ISSN 0885-6125, Volume 81, Number 1

Volume 81, Number 1

October 2010

Machine Learning

Special Issue: Selected Papers from ECML PKDD 2010
Guest Editors: José L Balcázar, Francesco Bonchi, Aristides Gionis, and Michèle Sebag

Springer

Available online
www.springerlink.com

ISSN 0885-6125

Mach Learn (2010) 81: 99–113
DOI 10.1007/s10994-010-5203-x

A geometric view of conjugate priors

Arvind Agarwal · Hal Daumé III

Received: 30 April 2010 / Accepted: 20 June 2010 / Published online: 22 July 2010

Abstract In Bayesian machine learning, conjugate priors are popular, mostly due to mathematical convenience. In this paper, we show that there are deeper reasons for choosing a conjugate prior. Specifically, we formulate the conjugate prior in the form of Bregman divergence and show that it is the inherent geometry of conjugate priors that makes them appropriate and intuitive. This geometric interpretation allows one to view the hyperparameters of conjugate priors as the *effective* sample points, thus providing additional intuition. We use this geometric understanding of conjugate priors to derive the hyperparameters and expression of the prior used to couple the generative and discriminative components of a hybrid model for semi-supervised learning.

Keywords Bregman divergence · Conjugate prior · Exponential families · Generative models

1 Introduction

In probabilistic modeling, a practitioner typically chooses a likelihood function (model) based on her knowledge of the problem domain. With limited training data, a simple maximum likelihood (ML) estimation of the parameters of this model will lead to overfitting and poor generalization. One can regularize the model by adding a prior, but the fundamental question is: which prior? We give a turn-key answer to this problem by analyzing the underlying *geometry* of the likelihood model and suggest choosing the unique prior with the same geometry as the likelihood. This unique prior turns out to be the *conjugate* prior, in the case of the exponential family. This provides justification beyond "computational convenience" for using the conjugate prior in machine learning and data mining applications.

Editors: José L Balcázar, Francesco Bonchi, Aristides Gionis, and Michèle Sebag.

A. Agarwal (✉) · H. Daumé III
School of Computing, University of Utah, Salt Lake City, UT 84112, USA
e-mail: arvind@cs.utah.edu

H. Daumé III
e-mail: hal@cs.utah.edu

In this work, we give a geometric understanding of the maximum likelihood estimation method and a geometric argument in the favor of using conjugate priors. Empirical evidence showing the effectiveness of the conjugate priors can be found in our earlier work (Agarwal and Daumé 2009). In Sect. 4.1, first we formulate the ML estimation problem into a completely geometric problem with no explicit mention of probability distributions. We then show that this geometric problem carries a geometry that is inherent to the structure of the likelihood model. For reasons given in Sects. 4.3 and 4.4, when considering the prior, it is important that one uses the same geometry as likelihood. Using the same geometry also gives the closed-form solution for the maximum-a-posteriori (MAP) problem. We then analyze the prior using concepts borrowed from the information geometry. We show that this geometry induces the *Fisher information metric* and *1-connection*, which are respectively, the natural metric and connection for the exponential family (Sect. 5). One important outcome of this analysis is that it allows us to treat the hyperparameters of the conjugate prior as the effective sample points drawn from the distribution under consideration. This analysis also allows us to extend the results of MAP estimation in the exponential family to the α-family (Sect. 5.1) because, similar to exponential families, α-families also carry an inherent geometry (Zhang 2004). We finally extend this geometric interpretation of conjugate priors to analyze the hybrid model given by Lasserre et al. (2006) in a purely geometric setting and justify the argument presented in Agarwal and Daumé (2009) (i.e. a *coupling prior* should be conjugate) using a much simpler analysis (Sect. 6). Our analysis couples the discriminative and generative components of hybrid model using the Bregman divergence which reduces to the coupling prior given in Agarwal and Daumé (2009). This analysis avoids the *explicit* derivation of the hyperparameters, rather automatically gives the hyperparameters of the conjugate prior along with the expression.

2 Motivation

Our analysis is driven by the desire to understand the geometry of the conjugate priors for the exponential families. This understanding has many advantages that are described in the remainder of the paper: an extension of notion of conjugacy beyond the exponential family (to α-family), and geometric analysis of models that use the conjugate priors (Agarwal and Daumé 2009).

We motivate our analysis by asking ourselves the following question: Given a parametric model $p(x;\theta)$ for the data likelihood, and a prior on its parameters θ, $p(\theta;\alpha,\beta)$; what should the hyperparameters α and β of the prior encode? We know that θ in the likelihood model is the estimation of the parameter using the given data points. In other words, the estimated parameter fits the model according to the given data while the prior on the parameter provides the generalization. This generalization is enforced by some prior belief encoded in the hyperparameters. Unfortunately, one does not know what is the likely value of the parameters; rather one might have some belief in what *data points* are likely to be sampled from the model. Now the question is: Do the hyperparameters encode this belief in the parameters in terms of the sampling points? Our analysis shows that the hyperparameters of the conjugate prior is nothing but the effective sampling points. In case of non-conjugate priors, the interpretation of hyperparameters is not clear.

A second motivation is the following geometric analysis. Before we go into the problem, consider two points in the *Euclidean* space which one would like to interpolate using a parameter $\gamma \in [0,1]$. A natural way to do so is to interpolate them linearly i.e., connect two points using a straight line, and then find the interpolating point at the desired γ, as shown in

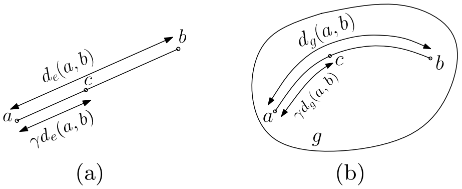

Fig. 1 Interpolation of two points a and b using (**a**) Euclidean geometry, and (**b**) non-Euclidean geometry. Here geometry is defined by the respective distance/divergence functions d_e and d_g. It is important to notice that the divergence is a generalized notion of the distance in the non-Euclidean spaces, in particular, in the spaces of the exponential family statistical manifolds. In these spaces, it is the divergence function that define the geometry

Fig. 1(a). This interpolation scheme does not change if we move to a non-Euclidean space. In other words, if we were to interpolate two points in the non-Euclidean space, we would find the interpolating point by connecting the two points by a geodesic (an equivalent to the straight line in the non-Euclidean space) and then finding the point at the desired γ, shown in Fig. 1(b).

