CVI Note

Will Tebbutt

April 2022

This technical note pertains to the interpretation of the work of Khan and Lin (2017) as standard natural gradient ascent in the natural parameters of a particular exponential family. Thanks to Wessel Bruinsma for insightful discussions about minimality.

1 Exponential Families

Consult Wainwright and Jordan (2008) for an excellent overview of exponential families and their properties. The important results are presented here for convenience.

Consider an exponential family prior distribution

$$p(\mathbf{u}) := h(\mathbf{u}) \exp(\langle \eta_0, \phi(\mathbf{u}) \rangle - A(\eta_0)) \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product, η_0 are natural parameters, h the base measure, ϕ the sufficient statistic function, and A the log partition function. This prior has a special relationship to the likelihoods of the form

$$p(\mathbf{y} \,|\, \mathbf{u}) \propto \exp(\langle \tilde{\eta}, \phi(\mathbf{u}) \rangle) \,, \tag{2}$$

in that the corresponding posterior has the same form as the prior:

$$p(\mathbf{u} \,|\, \mathbf{y}) = h(\mathbf{u}) \exp(\langle \eta_1, \phi(\mathbf{u}) \rangle - A(\eta_1)) \,, \quad \eta_1 := \eta_0 + \tilde{\eta}, \tag{3}$$

This is known as a *conjugacy* relationship between the prior and likelihood, and we say that *the likelihood is conjugate to the prior.*

Expectation Parameters Letting p_η be some exponential family, written in terms of its natural parameters η, we can equivalently represent it through its *expectation parameters*, which are defined as

$$\mu(\eta) := \mathbb{E}_{p_\eta}[\phi(\mathbf{u})] \,. \tag{4}$$

The expectation parameters have two important properties for what follows. Firstly, they are equal to the gradient of the log partition function w.r.t the natural parameters:

$$\mu = \frac{\mathrm{d}A}{\mathrm{d}\eta}, \tag{5}$$

so, bringing together Eq. (4) and Eq. (5):

$$\mu(\eta) = \mathbb{E}_{p_\eta}[\phi(\mathbf{u})] = \left. \frac{\mathrm{d}A}{\mathrm{d}\eta} \right|_{\theta := \theta_t} . \tag{6}$$

Fisher Information Secondly, the Fisher information matrix for an exponential family distribution p_η is

$$\mathbf{F} = \frac{\mathrm{d}\mu}{\mathrm{d}\eta}. \tag{7}$$

Minimality An exponential family is *minimal* if there is no vector γ s.t. $\langle \gamma, \phi(\mathbf{u}) \rangle$ is constant for all $\mathbf{u} \neq 0$. This property is important in this work because if an exponential family is minimal, then the mapping from $\mu(\eta)$ is a bijection, which we shall need in order to carry out a number of useful operations.

1.1 Affine Transformations of Exponential Families

It turns out that by parametrising the natural parameters η in terms of some other parameters α, we can construct a new exponential family for which α are the natural parameters. Specifically, consider

$$\eta(\alpha) := C\alpha + c \tag{8}$$

for some linear transformation C and vector c. The corresponding exponential family density can be expressed in terms of α:

$$\begin{aligned} p(\mathbf{u}) &= h(\mathbf{u}) \exp(\langle \eta(\alpha)\,, \phi(\mathbf{u}) \rangle - A(\eta(\alpha))) \\ &= h(\mathbf{u}) \exp(\langle C\alpha + c, \phi(\mathbf{u}) \rangle - A(C\alpha + c)) \\ &= h(\mathbf{u}) \exp(\langle c, \phi(\mathbf{u}) \rangle + \langle \alpha, C^* \phi(\mathbf{u}) \rangle - A(C\alpha + c))\,. \end{aligned} \tag{9}$$

where C^* is the adjoint of C. Now let

$$\begin{aligned} h_\alpha(\mathbf{u}) &:= h(\mathbf{u}) \exp(\langle c, \phi(\mathbf{u}) \rangle)\,, \\ \phi_\alpha(\mathbf{u}) &:= C^\top \phi(\mathbf{u})\,, \\ A_\alpha(\alpha) &:= A(C\alpha + c)\,. \end{aligned}$$

Expressing Eq. (9) in terms of these quantities yields a new exponential family density

$$p_\alpha(\mathbf{u}) := h_\alpha(\mathbf{u}) \exp(\langle \alpha, \phi_\alpha(\mathbf{u}) \rangle - A_\alpha(\alpha))\,. \tag{10}$$

The expectation parameters can be obtained directly from their definition in Eq. (4):

$$\beta = \mathbb{E}[\phi_\alpha(\mathbf{u})] = \mathbb{E}[C^* \phi(\mathbf{u})] = C^* \mathbb{E}[\phi(\mathbf{u})] = C^* \mu \tag{11}$$

where μ are the expectation parameters in the original parametrisation. The third equality follows from the linearity of expectations.

