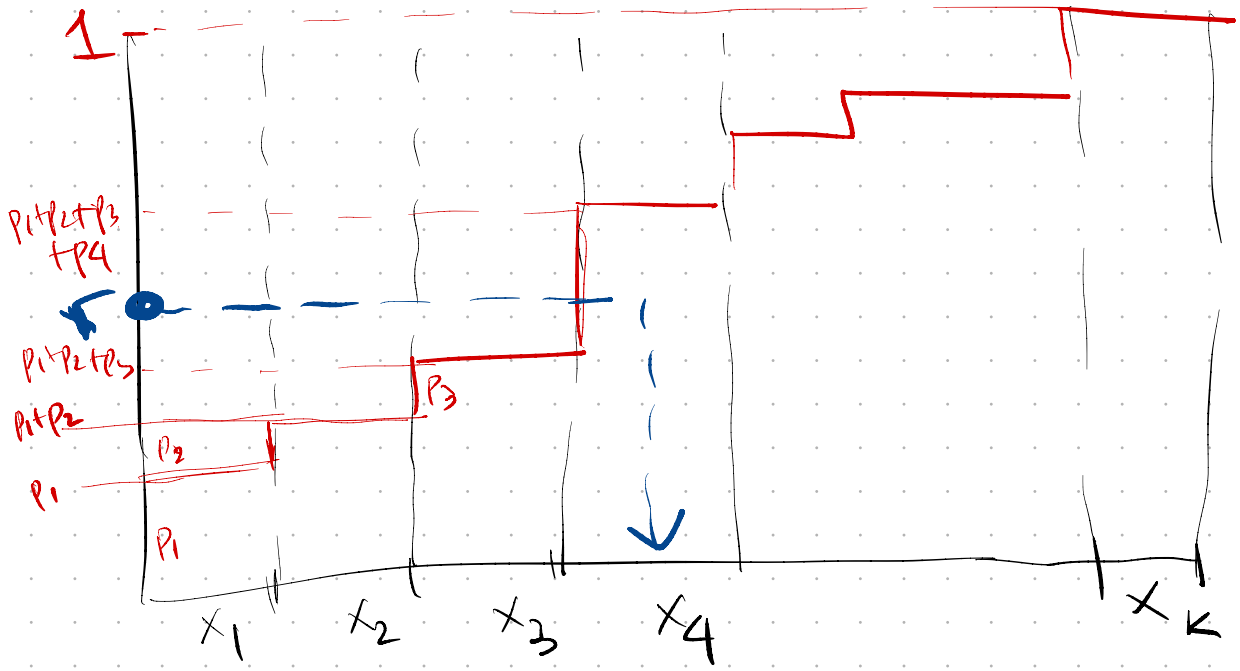
Markov Th: except regular erdos-renyi (cycles) very often $\pi \to \pi^*$ converges stationary dist

$\pi^* = \pi^* \cdot P \Rightarrow \pi^*$ eigen vector for P corresp to eigenval $\lambda = 1$