This situation arises when one has two models and wants to build a better model by interpolating them. This exact situation is encountered in Lasserre et al. (2006) where the objective is to build a hybrid model by interpolating (or coupling) discriminative and generative models. Agarwal and Daumé (2009) couples these two models using the conjugate prior, and empirically shows using a conjugate prior for the coupling outperforms the original choice (Lasserre et al. 2006) of a Gaussian prior. In this work, we find the hybrid model by interpolating the two models using the *inherent geometry*[1] of the space (interpolate along the geodesic in the space defined by the inherent geometry) which automatically results in the conjugate prior along with its hyperparameters. Our analysis and the analysis of Agarwal and Daumé lead to the same result, but ours is much simpler and naturally extends to the cases where one wants to couple more than two models. One big advantage of our analysis is that unlike prior approaches (Agarwal and Daumé 2009), we need not know the expression and the hyperparameters of the prior in advance. They are automatically derived by the analysis. Our analysis based on the geometric interpretation can also be used to interpolate the models using a polynomial of higher degree instead of just the straight line i.e., quadratic interpolation etc., and to derive the corresponding prior. Our analysis only requires the inherent geometry which is given by the models under the consideration and the interpolation parameters (parameters of the polynomial). No explicit expression of the coupling prior is needed.

3 Background

In this section we give the required background, specially, we revisit the concepts related to Legendre duality, exponential families and Bregman divergence.

[1] In exponential family statistical manifold, inherent geometry is defined by the divergence function because it is the divergence function that induces the metric structure and connection of the manifold. Refer Amari and Nagaoka (2001) for more details.

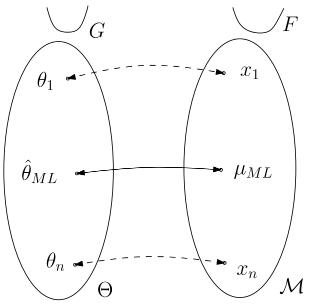

Fig. 2 Duality between mean parameters and natural parameters. Notice the convex functions defined over both spaces. these functions are dual of each other and so are the spaces

3.1 Legendre duality

Let $\mathcal{M} \subseteq \mathbb{R}^d$ and $\Theta \subseteq \mathbb{R}^d$ be two spaces and let $F : \mathcal{M} \to \mathbb{R}^+$ and $G : \Theta \to \mathbb{R}^+$ be two convex functions. F and G are said to be *conjugate duals* of each other if:

$$F(\mu) := \sup_{\theta \in \Theta}\{\langle \mu, \theta \rangle - G(\theta)\} \tag{1}$$

here $\langle a, b \rangle$ denotes the dot product of vectors a and b. The spaces (Θ and $\mathcal{M}$) associated with these dual functions are called *dual spaces*. We sometime use the standard notation to refer this duality i.e., $G = F^*$ and $F = G^*$. A particularly important connection between dual spaces is that: for each $\mu \in \mathcal{M}$, $\nabla F(\mu) = \theta \in \Theta$ (denoted as $\mu^* = \theta$)) and similarly, for each $\theta \in \Theta$, $\nabla G(\theta) = \mu \in \mathcal{M}$ (or $\theta^* = \mu$)). For more details, refer to Rockafellar (1996). Figure 2 gives a pictorial representation of this duality and the notations associated with it.

3.2 Bregman divergence

We now give a brief overview of Bregman divergence (for more details see Banerjee et al. 2005). Let $F : \mathcal{M} \to \mathbb{R}$ be a continuously-differentiable real-valued and strictly *convex function* defined on a closed *convex set* $\mathcal{M}$. The Bregman divergence associated with F for points $p, q \in \mathcal{M}$ is:

$$B_F(p\|q) = F(p) - F(q) - \langle \nabla F(q), (p - q) \rangle \tag{2}$$

If G is the *conjugate dual* of F then:

$$B_F(p\|q) = B_G(q^*\|p^*) \tag{3}$$

here p^* and q^* are the duals of p and q respectively. It is emphasized that Bregman divergence is not symmetric i.e., in general, $B_F(p\|q) \neq B_F(q\|p)$, therefore it is important what directions these divergences are measured in.

3.3 Exponential family

In this section, we review the exponential family. The exponential family is a set of distributions, whose probability density function can be expressed in the following form:

$$p(x; \theta) = p_o(x) \exp(\langle \theta, \phi(x) \rangle - G(\theta)) \tag{4}$$

here $\phi(x) : \mathcal{X}^m \to \mathbb{R}^d$ is a vector *potentials* or *sufficient statistics* and $G(\theta)$ is a normalization constant or *log-partition function*. With the potential functions $\phi(x)$ fixed, every θ induces a particular member $p(x;\theta)$ of the family. In our framework, we deal with exponential families that are *regular* and have the *minimal representation* (Wainwright and Jordan 2003).

The exponential family has a number of convenient properties and subsumes many common distributions. It includes the Gaussian, Binomial, Beta, Multinomial and Dirichlet distributions, hidden Markov models, Bayes nets, etc. One important property of the exponential family is the existence of conjugate priors. Given any member of the exponential family in (4), the *conjugate prior* is a distribution over its *parameters* with the following form:

$$p(\theta|\alpha,\beta) = m(\alpha,\beta)\ \exp(\langle\theta,\alpha\rangle - \beta G(\theta)) \tag{5}$$

here α and β are hyperparameters of the conjugate prior. Importantly, the function $G(\cdot)$ is the same between the exponential family member and its conjugate prior.

