Learning Rotations Online

Adam Smith (amsmith@cs.ucsc.edu)

June 11, 2008

Abstract

In this paper we show that the matrix von Mises-Fisher (vMF) distribution is a reasonable distribution upon which to build an online density estimation scheme for rotation matrices. We also consider a special case, the unit circle, for initial experimentation. The vector and matrix vMF distributions admit online algorithms in terms of their expectation parameters as a direct result of their exponential family nature, however proving regret bounds for these algorithms is arrested by the complexity of the cumulant functions for these distributions. In closing, we volunteer hazards and suggestions for immediate future work.

1 Introduction

The aim of this work is twofold. First, we aim to rigorously treat the machine learning problem of online density estimation for rotation matrices. Second, we aim to uncover general principles that lend intuition to the treatment of other matrix groups on compact spaces.

The primary motivation for this work is that rotation matrices are largely ignored in machine learning and we would like to change this. Secondarily, solutions to problems in several application domains could be greatly improved by proper treatment of rotations. For example, in computer vision, the problem of pose estimation involves determining the position and orientation of physical objects from noisy measurements. Solutions to this problem are often based on kernel methods which, while simplifying calculation via linearization, ignore the compact geometry of the space of orientations. In the realm of directional statistics, where distributions respect the geometry of space rotations, many problems are solved only by asymptotic analysis or other approximations and there are certainly no online algorithms with known regret bounds. Finally, a major payoff for intuition gained from working with rotation matrices could be had from treating members of the related balanced orthogonal group which can be used as a basis for representing many more matrix groups of interest by an embedding [DHSA93].

Informally, the problem we wish to solve is to predict the best rotation, in an online manner, given a stream of noisy orientation (rotation) estimates. This should be done in

such a way as to minimize the total regret of the algorithm with respect to the best fixed rotation chosen offline (with access to future observations).

Our approach is to adapt the online density estimation process from machine learning to the matrix von Mises-Fisher (vMF) distribution from directional statistics, and identify the space of rotation matrices with the space of n-frames. We will focus on the simplest case possible, that of the unit circle S^1 or $SO(2)$, a well understood compact Lie group.

2 Related Work

A rigorous treatment of online density estimation should find a foundation in existing machine learning theory. Much of the online learning machinery required has already been worked in general for distributions from the exponential family [AW01]. We will use this work as a template later on. Distributions on compact Lie groups have previously been considered in machine learning, however the general theory developed does not immediately produce distributions in the exponential family [Pen04].

On the other hand, exponential family distributions for spaces like that of the rotations have long been known in statistics. Downs introduced "orientation statistics" which deals with vMF distributions on orientations in the generalized case of Stiefel manifolds (of which $SO(n)$ is one) [Dow72]. The primary distribution we borrow from this work is the vMF distribution, for which maximum likelihood estimates are known to exist and are unique [JM79]. Finally, recent work by Chikuse mentions direct representation of rotations, several application areas for the vMF and related distributions as well as the relation to Procrustes analysis, a related problem which is normally not addressed probabilistically [Chi03].

3 von Mises-Fisher Distributions

The general matrix vMF distribution is defined over the Stiefel manifold $O(n, p)$ of $n \times p$ matrices such that for $\mathbf{X} \in O(n, p)$, $\mathbf{X}\mathbf{X}^T = \mathbf{I}_n$ (the space of orthonormal p-frames). The density of this distribution was given by Downs in the following manner for elements $\mathbf{X} \in O(n, p)$ and an $n \times p$ matrix parameter $\mathbf{F}$ [Dow72].

$$p(\mathbf{X}) = a(\mathbf{F}) \exp\left(\mathrm{tr}(\mathbf{F}\mathbf{X}')\right) \tag{1}$$

For the case of the unit circle, we can consider a vectorial representation of rotations that makes the notion of the "length" and "direction" of the distribution's parameter clear. We expect the intuition developed here to extend to higher dimensions by way of the polar decomposition for general $\mathbf{F}$. For now, we will work with the vector vMF distribution on elements $\mathbf{x}$ of the unit circle with the density for parameter vector θ given as follows.

