DELHI TECHNOLOGICAL UNIVERSITY

STOCHASTIC PROCESSES PROJECT REPORT

CHINESE RESTAURANT PROCESS

SUBMITTED BY:

KUMAR APURVA (DTU/2K18/MC/058)

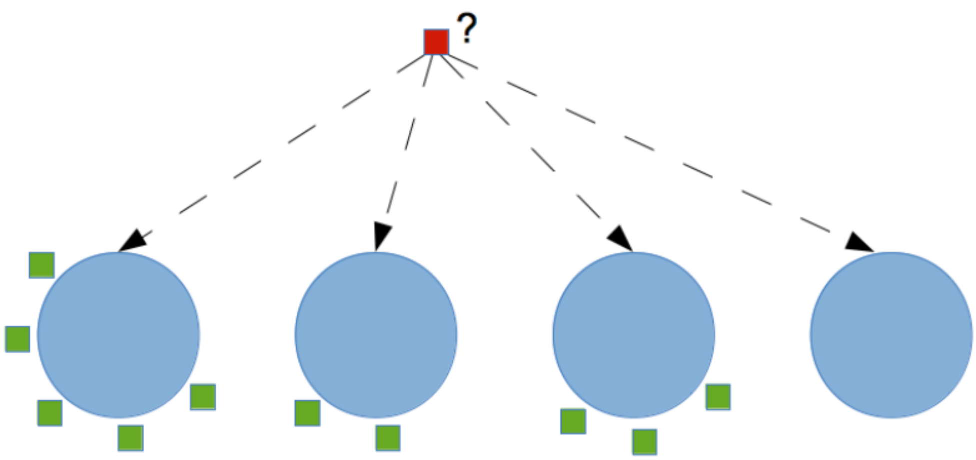
MADHURESH MAYANK (DTU/2K18/MC/062)

INDEX

INTRODUCTION

The Chinese restaurant process is a discrete-time stochastic process, analogous to seating customers at tables in a Restaurant.

Consider the following scenario:
Imagine a Restaurant with an infinite number of circular tables, each with infinite capacity. Customer 1 sits at the first table. The next customer either sits at the same table as customer 1, or the next table. Suppose after sometime, there are five customers at the first table, two customers at second table and three customers at third table. Now if a new customer walks in, what is the probability that he will sit either on any occupied table or the next unoccupied table?



Imagine the scenario as shown in Figure

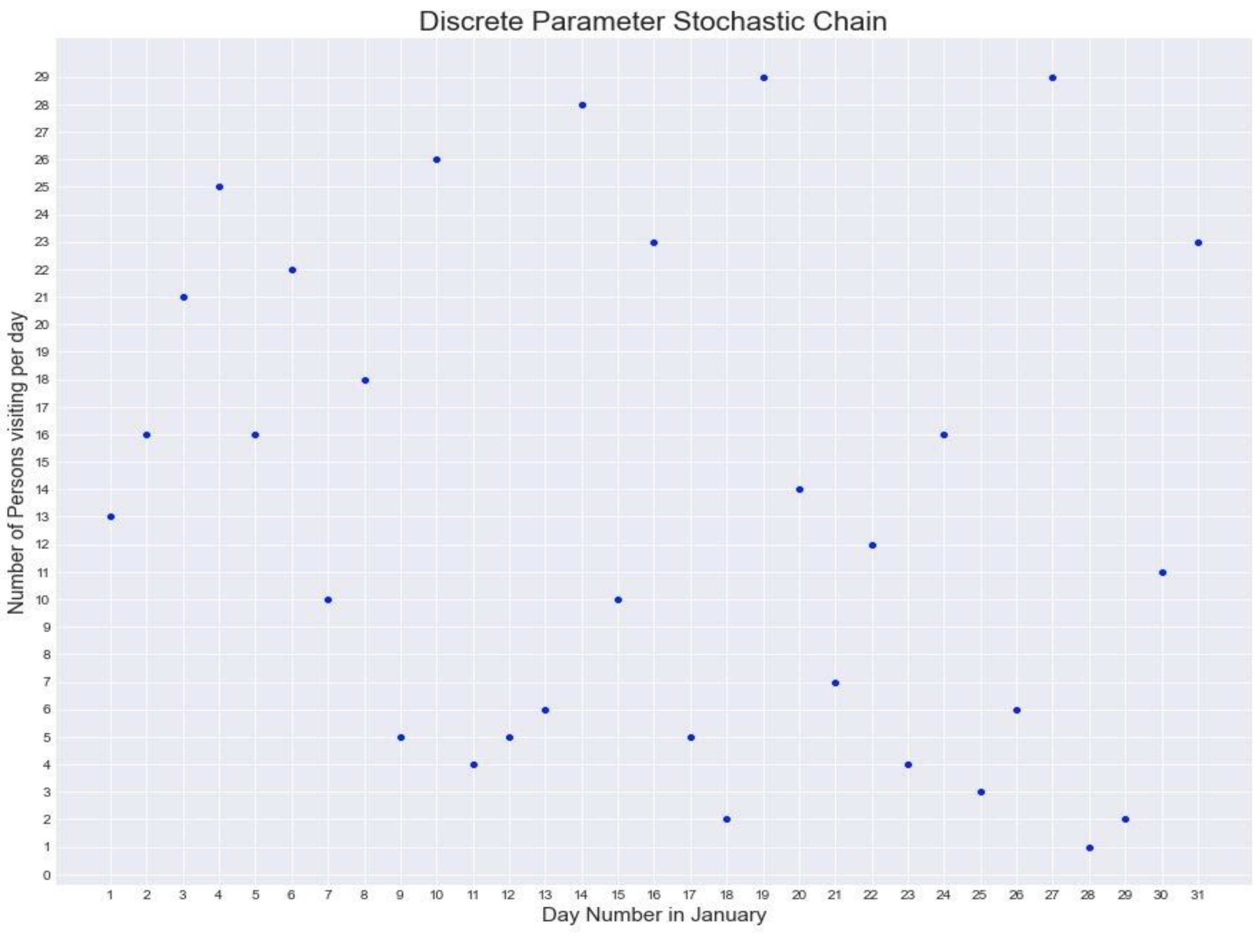
The above scenario correctly resembles the problem to which Chinese Restaurant Process provides the solution. This process continues, with each customer choosing to either sit at an occupied table with a probability proportional to the number of customers already there (i.e., they are more likely to sit at a table with many customers than few), or an unoccupied table. At time n, the n customers have been partitioned among m ≤ n tables (or blocks of the partition). The results of this process are exchangeable, meaning the order in which the customers sit does not affect the probability of the final distribution. This property greatly simplifies a number of problems in population genetics, linguistic analysis, and image recognition.