Sampling Techniques

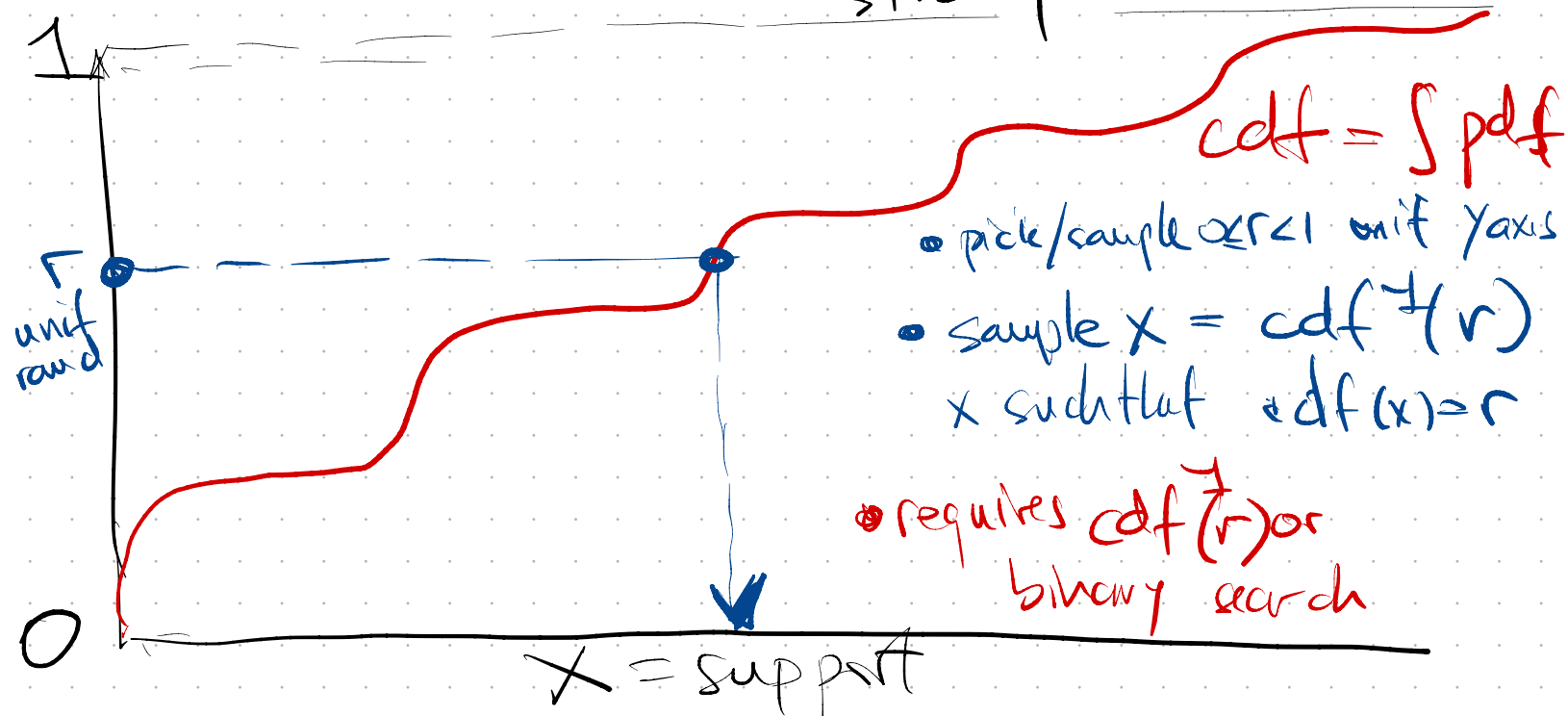
discrete dist outcomes x_1 x_2 — x_k
prob p_1 p_2 p_k

