Mathematical details of Dirichlet-categorical NG

Bas Cornelissen

August 8, 2017

This appendix develops the Dirichlet-categorical naming game in a more rigorous fashion. Please refer to chapter ?? for extensive motivation.

First, recall our notational conventions. The most precise notation would be of the form

$$\alpha_A^{(t)} = \left(\alpha_{A,1}^{(t)}, \dots, \alpha_{A,K}^{(t)}\right) \in \mathbb{R}^K, \tag{1}$$

and indicates the agent, the time and indices. I nearly always prefer a cleaner notation and often drop agents, or even time indices, whenever they are irrelevant. Also, vectors (boldface) get their time index in the subscript. Further recall that $\Sigma(\alpha) := \sum_k \alpha_k$ and that we write $[\![$ condition $]\!]$ for the indicator function evaluating to 1 if the condition holds and to 0 otherwise

Dirichlet and categorical distributions

The Dirichlet distribution is a continuous multivariate probability distribution defined over the interior of the $(K-1)$-simplex which we denote as $\Delta^{K-1} = \{x \in \mathbb{R}^K : \sum_k x_k = 1 \text{ and } 0 < x_i < 1\}$. We will only consider the $(K-1)$-simplex, so drop the superscript. Samples of a Dirichlet can thus be interpreted as K-dimensional probability-vectors. The Dirichlet is parametrised by a K-vector α and its density is given by

$$p(\theta \mid \alpha) = D(\alpha) \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}, \qquad D(\alpha) = \frac{\Gamma(\Sigma(\alpha))}{\prod_{k=1}^{K} \Gamma(\alpha_k)}, \tag{2}$$

where $D(\alpha)$ is the normalising constant, computed using the gamma function Γ, a continuous extension of the factorial with $\Gamma(n+1) = n!$ for $n \in \mathbb{N}$. If $\theta \sim \text{Dirichlet}(\alpha)$, then it has the following properties (e.g. Bishop 2006)

$$\mathrm{E}[\theta_k] = \frac{\alpha_k}{\Sigma(\alpha)}, \qquad \mathrm{Var}[\theta_k] = \frac{\alpha_k(\Sigma(\alpha) - \alpha_k)}{\Sigma(\alpha)^2(\Sigma(\alpha)+1)}, \qquad \mathrm{Mode}[\theta_k] = \frac{\alpha_k - 1}{\Sigma(\alpha) - K}. \tag{3}$$

It is often convenient to parametrise the Dirichlet differently, as $\alpha := \beta \cdot \mu$ with $\beta \in \mathbb{R}$ and $\mu \in \Delta$. Here β is the concentration parameter, a kind of inverse variance, and μ determines the location of the distribution. This translates into

$$\mathrm{E}[\theta_k] = \mu_k, \qquad \mathrm{Var}[\theta_k] = \frac{\mu_k(1-\mu_k)}{\beta+1} \tag{4}$$

from which we see that the mean is determined by μ and that larger β lead to smaller variance. We will use both parametrisations interchangeably.

The categorical distribution is a discrete probability distribution over K outcomes, described by a probability vector $\theta \in \Delta$. Recall that $c_k = \sum_i [\![x_i = k]\!]$ counts the number of k's in x. The joint distribution of b i.i.d. categorical variables $\mathrm{x} = (x_1, \ldots, x_b)$ is then given by

$$p(\mathrm{x} \mid \theta) = \prod_{i=1}^{b} \prod_{k=1}^{K} \theta_k^{[\![x_i=k]\!]} = \prod_{k=1}^{K} \theta_k^{\sum_i [\![x_k=k]\!]} = \prod_{k=1}^{K} \theta_k^{c_k}, \tag{5}$$

DIRICHLET-CATEGORICAL DISTRIBUTION To show that the Dirichlet is the *conjugate prior* of the categorical distribution, consider the following model

$$\theta \sim \mathrm{Dirichlet}(\alpha) \tag{6}$$

$$x_1, \ldots, x_b \sim \mathrm{Categorical}(\theta). \tag{7}$$

In this case, conjugacy means that the posterior distribution $p(\theta \mid \mathrm{x}, \alpha)$ is of the same parametric form as the prior $p(\theta \mid \alpha)$, namely a Dirichlet. More precisely, we have

$$p(\theta \mid \mathrm{x}, \alpha) \propto p(\theta \mid \alpha) \cdot p(\mathrm{x} \mid \theta) \tag{8}$$

$$\propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \cdot \prod_{k=1}^{K} \theta_k^{c_k} \tag{9}$$

$$= \prod_{k=1}^{K} \theta_k^{\alpha_k + c_k - 1}. \tag{10}$$

In the last line one can recognise a Dirichlet density with parameters $\alpha + \mathrm{c}$. We conclude that the the posterior is $\mathrm{Dirichlet}(\alpha + \mathrm{c})$-distributed, or, more explicitly,

$$\theta \mid \mathrm{x}, \alpha \sim \mathrm{Dirichlet}(\alpha_1 + c_1, \ldots, \alpha_K + c_K). \tag{11}$$

This result also illustrates the workings of the hyperparameter α. It is as if the model pretends to have observed α_k more instances of category k that it actually has. For that reason, the α_k's are often called *pseudo-counts*.