A second important property of exponential family member is that log-partition function G is convex and defined over the convex set $\Theta := \{\theta \in \mathbb{R}^d : G(\theta) < \infty\}$. Since the log-partition function G is convex over this set, it induces a Bregman divergence on the space Θ.

Another important property of the exponential family is the *one-to-one* mapping between the *canonical parameters* θ and the so-called "*mean parameters*" which we denote by μ. For each canonical parameter $\theta \in \Theta$, there exists a mean parameter μ, which belongs to the space $\mathcal{M}$ defined as:

$$\mathcal{M} := \left\{\mu \in \mathbb{R}^d : \mu = \int \phi(x)p(x;\theta)\,dx \quad \forall\theta \in \Theta\right\} \tag{6}$$

Our notation has been deliberately suggestive. Θ and $\mathcal{M}$ are dual spaces, in the sense of Legendre duality because of the following relationship between the log-partition function $G(\theta)$ and the expected value of the sufficient statistics $\phi(x)$: $\nabla G(\theta) = \mathbb{E}(\phi(x)) = \mu$.

In Legendre duality, we know that two spaces Θ and $\mathcal{M}$ are dual of each other if for each $\theta \in \Theta$, $\nabla G(\theta) = \mu \in \mathcal{M}$. Here G (the log partition function of the exponential family distribution) is the function defined on the space Θ. We call the function in the dual space $\mathcal{M}$ to be F i.e., $F = G^*$. A pictorial representation of the duality between canonical parameter space Θ and mean parameter space $\mathcal{M}$ is given in Fig. 2.

In our analysis, we will need the Bregman divergence over $\phi(x)$ which can be obtained by showing that an augmented $\mathcal{M}$ contains all possible $\phi(x)$. In order to define the Bregman divergence over all $\phi(x)$, we define a new set of mean parameters w.r.t. all probability distributions (*not* only w.r.t. exponential family distributions): $\mathcal{M}^+ := \{\mu \in \mathbb{R}^d : \mu = \int \phi(x)p(x)\,dx$ s.t. $\int p(x)\,dx = 1\}$.

Note that $\mathcal{M}^+$ is the convex hull of $\phi(x)$ thus containts all $\phi(x)$. We know from (see Theorem 3.3, Wainwright and Jordan 2008) that $\mathcal{M}$ is the *interior* of $\mathcal{M}^+$. Now we augment $\mathcal{M}$ with the boundary of $\mathcal{M}^+$ and Θ with the canonical parameters (limiting distributions) that will generate the mean parameters corresponding to this boundary. We know (see Theorem 2, Wainwright and Jordan 2003) that such parameters exist. Call these new sets $\mathcal{M}^+$ and Θ^+ respectively. We also know (Wainwright and Jordan 2003) that Θ^+ and $\mathcal{M}^+$ are conjugate dual of each other (for boundary, duality exists in the limiting sense) i.e., Bregman divergence is defined over the entire $\mathcal{M}^+$.

In the following discussion, $\mathcal{M}$ and Θ will denote the closed sets i.e. $\mathcal{M}^+$ and Θ^+ respectively.

4 Likelihood, prior and geometry

In this section, we first formulate the ML problem into a Bregman median problem (Sect. 4.1) and then show that corresponding MAP problem can also be converted into a Bregman median problem (Sect. 4.3). The MAP Bregman median problem consists of two parts: a likelihood model and a prior. We argue (Sects. 4.3 and 4.4) that a Bregman median problem makes sense only when both of these parts have the same geometry. Having the same geometry amounts to having the same log-partition function leading to the property of conjugate priors.

4.1 Likelihood in the form of Bregman divergence

Following Collins et al. (2001), we can write the distributions belonging to the exponential family in terms of Bregman divergence. Let $p(x;\theta)$ be the exponential family distribution as defined in (4), the log of which (likelihood) can be written as[2]:

$$\log p(x;\theta) = \log p_o(x) + F(x) - B_F(x\|\nabla G(\theta)) \tag{7}$$

This relationship depends on two observations: $F(\nabla G(\theta)) + G(\theta) = \nabla G(\theta)\theta$ and $(\nabla F)^{-1}(\theta) = \nabla G(\theta) \Rightarrow (\nabla F)(\nabla G(\theta)) = \theta$. These two observations can be used with (2) to see that (7) is equivalent to the probability distribution defined in (4). This representation of likelihood in the form of Bregman divergence gives insight in the geometry of the likelihood function. Gaining the insight into the exponential family distributions and establishing a meaningful relationship between likelihood and prior is the primary objective of this work.

In learning problems, one is interested in estimating the parameters θ of the model which results in low generalization error. Perhaps the most standard estimation method is *maximum likelihood* (ML). The ML estimate, $\hat{\theta}_{ML}$, of a set of n i.i.d. training data points $\mathcal{X} = \{x_1,\dots,x_n\}$ drawn from the exponential family is obtained by solving the following problem: $\hat{\theta}_{ML} = \max_{\theta\in\Theta} \log p(\mathcal{X};\theta)$.

Theorem 1 *Let $\mathcal{X} = \{x_1,\dots,x_n\}$ be a set of n i.i.d. training data points drawn from the exponential family distribution with the log partition function G, F be the dual function of G, then dual of ML estimate ($\hat{\theta}_{ML}$) of $\mathcal{X}$ under the assumed exponential family model solves the following Bregman median problem*:

$$\hat{\mu}_{ML} = \min_{\mu\in\mathcal{M}} \sum_{i=1}^{n} B_F(x_i\|\mu)$$

Proof The log-likelihood of $\mathcal{X}$ under the assumed exponential family distribution is given by $\log p(\mathcal{X};\theta) = \sum_{i=1}^{n} \log p(x_i;\theta)$ which along with (7) can be used to compute $\hat{\theta}_{ML}$:

$$\begin{aligned}\hat{\theta}_{ML} &= \max_{\theta\in\Theta} \sum_{i=1}^{n} (\log p_o(x_i) + F(x_i) - B_F(x_i\|\nabla G(\theta)))\\ &= \min_{\theta\in\Theta} \sum_{i=1}^{n} B_F(x_i\|\nabla G(\theta)) \end{aligned} \tag{8}$$

which using the expression $\nabla G(\theta) = \mu$ gives the desired result. □

[2] For the simplicity of the notations we will use x instead of $\phi(x)$ assuming that $x \in \mathbb{R}^d$. This does not change the analysis.