$$p(\mathbf{x}) = \exp\left(\theta \cdot \mathbf{x} - G(\theta)\right) \tag{2}$$

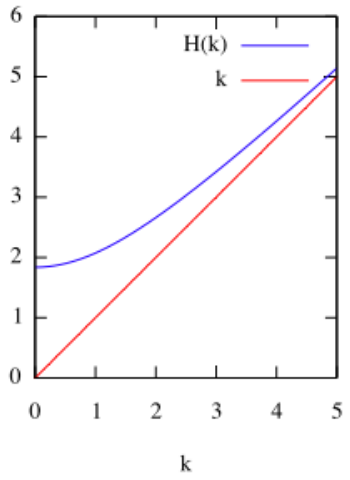

Figure 1: The cumulant of the vector vMF distribution on the unit circle as a function of concentration parameter k becomes linear for large k.

Clearly, this is an exponential family distribution. The parameter θ is a "natural parameter" and may take values from all of $\mathbb{R}^2$ (despite data living on the unit circle). Traditionally the parameter is factored into a scalar concentration parameter and modal direction (unit vector) such that $\theta = k\mathbf{d}$.

The cumulant function of this distribution is rotationally symmetric about the origin, depending only on the concentration parameter: $G(k\mathbf{d}) = H(k) = \log(2\pi I_0(k))$ where I_ν is the modified Bessel function of the first kind (order ν). As Figure 1 illustrates, the cumulant function begins quadratically at $\log(2\pi)$ but quickly approaches the line of unit slope as k grows.

It is well known that the cumulant function can be used to compute the expectation parameter from the natural parameter for a distribution. Consider the gradient of $G(\theta)$ with $h(k) = \frac{d}{dk}H(k)$.

$$
\begin{aligned}
g(\theta) &= \nabla_\theta G(\theta) && (3)\\
&= \nabla_\theta H(k) && (4)\\
&= h(k)\nabla_\theta k && (5)\\
&= h(k)\mathbf{d} && (6)\\
&= m\mathbf{d} && (7)\\
&= \mu && (8)
\end{aligned}
$$

From this we know that the expectation parameter shares the directional component of the natural parameter, however instead of roaming all of $\mathbb{R}^2$, the expectation parameter lives only within the unit disk. This restriction is apparent in the form of $h(k) = I_1(k)/I_0(k)$ and is illustrated in Figure 2. Intuitively, the average of several distinct points on the unit circle will live closer to the origin than any of the points.

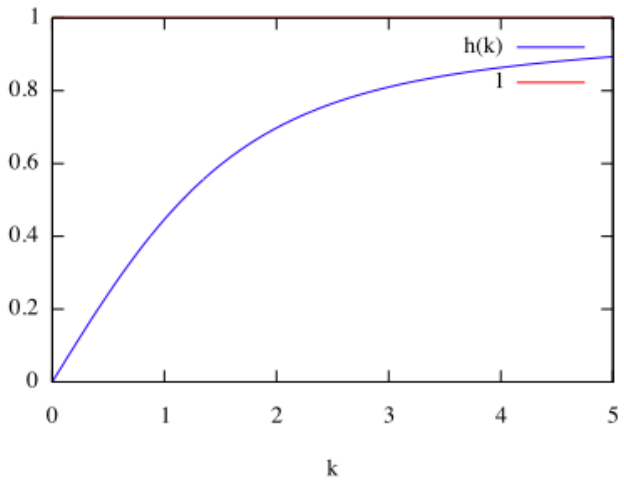

Figure 2: The length of the expectation parameter approaches unity (corresponding to the boundary of the unit disk) as k grows.

Armed with an understanding of the vector vMF distribution as an exponential family distribution, we can procede with the derivation of an online density estimation algorithm.

