# Latent Dirichlet Allocation

Handout Notes - Generative and Deep Learning (GDL)

Davide Bacciu - University of Pisa

---

**Notation.** A corpus is a dataset $\mathcal{D} = \{\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(N)}\}$ of $N$ documents. The vocabulary size is $V$, and the number of topics is $K$. For document $d$, $L_d$ denotes its length in tokens, while $n_{dv}$ denotes the count of vocabulary item $v \in \{1, \dots, V\}$ in bag-of-words form. Topic proportions for document $d$ are denoted by $\boldsymbol{\theta}_d \in \Delta^{K-1}$, topic assignments by $Z_{dn} \in \{1, \dots, K\}$, and observed words by $W_{dn} \in \{1, \dots, V\}$. The topic-word matrix is $\beta \in \mathbb{R}^{K \times V}$ with entries $\beta_{kv} = p(W = v \mid Z = k)$ and $\sum_{v=1}^{V} \beta_{kv} = 1$. The Dirichlet hyperparameter is $\boldsymbol{\alpha} \in \mathbb{R}_+^K$. For variational inference, document-specific variational parameters are $\boldsymbol{\gamma}_d \in \mathbb{R}_+^K$ and $\boldsymbol{\phi}_{dn} \in \Delta^{K-1}$ (or $\boldsymbol{\phi}_{dv}$ in count-based notation). We use $p(\cdot)$ consistently for probability mass functions/densities.

## 1 Why topic models?

High-dimensional discrete data often exhibit strong latent structure. Text is the canonical example: each document is made of words from a large vocabulary, yet only a small subset is present in any given document. A bag-of-words representation is therefore high-dimensional and sparse.

Topic models provide a compressed latent representation of such data. Instead of representing a document directly in the $V$-dimensional word space, one represents it through a smaller set of latent topics. A topic is not a label assigned externally; it is a probability distribution over words. A document then becomes a mixture of such topics.

This perspective is attractive for at least three reasons. First, it yields a lower-dimensional semantic representation of documents. Second, it is probabilistic: a document can partially belong to several topics rather than being forced into a single cluster. Third, the same modeling idea extends beyond text to any collection of multinomial observations, including image "visual words", graph motifs, and other count-based representations.

## 2 From bag-of-words to latent topics

A bag-of-words representation ignores word order and keeps only counts. For each document $d$, we may write either the token sequence

$$\mathbf{w}^{(d)} = (w_{d1}, \dots, w_{dL_d})$$

or the equivalent count vector

$$\mathbf{n}_d = (n_{d1}, \dots, n_{dV}), \qquad \sum_{v=1}^{V} n_{dv} = L_d.$$

At the corpus level, these counts form a document-term matrix.

A classical mixture model would assign one latent cluster to each whole document. Topic models are more flexible: they assign a latent topic to each token, so a single document may combine multiple themes. This is the essential conceptual move from *single-topic membership* to *mixed-topic membership*.

## 3 Latent Dirichlet Allocation as a Bayesian topic model

Latent Dirichlet Allocation (LDA) is one of the simplest and most influential Bayesian latent-variable models. It represents each document by a random topic-mixture vector $\boldsymbol{\theta}_d$, and each token in the document is generated by first choosing a topic and then choosing a word from that topic.

The model combines three distributions:

- a **Dirichlet prior** over document-specific topic proportions,
- a **categorical / multinomial distribution** over topics for each token,
- a **categorical / multinomial distribution** over words for each topic.

The crucial Bayesian ingredient is that the document-specific topic proportions are themselves random variables:

$$\boldsymbol{\theta}_d \sim \text{Dirichlet}(\boldsymbol{\alpha}).$$

Thus probabilities are no longer fixed unknowns only; they are drawn from another distribution.

## 4 The Dirichlet distribution and why it appears here

The Dirichlet distribution is the natural prior over vectors that lie on the simplex:

$$\Delta^{K-1} = \left\{ \boldsymbol{\theta} \in \mathbb{R}_+^K : \sum_{k=1}^{K} \theta_k = 1 \right\}.$$

Its density is

$$p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}, \qquad \boldsymbol{\theta} \in \Delta^{K-1}.$$

The parameters $\alpha_k$ act like prior pseudo-counts for the different topics.

The Dirichlet is especially convenient because it is conjugate to the multinomial / categorical likelihood. This means that if topic assignments were observed, the posterior over $\boldsymbol{\theta}_d$ would remain Dirichlet. That conjugacy is one reason LDA is mathematically elegant and computationally tractable enough to approximate efficiently.

