A principled approach to Expectation Maximisation and Latent Dirichlet Allocation using Jeffrey's update rule

Bart Jacobs

Institute for Computing and Information Sciences
Radboud University Nijmegen, NL
bart@cs.ru.nl

June 15, 2023

Abstract. Expectation Maximisation (EM) and Latent Dirichlet Allocation (LDA) are two frequently used inference algorithms, for finding an appropriate mixture of latent variables, and for finding an allocation of topics for a collection of documents. A recent insight in probabilistic learning is that Jeffrey's update rule gives a decrease of Kullback-Leibler divergence. Its logic is error correction. It is shown that this same rule and divergence decrease logic is at the heart of EM and LDA, ensuring that successive iterations are decreasingly wrong.

1 Introduction

Learning can happen via encouragement or via discouragement, that is by reinforcing what goes well, or by slowing down what is going the wrong way. Intuitively these differences are clear. In probabilistic learning one can distinguish rules of Pearl and Jeffrey for updating (conditioning, belief revision), see [22,17] and [3,19,8,12] for comparisons. In [14] the difference between these rules has been described mathematically: Pearl's rule gives an increase of validity (expected value), whereas Jeffrey's rule gives a decrease of (Kullback-Leibler) divergence. The latter may be understood as error correction (or reduction of free energy), like in predictive coding theory [23,9,11] — where the human mind is studied as a Bayesian prediction and correction engine.

This paper demonstrates the relevance of Jeffrey's update rule — with its divergence decrease — for two fundamental inference algorithms, namely Expectation Maximisation (EM) [7] and Latent Dirichlet Allocation (LDA) [2]. EM is used for uncovering mixtures of latent variables. It has many applications, for instance in natural language processing, computer vision, and genetics. LDA is used to get a big-picture of (large) collections of documents by discovering the topics that they cover. Both are unsupervised classification algorithms.

The paper gives an abstract reformulation of these two well-known algorithms in machine learning that brings out the logic of divergence reduction (or error correction) behind them. This reformulation is inspired by categorical probability theory (see *e.g.* [10,4]), in which conditional probabilities $p(y \mid x)$ are reinterpreted as probabilistic functions $X \to Y$, also known as Kleisli maps or channels, with a rich structure, among others for sequential composition, parallel composition, and reversal. This paper does not assume knowledge of category theory: the relevant constructions are described concretely, especially for reversal (Bayesian inversian, dagger) since it plays a crucial role in Jeffrey's rule.

Making explicit what these algorithms EM and LDA achieve, and how, is relevant in times with a rising need that algorithms in machine learning and AI explain their outcomes. A first requirement for such explanations is a clear semantical understanding, including the underlying logic. Thus, the aim here is to analyse (two) existing algorithms, in their basic forms. The algorithms are not extended or improved, but studied as such.

This paper is organised as follows. It first introduces notation and basic terminology for multisets and (discrete probability) distributions in Section 2, including channels (probabilistic functions) and their reversals. Section 3 recalls Jeffrey's update rule and the associated divergence decrease. It includes a (new) strengthened version of this rule, with multiple channels, that will be used for LDA. Subsequently, Section 4 describes the EM algorithm using channels and shows how its correctness can be proved in just four lines, see the proof of Theorem 3. Section 5 gives a similar reformulation and proof of correctness for LDA. Simple illustrations are included for Jeffrey's rule, EM and for LDA.

2 Multisets and distributions

A multiset (or bag) is a subset in which elements may occur multiple times. We borrow 'ket' notation $|\cdot\rangle$ from quantum theory and represent an urn with three red, two blue and one green ball as a multiset $3|R\rangle + 2|B\rangle + 1|G\rangle$. In the 'bag-of-words' model a document is understood as a multiset of words. A distribution is like a multiset but its multiplicities are not natural numbers but probabilities, from the unit interval $[0, 1]$, that add up to one, as in $\frac{1}{2}|R\rangle + \frac{1}{3}|B\rangle + \frac{1}{6}|G\rangle$.

More formally, a multiset on a set X is a function $\varphi\colon X \to \mathbb{N}$ with finite support $\text{supp}(\varphi) = \{x \in X \mid \varphi(x) > 0\}$. Similarly, a distribution on X is a function $\omega\colon X \to [0, 1]$ with finite support, with $\sum_x \omega(x) = 1$. We can equivalently write them in ket notation as $\varphi = \sum_x \varphi(x)|x\rangle$ and $\omega = \sum_x \omega(x)|x\rangle$. We write $\mathcal{M}(X)$ and $\mathcal{D}(X)$ for the sets of multisets and distributions on a set X.

Notice that we do not require that the set X is finite. But when X is finite, we can say that a multiset $\varphi \in \mathcal{M}(X)$, or a distribution $\omega \in \mathcal{D}(X)$, has full support if $\text{supp}(\varphi) = X$, or $\text{supp}(\omega) = X$. The unit multiset $\mathbf{1}_X \coloneqq \sum_x 1|x\rangle$ and the uniform distribution $\mathit{unif}_X \coloneqq \sum_x \frac{1}{n}|x\rangle$, for the size $n = |X|$ of the set X, are examples with full support. A fair coin $\frac{1}{2}|H\rangle + \frac{1}{2}|T\rangle$ and a fair dice $\frac{1}{6}|1\rangle + \frac{1}{6}|2\rangle + \frac{1}{6}|3\rangle + \frac{1}{6}|4\rangle + \frac{1}{6}|5\rangle + \frac{1}{6}|6\rangle$ are examples of uniform distributions, on the set $\{H, T\}$ and on $\{1, 2, 3, 4, 5, 6\}$.

The size $\|\varphi\| \in \mathbb{N}$ of a multiset $\varphi \in \mathcal{M}(X)$ is the total number of elements, including multiplicities: $\|\varphi\| \coloneqq \sum_x \varphi(x)$. We use special notation for the set of multisets of a particular size K.

$$\mathcal{M}[K](X) \coloneqq \{\varphi \in \mathcal{M}(X) \mid \|\varphi\| = K\}.$$

The only multiset on X with size 0 is the constant-zero function $\mathbf{0}\colon X \to \mathbb{N}$.

For a non-zero multiset $\varphi \in \mathcal{M}(X)$ we write $\mathit{Flrn}(\varphi) \in \mathcal{D}(X)$ for the distribution obtained via 'frequentist learning', that is via counting and normalisation:

$$\mathit{Flrn}(\varphi)(x) \coloneqq \frac{\varphi(x)}{\|\varphi\|} \qquad \text{that is} \qquad \mathit{Flrn}(\varphi) = \sum_{x \in X} \frac{\varphi(x)}{\|\varphi\|} \, \big| x \big\rangle.$$

The multinomial distribution describes draws with replacement from an urn filled with coloured balls, represented as a distribution $\omega \in \mathcal{D}(X)$, where X is the set of colours. The number $\omega(x) \in [0,1]$ is the probability/fraction of balls of colour $x \in X$ in the urn. For a fixed number K, the multinomial distribution $\mathit{mn}[K](\omega)$ assigns a probability to a draw of K balls, represented as a multiset $\varphi \in \mathcal{M}[K](X)$. It is defined as:

$$\mathit{mn}[K](\omega) \coloneqq \sum_{\varphi \in \mathcal{M}[K](X)} (\!(\varphi)\!) \cdot \prod_{x \in X} \omega(x)^{\varphi(x)} \, \big| \varphi \big\rangle \in \mathcal{D}\Big(\mathcal{M}[K](X)\Big), \tag{1}$$

where $(\!(\varphi)\!) \coloneqq \frac{\|\varphi\|!}{\prod_x \varphi(x)!}$ is the multinomial coefficient of φ, see *e.g.* [13,15] for more details. For instance for a distribution $\omega = \frac{1}{8}|a\rangle + \frac{1}{4}|b\rangle + \frac{5}{8}|c\rangle$ over the set of colours $X = \{a, b, c\}$ and for draws of size $K = 2$ we get:

$$\begin{aligned} \mathit{mn}[2](\omega) = \tfrac{1}{64} \Big| 2|a\rangle \Big\rangle + \tfrac{1}{16} \Big| 1|a\rangle + 1|b\rangle \Big\rangle + \tfrac{1}{16} \Big| 2|b\rangle \Big\rangle + \tfrac{5}{32} \Big| 1|a\rangle + 1|c\rangle \Big\rangle \\ + \tfrac{5}{16} \Big| 1|b\rangle + 1|c\rangle \Big\rangle + \tfrac{25}{64} \Big| 2|c\rangle \Big\rangle. \end{aligned}$$

In Section 5 a distribution on words will be used to assign a multinomial probability to a document, as a multiset (bag) of words.