$\sum p_i = 1$ $\Rightarrow$ look at CDF = cumulative dist



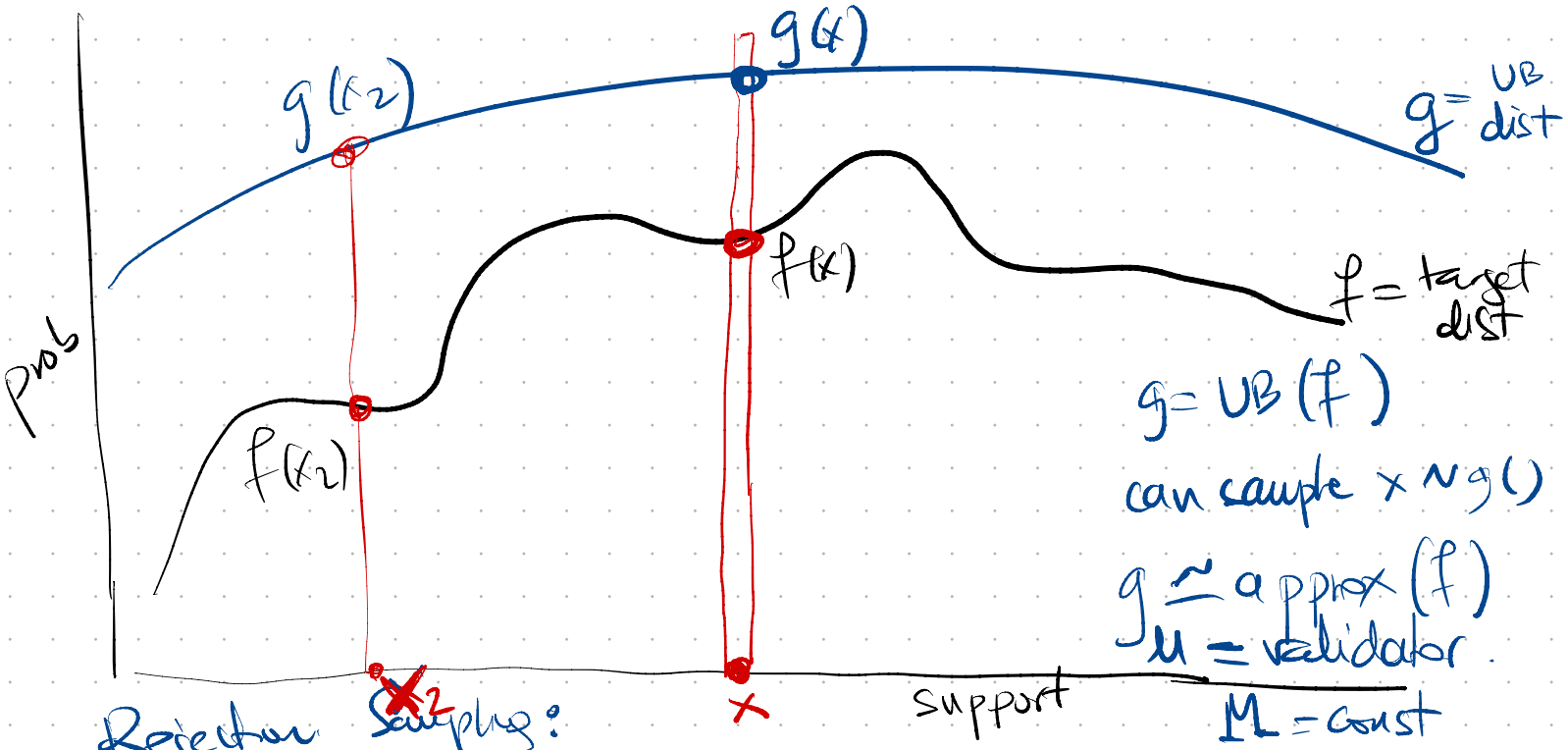
- **Inverse sampling**: - select on y axis random number $0 \le r < 1$
  - apply $cdf^{-1}(r)$ cdf-inverse
  - implement with binary search

- untrivial pdf $\Rightarrow$ continuous cdf strictly monotonic



Rejection Sampling $x \sim f(x)$ pdf

- can compute $f(\text{any } x)$
- cannot inverse $cdf(x)$ or computationally expensive.



Rejection Sampling:

