Conjugate Priors for Generalized MaxEnt Families

Brendan van Rooyen and Mark D. Reid

The Australian National University & NICTA

Abstract. In this work we show that a notion of a conjugate prior for non exponential family distributions can be recovered if one uses a slightly modified version of Bayesian updating. We prove some theorems concerning this new updating rule before giving a simple example of such a generalized conjugate prior.

Keywords: Loss Functions, Bregman Divergences, Convexity, Generalized Bayes Rules, Conjugate Priors

INTRODUCTION

Starting with a prior $P(\theta)$ over some unknown quantities/models Θ, Bayes rule dictates that upon seeing data $x \in X$ we update our prior to the posterior distribution $P(\theta|x) \propto P(x|\theta)P(\theta)$. If we assume that $P(x|\theta)$ is a member of an exponential family,

$$P(x|\theta) = \exp(-\langle \theta, \boldsymbol{F}(x)\rangle - \log(Z(\theta)))$$

where $\boldsymbol{F} : X \to \mathbb{R}^n$ is a vector of sufficient statistics and $Z(\theta)$ is a normalization constant, then there is a very natural conjugate family of priors [8]

$$P(\theta) \propto \exp(-\langle \theta, \boldsymbol{\beta}\rangle - \beta_0 \log(Z(\theta))).$$

This prior has the benefit that performing Bayesian updating becomes addition.

$$P(\theta|x) \propto \exp(-\langle \theta, \boldsymbol{F}(x) + \boldsymbol{\beta}\rangle - (\beta_0 + 1)\log(Z(\theta))).$$

As such using one of these priors can greatly reduce the amount of computation necessary in a Bayesian analysis. For $P(x|\theta)$ not in an exponential family there is in general no conjugate family of priors. However, if one uses a different update rule a notion of conjugate prior for some of these families can be achieved.

Generalized Updating

Here we consider generalized update rules of the form

$$P^*(\theta|x) = \operatorname*{arg\,min}_{Q(\theta)} \underbrace{\mathbb{E}_{\theta \sim Q(\theta)} L(x,\theta)}_{\text{Average Performance}} + \underbrace{D_{KL}(Q(\theta)||P(\theta))}_{\text{Closeness to Prior}} \quad (1)$$

where $L : X \times \Theta \to \mathbb{R}$ is a loss function that says how badly a particular model θ predicted x. The focus will be on loss functions of the form $L : X \times \mathscr{P}(X) \to \mathbb{R}$ with $L(x,\theta) = L(x, P(x|\theta))$ that are “proper" in a sense to made more precise later. We seek a posterior distribution Q that is close to the prior $P(\theta)$ with the caveat that if we draw $\theta \sim Q$, then on average $P(x|\theta)$ predicts x well. This update rule leads to posteriors of the form

$$P^*(\theta|x) \propto \exp(-L(x,\theta))P(\theta).$$

It can be shown that standard Bayesian updating is a special case of the above [10], when $L(x,\theta) = -\log(P(x|\theta))$. The obvious question is how does one match the loss function to the family of interest? Key to the exponential family is that

$$-\log(P(x|\boldsymbol{\theta})) = \langle \boldsymbol{\theta}, \boldsymbol{F}(x)\rangle + \log(Z(\boldsymbol{\theta})).$$

We seek other families $P(x|\boldsymbol{\theta})$ and loss functions L so that

$$L(x, P(x|\boldsymbol{\theta})) = \langle \boldsymbol{\theta}, \boldsymbol{F}(x) \rangle + \Psi(\boldsymbol{\theta}) \quad (2)$$

in which case taking a prior of the form

$$P(\boldsymbol{\theta}) \propto \exp(-\langle \boldsymbol{\theta}, \boldsymbol{\beta} \rangle - \beta_0 \Psi(\boldsymbol{\theta}))$$

yields

$$P^*(\boldsymbol{\theta}|x) \propto \exp(-\langle \boldsymbol{\theta}, \boldsymbol{F}(x) + \boldsymbol{\beta} \rangle - (\beta_0 + 1)\Psi(\boldsymbol{\theta}))$$

when using update rule (1). Finding a conjugate prior for non exponetial family distributions amounts to finding families $P(x|\theta)$ and loss functions so that condition (2) holds. We construct such families by

