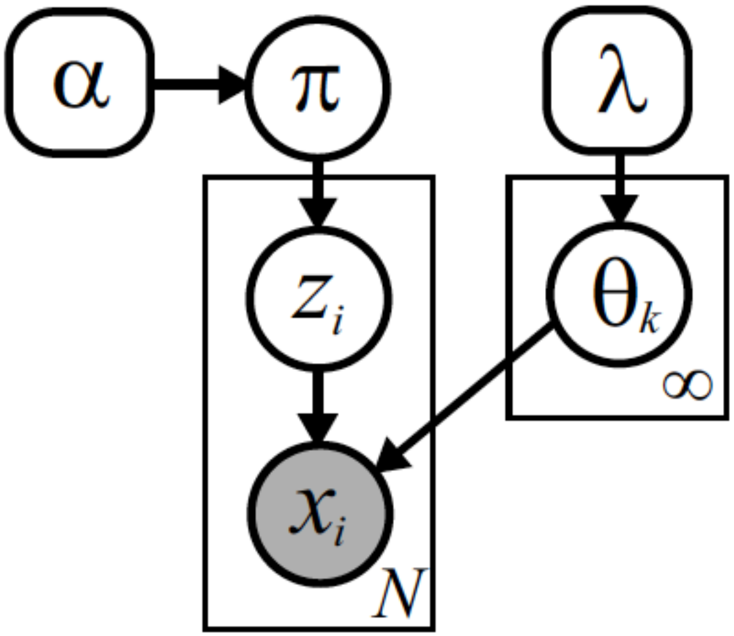
Applied Bayesian Nonparametrics

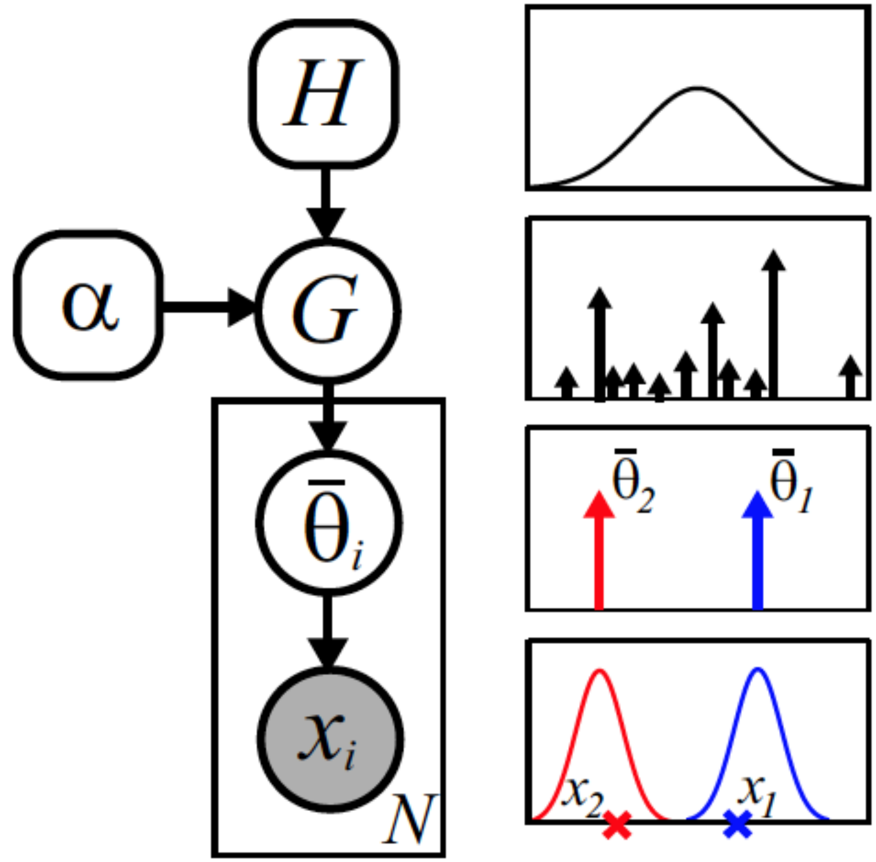
Special Topics in Machine Learning
Brown University CSCI 2950-P, Fall 2011