BACKGROUND KNOWLEDGE

Discrete Time Stochastic Process

A stochastic or random process can be defined as a collection of random variables that is indexed by some mathematical set, meaning that each random variable of the stochastic process is uniquely associated with an element in the set.
A stochastic process can be classified in different ways, for example, by its state space, its index set, or the dependence among the random variables. One common way of classification is by the cardinality of the index set and the state space.

When interpreted as time, if the index set of a stochastic process has a finite or countable number of elements, then the stochastic process is said to be in discrete time. If the index set is some interval of the real line, then time is said to be continuous. The two types of stochastic processes are respectively referred to as discrete-time and continuous-time stochastic processes. If the state space is n-dimensional Euclidean space, then the stochastic process is called a n-dimensional vector process or n-vector process.

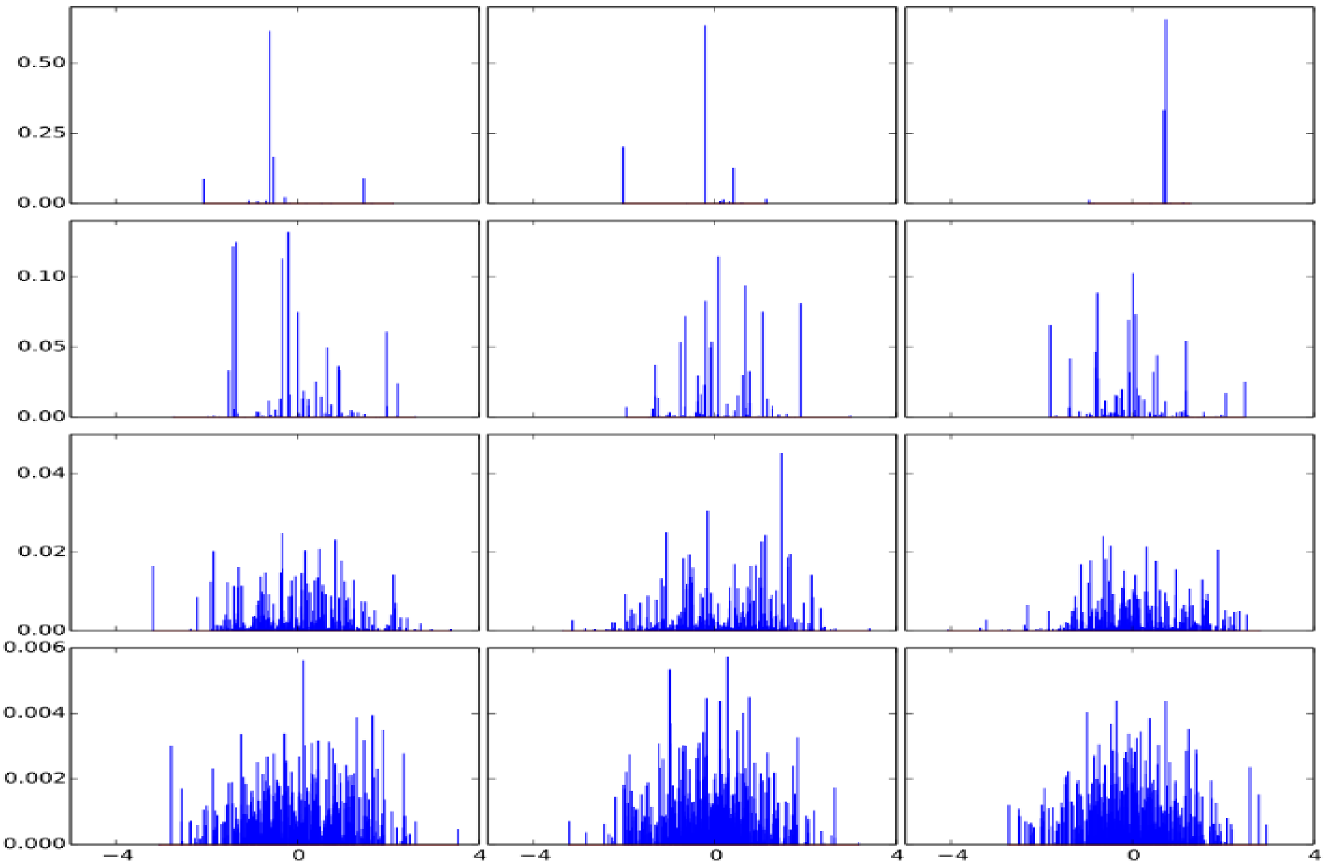


An example of Discrete Time stochastic process

Dirichlet Process

The Dirichlet processes (after Peter Gustav Lejeune Dirichlet) are a family of stochastic processes whose realizations are probability distributions. In other words, a Dirichlet process is a probability distribution whose range is itself a set of probability distributions. It is often used in Bayesian inference to describe the prior knowledge about the distribution of random variables.

The Dirichlet process is specified by a base distribution **'H'** and a positive real number **'α'** called the concentration parameter (also known as scaling parameter). The base distribution is the expected value of the process, i.e., the Dirichlet process draws distributions "around" the base distribution the way a normal distribution draws real numbers around its mean. The distributions drawn from the Dirichlet process are almost surely discrete. The scaling parameter specifies how strong this discretization is: in the limit of $\alpha \to 0$, the realizations are all concentrated at a single value, while in the limit of $\alpha \to \infty$ the realizations become continuous. Between the two extremes the realizations are discrete distributions with less and less concentration as α increases. The Dirichlet process is applied in data mining, machine learning, natural language processing, computer vision and bioinformatics.