What we describe in (1) is the so-called multivariate case, with multiple colours. When there are just two colours, that is, when the set X has two elements, say $X = \{0, 1\}$, we are in the bivariate case. It will be used in Example 4. Via the isomorphisms $\mathcal{D}(\{0,1\}) \cong [0,1]$ and $\mathcal{M}[K](\{0,1\}) \cong \{0, 1, \ldots, K\}$ one gets the binomial distribution as special case of (1), for a bias $r \in [0,1]$,

$$\mathit{bn}[K](r) \coloneqq \sum_{n \in \{0,\ldots,K\}} \binom{K}{n} \cdot r^n \cdot (1-r)^{K-n} \, \big| n \big\rangle \in \mathcal{D}\Big(\{0, \ldots, K\}\Big). \tag{2}$$

2.1 Channels and their daggers

An essential element of the principled categorical approach to probability is the use of channels, also known as Kleisli maps. For two sets X, Y, a channel from X to Y is a probabilistic function, written as $c\colon X \mathrel{\circ\!\!\rightarrow} Y$. It is an actual function of the form $c\colon X \to \mathcal{D}(Y)$ that assigns a distribution $c(x) \in \mathcal{D}(Y)$ to each element $x \in X$. In traditional notation it is written as a conditional probability distribution $p(y \mid x)$. These channels (probabilistic functions) can be composed, both sequentially and in parallel; moreover, they can be reversed, giving what is called a dagger channel [5,4,10], also known as the Bayesian inverse $p(x \mid y)$ of $p(y \mid x)$. This gives a useful calculus of channels.

For a distribution $\omega \in \mathcal{D}(X)$ on the domain of a channel $c\colon X \rightsquigarrow Y$ we may 'push forward' (or 'transform') the distribution along the channel, giving a distribution $c \gg\!= \omega \in \mathcal{D}(Y)$, on the codomain Y of the channel. This new distribution $c \gg\!= \omega$ is also called the 'prediction'. It is defined as:

$$\big(c \gg\!= \omega\big)(y) \coloneqq \sum_{x\in X} \omega(x)\cdot c(x)(y). \tag{3}$$

Using push forward $\gg\!=$ we can define composition of channels $c\colon X \rightsquigarrow Y$ and $d\colon Y \rightsquigarrow Z$ to a new channel $d \circledcirc c\colon X \rightsquigarrow Z$, namely as $(d \circledcirc c)(x) \coloneqq d \gg\!= c(x)$. Notice that we use special notation $\circledcirc$ for composition of channels.

We turn to the reversal of a channel $c\colon X \rightsquigarrow Y$ in presence of a 'prior' distribution $\omega \in \mathcal{D}(X)$. The result is a channel $c^{\dagger}_{\omega}\colon Y \rightsquigarrow X$, defined as:

$$c^{\dagger}_{\omega}(y) \coloneqq \sum_{x\in X} \frac{\omega(x)\cdot c(x)(y)}{(c \gg\!= \omega)(y)}\,\big|\,x\,\big\rangle. \tag{4}$$

For more details about this reversal we refer to the literature [5,4,10].

3 Jeffrey's update rule and its decrease of divergence

In probabilistic learning one can distinguish two different approaches to updating, namely following Pearl [22] (and Bayes) or following Jeffrey [17], see for comparisons *e.g.* [3,19,8,12,16]. The two approaches may produce completely different outcomes, but it is poorly understood when to use which approach. The distinction between the rules is characterised mathematically in [14]: Pearl's rule increases validity (expected value) and Jeffrey's rule decreases divergence.

In the present context we need only Jeffrey's rule and refer to [8,12] for Pearl's counterpart. In the theorem below we first repeat (from [14]) the formulation of Jeffrey's rule in terms of the dagger of a channel (4), together with the associated decrease of the divergence. The second item is new and contains a generalisation of Jeffrey's rule to multiple channels and data distributions. The latter are typically obtained via frequentist learning *Flrn*, see Sections 4 and 5. The appendix contains a proof.

Kullback-Leibler divergence D_{KL} is a standard comparison of distributions on the same set. It is defined, for $\omega, \rho \in \mathcal{D}(X)$, via the natural logarithm ln:

$$D_{KL}(\omega,\rho) \coloneqq \sum_{x\in X} \omega(x)\cdot \ln\left(\frac{\omega(x)}{\rho(x)}\right). \tag{5}$$

Jeffrey's rule reduces the divergence between data and prediction. In the cognitive context of predictive coding [23,9,11] this is called 'error correction'.

Theorem 1. *Let $\omega \in \mathcal{D}(X)$ be a distribution, used as prior.*

1. *("Jeffrey's divergence decrease") For a channel $c\colon X \mathrel{\circ\!\!\!\rightarrow} Y$ and a 'data' distribution $\tau \in \mathcal{D}(Y)$,*

$$D_{KL}\big(\tau,\, c \gg\!= \omega\big) \;\geq\; D_{KL}\big(\tau,\, c \gg\!= \omega'\big) \qquad \textit{where} \qquad \omega' \coloneqq c_{\omega}^{\dagger} \gg\!= \tau. \tag{6}$$

   This mapping $\omega \mapsto \omega' \coloneqq c_{\omega}^{\dagger} \gg\!= \tau$ is Jeffrey's update rule, giving ω' as updated, posterior distribution.

2. *("Mixture divergence decrease") Let $c_i\colon X \mathrel{\circ\!\!\!\rightarrow} Y_i$ be a finite collection of channels with distributions $\tau_i \in \mathcal{D}(Y_i)$ and probabilities $r_i \in [0,1]$ satisfying $\sum_i r_i = 1$. Then:*

$$\begin{aligned}\sum\nolimits_i r_i \cdot D_{KL}\big(\tau_i,\, c_i \gg\!= \omega\big) \;\geq\; \sum\nolimits_i r_i \cdot D_{KL}\big(\tau_i,\, c_i \gg\!= \omega'\big)& \\ \textit{where} \quad \omega' \;\coloneqq\; \sum\nolimits_i r_i \cdot \Big(\big(c_i\big)_{\omega}^{\dagger} \gg\!= \tau_i\Big).&\end{aligned} \tag{7}$$

Proof. We refer to [14] for the details of the (non-trivial) proof of the divergence decrease for Jeffrey's update rule (6). It crucially depends on (8) below. Let $\omega \in \mathcal{D}(X)$ be a distribution with predicates $p_1, \ldots, p_n \in [0,1]^X$ satisfying $\sum_i p_i = \mathbf{1}$, pointwise, and with probabilities $r_1, \ldots, r_n \in [0,1]$ satisfying $\sum_i r_i = 1$. Assuming non-zero validities $\omega \models p_i$, for each i, one has:

$$\sum\nolimits_i \frac{r_i \cdot (\omega \models p_i)}{\sum_j r_j \cdot (\omega|_{p_j} \models p_i)} \;\leq\; 1. \tag{8}$$

See [14] for details about the validity (expected value) $\omega \models p$ of a predicate p w.r.t. a distribution ω, and about the updated distribution $\omega|_p$.

We will use the inequality (8) to prove the second point of the theorem. We use the disjoint union $K \coloneqq \coprod_i Y_i$ as index set and with predicates and probabilities, for $(i,y) \in K$,

$$p_{(i,y)} \coloneqq c_i =\!\ll \mathbf{1}_y \;=\; c_i(-)(y) \;\in\; [0,1]^X \qquad\qquad s_{(i,y)} \coloneqq r_i \cdot \tau_i(y) \;\in\; [0,1].$$

The proof of (7) works as follows, basically as in [14], but with an extra level of indexing, via the index set K. Recall in the mixture case the updated distribution $\omega' \coloneqq \sum_i r_i \cdot \big((c_i)_{\omega}^{\dagger} \gg\!= \tau_i\big)$.

$$\begin{aligned}
&\sum\nolimits_i r_i \cdot D_{KL}\big(\tau_i,\, c_i \gg\!= \omega'\big) \;-\; \sum\nolimits_i r_i \cdot D_{KL}\big(\tau_i,\, c_i \gg\!= \omega\big) \\
&\overset{(5)}{=} \sum\nolimits_i r_i \cdot \sum_{y \in Y_i} \tau_i(y) \cdot \left[\ln\left(\frac{\tau_i(y)}{(c_i \gg\!= \omega')(y)}\right) \;-\; \ln\left(\frac{\tau_i(y)}{(c_i \gg\!= \omega)(y)}\right)\right] \\
&= \sum_{(i,y) \in K} s_{(i,y)} \cdot \ln\left(\frac{(c_i \gg\!= \omega)(y)}{(c_i \gg\!= \omega')(y)}\right) \\
&= \sum_{(i,y) \in K} s_{(i,y)} \cdot \ln\left(\frac{\omega \models c_i =\!\ll \mathbf{1}_y}{\omega' \models c_i =\!\ll \mathbf{1}_y}\right) \\
&\leq \ln\left(\sum_{(i,y) \in K} s_{(i,y)} \cdot \frac{\omega \models p_{(i,y)}}{\omega' \models p_{(i,y)}}\right) \qquad \text{by Jensen's inequality} \\
&\overset{(*)}{\leq} \ln\big(1\big) \;=\; 0.
\end{aligned}$$

The marked inequality $\overset{(*)}{\leq}$ uses (8). It applies since for $(i, y) \in K$,

$$\begin{aligned}
\omega' \models p_{(i,y)} &= \sum\nolimits_j r_j \cdot \big((c_j)^{\dagger}_{\omega} \mathrel{\gg\!\!=} \tau_j\big) \models p_{(i,y)} \\
&= \sum\nolimits_j r_j \cdot \sum_{x \in X} \big((c_j)^{\dagger}_{\omega} \mathrel{\gg\!\!=} \tau_j\big)(x) \cdot p_{(i,y)}(x) \\
&= \sum\nolimits_j r_j \cdot \sum_{x \in X} \sum_{z \in Y_j} (c_j)^{\dagger}_{\omega}(z)(x) \cdot \tau_j(z) \cdot p_{(i,y)}(x) \\
&\overset{(**)}{=} \sum_{(j,z) \in K} s_{(j,k)} \cdot \sum_{x \in X} \omega|_{c_j \mathrel{=\!\!\ll} \mathbf{1}_z}(x) \cdot p_{(i,y)}(x) \\
&= \sum_{(j,z) \in K} s_{(j,k)} \cdot \big(\omega|_{p_{(j,z)}} \models p_{(i,y)}\big).
\end{aligned}$$

The equation $\overset{(**)}{=}$ uses that the dagger definition (4) can equivalently be described as an update: $(c_j)^{\dagger}_{\omega}(z) = \omega|_{c_j \mathrel{=\!\!\ll} \mathbf{1}_z}$, see [14] for details. □

We include an illustration of Jeffrey's rule, as in the above first item.