The above theorem converts the problem of maximizing the log likelihood $\log p(\mathcal{X};\theta)$ into an equivalent problem of minimizing the corresponding Bregman divergences which is nothing but a *Bregman median* problem, the solution to which is given by $\hat{\mu}_{ML}=\sum_{i=1}^{n} x_i$. ML estimate $\hat{\theta}_{ML}$ can now be computed using the expression $\nabla G(\theta)=\mu$, $\hat{\theta}_{ML}=(\nabla G)^{-1}(\hat{\mu}_{ML})$.

Lemma 1 *If x is the sufficient statistics of the exponential family with the log partition function G, and F is the dual function of G defined over the mean parameter space $\mathcal{M}$ then* (1) $x\in\mathcal{M}$; (2) *there exists a $\theta\in\Theta$, such that $x^*=\theta$.*

Proof (1) By construction of $\mathcal{M}$, we know $x\in\mathcal{M}$. (2) From duality of $\mathcal{M}$ and Θ, for every $\mu\in\mathcal{M}$, there exists a $\theta\in\Theta$ such that $\theta=\mu^*$, and since $x\in\mathcal{M}$, which implies $x^*=\theta$. □

Corollary 1 (ML as Bregman Median) *Let $G(\theta)$ be the log partition function of the exponential family defined over the convex set Θ, $\mathcal{X}=\{x_1,\dots,x_n\}$ be set of n i.i.d. data points drawn from this exponential family, and θ_i be the dual of x_i, then ML estimation, $\hat{\theta}_{ML}$ of $\mathcal{X}=\{x_1,\dots,x_n\}$ solves the following optimization problem*:

$$\hat{\theta}_{ML}=\min_{\theta\in\Theta}\sum_{i=1}^{n} B_G(\theta\|\theta_i) \tag{9}$$

Proof Proof directly follows from Lemma 1 and Theorem 1. From Lemma 1, we know that $x_i^*=\theta_i$. Now using Theorem 1 and (3), $B_F(x_i\|\mu)=B_G(\theta\|x_i^*)=B_G(\theta\|\theta_i)$. One can also reduce the above result without using Lemma 1. It is known from Banerjee et al. (2005) that (8) holds for all x which using the duality gives the desired result. □

The above expression requires us to find a θ so that divergence from θ to other θ_i is minimized. Now note that G is what defines this divergence and hence the geometry of the Θ space (as discussed earlier in Sect. 2). Since G is the log partition function of an exponential family, *it is the log-partition function that determines the geometry of the space.* We emphasize that divergence is measured from the parameter being estimated to other parameters θ_i(s), as shown in Fig. 3.

Example 1 (1-D Gaussian) The exponential family representation of a 1-d Gaussian is $p(x;\theta)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(x-a)^2}{2\sigma^2})$ with $\theta=\frac{a}{\sigma^2}$ and $G(\theta)=\frac{\sigma^2}{2}\theta^2$ whose ML estimation is just $(\nabla G)^{-1}(\mu)=\frac{\mu}{\sigma^2}$ which gives $a=\mu=\frac{1}{n}\sum_i x_i$ i.e. data mean.

Example 2 (1-D Bernoulli) The exponential family representation of a Bernoulli distribution $p=a^x(1-a)^{1-x}$ is the distribution with $\theta=\log\frac{a}{1-a}$ with $G(\theta)=\log(1+e^\theta)$ whose ML estimation is $(\nabla G)^{-1}(\mu)=\log\frac{\mu}{1-\mu}$. Comparing it with θ gives $a=\mu=\frac{1}{n}\sum_i x_i$ which is the estimated probability of the event in n trials.

4.2 Conjugate prior in the form of Bregman divergence

We now give an expression similar to the likelihood for the conjugate prior:

$$\log p(\theta|\alpha,\beta)=\log m(\alpha,\beta)+\beta\left(\left\langle\theta,\frac{\alpha}{\beta}\right\rangle-G(\theta)\right) \tag{10}$$

Equation (10) can be written in the form of Bregman divergence by a direct comparison to (4), replacing x with α/β.

$$\log p(\theta|\alpha,\beta) = \log m(\alpha,\beta) + \beta\left(F\left(\frac{\alpha}{\beta}\right) - B_F\left(\frac{\alpha}{\beta}\|\nabla G(\theta)\right)\right) \tag{11}$$

The expression for the joint probability of data and parameters is given by:

$$\begin{aligned}\log p(x,\theta|\alpha,\beta) = \log p_o(x) + \log m(\alpha,\beta) + F(x) + \beta F\left(\frac{\alpha}{\beta}\right)\\ - \left(B_F(x\|\nabla G(\theta)) + \beta B_F\left(\frac{\alpha}{\beta}\|\nabla G(\theta)\right)\right)\end{aligned}$$

Combining all terms that do not depend on θ:

$$\log p(x,\theta|\alpha,\beta) = \text{const} - B_F(x\|\mu) - \beta B_F\left(\frac{\alpha}{\beta}\|\mu\right) \tag{12}$$

4.3 Geometric interpretation of conjugate prior

In this section we give a geometric interpretation of the term $B_F(x\|\mu) + \beta B_F(\frac{\alpha}{\beta}\|\mu)$ from (12).