Draws from the Dirichlet process DP (N (0,1), α). The four rows use different α (top to bottom: 1, 10, 100 and 1000) and each row contains three repetitions of the same experiment

APPROACH AND IMPLEMENTATION

We will define a distribution on the space of partitions of the positive integers, N. This would induce a distribution on the partitions of the first n integers, for every n ∈ N. Now reimagine the scenario presented on the first page in Introduction section. The tables are chosen according to the following random process:

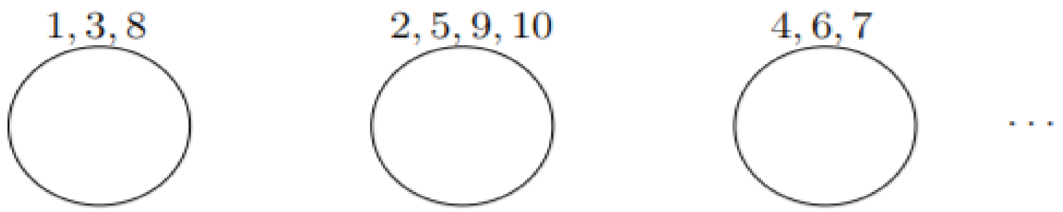
1. The first customer always chooses the first table.
2. The nth customer chooses the first unoccupied table with probability α/(n−1+α), and an occupied table with probability c/(n−1+α), where c is the number of people sitting at that table.

$$P(cust_n = table\ c/cust_1, \ldots, cust_{n-1}) = \frac{n_c}{\alpha + \sum_{c \in \varrho} n_c}$$

$$P(cust_n = new\ table/cust_1, \ldots, cust_{n-1}) = \frac{\alpha}{\alpha + \sum_{c \in \varrho} n_c}$$

$$where \sum_{c \in \varrho} n_c = n - 1$$

In the above, α is a scalar parameter of the process. We can see that the above condition defines a probability distribution. Let us denote by k_n the number of different tables occupied after n customers have walked in. Then $1 \leq k_n \leq n$ and it follows from the above description that precisely tables 1, . . ., k_n are occupied.



Circles represent tables and the numbers around them are the customers sitting at that table.

The following equation represents the probability of having distribution of customers shown in figure above.

$$\begin{aligned} \Pr(z_1, \ldots, z_{10}) &= \Pr(z_1)\Pr(z_2|z_1)\ldots\Pr(z_{10}|z_1, \ldots, z_9) \\ &= \frac{\alpha}{\alpha}\frac{\alpha}{1+\alpha}\frac{1}{2+\alpha}\frac{\alpha}{3+\alpha}\frac{1}{4+\alpha}\frac{1}{5+\alpha}\frac{2}{6+\alpha}\frac{2}{7+\alpha}\frac{2}{8+\alpha}\frac{3}{9+\alpha} \end{aligned}$$

Now, we will sample the Chinese Restaurant process using Python3. We will take three values of α that are: 1, 10, 100. For every α, we will take five samples. The sampling will be done for thousand customers. For sampling purposes, we will use Uniform Distribution to generate random variables with the range [0, 1].

