Maximum Likelihood Estimation of Dirichlet Distribution Parameters

Jonathan Huang

Abstract. Dirichlet distributions are commonly used as priors over proportional data. In this paper, I will introduce this distribution, discuss why it is useful, and compare implementations of 4 different methods for estimating its parameters from observed data.

1. Introduction

The Dirichlet distribution is one that has often been turned to in Bayesian statistical inference as a convenient prior distribution to place over proportional data. To properly motivate its study, we will begin with a simple coin toss example, where the task will be to find a suitable distribution P which summarizes our beliefs about the probability that the toss will result in heads, based on all prior such experiments.

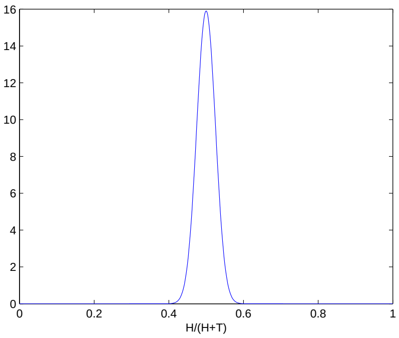

Figure 1. A distribution over possible probabilities of obtaining heads

We will want to convey several things via such a distribution. First, if we have an idea of what the odds of heads are, then we will want P to reflect this. For example, if we associate P with the experiment of flipping a penny, we would hope that P gives strong probability to 50-50 odds. Second, we will want the distribution to somehow reflect confidence by expressing how many coin flips we have witnessed in the past, the idea being that the more coin flips one has seen, the more confident one is about how a coin must behave. In the case where we have never seen a coin flip experiment, then P should assign uniform probability to all odds. On the other hand, if we have seen many experiments before, then we will have a good idea of what the odds are, and P will be strongly peaked at this value.

Figure 1 shows one possibility for P where probability density is plotted against probability of flipping heads. Here, the prior belief is fairly certain that the odds of obtaining heads is about 50-50. The form of the distribution for this particular graph is given by:

$$p(x) \propto x^{199} \left(1 - x\right)^{199}$$

and is an example of the so-called *beta distribution*.

2. The Dirichlet Distribution

This section will show that a generalization of the beta distribution to higher dimensions leads to the Dirichlet. In the coin toss example, we only considered the odds of getting heads (or tails) and placed a distribution on these odds. An m-dimensional Dirichlet will be defined as a distribution over *multinomials*, which are m-tuples $\mathbf{p} = (p_1, \ldots, p_m)$ that sum to unity. For the two dimensional case, this is just pairs (H, T) such that $H + T = 1$. The space of all m-dimensional multinomials is an $(m-1)$-simplex by definition, and so the Dirichlet distribution can also be thought of as a distribution over a simplex.

Algebraically, the distribution is given by

$$Dir(\mathbf{p}|\alpha_1, \ldots, \alpha_m) = \frac{1}{Z} \prod_k p_k^{\alpha_k - 1}$$

where $Z = \frac{\prod_{k=1}^m \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^m \alpha_k\right)}$ is a normalization factor. [1] There are m parameters α_k which are assumed to be positive. Figure 2 plots several examples of a three-dimensional Dirichlet.

Yet another way to think about the Dirichlet distribution is in terms of measures. Essentially, a Dirichlet is a measure over the space of all measures over a set of m elements. This is interesting because the idea can be extended in a rigorous way to the concept of *Dirichlet processes*, which are measures over measures on more general sets. The Dirichlet process is, in some sense, an infinite dimensional version of the Dirichlet distribution. This is a useful prior to put over mixing weights of a Gaussian mixture model and is used for automatically picking out the number of necessary clusters as opposed to the approach of trying to fit the data several times to different numbers of clusters to find the best number [**4**].

2.1. An Intuitive Reparameterization.

A simple reparameterization of the Dirichlet is given by setting:

$$s = \sum_{k=1}^m \alpha_k$$

[1] $\Gamma(x)$ denotes the Gamma function and is defined to be: $\int_0^\infty t^{x-1} e^{-t} dt$. Integrating this parts gives the functional definition: $\Gamma(x+1) = x\Gamma(x)$. Since $\Gamma(1) = 1$, we see that this function satsifies $\Gamma(n+1) = n!$ for $n \in \mathbb{N}$ and is a generalization of the factorial to the real line.

