Latent Dirichlet Allocation - Derivations & Proofs

Paul F. Roysdon, Ph.D.

Contents

1 Mathematical Derivations & Proofs

1.1 Introduction

Latent Dirichlet Allocation (LDA) is a generative probabilistic model of discrete data (e.g., text corpora), in which each document is modeled as a mixture over K latent topics, and each topic is a distribution over a vocabulary of size V. The key Bayesian ingredients are Dirichlet priors on (i) per-document topic proportions and (ii) per-topic word distributions, yielding conjugacy with categorical/multinomial likelihoods. We derive the full generative model, show Dirichlet–multinomial conjugacy, develop two standard inference/learning procedures—*mean-field variational Bayes* and *collapsed Gibbs sampling*—and provide complete algorithmic steps, proofs, and dimensions.

1.2 Data and Notation

Let the corpus consist of D documents. Document d has N_d tokens:

$$\mathbf{w}_d = (w_{d1}, \ldots, w_{dN_d}), \qquad w_{dn} \in \{1, \ldots, V\}.$$

LDA posits K topics. For each token there is a latent *topic assignment*

$$\mathbf{z}_d = (z_{d1}, \ldots, z_{dN_d}), \qquad z_{dn} \in \{1, \ldots, K\}.$$

Per-document topic proportions $\boldsymbol{\theta}_d \in \Delta^{K-1}$, per-topic word distributions $\boldsymbol{\beta}_k \in \Delta^{V-1}$. Stack topics as a matrix

$$\mathbf{B} \;=\; \big[\boldsymbol{\beta}_1 \;\cdots\; \boldsymbol{\beta}_K\big] \in [0,1]^{V\times K}, \quad \sum_{v=1}^{V} \beta_{kv} = 1 \;\forall k.$$

Priors:

$$\boldsymbol{\alpha} \in \mathbb{R}^K_{>0} \quad \text{(Dirichlet on } \boldsymbol{\theta}_d\text{)}, \qquad \boldsymbol{\eta} \in \mathbb{R}^V_{>0} \quad \text{(Dirichlet on each } \boldsymbol{\beta}_k\text{)}.$$

Dimensions: D (docs), K (topics), V (vocab), N_d (# tokens in d), $N = \sum_d N_d$ (# tokens).

1.3 Model Formulation (Generative Process)

For each topic $k = 1, \dots, K$:

$$\boldsymbol{\beta}_k \sim \text{Dirichlet}(\boldsymbol{\eta}).$$

For each document $d = 1, \dots, D$:

$$\boldsymbol{\theta}_d \sim \text{Dirichlet}(\boldsymbol{\alpha}), \qquad z_{dn} \mid \boldsymbol{\theta}_d \sim \text{Categorical}(\boldsymbol{\theta}_d), \quad w_{dn} \mid z_{dn}{=}k, \mathbf{B} \sim \text{Categorical}(\boldsymbol{\beta}_k), \quad n = 1, \dots, N_d.$$

Joint density (probability). With $\alpha_0 = \sum_k \alpha_k$, $\eta_0 = \sum_v \eta_v$, the complete joint density (probability) is

$$p(\mathbf{w}, \mathbf{z}, \{\boldsymbol{\theta}_d\}, \{\boldsymbol{\beta}_k\} \mid \boldsymbol{\alpha}, \boldsymbol{\eta}) = \Big[\prod_{k=1}^{K} \text{Dir}(\boldsymbol{\beta}_k \mid \boldsymbol{\eta}) \Big] \Big[\prod_{d=1}^{D} \text{Dir}(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}) \prod_{n=1}^{N_d} \theta_{d,z_{dn}} \, \beta_{z_{dn},w_{dn}} \Big], \tag{1}$$

where $\text{Dir}(\cdot)$ is the Dirichlet function.

1.4 Conjugacy and Collapsed Marginals

Dirichlet–multinomial integral (proof). For counts $\mathbf{n} = (n_1, \dots, n_K)$ with $\sum_k n_k = N$,

$$\int_{\Delta^{K-1}} \Big(\prod_{k=1}^{K} \theta_k^{n_k} \Big) \, \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \, d\boldsymbol{\theta} = \frac{B(\boldsymbol{\alpha} + \mathbf{n})}{B(\boldsymbol{\alpha})} = \frac{\Gamma(\alpha_0)}{\prod_k \Gamma(\alpha_k)} \cdot \frac{\prod_k \Gamma(\alpha_k + n_k)}{\Gamma(\alpha_0 + N)}.$$

This is the Dirichlet–multinomial (a.k.a. Polya) compound; $B(\cdot)$ is the multivariate Beta. ■