Theorem 2 (MAP as Bregman median) *Given a set $\mathcal{X}$ of n i.i.d. examples drawn from the exponential family distribution with the log partition function G and a conjugate prior as in* (11), *MAP estimation of parameters is $\hat{\theta}_{MAP} = \hat{\mu}^*_{MAP}$ where $\hat{\mu}_{MAP}$ solves the following problem*:

$$\hat{\mu}_{MAP} = \min_{\mu\in\mathcal{M}} \sum_{i=1}^{n} B_F(x_i\|\mu) + \beta B_F\left(\frac{\alpha}{\beta}\|\mu\right) \tag{13}$$

which admits the following solution:

$$\hat{\mu}_{MAP} = \frac{\sum_{i=1}^{n} x_i + \alpha}{n+\beta}$$

Proof MAP estimation by definition maximizes (12) for all data points $\mathcal{X}$ which is equivalent to minimizing $B_F(x_i\|\mu) + \beta B_F(\frac{\alpha}{\beta}\|\mu)$. One can expand this expression using (2) and use conditions $F(\nabla G(\theta)) + G(\theta) = \nabla G(\theta)\theta$ and $(\nabla F)^{-1}(\theta) = \nabla G(\theta)$ to obtain the desired solution. □

The above solution gives a natural interpretation of MAP estimation. One can think of prior as β number of extra points at position α/β. β works as the effective sample size of the prior which is clear from the following expression of the dual of the $\hat{\theta}_{MAP}$:

$$\hat{\mu}_{MAP} = \frac{\sum_{i=1}^{n} x_i + \sum_{i=1}^{\beta}\frac{\alpha}{\beta}}{n+\beta} \tag{14}$$

The expression (13) is analogous to (8) in the sense that both are defined in the dual space $\mathcal{M}$. One can convert (13) into an expression similar to (9) in the dual space which is again a Bregman median problem in the parameter space.

$$\hat{\theta}_{MAP} = \min_{\theta\in\Theta} \sum_{i=1}^{n} B_G(\theta \| \theta_i) + \sum_{i=1}^{\beta} B_G\left(\theta \| \left(\frac{\alpha}{\beta}\right)^*\right) \tag{15}$$

here $(\frac{\alpha}{\beta})^* \in \Theta$ is dual of $\frac{\alpha}{\beta}$. The above problem is a Bregman median problem of $n+\beta$ points, $\{\theta_1, \theta_2 \ldots \theta_n, \underbrace{(\alpha/\beta)^*, \ldots, (\alpha/\beta)^*}_{\beta \text{ times}}\}$, as shown in Fig. 3 (left).

A geometric interpretation is also shown in Fig. 3. When the prior is conjugate to the likelihood, they both have the same log-partition function (Fig. 3, left). Therefore they induce the same Bregman divergence. Having the same divergence means that distances from θ to θ_i (in likelihood) and the distances from θ to $(\alpha/\beta)^*$ are measured with the same divergence function, yielding the same geometry for both spaces.

It is easier to see using the median formulation of the MAP estimation problem that one must choose a prior that is conjugate. If one chooses a conjugate prior, then the distances among all points are measured using the same function. It is also clear from (14) that in the conjugate prior case, the point induced by the conjugate prior behaves as a sample point $(\alpha/\beta)^*$. A median problem over a space that have different geometries is an ill-formed problem, as discussed further in the next section.

4.4 Geometric interpretation of non-conjugate prior

We derived expression (15) because we considered the prior conjugate to the likelihood function. Had we chosen a non-conjugate prior with log-partition function Q, we would have obtained:

$$\hat{\theta}_{ML} = \min_{\theta\in\Theta} \sum_{i=1}^{n} B_G(\theta \| \theta_i) + \sum_{i=1}^{\beta} B_Q\left(\theta \parallel \left(\frac{\alpha}{\beta}\right)^*\right). \tag{16}$$

Here G and Q are different functions defined over Θ. Since these are the functions that define the geometry of the space parameter, having $G \neq Q$ is equivalent to consider them as being defined over different (metric) spaces. Here, it should be noted that distance between the sample point (θ_i) and the parameter θ is measured using the Bregman divergence B_G. On the other hand, the distance between the point induced by the prior $(\alpha/\beta)^*$ and θ is measured using the divergence function B_Q. This means that $(\alpha/\beta)^*$ can *not* be treated as one of the sample points. This tells us that, unlike the conjugate case, belief in the non-conjugate prior can not be encoded in the form of the sample points.

Another problem with considering a non-conjugate prior is that the dual space of Θ under different functions would be different. Thus, one will not be able to find the alternate expression for (16) equivalent to (13), and therefore not be able to find the closed-form expression similar to (14). This tells us why non-conjugate does not give us a closed form solution for $\hat{\theta}_{MAP}$.

A pictorial representation of this is also shown in Fig. 3. Note that, unlike the conjugate case, in the non-conjugate case, the data likelihood and the prior both belong to different spaces.

We emphasize that it is possible to find the solution of (16) i.e., in practice, there is nothing that prohibits the use of non-conjugate prior, using the conjugate prior is intuitive, and allows one to treat the hyper-parameters as pseudo data points.

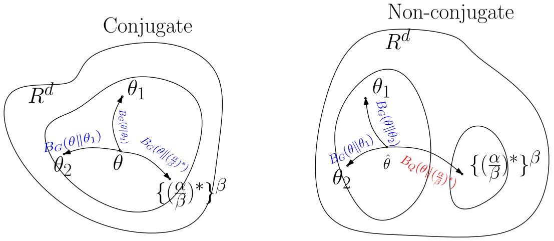

Fig. 3 Prior in the conjugate case has the same geometry as the likelihood while in the non-conjugate case, they have different geometries

5 Information geometric view

In this section, we show the appropriateness of the conjugate prior from the information geometric angle. In information geometry, Θ is a statistical manifold such that each $\theta \in \Theta$ defines a probability distribution. This statistical manifold has an inherent geometry, given by a *metric* and an *affine connection*. One natural metric is the Fisher information metric because of its many attractive properties: it is Riemannian and is invariant under reparameterization (for more details refer Amari and Nagaoka 2001).

