Predictive Entropy Search for Bayesian Optimization with Unknown Constraints *Supplementary Material*

José Miguel Hernández-Lobato*, Harvard University, USA
Michael A. Gelbart*, Harvard University, USA
Matthew W. Hoffman, University of Cambridge, UK
Ryan P. Adams, Harvard University, USA
Zoubin Ghahramani, University of Cambridge, UK

May 12, 2015

1 Description of Expectation Propagation

PESC computes a Gaussian approximation to the NFCPD (main text, Eq. (11)) using *Expectation Propagation* (EP) (Minka, 2001). EP is a method for approximating a product of factors (often a single prior factor and multiple likelihood factors) with a tractable distribution, for example a Gaussian. EP generates a Gaussian approximation by approximating each individual factor with a Gaussian. The product all these Gaussians results in a single Gaussian distribution that approximates the product of all the exact factors. This is in contrast to the Laplace approximation which fits a single Gaussian distribution to the whole posterior. EP can be intuitively understood as fitting the individual Gaussian approximations by minimizing the Kullback-Leibler (KL) divergences between each exact factor and its corresponding Gaussian approximation. This would correspond to matching first and second moments between exact and approximate factors. However, EP does this moment matching in the *context* of all the other approximate factors, since we are ultimately interested in having a good approximation in regions where the overall posterior probability is high. Concretely, assume we wish to approximate the distribution

$$q(\mathbf{x}) = \prod_{n=1}^{N} q_n(\mathbf{x})$$

with the approximate distribution

$$\tilde{q}(\mathbf{x}) = \prod_{n=1}^{N} \tilde{q}_n(\mathbf{x}) \,, \tag{1}$$

where each $\tilde{q}_n(\mathbf{x})$ is Gaussian with specific parameters. Consider now that we wish to tune the parameters of a particular approximate factor $\tilde{q}_n(\mathbf{x})$. Then, we define the *cavity* distribution $\tilde{q}^{\setminus n}(\mathbf{x})$ as

$$\tilde{q}^{\setminus n}(\mathbf{x}) = \prod_{n' \neq n}^{N} \tilde{q}_{n'}(\mathbf{x}) = \frac{\tilde{q}(\mathbf{x})}{\tilde{q}_n(\mathbf{x})} \,. \tag{2}$$

Instead of matching the moments of $q_n(\mathbf{x})$ and $\tilde{q}_n(\mathbf{x})$, EP matches the moments (minimizes the KL divergence) of $q_n(\mathbf{x})\tilde{q}^{\setminus n}(\mathbf{x})$ and $\tilde{q}_n(\mathbf{x})\tilde{q}^{\setminus n}(\mathbf{x}) = \tilde{q}(\mathbf{x})$. This causes the approximation quality to be higher in places where the

*Authors contributed equally.

entire distribution $\tilde{q}(\mathbf{x})$ is high, at the expense of approximation quality in less relevant regions where $\tilde{q}(\mathbf{x})$ is close to zero. To compute the moments of $q_n(\mathbf{x})\tilde{q}^{\setminus n}(\mathbf{x})$ we use Eqs. (5.12) and (5.13) in Minka (2001), which give the first and second moments of a distribution $p(\mathbf{x})\mathcal{N}(\mathbf{x} \mid \mathbf{m}, \mathbf{V})$ in terms of the derivatives of its log partition function.

Thus, given some initial value for the parameters of all the $\tilde{q}_n(\mathbf{x})$, the steps performed by EP are

1. Choose an n.
2. Compute the cavity distribution $\tilde{q}^{\setminus n}(\mathbf{x})$ given by Eq. (2) using the formula for dividing Gaussians.
3. Compute the first and second moments of $q_n(\mathbf{x})\tilde{q}^{\setminus n}(\mathbf{x})$ using Eqs. (5.12) and (5.13) in Minka (2001). This yields an updated Gaussian approximation $\tilde{q}(\mathbf{x})$ to $q(\mathbf{x})$ with mean and variance given by these moments.
4. Update $\tilde{q}_n(\mathbf{x})$ as the ratio between $\tilde{q}(\mathbf{x})$ and $\tilde{q}^{\setminus n}(\mathbf{x})$, using the formula for dividing Gaussians.
5. Repeat steps 1 to 4 until convergence.

The EP updates for all the $q_n(\mathbf{x})$ maybe be done in parallel by performing steps 1 to 4 for $n = 1, \ldots, N$ using the same $\tilde{q}(\mathbf{x})$ for each n. After his, the new $\tilde{q}(\mathbf{x})$ is computed, according to Eq. (1), as the product of all the newly updated $q_n(\mathbf{x})$. For these latter computations, one uses the formula for multiplying Gaussians.