October 20: Binary Latent Feature Models, Indian Buffet Process

Dirichlet Process Mixtures

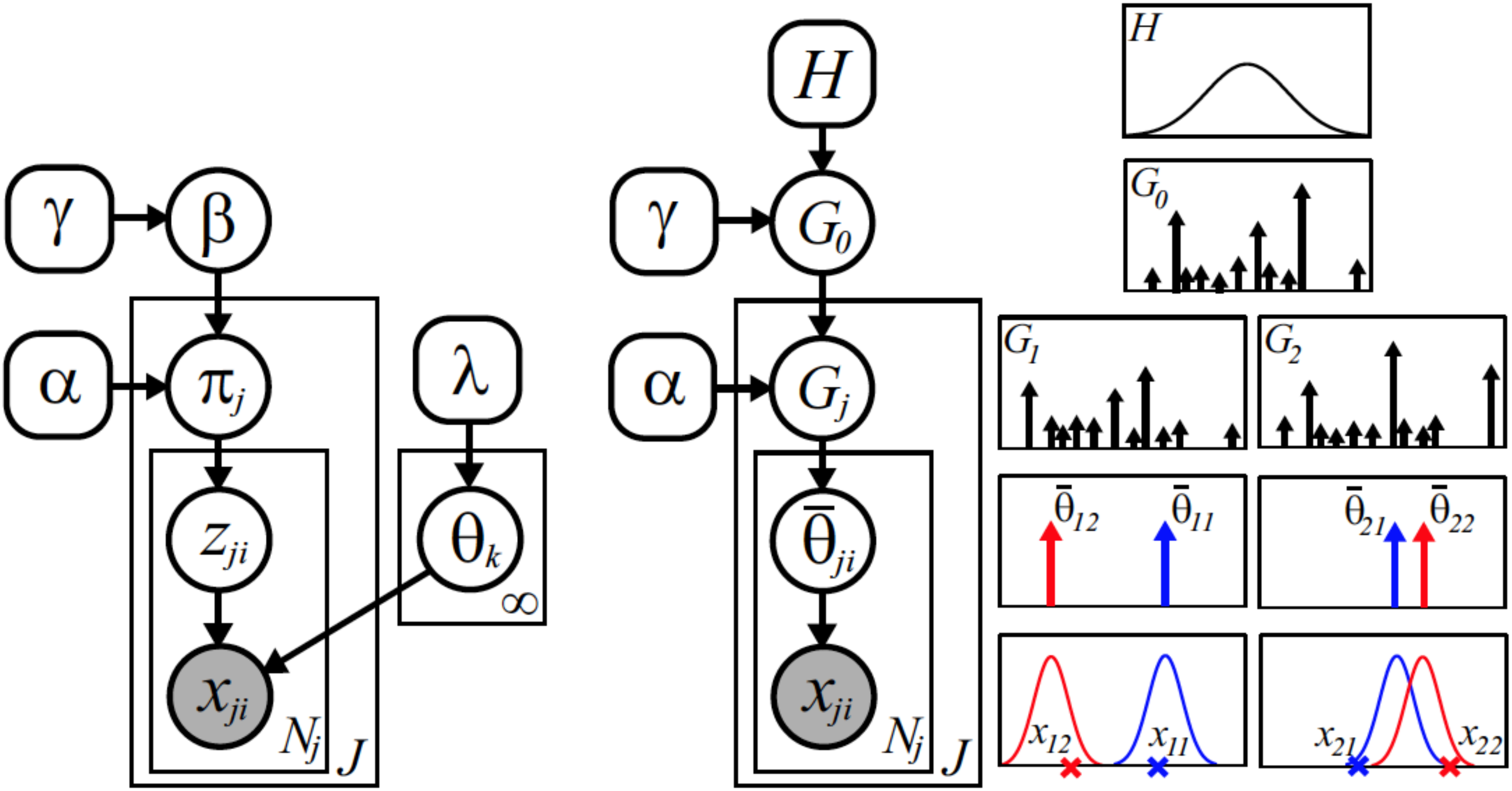
$$p(x \mid \pi, \theta_1, \theta_2, \ldots) = \sum_{k=1}^{\infty} \pi_k f(x \mid \theta_k)$$





