Latent Dirichlet Allocation

Dan Boley

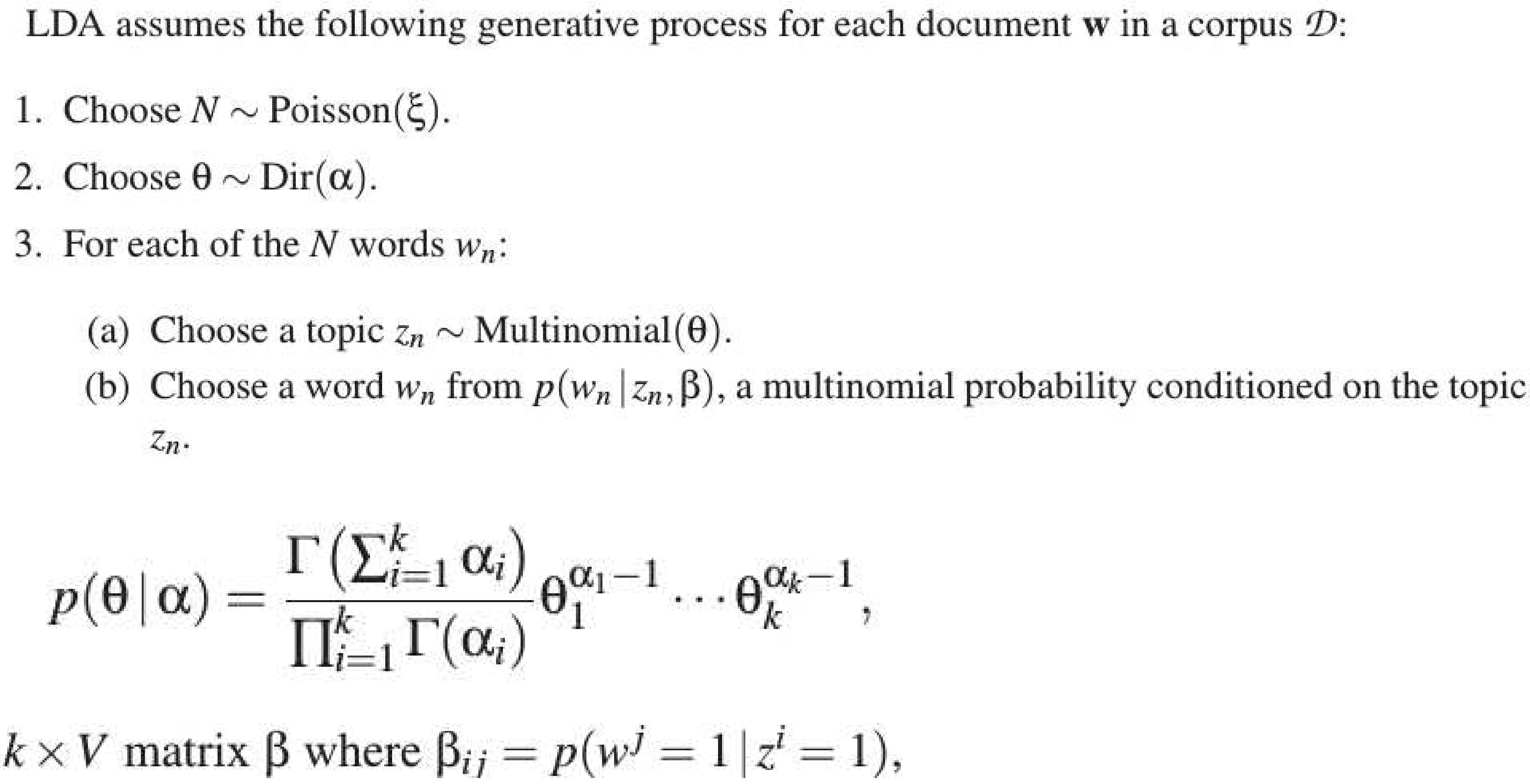
pLSI

- Each document is a mixture of topics.
- Each topic is a mixture of words.
- Must learn each document mixture individually.

LDA

- More general than pLSI
- Each document is represented as a mixture of topics
- Each document is generated from a process i.i.d.
- Each word is generated i.i.d. depending on the choice of topic and choice of word within each topic.

