Black-Box α-Divergence Minimization: Supplementary

José Miguel Hernández-Lobato[1*] jmh@seas.harvard.edu

Yingzhen Li[2*] yl494@cam.ac.uk

Mark Rowland[2] mr504@cam.ac.uk

Daniel Hernández-Lobato[3] daniel.hernandez@uam.es

Richard E. Turner[2] ret26@cam.ac.uk

[1]Harvard University, [2]University of Cambridge, [3]Universidad Autónoma de Madrid,

[*]Both authors contributed equally.

A The Min-Max Problem of EP

This section revisits the original EP algorithm as a min-max optimization problem. Recall in the main text that we approximate the true posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$ with a distribution in exponential family form given by $q(\boldsymbol{\theta}) \propto \exp\{\mathbf{s}(\boldsymbol{\theta})^T \boldsymbol{\lambda}_q\}$. Now we define a set of *unnormalized* cavity distributions $q^{\backslash n}(\boldsymbol{\theta}) = \exp\{\mathbf{s}(\boldsymbol{\theta})^T \boldsymbol{\lambda}_{\backslash n}\}$ for every data point $\boldsymbol{x}_n$. Then according to [Minka, 2001], the EP energy function is

$$E(\boldsymbol{\lambda}_q, \{\boldsymbol{\lambda}_{\backslash n}\}) = \log Z(\boldsymbol{\lambda}_0) + (N-1)\log Z(\boldsymbol{\lambda}_q) - \textstyle\sum_{n=1}^N \log \int p(\boldsymbol{x}_n|\boldsymbol{\theta}) q^{\backslash n}(\boldsymbol{\theta}) d\boldsymbol{\theta}\,. \tag{1}$$

In practice EP finds a stationary solution to the constrained optimization problem

$$\min_{\boldsymbol{\lambda}_q} \max_{\{\boldsymbol{\lambda}_n\}} E(\boldsymbol{\lambda}, \{\boldsymbol{\lambda}_n\}) \quad \text{subject to} \quad (N-1)\boldsymbol{\lambda}_q + \boldsymbol{\lambda}_0 = \textstyle\sum_{n=1}^N \boldsymbol{\lambda}_{\backslash n}\,, \tag{2}$$

where the constraint in (2) guarantees that the $\{\boldsymbol{\lambda}_{\backslash n}\}$ are valid cavity parameters that are consistent with the approximate posterior. Similarly for power EP the energy function has the following form:

$$E(\boldsymbol{\lambda}_q, \{\boldsymbol{\lambda}_{\backslash n}\}) = \log Z(\boldsymbol{\lambda}_0) + \left(\tfrac{N}{\alpha} - 1\right)\log Z(\boldsymbol{\lambda}_q) - \tfrac{1}{\alpha}\textstyle\sum_{n=1}^N \log \int p(\boldsymbol{x}_n|\boldsymbol{\theta})^\alpha q^{\backslash n}(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{3}$$

where the constraint of the optimization problem changes to $(N-\alpha)\boldsymbol{\lambda}_q + \alpha\boldsymbol{\lambda}_0 = \sum_{n=1}^N \boldsymbol{\lambda}_{\backslash n}$.

The optimization problem in (2) can be solved using a double-loop algorithm [Heskes and Zoeter, 2002, Opper and Winther, 2005]. This algorithm alternates between an optimization of the cavity parameters $\{\boldsymbol{\lambda}_{\backslash n}\}$ in the inner loop and an optimization of the parameters of the posterior approximation $\boldsymbol{\lambda}_q$ in the outer loop. Each iteration of the double-loop algorithm is guaranteed to minimize the energy in (1). However, the alternating optimization of $\boldsymbol{\lambda}_q$ and $\{\boldsymbol{\lambda}_{\backslash n}\}$ is very inefficient to be useful in practice.