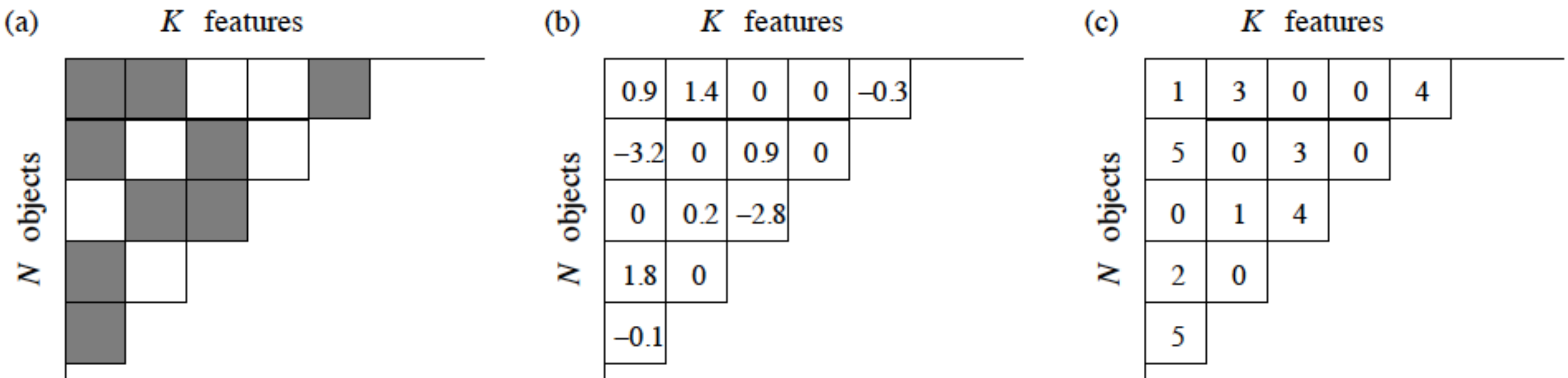
- ***Mixture*** model: Each observation is associated with exactly one underlying cluster or class
- Complex datasets may require a huge number of classes

Hierarchical Dirichlet Processes



- ***Admixture*** model: Each group of observations is associated with a *distribution* over latent clusters/classes/topics
- Conservation: Increasing mass of one topic decreases others

Latent Feature Models

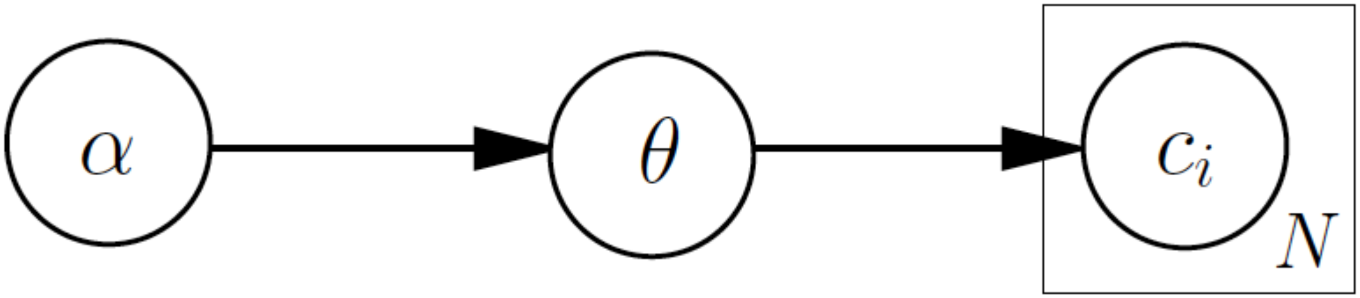


Focus today: Distributions on binary matrices indicating feature presence/absence

Depending on application, features can be associated with any parameter value of interest

- ***Latent Feature*** model: Each group of observations is associated with a *subset* of the possible latent features
- Factorial power: There are 2^K combinations of K features
- Question: What is the analog of the DP for feature modeling?

Finite Dirichlet Mixtures



$$p(\theta) = \frac{\prod_{k=1}^{K} \theta_k^{\alpha_k - 1}}{D(\alpha_1, \alpha_2, \ldots, \alpha_K)} \qquad P(\mathbf{c}|\theta) = \prod_{i=1}^{N} P(c_i|\theta) = \prod_{i=1}^{N} \theta_{c_i}$$

