Lower Bounds for the Gibbs Sampler over Mixtures of Gaussians

Christopher Tosh CTOSH@CS.UCSD.EDU
Sanjoy Dasgupta DASGUPTA@CS.UCSD.EDU
Department of Computer Science and Engineering, UCSD, 9500 Gilman Drive, La Jolla, CA 92093-0404

Abstract

The *mixing time* of a Markov chain is the minimum time t necessary for the total variation distance between the distribution of the Markov chain's current state X_t and its stationary distribution to fall below some $\epsilon > 0$. In this paper, we present lower bounds for the mixing time of the Gibbs sampler over Gaussian mixture models with Dirichlet priors.

1. Introduction

Inferring the parameters of a mixture model based on observed data is a classical problem in machine learning that has received much attention from computer scientists and statisticians alike. One of the first computational approaches to this problem was given by Dempster, Laird, and Rubin (1977) in the form of the popular EM algorithm. The goal of their algorithm was to find the parameters which maximized the likelihood of the observed data. While their algorithm is only guaranteed to converge to a local maximum (Wu, 1983), others have demonstrated efficient algorithms that recover the true parameters of mixtures of various distributions (Moitra & Valiant, 2010; Belkin & Sinha, 2010; Hsu & Kakade, 2013).

Alternatively, the Bayesian approach to the problem views the parameters to be inferred not as fixed quantities, but as random variables generated by some underlying distribution. By placing a prior distribution on the unknown quantities and using Bayes' rule to update the prior with the likelihood of the observed data, a posterior distribution is obtained for the desired parameters. Since in many cases of interest, the normalizing factor of this posterior distribution is not available in closed form, the typical approach to inference is via Monte Carlo methods, specifically Markov chain Monte Carlo methods (MCMC) (Gelman et al., 1995). One of the most popular MCMC methods is Gibbs sampling (Geman & Geman, 1984).

Gibbs sampling, often referred to as Glauber dynamics (Levin et al., 2008) or alternating conditional sampling (Gelman et al., 1995), is a generic MCMC method that relies on knowing only the conditional marginal probabilities of the unknown parameters. More concretely, if $X_1, \ldots, X_k$ are the random variables of interest and the current state of the Gibbs sampler is $(x_1, \ldots, x_k)$, then the Gibbs sampler takes a step by choosing an index i and updating the value of X_i according to $Pr(X_i = x \,|\, X_{-i} = x_{-i})$, where the negative subscript refers to the vector of values missing that index. There are two ways of selecting the indices to be updated. One can choose the indices sequentially, the systematic scan method, or update them randomly, the random scan method. It is an open question by how much these two methods can differ in their rates of convergence (Diaconis, 2013). In this paper, we will only consider the random scan Gibbs sampler.

Regardless of whether the systematic or random scan method is used, it is well known that in the limit, as the number of updates goes to infinity, the distribution of the state of the Gibbs sampler converges to the desired posterior distribution (Geman & Geman, 1984; Diebolt & Robert, 1994). What is less understood is how long it takes for the distribution of the Gibbs sampler's state to be within a reasonable distance of the posterior, the so-called *mixing time* of the Markov chain (Levin et al., 2008), sometimes referred to as the *burn-in period* (Brooks, 1998). The mixing rate of the Gibbs sampler can vary wildly depending on the application, from nearly linear (Jerrum, 1995; Luby & Vigoda, 1999) to exponential (Randall, 2006; Galvin & Randall, 2007). Thus, to get meaningful bounds on the mixing rate, we need to consider the specific application we have in mind. Here, it is learning mixture models.

The Bayesian approach to specifying mixture models is to provide a generative process by which the observable quantities are created. We follow a widely used generative model seen, for example, in Neal (2000). For

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

a mixture model of k components with density function $P(x) = w_1P_1(x) + \ldots + w_kP_k(x)$, we assume that our mixing weights $(w_1, \ldots, w_k)$ are drawn from a symmetric k-dimensional Dirichlet distribution with single parameter $\alpha > 0$. Each of the P_i is parameterized by $\theta_i \in \Theta$ which is drawn from a prior distribution with parameter β. We call this $\mathcal{Q}(\beta)$ The label z_i for each data point is drawn from the categorical distribution with parameter $w = (w_1, \ldots, w_k) \in \Delta_k$. Finally, the point x_i is drawn from the mixture component parameterized by θ_{z_i}. We call this $\mathcal{P}(\theta_{z_i})$. This is summarized in (1).

$$\begin{aligned} (w_1, \ldots, w_k) &\sim \text{Dirichlet}(\alpha, \ldots, \alpha) \\ \theta_1, \ldots, \theta_k &\sim \mathcal{Q}(\beta) \\ z_i &\sim \text{Categorical}(w_1, \ldots, w_k) \\ x_i &\sim \mathcal{P}(\theta_{z_i}) \end{aligned} \tag{1}$$

Diebolt and Robert (1994) showed that given the sequence $(x_1, \ldots, x_n)$, the number of generating distributions k, the prior distribution $\mathcal{Q}$, and the likelihood distribution $\mathcal{P}$, the Gibbs sampler could be used to sample from the posterior distribution $Pr(w, \theta, z \mid x)$. To do this, their Gibbs sampler alternated between sampling the distributional parameters w and θ and the labellings z. This is known as the *naïve Gibbs sampler*. Often it is sufficient to sample only the label variables z. In the case where the prior is conjugate to the likelihood, it is possible to integrate out the distributional parameters w and θ and to only sample the labellings z. This is called the *collapsed Gibbs sampler*.

Both of the above Gibbs samplers on mixture models suffer from the issue of *identifiability*. That is, any permutation of the label variable values z_i identifies to exactly the same underlying structure, but the Gibbs sampler still considers them to be distinct. Thus, any grouping of points into k clusters is represented in the space of labellings $k!$ times, which can lead to slow mixing. In the statistics literature, this is sometimes referred to as the 'label switching problem' (Jasra et al., 2005). Since all the permutations contain the same basic information, the particular labellings are not important to us. Can something be said about how well the Gibbs sampler performs when we only consider the structure underlying all permutations of the labellings?