We can also derive the compound distribution $p(x \mid \alpha) = \int_\Delta p(x \mid \theta) \cdot p(\theta \mid \alpha)\, d\theta$ by marginalizing out all probability vectors θ. To do this, we have to use a trick, which exploits the fact that the Dirichlet distribution is normalized,

$$\int_\Delta D(\alpha) \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}\, d\theta = 1. \tag{12}$$

Moving the normalising constant out of the integral, we see that

$$\int_\Delta \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \, d\theta = \frac{1}{D(\alpha)} \tag{13}$$

Using that trick we can compute the marginal probability of x as

$$p(\mathrm{x} \mid \alpha) = \int_\Delta \prod_{i=1}^{b} p(x_i \mid \theta) \cdot p(\theta \mid \alpha) \, d\theta \tag{14}$$

$$= \int_\Delta \prod_{i=1}^{b} \theta_{x_i} \cdot D(\alpha) \cdot \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \, d\theta \tag{15}$$

$$= D(\alpha) \int_\Delta \prod_{k=1}^{K} \theta_k^{\alpha_k + c_k - 1} \, d\theta \tag{16}$$

$$= \frac{D(\alpha)}{D(\alpha + \mathrm{c})} \tag{17}$$

$$= \frac{\Gamma(\Sigma(\alpha))}{\Gamma(\Sigma(\alpha + \mathrm{c}))} \prod_{k=1}^{K} \frac{\Gamma(\alpha_k + c_k)}{\Gamma(\alpha_k)} \tag{18}$$

Note that when $b = 1$, hence $\mathrm{x} = (x)$, the (almost defining) relation $\Gamma(n+1) = n\Gamma(n)$ can be exploited to further simplify the distribution. Concretely, note that $\Gamma(\Sigma(\alpha + \mathrm{c})) = \Gamma(\Sigma(\alpha) + 1) = \Sigma(\alpha)\Gamma(\Sigma(\alpha))$. This simplifies the first term in equation 18, so we can simplify the marginal probability to

$$p(x \mid \alpha) = \frac{1}{\Sigma(\alpha)} \cdot \frac{\Gamma(\alpha_x + 1)}{\Gamma(\alpha_x)} \cdot \prod_{k \neq x} \frac{\Gamma(\alpha_k)}{\Gamma(\alpha_k)} \tag{19}$$

$$= \frac{1}{\Sigma(\alpha)} \cdot \frac{\alpha_x \Gamma(\alpha_x)}{\Gamma(\alpha_x)} = \frac{\alpha_x}{\Sigma(\alpha)}. \tag{20}$$

This also gives the posterior predictive distribution $p(y \mid \mathrm{x}, \alpha)$, since that is just $p(y \mid \alpha')$ for the updated parameters $\alpha' := \alpha + \mathrm{c}$:

$$p(y \mid \mathrm{x}, \alpha) = \frac{\alpha_y + c_y}{\Sigma(\alpha + \mathrm{c})} \tag{21}$$

This is a remarkably simple result, indeed: the probability of observing y is proportional to the number of times it has been observed already, including the pseudo-observations.

Exponentiated distributions

Different strategies can be used for selecting languages or words in the Bayesian Naming Game. This is done by exponentiating the distributions $p(\theta \mid \alpha)$ and $p(\mathrm{x} \mid \theta)$ by two parameters, η and ζ respectively. The resulting distributions again take simple form:

$$p(\theta \mid \alpha)^\eta \propto \left(\prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \right)^\eta = \prod_{k=1}^{K} \theta_k^{[\eta(\alpha_k - 1) + 1] - 1} \tag{22}$$

$$\propto \mathrm{Dirichlet}(\theta \mid \eta(\alpha - 1) + 1). \tag{23}$$

The case of the categorical is obvious,

$$p(x \mid \theta)^{\zeta} = \frac{\theta_x^{\zeta}}{\Sigma(\theta^{\zeta})}. \tag{24}$$