1. Showing the connection between Shannon entropy and log loss.
2. We generalize this to other entropies and their loss functions.
3. We then show the connection between maximum Shannon entropy with moment constraints and standard exponential families.
4. Finally we tie this together with families of distributions produced by maximizing a generalized entropy under moment constraints.

The main results are that for each generalized entropy there is a corresponding loss and that if we take $P(x|\theta)$ to be a maximum generalized entropy family then we can satisfy condition (2) if we use this loss.

NOTATION/PRELIMINARIES

Let $\mathbb{R}^n_+$ be the set of all vectors in $\mathbb{R}^n$ with non negative entries. Denote the 1-norm of a vector v by $||v||_1$. For a set X let $\mathscr{P}(x)$ be the set of probability distributions on X, and $|X|$ its cardinality.

We take the general view that statistical estimation can be seen as a game between a player (the statistician) and nature. As such, let X be a finite set comprising of natures actions, A the set of the player's actions and $L : X \times A \to \mathbb{R}$ a loss function where $L(x,a)$ is how bad it is for the player if they play action a and nature plays x. As shorthand denote by $L_a : X \to \mathbb{R}$ where $L_a(x) = L(x,a)$.

Vector quantities will be denoted by bold text. Both inner products and expectations will be denoted by angled brackets, $\mathbb{E}_{x \sim P} f(x) = \langle P, f \rangle$, meaning $\mathbb{E}_{x \sim P} L(x,a) = \langle P, L_a \rangle$ and $\langle \delta_x, L_a \rangle = L(x,a)$.

Recall for a concave function $\phi : X \subseteq \mathbb{R}^n \to \mathbb{R}$, that $v \in \mathbb{R}^n$ is a super gradient at the point x if

$$\langle y - x, v \rangle + \phi(x) \geq \phi(y), \ \forall y \in X.$$

Denote the set of all super gradients of ϕ by $\partial \phi(x)$, and a particular super gradient by $\nabla \phi(x)$ [6]. For smooth ϕ super gradients and regular gradients coincide. Denote by D_ϕ the ϕ's induced Bregman divergence

$$D_\phi(x||y) = \langle y - x, \nabla \phi(x) \rangle + \phi(x) - \phi(y), \ \nabla \phi(x) \in \partial \phi(x).$$

A function $\phi : \mathbb{R}^n_+ \to \mathbb{R}$ is 1-homogeneous if for all $\alpha > 0$, $\underline{L}(\alpha x) = \alpha \underline{L}(x)$. Any function $\phi : \mathscr{P}(x) \to \mathbb{R}$ has a 1-homogeneous extension $\hat{\phi}(x) = ||x||_1 \phi(\frac{x}{||x||_1})$

LOSS FUNCTIONS AND ENTROPY

Shannon entropy and log loss are connected by the following identity [4]

$$\langle P, -\log Q \rangle \geq \langle P, -\log P \rangle = H(P), \forall P, Q \in \mathscr{P}(X) \quad (3)$$

with the interpretation of, if the statistician believes nature is drawing $x \sim P$, then the best the statistician can do on average is by playing the distribution P. In this section we seek to generalize this identity to other pairs of entropies and loss functions. We follow [4, 2, 7], the proofs of lemmas occur elsewhere and are presented here for illustrative purposes.