In exponential family distributions, the Fisher metric $M(\theta)$ is induced by the KL-divergence $KL(\cdot\|\theta)$, which is equivalent to the Bregman divergence defined by the log-partition function. Thus, it is the log-partition function G that induces the Fisher metric, and therefore determines the *natural* geometry of the space. It justifies our earlier argument of choosing the log-partition function to define the geometry. Now if we were to treat the prior as a point on the statistical manifold defined by the likelihood model, the Fisher information metric on the point given by the prior must be same as the one defined on likelihood manifold. This means that the prior must have the same log-partition function as the likelihood i.e., it must be conjugate.

5.1 Generalization to α-affine manifold

Not all probability distributions belong to the exponential family (although many do). A broader family of distributions is the "α-family" (Amari and Nagaoka 2001). Although a full treatment of this family is beyond the scope of the work, we briefly discuss an extension of our results to the α-family. An α-family distribution is defined as:

$$\log p_\alpha(x;\theta) = \begin{cases} \frac{2}{1-\alpha} p(x;\theta)^{(1-\alpha)/2} & \alpha \neq 1 \\ \log p(x;\theta) & \alpha = 1 \end{cases}$$

where $p(x;\theta)$ defined as in (4). Note that the exponential family is a special case of α-family for $\alpha = 1$.

MAP estimation of the parameters in the exponential family can be cast as a median problem, where an appropriate Bregman divergence is used to define the geometry. In other words, for exponential family, a Bregman-median problem naturally arose as an estimation method.

By using an appropriately defined, "natural," divergence for the α-family, one can actually obtain a similar result for this broader family of distributions. Using such a natural divergence, one can also define a "conjugate prior" for the α-family. Zhang et al. (2004)

shows that such a natural divergence exist for α-family and is given by:

$$D_G^{\alpha}(\theta_1, \theta_2) = \frac{4}{1-\alpha^2}\left(\frac{1-\alpha}{2}G(\theta_1) + \frac{1+\alpha}{2}G(\theta_2) - G\left(\frac{1-\alpha}{2}\theta_1 + \frac{1+\alpha}{2}\theta_2\right)\right)$$

Like the exponential family, this divergence also induces the Fisher information metric.

6 Hybrid model

In this section, we show an application of our analysis to a common supervised and semi-supervised learning framework. In particular, we consider a generative/discriminative hybrid model (Agarwal and Daumé 2009; Druck et al. 2007; Lasserre et al. 2006) that has been shown to be successful in many application.

The hybrid model is a mixture of discriminative and generative models, each of which has its own separate set of parameters. These two sets of parameters (hence two models) are combined using a prior called the *coupling prior*. Let $p(y|\mathbf{x}, \theta_d)$ be the discriminative component, $p(\mathbf{x}, y|\theta_g)$ be the generative component and $p(\theta_d, \theta_g)$ be the prior that couples discriminative and generative components. The joint likelihood of the data and parameters is:

$$\begin{aligned} p(\mathbf{x}, y, \theta_d, \theta_g) &= p(\theta_g, \theta_d)p(y|\mathbf{x}, \theta_d)p(\mathbf{x}|\theta_g) \\ &= p(\theta_g, \theta_d)p(y|\mathbf{x}, \theta_d)\sum_{y'} p(\mathbf{x}, y'|\theta_g) \end{aligned} \tag{17}$$

Here θ_d is a set of discriminative parameters, θ_g a set of generative parameters, and $p(\theta_g, \theta_d)$ provides the natural coupling between these two sets of parameters.

The most important aspect of this model is the *coupling prior* $p(\theta_g, \theta_d)$, which *interpolates* the hybrid model between two extremes: fully generative when the prior forces $\theta_d = \theta_g$, and fully discriminative when the prior renders θ_d and θ_g independent. In non-extreme cases, the goal of the coupling prior is to encourage the generative model and the discriminative model to have similar parameters. It is easy to see that this effect can be induced by many functions. One obvious way is to *linearly* interpolate them as done by Lasserre et al. (2006), Druck et al. (2007) using a Gaussian prior (or the Euclidean distance) of the following form:

$$p(\theta_g, \theta_d) \propto \exp\left(-\lambda\|\theta_g - \theta_d\|^2\right) \tag{18}$$

where, when $\lambda = 0$, model is purely discriminative while for $\lambda = \infty$, model is purely generative. Thus λ in the above expression is the interpolating parameter, and is same as the γ in Sect. 2. Note that the log of the prior is nothing but the squared Euclidean distance between two sets of parameters.

It has been noted multiple times (Bouchard 2007; Agarwal and Daumé 2009) that a Gaussian prior is not always appropriate, and the prior should instead be chosen according to models being considered. Agarwal and Daumé (2009) suggested using a prior that is conjugate to the generative model. Their main argument for choosing the conjugate prior came from the fact that this provides a closed form solution for the generative parameters and therefore is mathematically convenient. We will show that it is more than convenience that makes conjugate prior appropriate. We show that choosing a non-conjugate prior is not only not convenient but also not appropriate. Moreover, our analysis does not assume anything about the expression and the hyperparameters of the prior beforehand, rather derive them automatically.

6.1 Generalized hybrid model

In order to see the effect of the geometry, we first present the generalized hybrid model for distributions that belong to the exponential family and present them in form of Bregman divergences. Following the expression used in Agarwal and Daumé (2009), the generative model can be written as:

$$p(\mathbf{x}, y|\theta_g) = h(\mathbf{x}, y)\exp(\langle\theta_g, T(\mathbf{x}, y)\rangle - G(\theta_g)) \tag{19}$$

where $T(\cdot)$ is the potential function similar to ϕ in (4), now only defined on $(\mathbf{x}, y)$.

