Lower Bounds for the Gibbs Sampler over the Mixtures of Gaussians

Short Project Writeup : Mehrdad Farajtabar, Tejas Shinde

Main Idea:

Gibbs sampler over Gaussian mixture models with Dirichlet priors is a Markov chain. The authors present the lower bound for the *mixing time* of this Markov chain under two different settings.

Brief Overview:

1. **Mixture models of Gaussians:** Basically mixture of (spherical) gaussians can be described by the following distribution,

$$P(x) = \sum_{i=1}^{k} w_i N(x|\mu_i, \sigma_i^2) \text{ s.t. } \sum_{i=1}^{k} w_i = 1 .$$

The Bayesian approach to specify such mixture models is to provide a *generative process* by which observable quantities e.g. $(x_1, x_2, ...x_n)$ are created and to obtain a posterior distribution for them. One of the popular application of this approach in machine learning domain is to use such posterior distribution for drawing inference upon unobserved quantities. Unfortunately, the solution to the posterior distribution is not usually closed form, thus we have to *sample* from this distribution. Gibbs sampler is used for such purposes.

2. **Mixing Rates:** Although Gibbs sampler is used for sampling from the posterior distribution, *mixing time* of the chain tells us *how long till the samples start coming from the desired distribution (within an approximate bound).*

To answer this, the authors use the theorem (*Sinclair 1988*) which provides a lower bound for τ_{mix} in terms of the inverse of the conductance of a markov chain. The key idea behind it is,

   a. We bound the conductance of our Markov chain (Gibbs sampler) from above.
   b. We bound the mixing time from below by taking inverse of the conductance.

3. **Lower Bounds:** The basic setting assumes mixture of spherical Gaussian clusters from which we will draw the data points. The lower bounds are calculated in two settings, first when number of clusters are misspecified. And second, when number of clusters are correctly specified. This yields the primary contribution by the authors, described as the two theorems stated in the following section.

Primary Contribution:

Theorem 1: For proper setting of δ, the mixing time of the induced Gibbs sampler with a misspecified number of mixtures is bounded as,

$$\tau_{mix} \geq \frac{1}{24} e^{\frac{r^2}{8\sigma^2}} .$$

Theorem 2: For 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,

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

Details:

I. Mixture Model of Gaussians and Gibbs Sampling:

The mixing rate of Gibbs sampler can vary wildly depending on the application, from nearly linear to exponential. Thus, to get meaningful bounds on the mixing rate, we need to consider the specific application, which in this case is learning mixture models.

The Bayesian approach to specify mixture models can be summarized in three steps,

a. Introduce hidden variables that associate to each data point the mixture distribution it was generated from.
b. Formulate generative model which describes all the parameter.
c. Derive posterior distribution from which we can infer.

Following figure takes closer look at the generative process.

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

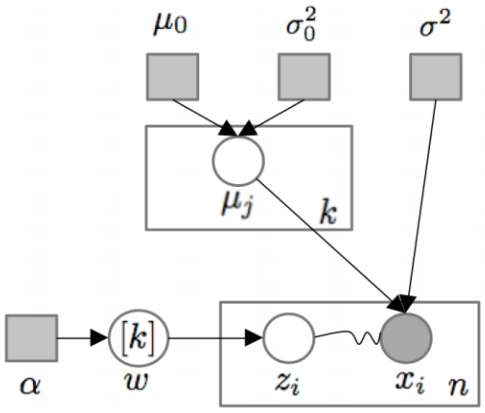

Fig: Generative Process

Walkthrough of the process: We generate a vector of k weights from a symmetric k-dimensional Dirichlet distribution with single parameter $\alpha > 0$ and another vector of k means from spherical Gaussian. Now to generate data points $(x_1, x_2 \ldots x_n)$ we first generate n labels $(z_1..z_n)\varepsilon\{1..k\}^n$ which will describe these data points. And then we generate the data points from the spherical Gaussian distribution using one of the k means which corresponds to one of the labels.

This process brings us to the posterior distribution of data points, which is mostly not in a closed form. Thus, we use Gibb's sampler to sampling from this distribution.