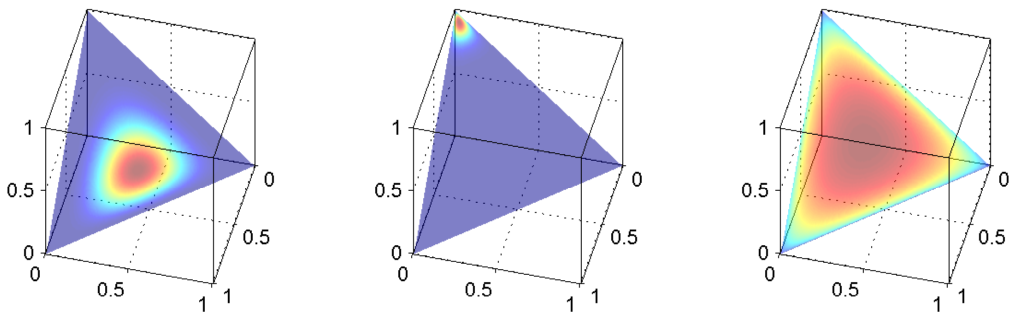

FIGURE 2

and

$$\mathbf{m} = \left(\frac{\alpha_1}{s}, \ldots, \frac{\alpha_m}{s}\right)$$

The vector $\mathbf{m}$ sums to unity and hence is a point on the simplex. It turns out to be exactly the mean of the Dirichlet distribution. s is commonly referred to as the *precision* of the Dirichlet (and sometimes as the *concentration parameter*) and as its name implies, controls how concentrated the distribution is around its mean. For example, on the right hand side of Figure 2, s is small and hence yields a diffuse distribution, whereas the center plot on Figure 2 has a large s and is hence concentrated tightly about the mean. As will be discussed later, it is sometimes useful to estimate mean independently of precision or vice-versa.

2.2. The Exponential Family. It is illuminating to study the Dirichlet as a special case of a larger class of distributions called the *exponential family*, which is defined to be all distributions which can be written as

$$p(x|\eta) = h(x)\exp\{\eta^T T(x) - A(\eta)\}$$

where η is called the *natural* or *canonical parameter*, $T(x)$ the *sufficient statistic*, and $A(\eta)$ the *log normalizer*. Some common distributions which belong to this family are the Gaussian, Bernoulli and Multinomial distributions. It is easy to see that the Dirichlet also takes this form by writing:

$$\begin{aligned} h(x) &= 1 \\ \eta &= \alpha - 1 \\ T(x) &= \log p \\ A(\eta) &= N\left(\sum_k \log\Gamma(\alpha_k) - \log\Gamma\left(\sum_k \alpha_k\right)\right) \end{aligned}$$

Besides being well understood, there are several reasons why distributions from this family are commonly employed in statistics. As shown by the Pitman-Koopman-Darmois theorem, it is only in this family that the dimension of the sufficient statistic is bounded even as the number of samples goes to infinity. This leads to efficient point estimation methods.

Bayesians are particularly indebted to the exponential family due to the fact that if a likelihood function belongs to it, then a *conjugate prior* must exist. [2] Existence of such a prior simplifies computations immensely and the lack of one often requires one to resort to numerical techniques for estimating a posterior.

A final noteworthy point is that $A(\eta)$ is the cumulant generating function for the sufficient statistic, so in particular, $A'(\eta)$ is the expectation, and $A''(\eta)$ is the variance. This implies that A is convex, which further implies that the log-likelihood function of data drawn from these distributions is convex in η.