Collapsed joint $p(\mathbf{z}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\eta})$. Let n_{dk} be the # of tokens in document d assigned to topic k; n_{kv} the # of tokens using word v assigned to topic k; $n_k = \sum_v n_{kv}$. Integrating Eqn. (1) over $\{\boldsymbol{\theta}_d\}$ and $\{\boldsymbol{\beta}_k\}$ yields

$$p(\mathbf{z}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\eta}) = \prod_{d=1}^{D} \frac{\Gamma(\alpha_0)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \cdot \frac{\prod_{k=1}^{K} \Gamma(\alpha_k + n_{dk})}{\Gamma(\alpha_0 + N_d)} \;\times\; \prod_{k=1}^{K} \frac{\Gamma(\eta_0)}{\prod_{v=1}^{V} \Gamma(\eta_v)} \cdot \frac{\prod_{v=1}^{V} \Gamma(\eta_v + n_{kv})}{\Gamma(\eta_0 + n_k)}. \tag{2}$$

1.5 Inference & Learning I: Mean-Field Variational Bayes

We approximate the posterior with a factorized family

$$q(\{\boldsymbol{\beta}_k\}, \{\boldsymbol{\theta}_d\}, \mathbf{z}) = \Big[\prod_{k=1}^{K} \text{Dir}(\boldsymbol{\beta}_k \mid \boldsymbol{\lambda}_k) \Big] \Big[\prod_{d=1}^{D} \text{Dir}(\boldsymbol{\theta}_d \mid \boldsymbol{\gamma}_d) \prod_{n=1}^{N_d} \text{Categorical}(z_{dn} \mid \boldsymbol{\phi}_{dn}) \Big],$$

with variational parameters $\boldsymbol{\lambda}_k \in \mathbb{R}^V_{>0}$, $\boldsymbol{\gamma}_d \in \mathbb{R}^K_{>0}$, and $\boldsymbol{\phi}_{dn} \in \Delta^{K-1}$.

Evidence lower bound (ELBO).

$$\mathcal{L}(q) = \mathbb{E}_q[\log p(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\beta} \mid \boldsymbol{\alpha}, \boldsymbol{\eta})] - \mathbb{E}_q[\log q(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\beta})]. \tag{3}$$

Coordinate ascent updates follow from setting derivatives of $\mathcal{L}$ to zero.

Expected logs. For digamma $\psi(\cdot)$,

$$\mathbb{E}[\log \theta_{dk}] = \psi(\gamma_{dk}) - \psi\big(\sum_{t=1}^{K} \gamma_{dt}\big), \qquad \mathbb{E}[\log \beta_{kv}] = \psi(\lambda_{kv}) - \psi\big(\sum_{u=1}^{V} \lambda_{ku}\big).$$

Coordinate ascent (CAVI) updates. For each document d and token n with word $v = w_{dn}$,

$$\phi_{dnk} \propto \exp\Big(\mathbb{E}[\log \theta_{dk}] + \mathbb{E}[\log \beta_{kv}]\Big) = \exp\Big(\psi(\gamma_{dk}) - \psi(\textstyle\sum_t \gamma_{dt}) + \psi(\lambda_{kv}) - \psi(\sum_u \lambda_{ku})\Big), \tag{4}$$

$$\gamma_{dk} = \alpha_k + \sum_{n=1}^{N_d} \phi_{dnk}, \tag{5}$$

$$\lambda_{kv} = \eta_v + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{dnk}\ \mathbf{1}\{w_{dn} = v\}. \tag{6}$$

Normalize $\boldsymbol{\phi}_{dn}$ after Eqn. (4). Iterate Eqns. (4)–(6) until ELBO convergence.

Variational EM (ML topics). If we treat $\mathbf{B}$ as *parameters* (not random) and maximize w.r.t. $\boldsymbol{\beta}_k$ under $\sum_v \beta_{kv} = 1$, then the M-step yields

$$\beta_{kv} \ \propto \ \sum_{d,n} \phi_{dnk}\, \mathbf{1}\{w_{dn} = v\} \qquad \text{(optionally smoothed: } +\eta_v - 1), \tag{7}$$

followed by normalization across v.