2 The Gaussian approximation to the NFCPD

In this section we fill in the details of computing a Gaussian approximation to $q(\mathbf{f}, \mathbf{c}_1, \ldots, \mathbf{c}_K)$, given by Eq. (12) of the main text. Recall that $\mathbf{f} = (f(\mathbf{x}_\star), f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))$ and $\mathbf{c}_k = (c(\mathbf{x}_\star), c(\mathbf{x}_1), \ldots, c(\mathbf{x}_n))$, where $k = 1, \ldots, K$, and the first element in these vectors is accessed with the index number 0 and the last one with the index number n. The expression for $q(\mathbf{f}, \mathbf{c}_1, \ldots, \mathbf{c}_K)$ can be written as

$$q(\mathbf{f}, \mathbf{c}_1, \ldots, \mathbf{c}_K) = \frac{1}{Z_q} \mathcal{N}\left(\mathbf{f} \mid \mathbf{m}^{\mathbf{f}}_{\text{pred}}, \mathbf{V}^{\mathbf{f}}_{\text{pred}}\right) \left[\prod_{k=1}^{K} \mathcal{N}\left(\mathbf{c}_k \mid \mathbf{m}^{\mathbf{c}_k}_{\text{pred}}, \mathbf{V}^{\mathbf{c}_k}_{\text{pred}}\right)\right] \times \left[\prod_{k=1}^{K} \Theta[c_{k,0}]\right] \prod_{n=1}^{N} \left[\left\{\prod_{k=1}^{K} \Theta[c_{k,n}]\right\} \Theta[f_n - f_0] + \left\{1 - \prod_{k=1}^{K} \Theta[c_{k,n}]\right\}\right], \tag{3}$$

where $\mathbf{m}^{f}_{\text{pred}}$ and $\mathbf{V}^{f}_{\text{pred}}$ are the mean and covariance matrix of the posterior distribution of $\mathbf{f}$ given the data in $\mathcal{D}^f$ and $\mathbf{m}^{\mathbf{c}_k}_{\text{pred}}$ and $\mathbf{V}^{\mathbf{c}_k}_{\text{pred}}$ are the mean and covariance matrix of the posterior distribution of $\mathbf{c}_k$ given the data in $\mathcal{D}^k$. In particular, from Rasmussen & Williams (2006) Eqs. 2.22-2.24, we have that

$$\begin{aligned} \mathbf{m}^{\mathbf{f}}_{\text{pred}} &= \mathbf{K}^f_\star (\mathbf{K}^f + \nu_f^2 \mathbb{I})^{-1} \mathbf{y}^f , \\ \mathbf{V}^{\mathbf{f}}_{\text{pred}} &= \mathbf{K}^f_{\star,\star} - \mathbf{K}^f_\star (\mathbf{K}^f + \nu_f^2 \mathbb{I})^{-1} [\mathbf{K}^f_\star]^\top , \end{aligned}$$

where $\mathbf{K}^f_\star$ is an $(N+1) \times N$ matrix with the prior cross-covariances between elements of $\mathbf{f}$ and $f_1, \ldots, f_n$ and $\mathbf{K}^f_{\star,\star}$ is an $(N+1) \times (N+1)$ matrix with the prior covariances between elements of $\mathbf{f}$ and ν_f is the standard deviation of the additive Gaussian noise in the evaluations of f. Similarly, we have that

$$\begin{aligned} \mathbf{m}^{\mathbf{c}_k}_{\text{pred}} &= \mathbf{K}^k_\star (\mathbf{K}^k + \nu_k^2 \mathbb{I})^{-1} \mathbf{y}^k , \\ \mathbf{V}^{\mathbf{c}_k}_{\text{pred}} &= \mathbf{K}^k_{\star,\star} - \mathbf{K}^k_\star (\mathbf{K}^k + \nu_k^2 \mathbb{I})^{-1} [\mathbf{K}^k_\star]^\top , \end{aligned}$$

where $\mathbf{K}^k_\star$ is an $(N+1) \times N$ matrix with the prior cross-covariances between elements of $\mathbf{c}_k$ and $c_{k,1}, \ldots, c_{k,n}$ and $\mathbf{K}^k_{\star,\star}$ is an $(N+1) \times (N+1)$ matrix containing the prior covariances between the elements of $\mathbf{c}_k$ and ν_k is the standard deviation of the additive Gaussian noise in the evaluations of c_k. We will refer to the non-Gaussian factors in Eq. (3) as

$$h_n(f_n, f_0, c_{1,n}, \ldots, c_{k,n}) = \left\{\prod_{k=1}^{K} \Theta\left[c_{k,n}\right]\right\} \Theta\left[f_n - f_0\right] + \left\{1 - \prod_{k=1}^{K} \Theta\left[c_{k,n}\right]\right\} \tag{4}$$

(this is called $\Psi(\mathbf{x}_n, \mathbf{x}_\star, \mathbf{f}, \mathbf{c}_1, \ldots, \mathbf{c}_K)$ in the main text) and

$$g_k(c_{k,0}) = \Theta[c_{k,0}]\,, \tag{5}$$

such that Eq. (3) can be written as

$$\begin{aligned} q(\mathbf{f}, \mathbf{c}_1, \ldots, \mathbf{c}_K) \propto \mathcal{N}(\mathbf{f} \,|\, \mathbf{m}^{\mathbf{f}}_{\text{pred}}, \mathbf{V}^{\mathbf{f}}_{\text{pred}}) \left[\prod_{k=1}^{K} \mathcal{N}(\mathbf{c}_k \,|\, \mathbf{m}^{\mathbf{c}_k}_{\text{pred}}, \mathbf{V}^{\mathbf{c}_k}_{\text{pred}})\right] \\ \left[\prod_{n=1}^{N} h_n(f_n, f_0, c_{1,n}, \ldots, c_{k,n})\right] \left[\prod_{k=1}^{K} g_k(c_{k,0})\right]\,. \end{aligned} \tag{6}$$