In this paper, we show that the collapsed Gibbs sampler induces another Markov chain over the space of equivalence classes of the labelling permutations. This space contains a unique representative for each of the equivalence classes and therefore does not suffer from the issue of identifiability. Further, the induced Markov chain has as its stationary distribution the exact posterior distribution that we wish to sample from. However, even with these advantages, the induced Markov chain is not guaranteed to mix rapidly.

We provide lower bounds on the rate of convergence for the Gibbs sampler over two instances. In both examples, the clusters look Gaussian. In one of the cases the number of Gaussians is misspecified, and we show that the Markov chain is still very far from stationarity after exponentially many steps. In the other case the number of Gaussians is correctly specified, and we show that the Markov chain must take at least $n^{\Omega(\alpha)}$ steps before it approaches stationarity, where $\alpha > 0$ is a sparsity parameter of the prior distribution on the mixing weights.

1.1. High-Level Overview

We are particularly interested in mixtures of spherical Gaussians with fixed variance σ^2. To generate our data point sequence $x = (x_1, \ldots, x_n)$, where each $x_i \in \mathbb{R}^d$, we follow the generative process described in (1). Here, our prior distribution is a spherical Gaussian $\mathcal{N}(\mu_0, \sigma_0^2 I_d)$. From this distribution, we draw means $\mu_1, \ldots, \mu_k$, which are the parameters of our likelihood distribution, $\mathcal{N}(\cdot, \sigma^2 I_d)$. The process is summarized in (2).

$$\begin{aligned} (w_1, \ldots, w_k) &\sim \text{Dirichlet}(\alpha, \ldots, \alpha) \\ \mu_1, \ldots, \mu_k &\sim \mathcal{N}(\mu_0, \sigma_0^2 I_d) \\ z_i &\sim \text{Categorical}(w_1, \ldots, w_k) \\ x_i &\sim \mathcal{N}(\mu_{z_i}, \sigma^2 I_d) \end{aligned} \tag{2}$$

The above process has several variables that need to be specified: the number of clusters k, the sparsity of the Dirichlet distribution α, the mean and variance of the Gaussian prior (μ_0, σ_0^2), and the variance of the mixing Gaussians σ^2. This flexibility means that the above process can be used to describe a wide variety of data sets.

Given a setting of these variables, the collapsed Gibbs sampler is a random walk on the space of labellings $(z_1, \ldots, z_n)$. At each step, it randomly chooses an index i and sets z_i to label j with probability $Pr(z_i = j \mid z_{-i}, x)$ where z_{-i} indicates the vector z minus the ith coordinate. Starting with an initial configuration $z^{(0)} \in \{1, \ldots, k\}^n$, this process generates successive labellings $z^{(0)}, z^{(1)}, \ldots$ that distributionally approach the probability distribution $Pr(z \mid x)$. That is, if the distribution of the sample $z^{(t)}$ is given by π_t, then $\lim_{t \to \infty} \pi_t(\cdot) = Pr(\cdot \mid x)$.

However, the above does not tell us about the rate at which convergence occurs. For this, we need a notion of distance for probability distributions. If ν and μ are probability measures over a space Ω, the total variation distance is

$$\|\mu - \nu\|_{TV} := \max_{A \subset \Omega} |\mu(A) - \nu(A)| = \frac{1}{2} \sum_{\omega \in \Omega} |\mu(\omega) - \nu(\omega)|.$$

If π_t is defined as above, then the mixing rate of the Gibbs sampler is the minimum number of step τ_{mix} to lower the total variation distance between π_t and $Pr(\cdot \mid x)$ below 1/4.

To establish lower bounds on the mixing rate of the Gibbs sampler, it is sufficient to establish the result for any family

of point sets. However, such results are particularly interesting when the data points really do approximately conform to the model. As an approximation to a Gaussian, we will consider a cluster of radius r centered about $x \in \mathbb{R}^d$ to be a collection of points within distance r of their mean, x. We consider two cases of point sets and Gibbs samplers and provide lower bounds on the mixing rate of both.

In the first case, we consider a collection of 6 identical clusters of n points each lying in d-dimensional space with $d \geq 3$. The clusters are positioned such that no two cluster means are within distance r of each other and the diameter of each cluster is bounded above by δr. However, when we specify the number of Gaussians for the Gibbs sampler, we use 3 instead of 6. In Section 6.1, we prove the following.

Theorem 6.1 *For a proper setting of δ, the mixing time of the induced Gibbs sampler with a misspecified number of mixtures is bounded as $\tau_{mix} \geq (1/24)\exp(r^2/8\sigma^2)$.*

It is worth mentioning that the ratio r/σ can be arbitrarily large. In fact, when the points comprising the individual clusters are drawn from a Gaussian with variance σ^2, a larger value for r/σ actually corresponds to a more well-separated instance, which should be easier to learn, not harder. It is interesting that the Gibbs sampler should perform worse as the problem instance gets more tractable.

In the second case, where we correctly specify the number of clusters, we consider a collection of 3 identical clusters of n points each lying in d-dimensional space with $d \geq 2$. The means of these clusters lie at a distance r from each other and the diameter of each cluster is bounded above by δr. In Section 6.2, we prove the following lower bound.

Theorem 6.2 *For a proper setting of δ and α, the mixing time of the induced Gibbs sampler with a correctly specified number of mixtures is bounded below as*

$$\tau_{mix} \geq \frac{1}{8}\min\left(\frac{1}{6}e^{\left(\frac{r^2}{96\sigma^2}\right)}, \frac{n^{\alpha-d/2}\left(\frac{\sigma}{\sigma_0}\right)^d \exp\left(\frac{\alpha-\alpha^2}{n}\right)}{2^{3(\alpha-1/2)}\Gamma(\alpha)\exp\left(\frac{r^2}{\sigma_0^2}\right)}\right).$$

If the means are drawn from a d-dimensional normal distribution with variance σ_0^2, then with high probability $r^2 \approx d\sigma_0^2$. In this case, the dominant term in Theorem 6.2 is the minimum of $\exp(r^2/96\sigma^2) \approx \exp(d\sigma_0^2/\sigma^2)$ and $n^{\alpha-d/2}\left(\sigma/\sigma_0\right)^d$. Additionally, to get clusters that are even, or nearly even, in cardinality like the ones we consider, the parameter α must be relatively large. Thus, even though we are taking the minimum of two quantities, we can still expect the lower bound to be fairly large for reasonable n.