- sample $x \sim g()$
- sample validator $u = unif(0,1)$ for this sample, i.m

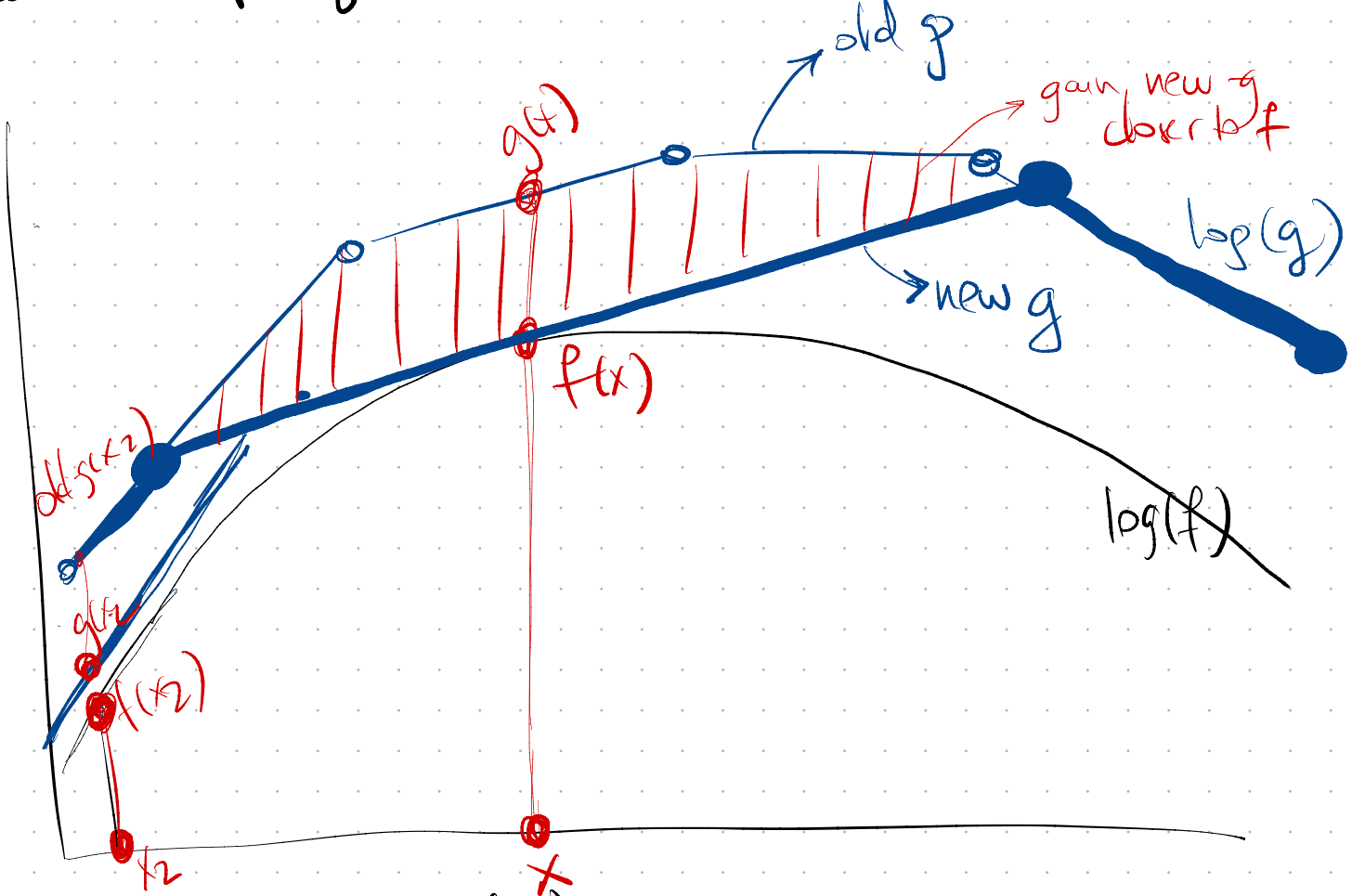
[if $u < \frac{f(x)}{g(x) \cdot M} \Rightarrow$ accept x as coming from $f()$ dist

Else $u > \frac{f(x)}{g(x) M} \Rightarrow$ reject x, repeat

Adaptive Rejection Sampling

- f = log concave $\Leftrightarrow$ $\log(f)$ anticonvex | easy
- $g = UB(f)$ = cover of line segments
- can sample g

f is not growing faster than exp



- sample $X \sim g(x)$
- u = validator $\sim unif(0,1)$
- if $u \leq \frac{f(x)}{g(x) \cdot M}$ $\Rightarrow$ accept
- if $u > \frac{f(x)}{g(x) \cdot M}$ $\Rightarrow$ reject; update g-segment(x) = segment tangent to $f(x)$