So we conclude that

$$p_{\text{LA}}(\theta \mid \alpha) = \text{Dirichlet}\big(\theta \mid \eta(\alpha - 1) + 1\big) \tag{25}$$

$$p_{\text{PA}}(x \mid \theta) = \text{Categorical}\big(x \mid \theta^{\zeta}/\Sigma(\theta^{\zeta})\big) \tag{26}$$

THE DIFFICULT CASE $\zeta = \infty$ Whenever $\zeta \neq 1$, Bayesian agents are facing a different inference problem and should infer the posterior distribution $p(\theta \mid x, \alpha) \propto p(x \mid \theta)^{\zeta} \cdot p(\theta \mid \alpha)$. However, this is no longer a Dirichlet distribution. To see this, we compute the joint

$$p(\theta, x \mid \alpha) = p(\theta \mid \alpha) \cdot \prod_{i=1}^{b} p(x_i \mid \theta)^{\zeta} \tag{27}$$

$$= D(\alpha) \cdot \prod_k \theta_k^{\alpha_k - 1} \cdot \frac{1}{\Sigma(\theta^{\zeta})} \cdot \prod_k \theta_k^{\zeta \cdot c_k} \tag{28}$$

$$= \frac{1}{\Sigma(\theta^{\zeta})} \cdot D(\alpha) \cdot \prod_k \theta_k^{\alpha_k + \zeta c_k - 1} \tag{29}$$

Although this is reminiscent of the Dirichlet density, it is not proportional to it, since the first term depends on θ. Consequently, deriving a closed-form expression of the posterior seems hard, as it involves solving the integral

$$\int_{\Delta} \frac{1}{\Sigma(\theta^{\zeta})} \prod_k \theta_k^{\alpha_k + \zeta c_k - 1} \, d\theta. \tag{30}$$

The reciprocal of the sum hindered any progress on this point and all suggestions would be more than welcome.[1]

This might be an interesting problem in its own, partly because we *can* relatively easily identify the extreme cases. When $\zeta = 1$ we trivially get the normal posterior, but when $\zeta = \infty$ we can also get an idea of the posterior. After observing x, the language θ used to generate it can only be one where x gets maximum probability. In other words, θ must have been in $\Delta_x := \{\theta \in \Delta : \theta_x \geq \theta_k \text{ for all } k\}$. Consequently,

$$p(\theta \mid x, \alpha) = p(\theta \mid \theta \in \Delta_x, \alpha) \tag{31}$$

$$= \frac{[\![\theta \in \Delta_x]\!] \cdot \text{Dirichlet}(\theta \mid \alpha)}{\int_{\Delta_x} \text{Dirichlet}(\theta \mid \alpha) \, d\theta} \tag{32}$$

That is, the posterior is proportional to the prior, restricted to the section Δ_x of the simplex where the largest component is x. I have not tried integrating the Dirichlet over Δ_x yet, other than the symmetric case, i.e. $\alpha = \beta/K \cdot 1$, when the integral is simply $1/K$. In any case, one immediately sees that this cannot be a Dirichlet distribution: the posterior is discontinuous at the boundary of Δ_x, or at least at the part of it that lies inside Δ.

1 I also posted the problem at math.stackexchange.com/q/2360468.

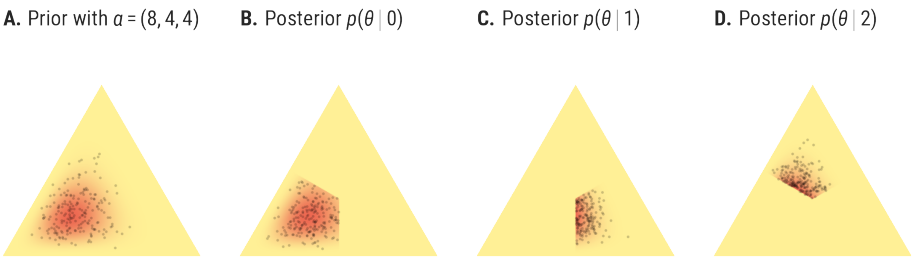

FIGURE 1 The posterior distribution $p(\theta \mid x)$ for various x if $\zeta = \infty$, that is, if agents always pick the most likely word. The posterior restricts the prior to the area of the simplex where $\arg\max_k \theta_k = x$.

FIG02

All this is illustrated in figure 1. It should be clear from that figure that agents who update their beliefs like this very quickly run into serious problems. After observing $x = 0$ all mass is restricted to Δ_0 and if the agent next observes $x = 1$, it has the problem that $p(1 \mid \alpha) = \int_{\Delta_0} p(\theta, 1)\, d\theta = 0$. This is one of the reasons for assuming that agents use exaggerated distributions only during production, and do *not* account for them during posterior inference.

