- Clustering - How to choose K?
  - Ad hoc approaches: data-driven, but little theory
  - Methods based on maximizing objective fns (M-estimators)

    e.g. clustering via K-means, spectral clustering has frequentist guarantees but hard to be data-driven.
  - Parametric methods such as AIC, BIC requires hidden assumptions

- Model-based approach:

$$P(X|\phi) = \sum_{k=1}^{K} \pi_k \, p(X|\phi_k)$$

Let $G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k}$ mixing measure

<u>Model</u>:

$$\phi \mid G \sim G$$
$$x \mid \phi \sim p(x|\phi)$$

- G Random. Need a prior dist on G.

want K unbounded ($K = \infty$) $\Rightarrow$ Stick-breaking prior.

This is a dist over G of form:

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \qquad \left\{ \begin{array}{l} \pi \text{ Random} \\ \phi \text{ Random} \\ \sum_{1}^{\infty} \pi_k = 1. \end{array} \right.$$

- <u>Chinese Restaurant Process</u> (CRP)

A Random process in which customers sit down in a Chinese Restaurant with infinite number of tables.

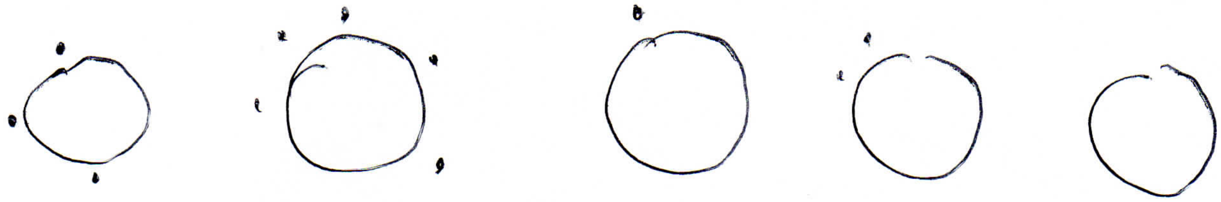
CRP:

- first customer sits at the first table
- m-th subsequent customer sits at a table drawn from the following dist:

$\mathbb{P}$(sits at previously occupied table i) $\propto n_i$

$\mathbb{P}$(sits in a new (unoccupied) table) $\propto \alpha_0$

n_i : # customers sitting at table i (among previous $m-1$ customers)



CRP and Clustering:

CRP induces a dist on disjoint partition of customers into clusters (subsets) and # of clusters (tables).

Let π_K be the proportion of customers sitting at table K.

$\phi_K \overset{iid}{\sim} G_0$ (because π_K's, ϕ_K's are).

Then we obtained a $\boxed{\text{Random}}$ measure: $G = \sum_{k=1}^{\infty} \pi_K \delta_{\phi_K}$

what can we say about G?

Polya urn model CRP is a particular example of a general class of prob. models known as Pólya urn.

~~Let $\theta_1, \ldots, \theta_n \mid G \overset{iid}{\sim} G$. there are duplicates among the θ_i's $\Rightarrow$~~ The Polya urn model is as follows:

$$\theta_1 \sim G_0$$

$$\theta_i \mid \theta_1, \ldots, \theta_{i-1} \sim \alpha_0 G_0 + \sum_{j=1}^{i-1} \delta_{\theta_j}$$

$$P(\theta_1, \theta_2) = P(\theta_1) P(\theta_2 | \theta_1)$$
$$\propto P(\theta_1) \left(\alpha_0 G_0(d\theta_2) + \delta_{\theta_1}(\theta_2) \right)$$
$$\propto \alpha_0 G_0(d\theta_1) G_0(d\theta_2) + G_0(d\theta_1) \mathbb{1}(\theta_1 = \theta_2).$$

$$\Rightarrow P(\theta_1, \theta_2) = P(\theta_2, \theta_1).$$

$$\mathbb{P}(\theta_1, \theta_2, \theta_3) = \mathbb{P}(\theta_1, \theta_2) \mathbb{P}(\theta_3 | \theta_1, \theta_2)$$
$$= \mathbb{P}(\theta_2, \theta_1) \mathbb{P}(\theta_3 | \theta_2, \theta_1) = \mathbb{P}(\theta_2, \theta_1, \theta_3)$$

in general $\theta_1, \theta_2, \dots, \theta_n, \dots$ are infinitely exchangeable!

By De Finetti's thm $\exists$ a random mixing dist G s.t

$$\theta_1, \dots \theta_n | G \overset{iid}{\sim} G.$$

So that $$P(\theta_1 \dots \theta_n) = \int \prod_{i=1}^{n} P(\theta_i | G) \, d\Pi(G)$$

What is G?