Algorithm 1 : Gibbs Sampler

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

$z^{(0)}, z^{(1)}, \ldots, z^{(t)}, z^{(t+1)}, \ldots$

[Algorithm 1] gives outline of a **collapsed** Gibbs sampler which provides the outcomes $z^{(0)}, z^{(1)}, \ldots, z^{(t)}, z^{(t+1)}$ according to the probability of labeling z given data points x, denoted as $Pr(z|x)$. In many contexts (including the learning of mixture models) we are interested in the distribution given by $Pr(z|x)$. In particular if we are able to relate this to the stationary distribution of the Gibbs sampler, we will have an effective tool for deriving the mixing time. The authors state and prove the following Lemma, which does this very same thing.

Lemma : *Let P denote the collapsed Gibbs sampler, π denote the conditional probability distribution of the labels given by $Pr(z|x)$ and assume that $P(\theta) > 0$ everywhere. Then P is irreducible, aperiodic, and reversible with respect to π. In particular, π is the stationary distribution of P.*

Proof (*sketch*) : $P(\theta)$ denotes the distribution from which we draw the data points. And since it is greater than 0 everywhere, we conclude that each index has a non-negative probability of being chosen. Thus, graph induced by P is strongly connected, implying P is irreducible. Also the graph has self-loops, which makes P aperiodic. To establish reversibility, for any two states σ, τ in state space, we can trivially assume them to differ in exactly one label. This assumption along with the definition of π distribution helps us prove, $P(\sigma, \tau)\, \pi(\sigma) = P(\tau, \sigma)\, \pi(\tau)$. This establishes the stationarity and uniqueness of the distribution.

II. Mixing Rates

Now with the knowledge of the chain (Gibbs sampler) and it's stationary distribution ($P(z|x)$) we can study the mixing properties of the chain.

Definition : *Mixing rate of the Gibbs sampler is the minimum number of steps τ_{mix} to lower the total variation distance between the distribution of $z^{(t)}$ and posterior distribution $Pr(z|x)$ below $\frac{1}{4}$.*

However, the authors make use of the following theorem to bound τ_{mix} which relates mixing time to the conductance of the chain.

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

Definition : *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 is Φ^* which is the minimum conductance of any set S with $\pi(S) \leq \frac{1}{2}$.*

This does provide the preliminaries for bounding the mixing time, however the identifiability makes it difficult to analyze the mixing time of P.

Identifiability: If σ is a permutation over $\{1 \cdots k\}$, then z and $\sigma(z) = (\sigma(z_1), \cdots, \sigma(z_n))$ 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 labelling space may not be true for the space we care about. But how to capture the desired space and what should be the markov chain? The authors answer this question by stating an equivalence relation and the corresponding markov chains.

Equivalence Classes of Markov Chains: If we have a state space of Ω and an equivalence relation ~ on Ω, then consider a sequence over the equivalence classes $([X_1], [X_2], \cdots)$ defined from the states of markov chain $(X_1, X_2, \cdots)$. Then due to [Levin, Peres, Wilmer] lemma we can show when this sequence will be a markov chain.

Lemma[*Levin et.al*. 2008] : *Let* $(X_1, X_2 \cdots)$ *be a markov chain with state space* Ω *and transition matrix* P *and let* $\sim$ *be an equivalence relation over* Ω *with equivalence classes* $\Omega^{\#} = \{[x] : x \in \Omega\}$. *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* $([X_1], [X_2]. \cdots)$ *is a Markov chain with state space* $\Omega^{\#}$ *and transition function* $P^{\#}([x], [y]) = P(x, [y])$.

Lemma : *If* P *reversible with respect to* π, *then* $P^{\#}$ *is reversible with respect to* $\pi^{\#}([x]) = \pi(x) = \sum_{x' \sim x} \pi(x)$.

Thus, 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. Let S be $\{1 \cdots n\}$ and define $\Omega_t(x)$ to be set of all t-partitions of S and $\Omega_{<k}(x) = \cup_{t=1}^{k} \Omega_t(x)$. So we need to establish the Markov chain over this state space. But before that let us state the Projected Gibbs Sampler which will follow this Markov chain.

Algorithm 1 : Projected Gibbs Sampler

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 : *The state space* $\Omega_{\leq k}(x)$ *is isomorphic to the set of equivalence classes induced by* $\sim$ *over* $\{1, \cdots k\}^n$, $\Omega^{\#}$. *Furthermore, the* P^b *specified in Algorithm* 2 *is the induced Markov chain of* P *on* $\Omega_{\leq k}(x)$, $P^{\#}$. *Finally* P^b *is reversible with respect to,* $\pi^b(C) \propto 1/(k - |C|)! \prod_{S \in C} \frac{\Gamma(|S| + \alpha)}{\Gamma(\alpha)} q(S)$

This brings us to conclusion that $P^{\#}$ and P^b are the same Markov chain.