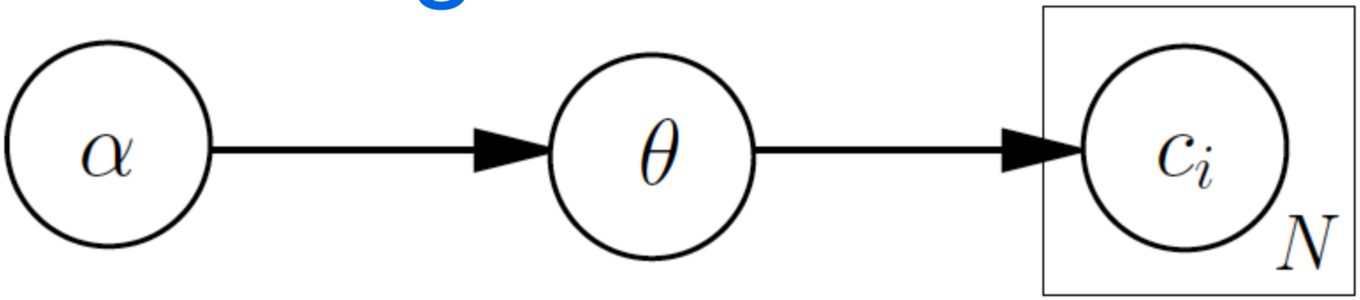
$$D(\alpha_1, \alpha_2, \ldots, \alpha_K) = \int_{\Delta_K} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \, d\theta = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)}$$

$$D(\tfrac{\alpha}{K}, \tfrac{\alpha}{K}, \ldots, \tfrac{\alpha}{K}) = \frac{\Gamma(\frac{\alpha}{K})^K}{\Gamma(\alpha)}$$

$$P(\mathbf{c}) = \int_{\Delta_K} \prod_{i=1}^{n} P(c_i|\theta) p(\theta) \, d\theta = \frac{\prod_{k=1}^{K} \Gamma(m_k + \frac{\alpha}{K})}{\Gamma(\frac{\alpha}{K})^K} \frac{\Gamma(\alpha)}{\Gamma(N + \alpha)}$$

Marginal likelihoods generally expressed as ratios of normalizers

From Assignments to Partitions



$$P(\mathbf{c}) = \int_{\Delta_K} \prod_{i=1}^{n} P(c_i|\theta)p(\theta)\,d\theta = \frac{\prod_{k=1}^{K}\Gamma(m_k + \frac{\alpha}{K})}{\Gamma(\frac{\alpha}{K})^K}\frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}$$

$$= \left(\frac{\alpha}{K}\right)^{K_+}\left(\prod_{k=1}^{K^+}\prod_{j=1}^{m_k-1}(j+\tfrac{\alpha}{K})\right)\frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}$$

K_+ is the number of classes for which $m_k > 0$

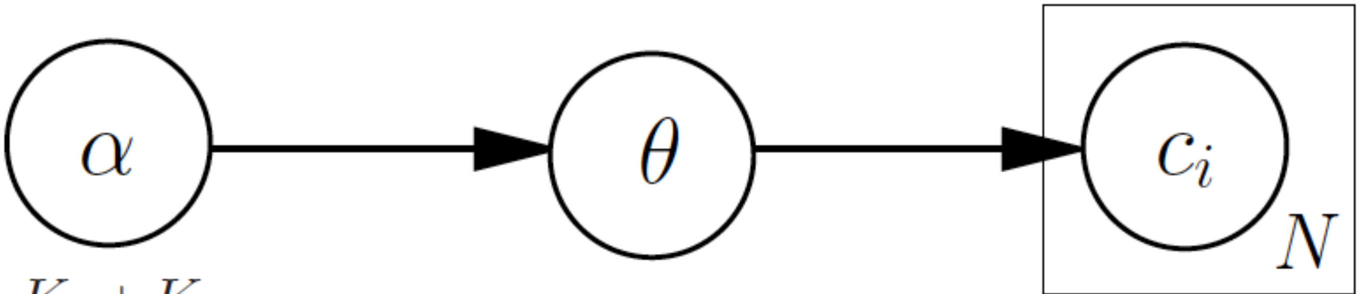
There are K^N possible values for $\mathbf{c}$

$P(\mathbf{c}) \rightarrow 0$ as $K \rightarrow \infty$

Instead look at label equivalence classes: $K = K_0 + K_+$

$$P([\mathbf{c}]) = \sum_{\mathbf{c}\in[\mathbf{c}]} P(\mathbf{c}) = \frac{K!}{K_0!}\left(\frac{\alpha}{K}\right)^{K_+}\left(\prod_{k=1}^{K^+}\prod_{j=1}^{m_k-1}(j+\tfrac{\alpha}{K})\right)\frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}$$