Proof (CAVI). *Proof.* Taking $\partial \mathcal{L}/\partial \phi_{dnk}$ with normalization constraint gives $\log \phi_{dnk} \propto \mathbb{E}[\log \theta_{dk}] + \mathbb{E}[\log \beta_{kv}]$. Dirichlet–multinomial conjugacy in the complete conditionals yields Eqn. (5) and Eqn. (6). ■

1.6 Inference & Learning II: Collapsed Gibbs Sampling

Integrate out $\{\boldsymbol{\theta}_d\}$ and $\{\boldsymbol{\beta}_k\}$ to sample topic assignments $\mathbf{z}$ from Eqn. (2). Let $v = w_{dn}$. Define "leave-one-out" counts that exclude token (d, n):

$$n_{dk}^{\neg dn}, \quad n_{kv}^{\neg dn}, \quad n_{k}^{\neg dn} = \sum_{v} n_{kv}^{\neg dn}.$$

Collapsed conditional (derivation). From Eqn. (2), the ratio of the joint with $z_{dn}{=}k$ versus not depends only on factors involving n_{dk} and n_{kv}. Using $\Gamma(x+1) = x\,\Gamma(x)$,

$$\begin{aligned} p(z_{dn}{=}k \mid \mathbf{z}^{\neg dn}, \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\eta}) &\propto \frac{\Gamma(\alpha_k + n_{dk}^{\neg dn} + 1)}{\Gamma(\alpha_k + n_{dk}^{\neg dn})} \cdot \frac{\Gamma(\eta_v + n_{kv}^{\neg dn} + 1)}{\Gamma(\eta_v + n_{kv}^{\neg dn})} \cdot \frac{\Gamma(\eta_0 + n_{k}^{\neg dn})}{\Gamma(\eta_0 + n_{k}^{\neg dn} + 1)} \\ &= (\alpha_k + n_{dk}^{\neg dn}) \cdot \frac{\eta_v + n_{kv}^{\neg dn}}{\eta_0 + n_{k}^{\neg dn}}. \end{aligned} \tag{8}$$

After normalization over k,

$$p(z_{dn}{=}k \mid \cdot) \ \propto \ (\alpha_k + n_{dk}^{\neg dn}) \cdot \frac{\eta_v + n_{kv}^{\neg dn}}{\eta_0 + n_{k}^{\neg dn}}. \tag{9}$$

Sampling scheme. Sweep all tokens, update counts in-place; after burn-in, thin/summarize $\{n_{kv}\}$ to estimate topics:

$$\widehat{\beta}_{kv} \ = \ \frac{\eta_v + n_{kv}}{\eta_0 + n_k}, \qquad \widehat{\theta}_{dk} \ = \ \frac{\alpha_k + n_{dk}}{\alpha_0 + N_d}.$$

1.7 Hyperparameter Estimation (Optional)

Updating $\boldsymbol{\alpha}$ (per-document Dirichlet). Maximize $\sum_d \log p(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha})$ under variational posterior by replacing $\log \theta_{dk}$ with $\mathbb{E}[\log \theta_{dk}]$:

$$\mathcal{J}(\boldsymbol{\alpha}) = D\Big(\log\Gamma(\alpha_0) - \sum_k \log\Gamma(\alpha_k)\Big) + \sum_d \sum_k (\alpha_k - 1)\Big(\psi(\gamma_{dk}) - \psi(\textstyle\sum_t \gamma_{dt})\Big).$$

Gradient and Hessian:

$$g_k = D\big(\psi(\alpha_0) - \psi(\alpha_k)\big) + \sum_d \Big(\psi(\gamma_{dk}) - \psi(\textstyle\sum_t \gamma_{dt})\Big), \quad H_{kl} = D\big(\psi'(\alpha_0) - \delta_{kl}\psi'(\alpha_k)\big).$$

Perform Newton updates in the positive orthant (e.g. Minkas method), with line search to ensure ascent.

Updating $\boldsymbol{\eta}$ (per-topic Dirichlet). Analogously, with variational $q(\boldsymbol{\beta}_k) = \mathrm{Dir}(\boldsymbol{\lambda}_k)$,

$$\mathcal{J}(\boldsymbol{\eta}) = K\Big(\log\Gamma(\eta_0) - \sum_v \log\Gamma(\eta_v)\Big) + \sum_k \sum_v (\eta_v - 1)\Big(\psi(\lambda_{kv}) - \psi(\textstyle\sum_u \lambda_{ku})\Big),$$

and compute Newton steps with gradient g_v and Hessian $H_{uv} = K\big(\psi'(\eta_0) - \delta_{uv}\psi'(\eta_v)\big)$.