2.3. The Dirichlet as a Prior.

The most common reason for using a Dirichlet distribution is as a prior on the parameters to a multinomial distribution. The multinomial distribution also happens to be a member of the exponential family, and accordingly, has an associated conjugate prior. The multinomial distribution is a generalization of the binomial distribution and is defined over m-tuples of "counts", which are just nonnegative integers:

$$Mult(x|\theta) = \frac{(\sum_k x_k)!}{\prod_{k=1}^m (x_k!)} \prod_{k=1}^m \theta_k^{x_k}$$

where the parameters θ are probabilities of falling into one of m classes and hence θ is a point on an $(m-1)$-simplex. It is not difficult to explicitly show that the Multinomial and Dirichlet distributions form a conjugate prior pair:

$$\begin{aligned} p(x|\theta)p(\theta) &= Mult(x|\theta)Dir(\theta|\alpha) \\ &\sim \prod_{k=1}^m \theta_k^{x_k} \prod_{k=1}^m \theta_k^{\alpha_k - 1} \\ &\sim \prod_k \theta^{x_k + \alpha_k - 1} \\ &= Dir(x+\alpha) \end{aligned}$$

The last line follows by observing that the posterior is a distribution, so when normalized, must yield an actual Dirichlet. What is very nice about this expression is that it mathematically formalizes the intuition that the parameters to the prior, α, can be thought of as pseudocounts. Going back to the two dimensional case, we see that α encodes a tally of the results of all prior coin flips.

3. Estimating Parameters

Given a set of observed multinomial data, $\mathcal{D} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N\}$, the parameters for a Dirichlet distribution can be estimated by maximizing the log-likelihood

[2] A conjugate prior for a likelihood function is defined to be a prior for which posterior and prior are of the same distribution type.

function of the data, which is given by:

$$\begin{aligned} F(\alpha) = \log p(D|\alpha) &= \log \prod_i p(\mathbf{p}_i|\alpha) \\ &= \log \prod_i \frac{\Gamma\left(\sum_k \alpha_k\right)}{\prod_k \Gamma(\alpha_k)} \prod_k p_{ik}^{\alpha_k - 1} \\ &= N \left(\log \Gamma \left(\sum_k \alpha_k \right) - \sum_k \log \Gamma\left(\alpha_k\right) + \sum_k (\alpha_k - 1) \log \hat{p}_k \right) \end{aligned}$$

where $\log \hat{p}_k = \frac{1}{N} \sum_i \log p_{ik}$ and are the *observed sufficient statistics.*

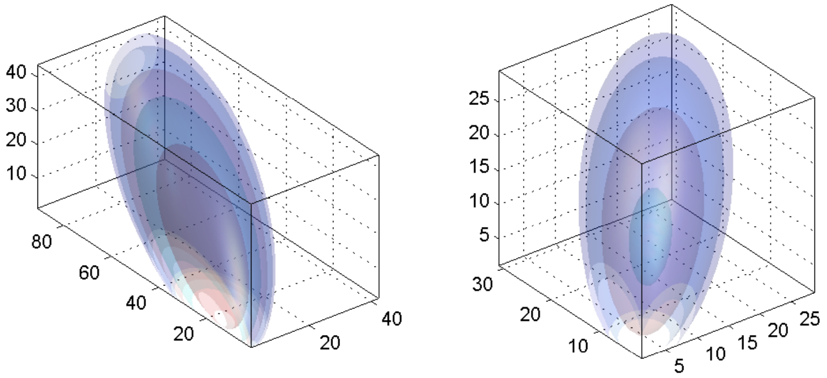

FIGURE 3. Examples of log-likelihood functions of a three dimensional Dirichlet

The following sections will provide an overview of several methods for numerically maximizing this objective function, F as there is no closed form solution to this. As discussed above, they will all use the fact that the log-likelihood is convex in α to guarantee a unique optimum.

3.1. Gradient Ascent. The first method to try is Gradient Ascent, which iteratively steps along positive gradient directions of F until convergence. The gradient of the objective is given by differentiating F:

$$(\nabla F)_k = \frac{\partial F}{\partial \alpha_k} = N \left(\Psi \left(\sum_k \alpha_k \right) - \Psi(\alpha_k) + \log \hat{p}_k \right)$$

where $\Psi = \frac{d \log \Gamma(x)}{dx}$ is the *digamma* function. There is no analytic expression for doing a line search; one can always continue to step along a constant fraction of the gradient, but care must be taken that the constraints of the problem be enforced (e.g. the α_k must always be positive.)