Loss from Entropy

For us an *entropy* will be any concave function 1-homogeneous function

$$\underline{L} : \mathbb{R}_+^{|X|} \to \mathbb{R}$$

This allows for a very large class of entropy functions. In particular for any loss L, its Bayes Risk $\underline{L}(\mu) = \inf_a \langle \mu, L_a \rangle$ will be an entropy function. Here we show that by starting with such an entropy, and taking $L_P = \nabla \underline{L}(P)$ one obtains a loss $L : X \times \mathscr{P}(X) \to \mathbb{R}$. In words, the player is allowed to specify a distribution over natures actions X, and is penalized according to how much mass this distribution assigns to x. This loss is proper meaning

$$\langle P, L_Q \rangle \geq \langle P, L_P \rangle, \; \forall P, Q \in \mathscr{P}(X)$$

furthermore

$$\langle P, L_P \rangle = \underline{L}(P)$$

meaning these losses give a generalization of (3) to arbitrary entropies. If $\underline{L}$ is strictly concave then equality holds only when $P = Q$. To verify these claims we require two easily proved lemmas.

Lemma. *Let $\underline{L}$ be an entropy. If $v \in \partial \underline{L}(\mu)$ then $v \in \partial \underline{L}(\alpha\mu)$ for all $\alpha > 0$.*

Proof. We have for all μ_1, $\langle \mu_1 - \mu, v \rangle + \underline{L}(\mu) \geq \underline{L}(\mu_1)$. Multiplying by $\alpha > 0$ and using the 1-homogeneity of $\underline{L}$ gives

$$\begin{aligned} \langle \alpha\mu_1 - \alpha\mu, v \rangle + \alpha\underline{L}(\mu) &\geq \alpha\underline{L}(\mu_1) \\ \Rightarrow \langle \alpha\mu_1 - \alpha\mu, v \rangle + \underline{L}(\alpha\mu) &\geq \underline{L}(\alpha\mu_1) \\ \Rightarrow \langle \mu_1 - \alpha\mu, v \rangle + \underline{L}(\alpha\mu) &\geq \underline{L}(\mu_1) \end{aligned}$$

Hence $v \in \partial \underline{L}(\alpha\mu)$

□

Therefore the gradient $\nabla \underline{L}$ of a 1-homogeneous function is 0-homogeneous.

Lemma. *For all entropies $\underline{L}$ and all $\mu \in \mathbb{R}_+^{|X|}$, $\underline{L}(\mu) = \langle \mu, \nabla \underline{L}(\mu) \rangle$*

Proof. By the supergradient property and the 1-homogeneity of $\underline{L}$, we have for all $\mu \in \mathbb{R}_+^{|X|}, \alpha > 0$

$$\begin{aligned} \langle \alpha\mu - \mu, \nabla \underline{L}(\mu) \rangle + \underline{L}(\mu) &\geq \underline{L}(\alpha\mu) \\ \Rightarrow \langle \alpha\mu - \mu, \nabla \underline{L}(\mu) \rangle + \underline{L}(\mu) &\geq \alpha\underline{L}(\mu) \end{aligned}$$

letting $\alpha \to 0$ gives $\underline{L}(\mu) \geq \langle \mu, \nabla \underline{L}(\mu) \rangle$. Similarly, by the supergradient property, the 1-homogeneity of $\underline{L}$ and the 0-homogeneity of ∇L we have for all $\mu_1 \in \mathbb{R}_+^{|X|}, \alpha > 0$

$$\begin{aligned} \langle \mu - \alpha\mu, \nabla \underline{L}(\alpha\mu) \rangle + \underline{L}(\alpha\mu) &\geq \underline{L}(\mu) \\ \Rightarrow \langle \mu - \alpha\mu, \nabla \underline{L}(\alpha\mu) \rangle + \alpha\underline{L}(\mu) &\geq \underline{L}(\mu) \\ \Rightarrow \langle \mu - \alpha\mu, \nabla \underline{L}(\mu) \rangle + \alpha\underline{L}(\mu) &\geq \underline{L}(\mu) \end{aligned}$$

letting $\alpha \to 0$ gives $\underline{L}(\mu) \leq \langle \mu, \nabla \underline{L}(\mu) \rangle$. Combining gives $\underline{L}(\mu) = \langle \mu, \nabla \underline{L}(\mu) \rangle$.

