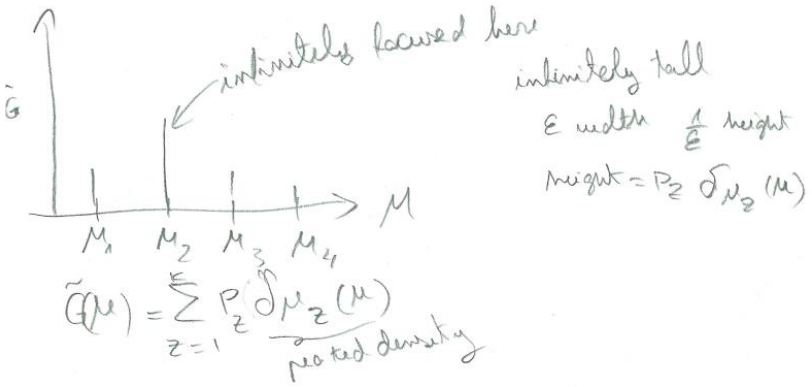
Alin Tomescu, http://people.csail.mit.edu/~alinush

Lecture 20

$$P(x|\theta) = \sum_{z=1}^{k} N(x, \mu_z, \sigma^2)$$

We wrote this in a different way focusing on the means. See Figure 1 below:

$$P(x|\theta) = E_{\mu \in \tilde{G}}\{N(x; \mu, \sigma^2\}$$

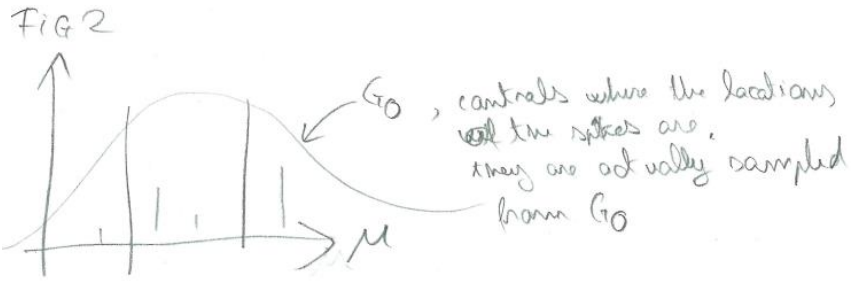


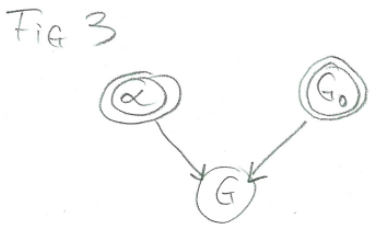
Density is the random variable we are interested in.

Dirichlet processes

Samples from the process they look like the graph in Figure 1, spikey densities with a finite number of spikes.

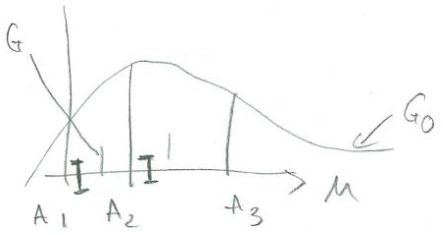
$G \sim DP(\alpha, G_0)$ where G_0 is the prior density (not necessarily Dirichlet)





Why is this even called a Dirichlet process? We'll see later.

FIG 4



let's partition the means into partitions A_1, A_2, A_3 ...

$$A_i \cap A_j = \emptyset, i \neq j$$

$$\cup_{i=1}^{k} A_i = (-\infty, \infty)$$

What is $G(A_i)$?

Why?

$$G(A_i) = \int_{A_i} G(\mu) d\mu$$

How are these distributed in A_i? Exactly like a Dirichlet, for any partition.

$$\begin{bmatrix} G(A_1) \\ \vdots \\ G(A_k) \end{bmatrix} \sim \text{Dirichlet}\big(\alpha G_0(A_1), \ldots, \alpha G_0(A_k)\big)$$

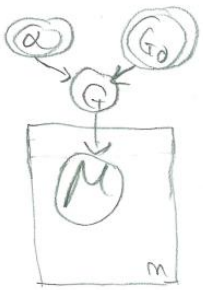
If I now have three partitions A_1, A_2, A_3 and I combine them as $A_1 \cup A_2$ and A_3, it corresponds to:

$$\begin{bmatrix} G(A_1) + G(A_2) \\ G(A_3) \end{bmatrix} \sim \text{Dirichlet}\big(\alpha G_0(A_1) + \alpha G_0(A_2), \alpha G_0(A_3)\big)$$

$$\begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}, \begin{bmatrix} \theta_1 + \theta_2 \\ \theta_3 \end{bmatrix}$$

Let's draw samples from G (the dirichlet process)

FIG 5



$G \sim DP(\alpha, G_0)$

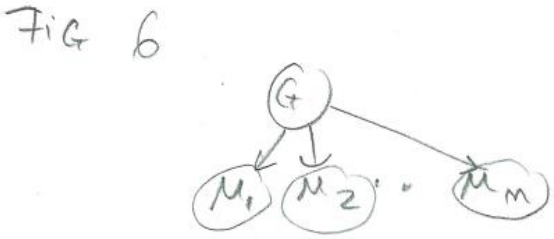
For $i = 1, \ldots, n$

$M_i \sim G$

In the finite mixture model, we will now:

- Select $\tilde{G}$
- Sample n means $\mu_i \sim \tilde{G}$

- Sample the x_i's from $N(x; \mu_i, \sigma^2)$



How do these means co-vary? How does the choice of a mean for a particular sample correlate with another choice for a mean. Many of them are exactly the same, according to the Chinese Restaurant process.

$$P(\mu_i = \mu \mid \mu^{-i}) = \frac{1}{\alpha + n - 1}\sum_{j \neq i} \delta_{\mu_j}(\mu) + \frac{\alpha}{\alpha + n - 1} G_0(\mu)$$

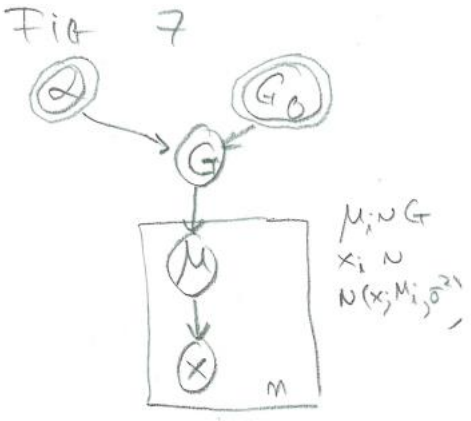
With high probability I'm going to reuse means that I've already sampled. Any new spike that I generate is generated from the G_0.

Expected number of spikes: $\sum_{n=1}^{N} \frac{\alpha}{\alpha+n-1} \sim \log(n)$

I'm going to resist including a lot of mixture components as the number of data points increases.

If α is vey large, if I have n means to sample, I will get exactly n unique spikes.

Only by reusing spikes I'm starting to cluster the points. So it's important that this grows very slowly.



Let's make this a mixture model and add observations

If I know the samples, what can I say about the means? How are the data points clustered? I have to infer how the generative process works given that I see the data (the x values).

How do I modify the Gibbs sampling so I can sample from the posterior distribution of the means given the data?

Now it's important to understand the independence properties in the graph.

If I know $n-1$ means does that evidence impact the 1st mean? No, because they are independent in the graph.

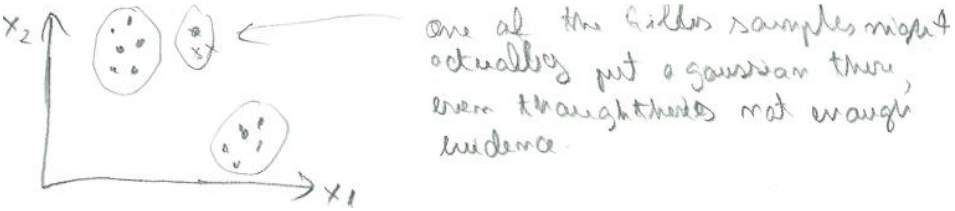
Alin Tomescu, http://people.csail.mit.edu/~alinush

Gibbs sampling procedure

I wish to sample a mean, fix all the values for the rest, and I condition on all the observation:

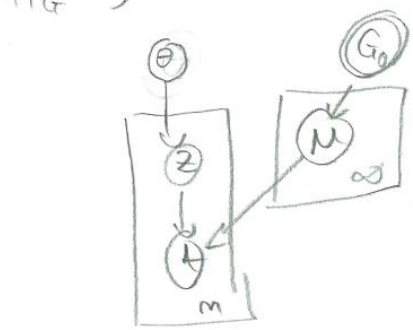
$$P(\mu_i = \mu | \mu^{-i}, x_i) \propto P(\mu_i = \mu | \mu^{-i}) P(x_i | \mu)$$

$$P(x_i | \mu) = N(x_i, \mu, \sigma^2)$$



This Dirichlet process model adapts to the data. You don't have to specify the number of Gaussians for instance, it just comes out of the mean sampling as the number of unique means and is expected to be $\sim \log n$ for n data points.

Fig 9



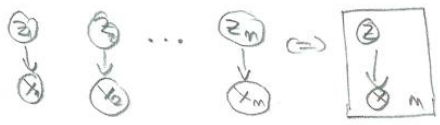
$$P(z_i = l | z^{-i}) = \begin{cases} \dfrac{n^{-i}(l)}{\alpha + n - 1}, if\ n^{-i}(l) > 0 \\ \dfrac{\alpha}{\alpha + n - 1}, if\ l = new \end{cases}$$

The infinite number of means in figure 9 is really at most n since we only need n distinct means at most.

Structured probability models, inference problems, Hidden Markov Models

One property of all these models is that if I change the order I get exactly the same distribution. It's exchangeable. We are not modelling anything that has dependence as a sequence.

Fig 10



Look at models…

I want to look at how the topic choices evolve as I choose words?

Example of a HMM: Think of modelling speeds. Every 10 milliseconds I take that waveform and I take a FTT of that and I get what frequencies are present within that time window. I get frequency spectrum as a function of time (spectrogram). I will need to compute some properties of those frequencies.