- <u>Stick-breaking process</u>: $\boxed{\{\pi_k\}_{k=1}^{\infty}}$

Let $\beta_k \overset{iid}{\sim} \text{Beta}(1, \alpha_0)$ $\quad k = 1, 2, \dots$

Define
$$\pi_1 = \beta_1$$
$$\pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_l) \qquad k = 2, 3, \dots$$

Easy to check that $\sum_{k=1}^{\infty} \pi_k = 1$. (Breaking a stick to small pieces using beta proportions)

$$1 - \sum_{k=1}^{K} \pi_k = 1 - \beta_1 - \beta_2(1-\beta_1) - \beta_3(1-\beta_1)(1-\beta_2)$$
$$= (1-\beta_1)(1-\beta_2) - \beta_3(1-\beta_1)(1-\beta_2)$$
$$= \dots$$
$$= (1-\beta_1) \dots (1-\beta_K) \to 0 \text{ as } K \to \infty$$

<u>Thm</u>: Let $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$ where $\phi_k \overset{iid}{\sim} G_0$ then this is the dist G that defines Polya's urn

in fact the random measure G is called a Dirichlet Process
where does Dirichlet come in? Hint: Beta. dist was used

Recall Dirichlet Dist (finite) & Conjugacy (Dir - Multinomial).

$$P(\pi|\alpha) = \frac{\Gamma(\sum \alpha_i)}{\prod \Gamma(\alpha_i)} \pi_1^{\alpha_1 - 1} \cdots \pi_K^{\alpha_K - 1} \qquad (K < \infty)$$

$$\pi|\alpha \sim \mathrm{Dir}(\alpha)$$

Let $\theta_1, \dots, \theta_n | \pi \overset{iid}{\sim} \mathrm{Mult}(\pi) \qquad Y_i \in \{1, \dots, k\}$

Then $$P(\theta_1 \dots \theta_n) = \int \prod_{i=1}^{n} P(\theta_i|\pi)\, dP(\pi|\alpha)$$

$$\propto \int \prod_{k=1}^{K} \pi_k^{n_k} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} d\pi \propto \int \prod_{k=1}^{K} \pi_k^{n_k + \alpha_k - 1} d\pi$$

$$= \frac{\Gamma(\alpha_1 + \dots + \alpha_K)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_K)} \frac{\Gamma(\alpha_1 + n_1) \dots \Gamma(\alpha_K + n_K)}{\Gamma(\alpha_1 + \dots + \alpha_K + n)}$$

$$P(\theta_n | \theta_1 \dots \theta_{n-1}) = \frac{\Gamma(\alpha_1 + n_1) \dots \Gamma(\alpha_K + n_K)}{\Gamma(\alpha_1 + \dots + \alpha_K + n)} \frac{\Gamma(\alpha_1 + \dots + \alpha_K + n - 1)}{\Gamma(\alpha_1 + \tilde{n}_1) \dots \Gamma(\alpha_K + \tilde{n}_K)}$$

$$= \frac{\alpha_k + n_k}{\sum \alpha_k + n} \quad \text{if } \theta_n = k.$$

$$\theta_n | \theta_1 \dots \theta_{n-1} \propto \alpha_k + n_k \quad \text{if } \theta_n = k$$

Suppose that we can somehow let $K \to \infty$, yet $\alpha_k = \frac{\alpha}{K}$
so that $\sum \alpha_k = \alpha$ fixed, then

$$\theta_n | \theta_1 \dots \theta_{n-1} \propto n_k \quad \text{if } \theta_n = k.$$

there is a remaining probability that $\theta_n \neq \theta_1 \dots \theta_{n-1}$

that probability $\longrightarrow \frac{\alpha}{\alpha + n}$.

How is this related to ~~DP~~ stick-breaking? ?

- ~~As $k \to \infty$, $\pi \longrightarrow$ Dirichlet process~~
- Reorder $(\pi_1, \dots, \pi_K)$ in decreasing order then $(\pi_1, \dots, \pi_K) \longrightarrow$ stick breaking process

- Now we're ready to state original def of Dirichler process.
- Stochastic process vs. RV's.

Elementary prob. thy : RV's are real-valued function $X: \Omega \to \mathbb{R}$

Ω : set of events

$$\mathbb{P}(X \le t) = \mathbb{P}(\{X(\omega) < t\}).$$

A stochastic process can be viewed as a dist/measure on space of more complex objects such as functions or measures.

therefore, Random function or random measures.