□

From the previous two lemmas we have

$$\begin{aligned} D_{\underline{L}}(\mu_1 || \mu_2) &= \underline{L}(\mu_2) + \langle \mu_1 - \mu_2, \nabla \underline{L}(\mu_2) \rangle - \underline{L}(\mu_1) \\ &= \langle \mu_1, \nabla \underline{L}(\mu_2) \rangle - \langle \mu_1, \nabla \underline{L}(\mu_1) \rangle \end{aligned}$$

and as Bregman divergences are always non negative, we have $\langle \mu_1, \nabla \underline{L}(\mu_2) \rangle \geq \langle \mu_1, \nabla \underline{L}(\mu_1) \rangle$.

Theorem. *If $\underline{L}$ is an entropy then $\nabla \underline{L} : \mathscr{P}(X) \to \mathbb{R}^{|X|}$ is a proper loss.*

This suggests a means to construct proper loss functions, first choose a concave 1-homogeneous entropy function $\underline{L}$, and then differentiate (or take supergradients if the function is not smooth).

Example (Shannon Entropy)**.** Let $\underline{L}(P) = \sum_{i=1}^{|X|} -P_i \log(P_i)$ when $P \in \mathscr{P}(X)$ and define

$$\underline{L}(\mu) = \|\mu\|_1 \underline{L}(\frac{\mu}{\|\mu\|_1}) = \sum_{i=1}^{|X|} -\mu_i \log(\frac{\mu_i}{\|\mu\|_1})$$

the 1-homogeneous extension of Shannon Entropy. It is easy to verify that $\frac{\partial \underline{L}(\mu)}{\partial \mu_i} = -\log(\frac{\mu_i}{\|\mu\|_1})$ hence

$$L(x,P) = -\log(P(x)).$$

Example (Tsallis Entropy)**.** Let $\underline{L}_\alpha(P) = \frac{1}{1-\alpha} \sum_{i=1}^{|X|} P_i^\alpha,\ P \in \mathscr{P}(X)$. Once again take the 1-homogeneous extension and differentiate

$$\frac{\partial \underline{L}_\alpha(\mu)}{\partial \mu_i} = \frac{\alpha}{1-\alpha}(\frac{\mu_i}{\|\mu\|_1})^{\alpha-1} + \sum_{j=1}^{|X|}(\frac{\mu_j}{\|\mu\|_1})^\alpha$$

giving

$$L(x,P) = \frac{\alpha}{1-\alpha}(P(x))^{\alpha-1} + \sum_x (P(x))^\alpha.$$

THE GENERALIZED MAXENT PROBLEM AND UPDATING

One route to exponential family distributions is via the maximization of Shannon entropy with moment constraints [5]. One chooses functions $F_i : X \to \mathbb{R}$ as well as the means of these functions f_i and finds the distribution P that maximizes Shannon entropy subject to these mean constraints. As the mean constraints f_i we obtain an exponential family of distributions. Here we change the entropy while keeping the same constraints.

$$\begin{aligned} &\max \underline{L}(\mu),\ \ \mu \in \mathbb{R}_+^{|X|} \\ &\text{subject to } \langle \mu, \boldsymbol{F} \rangle = \boldsymbol{f} \\ &\langle \mu, 1 \rangle = 1. \end{aligned}$$