We then approximate the exact non-Gaussian factors $h_n(f_n, f_0, c_{1,n}, \ldots, c_{k,n})$ and $g_k(c_{k,0})$ with corresponding Gaussian approximate factors $\tilde{h}_n(f_n, f_0, c_{1,n}, \ldots, c_{k,n})$ and $\tilde{g}_k(c_{k,0})$, respectively. By the assumed independence of the objective and constraints, these approximate factors take the form

$$\begin{aligned} \tilde{h}_n(f_n, f_0, c_{1,n}, \ldots, c_{k,n}) \propto \exp\left\{-\frac{1}{2}[f_n\, f_0]\mathbf{A}_{h_n}[f_n\, f_0]^\top + [f_n\, f_0]\mathbf{b}_{h_n}\right\} \\ \prod_{k=1}^{K} \exp\left\{-\frac{1}{2}a_{h_n}c_{k,n}^2 + b_{h_n}c_{k,n}\right\} \end{aligned} \tag{7}$$

and

$$\tilde{g}_k(c_{k,0}) \propto \exp\left\{-\frac{1}{2}a_{g_k}c_{k,0}^2 + b_{g_k}c_{k,0}\right\}\,, \tag{8}$$

where $\mathbf{A}_{h_n}, \mathbf{b}_{h_n}, a_{h_n}, b_{h_n}, a_{g_k}$ and b_{g_k} are the natural parameters of the respective Gaussian distributions.

The right-hand side of Eq. (6) is then approximated by

$$\begin{aligned} \tilde{q}(\mathbf{f}, \mathbf{c}_1, \ldots, \mathbf{c}_K) \propto \mathcal{N}\left(\mathbf{f} \,|\, \mathbf{m}^{f}_{\text{pred}}, \mathbf{V}^{f}_{\text{pred}}\right) \left[\prod_{k=1}^{K} \mathcal{N}\left(\mathbf{c}_k \,|\, \mathbf{m}^{\mathbf{c}_k}_{\text{pred}}, \mathbf{V}^{\mathbf{c}_k}_{\text{pred}}\right)\right] \\ \left[\prod_{n=1}^{N} \tilde{h}_n(f_n, f_0, c_{1,n}, \ldots, c_{k,n})\right] \left[\prod_{k=1}^{K} \tilde{g}_k(c_{k,0})\right]\,. \end{aligned} \tag{9}$$

Because the $\tilde{h}_n(f_n, f_0, c_{1,n}, \ldots, c_{k,n})$ and $\tilde{g}_k(c_{k,0})$ are Gaussian, they can be combined with the Gaussian terms in the first line of Eq. (9) and written as

$$\tilde{q}(\mathbf{f}, \mathbf{c}_1, \ldots, \mathbf{c}_K) \propto\ \mathcal{N}(\mathbf{f} \,|\, \mathbf{m}^{\mathbf{f}}, \mathbf{V}^{\mathbf{f}}) \left[\prod_{k=1}^{K} \mathcal{N}(\mathbf{c}_k \,|\, \mathbf{m}^{\mathbf{c}_k}, \mathbf{V}^{\mathbf{c}_k})\right]\,, \tag{10}$$

where, by applying the formula for products of Gaussians, we obtain

$$\begin{aligned} \mathbf{V}^{\mathbf{f}} &= \left[\left(\mathbf{V}^{\mathbf{f}}_{\text{pred}}\right)^{-1} + \widetilde{\mathbf{V}}^{\mathbf{f}}\right]^{-1}\,, & \mathbf{m}^{\mathbf{f}} &= \mathbf{V}^{\mathbf{f}}\left[\left(\mathbf{V}^{\mathbf{f}}_{\text{pred}}\right)^{-1}\mathbf{m}^{\mathbf{f}}_{\text{pred}} + \widetilde{\mathbf{m}}^{\mathbf{f}}\right]\,, \\ \mathbf{V}^{\mathbf{c}_k} &= \left[\left(\mathbf{V}^{\mathbf{c}_k}_{\text{pred}}\right)^{-1} + \widetilde{\mathbf{V}}^{\mathbf{c}_k}\right]^{-1}\,, & \mathbf{m}^{\mathbf{c}_k} &= \mathbf{V}^{\mathbf{c}_k}\left[\left(\mathbf{V}^{\mathbf{c}_k}_{\text{pred}}\right)^{-1}\mathbf{m}^{\mathbf{c}_k}_{\text{pred}} + \widetilde{\mathbf{m}}^{\mathbf{c}_k}\right]\,, \end{aligned} \tag{11}$$

with the following definitions for $\widetilde{\mathbf{V}}, \widetilde{\mathbf{m}}^{\mathbf{f}}, \widetilde{\mathbf{V}}^{\mathbf{c}_k}$, and $\widetilde{\mathbf{m}}^{\mathbf{c}_k}$:

- $\widetilde{\mathbf{V}}^{\mathbf{f}}$ is an $(N+1)\times(N+1)$ precision matrix with entries given by
  - $\tilde{v}^{\mathbf{f}}_{n,n} = [\mathbf{A}_{h_n}]_{1,1}$ for $n = 1, \ldots, N$,
  - $\tilde{v}^{\mathbf{f}}_{0,n} = \tilde{v}^{\mathbf{f}}_{n,0} = [\mathbf{A}_{h_n}]_{1,2}$ for $n = 1, \ldots, N$,

- $\tilde{v}^{\mathbf{f}}_{0,0} = \sum_{n=1}^{N} [\mathbf{A}_{h_n}]_{2,2}$,
- and all other entries are zero.