**Worked example — Dirichlet intuition and sparsity**

Suppose $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ is drawn from a symmetric Dirichlet with $\alpha_k = \alpha$ for all $k$. Then:

- if $\alpha \gg 1$, the mass concentrates near uniform topic mixtures, so documents tend to use many topics with similar proportions;
- if $\alpha \approx 1$, topic mixtures are fairly diffuse but variable;
- if $\alpha \ll 1$, the mass concentrates near the corners of the simplex, so documents tend to be sparse mixtures dominated by a few topics.

Thus $\boldsymbol{\alpha}$ controls not only the mean shape of topic proportions, but also how concentrated or sparse those proportions tend to be.

## 5 The LDA generative process

For each topic $k \in \{1, \ldots, K\}$, LDA assumes a topic-word distribution

$$\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kV}), \qquad \sum_{v=1}^{V} \beta_{kv} = 1.$$

For each document $d$, the generative process is:

1. draw topic proportions

$$\boldsymbol{\theta}_d \sim \text{Dirichlet}(\boldsymbol{\alpha});$$

2. for each token position $n = 1, \ldots, L_d$:

   (a) draw a topic assignment

$$Z_{dn} \mid \boldsymbol{\theta}_d \sim \text{Categorical}(\boldsymbol{\theta}_d);$$

   (b) draw a word

$$W_{dn} \mid Z_{dn} = k, \beta \sim \text{Categorical}(\boldsymbol{\beta}_k).$$

Equivalently, the joint distribution for one document is

$$p(\boldsymbol{\theta}_d, \mathbf{z}_d, \mathbf{w}_d \mid \boldsymbol{\alpha}, \beta) = p(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}) \prod_{n=1}^{L_d} p(z_{dn} \mid \boldsymbol{\theta}_d)\, p(w_{dn} \mid z_{dn}, \beta).$$

For the whole corpus, assuming documents are conditionally independent given the global parameters,

$$p(\mathcal{D} \mid \boldsymbol{\alpha}, \beta) = \prod_{d=1}^{N} p(\mathbf{w}_d \mid \boldsymbol{\alpha}, \beta).$$

# 6 What is learned in LDA?

The main global parameters are the topic-word distributions $\beta$. Depending on the setup, the Dirichlet hyperparameter $\boldsymbol{\alpha}$ may either be fixed by the practitioner or optimized from data.

The latent variables are document-specific:

- $\boldsymbol{\theta}_d$: the topic proportions of document $d$,
- $\mathbf{z}_d = (z_{d1}, \ldots, z_{dL_d})$: the token-level topic assignments.

Hence, learning in LDA has the typical structure of latent-variable learning:

> estimate global parameters while simultaneously inferring latent document-level variables.

# 7 Why exact inference is intractable

For a single document, the marginal likelihood is

$$p(\mathbf{w}_d \mid \boldsymbol{\alpha}, \beta) = \int p(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}) \prod_{n=1}^{L_d} \sum_{z_{dn}=1}^{K} p(z_{dn} \mid \boldsymbol{\theta}_d)\, p(w_{dn} \mid z_{dn}, \beta)\, d\boldsymbol{\theta}_d.$$

The corresponding posterior over latent variables is

$$p(\boldsymbol{\theta}_d, \mathbf{z}_d \mid \mathbf{w}_d, \boldsymbol{\alpha}, \beta) = \frac{p(\boldsymbol{\theta}_d, \mathbf{z}_d, \mathbf{w}_d \mid \boldsymbol{\alpha}, \beta)}{p(\mathbf{w}_d \mid \boldsymbol{\alpha}, \beta)}.$$

The difficulty is the denominator. Because the topic proportions $\boldsymbol{\theta}_d$ and the topic assignments $\mathbf{z}_d$ are coupled through the marginalization, the posterior cannot be simplified into a tractable closed form. This is exactly the setting where variational inference becomes useful.

## 8 Mean-field variational approximation for LDA

The central idea is to replace the exact posterior by a tractable approximation. For each document $d$, introduce a variational distribution of the form

$$q(\boldsymbol{\theta}_d, \mathbf{z}_d \mid \boldsymbol{\gamma}_d, \boldsymbol{\phi}_d) = q(\boldsymbol{\theta}_d \mid \boldsymbol{\gamma}_d) \prod_{n=1}^{L_d} q(z_{dn} \mid \boldsymbol{\phi}_{dn}),$$

where

$$q(\boldsymbol{\theta}_d \mid \boldsymbol{\gamma}_d) = \text{Dirichlet}(\boldsymbol{\gamma}_d), \qquad q(z_{dn} \mid \boldsymbol{\phi}_{dn}) = \text{Categorical}(\boldsymbol{\phi}_{dn}).$$

This is a *mean-field* factorization: we deliberately break posterior dependencies in order to obtain a tractable family.