For every new customer, we calculate the probability of sitting at any occupied table or an unoccupied table according to the above process. Then, we compare the value of random variable generated through Uniform Distribution to the probabilities and find the most probable output for that customer. We sample the Number of Occupied Tables against the Number of Customers in the Restaurant.

We have function 'compute_cum_prob' that calculates the probability according to the Chinese Restaurant Process. This function returns an array containing the cumulative sum of probabilities of sitting at any available table. The function 'new_customer' compares the cumulative sum of probabilities returned by above function to the randomly generated value from uniform distribution between [0, 1]. According to this comparison, the table count is updated and returned to the 'chinese_restaurant_process'. This process is completed for all thousand customers. This entire process of table allocation is repeated five times for every value of α.

```
def compute_cum_prob(counts, alpha):
    probs = np.array(counts)
    norm = probs.sum() + alpha
    probs = probs/norm
    probs = list(probs)
    probs.append(alpha/norm)

    cum = np.zeros(len(probs))
    for i in range(len(probs)):
        cum[i] = np.sum(probs[:i+1])
    return cum

def new_customer(counts, alpha):
    unif = stats.uniform()
    u = unif.rvs(1)
    cum = compute_cum_prob(counts, alpha)
    for i, prob_c in enumerate(cum):
        if u < prob_c:
            if i == len(cum)-1:
                counts.append(1)
            else:
                counts[i] +=1
            break
    return counts

def chinese_restaurant_process(counts, alpha, n_cust):
    for j in range(n_cust):
        counts = new_customer(counts, alpha)
    return counts
```

Code Cell containing the implementation of Chinese Restaurant Process

RESULT & ANALYSIS

With help of all the functions defined to implement Chinese Restaurant Process, we finally call the 'main' function to get the results of sampling. Through this sampling, we will be able to understand the impact of concentration parameter α on the Distribution taken up in the Process such as the discretization of the Distribution Curve and the the shape of the curve obtained through Graphical Analysis.

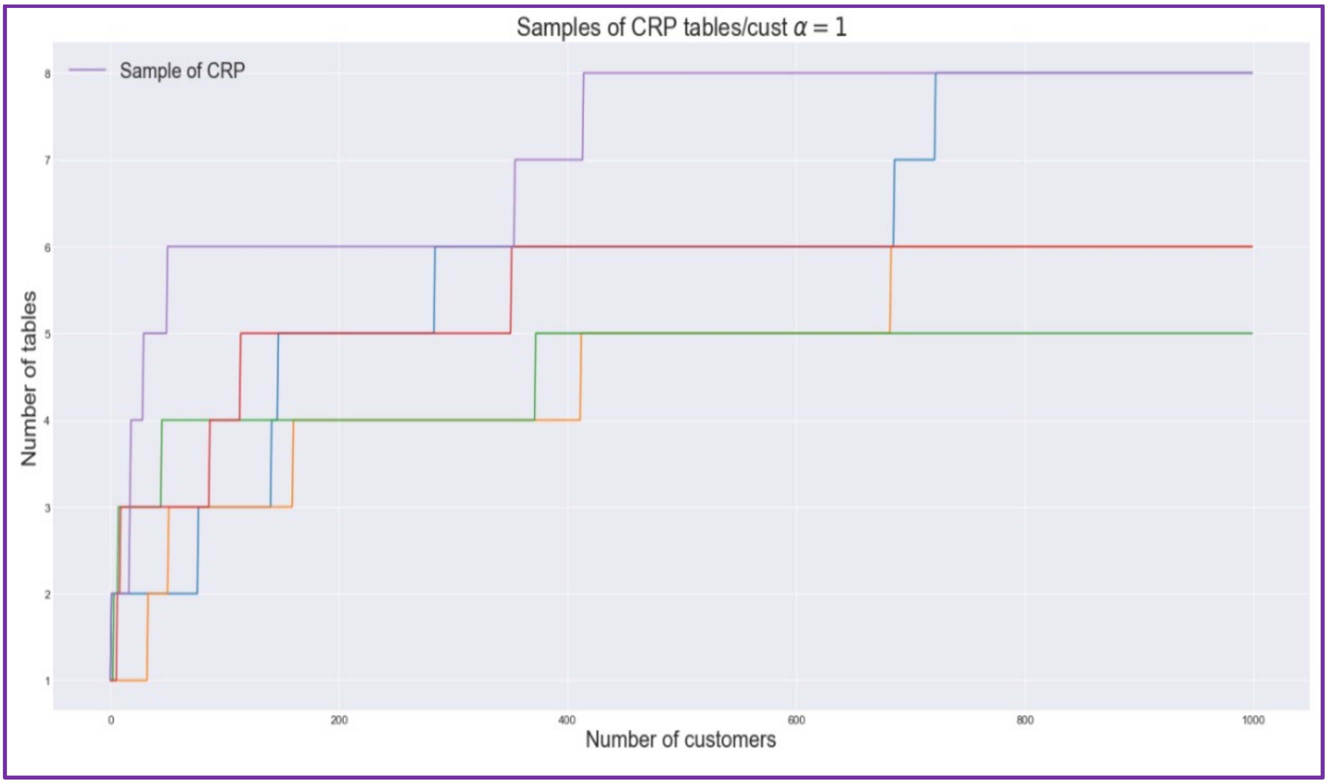
We will plot three graphs for different values of α that are: [1, 10, 100]. For each value of α, we will take five sample which will give us five subplots in every plot. The horizontal axis in the graph represents the Number of Customers present and the vertical axis represents the Number of Table Occupied correspondingly. For plotting purposes, we will use 'matplotlib' and 'seaborn' modules available in Python3 framework.

```
def main():
    # Plot parameters
    total_customers = 1000
    samples_crp = 5

    for alpha in [1, 10, 100]:
        fig = plt.figure(figsize=(20, 10))
        ax = fig.add_subplot(111)

        for j in range(samples_crp):
            counts = list()
            num_tables = list()
            for i in range(total_customers):
                counts = chinese_restaurant_process(counts, alpha, n_cust=1)
                num_tables.append(len(counts))
                print('Processed customer: Alpha %s Sample %s %s/%s' % (alpha, j, i+1, total_customers))
                clear_output(wait=True)
            if j != samples_crp-1:
                ax.plot(range(total_customers), num_tables)
        ax.plot(range(total_customers), num_tables, label='Sample of CRP')
        ax.set_title(r'Samples of CRP tables/cust $\alpha=%s$' % alpha, size=title_size)
        ax.set_ylabel('Number of tables', size=axis_size)
        ax.set_xlabel('Number of customers', size=axis_size)
        ax.tick_params(labelsize=axis_size-10)
        plt.legend(prop={'size': legend_size})
        plt.tight_layout()
        plt.savefig('images/crp/chinese_restaurant_process_tables_alpha_%s.png' % alpha, dpi=100)

main()
```