- $\widetilde{\mathbf{m}}^{\mathbf{f}}$ is an $(N+1)$-dimensional vector with entries given by
  - $\widetilde{m}^{\mathbf{f}}_n = [\mathbf{b}_{h_n}]_1$ for $n = 1, \ldots, N$,
  - $\widetilde{m}^{\mathbf{f}}_0 = \sum_{n=1}^{N} [\mathbf{b}_{h_n}]_2$.
- $\widetilde{\mathbf{V}}^{\mathbf{c}_k}$ is a $(N+1) \times (N+1)$ diagonal precision matrix with entries given by
  - $\tilde{v}^{\mathbf{c}_k}_{n,n} = a_{h_n}$ for $n = 1, \ldots, N$,
  - $\tilde{v}^{\mathbf{c}_k}_{0,0} = a_{g_k}$.
- $\tilde{\mathbf{m}}^{\mathbf{c}_k}$ is an $(N+1)$-dimensional vector such that
  - $\widetilde{m}^{\mathbf{c}_k}_n = b_{h_n}$ for $n = 1, \ldots, N$,
  - $\widetilde{m}^{\mathbf{c}_k}_0 = b_{g_k}$.

In the next section we explain how to actually obtain the values of $\mathbf{A}_{h_n}, \mathbf{b}_{h_n}, a_{h_n}, b_{h_n}, a_{g_k}$ and b_{g_k} by running EP.

3 The EP approximation to h_n and g_k

In this section we explain how to iteratively refine the approximate Gaussian factors $\tilde{h}_n$ and $\tilde{g}_k$ with EP.

3.1 EP update operations for $\tilde{h}_n$

EP adjusts each $\tilde{h}_n$ by minimizing the KL divergence between

$$h_n(f_n, f_0, c_{1,n}, \ldots, c_{k,n}) \tilde{q}^{\setminus n}(\mathbf{f}, \mathbf{c}_1, \ldots, \mathbf{c}_K) \tag{12}$$

and

$$\tilde{h}_n(f_n, f_0, c_{1,n}, \ldots, c_{k,n}) \tilde{q}^{\setminus n}(\mathbf{f}, \mathbf{c}_1, \ldots, \mathbf{c}_K)\,, \tag{13}$$

where we define the cavity distribution $\tilde{q}^{\setminus n}(\mathbf{f}, \mathbf{c}_1, \ldots, \mathbf{c}_K)$ as

$$\tilde{q}^{\setminus n}(\mathbf{f}, \mathbf{c}_1, \ldots, \mathbf{c}_K) = \frac{\tilde{q}(\mathbf{f}, \mathbf{c}_1, \ldots, \mathbf{c}_K)}{\tilde{h}_n(f_n, f_0, c_{1,n}, \ldots, c_{k,n})}\,.$$

Because $\tilde{q}(\mathbf{f}, \mathbf{c}_1, \ldots, \mathbf{c}_K)$ and $\tilde{h}_n(f_n, f_0, c_{1,n}, \ldots, c_{k,n})$ are both Gaussian, $\tilde{q}^{\setminus n}(\mathbf{f}, \mathbf{c}_1, \ldots, \mathbf{c}_K)$ is also Gaussian. If we marginalize out all variables except those which $\tilde{h}_n$ depends on, namely $f_n, f_0, c_{1,n}, \ldots, c_{k,n}$, then we have that $\tilde{q}^{\setminus n}(f_n, f_0, c_{1,n}, \ldots, c_{k,n})$ takes the form

$$\tilde{q}^{\setminus n}(f_n, f_0, c_{1,n}, \ldots, c_{k,n}) \propto \mathcal{N}\left([f_n f_0] \,|\, \mathbf{m}^{f_n,f_0}_{n,\text{old}}, \mathbf{V}^{f_n,f_0}_{n,\text{old}}\right) \left[\prod_{k=1}^{K} \mathcal{N}\left(c_{k,n} \,|\, m^{c_{k,n}}_{n,\text{old}}, v^{c_{k,n}}_{n,\text{old}}\right)\right]\,, \tag{14}$$

where, by applying the formula for dividing Gaussians, we obtain

$$\mathbf{V}^{f_n,f_0}_{n,\text{old}} = \left\{\left[\mathbf{V}^{\mathbf{f}}_{f_n,f_0}\right]^{-1} - \mathbf{A}_{h_n}\right\}^{-1}\,, \tag{15}$$

$$\mathbf{m}^{f_n,f_0}_{n,\text{old}} = \mathbf{V}^{f_n,f_0}_{n,\text{old}} \left\{\left[\mathbf{V}^{\mathbf{f}}_{f_n,f_0}\right]^{-1} \mathbf{m}^{\mathbf{f}}_{f_n,f_0} - \mathbf{b}_{h_n}\right\}\,, \tag{16}$$