The key idea in the proofs of both of these lower bounds is that there exists a subset of states of the Gibbs sampler that has relatively low probability mass but is difficult for the Gibbs sampler to transition out of. In Section 7 we experimentally verify that this is the case by showing that most runs of the Gibbs sampler spend the majority of their time in local optima with relatively low probability mass.

2. Preliminaries

In this section, we give a more general review of the theory of Markov chains. A Markov chain is a sequence of random variables $X_0, X_1, \ldots$ taking values in a state space Ω s.t.

$$Pr(X_t = x \,|\, X_{t-1}, \ldots, X_0) = Pr(X_t = x \,|\, X_{t-1})$$

for all $x \in \Omega$ and $t \geq 1$. We can view the transition probability as a matrix P indexed by elements of Ω s.t.

$$P(x, y) = Pr(X_t = y \,|\, X_{t-1} = x).$$

Note that if X_0 is some fixed state $x \in \Omega$, the distribution of X_t is $P^t(x, \cdot)$. A Markov chain is said to be *irreducible* or *strongly connected* if, for all $x, y \in \Omega$, there exists a $t > 0$ s.t. $P^t(x, y) > 0$. It is *aperiodic* if for all $x, y \in \Omega$,

$$\gcd(\{t : P^t(x, y) > 0\}) = 1.$$

A distribution π over Ω is said to be a *stationary distribution* of P if, when π and P are viewed as matrices indexed by Ω, then π is a left eigenvector of P with corresponding eigenvalue 1 or, equivalently, $\pi = \pi P$. The natural interpretation of this is that if the current state of the Markov chain is distributed according to π, then the next state will also be distributed according to π. A sufficient, but not necessary, condition for a distribution to be stationary with respect to a Markov chain is *reversability*. A Markov chain P is reversible with respect to a distribution π if for all $x, y \in \Omega$, we have $\pi(x)P(x, y) = \pi(y)P(y, x)$.

How do we measure the 'closeness' of two distributions? One measure is the *total variation distance*. The total variation distance between two distributions μ, ν over Ω is

$$\|\mu - \nu\|_{TV} := \max_{A \subset \Omega} |\mu(A) - \nu(A)| = \frac{1}{2}\sum_{\omega \in \Omega} |\mu(\omega) - \nu(\omega)|.$$

For $\epsilon > 0$, the *mixing time* of a Markov chain P with unique stationary distribution π is

$$\tau(\epsilon) = \min\{t : \max_{x \in \Omega} \|P^t(x, \cdot) - \pi\|_{TV} < \epsilon\}.$$

Taking ϵ to be any constant less than $1/2$ gives us nontrivial results, but by convention ϵ is often taken to be $1/4$. Thus, where it will cause no confusion, we refer to the mixing time interchangeably with the quantity $\tau_{mix} := \tau(1/4)$. Given a Markov chain P, its stationary distribution π, and a subset $S \subset \Omega$, the *conductance of S* is

$$\Phi(S) := \frac{1}{\pi(S)} \sum_{x \in S, y \in S^c} \pi(x)P(x, y)$$

and the *conductance of* P, Φ^*, is the minimum conductance of any set S with $\pi(S) \leq 1/2$. The following theorem relates the mixing time and conductance of a Markov chain.

Theorem 2.1 (Sinclair (1988))**.** *Let P be an aperiodic, irreducible, and reversible Markov chain with conductance Φ^* and mixing time τ_{mix}. Then, $\tau_{mix} \geq \frac{1}{4\Phi^*}$.*

3. Mixture Models and Gibbs Sampling

In this paper, we are concerned with Bayesian inference of the parameters of a mixture distribution. Generally speaking, a *mixture distribution* is specified by k probability density functions $P_1, \ldots, P_k : \Omega \to \mathbb{R}$ and k corresponding weights $w_1, \ldots, w_k$ s.t. $w_1 + \cdots + w_k = 1$ and has probability density function $P(x) = w_1 P_1(x) + \ldots + w_k P_k(x)$.

We are particularly interested in a certain class of mixture models: finite mixtures of exponential families of distributions. For this section and the next, we will work in this generality before addressing Gaussian mixture models.

3.1. Generative Model

We now summarize the generative model we consider in this paper, which can be seen, for example, in Neal (2000). For a mixture model of k components, we assume that our mixing weights $(w_1, \ldots, w_k)$ are drawn from a symmetric k-dimensional Dirichlet distribution with single parameter $\alpha > 0$. This is a distribution over the k-simplex,

$$\Delta_k = \left\{ (w_1, \ldots, w_k) \in \mathbb{R}^k \;\middle|\; \sum_{i=1}^{k} w_i = 1 \right\},$$

and has probability density function

$$D_\alpha(w_1, \cdots, w_k) = \frac{\Gamma(k\alpha)}{\Gamma(\alpha)^k} \prod_{i=1}^{k} w_i^{\alpha-1}.$$

Here, $\Gamma(\cdot)$ is the gamma function. Each of the parameters, $\theta_i \in \Theta$, of the mixing distributions is drawn from the same prior distribution parameterized by some vector $\beta \in \mathbb{R}^s$. Call this $\mathcal{Q}(\beta)$ and its probability density function $Q_\beta : \Theta \to \mathbb{R}$. The label z_i for each data point is drawn from the categorical distribution with parameter $w = (w_1, \ldots, w_k)$. The k-dimensional categorical distribution is defined over the discrete set $\{1, \ldots, k\}$ and has probability mass function

$$C_w(i) = w_i \text{ for } i \in \{1, \ldots, k\}.$$

Finally, the point x_i is drawn from the distribution parameterized by θ_{z_i}. Call this $\mathcal{P}(\theta_{z_i})$ and its probability density function $P_{\theta_{z_i}}$. This can be summarized as the following.

$$\begin{aligned} (w_1, \ldots, w_k) &\sim \text{Dirichlet}(\alpha, \ldots, \alpha) \\ \theta_1, \ldots, \theta_k &\sim \mathcal{Q}(\beta) \\ z_i &\sim \text{Categorical}(w_1, \ldots, w_k) \\ x_i &\sim \mathcal{P}(\theta_{z_i}) \end{aligned} \tag{1}$$

Algorithm 1 The collapsed Gibbs sampler, P.