Code Cell containing 'main' function which calls for above mentioned funtions

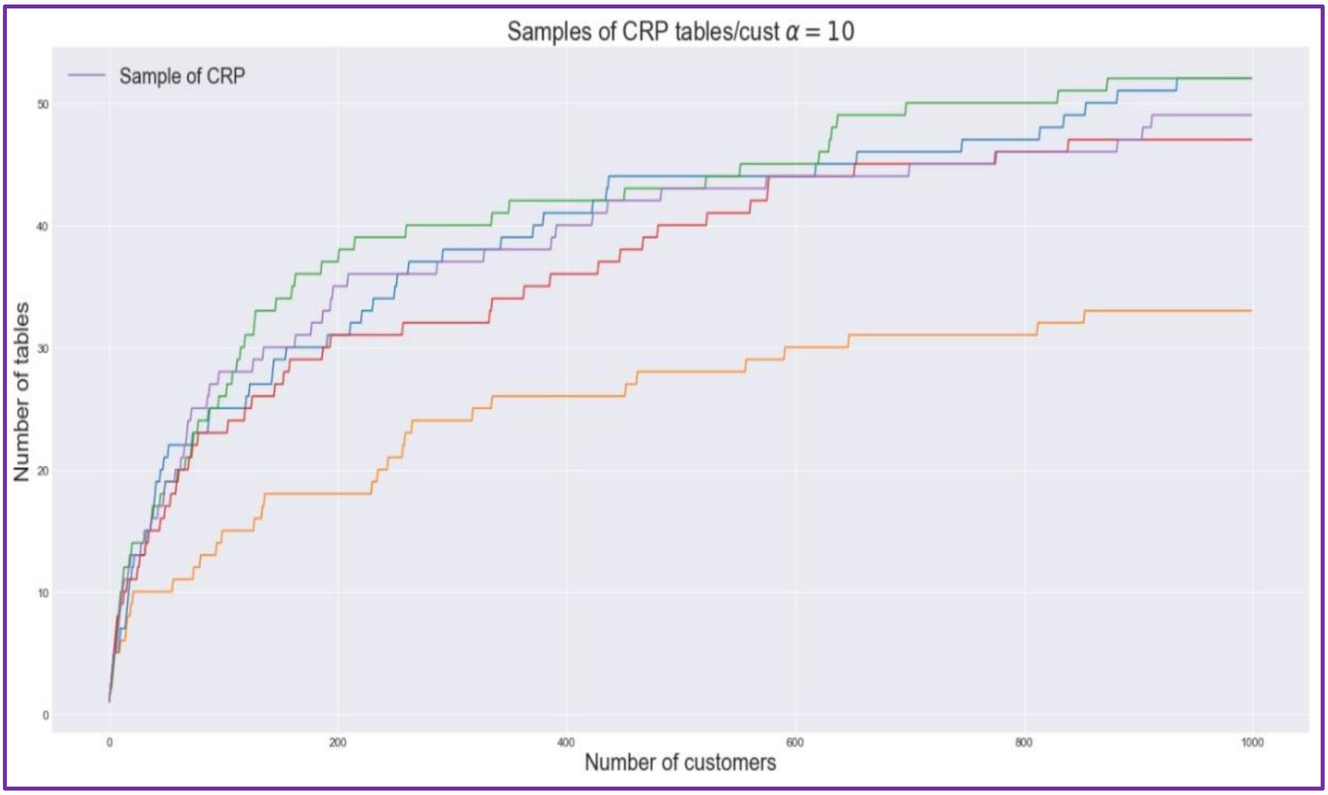
Now, we will see the output for each value of α, one by one. With increase in value of α, we will analyse the behaviour of the Distribution along with the plot. This sampling establishes the Chinese Restaurant Process Graphically.