Example 2. The following update question is attributed to Jeffrey, and reproduced for instance in [3,6]. It involves three colors of clothes: green (g), blue (b) and violet (v), in a space $C = \{g, b, v\}$. Clothes can be sold or not, as represented by $S = \{s, s^{\perp}\}$. The prior sales distribution $\omega \in \mathcal{D}(S)$ is $\omega = \frac{14}{25}|s\rangle + \frac{11}{25}|s^{\perp}\rangle$; it tells that a bit more than half of the clothes are sold. The colour distributions for sales and non-sales are provided via a channel $c \colon S \to \mathcal{D}(C)$, of the form:

$$c(s) = \tfrac{3}{14}|g\rangle + \tfrac{3}{14}|b\rangle + \tfrac{4}{7}|v\rangle \qquad c(s^{\perp}) = \tfrac{9}{22}|g\rangle + \tfrac{9}{22}|b\rangle + \tfrac{2}{11}|v\rangle.$$

A cloth is inspected by candlelight and the following likelihoods are reported per color: 70% certainty that it is green, 25% that it is blue, and 5% that it is violet. This gives a data/evidence distribution $\tau = \frac{7}{10}|g\rangle + \frac{1}{4}|b\rangle + \frac{1}{20}|v\rangle \in \mathcal{D}(C)$. We ask: what is the likelihood that the observed cloth will be sold?

The push-forward colour distribution $c \mathrel{\gg\!\!=} \omega$ with its prior divergence from the data are:

$$c \mathrel{\gg\!\!=} \omega \overset{(3)}{=} \tfrac{3}{10}|g\rangle + \tfrac{3}{10}|b\rangle + \tfrac{2}{5}|v\rangle \qquad D_{KL}\big(\tau,\, c \mathrel{\gg\!\!=} \omega\big) \overset{(5)}{=} 0.444.$$

The formula (4) determines the dagger channel $d \coloneqq c^{\dagger}_{\omega} \colon C \to \mathcal{D}(S)$ as:

$$d(g) = \tfrac{2}{5}|s\rangle + \tfrac{3}{5}|s^{\perp}\rangle \qquad d(b) = \tfrac{2}{5}|s\rangle + \tfrac{3}{5}|s^{\perp}\rangle \qquad d(v) = \tfrac{4}{5}|s\rangle + \tfrac{1}{5}|s^{\perp}\rangle.$$

We then get as updated (posterior) sales distribution $\omega' \coloneqq d \mathrel{\gg\!\!=} \tau \in \mathcal{D}(S)$ with decreased divergence:

$$\omega' \coloneqq d \mathrel{\gg\!\!=} \tau = \tfrac{21}{50}|s\rangle + \tfrac{29}{50}|s^{\perp}\rangle \quad \text{now with} \quad D_{KL}\big(\tau,\, c \mathrel{\gg\!\!=} \omega'\big) = 0.368.$$

The posterior sale probability $\frac{21}{50}$ for the inspected cloth is lower than the prior probability $\frac{14}{25} = \frac{28}{50}$. This outcome also occurs in [3, Ex. 1], [6, p.41] (as marginal), but without the above dagger-channel and the divergence decrease.

4 Expectation Maximisation (EM)

Expectation Maximisation (EM) is an algorithm where two steps, called E-step and M-step are alternated and iterated, as in E-M-E-M-E-M-$\cdots$, until some fixed point is reached. Its first general formulation occurs in [7], but it was used in more specialised forms before, see [18, 1.8] for historical details. In general, EM seeks an appropriate mixture of hidden/latent variables together with appropriate parameter values in a statistical model, see [20].

Here we describe the model and algorithm in channel-based form, where the divergence between data and predictions decreases with every iteration. The setting involves a channel, with a 'mixture' distribution on its domain and a 'data' multiset on its codomain. The channel will have type $Z \rightsquigarrow Y$, where Z is the space of classifications, and Y is the data space. Typically, the channel is determined by a parameter θ, which we write as $c[\theta]\colon Z \rightsquigarrow Y$. This θ may be a single number, a list of numbers, or even a matrix, of some dimension.

Theorem 3. *Let a 'data' multiset $\psi \in \mathcal{M}(Y)$ be given. We consider an initial 'mixture' distribution $\omega^{(0)} \in \mathcal{D}(Z)$ and a family of channels $c[\theta]\colon Z \rightsquigarrow Y$, with parameter θ, having an initial value $\theta^{(0)}$.*

Consider the following two steps at stage $n \in \mathbb{N}$, to produce new distributions and channels, assuming that we already have a distribution $\omega^{(n)} \in \mathcal{D}(Z)$ and channel $c^{(n)} := c[\theta^{(n)}]\colon Z \rightsquigarrow Y$, for parameter value $\theta^{(n)}$.

E-step *Using Jeffrey's update rule, from Theorem 1* (1)*, we obtain a next mixture distribution as:*

$$\omega^{(n+1)} := c[\theta^{(n)}]^{\dagger}_{\omega^{(n)}} \gg\!= \mathit{Flrn}(\psi) \in \mathcal{D}(Z). \tag{9}$$

M-step *We pick as next channel-parameter value the one with minimal Kullback-Leibler divergence in:*

$$\theta^{(n+1)} \in \operatorname*{argmin}_{\theta} D_{KL}\Big(\mathit{Flrn}(\psi),\, c[\theta] \gg\!= \omega^{(n)}\Big). \tag{10}$$

Take $c^{(n+1)} := c[\theta^{(n+1)}]$ as next channel.

These two steps result in decreasing divergences.

1. *Each iteration yields a decrease of Kullback-Leibler divergence:*

$$D_{KL}\Big(\mathit{Flrn}(\psi),\, c^{(n+1)} \gg\!= \omega^{(n+1)}\Big) \le D_{KL}\Big(\mathit{Flrn}(\psi),\, c^{(n)} \gg\!= \omega^{(n)}\Big). \tag{11}$$

   This means that the predicted data distribution is decreasingly wrong.

2. *A next parameter $\theta^{(n+1)}$ can (often) be found as solution to the equation:*

$$\sum_{z\in Z,\, y\in Y} \psi(y) \cdot \big(c^{(n)}\big)^{\dagger}_{\omega^{(n)}}(y)(z) \cdot \frac{\mathrm{d}}{\mathrm{d}\theta} \ln\Big(c[\theta](z)(y)\Big) = 0. \tag{12}$$

   This solution is not the minimal one in (10)*, but it still yields the relevant decrease of divergence in* (11)*.*

The word ‘often’ is inserted because finding a minimal parameter value via a solution of (12) only works when the channel has suitable (partial) derivatives, see Example 4 below.