B Linear regression example

In this section we demonstrate several properties of BB-α on a toy linear regression problem; in particular, we compare the BB-α optimal distribution to the true posterior in the cases where the true posterior lies in the variational family considered, and in the mean-field case where the variational family is Gaussian with diagonal covariance matrix.

More specifically, we consider a linear regression model of the form

$$y_n = \boldsymbol{\theta}^\top \mathbf{x}_n + \sigma \varepsilon_n \,,$$

where $y_n \in \mathbf{R}$, $\mathbf{x}_n \in \mathbf{R}^2$ for $n = 1, \ldots, N$, $\boldsymbol{\theta} \in \mathbf{R}^2$, $\sigma \in \mathbf{R}_+$, and $(\varepsilon_n)_{n=1}^N \overset{\text{iid}}{\sim} N(0,1)$. This model is simple enough to analytically optimize the BB-α energy function, and provides intuition for BB-α in more general contexts. We specify a prior $p_0(\boldsymbol{\theta}) \propto \exp(-\boldsymbol{\theta}^\top \boldsymbol{\theta}/2)$ on the model weights; in this case, the posterior distribution for the model weights $\boldsymbol{\theta}$ is given by

$$\begin{aligned} p(\boldsymbol{\theta}) &\propto p_0(\boldsymbol{\theta}) \prod_{n=1}^N p(y_n | \boldsymbol{\theta}, \mathbf{x}_n) \\ &\propto \exp\left(-\frac{1}{2} \boldsymbol{\theta}^\top \boldsymbol{\theta} - \sum_{n=1}^N \frac{1}{2\sigma^2} \left(y_n - \mathbf{x}_n^\top \boldsymbol{\theta} \right)^2 \right) \\ &\propto \exp\left(-\frac{1}{2} \boldsymbol{\theta}^\top (I + \frac{1}{\sigma^2} \sum_{n=1}^N \mathbf{x}_n \mathbf{x}_n^\top) \boldsymbol{\theta} + (\frac{1}{\sigma^2} \sum_{n=1}^N y_n \mathbf{x}_n^\top) \boldsymbol{\theta} \right) \end{aligned}$$

and so the posterior is Gaussian with covariance matrix and mean given by

$$\Sigma = \left(I + \frac{1}{\sigma^2} \sum_{n=1}^N \mathbf{x}_n \mathbf{x}_n^\top \right)^{-1} , \tag{4a}$$

$$\mu = \left(I + \frac{1}{\sigma^2} \sum_{n=1}^N \mathbf{x}_n \mathbf{x}_n^\top \right)^{-1} \frac{1}{\sigma^2} \sum_{n=1}^N y_n \mathbf{x}_n \,. \tag{4b}$$

B.1 Non-recovery of posterior distribution

We first consider the case of a variational family of distributions which contains the true posterior distribution for this model. We let $f(\boldsymbol{\theta})$ be an arbitrary $2-$dimensional Gaussian distribution

$$f(\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2} \boldsymbol{\theta}^\top \Lambda \boldsymbol{\theta} + \eta^\top \boldsymbol{\theta} \right)$$

parameterized by its natural parameters $\Lambda \in \mathbf{R}^{2\times 2}$ and $\eta \in \mathbf{R}^2$, and consider the variational family of the form $q(\boldsymbol{\theta}) \propto p_0(\boldsymbol{\theta}) f(\boldsymbol{\theta})^N$. As described in the main text, the BB-α optimality equations for the variational distribution q are given by

$$\mathbf{E}_q \left[\mathbf{s}(\boldsymbol{\theta}) \right] = \frac{1}{N} \sum_{n=1}^N \mathbf{E}_{\widetilde{p}_n} \left[\mathbf{s}(\boldsymbol{\theta}) \right] \,, \tag{5}$$

where $\mathbf{s}$ is a vector of sufficient statistics for q, and $\widetilde{p}_n(\boldsymbol{\theta}) \propto q(\boldsymbol{\theta}) (p(y_n | \boldsymbol{\theta}, \mathbf{x}_n) / f(\boldsymbol{\theta}))^\alpha$ is the tilted distribution for data point n.