Let G^* be the dual function of G; the corresponding Bregman divergence (retaining only the terms that depend on the parameter θ) is given by:

$$B_{G^*}\big((\mathbf{x}, y)\|\nabla G(\theta_g)\big). \tag{20}$$

Solving the generative model independently reduces to choosing a θ_g from the space of all generative parameters Θ_g which has a geometry defined by the log-partition function G. Similarly to the generative model, the exponential form of the discriminative model is given as:

$$p(y|\mathbf{x}, \theta_d) = \exp(\langle\theta_d, T(\mathbf{x}, y)\rangle - M(\theta_d, \mathbf{x})) \tag{21}$$

Importantly, the sufficient statistics T are the *same* in the generative and discriminative models; such generative/discriminative pairs occur naturally: logistic regression/naive Bayes and hidden Markov models/conditional random fields are examples. However, observe that in the discriminative case, the log partition function M depends on both $\mathbf{x}$ and θ_d which makes the analysis of the discriminative model harder. Unlike the generative model, one does not have the explicit form of the log-partition function M that is independent of $\mathbf{x}$. This means that the discriminative component (21) can not be converted into an expression like (20), and the MLE problem can not be reduced to the Bregman median problem like the one given in (9).

6.2 Geometry of the hybrid model

We simplify the analysis of the hybrid model by writing the discriminative model in an alternate form. This alternate form makes obvious the underlying geometry of the discriminative model. Note that the only difference between the two models is that discriminative model models the conditional distribution while the generative model models the joint distribution. We can use this observation to write the discriminative model in the following alternate form using the expression $p(y|x, \theta) = \frac{p(y,x|\theta)}{\sum_{y'} p(y'x|\theta)}$ and (19):

$$p(y|x, \theta_d) = \frac{h(\mathbf{x}, y)\exp(\langle\theta_d, T(\mathbf{x}, y)\rangle - G(\theta_d))}{\sum_{y'} h(\mathbf{x}, y')\exp(\langle\theta_d, T(\mathbf{x}, y')\rangle - G(\theta_d))} \tag{22}$$

Denote the space of parameters of the discriminative model by Θ_d. It is easy to see that geometry of Θ_d is defined by G since function G is defined over θ_d. This is same as the geometry of the parameter space of the generative model Θ_g. Now let us define a new space Θ_H which is the *affine* combination of Θ_d and Θ_g. Now, Θ_H will have the same geometry as Θ_d and Θ_g i.e., geometry defined by G. Now the goal of the hybrid model is to find a $\theta \in \Theta_H$ that maximizes the likelihood of the data under the hybrid model. These two spaces are shown pictorially in Fig. 4.

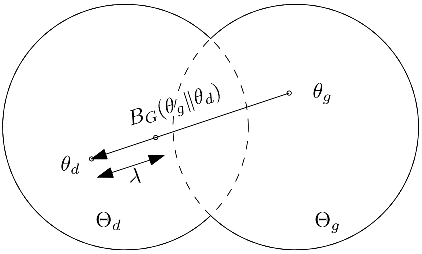

Fig. 4 Parameters θ_d and θ_g are interpolated using the Bregman divergence

6.3 Prior selection

As mentioned earlier, the coupling prior is the most important part of the hybrid model, which controls the amount of coupling between the generative and discriminative models. There are many ways to do this, one of which is given by Lasserre et al. (2006), Druck et al. (2007). By their choice of Gaussian prior as coupling prior, they implicitly couple the discriminative and generative parameters by the squared Euclidean distance. We suggest coupling these two models by a general prior, of which the Gaussian prior is a special case.

6.3.1 Bregman divergence and coupling prior

Let a general coupling be given by $B_S(\theta_g\|\theta_d)$. Notice the direction of the divergence. We have chosen this direction because the prior is induced on the generative parameters, and it is clear from (15) that parameters on which prior is induced, are placed in the first argument in the divergence function. The direction of the divergence is also shown in Fig. 4.

Now we recall the relation (11) between the Bregman divergence and the prior. Ignoring the function m (this is consumed in the measure defined on the probability space) and replacing $\nabla G(\theta)$ by θ^*, we get the following expression:

$$\log p(\theta_g|\alpha,\beta)=\beta\left(F\left(\frac{\alpha}{\beta}\right)-B_F\left(\frac{\alpha}{\beta}\|\theta_g^*\right)\right) \tag{23}$$

Now taking the $\alpha=\lambda\theta_d^*$ and $\beta=\lambda$, we get:

$$\log p(\theta_g|\lambda\theta_d^*,\lambda)=\lambda(F(\theta_d^*)-B_F(\theta_d^*\|\theta_g^*)) \tag{24}$$

$$p(\theta_g|\lambda\theta_d^*,\lambda)=\exp(\lambda(F(\theta_d^*)))\exp(-\lambda B_F(\theta_d^*\|\theta_g^*)) \tag{25}$$

For the general coupling divergence function $B_S(\theta_g\|\theta_d)$, the corresponding coupling prior is given by:

$$\exp(-\lambda B_{S^*}(\theta_d^*\|\theta_g^*))=\exp(-\lambda(F(\theta_d^*)))p(\theta_g|\lambda\theta_d^*,\lambda) \tag{26}$$

The above relationship between the divergence function (left side of the expression) and coupling prior (right side of the expression) allows one to define a Bregman divergence for a given coupling prior and vise versa.