Assuming that $\underline{L}$ is smooth and that the maximum is attained in the interior of $\mathscr{P}(X)$, by standard techniques from constrained optimization [1] the maximum $\mu_{\boldsymbol{f}}$ occurs when

$$\nabla \underline{L}(\mu_{\boldsymbol{f}}) = \sum_{i=1}^{n} \theta_i F_i + \psi$$

where θ_i is the Lagrange multiplier from the constraint $\langle \mu, F_i \rangle = f_i$ and ψ is the Lagrange multiplier from the constraint $\langle \mu, 1 \rangle = 1$. It is easy to show that ψ can be rewritten as a function of the other Langrange multipliers, $\psi = \Psi(\theta)$. Taking expectations with respect to point mass distributions gives

$$\langle \delta_x, \nabla \underline{L}(\mu_{\boldsymbol{f}}) \rangle = L(x, P(x|\theta)) = \langle \boldsymbol{\theta}, \boldsymbol{F}(x) \rangle + \Psi(\theta)$$

which is exactly condition (2). What we have shown is if you use update rule (1) along with a maximum entropy family and the entropy's corresponding loss then a form of conjugate prior is recovered.

Theorem. *If $P(x|\boldsymbol{\theta})$ is a maximum generalized entropy under moment constraints family for the smooth entropy $\underline{L}$ and with $P(x|\boldsymbol{\theta})$ in the interior of $\mathscr{P}(X)$, then*

$$L(x, P(x|\theta)) = \langle \boldsymbol{\theta}, \boldsymbol{F}(x) \rangle + \Psi(\theta).$$

Legendre Duals

Define $\underline{L}(\boldsymbol{f}) = \underline{L}(\mu_{\boldsymbol{f}})$. By taking expectations of $\nabla \underline{L}(\mu_{\boldsymbol{f}})$ with respect to $\mu_{\boldsymbol{f}}$ we have

$$\underline{L}(\boldsymbol{f}) = \underline{L}(\mu_{\boldsymbol{f}}) = \langle \boldsymbol{\theta}, \boldsymbol{f} \rangle + \Psi(\theta)$$

Once again by standard results of constrained minimization [1] $\partial_{f_i} \underline{L}(\boldsymbol{f}) = \theta_i$ giving

$$\begin{aligned} \partial_{\theta_i} \Psi(\theta) &= \partial_{\theta_i} \underline{L}(\boldsymbol{f}) - \partial_{\theta_i} \langle \theta, \boldsymbol{f} \rangle \\ &= -f_i + \sum_{j=i}^{n} \left(\frac{\partial \underline{L}(\boldsymbol{f})}{\partial f_j} - \theta_j \right) \frac{\partial f_j}{\partial \theta_i} \\ &= -f_i \end{aligned}$$

indicating that $\Psi(\theta)$ serves a role analogous to the log partition function for standard exponential families. In fact, much like the log partition function is the Legendre dual to negative Shannon entropy, $\Psi(\theta)$ is the Legendre dual of $-\underline{L}(\boldsymbol{f})$.

More Properties of Generalized Updating

We have shown that using update rule (1) for a generalized maximum entropy family gives a notion of a conjugate prior distribution. One question is whether or not using this prior and generalized updating rule (1) is consistent. Suppose we have data $\boldsymbol{x} = (x_1, \ldots, x_n)$ drawn iid from some distribution P, what does our generalized posterior distribution converge to? For an arbitrary family $P(x|\theta)$ it can be show that for large n,

$$P^*(\theta|\boldsymbol{x}) \propto \exp(-nD_{\underline{L}}(P||P(x|\theta))P(\theta).$$

Assuming that the prior does not assign zero probability to any θ, $P^*(\theta|\boldsymbol{x})$ converges to a point mass on
$\theta^* = \arg\min_\theta D_{\underline{L}}(P||P(x|\theta))$. Furthermore if $\underline{L}$ is strictly concave and $P = P(x|\theta^*)$ for some θ^* then our generalized updating scheme is consistent. If $P(x|\theta)$ happens to be a generalized maximum entropy family then

$$D_{\underline{L}}(P||P(x|\boldsymbol{\theta})) = \sum_{i=i}^{n} \langle P, F_i \rangle \theta_i + \Psi(\theta) - \underline{L}(P)$$

and this function is minimized when

$$-\partial_{\theta_i} \Psi(\theta) = \langle P, F_i \rangle, \ \forall i$$

which by the previous is when the expectation of F_i is the same under $P(x|\theta)$ and P.