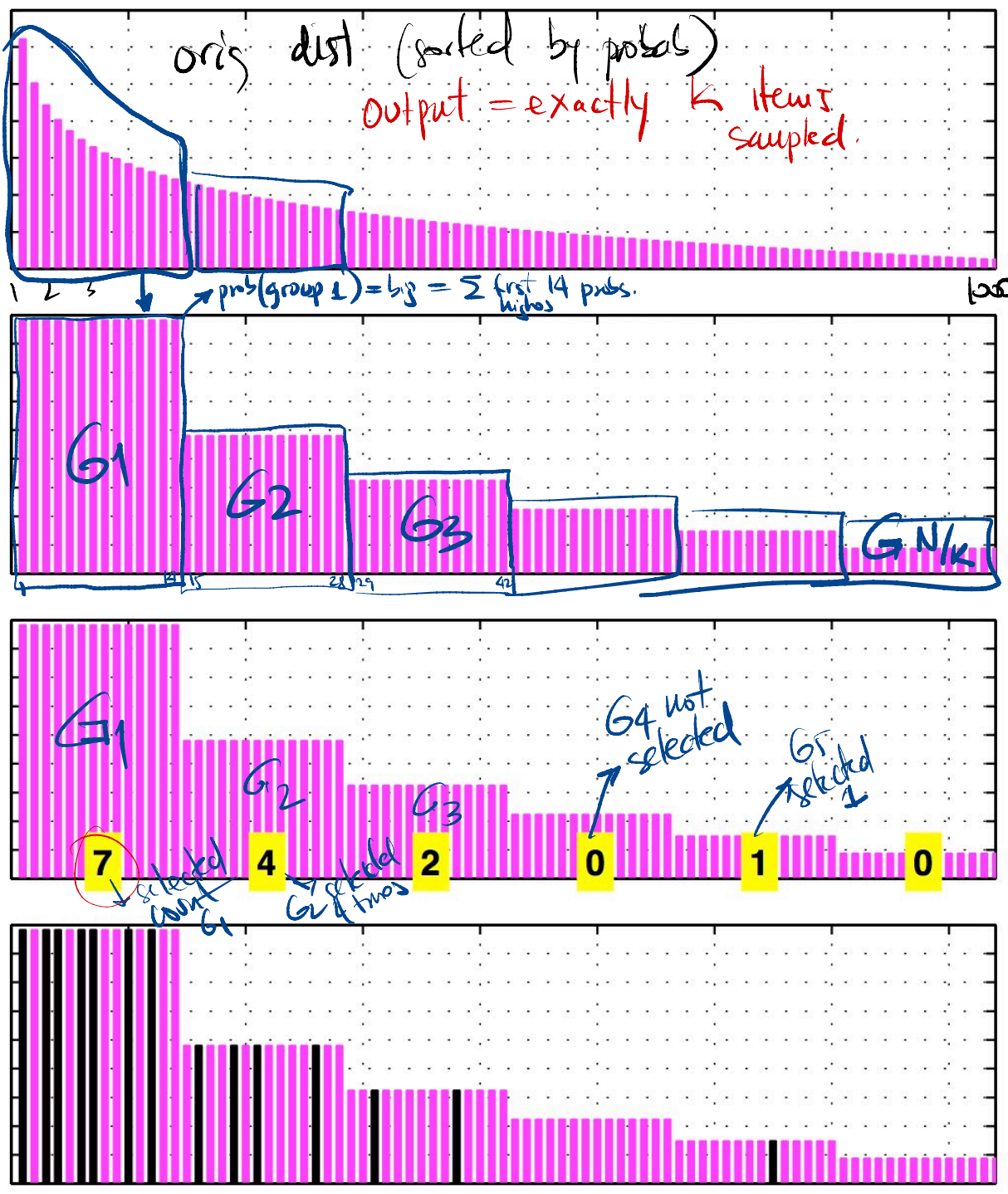
- intersect that segment with old-g
- new g is better approx for f
  ⇒ closer, less samples x are rejected from here
- new segment (update-g) still UB to f
- lower bound glow extrude
- skip some f(x) computation (g^{UB} / g^{LB})

- Pb sample K=14 items from a dist non-uniform over N=1000 items, without repetion
- if possible? $prob_{selection}(item) \simeq prob_{given}(item)$
- not possible more when K→large.
  - if $K \simeq N$ all item ⇒ selection prob $\simeq 1$
- feasible? if dist (sorted) is rather smooth

(1) in sorted order make groups of size K=14
in each group ignore individual-item probs, just use
group-prob = Σ item-probs.

(2) sample K times groups ~ group-probs (non-unif) with repetition

(3) sample each group without rep, unif for counts at (2)


orig dist (sorted by probs)
output = exactly k items sampled.
prob(group 1) = b_1 = Σ first 14 probs.
G1
G2
G3
G N/k
G4 not selected
G5 selected 1
7
4
2
0
1
0