Proof. 1. The claimed decrease of divergence arises as follows.

$$\begin{array}{ll}
D_{KL}\Big(\mathit{Flrn}(\psi),\, c[\theta^{(n+1)}] \gg= \omega^{(n+1)}\Big) & \\
\quad\leq\, D_{KL}\Big(\mathit{Flrn}(\psi),\, c[\theta^{(n)}] \gg= \omega^{(n+1)}\Big) & \text{since } \theta^{(n+1)} \text{ is argmin} \\
\quad\leq\, D_{KL}\Big(\mathit{Flrn}(\psi),\, c[\theta^{(n)}] \gg= \big(c[\theta^{(n)}]^{\dagger}_{\omega^{(n)}} \gg= \mathit{Flrn}(\psi)\big)\Big) & \text{by definition of } \omega^{(n+1)} \\
\quad\leq\, D_{KL}\Big(\mathit{Flrn}(\psi),\, c[\theta^{(n)}] \gg= \omega^{(n)}\Big) & \text{by Theorem 1 (1).}
\end{array}$$

2. The minimum parameter value θ in the expression $D_{KL}\Big(\mathit{Flrn}(\psi),\, c[\theta] \gg= \omega^{(n)}\Big)$ in (10) is located where the derivative $\frac{\mathrm{d}}{\mathrm{d}\theta}$ is zero. We thus calculate:

$$\begin{array}{l}
\frac{\mathrm{d}}{\mathrm{d}\theta}\, D_{KL}\Big(\mathit{Flrn}(\psi),\, c[\theta] \gg= \omega^{(n)}\Big) \\
\overset{(5)}{=}\; \frac{\mathrm{d}}{\mathrm{d}\theta} \sum_{y\in Y} \mathit{Flrn}(\psi)(y)\cdot \ln\left(\frac{\mathit{Flrn}(\psi)(y)}{(c[\theta] \gg= \omega^{(n)})(y)}\right) \\
=\; \frac{\mathrm{d}}{\mathrm{d}\theta} \sum_{y\in Y} \mathit{Flrn}(\psi)(y)\cdot \ln\Big(\mathit{Flrn}(\psi)(y)\Big) - \sum_{y\in Y} \mathit{Flrn}(\psi)(y)\cdot \ln\Big((c[\theta] \gg= \omega^{(n)})(y)\Big) \\
=\; \frac{-1}{\|\psi\|}\cdot \sum_{y\in Y} \psi(y)\cdot \frac{\mathrm{d}}{\mathrm{d}\theta} \ln\Big((c[\theta] \gg= \omega^{(n)})(y)\Big) \\
=\; \frac{-1}{\|\psi\|}\cdot \sum_{y\in Y} \frac{\psi(y)}{(c[\theta] \gg= \omega^{(n)})(y)}\cdot \frac{\mathrm{d}}{\mathrm{d}\theta}(c[\theta] \gg= \omega^{(n)})(y) \\
=\; \frac{-1}{\|\psi\|}\cdot \sum_{y\in Y} \frac{\psi(y)}{(c[\theta] \gg= \omega^{(n)})(y)}\cdot \frac{\mathrm{d}}{\mathrm{d}\theta} \sum_{z\in Z} c[\theta](z)(y)\cdot \omega^{(n)}(z) \\
=\; \frac{-1}{\|\psi\|}\cdot \sum_{z\in Z,\, y\in Y} \frac{\psi(y)\cdot \omega^{(n)}(z)}{(c[\theta] \gg= \omega^{(n)})(y)}\cdot \frac{\mathrm{d}}{\mathrm{d}\theta}\, c[\theta](z)(y) \\
=\; \frac{-1}{\|\psi\|}\cdot \sum_{z\in Z,\, y\in Y} \frac{\psi(y)\cdot \omega^{(n)}(z)\cdot c[\theta](z)(y)}{(c[\theta] \gg= \omega^{(n)})(y)}\cdot \frac{\mathrm{d}}{\mathrm{d}\theta} \ln\Big(c[\theta](z)(y)\Big) \\
\overset{(4)}{=}\; \frac{-1}{\|\psi\|}\cdot \sum_{z\in Z,\, y\in Y} \psi(y)\cdot c[\theta]^{\dagger}_{\omega^{(n)}}(y)(z)\cdot \frac{\mathrm{d}}{\mathrm{d}\theta} \ln\Big(c[\theta](z)(y)\Big). \qquad (*)
\end{array}$$

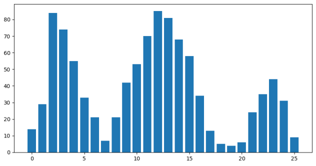
At this stage we need two more observations to see why it suffices to solve the equation (12).

(a) The leading factor $\frac{-1}{\|\psi\|}$ can be dropped from the above last line $(*)$ when we seek a solution via setting it to zero; because of the minus sign -1, we are not looking for a minimum, but for a maximum.

(b) The first θ in the dagger expression $c[\theta]^{\dagger}_{\omega^{(n)}}$ in $(*)$ can be replaced by $\theta^{(n)}$, which turns $c[\theta]^{\dagger}_{\omega^{(n)}}$ into $\big(c^{(n)}\big)^{\dagger}_{\omega^{(n)}}$, as in (12). This is a subtle point.

As we can see in the four line proof of Theorem 3 (1), we only need that the solution $\theta^{(n+1)}$ yields a divergence that is less than the divergence for $\theta^{(n)}$. Hence if one of the θ's in $(*)$ equals $\theta^{(n)}$, we do not get the real minimum divergence for $\theta^{(n+1)}$, but we still get a divergence that is less than the one for $\theta^{(n)}$. □

Example 4.



Consider the histogram of 1000 data elements on the right, on the space $\{0, 1, \ldots, N\}$ for $N = 25$. The shape of the data suggests that we have a mixture of three binomials at hand. Indeed, we have obtained these data by sampling 1000 times from the mixture of binomials:

$$\tfrac{1}{2} \cdot bn[N]\big(\tfrac{1}{2}\big) \;+\; \tfrac{1}{3} \cdot bn[N]\big(\tfrac{1}{8}\big) \;+\; \tfrac{1}{6} \cdot bn[N]\big(\tfrac{9}{10}\big). \tag{13}$$

Our aim in this example is to see if we can recover the mixture weights $(\frac{1}{2}, \frac{1}{3}, \frac{1}{6})$ and the biases $(\frac{1}{2}, \frac{1}{8}, \frac{9}{10})$ in (13) from these sampled data, via EM as described in Theorem 3. Formally, the above plot is used as a multiset $\psi = 14|0\rangle + 29|1\rangle + \cdots + 9|25\rangle \in \mathcal{M}[1000]\big(\{0, \ldots, 25\}\big)$.

We take a three element latent space, say $Z = \{1, 2, 3\}$, together with a parameterised channel $c[\boldsymbol{\theta}] \colon Z \rightsquigarrow \{0, 1, \ldots, N\}$, in this situation with $N = 25$. The channel $c[\boldsymbol{\theta}]$ consists of three binomial distributions, with a 3-tuple $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3) \in [0, 1]^3$ as parameter, via $c[\boldsymbol{\theta}](i) \coloneqq bn[N](\theta_i)$.

In this situation we illustrate how to solve Equation (12), where, for convience we abbreviate the dagger channel as $d_n \coloneqq \big(c^{(n)}\big)^{\dagger}_{\omega^{(n)}} \colon \{0, \ldots, N\} \rightsquigarrow Z$. We look at the solution for partial derivatives, for each $i \in Z$, using the familiar equation $\frac{\partial}{\partial x} \ln(x) = \frac{1}{x}$, plus the fact that the logarithm turns products into sums:

$$\begin{aligned}
0 &\overset{(12)}{=} \sum_{z \in \{1,2,3\}} \; \sum_{k \in \{0,\ldots,N\}} \psi(k) \cdot d_n(k)(z) \cdot \frac{\partial}{\partial \theta_i} \ln\Big(c[\boldsymbol{\theta}](z)(k)\Big) \\
&\overset{(2)}{=} \sum_{z \in \{1,2,3\}} \; \sum_{k \in \{0,\ldots,N\}} \psi(k) \cdot d_n(k)(z) \cdot \frac{\partial}{\partial \theta_i} \ln\Big(\tbinom{N}{k} \cdot \theta_z^k \cdot (1 - \theta_z)^{N-k}\Big) \\
&= \sum_{k \in \{0,\ldots,N\}} \psi(k) \cdot d_n(k)(i) \cdot \left[\frac{k}{\theta_i} - \frac{N-k}{1 - \theta_i}\right].
\end{aligned}$$

Via some elementary arithmethic we now get as solution:

$$\theta_i \;=\; \frac{\sum_k \psi(k) \cdot k \cdot d_n(k)(i)}{N \cdot \sum_k \psi(k) \cdot d_n(k)(i)}. \tag{$*$}$$

At this stage we can put things together and give a concrete description of the EM-algorithm (for the current example).