$$v^{c_{k,n}}_{n,\text{old}} = \left\{\left[v^{\mathbf{c}_k}_{n,n}\right]^{-1} - a^{h_n}_{c_{k,n}}\right\}^{-1}\,, \tag{17}$$

$$m^{c_{k,n}}_{n,\text{old}} = \left\{m^{\mathbf{c}_k}_n \left[v^{\mathbf{c}_k}_{n,n}\right]^{-1} - b_{h_n}\right\}^{-1}\,, \tag{18}$$

where $\mathbf{V}^{\mathbf{f}}_{f_n,f_0}$ is the 2×2 covariance matrix obtained by taking the entries corresponding to f_n and f_0 from $\mathbf{V}^{\mathbf{f}}$, and $\mathbf{m}^{\mathbf{f}}_{f_n,f_0}$ is the corresponding 2-dimensional mean vector. Similarly, $v^{\mathbf{c}_k}_{n,n}$ is the variance for $c_{k,n}$ in $\tilde{q}$ and $m^{\mathbf{c}_k}_n$ is the corresponding mean.

To minimize the KL divergence between Eqs. (12) and (13), we match the first and second moments of these two distributions. The moments of Eq. (12) can be obtained from the derivatives of its normalization constant. This normalization constant is given by

$$\begin{aligned} Z &= \int h_n(f_n, f_0, c_{1,n}, \ldots, c_{k,n}) \tilde{q}^{\setminus n}(f_n, f_0, c_{1,n}, \ldots, c_{k,n})\, df_n, df_0, dc_{1,n}, \ldots, dc_{k,n} \\ &= \left(\left\{ \prod_{k=1}^{K} \Phi\left[\alpha_n^k\right] \right\} \Phi(\alpha_n) + \left\{ 1 - \prod_{k=1}^{K} \Phi\left[\alpha_n^k\right] \right\} \right), \end{aligned} \tag{19}$$

where $\alpha_n^k = m^{c_{k,n}}_{n,\text{old}} / \sqrt{v^{c_{k,n}}_{n,\text{old}}}$ and $\alpha_n = [1, -1]\mathbf{m}^{f_n,f_0}_{n,\text{old}} / \sqrt{[1, -1]\mathbf{V}^{f_n,f_0}_{n,\text{old}}[1, -1]^\top}$. We follow Eqs. (5.12) and (5.13) Minka (2001) to update a_{h_n} and b_{h_n}; however, we use the second partial derivatives with respect to $m^{c_{k,n}}_{n,\text{old}}$ rather than first partial derivative with respect to $v^{c_{k,n}}_{n,\text{old}}$ for numerical robustness. These derivatives are given by

$$\frac{\partial \log Z}{\partial m^{c_{k,n}}_{n,\text{old}}} = \frac{(Z-1)\phi(\alpha_n^k)}{Z\Phi(\alpha_n^k)\sqrt{v^{c_{k,n}}_{n,\text{old}}}}, \tag{20}$$

$$\frac{\partial^2 \log Z}{\partial [m^{c_{k,n}}_{n,\text{old}}]^2} = -\frac{(Z-1)\phi(\alpha_n^k)}{Z\Phi(\alpha_n^k)} \cdot \frac{\left[\alpha_n^k + \frac{(Z-1)\phi(\alpha_n^k)}{Z\Phi(\alpha_n^k)}\right]}{v^{c_n^n}_{l,\text{old}}}, \tag{21}$$

where ϕ and Φ are the standard Gaussian pdf and cdf, respectively. The update equations for the parameters a_{h_n} and b_{h_n} of the approximate factor $\tilde{h}$ are then

$$\begin{aligned} a^{\text{new}}_{h_n} &= -\left(\left(\frac{\partial^2 \log Z}{\partial [m^{c_{k,n}}_{n,\text{old}}]^2} \right)^{-1} + v^{c_{k,n}}_{n,\text{old}} \right)^{-1}, \\ b^{\text{new}}_{h_n} &= \left\{ m^{c_{k,n}}_{n,\text{old}} - \left[\frac{\partial^2 \log Z}{\partial [m^{c_{k,n}}_{n,\text{old}}]^2} \right]^{-1} \frac{\partial \log Z}{\partial m^{c_{k,n}}_{n,\text{old}}} \right\} a^{\text{new}}_{h_n} \end{aligned} \tag{22}$$