4 Divergence

The first step in developing an online density estimation algorithm for our case is to identity a divergence between two distribution parameters that can be used in optimization. A Bregman divergence can easily be constructed for an exponential family distribution following previous work [AW01]. Recall the canonical form for Bregman divergences.

$$\Delta_G(\tilde{\theta}, \theta) = G(\tilde{\theta}) - G(\theta) - (\tilde{\theta} - \theta) \cdot \nabla_\theta G(\theta) \tag{9}$$

In our case, we make take $G(\theta)$ to be the cumulant function from the vector vMF distribution (as it is convex and defined for all $\mathbb{R}^2$). Plugging in the definitions, we have the following divergence in terms of the factored parameters.

$$\Delta(\tilde{k}\tilde{\mathbf{d}}, k\mathbf{d}) = H(\tilde{k}) - H(k) - (\tilde{k}\tilde{\mathbf{d}} - k\mathbf{d}) \cdot h(k)\mathbf{d} \tag{10}$$

The length and direction dependence of the divergence are illustrated in Figures 3 and 4. With the reader's imagination, the divergence between arbitrary parameters can be seen as a cone with a softened point that is always pointed at the reference parameter, but is tipped toward the modal direction of the the reference parameter with the angle of tipping controlled by the concentration. The cone-nature of the divergence comes from the asymptotically linear behavior of $G(\theta)$ and the soft tip comes from the quadratic behavior

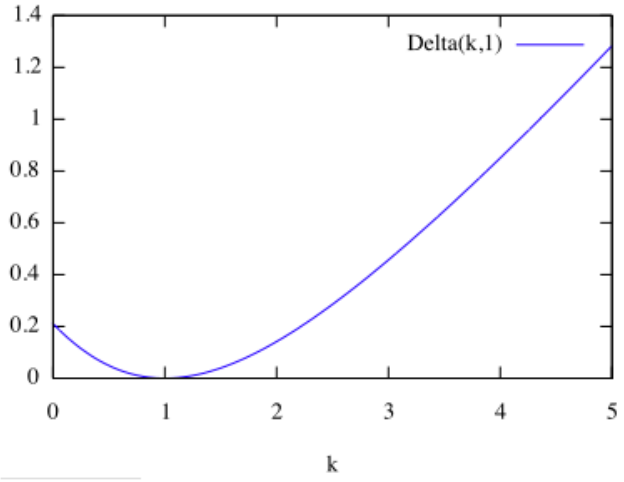

Figure 3: The divergence between two distributions with a common modal direction, the first with arbitrary $\tilde{k}$ and the second with $k = 1$, is convex in $\tilde{k}$.

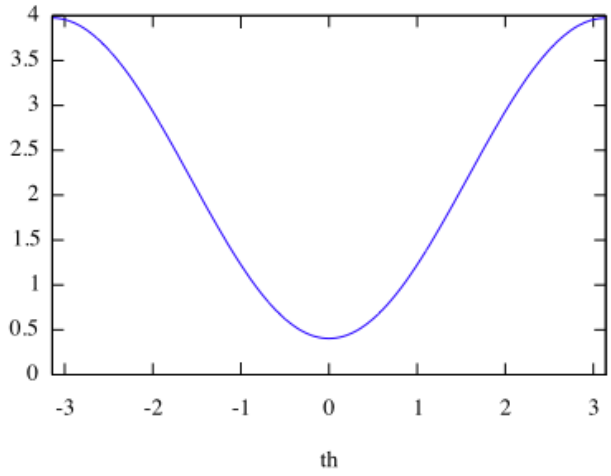

Figure 4: The divergence between two distributions that vary in modal direction, the first with $\tilde{k} = 4$ and the second with $k = 1$, is a shifted cosine with respect to the angle between the modal directions. Note that the divergence with zero angular difference agrees with the divergence for $\tilde{k} = 4$ in Figure 3.

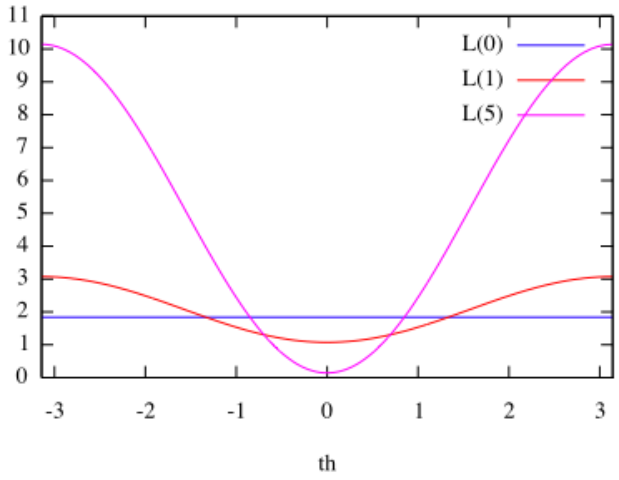

Figure 5: The loss of a parameter on a particular example, as a function of angular difference between the example and the modal direction, is flat for the uniform distribution but approaches zero for the case of no angular difference and diverges elsewhere as the concentration grows.

near $k = 0$. For the most part, the divergence encodes a tradeoff between accuracy of the length and direction, with a preference for accuracy of the direction as the length increases.