1. Pick arbitrary $\omega^{(0)} \in \mathcal{D}(Z) = \mathcal{D}(\{1, 2, 3\})$ and $\boldsymbol{\theta}^{(0)} \in [0, 1]^3$;

2. Assume $\omega^{(n)} \in \mathcal{D}(Z)$ and $\boldsymbol{\theta}^{(n)} \in [0,1]^3$ are already computed and use them first to form the dagger channel $d_n \coloneqq c\big[\boldsymbol{\theta}^{(n)}\big]^{\dagger}_{\omega^{(n)}} \colon \{0,\ldots,N\} \mathrel{\circ\!\!\!\rightarrow} Z$ as in (4).
   - **(E)** Take the next mixture distribution $\omega^{(n+1)} \in \mathcal{D}(Z)$ via Jeffrey's rule:
   $$\omega^{(n+1)}(i) \overset{(9)}{=} \Big(d_n \gg\!\!= \mathit{Flrn}(\psi)\Big)(i) \;=\; \frac{1}{\|\psi\|} \cdot \sum_{k\in\{0,\ldots,N\}} d_n(k)(i) \cdot \psi(k).$$
   - **(M)** Take the next parameters $\boldsymbol{\theta}^{(n+1)} \in [0,1]^3$ as:
   $$\theta_i^{(n+1)} \overset{(*)}{=} \frac{\sum_k \psi(k) \cdot k \cdot d_n(k)(i)}{N \cdot \sum_k \psi(k) \cdot d_n(k)(i)} \;=\; \frac{\sum_k \mathit{Flrn}(\psi)(k) \cdot k \cdot d_n(k)(i)}{N \cdot \omega^{(n+1)}(i)}.$$

The table below gives an overview of five runs of this algorithm, starting from arbitrary values. Clearly, the divergences are decreasing, as prescribed in (11).

round	**KL-div**	**mixtures $\omega^{(n)}$**	**biases $\boldsymbol{\theta}^{(n)}$**
0	0.853	$0.477\vert 1\rangle + 0.354\vert 2\rangle + 0.169\vert 3\rangle$	0.235, 0.389, 0.691
1	0.326	$0.353\vert 1\rangle + 0.35\vert 2\rangle + 0.297\vert 3\rangle$	0.159, 0.46, 0.754
2	0.132	$0.321\vert 1\rangle + 0.454\vert 2\rangle + 0.225\vert 3\rangle$	0.128, 0.478, 0.812
3	0.029	$0.311\vert 1\rangle + 0.515\vert 2\rangle + 0.174\vert 3\rangle$	0.122, 0.488, 0.872
4	0.011	$0.309\vert 1\rangle + 0.535\vert 2\rangle + 0.156\vert 3\rangle$	0.121, 0.493, 0.898

We see that in five rounds we already get quite close to the original mixture and biases in (13). The order is different, but this is because the classification labels in $Z = \{1, 2, 3\}$ are meaningless and cannot be distinguished by the algorithm.

5 Latent Dirichlet Allocation (LDA)

The second model and inference algorithm in this paper was introduced in [2] under the name Latent Dirichlet Allocation, commonly abbreviated as LDA. It is used for what is called topic modeling: classifying documents according to their topics. The set-up of the algorithm is more complicated then EM and involves continuous Dirichlet distributions. In our analysis we show that LDA is essentially about divergence reduction via Jeffrey's rule — in multi-channel form, as in Theorem 1 (2). The Dirichlet distributions introduce a certain level of complexity, but turn out to play a limited role in the algorithm itself. We cover the essentials and refer to the literature for further information (see *e.g.* [21,20]).

Dirichlet is a continuous distribution on discrete distributions. Writing $\mathcal{G}$ for the Giry monad of continous distributions, we have $\mathit{Dir}(\alpha) \in \mathcal{G}\big(\mathcal{D}(X)\big)$, where X is a finite set and $\alpha \in \mathcal{M}(X)$ is a multiset with full support[1]. This $\mathit{Dir}(\alpha)$ is

[1] In the current paper we use multisets with natural numbers as multiplicities; this can be generalised to non-negative real numbers as multiplicities. The Dirichlet distribution $\mathit{Dir}(\alpha)$ can be defined for such more general multisets. But we shall not do so here since it does not affect the LDA algorithm.

defined via a probability density function (pdf) $\mathit{dir}(\alpha)\colon \mathcal{D}(X) \to \mathbb{R}_{\geq 0}$, namely:

$$\mathit{dir}(\alpha)(\omega) \coloneqq \frac{(\|\alpha\| - 1)!}{\prod_x (\alpha(x) - 1)!} \cdot \prod_{x \in X} \omega(x)^{\alpha(x)-1}.$$

The continuous Dirichlet distribution $\mathit{Dir}(\alpha) \in \mathcal{G}\big(\mathcal{D}(X)\big)$ is the function that assigns to a measurable subset $M \subseteq \mathcal{D}(X)$ the probability $\int_{\omega \in M} \mathit{dir}(\alpha)(\omega)\,\mathrm{d}\omega$.

We assume a finite set W of words and use the bag-of-words model for documents, so that a document is a multiset $\psi \in \mathcal{M}(W)$ over words. As data we use a collection of such documents/multisets, written as $\boldsymbol{\psi} = \big(\psi_i\big)_{i \in I}$, for some finite index set I. We shall write it as $\boldsymbol{\psi} \in \mathcal{M}(W)^I$ and call it a corpus.

We also assume a finite set T of topics. This may simply be a set $\boldsymbol{n} \coloneqq \{0, 1, \ldots, n-1\}$, since topics do not have an interpretation.

We shall use multisets $\alpha \in \mathcal{M}(T)$ and $\beta \in \mathcal{M}(W)$ as parameters for Dirichlet distributions $\mathit{Dir}(\alpha) \in \mathcal{G}\big(\mathcal{D}(T)\big)$ and $\mathit{Dir}(\beta) \in \mathcal{G}\big(\mathcal{D}(W)\big)$, where $\alpha \in \mathcal{M}(T)$ and $\beta \in \mathcal{M}(W)$ are multisets with full support. We put them in parallel, using the tensor $\otimes$ for continuous distributions, and thus get:

$$\mathit{Dir}(\alpha)^I \coloneqq \underbrace{\mathit{Dir}(\alpha) \otimes \cdots \otimes \mathit{Dir}(\alpha)}_{|\,I\,|\text{ times}} \in \mathcal{G}\Big(\underbrace{\mathcal{D}(T) \times \cdots \times \mathcal{D}(T)}_{|\,I\,|\text{ times}}\Big) = \mathcal{G}\big(\mathcal{D}(T)^I\big).$$

Similarly, we use $\mathit{Dir}(\beta)^T \in \mathcal{G}\big(\mathcal{D}(W)^T\big)$. These parallel products $\otimes$ of continuous distributions work via the multiplication of the pdf's involved, see *e.g.* [21].

These parallel Dirichlet's are used as (continuous) distributions on $\boldsymbol{\theta} \in \mathcal{D}(T)^I$ and $\boldsymbol{\zeta} \in \mathcal{D}(W)^T$, that is on a document-topic channel $\boldsymbol{\theta}\colon I \to \mathcal{D}(T)$ and a topic-word channel $\boldsymbol{\zeta}\colon T \to \mathcal{D}(W)$. This $\boldsymbol{\theta}$ sends a document (index) $i \in I$ to the topic distribution $\boldsymbol{\theta}(i) \in \mathcal{D}(T)$ for document $\psi_i \in \mathcal{M}(W)$. Similarly, $\boldsymbol{\zeta}$ sends a topic $t \in T$ to the distribution $\boldsymbol{\zeta}(t) \in \mathcal{D}(W)$ of words, for the topic t.

The LDA model consists of the following composite, where *mn* is multinomial.

$$\mathcal{D}(T)^I \times \mathcal{D}(W)^T \xrightarrow{\mathit{comp}} \mathcal{D}(W)^I \xrightarrow{\mathit{mn}^I} \mathcal{D}\big(\mathcal{M}(W)^I\big)$$

The function *comp* performs channel composition: $\mathit{comp}\big(\boldsymbol{\theta}, \boldsymbol{\zeta}\big) = \boldsymbol{\zeta} \mathbin{\circ\!\!\cdot} \boldsymbol{\theta}\colon I \to \mathcal{D}(W)$. The likelihood for document data $\boldsymbol{\psi} \in \mathcal{M}(W)^I$, given hyperparameters α, β is expressed by the (continuous) push forward:

$$\Big((\mathit{mn}^I \circ \mathit{comp}) \gg\!= \big(\mathit{Dir}(\alpha)^I \otimes \mathit{Dir}(\beta)^T\big)\Big)(\boldsymbol{\psi}) \in [0,1]. \tag{14}$$

We can write the expression (14) in terms of integrals:

$$\int_{\boldsymbol{\theta} \in \mathcal{D}(T)^I} \int_{\boldsymbol{\zeta} \in \mathcal{D}(W)^T} \prod_{i \in I} \mathit{dir}(\alpha)\big(\boldsymbol{\theta}(i)\big) \cdot \prod_{t \in T} \mathit{dir}(\beta)\big(\boldsymbol{\zeta}(t)\big) \cdot \prod_{i \in I} \mathit{mn}\big(\boldsymbol{\zeta} \gg\!= \boldsymbol{\theta}(i)\big)(\psi_i)\,\mathrm{d}\boldsymbol{\zeta}\,\mathrm{d}\boldsymbol{\theta}.$$

We are interested in the likelihood expression with $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ as free variables:

$$\mathcal{L}_{\alpha,\beta,\boldsymbol{\psi}}\big(\boldsymbol{\theta}, \boldsymbol{\zeta}\big) \coloneqq \prod_{i \in I} \mathit{dir}(\alpha)\big(\boldsymbol{\theta}(i)\big) \cdot \prod_{t \in T} \mathit{dir}(\beta)\big(\boldsymbol{\zeta}(t)\big) \cdot \prod_{i \in I} \mathit{mn}\big(\boldsymbol{\zeta} \gg\!= \boldsymbol{\theta}(i)\big)(\psi_i). \tag{15}$$

The LDA aim is to find the document-topic channel $\boldsymbol{\theta}\colon I \to \mathcal{D}(T)$ and the topic-word $\boldsymbol{\zeta}\colon T \to \mathcal{D}(W)$ that maximise this likelihood expression (15).