Since the variational family considered is 2-dimensional Gaussian, the sufficient statistics for q are $\mathbf{s}(\boldsymbol{\theta}) = ((\theta_i)_{i=1}^2, (\theta_i \theta_j)_{i,j=1}^{2,2})$. We denote the solution of Equation (5) by $q^\star$, and denote its mean and variance by $\mu_{q^\star}$ and $\Sigma_{q^\star}$ respectively; we denote the corresponding quantities for the tilted distribution $\widetilde{p}_n$ by $\widetilde{\mu}_n$ and $\widetilde{\Sigma}_n$. Equation (5) then becomes

$$\mu_{q^\star} = \frac{1}{N} \sum_{n=1}^N \widetilde{\mu}_n \,, \tag{6a}$$

$$\Sigma_{q^\star} + \mu_{q^\star} \mu_{q^\star}^\top = \frac{1}{N} \left(\sum_{n=1}^N \widetilde{\Sigma}_n + \widetilde{\mu}_n \widetilde{\mu}_n^\top \right) . \tag{6b}$$

Or, in terms of natural parameters,

$$\Lambda_{q^\star}^{-1}\eta_{q^\star} = \frac{1}{N}\sum_{n=1}^{N}\widetilde{\Lambda}_n^{-1}\widetilde{\eta}_n\,, \tag{7a}$$

$$\Lambda_{q^\star}^{-1} + (\Lambda_{q^\star}^{-1}\eta_{q^\star})(\Lambda_{q^\star}^{-1}\eta_{q^\star})^\top = \frac{1}{N}\sum_{n=1}^{N}\left(\widetilde{\Lambda}_n^{-1} + (\widetilde{\Lambda}_n^{-1}\widetilde{\eta}_n)(\widetilde{\Lambda}_n^{-1}\eta_n)^\top\right)\,. \tag{7b}$$

But the natural parameters of the true posterior are the mean of the natural parameters of the tilted distributions; i.e. power EP returns the true posterior when it lies in the approximating family. Since the map from moments to natural parameters is non-linear, we deduce that the distribution q fitted according the moment matching equations (7) is not equal to the true posterior for non-zero α.

B.2 Example 1

We now provide a concrete example of this phenomenon, by performing the explicit calculations for a toy dataset consisting of two observations. We select our dataset to given by $\mathbf{x}_1 = (1,0)^\top$ and $\mathbf{x}_2 = (0,1)^\top$, for now letting the output points $y_{1:2}$ be arbitrary. In this case, we can read off the mean and covariance of the true posterior from (4):

$$\Sigma = \frac{\sigma^2}{1+\sigma^2}I_2\,, \qquad \mu = \frac{1}{1+\sigma^2}(y_1, y_2)^\top\,.$$

Note that in this case the true posterior has a diagonal covariance matrix. We consider fitting a variational family of distributions containing the true posterior via BB-α, to demonstrate that the true posterior is generally not recovered for non-zero α.

We now suppose that

$$f(\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top\Lambda\boldsymbol{\theta} + \eta^\top\boldsymbol{\theta}\right)$$

is constrained to have diagonal precision matrix Λ (and hence diagonal covariance matrix). The optimality equation is still given by (5), but the sufficient statistics are now given by the reduced vector $\mathbf{s}(\boldsymbol{\theta}) = ((\theta_i)_{i=1}^2, (\theta_i^2)_{i=1}^2)$. Now, we have $q(\boldsymbol{\theta}) \propto p_0(\boldsymbol{\theta})f(\boldsymbol{\theta})^N$, and so

$$\begin{aligned} q(\boldsymbol{\theta}) &\propto \exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top\boldsymbol{\theta}\right)\exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top(N\Lambda)\boldsymbol{\theta} + N\eta^\top\boldsymbol{\theta}\right) \\ &= \exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top(I+N\Lambda)\boldsymbol{\theta} + N\eta^\top\boldsymbol{\theta}\right)\,. \end{aligned}$$