Initialize $z_1, \ldots, z_n \in \{1, \ldots, k\}$
while *true* **do**
  Choose i u.a.r. from $\{1, \ldots, n\}$
  Update z_i according to $Pr(z_i = j \mid z_{-i}, x_1, \ldots, x_n)$
end while

Suppose that we produce a sequence $x = (x_1, \ldots, x_n)$ from the above generative process. Then for any $z = (z_1, \ldots, z_n) \in \{1, \ldots, k\}^n, \theta = (\theta_1, \ldots, \theta_k) \in \Theta^k$, and $w = (w_1, \ldots, w_k) \in \Delta_k$, we have the joint distribution

$$Pr(x, z, \theta, w) = \frac{\Gamma(k\alpha)}{\Gamma(\alpha)^k} \prod_{j=1}^{k} w_j^{\alpha-1} Q_\beta(\theta_j) \prod_{i=1}^{n} (w_{z_i} P_{\theta_{z_i}}(x_i)).$$

We denote by $C_j(z)$ the set of indices i for which $z_i = j$ and by $n(z)$ the vector whose jth element is $|C_j(z)|$. Here we think about $C_j(z)$ as being the jth 'cluster.'

For a subset S of $\{1, \ldots, n\}$, we let $P_\theta(S)$ denote the probability of S under the specific model $\theta \in \Theta$:

$$P_\theta(S) := \prod_{i \in S} P_\theta(x_i)$$

and let $q(S)$ denote the probability of S given $\theta \sim \mathcal{Q}(\beta)$:

$$q(S) := \int_\Theta Q_\beta(\theta) P_\theta(S)\, d\theta.$$

In many contexts, we are interested in the probability of a labelling z given a data sequence x. By Bayes' theorem,

$$Pr(z|x) \propto \prod_{j=1}^{k} \left(\frac{\Gamma(n_j(z) + \alpha)}{\Gamma(\alpha)} q(C_j(z)) \right). \tag{3}$$

Denote $Pr(z|x)$ by $\pi(z)$. Even when q is computable in closed form, there are no known exact methods for computing the normalizing factor of π. However, it is often enough to approximately sample from π with the Gibbs sampler.

3.2. The Collapsed Gibbs Sampler

The traditional collapsed Gibbs sampler is shown in Algorithm 1. The following lemma establishes that Algorithm 1 does indeed converge to the desired distribution.

Lemma 3.1. *Let P denote the collapsed Gibbs sampler, π denote the conditional probability distribution in (3), and assume that $\mathcal{P}(\theta) > 0$ everwhere. Then P is irreducible, aperiodic, and reversible with respect to π. In particular, π is the unique stationary distribution of P.*

We still need to compute $Pr(z_i = j \mid z_{-i}, x)$. Let S be a subset of indices and i be an index. Then we define

$$\Delta(S, i) := \frac{q(S \cup \{i\})}{q(S \setminus \{i\})}.$$

With this notation, we can prove the following lemma.

Algorithm 2 The projected Gibbs sampler, $P^\flat$.

Initialize a clustering $\mathbb{C} \in \Omega_{\leq k}(x)$
while *true* **do**
  Choose i u.a.r. from $\{1, \ldots, n\}$
  Move i to $S \in \mathbb{C}$ with probability proportional to $(\alpha + |S \setminus \{i\}|)\Delta(S, i)$
  Move i to own set with probability proportional to $(k - |\mathbb{C}|) \cdot \alpha \cdot q(\{i\})$[1]
end while

Lemma 3.2. *$Pr(z_i = j \mid z_{-i}, x)$ is proportional to $(\alpha + n_j(z_{-i}))\Delta(C_j(z), i)$.*

4. Markov Chains and Equivalence Classes

Identifiability makes it difficult to analyze the mixing time of P. If σ is a permutation over $\{1, \ldots, k\}$, then z and $\sigma(z) = (\sigma(z_1), \ldots, \sigma(z_n))$ hold the same information for us. We are interested in the clustering of the points, not the specific number assigned to each cluster. However, P views z and $\sigma(z)$ as separate states. Thus, mixing results proved over the labelling space may not hold true for the space we care about. We will now see how to factor out this extraneous information by a suitable projection.

4.1. Equivalence Classes of Markov Chains

Consider the following setting: we have a state space Ω and an equivalence relation $\sim$ on Ω. Let $(X_1, X_2, \ldots)$ be a Markov chain and consider the sequence over the equivalence classes $([X_1], [X_2], \ldots)$. Under what conditions is this a Markov chain? The following lemma (Levin et al., 2008) answers this question.

Lemma 4.1 (Levin, Peres, Wilmer - Lemma 2.5)**.** *Let $(X_1, X_2, \ldots)$ be a Markov chain with state space Ω and transition matrix P and let $\sim$ be an equivalence relation over Ω with equivalence classes $\Omega^\sharp = \{[x] : x \in \Omega\}$. Assume P satisfies $P(x, [y]) = P(x', [y])$ for all $x \sim x'$, where $P(x, [y]) := \sum_{y' \sim y} P(x, y)$. Then $([X_1], [X_2], \ldots)$ is a Markov chain with state space $\Omega^\sharp$ and transition function $P^\sharp([x], [y]) = P(x, [y])$.*

What is the form of the stationary distribution for $P^\sharp$?