3.2. A Fixed Point Iteration. Minka [**1**] provides a convergent fixed point iteration technique for estimating parameters. The idea behind this is to guess an initial α, find a function that bounds F from below which is tight at α, then to optimize this function to arrive at a new guess at α.

There are many inequalities associated to the ratio $\frac{\Gamma(x+\beta)}{\Gamma(x)}$ which have been extensively studied by many mathematicians ([**5**],[**6**],[**8**]). One commonly cited one is:

$$\Gamma(x) \geq \Gamma(\hat{x}) \exp((x-\hat{x})\Psi(\hat{x}))$$

which leads to a lower bound on the log likelihood, $F(\alpha)$:

$$F(\alpha) \geq N\left(\left(\sum_k \alpha_k\right) \Psi\left(\sum_k \alpha_k^{old}\right) - \sum_k \log\Gamma(\alpha_k) + \sum_k \alpha_k \log \hat{p}_k + C\right)$$

where C is a constant with respect to α. Now this expression is maximized by setting the gradient to zero and solving for α. The update step is given by:

$$\alpha_k^{new} = \Psi^{-1}\left(\Psi\left(\sum_k \alpha_k^{old}\right) + \log \hat{p}_k\right)$$

The digamma function Ψ can be inverted efficiently by using a Newton-Raphson update procedure to solve $\Psi(x) = y$.

3.3. The Newton-Raphson Method.

Newton-Raphson provides a quadratically converging method for parameter estimation. The general update rule can be written as:

$$\alpha^{new} = \alpha^{old} - H^{-1}(F) \cdot \nabla F$$

where H is the Hessian matrix.

For this particular log likelihood function, there is no problem applying Newton-Raphson to high dimensional data, because the inverse of the Hessian matrix can be computed in linear time. In particular, the Hessian of F is the sum of a matrix whose elements are all the same and a diagonal matrix. It is given by:

$$\frac{\partial^2 F}{\partial \alpha_k^2} = N\left(\Psi'\left(\sum_k \alpha_k\right) - \Psi'(\alpha_k)\right)$$

$$\frac{\partial^2 F}{\partial \alpha_j \partial \alpha_k} = N\Psi'\left(\sum_k \alpha_k\right)$$

We can rewrite this as

$$\begin{aligned} H &= Q + c11^T \\ q_{jk} &= -N\Psi'(\alpha_k)\delta(j-k) \\ c &= N\Psi'\left(\sum_k \alpha_k\right) \end{aligned}$$

To invert the Hessian, we observe that for any invertible matrix Q and non-zero scalar c:

$$\begin{aligned}\left(Q+c11^T\right)\left(Q^{-1}-\frac{Q^{-1}11^TQ^{-1}}{1/c+1^TQ^{-1}1}\right) &= QQ^{-1}-\frac{QQ^{-1}11^TQ^{-1}}{1/c+1^TQ^{-1}1}+c11^TQ^{-1}\\ &\quad -\frac{c11^TQ^{-1}11^TQ^{-1}}{1/c+1^TQ^{-1}1}\\ &= QQ^{-1}+\frac{1}{1/c+1^TQ^{-1}1}\left(-11^TQ^{-1}+11^TQ^{-1}\right.\\ &\quad \left.+c11^TQ^{-1}(1^TQ^{-1}1)-c1(1^TQ^{-1}1)1^TQ^{-1}\right)\\ &= 1\end{aligned}$$

Since Q is diagonal, Q^{-1} is easily computed and the update rule for Newton-Raphson can be rewritten in terms of each coordinate:

$$\alpha_k^{new}=\alpha_k^{old}-\frac{(\nabla F)_k-b}{q_{kk}}$$

where $b=\frac{Q^{-1}11^TQ^{-1}}{1/c+1^TQ^{-1}1}=\frac{\sum_j(\nabla F)_j/q_{jj}}{1/z+\sum_j 1/q_{jj}}$

3.4. Estimating Mean and Precision Separately. The fourth way for estimating a Dirichlet is to estimate mean and precision separately leaving the other fixed. Sometimes it may be enough to just know one of these parameters, but if all of them are desired, then one can alternate between estimating mean and precision (as would be done in a coordinate ascent method) until convergence.