Similarly, note that the tilted distribution $\widetilde{p}_n(\boldsymbol{\theta}) \propto p_0(\boldsymbol{\theta})f(\boldsymbol{\theta})^{N-\alpha}p(y_n|\boldsymbol{\theta},\mathbf{x}_n)^\alpha$. We note that

$$\begin{aligned} p(y_n|\boldsymbol{\theta},\mathbf{x}_n) &\propto \exp\left(-\frac{1}{2\sigma^2}(y_n - \mathbf{x}_n^\top\boldsymbol{\theta})^2\right) \\ &\propto \exp\left(y_n\mathbf{x}_n^\top\boldsymbol{\theta} - \frac{1}{2\sigma^2}\boldsymbol{\theta}^\top(\mathbf{x}_n\mathbf{x}_n^\top)\boldsymbol{\theta}\right)\,. \end{aligned}$$

So we have

$$\widetilde{p}_n(\boldsymbol{\theta}) \propto \exp(-\frac{1}{2}\boldsymbol{\theta}^\top\boldsymbol{\theta})\exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top((N-\alpha)\Lambda)\boldsymbol{\theta} + (N-\alpha)\eta^\top\boldsymbol{\theta}\right)\exp\left(\frac{\alpha}{\sigma^2}y_n\mathbf{x}_n^\top\boldsymbol{\theta} - \frac{\alpha}{2\sigma^2}\boldsymbol{\theta}^\top(\mathbf{x}_n\mathbf{x}_n^\top)\boldsymbol{\theta}\right) \tag{8a}$$

$$\propto \exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top\left(I + (N-\alpha)\Lambda + \frac{\alpha}{\sigma^2}\mathbf{x}_n\mathbf{x}_n^\top\right)\boldsymbol{\theta} + \left((N-\alpha)\eta^\top + \frac{\alpha}{\sigma^2}y_n\mathbf{x}_n^\top\right)\boldsymbol{\theta}\right)\,. \tag{8b}$$

We can now use these expressions for the natural parameters, together with the optimality conditions (5) to fit the variational family of distributions.

B.2.1 Matching first moments of the variational distribution

Denoting $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2)$, setting $N = 2$ and using the specific values of $\mathbf{x}_{1:2}$ mentioned above, the optimality equation (7a) implies that we have

$$\frac{2\eta_i}{1+2\lambda_i} = \frac{1}{2}\left(\frac{(2-\alpha)\eta_i + \frac{\alpha}{\sigma^2}y_i}{1+(2-\alpha)\lambda_i + \frac{\alpha}{\sigma^2}}\right) + \frac{1}{2}\left(\frac{(2-\alpha)\eta_i}{1+(2-\alpha)\lambda_i}\right), \qquad i = 1, 2\,.$$

These equations are linear in the components of η, so we have that

$$\eta_i = \frac{(y_i\frac{\alpha}{\sigma^2})/(1+(2-\alpha)\lambda_i + \frac{\alpha}{\sigma^2})}{\frac{4}{1+2\lambda_i} - \frac{2-\alpha}{1+(2-\alpha)\lambda_i+\frac{\alpha}{\sigma^2}} - \frac{2-\alpha}{1+(2-\alpha)\lambda_i}}, \qquad i = 1, 2\,,$$

yielding the natural parameter η_i in terms of λ_i.