The variational parameters have intuitive meanings:

- $\boldsymbol{\gamma}_d$ approximates the posterior Dirichlet parameters of the document-topic proportions;
- $\boldsymbol{\phi}_{dn}$ approximates the posterior topic assignment probabilities for token $n$ in document $d$.

## 9 Variational EM objective

Variational learning in LDA maximizes the evidence lower bound (ELBO). For one document,

$$\mathcal{L}(\mathbf{w}_d, \boldsymbol{\gamma}_d, \boldsymbol{\phi}_d, \beta) = \mathbb{E}_q[\log p(\boldsymbol{\theta}_d, \mathbf{z}_d, \mathbf{w}_d \mid \boldsymbol{\alpha}, \beta)] - \mathbb{E}_q[\log q(\boldsymbol{\theta}_d, \mathbf{z}_d)]\,.$$

Summing over documents gives the corpus ELBO.

Variational EM alternates:

- **E-step:** fix $\beta$ (and usually $\boldsymbol{\alpha}$), then optimize document-specific variational parameters $(\boldsymbol{\gamma}_d, \boldsymbol{\phi}_d)$;
- **M-step:** fix the variational parameters and update the global parameters $\beta$ (and optionally $\boldsymbol{\alpha}$).

## 10 Document-level E-step: explicit variational updates

The mean-field factorization yields simple fixed-point updates. A key quantity is the expectation of the log topic proportion under a Dirichlet:

$$\mathbb{E}_q[\log \theta_{dk}] = \psi(\gamma_{dk}) - \psi\left(\sum_{j=1}^{K} \gamma_{dj}\right),$$

where $\psi(\cdot)$ is the digamma function, i.e. the derivative of $\log \Gamma(\cdot)$.

**Worked example — Variational E-step updates for one document**

For each token position $n$ and topic $k$, the topic responsibility is updated as

$$\phi_{dnk} \propto \beta_{k,w_{dn}} \exp\left(\psi(\gamma_{dk}) - \psi\left(\sum_{j=1}^{K} \gamma_{dj}\right)\right).$$

After computing these unnormalized values, normalize over $k$ so that

$$\sum_{k=1}^{K} \phi_{dnk} = 1.$$

Then update the Dirichlet parameters:

$$\gamma_{dk} = \alpha_k + \sum_{n=1}^{L_d} \phi_{dnk}.$$

Hence each $\gamma_{dk}$ is the prior count $\alpha_k$ plus the expected number of tokens in document $d$ assigned to topic $k$.
If bag-of-words counts are used rather than individual tokens, the same update becomes

$$\phi_{dvk} \propto \beta_{kv} \exp\left(\psi(\gamma_{dk}) - \psi\left(\sum_{j=1}^{K} \gamma_{dj}\right)\right),$$

for each vocabulary item $v$ present in the document, followed by

$$\gamma_{dk} = \alpha_k + \sum_{v=1}^{V} n_{dv}\, \phi_{dvk}.$$

These updates are iterated until document-level variational convergence.

These equations have a clear interpretation. The factor $\beta_{kv}$ favors topics that assign high probability to the observed word $v$. The exponential term favors topics that are currently prominent in the document according to $\boldsymbol{\gamma}_d$. The update for $\boldsymbol{\phi}$ therefore balances *word-level evidence* and *document-level topic proportions*.

## 11 Global M-step: updating the topics

Once variational responsibilities are available for all documents, the M-step updates the topic-word distributions. The sufficient statistic is the expected topic-word count:

$$\widehat{n}_{kv} = \sum_{d=1}^{N} n_{dv}\, \phi_{dvk}.$$

This is the expected number of times vocabulary word $v$ is assigned to topic $k$ across the corpus.