LDA Story



LDA assumes the following generative process for each document $\mathbf{w}$ in a corpus $\mathcal{D}$:

1. Choose $N \sim \text{Poisson}(\xi)$.
2. Choose $\theta \sim \text{Dir}(\alpha)$.
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
   (b) Choose a word w_n from $p(w_n \,|\, z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

$$p(\theta \,|\, \alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1},$$

$k \times V$ matrix β where $\beta_{ij} = p(w^j = 1 \,|\, z^i = 1)$,

Latent Dirichlet Alloc. - Plate Diagram

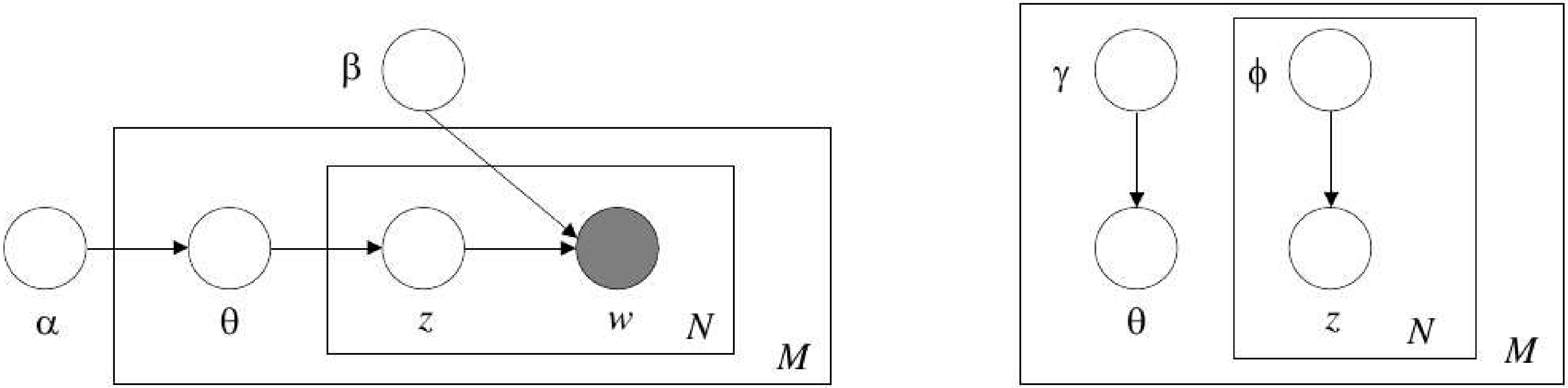


Figure 5: (Left) Graphical model representation of LDA. (Right) Graphical model representation of the variational distribution used to approximate the posterior in LDA.