B.2.2 Matching second moments of the variational distribution

The optimality equation (7b) can now be used to recover the precision matrix parameters $\lambda_{1:2}$, and hence also the parameters $\eta_{1:2}$ using the formula derived in the above section. Importantly, recall that since we are dealing only with Gaussian variational distributions with diagonal covariance matrices, we need only the diagonal elements of this matrix equation. The resulting equations for $\lambda_{1:2}$ are quite involved, so we begin by evaluating the left-hand side of Equation (7b). Evaluating the i^{th} diagonal element (where $i = 1$ or 2) gives

$$\frac{1}{1+2\lambda_i} + \frac{4\eta_i^2}{(1+2\lambda_i)^2}\,.$$

We now turn our attention to the right-hand side. We again evaluate the i^{th} diagonal element, which yields

$$\frac{1}{2}\left(\frac{1}{1+(2-\alpha)\lambda_i + \frac{\alpha}{\sigma^2}} + \left(\frac{(2-\alpha)\eta_i + \frac{\alpha}{\sigma^2}y_i}{1+(2-\alpha)\lambda_i + \frac{\alpha}{\sigma^2}}\right)^2\right) + \frac{1}{2}\left(\frac{1}{1+(2-\alpha)\lambda_i} + \left(\frac{(2-\alpha)\eta_i}{1+(2-\alpha)\lambda_i}\right)^2\right).$$

Equating the above two expressions in general yields λ_i as a zero of a high-order polynomial, and therefore does not have an analytic solution. However, taking the data points $y_{1:2}$ to be zero simplifies the algebra considerably, and allows for an analytic solution for the natural parameters to be reached. In this case, the optimality equation yields

$$\frac{1}{1+2\lambda_i} = \frac{1}{2}\left(\frac{1}{1+(2-\alpha)\lambda_i + \frac{\alpha}{\sigma^2}} + \frac{1}{1+(2-\alpha)\lambda_i}\right), \qquad i = 1, 2\,.$$

Solving this fixed point equation (with the constraints that $1+2\lambda_i > 0$, $1+(2-\alpha)\lambda_i + \frac{\alpha}{\sigma^2} > 0$ and $1+(2-\alpha)\lambda_i > 0$ to make sure q and $\tilde{p}_n$ are valid distributions) gives

$$\lambda_i = \frac{\sqrt{\alpha^2 - 2\alpha + (\sigma^2+1)^2} - \alpha - \sigma^2 + 1}{2\sigma^2(2-\alpha)}, \qquad i = 1, 2\,.$$

Plotting a diagonal element of the variational covariance matrix as a function of α gives the curve plotted in Figure 1. This plot demonstrates that the variance of the fitted distribution increases continuously and monotonically with α in the range $(0, 2)$.

With the fitted mean and covariance matrix, we can plot the fitted approximate posterior distributions as α varies, along with the true posterior, see Figure 2. Note that in the limit as $\alpha \to 0$, the true posterior is recovered by BB-α, as the corresponding α-divergence converges to the KL-divergence; in fact, the true posterior and the fitted approximation for $\alpha = 10^{-6}$ are indistinguishable at this scale. Note that for non-zero α, the fitted covariance parameters differ from those of the true posterior.

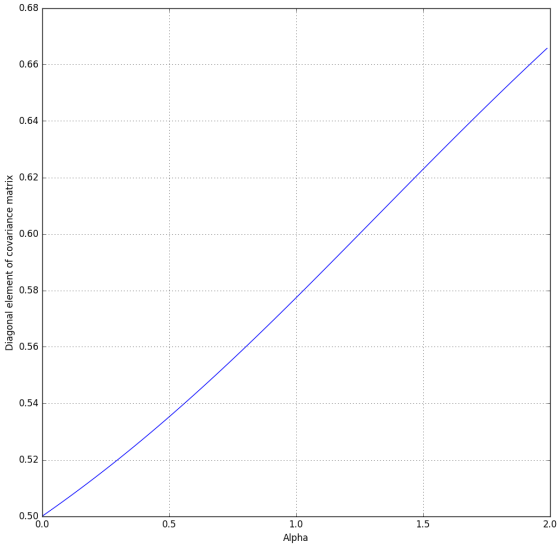

Figure 1: Diagonal element of fitted covariance matrix against α, using output data points $y_1 = y_2 = 0$ and $\sigma^2 = 1$