We now perform the analogous operations to update $\mathbf{A}_{h_n}$ and $\mathbf{b}_{h_n}$. We need to compute

$$\begin{aligned} \frac{\partial \log Z}{\partial \mathbf{m}^{f_n,f_0}_{n,\text{old}}} &= \frac{\left\{ \prod_{k=1}^{K} \Phi[\alpha_n^k] \right\} \phi(\alpha_n)}{Z\sqrt{s}} [1, -1], \\ \frac{\partial \log Z}{\partial \mathbf{V}^{f_n,f_0}_{n,\text{old}}} &= -\frac{1}{2}[1, -1]^\top [1, -1] \frac{\left\{ \prod_{k=1}^{K} \Phi[\alpha_n^k] \right\} \phi(\alpha_n)\alpha_n}{Zs}, \end{aligned}$$

where

$$s = [-1\, 1]\mathbf{V}^{f_n,f_0}_{n,\text{old}}[-1\, 1]^\top. \tag{23}$$

We then compute the optimal mean and variance (those that minimize the KL-divergence) of the product in Eq. (13) by using Eqs. (5.12) and (5.13) from Minka (2001):

$$\begin{aligned} [\mathbf{V}^{\mathbf{f}}_{f_n,f_0}]_{\text{new}} &= \mathbf{V}^{f_n,f_0}_{n,\text{old}} - \mathbf{V}^{f_n,f_0}_{n,\text{old}} \left[\frac{\partial \log Z}{\partial \mathbf{m}^{f_n,f_0}_{n,\text{old}}} \left(\frac{\partial \log Z}{\partial \mathbf{m}^{f_n,f_0}_{n,\text{old}}} \right)^\top - 2\frac{\partial \log Z}{\partial \mathbf{V}^{f_n,f_0}_{n,\text{old}}} \right] \mathbf{V}^{f_n,f_0}_{n,\text{old}}, \\ [\mathbf{m}^{\mathbf{f}}_{f_n,f_0}]_{\text{new}} &= \mathbf{m}^{f_n,f_0}_{n,\text{old}} + \mathbf{V}^{f_n,f_0}_{n,\text{old}} \frac{\partial \log Z}{\partial \mathbf{m}^{f_n,f_0}_{n,\text{old}}}. \end{aligned} \tag{24}$$

Next we need to divide the Gaussian with parameters given by Eq. (24) by the Gaussian cavity distribution $\tilde{q}^{\setminus n}(f_n, f_0, c_{1,n}, \ldots, c_{k,n})$. Therefore the new parameters $\mathbf{A}_{h_n}$ and $\mathbf{b}_{h_n}$ of the approximate factor $\tilde{h}$ are obtained using the formula for the ratio of two Gaussians:

$$
\begin{aligned}
[\mathbf{A}_{h_n}]^{\text{new}} &= \left[\mathbf{V}^{\mathbf{f}}_{f_n,f_0}\right]^{-1}_{\text{new}} - \left[\mathbf{V}^{f_n,f_0}_{n,\text{old}}\right]^{-1} . \\
[\mathbf{b}_{h_n}]^{\text{new}} &= \left[\mathbf{m}^{\mathbf{f}}_{f_n,f_0}\right]_{\text{new}} \left[\mathbf{V}^{\mathbf{f}}_{f_n,f_0}\right]^{-1}_{\text{new}} - \mathbf{m}^{f_n,f_0}_{n,\text{old}} \left[\mathbf{V}^{f_n,f_0}_{n,\text{old}}\right]^{-1} .
\end{aligned} \tag{25}
$$

3.2 EP approximation to g_k

We now show how to refine the $\{\tilde{g}_k\}$. We adjust each $\tilde{g}_k$ by minimizing the KL divergence between

$$
g_k(c_{k,0})\tilde{q}^{\setminus k}(\mathbf{f}, \mathbf{c}_1, \ldots, \mathbf{c}_K) \tag{26}
$$

and

$$
\tilde{g}_k(c_{k,0})\tilde{q}^{\setminus k}(\mathbf{f}, \mathbf{c}_1, \ldots, \mathbf{c}_K) , \tag{27}
$$

where

$$
\tilde{q}^{\setminus k}(\mathbf{f}, \mathbf{c}_1, \ldots, \mathbf{c}_K) = \frac{\tilde{q}(\mathbf{f}, \mathbf{c}_1, \ldots, \mathbf{c}_K)}{\tilde{g}_k(c_{k,0})} .
$$

Because $\tilde{q}(\mathbf{f}, \mathbf{c}_1, \ldots, \mathbf{c}_K)$ and $\tilde{g}_k(c_{k,0})$ are both Gaussian, $\tilde{q}^{\setminus k}(\mathbf{f}, \mathbf{c}_1, \ldots, \mathbf{c}_K)$ is also Gaussian. Furthermore, if we marginalize out all variables except those which $\tilde{g}_k$ depends on, namely $c_{k,0}$, then $\tilde{q}^{\setminus k}(\mathbf{f}, \mathbf{c}_1, \ldots, \mathbf{c}_K)$ takes the form $\tilde{q}^{\setminus k}(c_{k,0}) = \mathcal{N}\left(c_{k,0} \,|\, m^{c_{k,n}}_{k,\text{old}}, v^{c_{k,n}}_{k,\text{old}}\right)$, where, by applying the formula for the ratio of Gaussians, $m^{c_{k,n}}_{k,\text{old}}$ and $v^{c_{k,n}}_{k,\text{old}}$ are given by

$$
\begin{aligned}
v^{c_{k,n}}_{k,\text{old}} &= \left\{[v^{\mathbf{c}_k}_{n,n}]^{-1} - a_{g_k}\right\}^{-1} , \\
m^{c_{k,n}}_{k,\text{old}} &= \left\{m^{\mathbf{c}^k}_n [v^{\mathbf{c}^k}_{n,n}]^{-1} - b_{g_k}\right\}^{-1} .
\end{aligned} \tag{28}
$$