III. Lower Bounds

We can now revisit the main contribution by the author given by the two theorems (Theorem 1 and Theorem 2) in more details. As stated in the theorems, the mixing time is analyzed for two cases, first when the number of Gaussians is misspecified and second when it is correctly specified.

Case 1] Misspecified Number of Clusters

The sequence of points observed correspond to 6 spherical clusters, $T_1, T_2 \cdots T_6$ of n points each with diameter δr whose means are located at the vertices of a triangular prism whose edge lengths are identically r.

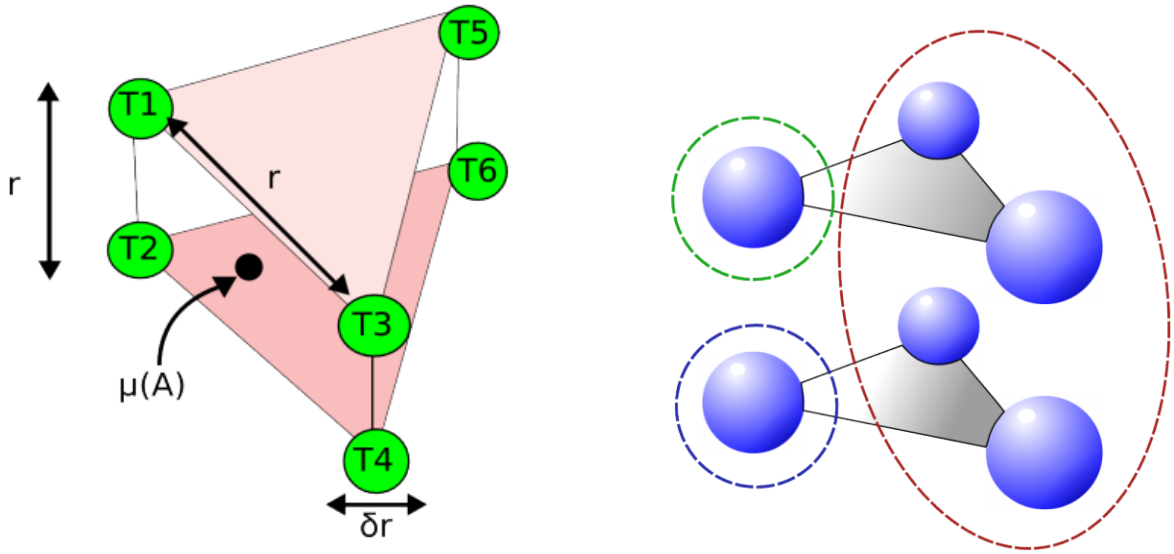

Fig: The sequence of points X_m projected to R^3 (left) and mis-specified number of clusters = 3 (right)

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 can state the following result for the Gibbs sampler P over Ω.

Result: Let $0 < \delta \leq 1/32,\ \alpha > 0,\ 0 < \sigma \leq \sigma_0$ and $k = 3$. Then there is a constant $n_0 = \Omega(max\{\alpha,\ \sigma^2,\ d\})$ such that for $n \geq n_0$ the mixing rate of the induced Gibbs sampler P with parameters $\alpha,\ \sigma,\ \sigma_0$ & k over Ω is bounded below as $\tau_{mix} \geq (1/24)exp(\frac{r^2}{8\sigma^2})$.

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

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: 1. we can move an index in A to one of $S_1 or\ S_2$ 2. we can move an index in $S_1 or\ S_2$ to A 3. we can move an index between S_1 and S_2.

Using the transition probabilities from projected Gibbs Sampler and $\Delta(\cdot,\cdot)$ we can see that the likelihood of moving a point i from cluster S in C to another cluster T in C is roughly of the form,

$$Pr(move\ i\ to\ T) \leq exp(\frac{|x_i - \mu(S)|^2}{\sigma^2} - \frac{|x_i - \mu(T)|^2}{\sigma^2})$$

Note that sizes of S & T are within a constant fraction of each other. The authors prove using a prior-posterior conjugacy lemma (Murphy 2012), that in such case we can have the terms in exponential approaching constants as the number of points grow. And since all the points (i.e. x_i's) are closer to their own cluster mean than others (by assumption), the above term is actually an exponential in $-r^2/\sigma^2$.