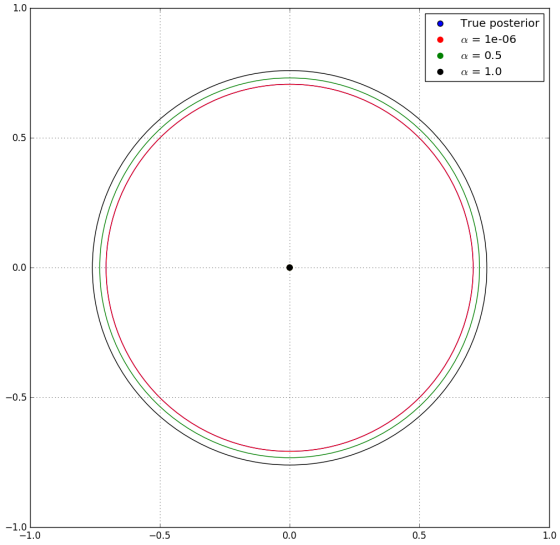

Figure 2: Plot of the mean and one standard deviation's confidence region for the true posterior and several BB-α approximations. Output data is set to $y_1 = 0$, $y_2 = 0$ and $\sigma^2 = 1$

B.3 Mean-field approximations

We now demonstrate the behavior of BB-α in fitting mean-field Gaussian distributions, showing that smooth interpolation between KL-like and EP-like behavior can be obtained by varying α.

B.4 Example 2

We now consider the case where the true posterior has non-diagonal covariance matrix; we set the input values in our toy dataset to be $\mathbf{x}_1 = (1, -1)^\top$, $\mathbf{x}_2 = (-1, 1)^\top$ (again leaving the output data $y_{1:2}$ arbitrary for now), and note from Equation (4) that this implies the mean and covariance of the true posterior are given by

$$\Sigma = \frac{1}{4+\sigma^2}\begin{pmatrix}\sigma^2+2 & 2\\ 2 & \sigma^2+2\end{pmatrix},$$
$$\mu = \left(\frac{y_1-y_2}{4+\sigma^2}, \frac{y_2-y_1}{4+\sigma^2}\right)^\top.$$

We fit the same variational family considered in Example 1; namely the family in which

$$f(\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top \Lambda \boldsymbol{\theta} + \eta^\top \boldsymbol{\theta}\right)$$

is constrained to have diagonal precision matrix Λ (and hence covariance matrix). We can now use this information, together with the form of the tilted distributions in Equation (8) and the optimality conditions in Equation (5) to fit the variational family of distributions.

B.4.1 Matching first moments of the variational distribution

Denoting $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2)$, substituting in $N = 2$ and using our specific choice of data points $\mathbf{x}_{1:2}$, Equation (7a) yields the linear system

$$\begin{pmatrix} \frac{2}{1+2\lambda_1} & 0 \\ 0 & \frac{2}{1+2\lambda_2} \end{pmatrix} \eta = \begin{pmatrix} 1+(2-\alpha)\lambda_1 + \frac{\alpha}{\sigma^2} & -\frac{\alpha}{\sigma^2} \\ -\frac{\alpha}{\sigma^2} & 1+(2-\alpha)\lambda_2 + \frac{\alpha}{\sigma^2} \end{pmatrix}^{-1} \left((2-\alpha)\eta + \frac{\alpha}{2\sigma^2} \begin{pmatrix} y_1 - y_2 \\ y_2 - y_1 \end{pmatrix} \right) .$$

This linear system can be solved for η, although in general the solution is a complicated rational function of (λ_1, λ_2). However, taking $y_1 = y_2$ yields $\eta = 0$.