$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta),$$

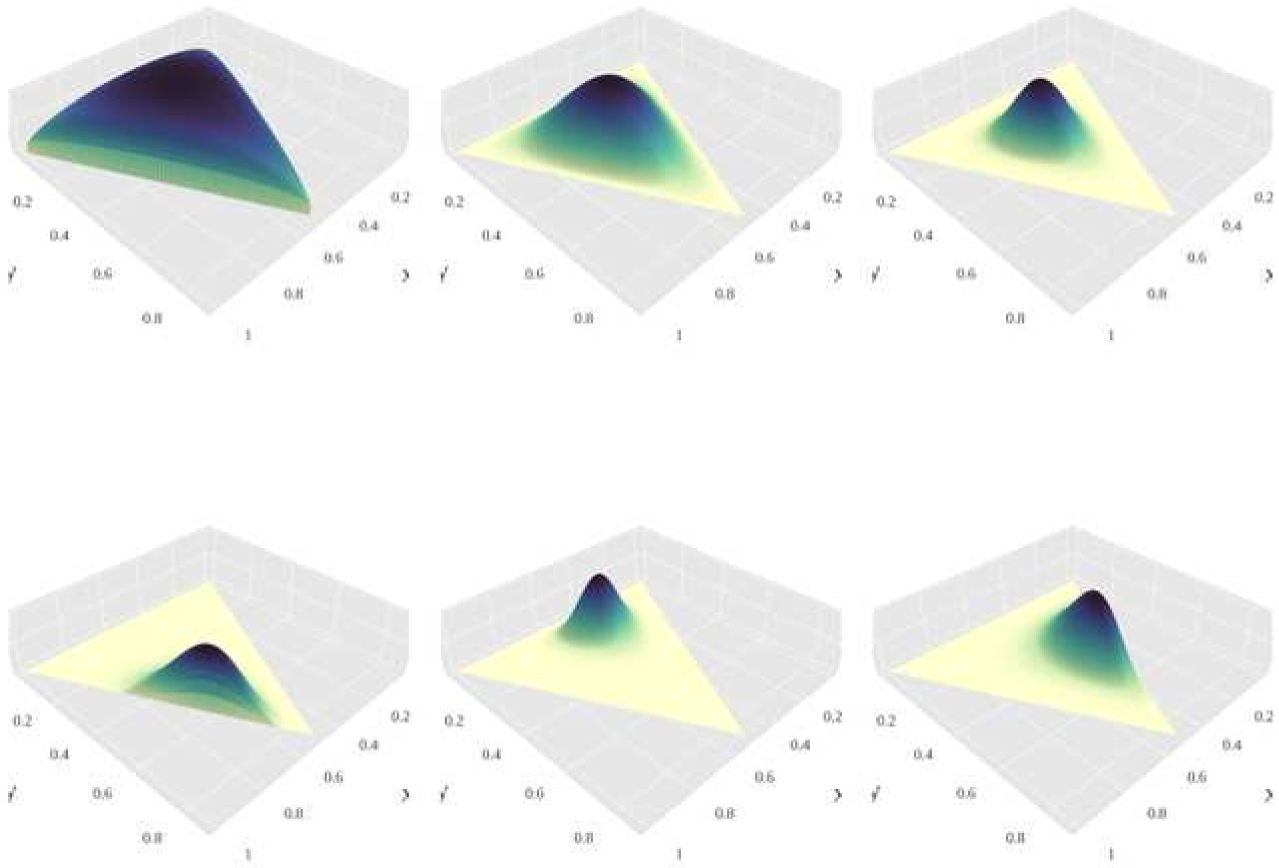
Dirichlet Distribution

- Generalization of Beta Distribution, when there are more than 2 parameters.
- Dir($\boldsymbol{\alpha}$) for $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K) > 0$:

  $$\rho(x_1, \ldots, x_K) = \frac{1}{B(\boldsymbol{\alpha})}\, x_1^{\alpha_1 - 1} \cdot x_2^{\alpha_2 - 1} \cdots x_K^{\alpha_K - 1}$$

  where $0 < x_i < 1$, $x_1 + \cdots x_K = 1$.
- $B(\boldsymbol{\alpha}) = \dfrac{\prod \Gamma(\alpha_i)}{\Gamma(\sum \alpha_i)}$.
- Gives variety of distributions over simplex $0 < x_i < 1$, $x_1 + \cdots x_K = 1$.