The above means that we can straightforwardly find the expectation parameters β given the natural parameters α via the sequence of transformations

$$\alpha \to \eta \to \mu \to \beta. \tag{12}$$

It should in principle be possible to go in the opposite direction provided that the parametristaion of the new exponential family is minimal. Whether or not this is straightforward to do in practice will depend upon the properties of C and C^*.

Minimality So when is the new exponential family minimal? Consider that

$$\langle \gamma_\alpha, \phi_\alpha(\mathbf{u}) \rangle = \langle \gamma_\alpha, C^* \phi(\mathbf{u}) \rangle = \langle C\gamma_\alpha, \phi(\mathbf{u}) \rangle. \tag{13}$$

If the original exponential family is not minimal, it is unclear that much can be concluded. Fortunately this case is not the interesting one, and is considered no further. On the other hand, if the original exponential family is minimal, then there does not exist $\gamma \neq 0$ such that $\langle \gamma, \phi(\mathbf{u}) \rangle$ is constant for all $\mathbf{u}$. Therefore, the only way in which it is possible to make $\langle \gamma_\alpha, \phi_\alpha(\mathbf{u}) \rangle$ constant for all $\mathbf{u}$ is to find γ_α such that $C\gamma_\alpha = 0$. That is, we must find γ_α which lies in the kernel (nullspace) of C, denoted $\ker(C)$. This gives us a simple condition to determine minimality of the new exponential family: the new exponential family retains minimality iff the kernel of C contains only 0. In the following, C is a bijection, so this condition holds.

2 2-Tuples

It will be helpful in what follows to define 2-tuples that are convex cones with inner products, in order to discuss the parameters of a multivariate Normal distribution as a single object, enabling us to work within our existing abstractions for exponential families.[1] Specifically, we denote a 2-tuple (a, b) where a and b are elements of convex cones A and B respectively, each of which have inner products. We denote the space of all such 2-tuples as $\mathbb{T}(A, B)$. In the case of the parameters of a multivariate Normal, A would be the space of N-dimensional vectors with real-valued elements, and B the set of $N \times N$ positive-definite matrices.

For $\alpha, \beta \in \mathbb{R}_+$ and $t := (a, b), t' := (a', b') \in \mathbb{T}(A, B)$ let

$$\alpha t + \beta t' := (\alpha a + \beta a', \alpha b + \beta b'), \tag{14}$$

$$\langle t, t' \rangle := \langle a, a' \rangle + \langle b, b' \rangle. \tag{15}$$

3 The Conjugate Computation VI (CCVI) Parametrisation

Khan and Lin (2017) present a particular parametristaion of variational inference that they term Conjugate Computation Variational Inference. In order to derive their inference procedure, they rely on details of the natural gradient algorithm, in addition to some assumptions on the form of the approximate posterior. We instead show that it's possible to use the affine transformation trick discussed above to view their algorithm as performing natural gradient descent in a modified exponential family. This is valuable pedagogically in that it provides a clear explanation of what their technique is doing. Furthermore, it clarifies routes by which optimisation algorithms other than vanilla natural gradient descent can be applied, such a Nesterov Acceleration (Nesterov 1983).

Consider an exponential family prior distribution

$$p(\mathbf{u}) = h(\mathbf{u}) \exp(\langle \eta, \phi(\mathbf{u}) \rangle - A(\eta)), \tag{16}$$

and likelihood which factorises:

$$p(\mathbf{y} \,|\, \mathbf{u}) = \prod_{n=1}^{N} p(\mathbf{y}_n \,|\, \mathbf{u}_n). \tag{17}$$

For the sake of concreteness, suppose that the prior is Gaussian, so its natural parameters can be treated as a 2-tuple $\eta = (\eta^{(1)}, \eta^{(2)})$ and its sufficient statistics are another 2-tuple, $\phi(\mathbf{u}) = (\mathbf{u}, \mathbf{u}\mathbf{u}^\top)$.