3.4.1. *Mean.* First consider estimating the mean $\mathbf{m}$ with a fixed precision s. The likelihood for $\mathbf{m}$ is

$$p(D|\mathbf{m})\propto\left(\frac{\exp(sm_k\log\hat{p}_k)}{\Gamma(sm_k)}\right)^N$$

We now reparametrize this by an unconstrained vector z which is defined by $\frac{z_k}{\sum_k z_k}$ and the log-likelihood function is now rewritten as:

$$\log p(D|\mathbf{m})=N\sum_k\left[\frac{z_k}{\sum_k z_k}\log\hat{p}_k-\log\Gamma\left(s\frac{z_k}{\sum_k z_k}\right)\right]$$

Differentiate to obtain a gradient which can be used in a gradient ascent update rule:

$$\begin{aligned}\frac{d\log p(D|\mathbf{m})}{dz_i} &= N\sum_k\left[\frac{\sum_k z_k-z_i}{\left(\sum_k z_k\right)^2}s\log\hat{p}_k-s\left(\frac{\sum_k z_k-z_i}{\left(\sum_k z_k\right)^2}\right)\Psi\left(s\frac{z_k}{\sum_k z_k}\right)\right]\\ &= \frac{Ns}{\sum_k z_k}\left(\log\hat{p}_k-\Psi(sm_k)-\sum_k m_k(\log\hat{p}_k-\Psi(sm_k))\right)\end{aligned}$$

An alternative would be the following fixed point update which converges very rapidly:

$$\Psi(\alpha_k)=\log\hat{p}_k-\sum_k m_k^{old}(\log\hat{p}_k-\Psi(sm_k^{old}))$$

$$m_k^{new}=\frac{\alpha_k}{\sum_k\alpha_k}$$

3.4.2. *Precision.* We now estimate the precision for a fixed mean vector. The appropriate likelihood function here is:

$$p(D|s) \propto \left(\frac{\Gamma(s) \exp\left(s \sum_k m_k \log \hat{p}_k\right)}{\prod_k \Gamma(sm_k)} \right)^N$$

And the first and second derivatives of the log-likelihood are given by:

$$\frac{d \log p(D|s)}{ds} = N \left(\Psi(s) - \sum_k m_k \left(\Psi(sm_k) + \log \hat{p}_k \right) \right)$$

$$\frac{d^2 \log p(D|s)}{ds^2} = N \left(\Psi'(s) - \sum_k m_k^2 \Psi'(sm_k) \right)$$

[**2**] provides a Generalized Newton iteration for maximizing this function, [3] which yields an update rule which looks a lot like a Newton-Raphson update, but has faster convergence:

$$\frac{1}{s^{new}} = \frac{1}{s} + \frac{1}{s^2} \left(\frac{d^2 \log p(D|s)}{ds^2} \right)^{-1} \left(\frac{d \log p(D|s)}{ds} \right)$$

4. Results

To compare the four methods, I implemented each one in C along with routines for random sampling from a Dirichlet. [4]

To first test that the methods worked, 100,000 multinomials were drawn from a Dirichlet with known parameters, and the output of each method was compared to the ground truth. To compare speeds, I repeated this process with 10000 multinomials, 50 trials for each method and recorded averaged times to run this test (summarized in the figure). The algorithms were deemded to have converged when the step sizes dipped below 10^{-9}. They show that the Newton-Raphson method and the method of alternating mean/precision estimations were the fastest on average, that the methods scale approximately linearly according to dimension and precision.

A nontrivial issue in implementation is that of enforcing the inequality constraints that $\alpha_k > 0 \ \forall k$. With the fixed point iteration, this was never an issue, but the other three methods were prone to being carried out of bounds and had to be brought back inside, the worst of the three being Newton-Raphson. In my code, I simply check for this at each iteration, but another idea for future work will be to place a log barrier function at these functions [**9**]. For several of the methods, there are several clever methods for initalizing the iteration by approximating the log likelihood by a simpler function. Using these often alleviates the issues of outstepping the bounds since they encourage the algorithms to converge in fewer steps.