Measuring the distance between languages

Most measures introduced to analyse the Dirichlet-categorical naming game, measure distances between languages. As a distance measure for distributions, the Jensen-Shannon divergence (JSD) can be used. The JSD is a symmetric version of the more common Kullback-Leibler divergence and measures the similarity between probability distributions. Figure 2 illustrates the JSD of different discrete distributions to the uniform distribution. Formally, if $\pi_1, \ldots, \pi_N$ are probability distributions, their divergence is

$$\text{JSD}(\pi_1, \ldots, \pi_N) := H\Big(\frac{1}{N}\sum_{i=1}^{N} \pi_i\Big) - \frac{1}{N}\sum_{i=1}^{N} H(\pi_i), \tag{33}$$

where H is the Shannon-entropy, a measure for the uncertainty in a distribution. As one can see, the JSD measures the difference between the entropy of the average distribution and the average entropy. If the divergence is zero, the distributions are identical since the JSD is the square of a metric (Endres and Schindelin 2003; Briët and Harremoës 2009). The divergence is moreover bounded,

$$0 \leq \text{JSD}(\pi_1, \ldots, \pi_N) \leq \log_2(N) \tag{34}$$

so the normalised divergence, between 0 and 1, is obtained by dividing by $\log_2(N)$.

Bayesian updating and lateral inhibition

Intuitively, Bayesian updating implements a kind of lateral inhibition — but how exactly? We derive the 'update' rules in the Dirichlet-categorical naming game. Recall that every word is assigned a score $s(x) = p(x \mid \alpha)$. The question is how score of word y changes after observing x. That is, what is $s_{t+1}(x)$ in terms of $s_t(x)$? For a sampler,

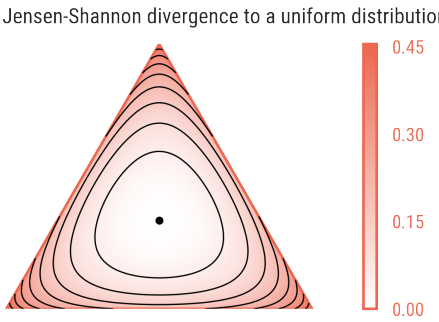

FIGURE 2 The divergence between distributions in the 2-simplex and the uniform distribution $(1/3, 1/3, 1/3)$ (indicated by a dot) under the Jensen-Shannon divergence. Points on the solid lines have the same distance to the uniform

FIG02 Figure inspired by a blogpost of Lior Pachter liorpachter.wordpress.com/tag/jensen-shannon-metric/

this follows directly from eq. 21:

$$s_{t+1}(y) = p(y \mid x, \alpha) = \frac{\alpha_y + [\![y = x]\!]}{\Sigma(\alpha) + 1} \tag{35}$$

$$= \frac{\alpha_y}{\Sigma(\alpha)} \cdot \frac{\Sigma(\alpha)}{\Sigma(\alpha) + 1} + \frac{[\![y = x]\!]}{\Sigma(\alpha) + 1} \tag{36}$$

$$= s_t(y) \cdot \frac{\Sigma(\alpha)}{\Sigma(\alpha) + 1} + \frac{[\![y = x]\!]}{\Sigma(\alpha) + 1} \tag{37}$$

For the MAP language strategy ($\eta = \infty$), the agent always chooses the mode ν of the distribution Dirichlet(α), i.e.,

$$\nu = \frac{\alpha - 1}{\Sigma(\alpha) - K} \tag{38}$$

The score $s(y) = p(y \mid x, \alpha)$ is therefore the y'th component of the mode. A similar argument as above shows that

$$s_{t+1}(y) = s_t(y) \cdot \frac{\Sigma(\alpha) - K}{\Sigma(\alpha) - K + 1} + \frac{[\![y = x]\!]}{\Sigma(\alpha) - K + 1}. \tag{39}$$

This is a similar lateral inhibition mechanism as the one used by samplers.

Bibliography

Bishop, Christopher M. (2006). *Pattern Recognition and Machine learning*. Information Science and Statistics. Springer.

Briët, Jop and Peter Harremoës (2009). "Properties of classical and quantum Jensen-Shannon divergence". In: *Physical Review A* 79.5, p. 052311. DOI: 10.1103/PhysRevA.79.052311. arXiv: 0806.4472.

Endres, D.M. and J.E. Schindelin (2003). "A new metric for probability distributions". In: *IEEE Transactions on Information Theory* 49.7, pp. 1858–1860. DOI: 10.1109/TIT.2003.813506.