B.4.2 Matching the second moments of the variational distribution

With the above choices for y_1 and y_2, it follows by symmetry that we must have $\lambda_1 = \lambda_2$. We denote this unknown variable by λ in what follows. Considering the diagonal elements of the Equation (7b) then yields

$$\frac{1}{1+2\lambda_i} = \frac{1+(2-\alpha)\lambda_i + \alpha/\sigma^2}{(1+(2-\alpha)\lambda_i + \alpha/\sigma^2) - \alpha^2/\sigma^2} + \frac{y^2\alpha^2}{\sigma^4} \frac{(1+(2-\alpha)\lambda_i)^2}{(1+(2-\alpha)\lambda_i + \alpha/\sigma^2)^2 - \alpha^2/\sigma^4} .$$

This can be re-arranged into a cubic for λ and thus solved analytically, at least in theory; the resulting expression for λ in terms of α, σ and y is lengthy in practice and is therefore omitted here. We instead consider the case $y = y_1 = y_2 = 0$, where the algebra is more tractable. This results in the following equation for λ:

$$\frac{1}{1+2\lambda_i} = \frac{1+(2-\alpha)\lambda_i + \alpha/\sigma^2}{(1+(2-\alpha)\lambda_i + \alpha/\sigma^2)^2 - \alpha^2/\sigma^4} , \qquad i = 1, 2 .$$

This equation is merely quadratic, and hence has a more easily expressible solution; solving this equation (with the constraint that both sides are positive, and the denominator of the RHS is also positive) gives the precision parameters as

$$\lambda_i = \frac{\sqrt{4\alpha^2 - 8\alpha + \sigma^4 + 4\sigma^2 + 4} - (2\alpha + \sigma^2 - 2)}{2\sigma^2(2-\alpha)} , \qquad i = 1, 2 .$$

Plotting a diagonal element of the corresponding fitted covariance matrix as a function of α gives the curve shown in Figure 3. Note that as before, the element exhibits a continuous, monotonic dependence on α in the range $(0, 2)$.

With the fitted mean and covariance matrix, we can plot the fitted approximate posterior distributions as α varies, along with the true posterior, see Figure 4. Note that in the limit as $\alpha \to 0$, the BB-α solution mimics the KL fit, exhibiting low variance relative to the true posterior, and as α increases, so does the spread of the distribution, consistent with Figure 3.

C Results on toy dataset with neural network regression

We evaluated the predictions obtained by neural networks trained with BB-α in the toy dataset described in [Hernández-Lobato and Adams, 2015]. This dataset is generated by sampling 20 inputs x uniformly at random in the interval $[-4, 4]$. For each value of x obtained, the corresponding target y is computed as $y = x^3 + \epsilon_n$, where $\epsilon_n \sim \mathcal{N}(0, 9)$. We fitted a neural network with one hidden layer with 100 hidden units and rectifier activation functions. In this neural network model, we fixed the variance of the output noise optimally to $\sigma^2 = 9$ and kept the prior variance for the weights fixed to value one. Figure 5 shows plots of the predictive distributions obtained for different values of α. We can see that smaller and negative values of α produce lower variance in the predictive distributions, while larger values of α result in higher predictive variance.

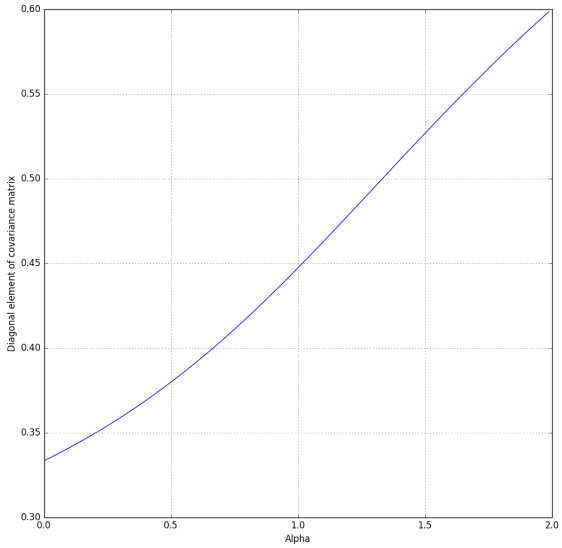

Figure 3: Diagonal element of fitted covariance matrix against α, using output data points $y_1 = 0$, $y_2 = 0$ and $\sigma^2 = 1$