**Worked example — M-step update for the topic-word distributions**

The topic-word vector $\boldsymbol{\beta}_k$ must satisfy

$$\beta_{kv} \geq 0, \qquad \sum_{v=1}^{V} \beta_{kv} = 1.$$

The ELBO terms involving $\beta$ reduce to

$$\sum_{k=1}^{K} \sum_{v=1}^{V} \widehat{n}_{kv} \log \beta_{kv}.$$

For each topic $k$, maximize

$$\sum_{v=1}^{V} \widehat{n}_{kv} \log \beta_{kv} \quad \text{subject to} \quad \sum_{v=1}^{V} \beta_{kv} = 1.$$

Using a Lagrange multiplier $\lambda_k$,

$$\mathcal{J}_k(\boldsymbol{\beta}_k, \lambda_k) = \sum_{v=1}^{V} \widehat{n}_{kv} \log \beta_{kv} + \lambda_k \left( \sum_{v=1}^{V} \beta_{kv} - 1 \right).$$

Set derivatives to zero:

$$\frac{\partial \mathcal{J}_k}{\partial \beta_{kv}} = \frac{\widehat{n}_{kv}}{\beta_{kv}} + \lambda_k = 0 \quad \Longrightarrow \quad \beta_{kv} = -\frac{\widehat{n}_{kv}}{\lambda_k}.$$

Enforce normalization:

$$1 = \sum_{v=1}^{V} \beta_{kv} = -\frac{1}{\lambda_k} \sum_{v=1}^{V} \widehat{n}_{kv} \quad \Longrightarrow \quad \lambda_k = -\sum_{v=1}^{V} \widehat{n}_{kv}.$$

Therefore,

$$\boxed{\beta_{kv} = \frac{\widehat{n}_{kv}}{\sum_{v'=1}^{V} \widehat{n}_{kv'}}.}$$

Thus each topic is updated by normalizing expected topic-word counts.

The hyperparameter $\boldsymbol{\alpha}$ can also be optimized. However, unlike the update for $\beta$, its optimization is not available as a simple closed-form normalization and is often carried out with Newton–Raphson or related methods.

## 12 Variational EM algorithm for LDA

Putting the pieces together, LDA learning alternates document-level variational inference with corpus-level topic re-estimation.

**Worked example — Variational EM pseudocode for LDA**

**Input:** corpus counts $\{n_{dv}\}$, number of topics $K$, initial $\boldsymbol{\alpha}$, initial $\beta$, tolerance $\varepsilon$, max iterations $T_{\max}$.
**Output:** learned topic-word matrix $\beta$ (and optionally updated $\boldsymbol{\alpha}$), together with document-level variational parameters $\{\boldsymbol{\gamma}_d, \boldsymbol{\phi}_d\}$.

**Initialize:**
Choose initial $\beta$ such that each row $\boldsymbol{\beta}_k$ lies on the simplex.
Set outer iteration counter $t \leftarrow 0$.

**Repeat until ELBO / likelihood convergence or $t = T_{\max} - 1$:**

**E-step: per-document variational inference.**
For each document $d = 1, \ldots, N$:

1. Initialize
$$\gamma_{dk} \leftarrow \alpha_k + \frac{L_d}{K} \quad \text{for } k = 1, \ldots, K,$$
and initialize each $\boldsymbol{\phi}_{dv}$ uniformly over topics.
2. Repeat until document-level convergence:
   (a) For each vocabulary item $v$ with $n_{dv} > 0$, update
$$\phi_{dvk} \propto \beta_{kv} \exp\left(\psi(\gamma_{dk}) - \psi\left(\sum_{j=1}^{K} \gamma_{dj}\right)\right), \qquad k = 1, \ldots, K,$$
   then normalize over $k$.
   (b) Update
$$\gamma_{dk} \leftarrow \alpha_k + \sum_{v=1}^{V} n_{dv}\phi_{dvk}, \qquad k = 1, \ldots, K.$$

**M-step: global topic update.**
For each topic $k$ and vocabulary item $v$, compute expected counts

$$\widehat{n}_{kv} \leftarrow \sum_{d=1}^{N} n_{dv}\phi_{dvk}.$$

Then update

$$\beta_{kv} \leftarrow \frac{\widehat{n}_{kv}}{\sum_{v'=1}^{V} \widehat{n}_{kv'}}.$$

**Optional hyperparameter update.**
Update $\boldsymbol{\alpha}$ using Newton–Raphson or another numerical procedure if desired.

**Convergence check.**
Evaluate the corpus ELBO (or a surrogate likelihood criterion). If improvement is below $\varepsilon$, stop. Otherwise increment $t$ and continue.

**Return** $\beta$, and optionally $\boldsymbol{\alpha}$, together with $\{\boldsymbol{\gamma}_d, \boldsymbol{\phi}_d\}$.