1.8 Predictive Inference and Perplexity

Held-out document likelihood (VB). Given topics $\{\boldsymbol{\lambda}_k\}$, infer $\boldsymbol{\gamma}_d, \boldsymbol{\phi}_{dn}$ on held-out document d^* via Eqns. (4)–(5); approximate predictive with expected logs:

$$\log p(\mathbf{w}_{d^*} \mid \text{model}) \approx \sum_n \log \sum_k \exp\Big(\psi(\gamma_{d^*k}) - \psi(\textstyle\sum_t \gamma_{d^*t})\Big)\, \exp\Big(\psi(\lambda_{k,w_{d^*n}}) - \psi(\textstyle\sum_u \lambda_{ku})\Big).$$

Perplexity.

$$\text{Perplexity}(\mathcal{D}_{\text{test}}) = \exp\Big(-\frac{\sum_{d\in\text{test}} \log p(\mathbf{w}_d)}{\sum_{d\in\text{test}} N_d}\Big).$$

1.9 Algorithmic Summary

Mean-field Variational Bayes (VB) for LDA.

1. **Initialize** $\boldsymbol{\lambda}_k$ (e.g. η + random), for each d: $\boldsymbol{\gamma}_d \leftarrow \boldsymbol{\alpha} + N_d/K$, $\boldsymbol{\phi}_{dn} \leftarrow 1/K$.
2. **Repeat** until convergence:
   (a) For each document d: iterate Eqn. (4), Eqn. (5) over tokens n.
   (b) Global update: Eqn. (6) for all (k, v).
   (c) (Optional) Update hyperparameters $\boldsymbol{\alpha}, \boldsymbol{\eta}$ by Newton steps.
3. **Outputs:** topics via $\widehat{\beta}_{kv} \propto \lambda_{kv}$, proportions $\widehat{\theta}_{dk} \propto \gamma_{dk}$.

Collapsed Gibbs Sampling for LDA.

1. **Initialize** z_{dn} uniformly at random; compute counts n_{dk}, n_{kv}, n_k.
2. **For** sweeps $t = 1, \ldots, T$:
   (a) For each token (d, n), decrement counts, sample z_{dn} from Eqn. (9), increment counts.
3. **Estimate** topics and proportions from posterior means (averaged over samples): $\widehat{\beta}_{kv} = (\eta_v + n_{kv})/(\eta_0 + n_k)$, $\widehat{\theta}_{dk} = (\alpha_k + n_{dk})/(\alpha_0 + N_d)$.

1.10 Computational Aspects

VB per-iteration cost is $O\big(\sum_d N_d K + KV\big)$ due to updates of all $\boldsymbol{\phi}_{dn}$ and $\boldsymbol{\lambda}_k$. Collapsed Gibbs has cost $O(NK)$ per sweep (or less with sparse/alias-based sampling). For large V, top-S vocab pruning or sparse $\boldsymbol{\lambda}$ updates reduce memory/time.

1.11 Summary of Variables and Their Dimensions

- D (docs), K (topics), V (vocab size), N_d (# tokens in doc d), $N = \sum_d N_d$.
- $\mathbf{w}_d \in \{1, \dots, V\}^{N_d}$: tokens of doc d; w_{dn} is the nth token index.
- $\mathbf{z}_d \in \{1, \dots, K\}^{N_d}$: topic assignments; z_{dn} is the topic for w_{dn}.
- $\boldsymbol{\theta}_d \in \Delta^{K-1}$: topic mixture for doc d; prior $\mathrm{Dir}(\boldsymbol{\alpha})$.
- $\boldsymbol{\beta}_k \in \Delta^{V-1}$: word distribution for topic k; prior $\mathrm{Dir}(\boldsymbol{\eta})$.
- $\mathbf{B} = [\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K] \in [0,1]^{V \times K}$: topic matrix.
- Counts: n_{dk} (# tokens in d assigned to k), n_{kv} (# tokens of word v assigned to k), $n_k = \sum_v n_{kv}$.
- Variational params: $\boldsymbol{\phi}_{dn} \in \Delta^{K-1}$, $\boldsymbol{\gamma}_d \in \mathbb{R}^K_{>0}$, $\boldsymbol{\lambda}_k \in \mathbb{R}^V_{>0}$.
- Hyperparameters: $\boldsymbol{\alpha} \in \mathbb{R}^K_{>0}$, $\boldsymbol{\eta} \in \mathbb{R}^V_{>0}$; sums α_0, η_0.

1.12 Summary

From first principles, LDA defines a hierarchical Bayesian model with Dirichlet priors on per-document topic mixtures and per-topic word distributions, yielding conjugate Dirichlet–multinomial structure. Integrating out the Dirichlet variables gives the collapsed joint Eqn. (2), from which the Gibbs conditional Eqn. (9) follows. Alternatively, mean-field variational Bayes maximizes an ELBO with closed-form coordinate updates Eqns. (4)–(6) (and ML/smoothed updates Eqn. (7)). We have stated all distributions, objectives, updates, proofs of conjugacy, and algorithmic procedures, together with explicit variable dimensions, in a notation consistent with prior sections.