An Infinite Limit



$$K = K_0 + K_+$$

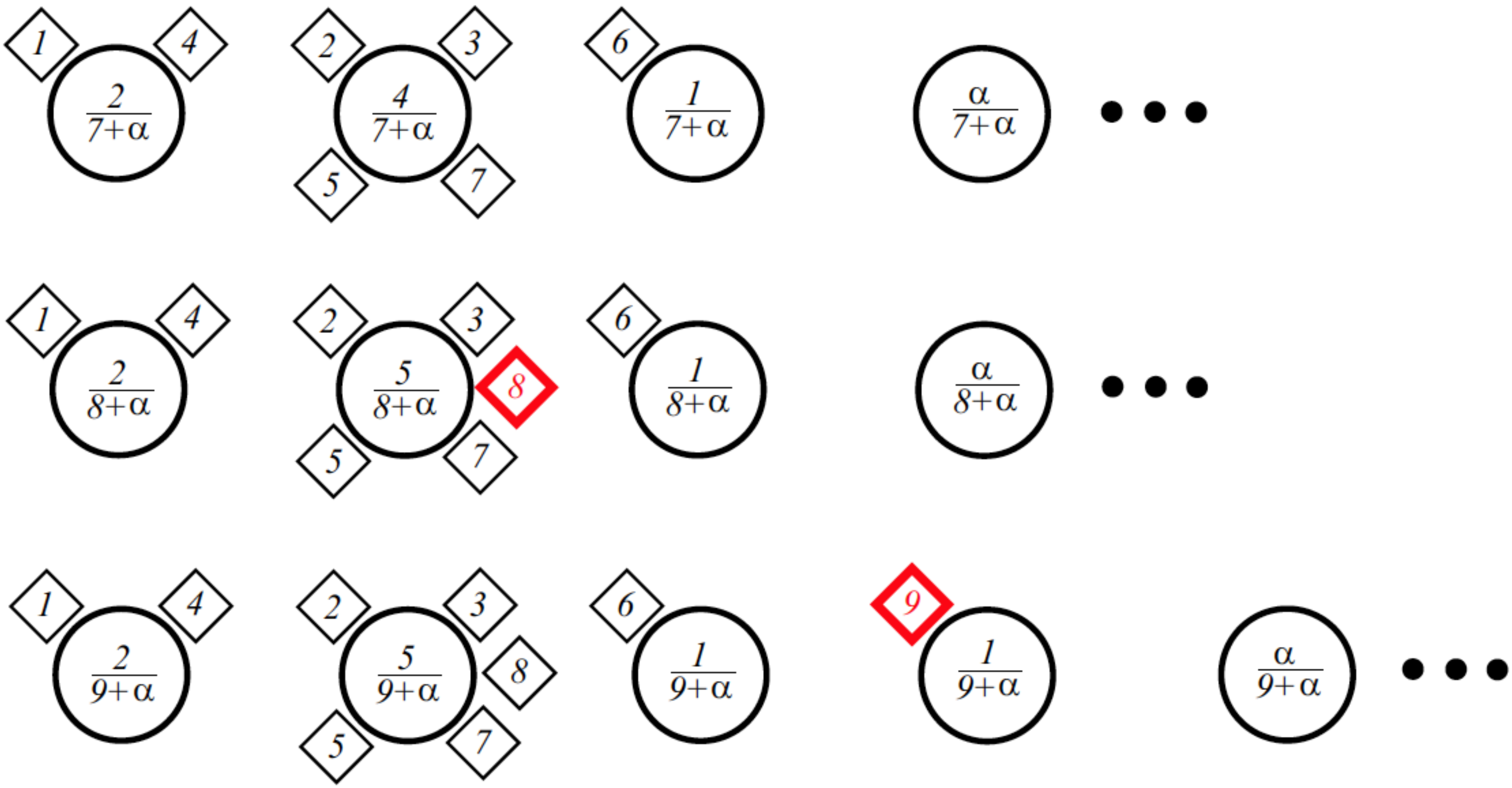
$$P([\mathbf{c}]) = \sum_{\mathbf{c}\in[\mathbf{c}]} P(\mathbf{c}) = \frac{K!}{K_0!}\left(\frac{\alpha}{K}\right)^{K_+}\left(\prod_{k=1}^{K^+}\prod_{j=1}^{m_k-1}(j+\tfrac{\alpha}{K})\right)\frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}$$

$$\lim_{K\to\infty} \alpha^{K_+}\cdot\frac{K!}{K_0!\,K^{K_+}}\cdot\left(\prod_{k=1}^{K_+}\prod_{j=1}^{m_k-1}(j+\tfrac{\alpha}{K})\right)\cdot\frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}$$

$$= \alpha^{K_+}\cdot 1 \cdot \left(\prod_{k=1}^{K_+}(m_k-1)!\right)\cdot\frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}$$