6.3.2 Coupling prior for the hybrid model

We know that the geometry of the space underlying the Gaussian prior is just Euclidean, which does not necessarily match the geometry of the likelihood space. The relationship between prior and divergence (26) allows one to first define the appropriate geometry for the model, and then define the prior that respects this geometry. In the above hybrid model, this geometry is given by the log partition function G of the generative model. This argument suggests to couple the hybrid model by the divergence of the form $B_G(\theta_g \| \theta_d)$. The coupling prior corresponding to this divergence function can be written using (26) as:

$$\exp(-\lambda B_G(\theta_g \| \theta_d)) = p(\theta_g | \lambda \theta_d^*, \lambda) \exp(-\lambda F(\theta_d^*)) \tag{27}$$

where $\lambda = [0, \infty]$ is the interpolation parameter, interpolating between the discriminative and generative extremes. In dual form, the above expression can be written as:

$$\exp(-\lambda B_G(\theta_g \| \theta_d)) = p(\theta_g | \lambda \theta_d^*, \lambda) \exp(-\lambda G(\theta_d)). \tag{28}$$

Here $\exp(-\lambda G(\theta_d))$ can be thought of as a prior on the discriminative parameters $p(\theta_d)$. In the above expression, $\exp(-\lambda B_G(\theta_g \| \theta_d)) = p(\theta_g | \theta_g) p(\theta_d)$ behaves as a joint coupling prior $P(\theta_d, \theta_g)$ as originally expected in the model (17). Note that hyperparameters of the prior α and β are naturally derived from the geometric view of the conjugate prior. Here $\alpha = \lambda \theta_d^*$ and $\beta = \lambda$.

6.3.3 Relation with Agarwal and Daumé

The prior we derived in the previous section turns out to be the exactly same as that proposed by Agarwal and Daumé (2009), even though theirs was not formally justified. In that work, the authors break the coupled prior $p(\theta_g, \theta_d)$ into two parts: $p(\theta_d)$ and $p(\theta_g | \theta_d)$. They then derive an expression for the $p(\theta_g | \theta_d)$ based on the intuition that the mode of $p(\theta_g | \theta_d)$ should be θ_d. Our analysis takes a different approach by coupling two models with the Bregman divergence rather than prior, and results in the expression and hyperparameters for the prior same as in Agarwal and Daumé (2009).

The two analyses diverge here, however. Our analysis derives the hyperparameters as: $\alpha = \lambda (\nabla G)^{-1}(\theta_d)$ and $\beta = \lambda$. However, the expression of the hyperparameters provided by Agarwal and Daumé (2009) was: $\alpha = \lambda \nabla G(\theta_d)$ and $\beta = \lambda$. Their derivation was the assumption that the mode of the coupling prior $p(\theta_g | \theta_d)$ should be θ_d. However, in the conjugate prior $p(\theta | \alpha, \beta)$, the mode is $\frac{\alpha}{\beta}$, and $\frac{\alpha}{\beta}$ behaves as the sufficient statistics for the prior. These terms have come from the data space, *not* from the parameter space. Therefore the mode of the coupling prior $p(\theta_g | \theta_d)$ should not be θ_d, but rather the dual of θ_d which is $(\nabla G)^{-1}(\theta_d) = \theta_d^*$. Therefore, $\alpha = \lambda \theta_d^*$ and $\beta = \lambda$ and our model gives exactly this.

7 Related work and conclusion

To our knowledge, there have been no previous attempts to understand Bayesian priors from a geometric perspective. One related piece of work (Snoussi and Mohammad-Djafari 2003) uses the Bayesian framework to find the best prior for a given distribution. It is noted that, in that work, the authors use the δ-geometry for the data space and the α-geometry for the prior space, and then show the different cases for different values (δ, α). We emphasize that even though it is possible to use different geometry for the both spaces, it always makes more sense to use the same geometry. As mentioned in *remark* 1 in Snoussi and Mohammad-Djafari (2003), useful cases are obtained only when we consider the same geometry.

We have shown that by considering the geometry induced by a likelihood function, the natural prior that results is exactly the conjugate prior. We have used this geometric understanding of conjugate prior to derive the coupling prior for the discriminative/generative hybrid model. Our derivation naturally gives us the expression and the hyperparameters of this coupling prior. Like the hybrid model, this analysis can be used to give the much simpler geometric interpretations of many models, and to extend the existing results to other models, i.e. we have used this analysis to extend the geometric formulation of MAP problem for the exponential family to α-family.

References

Agarwal, A., & Daumé, H. III (2009). Exponential family hybrid semi-supervised learning. In *IJCAI*. Pasadena, CA.

Amari, S. I., & Nagaoka, H. (2001). *Methods of information geometry*. *Translations of mathematical monographs*. Providence: American Mathematical Society.

Banerjee, A., Merugu, S., Dhillon, I. S., & Ghosh, J. (2005). Clustering with Bregman divergences. *Journal of Machine Learning Research*, *6*, 1705–1749.

Bouchard, G. (2007). Bias-variance tradeoff in hybrid generative-discriminative models. In *ICMLA '07* (pp. 124–129). Washington: IEEE Computer Society.

Collins, M., Dasgupta, S., & Schapire, R. E. (2001). *A generalization of principal component analysis to the exponential family*. *NIPS* (Vol. 14). Cambridge: MIT Press.

Druck, G., Pal, C., McCallum, A., & Zhu, X. (2007). Semi-supervised classification with hybrid generative/discriminative methods. In *KDD '07* (pp. 280–289). New York: ACM.

Lasserre, J. A., Bishop, C. M., & Minka, T. P. (2006). Principled hybrids of generative and discriminative models. In *CVPR '06* (pp. 87–94). Washington: IEEE Computer Society.

Rockafellar, R. T. (1996). *Convex analysis*. *Princeton mathematical series*. Princeton: Princeton University Press.

Snoussi, H., & Mohammad-Djafari, A. (2003). Information geometry and prior selection. In *Bayesian inference and maximum entropy methods in science and engineering*. *American institute of physics conference series* (Vol. 659, pp. 307–327). doi:10.1063/1.1570549.

Wainwright, M., & Jordan, M. (2003). *Graphical models, exponential families, and variational inference*. Tech. rep., University of California, Berkeley.

Wainwright, M. J., & Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, *1*(1–2), 1–305.

Zhang, J. (2004). Divergence function, duality, and convex analysis. *Neural Computation*, *16*(1), 159–195.