Case 2] Correctly specified number of clusters

In this case the observed sequence of points corresponds to 3 spherical clusters T_1, T_2, 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.

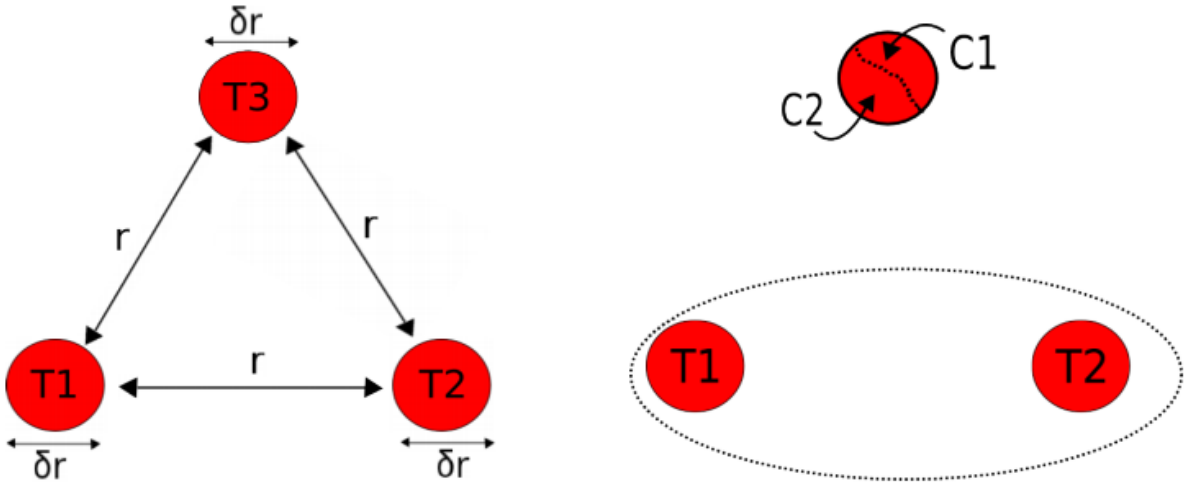

Fig: Point sequence X_G projected to R^2 (left) and typical clustering on them

Let S_k denote the indices of the points in cluster T_k and let $\Omega = \Omega_{\leq 3}(X_G)$ be our state space, we have the following result about the mixing time of P over Ω.

Result: For $\delta < \frac{1}{4}(\sqrt{\frac{7}{3}} - \frac{3}{2}),\ \alpha \geq 1,\ 0 < \sigma \leq \sigma_0,\ \&\ k = 3,$ there exists $n_0 = \Omega(max\{\alpha,\ \sigma^2,\ d\})$ such that, $n \geq n_0$ implies that the mixing rate of the induced Gibbs sampler P with parameters $\alpha,\ \sigma,\ \sigma_0\ \&\ k$ over Ω is bounded below as,

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

Sketch of the proof: Consider partition $V \subset \Omega$ such that $S_1 \& S_2$ are clustered together and their cluster contains no indices from S_3. A typical element of V is shown in the figure above (right). Because of the symmetric nature of Ω we know $\pi(V) \leq 1/2$. Thus we can use conductance of V to bound the mixing time.

Here to analyse $\Phi(V)$ we will consider V as the disjoint union of two sets $A = \{C\}\ \&\ B = V \setminus A$. Then by definition of conductance we have, $\Phi(V) \leq \frac{\pi(A)}{\pi(V)} + \frac{1}{\pi(V)} \sum_{x \in B,\ y \in V^c} \pi(x) P(x, y)$.

Thus, it will be sufficient to consider bounding the two right-hand side terms of the conductance separately. The approach to bound the first term will be to bound the relative probability mass of A under π against the entire set V. And for the second term, we bound the probability of transition from B to V^c, just like we saw in Case 1.

Experimental Result:

The authors present an interesting experimental result which helps us visualize the dynamics of the Markov chain at various time instances.