We shall use the (natural) logarithm ln of this expression, commonly called the log-likelihood; it turns the above products $\prod$ into sums $\sum$. Since ln is monotone, we might as well maximise the log-likelihood. A crucial observation is that this log-likelihood can be formulated in terms of Kullback-Leibler divergence. This opens the door to applying Jeffrey's update rule.

Lemma 5. *Let $\alpha \in \mathcal{M}(T)$ and $\beta \in \mathcal{M}(W)$ be multisets with full support and let $\boldsymbol{\psi} \in \mathcal{M}(W)^I$ be corpus of documents. The log-likelihood $\ln \mathcal{L}_{\alpha,\beta,\boldsymbol{\psi}}\big(\boldsymbol{\theta}, \boldsymbol{\zeta}\big)$ of the expression* (15) *can be written as:*

$$\begin{aligned}
&\ln \mathcal{L}_{\alpha,\beta,\boldsymbol{\psi}}\big(\boldsymbol{\theta}, \boldsymbol{\zeta}\big) \\
&= C - \sum_{i\in I} \big(\|\alpha - \mathbf{1}\| + \|\psi_i\|\big) \cdot \Bigg(\frac{\|\alpha-\mathbf{1}\|}{\|\alpha-\mathbf{1}\| + \|\psi_i\|} \cdot D_{KL}\Big(\mathit{Flrn}(\alpha-\mathbf{1}),\, \boldsymbol{\theta}(i)\Big) \\
&\qquad\qquad + \frac{\|\psi_i\|}{\|\alpha-\mathbf{1}\| + \|\psi_i\|} \cdot D_{KL}\Big(\mathit{Flrn}(\psi_i),\, \boldsymbol{\zeta} \gg\!= \boldsymbol{\theta}(i)\Big)\Bigg) \\
&\quad - \sum_{t\in T} \|\beta - \mathbf{1}\| \cdot D_{KL}\Big(\mathit{Flrn}(\beta-\mathbf{1}),\, \boldsymbol{\zeta}(t)\Big),
\end{aligned} \tag{16}$$

where C is a constant depending on the parameters $\alpha, \beta, \boldsymbol{\psi}$ but not on the variables $\boldsymbol{\theta}, \boldsymbol{\zeta}$. Recall that $\mathbf{1}$ is the multiset of singletons, so that $(\alpha - \mathbf{1})(x) = \alpha(x) - 1$. This subtraction is allowed since α has full support. The same holds for β.

Proof. (of Lemma 5). We apply the logarithm ln to (15), expand the Dirichlet and multinomial expressions, and write C for some constant, not depending on

$\boldsymbol{\theta}, \boldsymbol{\zeta}$.

$$
\begin{aligned}
&\ln \mathcal{L}_{\alpha,\beta,\boldsymbol{\psi}}\big(\boldsymbol{\theta}, \boldsymbol{\zeta}\big) \\
&= \sum_{i \in I} \ln\left(\frac{(\|\alpha\| - 1)!}{\prod_t (\alpha(t) - 1)!}\right) + \sum_{t \in T} (\alpha(t) - 1) \cdot \ln\big(\boldsymbol{\theta}(i)(t)\big) \\
&\qquad + \sum_{t \in T} \ln\left(\frac{(\|\beta\| - 1)!}{\prod_w (\beta(w) - 1)!}\right) + \sum_{w \in W} (\beta(w) - 1) \cdot \ln\big(\boldsymbol{\zeta}(t)(w)\big) \\
&\qquad + \sum_{i \in I} \ln\big(\boldsymbol{(}\psi_i\boldsymbol{)}\big) + \sum_{w \in W} \psi_i(w) \cdot \ln\Big(\big(\boldsymbol{\zeta} \gg\!= \boldsymbol{\theta}(i)\big)(w)\Big). \\
&= C + \sum_{i \in I} \|\alpha - \mathbf{1}\| \cdot \sum_{t \in T} \mathit{Flrn}(\alpha - \mathbf{1})(t) \cdot \ln\big(\boldsymbol{\theta}(i)(t)\big) \\
&\qquad\qquad - \|\alpha - \mathbf{1}\| \cdot \sum_{t \in T} \mathit{Flrn}(\alpha - \mathbf{1})(t) \cdot \ln\big(\mathit{Flrn}(\alpha - \mathbf{1})(t)\big) \\
&\qquad + \sum_{t \in T} \|\beta - \mathbf{1}\| \cdot \sum_{w \in W} \mathit{Flrn}(\beta - \mathbf{1})(w) \cdot \ln\big(\boldsymbol{\zeta}(t)(w)\big) \\
&\qquad\qquad - \|\beta - \mathbf{1}\| \cdot \sum_{w \in W} \mathit{Flrn}(\beta - \mathbf{1})(w) \cdot \ln\big(\mathit{Flrn}(\beta - \mathbf{1})(w)\big) \\
&\qquad + \sum_{i \in I} \|\psi_i\| \cdot \sum_{w \in W} \mathit{Flrn}(\psi_i)(w) \cdot \ln\Big(\big(\boldsymbol{\zeta} \gg\!= \boldsymbol{\theta}(i)\big)(w)\Big) \\
&\qquad\qquad - \|\psi_i\| \cdot \sum_{w \in W} \mathit{Flrn}(\psi_i)(w) \cdot \ln\big(\mathit{Flrn}(\psi_i)(w)\big) \\
&= C - \sum_{i \in I} \|\alpha - \mathbf{1}\| \cdot D_{KL}\Big(\mathit{Flrn}(\alpha - \mathbf{1}),\, \boldsymbol{\theta}(i)\Big) \;+\; \|\psi_i\| \cdot D_{KL}\Big(\mathit{Flrn}(\psi_i),\, \boldsymbol{\zeta} \gg\!= \boldsymbol{\theta}(i)\Big) \\
&\qquad - \sum_{t \in T} \|\beta - \mathbf{1}\| \cdot D_{KL}\Big(\mathit{Flrn}(\beta - \mathbf{1}),\, \boldsymbol{\zeta}(t)\Big) \\
&= C - \sum_{i \in I} r_i \cdot \Big(r_{i,1} \cdot D_{KL}\Big(\mathit{Flrn}(\alpha - \mathbf{1}),\, \boldsymbol{\theta}(i)\Big) + r_{i,2} \cdot D_{KL}\Big(\mathit{Flrn}(\psi_i),\, \boldsymbol{\zeta} \gg\!= \boldsymbol{\theta}(i)\Big)\Big) \\
&\qquad - \sum_{t \in T} \|\beta - \mathbf{1}\| \cdot D_{KL}\Big(\mathit{Flrn}(\beta - \mathbf{1}),\, \boldsymbol{\zeta}(t)\Big).
\end{aligned}
$$

In the last line we use the abbreviations (21). □

This lemma tells us that in order to *maximise* the log-likelihood we have to *minimise* the three Kullback-Leibler divergences in (16), because of the minus sign $-$ before the D_{KL} expressions.

Theorem 6. *Consider the LDA situation as described above, with multiset parameters $\alpha \in \mathcal{M}(T)$ and $\beta \in \mathcal{M}(W)$ and a corpus of documents $\boldsymbol{\psi} = \big(\psi_i\big)_{i \in I}$.*

An infinite series of channels $\boldsymbol{\theta}^{(n)} \in \mathcal{D}(T)^I$ and $\boldsymbol{\zeta}^{(n)} \in \mathcal{D}(W)^T$ with increasing likelihoods:

$$
\mathcal{L}_{\alpha,\beta,\boldsymbol{\psi}}\Big(\boldsymbol{\theta}^{(n+1)}, \boldsymbol{\zeta}^{(n+1)}\Big) \;\geq\; \mathcal{L}_{\alpha,\beta,\boldsymbol{\psi}}\Big(\boldsymbol{\theta}^{(n)}, \boldsymbol{\zeta}^{(n)}\Big) \tag{17}
$$

is obtained in the following manner.

At stage 0*, arbitrary channels* $\boldsymbol{\theta}^{(0)} \in \mathcal{D}(T)^I$ *and* $\boldsymbol{\zeta}^{(0)} \in \mathcal{D}(W)^T$ *are chosen. Subsequent stages are handled as follows.*

1. *The next document-topic channel* $\boldsymbol{\theta}^{(n+1)} \in \mathcal{D}(T)^I$ *is defined via the mixture version of Jeffrey's rule in Theorem 1 (2), as convex combination, at* $i \in I$*:*

$$\begin{aligned}\boldsymbol{\theta}^{(n+1)}(i) \coloneqq \; & \frac{\|\alpha - \mathbf{1}\|}{\|\alpha - \mathbf{1}\| + \|\psi_i\|} \cdot \mathit{Flrn}(\alpha - \mathbf{1}) \\ & \qquad + \frac{\|\psi_i\|}{\|\alpha - \mathbf{1}\| + \|\psi_i\|} \cdot \Big(\big(\boldsymbol{\zeta}^{(n)}\big)^{\dagger}_{\boldsymbol{\theta}^{(n)}(i)} \gg= \mathit{Flrn}(\psi_i) \Big).\end{aligned} \tag{18}$$

This rule is used with the identity channel together with the channel $\boldsymbol{\zeta}^{(n)}$.