Similarly as in Section 3.1, to minimize the KL divergence between Eqs. (26) and (27), we match the first and second moments of these two distributions. The moments of Eq. (26) can be obtained from the derivatives of its normalization constant. This normalization constant is given by $Z = \Phi(\alpha)$ where $\alpha = m^{c_{k,n}}_{k,\text{old}} / \sqrt{v^{c_{k,n}}_{k,\text{old}}}$. Then, the derivatives are

$$
\begin{aligned}
\frac{\partial \log Z}{\partial m^{c_{k,n}}_{k,\text{old}}} &= \frac{\phi[\alpha]}{\Phi[\alpha]\sqrt{v^{c_{k,n}}_{k,\text{old}}}} \\
\frac{\partial^2 \log Z}{\partial [m^{c_{k,n}}_{k,\text{old}}]^2} &= -\frac{\phi[\alpha]}{v^{c_{k,n}}_{k,\text{old}}\Phi[\alpha]} \cdot \left\{\alpha + \frac{\phi[\alpha]}{\Phi[\alpha]}\right\}
\end{aligned} \tag{29}
$$

The update rules for a_{g_k} and b_{g_k} are then

$$
\begin{aligned}
[a_{g_k}]^{\text{new}} &= -\left(\left[\frac{\partial \log Z}{\partial m^{c_{k,n}}_{k,\text{old}}}\right]^{-1} + v^{c_{k,n}}_{k,\text{old}}\right)^{-1} \\
[b_{g_k}]^{\text{new}} &= \left\{m^{c_{k,n}}_{k,\text{old}} - \left(\frac{\partial^2 \log Z}{\partial [m^{c_{k,n}}_{k,\text{old}}]^2}\right)^{-1} \frac{\partial \log Z}{\partial m^{c_{k,n}}_{k,\text{old}}}\right\} [a_{g_k}]^{\text{new}} .
\end{aligned} \tag{30}
$$

Once EP has converged we can approximate the NFCPD using Eq. (14) in the main text:

$$
\begin{aligned}
p\left(\mathbf{z} \,|\, \mathcal{D}, \mathbf{x}, \mathbf{x}_\star\right) \approx \int & p(z_0 \,|\, \mathbf{f}) p(z_1 \,|\, \mathbf{c}_1) \cdots p(z_K \,|\, \mathbf{c}_K)\, \tilde{q}(\mathbf{f}, \mathbf{c}_1, \ldots, \mathbf{c}_K) \\
& \left(\left\{\prod_{k=1}^{K} \Theta[c_k(\mathbf{x})]\right\} \Theta[f(\mathbf{x}) - f_0] + \left\{1 - \prod_{k=1}^{K} \Theta[c_k(\mathbf{x})]\right\}\right) d\mathbf{f}\, d\mathbf{c}_1 \cdots d\mathbf{c}_K ,
\end{aligned} \tag{31}
$$

where $\mathbf{z} = (f(\mathbf{x}), c_1(\mathbf{x}), \ldots, c_K(\mathbf{x}))$, the first element in $\mathbf{z}$ is accessed using the index 0 and we have used the assumed independence of the objective and constraints to split $p(\mathbf{z} \,|\, \mathbf{f}, \mathbf{c}_1, \ldots, \mathbf{c}_K)$ into the product of the factors $p(f(\mathbf{x}) \,|\, \mathbf{f})$ and $p(c_1(\mathbf{x}) \,|\, \mathbf{c}_1), \ldots, p(c_K(\mathbf{x}) \,|\, \mathbf{c}_K)$.

4 Performing the integration in Eq. (31)

Because the constraint factor on the right-hand side of Eq. (31) only depends on f_0 out of all the integration variables, we can rewrite Eq. (31) as

$$p\left(\mathbf{z} \,|\, \mathcal{D}^f, \mathcal{D}^1, \ldots, \mathcal{D}^K, \mathbf{x}, \mathbf{x}_\star^{(m)}\right) \approx \frac{1}{Z} \int \left(\left\{ \prod_{k=1}^K \Theta[z_k] \right\} \Theta[z_0 - f_0] + \left\{ 1 - \prod_{k=1}^K \Theta[z_k] \right\} \right) \left(\int p(z_0 \,|\, \mathbf{f})\, p(z_1 \,|\, \mathbf{c}_1) \cdots p(z_K \,|\, \mathbf{c}_K)\, \tilde{q}(\mathbf{f}, \mathbf{c}_1, \ldots, \mathbf{c}_K)\, df_1 \cdots df_N\, d\mathbf{c}_1 \cdots d\mathbf{c}_K \right) df_0 \,, \tag{32}$$

where Z is a normalization constant.