- Gaussian process as ~~deg~~ dist on $\{X: T \to \mathbb{R}\}$ is defined as a collection of RV's $\{X(t) : \Omega \to \mathbb{R}\}_{t \in T}$.

s.t. every finite collection of $\{X(t)\}_{t \in S}$ is a multivariate Gaussian and that they are consistent with each other thru marginalization

- Kolmogorov's thm says that then there exists a meaningful dist on the space of functions $\{X : T \to \mathbb{R}\}$ which is called a Gaussian process..

For Gaussians, consistency is obviously satisfied : $(X_1, X_2) \sim N$ then X_1 is also N.

- A probability measure (dist) G is a function that map ~~set~~ a probability space Ω to $[0,1]$

$A \in \Omega$ $\quad G(A) =$ probability $\omega \in A$ under G.

i.e. if $X \sim G$ then $\mathbb{P}(X \in A) = G(A)$.

- To define a Random measure ~~G~~ G means the collecting $\{G(A)\}_{A \subset \Omega}$ is ~~a~~ collecting RV's.

Def: Let $(\Omega, \mathcal{F})$ a prob. space.

if G is a Random measure on $(\Omega, \mathcal{F})$ s.t

for any partition $(A_1, A_2, \dots, A_K)$ of Ω, $K \in \mathbb{N}$

$$(G(A_1), G(A_2), \dots, G(A_K)) \sim \mathrm{Dir}(\alpha G_0(A_1), \dots, \alpha G_0(A_K)).$$

then G is said to be dist according to a Dirichlet proc

G_0 is called a base measure (mean measure)

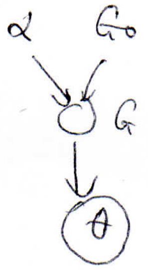
α concentration parameter.

Note Such DP exists due to Consistency property of finite dim Dir dist: if $(Y_1, \dots, Y_K) \sim \mathrm{Dir}(\alpha_1, \dots, \alpha_K)$, then $(Y_1 + \dots + Y_{r_1}, \dots, Y_{r_\ell} + \dots + Y_K) \sim \mathrm{Dir}(\alpha_1 + \dots + \alpha_{r_1}, \dots, \alpha_{r_\ell} + \dots + \alpha_K)$

Fact: if $G \sim DP(\alpha, G_0)$

then $G(A) \sim \mathrm{Beta}(\alpha G_0(A), \alpha(1 - G_0(A)))$

$$\Rightarrow \begin{cases} \mathbb{E}\, G(A) = G_0(A) \\ \mathrm{var}\, G(A) = \dfrac{\cancel{\alpha} G_0(A)(\alpha G_0(A) + 1)}{\cancel{\alpha}(\alpha + 1)} - G_0(A)^2 = \dfrac{G_0(A)(1 - G_0(A))}{\alpha + 1} \end{cases}$$



Question: if $\begin{cases} G \sim DP(\alpha, G_0) \\ \theta \mid G \sim G \end{cases}$

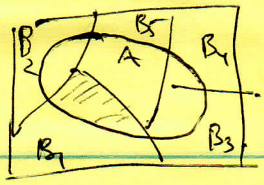
what is the posterior of $[G \mid \theta]$?

$\forall A \in \mathcal{F}$ $\quad P(X \in A \mid G(A)) = G(A)$ a.s.

$$\Rightarrow P(X \in A) = \mathbb{E}\, G(A) = G_0(A)$$

Take a measurable partition $(G(B_1) \dots G(B_K)) \sim \mathrm{Dir}(\alpha G_0(B_1) \dots \alpha G_0(B_K))$

Thm $\quad [G(B_1) \dots G(B_K) \mid X] \sim \mathrm{Dir}(\alpha G_0(B_1) + \delta_{B_1}(X) \dots\dots)$



$$P(X \in A;\ G(B_1) \leq y_1, \dots, G(B_k) \leq y_k)$$

$B_1' = A \cap B_1$

$$= \sum_{k=1}^{K} P(X \in B_i';\ G(B_1) \leq y_1 \dots, G(B_k) \leq y_k)$$

$$= \sum_{k=1}^{K} P(X \in B_i')\, P(G(B_1) \leq y_1, \dots, G(B_k) \leq y_k \mid X \in B_i')$$

$$= \sum_{k=1}^{K} G_0(B_i')\, P(G(B_i') + G(B_1'') \leq y_1 \dots \mid X \in B_i)$$

note that $\left(G(B_1'), G(B_1''), \dots, G(B_k'), G(B_k'')\right) \sim \mathrm{Dir}(\alpha G_0 \dots))$