a) α = 1



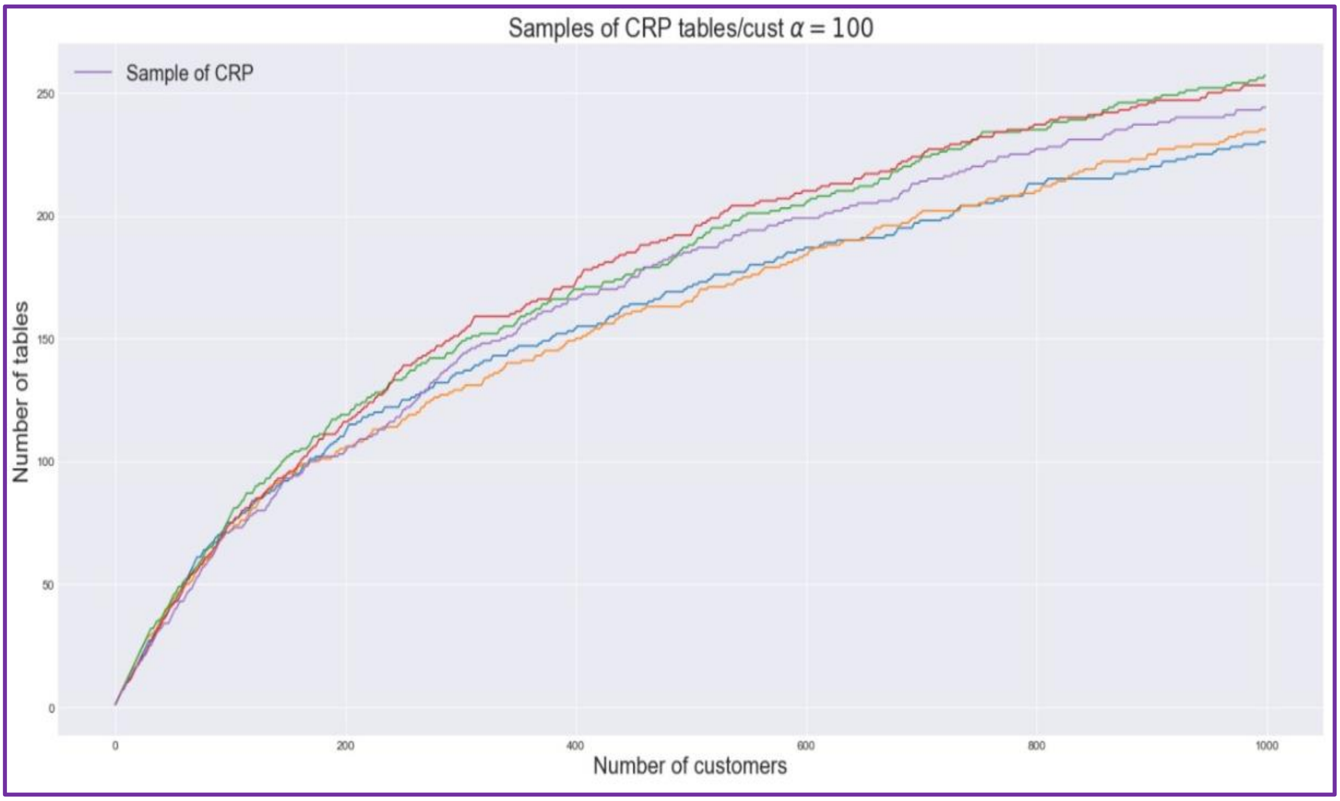
Analysis: Through this plot, we can analyse that for smaller values of α, we have very discrete nature of distribution. This can be clearly seen though the Graph obtained. The Number of table occupied resembles a step function which is accordance with the Discrete-Time Stochastic Process properties.

b) α = 10



Analysis: Now, for slightly higher value of α, we get a more continuous behaviour for the Distribution as shown in the graph. This show that the concentration parameter α is inversly proportional to the discrete nature of the distribution given by Chinese Restaurant Process. All the five samples show that the Distribution is headed towards a Continuous behaviour with increase in value of α.

c) α = 100



Analysis: Finally, in this plot, we have obtained almost a continuous curve. All five samples at this value of α confirms this observation. We can say that as the value of α increases the plot becomes more and more continuous. This is in accordance with the Dirichlet Process.