Clearly, for a given length, the worst case loss divergence occurs when the two parameters are opposite one another on the unit circle. For a given angular difference in direction, the worst case divergence occurs when the length is as large as possible. The implications of this behavior will be critical in the derivation for regret bounds in a later section.

5 Loss

This exponential family distribution comes with a natural loss function: the negative log-likelihood. The loss of the parameter θ on example $\mathbf{x}$ is given by

$$\begin{aligned} L_{\mathbf{x}}(\theta) &= -\log P(\mathbf{x}|\theta) && (11)\\ &= G(\theta) - \theta \cdot \mathbf{x} && (12)\\ &= \log(2\pi I_0(||\theta||)) - \theta \cdot \mathbf{x}. && (13) \end{aligned}$$

The behavior of the loss for parameters with varying length and direction is illustrated in Figure 5. It is worth noting the loss *is* convex on the parameter space, however it appears periodic in the figure only because the curves illustrate the loss over the non-convex sets of parameters on circles of varying radii.

6 Algorithm

With a divergence and a loss defined, the online algorithm for density estimation on the vector vMF of the unit circle is straightforward. To begin, we need an initial parameter. Setting $\theta_0 = 0$ is a suitable starting point as it uniquely identifies the uniform distribution (with $\mu_0 = 0$ accordingly). From the general results for exponential family distributions, we may immediately write down the update of the algorithm as follows, assuming a variable learning rate of $\eta_t^{-1} = \eta_0^{-1} + t$.

$$\mu_{t+1} = \mu_t - \eta_t(\mu_t - \mathbf{x}_t) \tag{14}$$

This update can be seen as forming a convex combination of the old expectation parameter and the most recent example, and it will clearly never produce a μ_t outside of the unit disk. The current expectation parameter $\mu_t = m_t\mathbf{d}_t = h(k_t)\mathbf{d}_t = g(\theta)$ can be readily computed from the current natural parameter, however it is quite difficult to find the inverse after the update: $\theta_{t+1} = f(\mu_{t+1}) = g^{-1}(\mu_{t+1})$. The difficulty in this computation was first brought to light over 100 years ago in Pearson's problem of the random walk which deals with resultant of fixed length steps in the plane with direction chosen at random [Pea05]. Discouragingly, the form of $G(\theta)$ becomes more complicated as the dimensionality of the problem increases (though asymptotic forms are known).

In exchange for the complexity of $G(\theta)$ the update in equation 14 comes with unexpected generality. The update handles changes to both the modal direction *and* concentration of the estimated distribution, whereas common density estimation for Gaussian distributions assumes a known, fixed covariance.

7 Regret Bounds

The online algorithm is simple enough, but proving that it cannot be fooled is quite a different story. Based on general results for exponential family distributions, we know that a bound on the total regret of our algorithm with respect to the best offline parameter hinges on an expression like the following.

$$L_{\text{algorithm}} - L_{\text{best}} = \sum_{t=1}^{T} \eta_t \Delta(\theta_t, \theta_{t+1}) \tag{15}$$

Our aim is to replace the divergence between successive parameters by a constant upper bound. If we can do this, the $\mathrm{O}(1/t)$-style learning rate ensures that the total regret of our algorithm will have a bound that grows only logarithmically in the total number of trials T.