2. *The next topic-word channel* $\boldsymbol{\zeta}^{(n+1)} \in \mathcal{D}(W)^T$ *at* $t \in T$ *and* $w \in W$ *is:*

$$\begin{aligned}\boldsymbol{\zeta}^{(n+1)}(t)(w) \coloneqq \operatorname*{argmin}_{\boldsymbol{\zeta} \in \mathcal{D}(W)^T} \sum_{i \in I} \; & D_{KL}\Big(\mathit{Flrn}(\psi_i),\, \boldsymbol{\zeta} \gg= \boldsymbol{\theta}^{(n+1)}(i) \Big) \\ & + D_{KL}\Big(\mathit{Flrn}(\beta - \mathbf{1}),\, \boldsymbol{\zeta}(t) \Big).\end{aligned} \tag{19}$$

Concretely, it can be chosen as:

$$\boldsymbol{\zeta}^{(n+1)}(t)(w) = \frac{\beta(w) - 1 + \sum_{i \in I} \; \psi_i(w) \cdot \big(\boldsymbol{\zeta}^{(n)}\big)^{\dagger}_{\boldsymbol{\theta}^{(n)}(i)}(w)(t)}{\|\beta - \mathbf{1}\| + \sum_{i \in I} \; \|\psi_i\| \cdot \big(\big(\boldsymbol{\zeta}^{(n)}\big)^{\dagger}_{\boldsymbol{\theta}^{(n)}(i)} \gg= \mathit{Flrn}(\psi_i) \big)(t)}. \tag{20}$$

Proof. For the first point it sufficies to prove this for the log-likelihood $\ln \mathcal{L}$. We drop the subscripts $\alpha, \beta, \boldsymbol{\psi}$ for convenience. Also, we abbreviate:

$$r_i \coloneqq \|\alpha - \mathbf{1}\| + \|\psi_i\| \qquad r_{i,1} \coloneqq \frac{\|\alpha - \mathbf{1}\|}{r} \qquad r_{i,2} \coloneqq \frac{\|\psi_i\|}{r} \tag{21}$$

Thus $r_{i,1} + r_{i,2} = 1$. Using the reformulation in Lemma 5 we get:

$$\begin{aligned}\ln \mathcal{L}\big(\boldsymbol{\theta}^{(n+1)}, \boldsymbol{\zeta}^{(n+1)}\big) &= C - \sum_{i \in I} r_i \cdot \Big(r_{i,1} \cdot D_{KL}\Big(\mathit{Flrn}(\alpha - \mathbf{1}),\, \boldsymbol{\theta}^{(n+1)}(i) \Big) \\ &\qquad\qquad + r_{i,2} \cdot D_{KL}\Big(\mathit{Flrn}(\psi_i),\, \boldsymbol{\zeta}^{(n+1)} \gg= \boldsymbol{\theta}^{(n+1)}(i) \Big) \Big) \\ &\qquad - \sum_{t \in T} \|\beta - \mathbf{1}\| \cdot D_{KL}\Big(\mathit{Flrn}(\beta - \mathbf{1}),\, \boldsymbol{\zeta}^{(n+1)}(t) \Big) \\ &\geq C - \sum_{i \in I} r_i \cdot \Big(r_{i,1} \cdot D_{KL}\Big(\mathit{Flrn}(\alpha - \mathbf{1}),\, \boldsymbol{\theta}^{(n+1)}(i) \Big) \\ &\qquad\qquad + r_{i,2} \cdot D_{KL}\Big(\mathit{Flrn}(\psi_i),\, \boldsymbol{\zeta}^{(n)} \gg= \boldsymbol{\theta}^{(n+1)}(i) \Big) \Big) \\ &\qquad - \sum_{t \in T} \|\beta - \mathbf{1}\| \cdot D_{KL}\Big(\mathit{Flrn}(\beta - \mathbf{1}),\, \boldsymbol{\zeta}^{(n)}(t) \Big) \\ &\geq C - \sum_{i \in I} r_i \cdot \Big(r_{i,1} \cdot D_{KL}\Big(\mathit{Flrn}(\alpha - \mathbf{1}),\, \boldsymbol{\theta}^{(n)}(i) \Big) \\ &\qquad\qquad + r_{i,2} \cdot D_{KL}\Big(\mathit{Flrn}(\psi_i),\, \boldsymbol{\zeta}^{(n)} \gg= \boldsymbol{\theta}^{(n)}(i) \Big) \Big) \\ &\qquad - \sum_{t \in T} \|\beta - \mathbf{1}\| \cdot D_{KL}\Big(\mathit{Flrn}(\beta - \mathbf{1}),\, \boldsymbol{\zeta}^{(n)}(t) \Big) \\ &= \ln \mathcal{L}\big(\boldsymbol{\theta}^{(n)}, \boldsymbol{\zeta}^{(n)}\big).\end{aligned}$$

The first inequality $\geq$ holds because $\boldsymbol{\zeta}^{(n+1)}$ is defined as argmin in (19). The second inequality $\geq$ follows from Jeffrey's divergence reduction, in mixture form, see Theorem 1 (2). We apply it with prior distribution $\omega \coloneqq \boldsymbol{\theta}^{(n)}(i)$, with two channels, namely the identity $c_1 \coloneqq \mathrm{id}\colon T \mathrel{\circ\!\!\rightarrow} T$ and $c_2 \coloneqq \boldsymbol{\zeta}^{(n)}\colon T \mathrel{\circ\!\!\rightarrow} W$, with two distributions $\tau_1 \coloneqq \mathit{Flrn}(\alpha{-}\mathbf{1}) \in \mathcal{D}(T)$ and $\tau_2 \coloneqq \mathit{Flrn}(\psi_i) \in \mathcal{D}(W)$, and with two probabilities $r_{1,1} \coloneqq \frac{\|\alpha-\mathbf{1}\|}{\|\alpha-\mathbf{1}\|+\|\psi_i\|}$ and $r_{i,2} \coloneqq \frac{\|\psi_i\|}{\|\alpha-\mathbf{1}\|+\|\psi_i\|}$. The dagger of the identity channel is the identity, so that $(c_1)^{\dagger}_{\omega} \gg\!= \tau_1 = \mathit{Flrn}(\alpha-\mathbf{1})$. The updated state ω' in Theorem 1 (2) is then $\boldsymbol{\theta}^{(n+1)}(i)$ as defined above.

We turn to formula (20). In order to find the argmin in (19) we use Lagrange's multiplier method, see *e.g.* [1, §2.2]. This method ensures that in the solution gives convex combinations

Thus, we first extend the relevant equation with additional parameters κ_t, for $t \in T$, in the function H defined as the log-likelihood plus an extra expression — typical for Lagrange:

$$H\big(\boldsymbol{\zeta}, \boldsymbol{\kappa}\big) \;=\; \ln \mathcal{L}_{\alpha,\beta,\boldsymbol{\psi}}\big(\boldsymbol{\theta}, \boldsymbol{\zeta}\big) + \sum_{t\in T} \kappa(t) \cdot \left(1 - \sum_{w\in W} \boldsymbol{\zeta}(t)(w)\right).$$

Thus we keep the hyperparameters $\alpha, \beta, \boldsymbol{\psi}$ and also the channel $\boldsymbol{\theta}$ fixed. We then consider the partial derivatives, for $s \in T$ and $v \in W$.

$$\begin{aligned}
\frac{\partial H}{\partial \boldsymbol{\zeta}(s)(v)}\big(\boldsymbol{\zeta}, \boldsymbol{\kappa}\big) \;&=\; \frac{(\beta(v) - 1)}{\boldsymbol{\zeta}(s)(v)} + \sum_{i\in I} \frac{\psi_i(v) \cdot \boldsymbol{\theta}(i)(s)}{\big(\boldsymbol{\zeta} \gg\!= \boldsymbol{\theta}(i)\big)(v)} \;-\; \kappa(s) \\
&=\; \frac{1}{\boldsymbol{\zeta}(s)(v)} \cdot \left(\beta(v) - 1 + \sum_{i\in I} \frac{\psi_i(v) \cdot \boldsymbol{\theta}(i)(s) \cdot \boldsymbol{\zeta}(s)(v)}{\big(\boldsymbol{\zeta} \gg\!= \boldsymbol{\theta}(i)\big)(v)}\right) \;-\; \kappa(s) \\
&=\; \frac{\beta(v) - 1 + \sum_{i\in I} \psi_i(v) \cdot \boldsymbol{\zeta}^{\dagger}_{\boldsymbol{\theta}(i)}(v)(s)}{\boldsymbol{\zeta}(s)(v)} \;-\; \kappa(s) \\
\frac{\partial H}{\partial \kappa(s)}\big(\boldsymbol{\zeta}, \boldsymbol{\kappa}\big) \;&=\; 1 - \sum_{w\in W} \boldsymbol{\zeta}(s)(w).
\end{aligned}$$