Lemma 4.2. *Let P, $P^\sharp$, Ω, $\Omega^\sharp$, and $\sim$ be as in Lemma 4.1. If P is reversible with respect to π, then $P^\sharp$ is reversible with respect to $\pi^\sharp([x]) = \pi([x]) := \sum_{x' \sim x} \pi(x)$.*

4.2. Induced Clusterings

Let x be a point sequence and consider the equivalence relation $\sim$ over labellings such that $z \sim z'$ if there exists a permutation σ s.t. $\sigma(z) = \sigma(z')$. Let P denote the Gibbs sampler from Algorithm 1. What does the corresponding Markov chain over the equivalence classes, $P^\sharp$, look like?

Given a labelling z, we know from (3) that $z', z'' \in [z]$ have the same probability mass under π. Thus, one way to describe $P^\sharp$ is that if the current state is $[z]$, it chooses any labelling $z' \in [z]$, moves to a neighboring labelling z'' according to P, and sets the new state to be $[z'']$.

While this is a concise way of describing $P^\sharp$, it offers little intuition on what the state space looks like. An alternative view is to consider the following notion of clustering.

Given an index set S, a *t-partition* or *t-clustering* of S, is a set of t nonempty, disjoint subsets whose union is S. Now let $S = \{1, \cdots, n\}$ and define $\Omega_t(x)$ to be the set of all t-partitions of S and $\Omega_{\leq k}(x) = \cup_{t=1}^k \Omega_t(x)$.

Lemma 4.3. *The state space $\Omega_{\leq k}(x)$ is isomorphic to the set of equivalence classes induced by $\sim$ over $\{1, \ldots, k\}^n$, $\Omega^\sharp$. Furthermore, the $P^\flat$ specified in Algorithm 2 is the induced Markov chain of P on $\Omega_{\leq k}(x)$, $P^\sharp$. Finally, $P^\flat$ is reversible with respect to*

$$\pi^\flat(\mathbb{C}) \propto \frac{1}{(k - |\mathbb{C}|)!} \prod_{S \in \mathbb{C}} \frac{\Gamma(|S| + \alpha)}{\Gamma(\alpha)} q(S).$$

The $1/(k - |\mathbb{C}|)!$ term appears because $\mathbb{C}$ has $k!/(k - |\mathbb{C}|)!$ counterparts in the labelling space. The upshot of Lemma 4.3 is that $P^\sharp$ and $P^\flat$ are the same Markov chain.

5. Mixtures of Gaussians

In this paper, we are particularly interested in mixtures of d-dimensional spherical Gaussians with known variance σ^2 and conjugate prior. One convenient conjugate prior of such a distribution is itself a d-dimensional spherical Gaussian. We will consider the following generative process.

$$\begin{aligned} (w_1, \ldots, w_k) &\sim \text{Dirichlet}(\alpha, \ldots, \alpha) \\ \mu_1, \ldots, \mu_k &\sim \mathcal{N}(\mu_0, \sigma_0^2 I_d) \\ z_i &\sim \text{Categorical}(w_1, \ldots, w_k) \\ x_i &\sim \mathcal{N}(\mu_{z_i}, \sigma^2 I_d) \end{aligned} \tag{2}$$

The following lemma seen, for example, in (Murphy, 2012) establishes the conjugacy of the prior and posterior in (2) and gives an explicit form for the posterior.

Lemma 5.1 ((Murphy, 2012))**.** *Suppose $\mathcal{P}(\theta)$ is a family of spherical Gaussians with fixed variance σ^2 and mean θ, and our prior on θ is another spherical Gaussian with mean μ_0 and variance σ_0^2. If we observe data $y = (y_1, \ldots, y_n)$ and let $S = \{1, \ldots, n\}$, then our posterior is also a spherical Gaussian with mean μ_S and variance σ_S^2 where*

[1] This is ambiguous if i is already its own cluster. In this case, the probability we keep i as its own set is proportional to $(k - |\mathbb{C}| + 1) \cdot \alpha \cdot q(\{i\})$.

$$\mu_S = \mu_0 \cdot \frac{\sigma^2}{\sigma^2 + \sigma_0^2|S|} + \mu(S) \cdot \frac{\sigma_0^2|S|}{\sigma^2 + \sigma_0^2|S|}$$
$$\sigma_S^2 = \sigma_0^2 \cdot \frac{\sigma^2}{\sigma^2 + \sigma_0^2|S|}$$

where $\mu(S) = \frac{1}{|S|}\sum_{i\in S} y_i$ *is the mean of* y*. Note that* σ_S *only depends on the cardinality of* S*. Further, if* $\sigma_0^2 \geq \sigma^2$*, the second equality immediate implies* $\sigma_S^2 \in \left[\frac{\sigma^2}{|S|+1}, \frac{\sigma^2}{|S|}\right]$.

Recall that for a set of indices S, $q(S)$ is the expected probability of S under $\theta \sim \mathcal{Q}(\beta)$. In the case of Gaussians, we can work out q in closed form.

Lemma 5.2. *Let* $\sigma^2, \mu_0, \sigma_0^2, Q_\beta, P_\theta, x$ *be as given above. Then for any set of indices* $S \subset \{1, \ldots, n\}$*, we have* $q(S) = L(S)R(S)$ *where* $L(S)$ *is the probability assigned to* S *by the max-likelihood model,*

$$L(S) = \left(\frac{1}{2\pi\sigma^2}\right)^{|S|d/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i\in S}\|x_i - \mu(S)\|^2\right),$$

and $R(S)$ *penalizes how far* $\mu(S)$ *is from* μ_0*:*

$$R(S) = \left(\frac{\sigma^2}{\sigma^2 + |S|\sigma_0^2}\right)^{d/2} \exp\left(-\frac{|S|\|\mu_0 - \mu(S)\|^2}{2(\sigma^2 + |S|\sigma_0^2)}\right).$$

The above derivation also gives us a nice expression for $\Delta(\cdot,\cdot)$, which is one of the factors in the transition probabilities from Lemma 3.2.