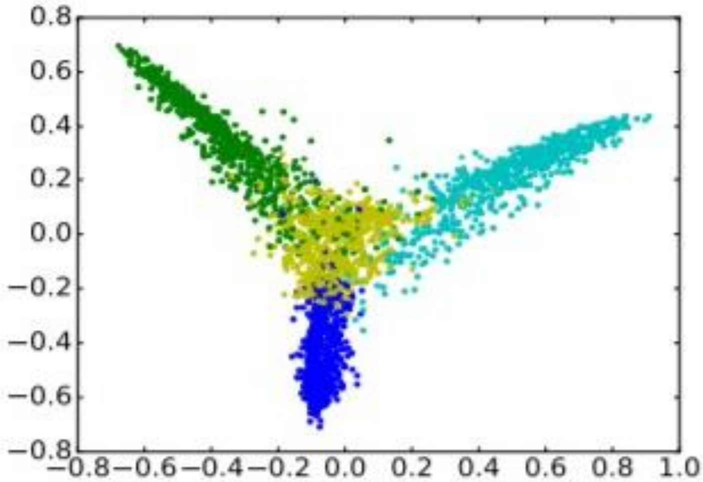
Though this sampling, we can observe that the CRP defines a prior over the partitioning of the sample space S, inducing a distribution over partitions of the sample space. The Chinese restaurant process takes the clustering property of the Dirichlet process without the base measure H, we can see the clustering property in the probability of table assignations, tables with more customers tend to get more customers $\boldsymbol{P(customer\ n\ at\ table\ c) \propto nC}$ **.**

The number of tables depend on the number of customers n and the concentration parameter α, both the mean and variance are logarithmically proportional to the the number of customers scaled by the concentration parameter.

APPLICATIONS

The Chinese Restaurant Process along with Dirichlet Process finds many applications in real life as well as for industrial purposes. Some of the applications include:

In a Distance-Based CRP, there is a slight modification when a new customer comes, that customer will sit on a table which he is closest to and not on a table where there are more number of people sitting. A very interesting application of this can be done to perform document clustering using word vectors.



An example of Distance Dependent CRP clustering

We can cluster (or classify) a document by using the Distance Dependent CRP. Scan through the document word by word, while doing so assign the first word with a probability 1/1+n, where n is number of clusters. Initially since there are no clusters n = 0, then the first word goes to table no. 1 i.e. cluster no. 1. Now use the vector of that word to represent the vector of the cluster. Now for the 2nd word do the same, if the cluster already has a word then add the vector of the new word with the vector of the previous word in the cluster and call it a cluster vector. Perform the operation until all the words in the document are clustered. This is an effective way of document clustering and has proved to be effective. One of the advantages of this method is its highly scalable and can perform very well in distributed environments.

Immediate use of the Chinese restaurant process is in the field of modeling human behaviors. The Chinese restaurant process can also be used for high-level "discriminative" work, as in image processing. Developing clusters of images according to a Chinese restaurant process can help machine learning programs to better adapt to sets of training rules and produce discriminative outcomes. So, in a sense, the Chinese restaurant process can be used for either behavioral modeling, or technical modeling, or both.

Distance-Dependent CRP can help in image over-segmentation. Image over-segmentation aims to partition an image into spatially adjacent and spectrally homogeneous regions. It could reduce the complexity of image representation and enhance the efficiency of subsequent image processing.

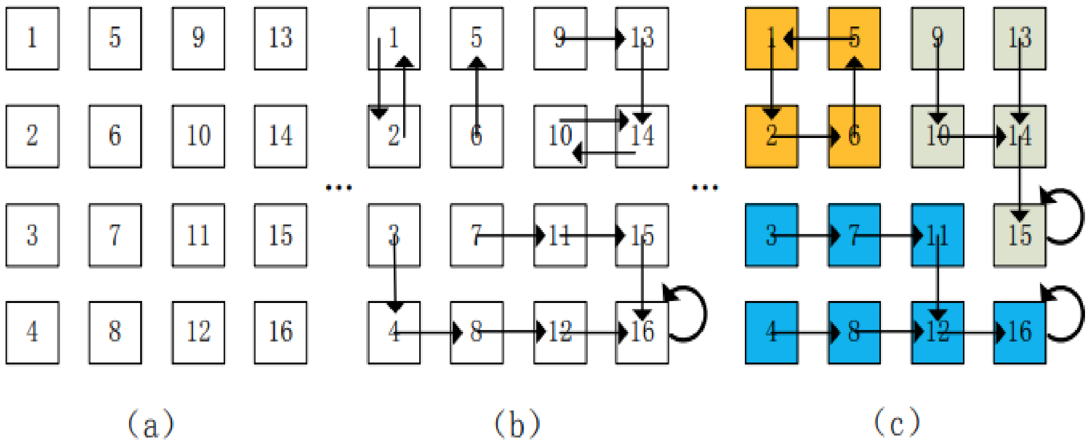
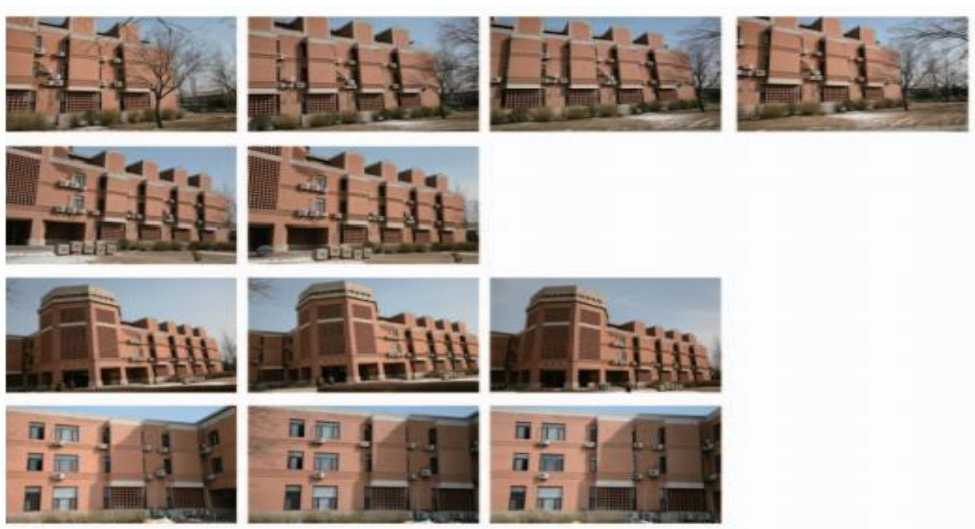