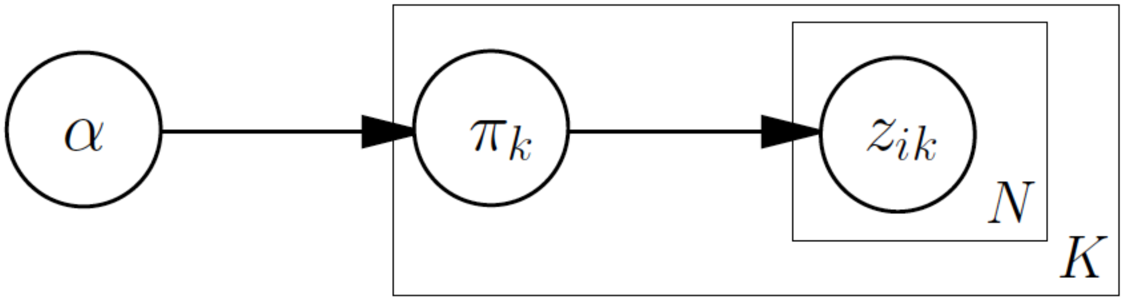
$$\begin{aligned}\frac{K!}{K_0!\,K^{K_+}} &= \frac{\prod_{k=1}^{K_+}(K-k+1)}{K^{K_+}}\\ &= \frac{K^{K_+} - \frac{(K_+-1)K_+}{2}K^{K_+-1} + \cdots + (-1)^{K_+-1}(K_+-1)!K}{K^{K_+}}\\ &= 1 - \frac{(K_+-1)K_+}{2K} + \cdots + \frac{(-1)^{K_+-1}(K_+-1)!}{K^{K_+-1}}.\end{aligned}$$

Chinese Restaurant Process



$$p(z_{N+1} = z \mid z_1, \ldots, z_N, \alpha) = \frac{1}{\alpha + N} \left(\sum_{k=1}^{K} N_k \delta(z, k) + \alpha \delta(z, \bar{k}) \right)$$

Finite Beta-Bernoulli Features



$$m_k = \sum_{i=1}^{N} z_{ik}$$

$$p(\pi_k) = \frac{\pi_k^{r-1}(1-\pi_k)^{s-1}}{B(r,s)}$$

$$P(\mathbf{Z}|\pi) = \prod_{k=1}^{K}\prod_{i=1}^{N} P(z_{ik}|\pi_k) = \prod_{k=1}^{K} \pi_k^{m_k}(1-\pi_k)^{N-m_k}$$

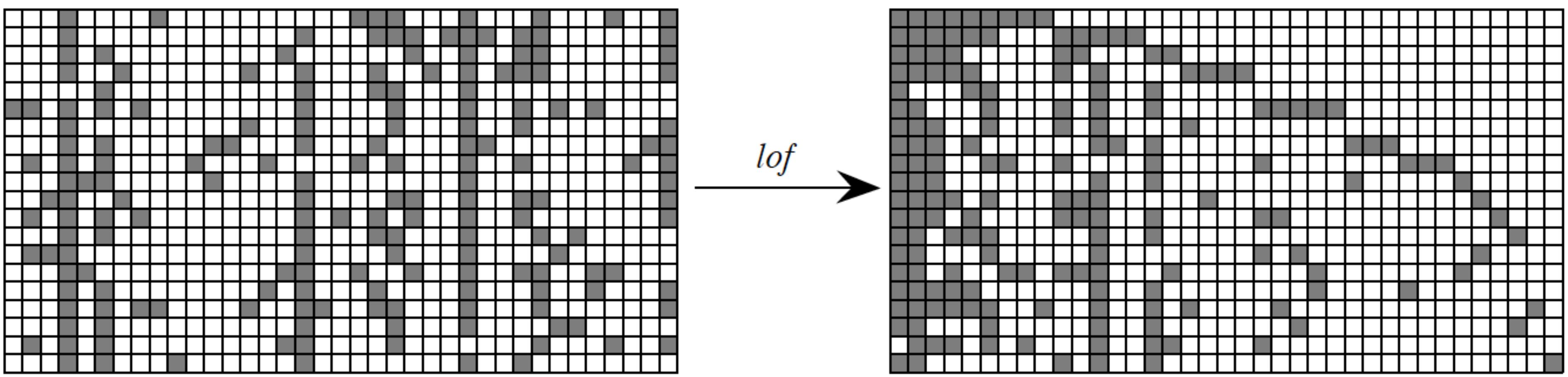
$$B(r,s) = \int_0^1 \pi_k^{r-1}(1-\pi_k)^{s-1}\, d\pi_k = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)}$$