Lemma 5.3. *Let* x *be as above and let* $S \subset \{1, \ldots, n\}$ *and* $i \in \{1, \ldots, n\} \setminus S$*, then*

$$\Delta(S, i) = \left(\frac{1}{2\pi(\sigma^2 + \sigma_S^2)}\right)^{d/2} \exp\left(-\frac{1}{2} \cdot \frac{\|x_i - \mu_S\|^2}{\sigma^2 + \sigma_S^2}\right).$$

In the Bayesian setting, we typically set σ_0^2 to be large, allowing flexibility in the placement of means. To enforce this, we will require that $\sigma_0 \geq \sigma$. Additionally, μ_0 is typically set to be the origin. This simplifies the form of μ_S:

$$\mu_S = \mu(S) \cdot \frac{\sigma_0^2|S|}{\sigma^2 + \sigma_0^2|S|}.$$

If the size of S is variable, then with an appropriate choice of $|S|$, the leading term of μ_S can be made arbitrarily close to the origin. To simplify things, however, we will only consider the case where μ_0 is the origin.

6. Mixing Rates

We analyze the mixing time of Algorithm 2 for two cases. In the first case, the number of Gaussians is misspecified. Even though we cannot expect the Gibbs sampler to recover the correct Gaussians in this case, it still makes sense to consider the samples generated by the Markov chain and evaluate how quickly these approach the stationary distribution. The lower bound we achieve is exponential in the ratio of the intercluster distances and the variance. It is worth noting that the larger this ratio is, the more well-separated the clusters are.

The second case is the more natural case where the number of Gaussians is correctly specified. We show the mixing time of the Gibbs sampler in this case is lower bounded by the minimum of two quantities, an exponential term much like the first case and a term of the form $n^{\Omega(\alpha)}$ where α is the sparsity parameter of the Dirichlet prior.

6.1. Misspecified Number of Clusters

The sequence of points we consider corresponds to 6 spherical clusters, $T_1, \ldots, T_6$, of n points each with diameter δr whose means are located at the vertices of a triangular prism whose edge lengths are identically r. Figure 1 displays our point configuration X_M when we project to $\mathbb{R}^3$.

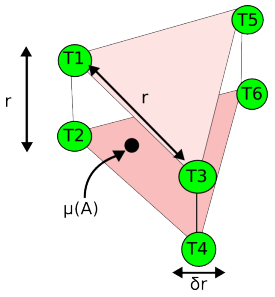


Figure 1. The sequence of points X_M projected to $\mathbb{R}^3$.

We let S_k denote the indices of the points in cluster T_k and let our state space be $\Omega = \Omega_{\leq 3}(X_M)$. Then we have the following result for the Gibbs sampler P over Ω.

Theorem 6.1. *Let* $0 < \delta \leq 1/32$*,* $\alpha > 0$*,* $0 < \sigma \leq \sigma_0$*, and* $k = 3$*. Then there is a constant* $n_0 = \Omega(\max\{\alpha, \sigma^2, d\})$ *s.t. for* $n \geq n_0$ *the mixing rate of the induced Gibbs sampler* P *with parameters* α*,* σ*,* σ_0*, and* k *over* Ω *is bounded below as* $\tau_{mix} \geq \frac{1}{24} \cdot e^{\frac{r^2}{8\sigma^2}}$.

Let $A = S_3 \cup \ldots \cup S_6$. Then we bound the conductance of the singleton set V whose only element is the partition $\mathbb{C} = \{S_1, S_2, A\}$. Because of the symmetric nature of Ω, we have that $\pi(V) \leq 1/2$.

Note two properties of $\mathbb{C}$. First, the number of points in each cluster of $\mathbb{C}$ is within a constant fraction of any other cluster of $\mathbb{C}$. Second, all the points in a cluster of $\mathbb{C}$ are closer to that cluster's mean than to any other cluster's mean by a constant fraction.

To bound the conductance of V, we will bound the probability that we transition out of V. This can happen in one of three ways: we can move an index in A to one of S_1 or S_2, we can move an index in S_1 or S_2 to A, or we can move an index between S_1 and S_2.

Recalling the transition probabilities from Algorithm 2 and

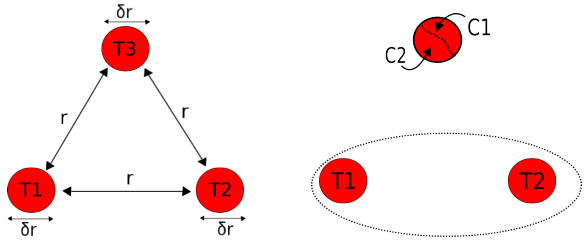
the form of $\Delta(\cdot,\cdot)$ from Lemma 5.3, we can see the likelihood of moving a point i in a cluster S in $\mathbb{C}$ to another cluster T in $\mathbb{C}$ is roughly of the form

$$\begin{aligned} Pr(\text{move } i \text{ to } T) &= \frac{(\alpha+|T|)\Delta(T,i)}{\sum_{T'\in\mathbb{C}}(\alpha+|T'\setminus\{i\}|)\Delta(T',i)} \\ &\leq \frac{(\alpha+|T|)\Delta(T,i)}{(\alpha+|S\setminus\{i\}|)\Delta(S,i)} \\ &\approx \left(\frac{\alpha+|T|}{\alpha+|S\setminus\{i\}|}\right)\left(\frac{\sigma^2+\sigma^2_{S\setminus\{i\}}}{\sigma^2+\sigma^2_T}\right)^{d/2} \\ &\exp\left(\frac{\|x_i-\mu(S)\|^2}{\sigma^2}-\frac{\|x_i-\mu(T)\|^2}{\sigma^2}\right). \end{aligned}$$

Note that since the sizes of S and T are within a constant fraction of each other, we have by Lemma 5.1 that the first two terms in the last line approach constants as the number of points grows. Since all the points are closer to their own cluster's mean than to any other cluster's mean by a constant fraction, the last term in the above is exponential in $-r^2/\sigma^2$. Theorem 6.1 follows by applying Theorem 2.1. The details of this proof are left to the Appendix.