Suppose that you wish to find approximate posterior parameters η_q, of the same form as η, such that the KL from the distribution with density

$$q(\mathbf{u}; \eta_q) = h(\mathbf{u}) \exp(\langle \eta_q, \phi(\mathbf{u}) \rangle - A(\eta_q)) \tag{18}$$

[1] Convex cones are required because the set of positive-definite matrices of a given size form a convex cone. If you prefer, just replace "convex cone" with "vector" and everything will basically be fine.

to the posterior $p(\mathbf{u} \,|\, \mathbf{y})$ is minimised. Some manipulation shows that this KL is

$$\mathcal{KL}[q \,\|\, p(\cdot \,|\, \mathbf{y})] = \langle \eta_q - \eta, \mu_q \rangle - A(\eta_q) + A(\eta) - \sum_{n=1}^{N} r_n(\eta_{q,n}) + \log p(\mathbf{y}) \tag{19}$$

where $\eta_{q,n} := ([\eta_q^{(1)}]_n, [\eta_q^{(2)}]_{nn})$, and $r_n(\eta_{q,n}) := \mathbb{E}_q[\log p(\mathbf{y}_n \,|\, \mathbf{u}_n)]$ is the reconstruction term, so the ELBO is

$$\mathcal{L}(\eta_q) = \langle \eta - \eta_q, \mu_q \rangle - A(\eta) + A(\eta_q) + \sum_{n=1}^{N} r_n(\eta_{q,n}) \tag{20}$$

Now utilise Sec. 1.1, and consider the affine parametrisation of the natural parameters given by

$$\eta(\alpha) := \alpha + \eta = (\eta^{(1)} + \alpha_1, \eta^{(2)} + \alpha_2), \tag{21}$$

where $\alpha_1 \in \mathbb{R}^N$ and $\alpha_2 \in \mathbb{R}^{N \times N}$ is positive definite. The ELBO becomes

$$\mathcal{L}(\alpha) = \langle -\alpha, \mu_q \rangle - A(\eta) + A(\alpha + \eta) + \sum_{n=1}^{N} r_n(\eta_{q,n}) \tag{22}$$

and the associated natural gradient is

$$\begin{aligned}\tilde{\nabla}_\alpha \mathcal{L}(\alpha) &= \frac{\mathrm{d}\alpha}{\mathrm{d}\mu_q} \nabla_\alpha \mathcal{L}(\alpha) \\ &= \frac{\mathrm{d}\alpha}{\mathrm{d}\mu_q} \left[-\mu_q - \frac{\mathrm{d}\mu_q}{\mathrm{d}\alpha} \alpha + \mu_q + \sum_{n=1}^{N} \nabla_\alpha r_n(\eta_{q,n}) \right] \\ &= \sum_{n=1}^{N} \nabla_{\mu_q} r_n(\eta_{q,n}) - \alpha. \end{aligned} \tag{23}$$

An iteration of natural gradient ascent in α is then

$$\alpha \leftarrow \alpha + \rho(\sum_{n=1}^{N} \nabla_{\mu_q} r_n(\eta_{q,n}) - \alpha) = (1 - \rho)\alpha + \rho \sum_{n=1}^{N} \nabla_{\mu_q} r_n(\eta_{q,n})\,, \tag{24}$$

which is exactly equal to the update in algorithm 1 of (Khan and Lin 2017).

The above establishes that CVI can indeed be interpreted as regular old natural gradient ascent, just in a slightly differently parametrised exponential family (the one involving α).

References

Khan, Mohammad and Wu Lin (2017). “Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models”. In: *Artificial Intelligence and Statistics*. PMLR, pp. 878–887.

Nesterov, Yurii E (1983). “A method for solving the convex programming problem with convergence rate O (1/kˆ 2)”. In: *Dokl. akad. nauk Sssr*. Vol. 269, pp. 543–547.

Wainwright, Martin J and Michael Irwin Jordan (2008). *Graphical models, exponential families, and variational inference*. Now Publishers Inc.