Setting all of these to zero yields:

$$\boldsymbol{\zeta}(s)(v) \;=\; \frac{\beta(v) - 1 + \sum_{i\in I} \psi_i(v) \cdot \boldsymbol{\zeta}^{\dagger}_{\boldsymbol{\theta}(i)}(v)(s)}{\kappa(s)}.$$

Thus:

$$\begin{aligned}
1 \;=\; \sum_{v\in W} \boldsymbol{\zeta}(s)(v) \;&=\; \sum_{v\in W} \frac{\beta(v) - 1 + \sum_{i\in I} \psi_i(v) \cdot \boldsymbol{\zeta}^{\dagger}_{\boldsymbol{\theta}(i)}(v)(s)}{\kappa(s)} \\
&=\; \frac{\|\beta{-}\mathbf{1}\| + \sum_{i\in I} \sum_{v\in W} \psi_i(v) \cdot \boldsymbol{\zeta}^{\dagger}_{\boldsymbol{\theta}(i)}(v)(s)}{\kappa(s)} \\
&=\; \frac{\|\beta{-}\mathbf{1}\| + \sum_{i\in I} \|\psi_i\| \cdot \big(\boldsymbol{\zeta}^{\dagger}_{\boldsymbol{\theta}(i)} \gg\!= \mathit{Flrn}(\psi_i)\big)(s)}{\kappa(s)}.
\end{aligned}$$

But then:

$$\boldsymbol{\zeta}(s)(v) \;=\; \frac{\beta(v) - 1 + \sum_{i\in I}\, \psi_i(v)\cdot \boldsymbol{\zeta}^{\dagger}_{\boldsymbol{\theta}(i)}(v)(s)}{\|\beta - \mathbf{1}\| + \sum_{i\in I}\, \|\psi_i\| \cdot \big(\boldsymbol{\zeta}^{\dagger}_{\boldsymbol{\theta}(i)} \gg\!= \mathit{Flrn}(\psi_i)\big)(s)}.$$

We may now use $\theta^{(n)}$ and $\zeta^{(n)}$ in the expression on the right-hand-side for the next-stage choice of $\zeta^{(n+1)}$, as in (20). □

We include a very simple example to illustrate LDA.

Example 7. We take a set $W = \{a, b, c, d, e, f\}$ with the first six letters of the alphabet as the set of words, and two topics: $T = \{1, 2\}$. We consider a corpus with 3 multisets of words, in the middle column in the table below. We see that the words b, d, f occur frequently in the first document, whereas the other words a, c, e occur often in the second one. The frequencies of letters in the third document is roughly equal. Hence we expect document 1 to be mostly on one topic, and document 2 on the other topic, and document 3 on both.

data	**document multiset**	**topic distribution**
1	$1\vert a\rangle + 6\vert b\rangle + 1\vert c\rangle + 7\vert d\rangle + 2\vert e\rangle + 8\vert f\rangle$	$0.831\vert 1\rangle + 0.169\vert 2\rangle$
2	$10\vert a\rangle + 1\vert b\rangle + 8\vert c\rangle + 2\vert d\rangle + 9\vert e\rangle + 1\vert f\rangle$	$0.132\vert 1\rangle + 0.868\vert 2\rangle$
3	$4\vert a\rangle + 3\vert b\rangle + 4\vert c\rangle + 5\vert d\rangle + 2\vert e\rangle + 3\vert f\rangle$	$0.512\vert 1\rangle \;+\; 0.488\vert 2\rangle$

The hyperparameter $\alpha \in \mathcal{M}(T)$ and $\beta \in \mathcal{M}(W)$ are chosen as constants, with multiplicity 2 for alpha and 1 for β. Running the LDA algorithm, as described in Theorem 6, 25 times yields a document-topic channel $\boldsymbol{\theta}$ with topic distributions for each document, in the column on the right in the above table. As expected, documents 1 and 2 are about different (opposite) topics.

The LDA-algorithm also produces a topic-word channel $\boldsymbol{\zeta}\colon T \mathrel{\circ\!\!\rightarrow} W$. It assigns in this simple example the following word probabilities to topics:

$$\begin{aligned} 1 &\mapsto 0.0000665|a\rangle + 0.278|b\rangle + 0.000707|c\rangle + 0.362|d\rangle + 0.0223|e\rangle + 0.337|f\rangle \\ 2 &\mapsto 0.362|a\rangle + 0.00277|b\rangle + 0.313|c\rangle + 0.0274|d\rangle + 0.295|e\rangle + 0.000435|f\rangle. \end{aligned}$$

This is consistent with what we saw above: topic 1 makes makes words b, d, f most likely, and topic 2 makes the other words a, c, e most likely.

6 Conclusions

EM and LDA are based on Jeffrey's update rule. Even if in actual implementations the formulations in terms of channels and their daggers may not be directly useful — for instance when results are approximated, typically via Gibbs sampling — having a crisp description of the mathematical essentials may be useful for understanding and reasoning about these fundamental EM and LDA algorithms in machine learning.

Acknowledgements Thanks to Dario Stein for many discussions about LDA.

References

1. C. Bishop. *Pattern Recognition and Machine Learning.* Information Science and Statistics. Springer, 2006.
2. D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journ. Machince Learning Research*, 3:993–1022, 2003. `doi:10.5555/944919.944937`.
3. H. Chan and A. Darwiche. On the revision of probabilistic beliefs using uncertain evidence. *Artif. Intelligence*, 163:67–90, 2005. `doi:10.1016/j.artint.2004.09.005`.
4. K. Cho and B. Jacobs. Disintegration and Bayesian inversion via string diagrams. *Math. Struct. in Comp. Sci.*, 29(7):938–971, 2019. `doi:10.1017/s0960129518000488`.
5. F. Clerc, F. Dahlqvist, V. Danos, and I. Garnier. Pointless learning. In J. Esparza and A. Murawski, editors, *Foundations of Software Science and Computation Structures*, number 10203 in Lect. Notes Comp. Sci., pages 355–369. Springer, Berlin, 2017. `doi:10.1007/978-3-662-54458-7_21`.
6. A. Darwiche. *Modeling and Reasoning with Bayesian Networks.* Cambridge Univ. Press, 2009.
7. A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journ. Royal Statistical Soc.*, 39(1):1–38, 1977.
8. F. Dietrich, C. List, and R. Bradley. Belief revision generalized: A joint characterization of Bayes' and Jeffrey's rules. *Journ. of Economic Theory*, 162:352–371, 2016. `doi:10.1016/j.jet.2015.11.006`.
9. K. Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010. `doi:10.1038/nrn2787`.
10. T. Fritz. A synthetic approach to Markov kernels, conditional independence, and theorems on sufficient statistics. *Advances in Math.*, 370:107239, 2020. `doi:10.1016/J.AIM.2020.107239`.
11. J. Hohwy. *The Predictive Mind.* Oxford Univ. Press, 2013. `doi:10.1093/acprof:oso/9780199682737.001.0001`.
12. B. Jacobs. The mathematics of changing one's mind, via Jeffrey's or via Pearl's update rule. *Journ. of Artif. Intelligence Research*, 65:783–806, 2019. `doi:10.1613/jair.1.11349`.
13. B. Jacobs. From multisets over distributions to distributions over multisets. In *Logic in Computer Science.* IEEE, Computer Science Press, 2021. `doi:10.1109/lics52264.2021.9470678`.
14. B. Jacobs. Learning from what's right and learning from what's wrong. In A. Sokolova, editor, *Math. Found. of Programming Semantics*, number 351 in Elect. Proc. in Theor. Comp. Sci., pages 116–133, 2021. `doi:10.4204/EPTCS.351.8`.
15. B. Jacobs. Urns & tubes. *Compositionality*, 4(4), 2022. `doi:10.32408/compositionality-4-4`.
16. B. Jacobs and D. Stein. Pearl's and Jeffrey's update as modes of learning in probabilistic programming. In M. Kerjean and P. Levy, editors, *Math. Found. of Programming Semantics*, Elect. Notes in Theor. Comp. Sci., 2023.
17. R. Jeffrey. *The Logic of Decision.* The Univ. of Chicago Press, 2^{nd} rev. edition, 1983.

18. G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Probability and Statistics. John Wiley & Sons, 2^{nd} edition, 2007. `doi:10.1002/9780470191613`.
19. A. Mrad, V. Delcroix, S. Piechowiak, P. Leicester, and M. Abid. An explication of uncertain evidence in Bayesian networks: likelihood evidence and probabilistic evidence. *Applied Intelligence*, 23(4):802–824, 2015.
20. K. Murphy. *Machine Learning. A Probabilistic Perspective*. MIT Press, Cambridge, MA, 2012.
21. P. Panangaden. *Labelled Markov Processes*. Imperial College Press, London, 2009.
22. J. Pearl. Jeffrey's rule, passage of experience, and neo-Bayesianism. In Jr. H. Kyburg, editor, *Knowledge Representation and Defeasible Reasoning*, pages 245–265. Kluwer Acad. Publishers, 1990. `doi:10.1007/978-94-009-0553-5_10`.
23. R. Rao and D. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2:79–87, 1999. `doi:10.1038/4580`.