$$B(\tfrac{\alpha}{K},1) = \frac{\Gamma(\frac{\alpha}{K})}{\Gamma(1+\frac{\alpha}{K})} = \frac{K}{\alpha}$$

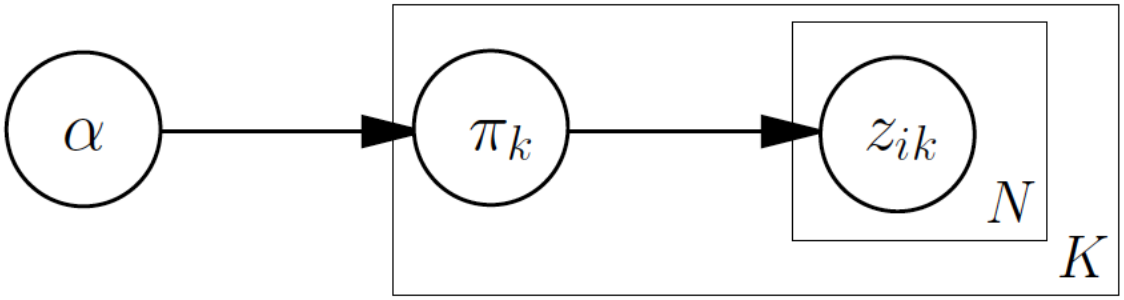
$$P(\mathbf{Z}) = \prod_{k=1}^{K}\int \left(\prod_{i=1}^{N} P(z_{ik}|\pi_k)\right) p(\pi_k)\, d\pi_k = \prod_{k=1}^{K} \frac{\frac{\alpha}{K}\Gamma(m_k+\frac{\alpha}{K})\Gamma(N-m_k+1)}{\Gamma(N+1+\frac{\alpha}{K})}$$

Marginal likelihoods generally expressed as ratios of normalizers

Left-Ordered Forms (LOFs)



Assignments to LOFs and a Limit



Consider histories *h*: All possible usage patterns for one feature across the N objects.

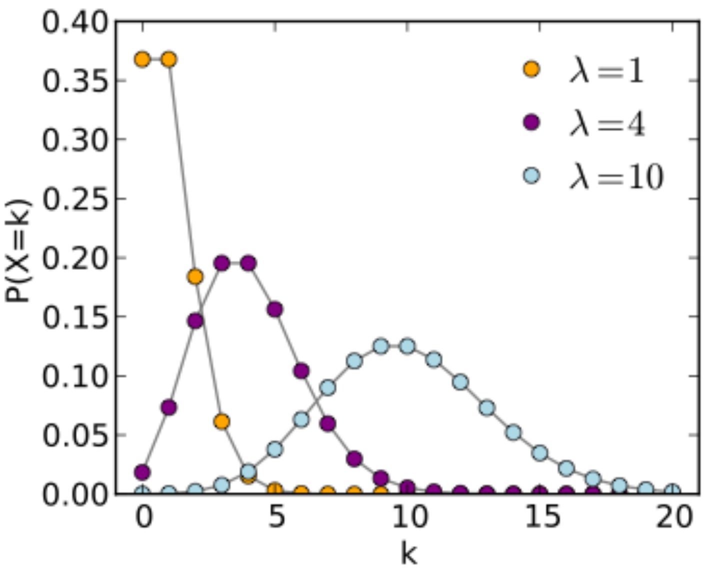
$$P([\mathbf{Z}]) = \sum_{\mathbf{Z}\in[\mathbf{Z}]} P(\mathbf{Z}) = \frac{K!}{\prod_{h=0}^{2^N-1} K_h!} \prod_{k=1}^{K} \frac{\frac{\alpha}{K}\Gamma(m_k + \frac{\alpha}{K})\Gamma(N - m_k + 1)}{\Gamma(N + 1 + \frac{\alpha}{K})}$$