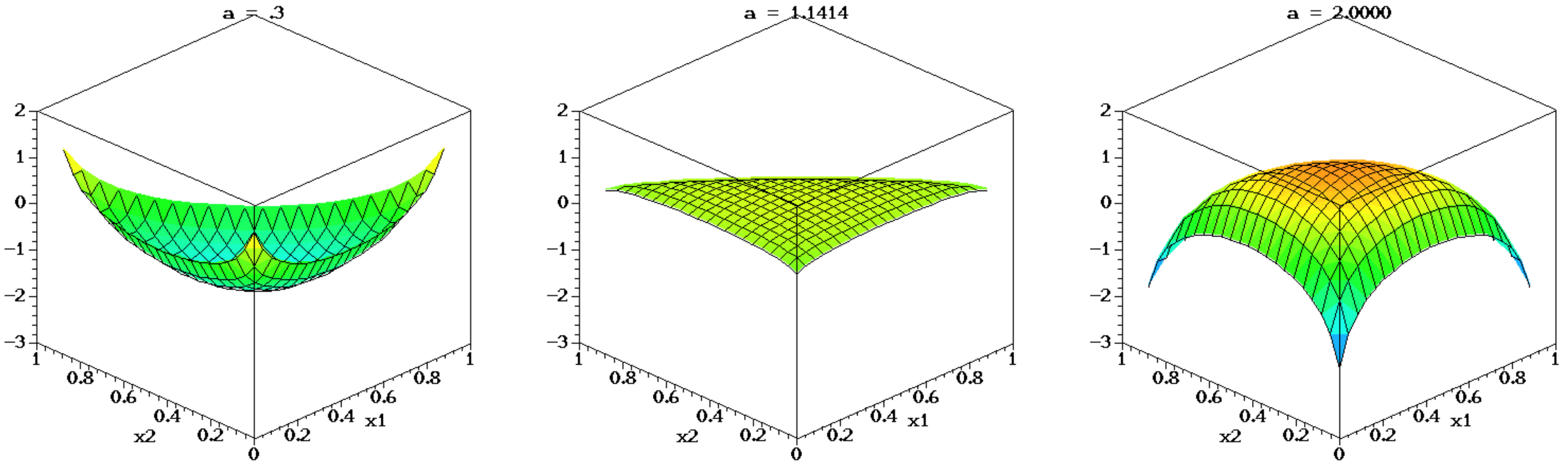
Dirichlet Distribution



probability density functions of a few Dirichlet distributions over a 2-simplex, for the following alpha vectors (clockwise, starting from the upper left corner): (1.3, 1.3, 1.3), (3,3,3), (7,7,7), (2,6,11), (14, 9, 5), (6,2,6).

`https://upload.wikimedia.org/wikipedia/commons/2/2b/Dirichlet-3d`

Dirichlet Distribution



Show how the log of the density function changes when K = 3 as we change the vector $\boldsymbol{\alpha}$ from (0.3, 0.3, 0.3) to (2.0, 2.0, 2.0), keeping all the individual α_i's equal.

`https://upload.wikimedia.org/wikipedia/commons/thumb/5/54/LogDirichletDensity-alpha_0.3_to_alpha_2.0.gif/250px-LogDirichletDensity-alpha_0.3_to_alpha_2.0.gif`

Multinomial Distribution

- Carry out n $i.i.d$ trials, each one with k possible outcomes.

  Record counts: how many times each outcome occurs, ignoring the order.

- Generalization of binomial distribution $(n > 1, k = 1)$

- Example: n coin tosses: $(k = 2)$. 6 sided die $(k = 6)$.

- 1 sample from this distribution $\mathbf{y} = (y_1, \ldots, y_k)$

  where $y_i \in \{1, \ldots, n\}$, $y_1 + \cdots + y_k = n$.

- Multi$(\mathbf{p})$: Probability of a specific observation $\mathbf{y} = (y_1, \ldots, y_k)$

  parametrized by $\mathbf{p} = (p_1, \ldots, p_k)$, $0 \le p_i \le 1$, $\sum p_i = 1$:

  $$P(\mathbf{y}) = \frac{n!}{y_1! \cdots y_k!} \, p_1^{y_1} \cdot p_2^{y_2} \cdots p_k^{y_k}$$

Conjugate

https://en.wikipedia.org/wiki/Conjugate_prior

The form of the conjugate prior can generally be determined by inspection of the probability density or probability mass function of a distribution. For example, consider a random variable which consists of the number of successes s in n Bernoulli trials with unknown probability of success q in [0,1]. This random variable will follow the binomial distribution, with a probability mass function of the form

$$P(s) = \binom{n}{s} q^s (1-q)^{n-s}$$

The usual conjugate prior is the beta distribution with parameters (α, β):

$$\rho(q) = \frac{q^{\alpha-1}(1-q)^{\beta-1}}{B(\alpha, \beta)}$$

where α and β are chosen to reflect any existing belief or information ($\alpha = 1$ and $\beta = 1$ would give a uniform distribution) and $B(\alpha, \beta)$ is the Beta function acting as a normalising constant.

Hyperparameters

In this context, α and β are called hyperparameters (parameters of the prior), to distinguish them from parameters of the underlying model (here q). It is a typical characteristic of conjugate priors that the dimensionality of the hyperparameters is one greater than that of the parameters of the original distribution. If all parameters are scalar values, then this means that there will be one more hyperparameter than parameter; but this also applies to vector-valued and matrix-valued parameters. (See the general article on the exponential family, and consider also the Wishart distribution, conjugate prior of the covariance matrix of a multivariate normal distribution, for an example where a large dimensionality is involved.)