To bound the divergence in question, recall the worst-case setting. We should consider the result of parameter update after seeing a single example opposite a run of examples all in the same direction. In terms of the length of the expectation parameter in the update, the

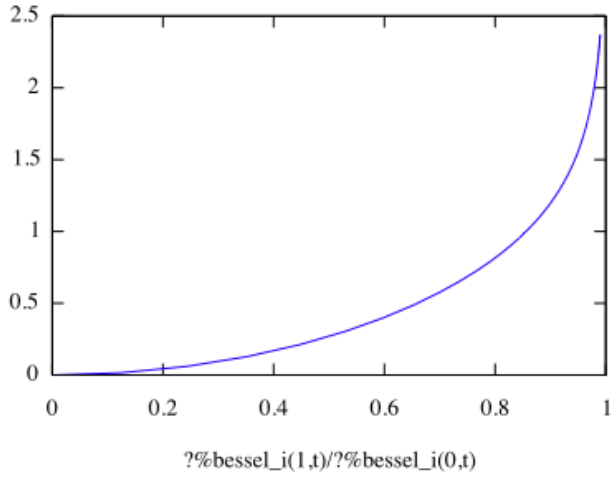

Figure 6: The dual function or convex conjugate $F(\mu)$ is convex on the expectation space with the m-dependence shown above. The slope of this graph is given by the transpose of the curve in Figure 2.

worst case occurs with $m_t = \frac{t-1}{t}$ ($T-1$ observations of $\mathbf{x}_t = \mathbf{d}$) followed by $m_{t+1} = \frac{t-2}{t+1}$ (a final observation of $\mathbf{x}'_t = -\mathbf{d}$). Ideally, these parameters could just be plugged into the definition of Δ, however we need to know the natural parameters that correspond to these expectation parameters, which in turn requires a computable version of $f(\mu)$. For any particular μ_t we can approximate θ_t by any number of root finding techniques, however it is not clear what approximation will yield the desired bound.

Experimentally, the worst-case divergence appears to be an increasing function of t that levels off as t goes to infinity. Using the `find_root` function of the Maxima computer algebra system to invert $g(\theta)$, the worst-case divergence can be seen to be bounded above by $\frac{1}{4}$ for t up to at least 100, but this is no proof.

8 Future Work

One aspect of this research is at an impasse and the another is open for immediate exploration. In the first case, the difficulty in computing $f(\mu)$ halts progress in finding a regret bound by the approach used above. In the second, the task of formally generalizing this analysis to rotations in n dimensions has not been considered here.

To make progress on finding a regret bound by the approach given here, it would be necessary to find suitable approximations for various functions involved. This is more involved than one might expect because $f(\mu)$, which grows very quickly (as can be imagined by the transpose of Figure 2), is used both positively and negatively in careful balance with $H(\theta)$ which makes use of the exponential-like modified Bessel functions. Figure 6 illustrates the m-dependence of $F(\mu)$ (plotted parametrically), with a derivative of 0 when $m = 0$ and quickly sliding up to infinity as m approaches unity as expected.

Another approach is to consider a potential based bound. Intuitively, the length of the parameter (natural or expectation) serves as a natural basis for the "cost" a target parameter. If the best offline parameter has a long parameter vector, we can expect the algorithm to make several high-cost mistakes before reaching the target. However, we would like the algorithm not to be easily tricked for distributions which are close to uniform.

Regarding the task of generalization, it may seem initially straightforward to use existing results for the vector vMF distribution to generalize the analysis presented here. In fact, the general form of the cumulant function is known for n dimensions.

$$G_n(\theta) = G_n(k\mathbf{d}) = H_n(k) = \frac{k^{n/2-1}}{(2\pi)^{n/2} I_{n/2-1}(k)} \tag{16}$$

Unfortunately, the vector vMF distribution in higher dimensions tells us only about the space of 1-frames, not the n-frames needed to represent rotation. On the unit circle, 2-frames happened to have the same number of degrees of freedom (namely one) so the vectorial representation was reasonable. This should not discourage the reader, however, because the notions of length and direction can again be ascribed to parameters of the matrix vMF distribution by means of the polar decomposition. In fact, in dealing with general members of the Stiefel manifold, Downs identifies the parameter of the distribution as a product of a diagonal matrix $\mathbf{K}$ encoding a generalized length and an orthonormal matrix $\mathbf{M}$ encoding a generalized direction (now the direction of several basis vectors).