[3] The idea behind the Newton method is to approximate a function locally by a quadratic by matching first and second derivatives, and optimizing this quadratic instead. In the generalized version of Newton, the idea is to approximate by a simpler function (not necessarily a quadratic) by matching the first and second derivatives and optimizing said simpler function.

[4] Sampling from an m-dimensional Dirichlet amounts to sampling from m different Gamma distributions with parameters depending on each α_k and then projecting the vector of these concatenated samples onto the simplex. In my implementation, I use a rejection sampling method for the Gamma distribution [**7**].

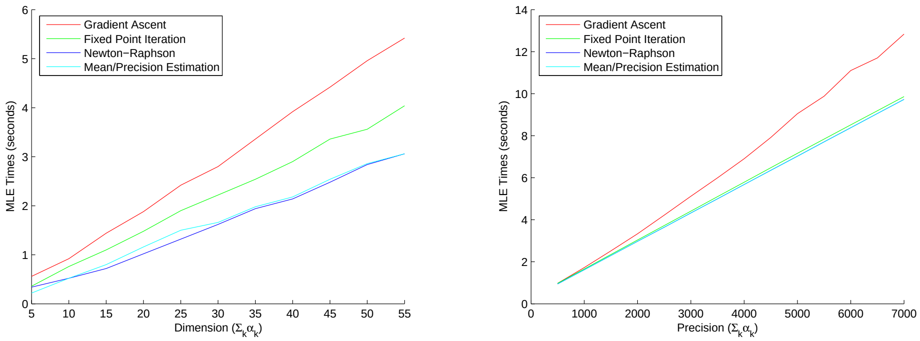

FIGURE 4. Times to estimate 50 Dirichlets plotted against dimension (a) and precision (b).

5. Conclusions

The example that motivated this project comes from latent semantic analysis in text modeling. In a commonly cited model, the Latent Dirichlet Allocation model [**3**], a Dirichlet prior is incorporated into a generative model for a text corpus, where every multinomial drawn from it represents how a document is mixed in terms of topics. For example, a document might spend 1/3 of its words discussing statistics, 1/2 on numerical methods, and 1/6 on algebraic topology $\left(\frac{1}{3}+\frac{1}{2}+\frac{1}{6}=1\right)$. For a large number of possible topics, fast maximum likelihood methods which work well for high dimensional data are essential, and I reviewed some alternatives to Gradient Ascent in this paper. Due to several important properties such as having sufficient statistics of bounded dimension, and a convex log-likelihood function this computation can be made quite efficient.

References

[1] T. Minka, *Estimating a Dirichlet Distribution*, (2000).
[2] T. Minka, *Beyond Newton's Method*, (2000).
[3] D. Blei, A. Ng, and M. Jordan, *Latent Dirichlet Allocation*, Journal of Machine Learning Research, 3:993-1022, (2003).
[4] D. Blei, T. Griffiths, M. Jordan, and J. Tenenbaum, *Hierarchical Topic Models and the Nested Chinese Restaurant Process*, Advances in Neural Information Processing Systems (NIPS) 16, Cambridge, MA, (2004). MIT Press.
[5] B.N. Guo and F. Qi, *Inequalities and Monotonicity for the Ratio of Gamma Functions*, Taiwanese Journal of Mathematics, Vol 19, No. 7. pp. 407-409. (1976).
[6] S.S. Dragomir, R.P. Agarwal, and N. Barnett, *Inequalities for Beta and Gamma Functions via some Classical and New Integral Inequalities*, Journal of Inequalities and Applications, (1999).
[7] G. Fishman, *Sampling from the Gamma Distribution on a Computer*, ACM Communications. Vol 19, No. 7. pp. 407-409. (1976).
[8] Milan Merkle, *Conditions for Convexity of a Derivative and Applications to the Gamma and Digamma Function*, Serb. Math. Inform. 16, pp. 13-20. (2001).
[9] S. Boyd and L. Vandenberghe, *Convex Optimization*, (2004).

Current address: Robotics Institute, Carnegie Mellon University
E-mail address: jch1@cs.cmu.edu