By ~~mult~~ Dir-Mult Conjugacy:

$$\left[(G(B_1'), G(B_1''), \dots) \mid X \in B_i' \right]$$

$$\sim \mathrm{Dir}(\alpha G_0(B_i') + \delta_{B_i'} \dots)$$

$$= \sum_{k=1}^{K} G_0(B_i')\, P(G(B_1) \leq y_1, \dots, G(B_k) \leq y_k) \mid \alpha G_0(B_1) + \delta_{B_1}, \dots, \alpha G(B_k) + \delta_{B_k}$$

So $P(G(B_1) \leq y_1 \dots, G(B_k) \leq y_k \mid X \in A)$

$$= \sum_{k=1}^{K} \frac{G_0(\underbrace{B_i'}_{\leftarrow A \cap B_i})}{G_0(A)}\, P(G(B_1) \leq y_1 \dots, G(B_k) \leq y_k) \mid \alpha G_0(B_1) + \delta_{B_1} \dots \alpha G_0(B_k) + \delta_{B_k})$$

if $A \subset B_i$ then $P(G(B_1) \leq y_1 \dots G(B_k) \leq y_k \mid X \in A)$

$$= P(G(B_1) \leq y_1 \dots G(B_k) \leq y_k \mid \alpha G_0(B_1) + \delta_{B_0} \dots)$$

Since $(G(B_1) \dots, G(B_k)) \mid_{X \in B} \sim \mathrm{Dir}(\alpha G_0(B_1) + \delta_{B_5} \dots)$

Sethuraman

Let $\beta_k \sim \text{Beta}(1, \alpha)$ $\qquad \theta_k \overset{iid}{\sim} G_0$

$\pi_1 = \beta_1$

$\pi_k = \beta_k (1-\beta_1) \dots (1-\beta_{k-1}) = \beta_k (1 - \cancel{\pi_1} - \dots - \cancel{\pi_{k-1}})$.

then let $G = \sum \pi_k \delta_{\theta_k}$

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

$$G = \beta_1 \delta_{\theta_1} + (1-\beta_1)\Big[\underbrace{\beta_2 \delta_{\theta_2} + (1-\beta_2)\beta_3 \delta_{\theta_3} + \dots}\Big]$$

$$= \beta_1 \delta_{\theta_1} + (1-\beta_1)\, G'$$

$G' \perp\!\!\!\perp \beta_1, \theta_1$.

So $G \overset{dist}{=} \beta_1 \delta_{\theta_1} + (1-\beta_1)\, G$.

Let $(B_1, \dots, B_k)$ be a partition

$P = (G(B_1) \dots G(B_k))$

then

$$P = \beta_1 (\delta_{\theta_1}(B_1), \dots, \delta_{\theta_1}(B_k)) + (1-\beta_1)\, P \qquad (*)$$

Can easily check that

$$\text{Dir}(\alpha G_0(B_1), \dots, \alpha G_0(B_k)) \overset{dist}{=} \beta_1 ((\delta_{\theta_1}(B_1), \dots \delta_{\theta_1}(B_k)) + (1-\beta_1)\, \text{Dir}(\alpha G_0(B_1), \dots, \alpha G_0(B_k))$$

if $\theta_1 \in B_1$ then $\beta_1 (1, 0, 0 \dots 0) + (1-\beta_1)\, \text{Dir}(\alpha G_0(B_1) \dots)$

$\overset{dist}{=} \text{Dir}(\alpha G_0 + \delta_{e_1})$.

$$\sum_{j=1}^{k} p(\theta_1 \in B_j)\, \text{Dir}(\alpha G_0 + \delta_{e_j}) = \text{Dir}(\alpha G_0).$$

next Lemma V, U, W RV's, $W \in [-1,1]$

U, V taking value in same vector space

and $\begin{cases} V \overset{dist}{=} U + WV \\ P(|W|=1) \neq 1 \end{cases}$ then V is unique.

if $V \overset{dist}{\neq} V'$ $\quad V_{n+1} := U_n + W_n V_n$

$V'_{n+1} = U_n + W_n V'_n$

$\to |V_{n+1} - V'_{n+1}| = |W_n||V_n - V'_n| \longrightarrow 0$ wp 1.

This thm proves automatically that DP distributed Random measures are discrete wp 1.

Alternative proof of discreteness

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

then $G|X \sim DP(\alpha G_0 + \delta_X)$.

$G | X_1, \dots X_n \sim DP(\alpha G_0 + \sum_{i=1}^n \delta_{X_i})$

$DP\left(\alpha + n, \frac{\alpha}{\alpha+n} G_0 + \sum_{i=1}^n \frac{1}{\alpha+n} \delta_{X_i}\right).$

as $n \to \infty$: $\begin{cases} \alpha + n \text{ concentration} \to \infty \\ \text{base measure} \xrightarrow{w} \frac{1}{n} \sum_{i=1}^n \delta_{X_i} \end{cases}$