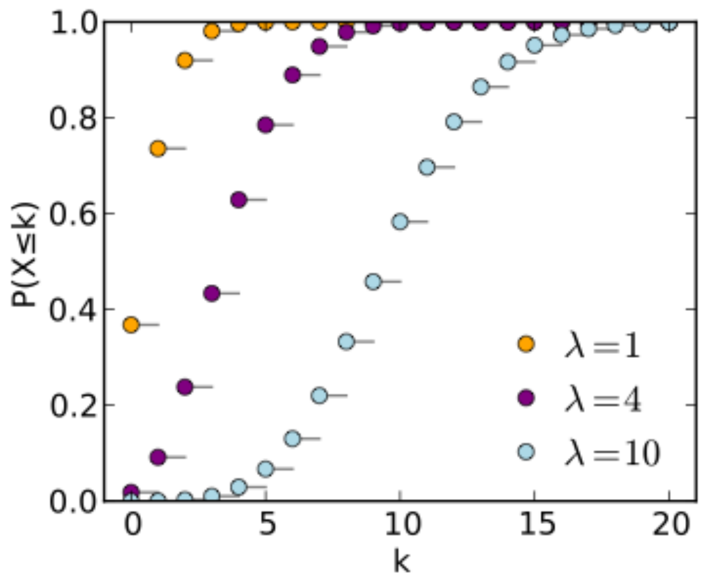
$$\lim_{K\to\infty} \frac{\alpha^{K_+}}{\prod_{h=1}^{2^N-1} K_h!} \cdot \frac{K!}{K_0!\, K^{K_+}} \cdot \left(\frac{N!}{\prod_{j=1}^{N}(j + \frac{\alpha}{K})}\right)^K \cdot \prod_{k=1}^{K_+} \frac{(N - m_k)! \prod_{j=1}^{m_k-1}(j + \frac{\alpha}{K})}{N!}$$

$$= \frac{\alpha^{K_+}}{\prod_{h=1}^{2^N-1} K_h!} \cdot 1 \cdot \exp\{-\alpha H_N\} \cdot \prod_{k=1}^{K_+} \frac{(N - m_k)!(m_k - 1)!}{N!},$$

$$H_N = \sum_{j=1}^{N} \frac{1}{j}$$

Poisson Distribution





$$p(x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

$$\lambda = \mathbb{E}[x] = \mathbb{V}[x]$$

If $X_i \sim \mathrm{Pois}(\lambda_i)$ follow a Poisson distribution with parameter λ_i and X_i are independent, then

$$Y = \sum_{i=1}^{N} X_i \sim \mathrm{Pois}\left(\sum_{i=1}^{N} \lambda_i\right)$$

Why Poisson? The Law of Rare Events

From Wikipedia:

We will prove that, for fixed λ, if

$$X_n \sim \mathrm{B}(n, \lambda/n); \qquad Y \sim \mathrm{Pois}(\lambda).$$

then for each fixed k

$$\lim_{n\to\infty} P(X_n = k) = P(Y = k).$$

To see the connection with the above discussion, for any Binomial random variable with large n and small p set $\lambda = np$. Note that the expectation $E(X_n) = \lambda$ is fixed with respect to n.

First, recall from calculus

$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda},$$

then since $p = \lambda / n$ in this case, we have

$$\begin{aligned}
\lim_{n\to\infty} P(X_n = k) &= \lim_{n\to\infty} \binom{n}{k} p^k (1-p)^{n-k} \\
&= \lim_{n\to\infty} \frac{n!}{(n-k)!k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
&= \lim_{n\to\infty} \underbrace{\left[\frac{n!}{n^k\,(n-k)!}\right]}_{A_n} \left(\frac{\lambda^k}{k!}\right) \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{\to\exp(-\lambda)} \underbrace{\left(1 - \frac{\lambda}{n}\right)^{-k}}_{\to 1} \\
&= \left[\lim_{n\to\infty} A_n\right] \left(\frac{\lambda^k}{k!}\right) \exp\left(-\lambda\right)
\end{aligned}$$

Like finite feature model, but there probabilities are random…