6.2. Correctly Specified Number of Clusters

The sequence of points we consider corresponds to 3 spherical clusters, T_1, T_2, and T_3, of n points each with diameter δr whose means are located at the vertices of an equilateral triangle of edge length r and centered about the origin. Figure 2(a) displays our point configuration X_G in $\mathbb{R}^2$.



(a) X_G projected to $\mathbb{R}^2$. (b) Typical clustering in V.

Figure 2. The setup for Section 6.2.

Letting $\Omega = \Omega_{\leq 3}(X_G)$ be our state space, we have the following result about the mixing time of P over Ω.

Theorem 6.2. *For $\delta < \frac{1}{4}\left(\sqrt{\frac{7}{3}}-\frac{3}{2}\right)$, $\alpha \geq 1$, $0<\sigma\leq\sigma_0$, and $k=3$, there exists $n_0 = \Omega(\max\{\alpha,\sigma^2,d\})$ s.t. $n \geq n_0$ implies that the mixing rate of the induced Gibbs sampler P with parameters α, σ, σ_0, and k over Ω is bounded below as*

$$\tau_{mix} \geq \frac{1}{8}\min\left(\frac{1}{6}e^{\left(\frac{r^2}{96\sigma^2}\right)}, \frac{n^{\alpha-d/2}\left(\frac{\sigma}{\sigma_0}\right)^d \exp\left(\frac{\alpha-\alpha^2}{n}\right)}{2^{3(\alpha-1/2)}\Gamma(\alpha)\exp\left(\frac{r^2}{\sigma_0^2}\right)}\right).$$

To establish this result, we consider the partitions $V \subset \Omega$ such that S_1 and S_2 are clustered together and their cluster contains no indices from S_3. A typical element of V is shown in Figure 2(b). Because of the symmetric nature of Ω, we know $\pi(V) \leq 1/2$. Thus we can use the conductance of V to bound the mixing time.

Ideally, we would like to proceed in the same manner as Section 6.1. However, there is a special case to consider. V contains a special clustering where the number of clusters is 2: $\mathbb{C} := \{S_1 \cup S_2, S_3\}$. The probability of transitioning from $\mathbb{C}$ to a clustering in V^c cannot be bounded from above in the same manner as before since we can choose a point in $S_1 \cup S_2$ and make it a singleton cluster with relatively high probability. Thus, to analyze $\Phi(V)$, we will consider V as the disjoint union of two sets $A = \{\mathbb{C}\}$ and $B = V \setminus A$. Then by the definition of conductance,

$$\Phi(V) \leq \frac{\pi(A)}{\pi(V)} + \frac{1}{\pi(V)}\sum_{x\in B, y\in V^c}\pi(x)P(x,y). \quad (4)$$

Thus, it will be sufficient to consider bounding the two right-hand side terms separately. Our approach to the first term, described in Section 6.2.1, will be to bound the relative probability mass of A under π against the entire set V. Our approach to the second term, described in Section 6.2.2, is similar to our approach in Section 6.1: we bound the probability of transitioning from B to V^c.

6.2.1. The Two Cluster Case

Our goal is to show A has small probability mass in comparison with the rest of V, giving us the following.

Lemma 6.3. *For $n \geq 2$ and $\alpha \geq 1$,*

$$\frac{\pi(A)}{\pi(V)} \leq \frac{2^{3(\alpha-1/2)}\Gamma(\alpha)\exp\left(\frac{\alpha^2-\alpha}{n}+\frac{r^2}{\sigma_0^2}\right)\sigma_0^d}{\sigma^d n^{\alpha-d/2}}.$$

Unfortunately, it is possible to partition S_3 into clusters C_1 and C_2 such that the quantity $q(C_1)q(C_2)$ is smaller than $q(S_3)$. How much smaller this quantity can be is controlled by the following lemma.

Lemma 6.4. *Let $\{C_1, C_2\}$ be a 2-partition of S_3 and suppose $n \geq 2$ and $\alpha \geq 1$, then*

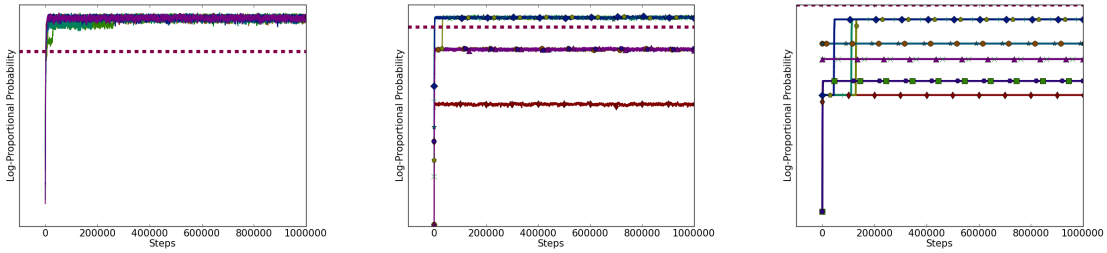
$$q(C_1)q(C_2) \geq \left(\frac{\sigma^2}{n\sigma_0^2}\right)^{d/2}\exp\left(-\frac{r^2}{\sigma_0^2}\right)q(S_3).$$

The proofs of these lemmas are left to the Appendix.

6.2.2. The Three Cluster Case

Bounding the probability that we move from B to V^c is done in the exact same way as the proof of Theorem 6.1. We prove the following lemma in the Appendix.

Lemma 6.5. *For $\delta \leq \frac{1}{4}\left(\sqrt{\frac{7}{3}}-\frac{3}{2}\right)$, there exists an $n_0 = \Omega(\max\{\alpha,\sigma^2,d\})$ s.t. for $n \geq n_0$,*



(a) $d = 10, \alpha = 0.5, \sigma_0 = 0.5$ (b) $d = 3, \alpha = 1.5, \sigma_0 = 10.0$ (c) $d = 3, \alpha = 1.5, \sigma_0 = 1000.0$

Figure 3. In all the above graphs, the dashed line represents the log-proportional probability of the generating clustering.

$$\frac{1}{\pi(V)} \sum_{\mathbb{C} \in B, \mathbb{C}' \in V^c} \pi(\mathbb{C}) P(\mathbb{C}, \mathbb{C}') \leq 6 \exp\left(-\frac{r^2}{96\sigma^2}\right).$$

To complete the proof of Theorem 6.2, we simply use (4) with Lemmas 6.3 and 6.5.