Bayes Theorem

$$p(a|b)p(b) = p(a\&b) = p(b|a)p(a)$$

$$p(a|b) = \frac{p(b|a)p(a)}{p(b)} =$$

$$[\text{Posterior probability of param } a \ \text{ after observing } b \ \ ] =$$
$$= \frac{[\text{Prob observation } b \ \text{ given param } a \ \ ] \cdot [\text{Prior prob of } a \ \ ]}{\text{normalized over all possible values of } b}$$

Update using observation

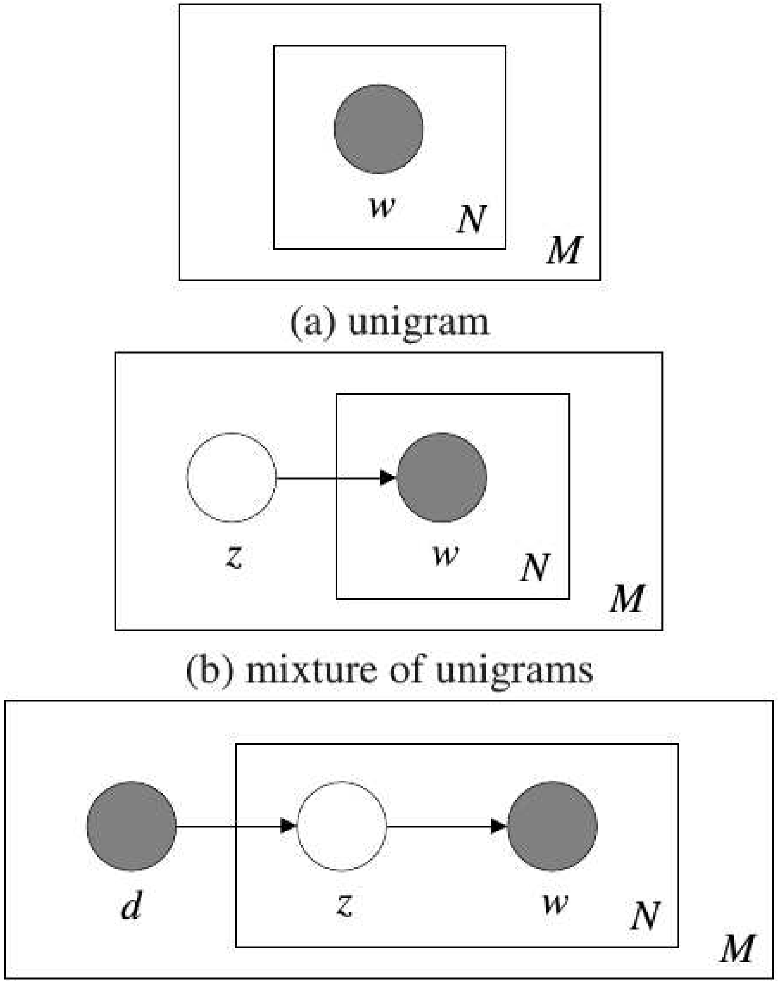
If we then sample this random variable and get s successes and f failures, we have

$$P(s, f \mid q = x) = \binom{s+f}{s} x^s (1-x)^f, \qquad \rho(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)},$$

$$\begin{aligned}
\rho(q = x \mid s, f) &= \frac{P(s, f \mid x)P(x)}{\int P(s, f \mid x)P(x)dx} \\
&= \frac{\binom{s+f}{s} x^{s+\alpha-1}(1-x)^{f+\beta-1}/B(\alpha, \beta)}{\int_{y=0}^{1} \left(\binom{s+f}{s} y^{s+\alpha-1}(1-y)^{f+\beta-1}/B(\alpha, \beta)\right) dy} \\
&= \frac{x^{s+\alpha-1}(1-x)^{f+\beta-1}}{B(s+\alpha, f+\beta)},
\end{aligned}$$

which is another Beta distribution with parameters $(\alpha + s, \beta + f)$. This posterior distribution could then be used as the prior for more samples, with the hyperparameters simply adding each extra piece of information as it comes.

Simpler Models



(a) unigram

(b) mixture of unigrams

(c) pLSI/aspect model