To guide exploration, we provide a few suggestions. For the representation of rotations, $\mathbf{K}$ should take the form of $k\mathbf{I}$ because there is no preferred axis of observations (as there is in the "orientation statistics" setup), so the scalar nature of $H_n(k)$ is retained from the vectorial case. This conveniently avoids the need to apply the Bessel functions to the parameter matrix directly. However, the matrix nature of the examples should manifest itself as a replacement of vector dot products by matrix trace products. Otherwise, the bulk of derivations involving Bregman divergences should follow similarly. With luck, the worst case divergence needed in regret bounds for an algorithm using this representation should also resolve into a co-linear situation where only opposing directions (whatever that may mean in the matrix case) need be considered.

Finally, a sketch other threads of the authors' research that attempt a similar goal is included in an appendix. Beyond this, good luck!

References

[AW01] K. Azoury and M. K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Journal of Machine Learning*, 43(3):211–246, June 2001. Special issue on *Theoretical Advances in On-line Learning, Game Theory and Boosting*, edited by Yoram Singer.

[Chi03] Yasuko Chickuse. *Statistics on Special Manifolds*. Springer, 2003.

[DHSA93] C. Doran, D. Hestenes, F. Sommen, and N. Van Acker. Lie groups as spin groups. *J. Math. Phys.*, 34(8):3642–3669, August 1993.

[Dow72] Thomas D. Downs. Orientation statistics. *Biometrika*, 59(3):665–676, 1972.

[JM79] P. E. Jupp and K. V. Mardia. Maximum likelihood estimators for the matrix von mises-fisher and bingham distributions. *The Annals of Statistics*, 7(3):599–606, May 1979.

[Pea05] Karl Pearson. The problem of the random walk. *Nature*, 72(1865):294, 1905.

[Pen04] Xavier Pennec. Probailities and statistics on riemannian manifolds: A geometric approach. *Rapport de recherche*, 1(5093), January 2004.

APPENDIX

A Various Approaches

A.1 Rotor Theory

(pursued during late Spring 2007 and early Winter 2008 quarters)

problem: predict a rotated version of the input vector ("regression")

parameter: a rotor from geometric algebra

loss: euclidean distance between output vectors

divergence: rotor norm, "euclidean distance between rotors"

method: constant η, optimization via geometric calculus

machine learning contribution: simple algorithm with geometric interpretation

solution: predict by applying best rotor input vector; update by lerping the old rotor and shortest rotor explaining the data (simply the geometric product of input and true output vectors)

lessons: linear combination is possible in rotor-space; approach is related to quaternionic methods in 3D

issues: no probabilistic interpretation; no regret bounds; only explainable in terms of geometric algebra

A.2 Lie Theory

(pursued during most of Winter 2008 and early Spring 2008 quarters)

problem: predict a rotation transformation ("estimation")

parameter: a rotation matrix

loss: Riemann metric (geodesic distance)

divergence: Riemann metric (geodesic distance)

method: constant η, optimization via matrix calculus

machine learning contribution: multi-value algorithm with algebraic interpretation

solution: predict with best known rotation directly; update with exp of lerp of logs of old and observed rotation (logs of rotation matrices are skew-symmetric, thus linearly combinable)

lessons: manifold of rotations shares structure with many other Lie groups; Riemann measure on rotations on circle is like angle squared (geodesic distance); the only analytic functions on the circle are constants (complex analysis)

issues: no probabilistic interpretation; no regret bounds; log is multi-valued, requires selection of "closest" log; log and exp have complicated form for balanced spaces; nondeterministic behavior when "opposite" example is observed

A.3 Probability Theory

(pursued during late Spring 2008 quarter, reported in this paper)

problem: density estimation for distribution over rotations

parameter: vMF distribution parameter (not restricted to manifold)

loss: negative log likelihood of distribution

divergence: Bregman divergence derived from cumulant of distribution

method: $O(1/t)$ η, optimization (in unit circle case) with vector calculus

machine learning contribution: implicit algorithm

solution: predict with rotation sampled from current distribution; update f of lerp of g of old parameter and observation

lessions: natural parameter lives in open, linear space; expectation space is contained inside the unit disk (or equivalent for higher dimensions); estimates exist and are unique; worst case divergence and loss occur along a common line, keep observing one rotation until the last step when the "opposite" is observed; expectation parameter is a minimally sufficient statistic

issues: nobody knows how to compute f exactly (with historical precident); impossible to ignore concentration in derivations; no regret bounds (though experiments are promising); – finally we have a clear probabilistic interpretation for the problem!