Figure 1. The illustration of distance dependent Chinese restaurant process (ddCRP) for image over-segmentation; (**a**) an image (i.e., a restaurant) with 16 pixels (i.e., customers) where each numbered square denotes a pixel; (**b**) each arrow indicates a customer choose to sit with another customer; and (**c**) a segment consists of pixels with a same color and every two pixels within a segment can reach to each other along the inferred customer assignments.

The Chinese restaurant process is closely connected to Dirichlet processes and Pólya's urn scheme, and therefore useful in applications of Bayesian statistics including nonparametric Bayesian methods. The Generalized Chinese Restaurant Process is closely related to Pitman–Yor process. These processes have been used in many applications, including modeling text, clustering biological microarray data, biodiversity modelling, and image reconstruction.



For reconstruction purposes the image data needs to be ordered. Key frame selection achieved using **Distance-dependent Chinese Restaurant Process (DdCRP)**.

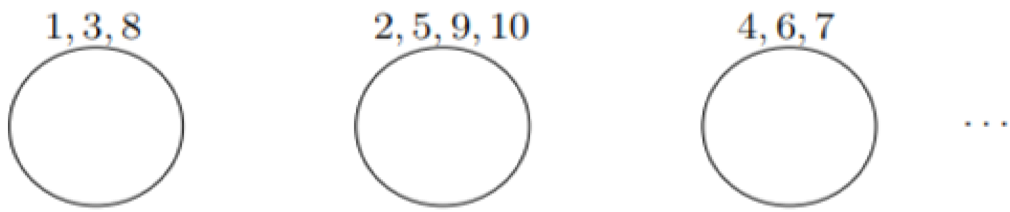
Image Regrouping using DdCRP

CONCLUSION & LEARNING

In this project, we have studied and sampled Chinese Restaurant Process. We have sampled the Chinese Restaurant Process for three values of α: [1, 10, 100]. Using the Graphical Analysis, we have established the results of Chinese Restaurant Process as well as Dirichlet Process.

We can make the following observations based on our sampling:

- The probability of a seating is invariant under permutations. Permuting the customers permutes the numerators in the above computation, while the denominators remains the same. This property is known as exchangeability.
- Any seating arrangement creates a partition. For example, the following seating arrangement partitions customers 1, . . . , 10 into the following three groups (1, 3, 8),(2, 5, 9, 10),(4, 6, 7).



Circles represent tables and the numbers around them are the customers sitting at that table.

  Exchangeability now implies that two seating arrangements whose partitions consist of the same number of components with identical sizes will have the same probability. For instance, the probability of any seating arrangement of ten customers where three tables are occupied, with three customers each on two of the tables and the remaining four on the third table, will have the same probability as the seating in our example. Thus we could define a distribution on the space of all partitions of the integer n, where n is the total number of customers. The number of partitions is given by the partition function p(n), which has no simple closed form. Asymptotically, $p(n) = \exp(O(\sqrt{n}))$.
- The expected number of occupied tables for n customers grows logarithmically. In particular $E[k_n|\alpha] = O(\alpha \log n)$.

Chinese Restaurant Process is very famous mathematical tool which can be used in various fields and domains. The ease of use and simplicity of Chinese Restaurant Process when combined with the powerful results makes it a very favoured and helpful tool for many applications as mentioned above.

REFERENCES

- https://en.wikipedia.org/wiki/Chinese_restaurant_process
- https://www.jmlr.org/papers/volume12/blei11a/blei11a.pdf
- Amir-Ahmadi et. al (2016). Choosing Prior Hyperparameters. Retrieved February 13, 2018 from: https://www.richmondfed.org/-/media/richmondfedorg/publications/research/working_papers/2016/pdf/wp16-09.pdf
- Stanford University (2016). Chinese Restaurant Viewpoint. Retrieved February 13, 2018 from: https://cs.stanford.edu/~ppasupat/a9online/1083.html
- Bello, J. et. al (2008). ISMIR 2008: Proceedings of the 9th International Conference of Music Information Retrieval. Lulu Press.
- Aldous, D. J. (1985). "Exchangeability and related topics". École d'Été de Probabilités de Saint-Flour XIII — 1983. Lecture Notes in Mathematics. 1117. pp. 1–1.