EXAMPLE

Here we consider the maximization of Tsallis entropy (introduced earlier), where $X = \mathbb{R}_+$ and $F(x) = x$ is the single sufficient statistic. We seek a family of distributions with

$$\begin{aligned} L(x, P(x|\theta)) &= \frac{\alpha}{1-\alpha} (P(x|\theta))^{\alpha-1} + \sum_x (P(x|\theta))^\alpha \\ &= \theta x + \Psi(\theta) \end{aligned}$$

rearranging gives

$$P(x|\theta) \propto (1 + \gamma x)^{\frac{1}{\alpha-1}}$$

with $\gamma = \frac{\theta}{\Psi(\theta) - \sum_x (P(x|\theta))^\alpha}$. Performing the relevant calculations of means, entropies and natural parameters using the tools from the previous section gives

$$P(x|\theta) = \theta^{1/\alpha} \left(1 - \frac{(\alpha-1)\theta^{1/\alpha}}{\alpha} x\right)^{\frac{1}{\alpha-1}}$$

for $\theta > 0$ and $\alpha \in (\frac{1}{2}, 1)$. For $\alpha < 1$ this family of distributions has fatter tails than the standard exponential distributions. It can be verified that

$$\Psi(\theta) = \frac{\alpha^2 \theta^{1-1/\alpha}}{(\alpha-1)(2\alpha-1)}$$

resulting in a conjugate prior of the form

$$\begin{aligned} P(\theta) &\propto \exp(-\beta_1 \theta - \beta_0 \frac{\alpha^2 \theta^{1-1/\alpha}}{(\alpha-1)(2\alpha-1)}) \\ &= \exp(-\beta_1 \theta + \beta_0 \theta^{1-1/\alpha}),\ \beta_0, \beta_1 > 0 \end{aligned}$$

where in the last line we have redefined β_0. Note the similarity with the standard conjugate prior for the standard exponential distribution,

$$P(x|\theta) \propto \exp(-\beta_1 \theta + \beta_0 \log(\theta)).$$

CONCLUDING REMARKS

Shannon entropy, Bayes rule, log loss and exponential family distributions are deeply related. Perhaps then it should be of little surprise that when one of these ingredients is changed then all must change in order for things to mesh together nicely. The potential computational benefits from making these changes should hopefully be evident. For example in situations where performing standard Bayesian updating is too difficult computationally due to the difficulty of computing the log loss, another loss can be used and update rule (1) used in its stead. In this work we have shown that one can retain the full computational benefits of conjugate priors even when not using a exponential family.

REFERENCES

1. Dimitri P Bertsekas. Constrained optimization and Lagrange multiplier methods. *Computer Science and Applied Mathematics, Boston: Academic Press, 1982*, 1, 1982.
2. AP Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, (April 2006):77–93, 2007.
3. Thomas Shelburne Ferguson. *Mathematical statistics: A decision theoretic approach*, volume 7. Academic Press New York, 1967.
4. Peter D Grünwald and A Philip Dawid. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *the Annals of Statistics*, 32(4):1367–1433, 2004.
5. Edwin T Jaynes. *Probability theory: the logic of science*. Cambridge university press, 2003.
6. Roberto Lucchetti. *Convexity and well-posed problems*. Springer, 2006.
7. Matthew Parry, A Philip Dawid, and Steffen Lauritzen. Proper local scoring rules. *The Annals of Statistics*, 40(1):561–592, 2012.
8. Christian Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer, 2007.
9. Constantino Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of statistical physics*, 52(1-2):479–487, 1988.
10. Peter M Williams. Bayesian conditionalisation and the principle of minimum information. *The British Journal for the Philosophy of Science*, 31(2):131–144, 1980.