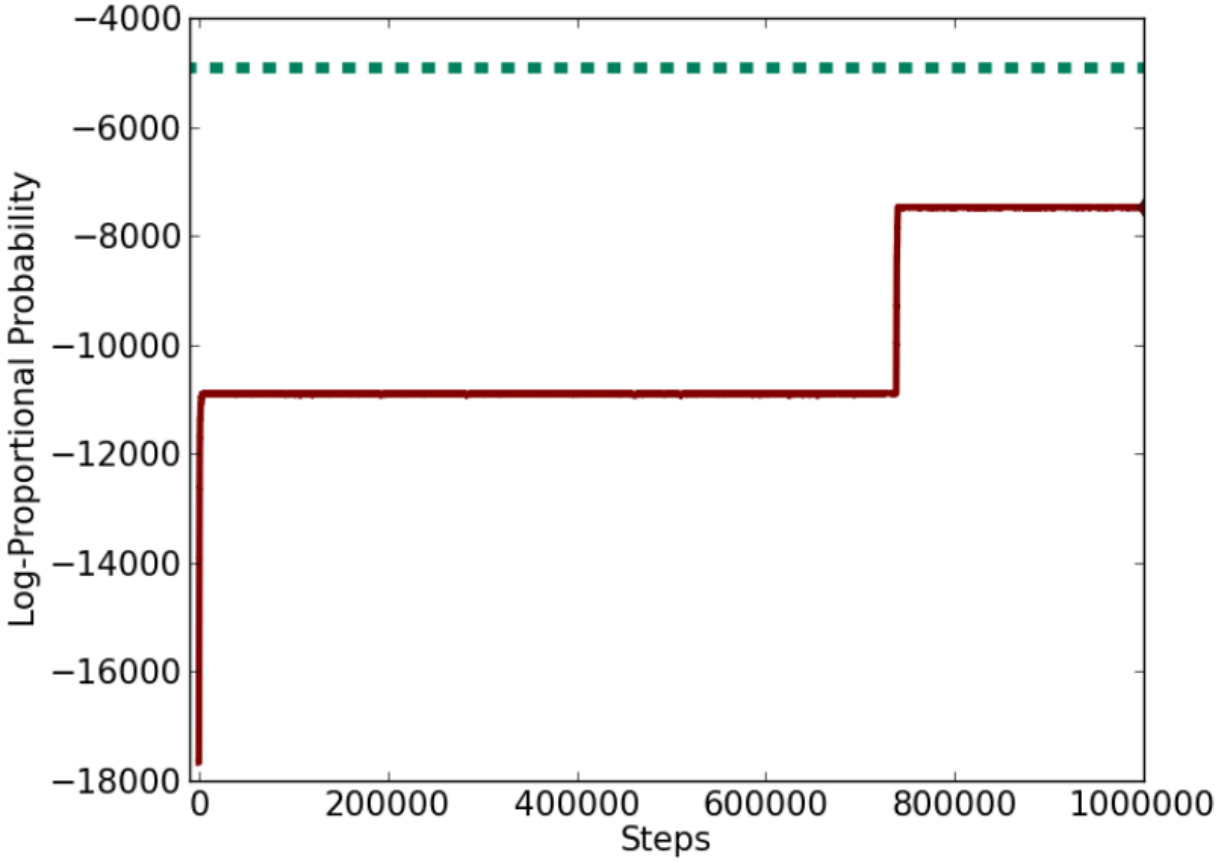

Fig: Simulated chain over clusters with parameters $k = 10,\ n = 500, d = 10, \alpha = 1.5, \sigma = 0.5, \sigma_0 = 5$

The result shows us the ground truth using dotted line at the top for the generated 500 data points. And the probability ($Pr(z|x)$ of the Gibbs sampler) shows the behavior of the Gibbs sampler with the parameters learned from the ground truth. So what we are looking at is essentially how the probability converges to the ground truth.

Bottleneck in conductance: From the graph we find that initially at around $10Kth$ time step, we see a sort of equilibrium in the chain probabilities. However this is due to the bottleneck in the chain state space, which prevents the chain from crossing to the states from the other side of the bottleneck. The bottleneck in the conductance is clearly visible from the jump at around 7.5Million th time step.

Although the chain crosses the bottleneck, it is still far from reaching the ground truth. Thus this result also shows us how the Gibbs sampler chain is **torpid**.