7. Experimental Results

For each experiment, we generated the point sequence by taking $k = 10$ draws from a d-dimensional spherical Gaussian $\mathcal{N}(0, \sigma_0^2 I_d)$ to get means $\mu_1, \ldots, \mu_{10}$. For each mean μ_i, we took $n = 50$ draws from $\mathcal{N}(\mu_i, \sigma^2 I_d)$ with $\sigma = 0.5$.

Recalling Algorithm 2, the Gibbs sampler requires parameters $k, \alpha, \sigma^2, \sigma_0^2$ and an initial clustering. For each set of experiments, we used the same k, σ^2, and σ_0^2 that generated the point sequence over which the sampler was run. We then fixed an α and performed 10 separate runs with different initial clusterings of the points. To generate our initial configurations, we randomly chose k centers and clustered the points together that were closest to a particular center.

Each run of the Gibbs sampler was done for 1,000,000 steps, and we plotted at each step the log of the relative probability of the current state $\mathbb{C}$.

In the experiments of Figure 3, we can see the importance of the ratio σ_0^2/σ^2. Figures 3(b) and 3(c) demonstrate that when all else is held constant, a higher value for σ_0^2/σ^2 will result in slower convergence times. Additionally, Figure 3(a) shows us that when α and σ_0^2/σ^2 are small, the Gibbs sampler will converge to a high probability state.

In the experiments of Figure 4, we can see the importance of α. There are many more phase changes when the value of α is lower. This is possibly due to the observation in Lemma 6.3 that the relative probability mass of an empty clustering is larger when α is smaller. This makes it possible for the Gibbs sampler to create empty clusters more often and thus to make more phase transitions.

Finally, Figure 5 gives us an idea of what these phase transitions look like. The confusion matrices compares the current clustering of the Gibbs sampler to the generating clustering.

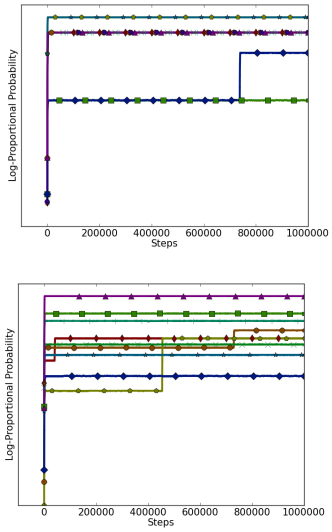


Figure 4. Above, $d = 10$, $\alpha = 1.0$, $\sigma_0 = 5.0$. Below, $d = 10$, $\alpha = 0.5$, $\sigma_0 = 5.0$

Acknowledgements

The authors are grateful to the National Science Foundation for support under grant IIS-1162581 and the Graduate Research Fellowship Program under grant DGE-1144086. We are also appreciative of the feedback given by the anonymous reviewers.

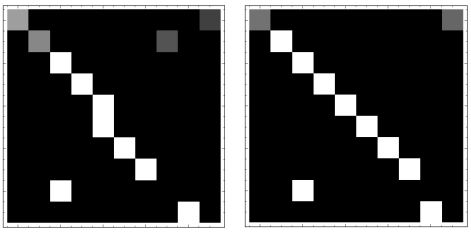


Figure 5. The confusion matrices of one of the runs from Figure 4 before and after a phase transition.

References

Belkin, Mikhail and Sinha, Kaushik. Polynomial learning of distribution families. In *FOCS 2010: Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science*, pp. 103–112, 2010.

Brooks, Stephen P. Markov chain sampling methods for Dirichlet process mixture models. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 47(1):69–100, 1998.

Dempster, A.P., Laird, N. M., and Rubin, D. B. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of Royal Statist. Soc. Ser. B*, 39:1–38, 1977.

Diaconis, Persi. Some things we've learned (about Markov chain Monte Carlo). *Bernoulli*, 19(4):1294–1305, 2013.

Diebolt, Jean and Robert, Christian P. Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(2):363–375, 1994.

Galvin, David and Randall, Dana. Torpid mixing of local Markov chains on 3-colorings of the discrete torus. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 376–384, 2007.

Gelman, Andrew, Robert, Christian, Chopin, Nicolas, and Rousseau, Judith. Bayesian data analysis, 1995.

Geman, Stuart and Geman, Donald. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 721–741, 1984.

Hsu, Daniel and Kakade, Sham M. Learning mixtures of spherical Gaussians: Moment methods and spectral decompositions. In *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science*, ITCS '13, pp. 11–20, 2013.

Jasra, A., Holmes, C. C., and Stephens, D. A. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statist. Sci.*, 20(1):50–67, 2005.

Jerrum, Mark. A very simple algorithm for estimating the number of k-colorings of a low-degree graph. *Random Struct. Alg.*, 7(2):157–165, 1995.

Levin, David A., Peres, Yuval, and Wilmer, Elizabeth L. *Markov Chains and Mixing Times*. American Mathematical Society, 2008.

Luby, Michael and Vigoda, Eric. Fast convergence of the Glauber dynamics for sampling independent sets. *Random Struct. Alg.*, 15:229–241, 1999.

Moitra, Ankur and Valiant, Gregory. Settling the polynomial learnability of mixtures of gaussians. In *FOCS 2010: Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science*, pp. 93–102, 2010.

Murphy, Kevin P. *Machine learning: a probabilistic perspective*. Cambridge, MA, 2012.

Neal, Radford M. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

Randall, Dana. Slow mixing of Glauber dynamics via topological obstructions. In *Proceedings of the 17th Symposium on Discrete Algorithms (SODA)*, pp. 870–879, 2006.

Wu, C. F. Jeff. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.