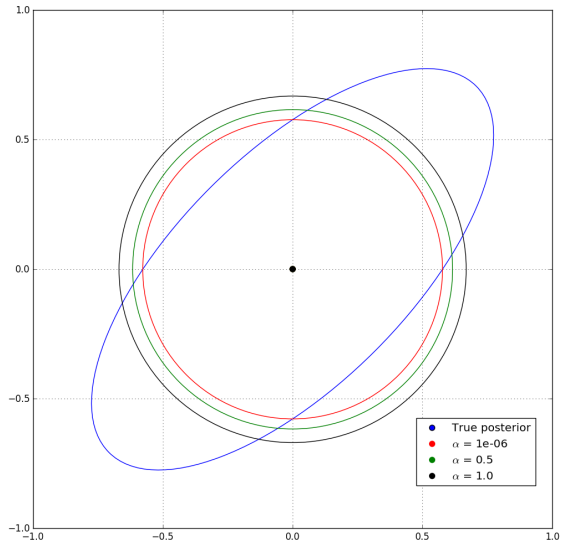

Figure 4: Plot of mean and one standard deviation's confidence region for the true posterior and several BB-α approximations. Output data is set to $y_1 = 0$, $y_2 = 0$ and $\sigma^2 = 1$

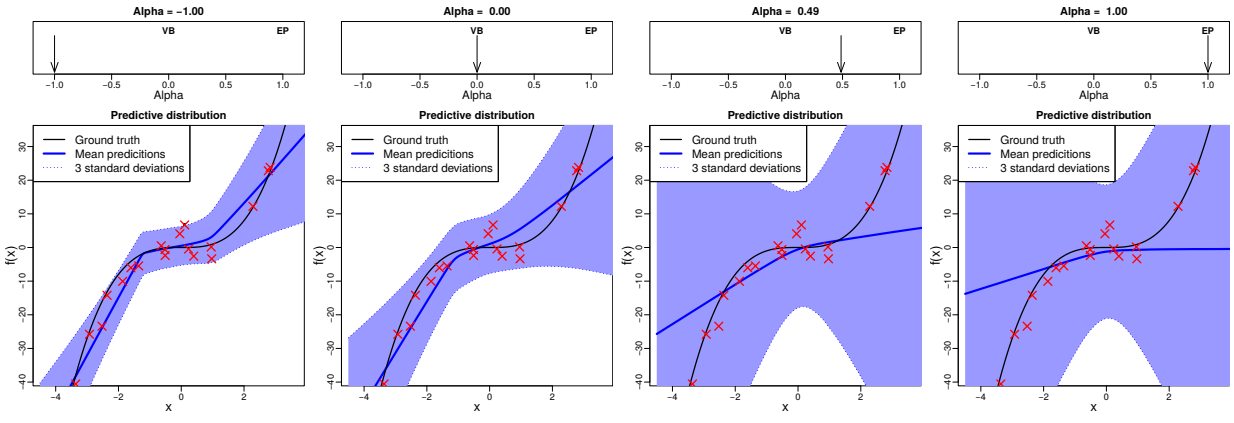

Figure 5: Predictions obtained for different values of α in the toy dataset with neural networks.

References

[Hernández-Lobato and Adams, 2015] Hernández-Lobato, J. M. and Adams, R. (2015). Probabilistic backpropagation for scalable learning of bayesian neural networks. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 1861–1869.

[Heskes and Zoeter, 2002] Heskes, T. and Zoeter, O. (2002). Expectation propagation for approximate inference in dynamic bayesian networks. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 216–223. Morgan Kaufmann Publishers Inc.

[Minka, 2001] Minka, T. P. (2001). Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc.

[Opper and Winther, 2005] Opper, M. and Winther, O. (2005). Expectation consistent approximate inference. *The Journal of Machine Learning Research*, 6:2177–2204.