- goal :
  1. Derive the ML estimator for θ
  2. " " MAP " " ", assuming Dirichlet prior

1. ML Estimator :
   - probability of a single sample : $P(z^n|\theta) = \prod_{k=1}^{K} \{ \theta^{(z_k^n)} \}$
   - likelihood function : (assuming independence)

$$
\begin{aligned}
: L(\theta|D) &= P(z^1, z^2, \dots, z^N | \theta) \\
&= \prod_{n=1}^{N} P(z^n|\theta) \\
&= \prod_{n=1}^{N} \prod_{k=1}^{K} \{ \theta^{(z_k^n)} \} \\
&= \prod_{k=1}^{K} \{ \theta_k^{(\sum_{n=1}^{N} z_k^n)} \} \\
&= \prod_{k=1}^{K} \{ \theta_k^{N_k} \} \quad \text{where } N_k = \sum_{n=1}^{N} z_k^n = \text{\# of samples in state } k
\end{aligned}
$$

   - log likelihood : $\log(L(\theta|D)) = \log(\prod_{k=1}^{K} \{\theta_k^{N_k}\}) = \sum_{k=1}^{K} N_k \log\theta_k$
   - constrained optimization :

: maximize $\log(L(\theta|D))$ subject to $\sum_{k=1}^{K} \theta_k = 1$

$\therefore \mathcal{L}(\theta, \lambda) = \sum_{k=1}^{K} N_k \log\theta_k + \lambda(\sum_{k=1}^{K}\{\theta_k\} - 1)$

$\therefore$ For each $k \in \{1, \dots, k\}$:

$$\partial\mathcal{L}/\partial\theta_k = N_k/\theta_k + \lambda = 0$$

$$\therefore \theta_k = -N_k/\lambda$$

$\Longrightarrow \sum_{k=1}^{K}\{\theta_k\} = 1 \Longrightarrow \sum_{k=1}^{K}\{-N_k/\lambda\} = 1$

$\Longrightarrow -1/\lambda \sum_{k=1}^{K} N_k = 1 \Longrightarrow -N/\lambda = 1$

$\Longrightarrow \lambda = -N$, where $N = \sum_k N_k$ is the total # of samples

$\boxed{\text{- ML Estimator : } \theta_k^{(ML)} = -N_k/(-N) = N_k/N}$

2. MAP Estimator :

: Given Prior : Dirichlet distribution w/ hyperparameter $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_K]^T$:

- $P(\theta|\alpha) = 1/B(\alpha) \cdot \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$
- $B(\alpha) = \{\prod_{k=1}^{K} \Gamma(\alpha_k)\} / \{\Gamma(\sum_{k=1}^{K} \alpha_k)\}$

- Posterior Distribution : $P(\theta|D) = P(D|\theta) \cdot P(\theta) / P(D)$

$= P(D|\theta) \cdot P(\theta)$ (b/c $P(D)$ doesn't depend on θ)

$= \prod_{k=1}^{K} \theta_k^{N_k} \cdot \prod_{k=1}^{K} \theta_k^{(\alpha_k - 1)}$

$= \prod_{k=1}^{K} \theta_k^{(N_k + \alpha_k - 1)}$

- log posterior : $\log(P(\theta|D)) = \sum_{k=1}^{K} (N_k + \alpha_k - 1) \log \theta_k$
- constrained optimization :

: maximize $\log(P(\theta|D))$ subject to $\sum_{k=1}^{K} \theta_k = 1$

$\therefore \mathcal{L}(\theta, \lambda) = \sum_{k=1}^{K} (N_k + \alpha_k - 1) \log \theta_k + \lambda(\sum_{k=1}^{K}\{\theta_k\} - 1)$

$\therefore$ For each $k \in \{1, \ldots, K\}$:

$$\partial \mathcal{L} / \partial \theta_k = (N_k + \alpha_k - 1)/\theta_k + \lambda = 0$$

$\therefore \theta_k = -(N_k + \alpha_k - 1)/\lambda$

$\Longrightarrow \sum_{k=1}^{K} [-(N_k + \alpha_k - 1)/\lambda] = 1 \Longrightarrow -1/\lambda \sum_{k=1}^{K} (N_k + \alpha_k - 1) = 1$

$\Longrightarrow -1/\lambda [\sum_k N_k + \sum_k \alpha_k - K] = 1 \Longrightarrow -1/\lambda [N + \sum_{k=1}^{K} \alpha_k - K] = 1$

$\Longrightarrow \lambda = -(N + \sum_{k=1}^{K} \alpha_k - K)$

- MAP Estimator :

: $\theta_k^{(MAP)} = -(N_k + \alpha_k - 1) / [-(N + \sum_{k=1}^{K} \alpha_k - K)]$

$= (N_k + \alpha_k - 1) / (N + \sum_{k=1}^{K} \alpha_k - K)$