We now compute the inner integral in Eq. (32) analytically. This is possible because the marginal posterior predictive distributions on $\mathbf{z}$ are Gaussian and we also have a Gaussian approximation $\tilde{q}(\mathbf{f}, \mathbf{c}_1, \ldots, \mathbf{c}_K)$. We first rewrite the inner integral in Eq. (32) by using the definition of $\tilde{q}$ in Eq. (10):

$$\int p(z_0 \,|\, \mathbf{f})\, p(z_1 \,|\, \mathbf{c}_1) \cdots p(z_K \,|\, \mathbf{c}_K) \mathcal{N}(\mathbf{f} \,|\, \mathbf{m}^{\mathbf{f}}, \mathbf{V}^{\mathbf{f}}) \prod_{k=1}^K \mathcal{N}(\mathbf{c}_k \,|\, \mathbf{m}^{\mathbf{c}_k}, \mathbf{V}^{\mathbf{c}_k}) df_1 \cdots df_N\, d\mathbf{c}_1 \cdots d\mathbf{c}_K \tag{33}$$

For each $\mathbf{c}_k$, the product of the conditional Gaussian distribution $p(c_k(\mathbf{x}) \,|\, \mathbf{c}_k)$ with the Gaussian $\mathcal{N}(\mathbf{c}_k \,|\, \mathbf{m}_k, \mathbf{V}_k)$ is an $(N+2)$-dimensional multivariate Gaussian distribution. All variables in $\mathbf{c}_k$ are then integrated out leaving only a univariate Gaussian on $c_k(\mathbf{x})$ with mean and variance m'_k and v'_k respectively. For the variables $f(\mathbf{x})$ and $\mathbf{f}$, the product of $p(f(\mathbf{x}) \,|\, \mathbf{f})$ and $\mathcal{N}(\mathbf{f} \,|\, \mathbf{m}_0, \mathbf{V}_0)$ yields an $(N+2)$-dimensional multivariate Gaussian distribution. The variables $f_1, \ldots, f_N$ are integrated out leaving only a bivariate Gaussian on the two-dimensional vector $\mathbf{f}' = (f(\mathbf{x}), f_0)$ with mean vector $\mathbf{m}'_{\mathbf{f}}$ and covariance matrix $\mathbf{V}'_{\mathbf{f}}$. Thus, the result of the inner integral in Eq. (32) is

$$\begin{aligned} &\int p(z_0 \,|\, \mathbf{f})\, p(z_1 \,|\, \mathbf{c}_1) \cdots p(z_K \,|\, \mathbf{c}_K) \mathcal{N}(\mathbf{f} \,|\, \mathbf{m}_0, \mathbf{V}_0) \prod_{k=1}^K \mathcal{N}(\mathbf{c}_k \,|\, \mathbf{m}_k, \mathbf{V}_k) df_1 \cdots df_N\, d\mathbf{c}_1 \cdots d\mathbf{c}_K \\ &= \mathcal{N}\left(\mathbf{f}' \,|\, \mathbf{m}'_{\mathbf{f}}, \mathbf{V}'_{\mathbf{f}}\right) \prod_{k=1}^K \mathcal{N}\left(c_k(\mathbf{x}) \,|\, m'_k, v'_k\right) \,, \end{aligned} \tag{34}$$

where, using Eqs. (3.22) and (3.24) of Rasmussen & Williams (2006) for the means and variances respectively, we have the definitions

$$\begin{aligned} [\mathbf{m}'_{\mathbf{f}}]_1 &= \mathbf{k}^f_{\text{final}}(\mathbf{x})^\top \left[\mathbf{K}^f_{\star,\star}\right]^{-1} \mathbf{m}^{\mathbf{f}} \,, \\ [\mathbf{m}'_{\mathbf{f}}]_2 &= \left[\mathbf{m}^{\mathbf{f}}\right]_0 \,, \\ [\mathbf{V}'_{\mathbf{f}}]_{1,1} &= k_f(\mathbf{x}, \mathbf{x}) - \mathbf{k}^f_{\text{final}}(\mathbf{x})^\top \left\{ \left[\mathbf{K}^f_{\star,\star}\right]^{-1} + \left[\mathbf{K}^f_{\star,\star}\right]^{-1} \mathbf{V}^{\mathbf{f}} \left[\mathbf{K}^f_{\star,\star}\right]^{-1} \right\} \mathbf{k}^f_{\text{final}}(\mathbf{x}) \,, \\ [\mathbf{V}'_{\mathbf{f}}]_{2,2} &= \left[\mathbf{V}^{\mathbf{f}}\right]_{0,0} \,, \\ [\mathbf{V}'_{\mathbf{f}}]_{1,2} &= k_f(\mathbf{x}, \mathbf{x}_\star^{(m)}) - \mathbf{k}^f_{\text{final}}(\mathbf{x})^\top \left\{ [\mathbf{K}^f_{\star,\star}]^{-1} + [\mathbf{K}^f_{\star,\star}]^{-1} \mathbf{V}^{\mathbf{f}} [\mathbf{K}^f_{\star,\star}]^{-1} \right\} \mathbf{k}^f_{\text{final}}(\mathbf{x}_\star^{(m)}) \,, \\ m'_k &= \mathbf{k}^k_{\text{final}}(\mathbf{x})^\top \left[\mathbf{K}^k_{\star,\star}\right]^{-1} \mathbf{m}^{\mathbf{c}_k} \,, \\ v'_k &= k_k(\mathbf{x}, \mathbf{x}) - \mathbf{k}^k_{\text{final}}(\mathbf{x})^\top \left\{ \left[\mathbf{K}^k_{\star,\star}\right]^{-1} + \left[\mathbf{K}^k_{\star,\star}\right]^{-1} \mathbf{V}^{\mathbf{c}_k} \left[\mathbf{K}^k_{\star,\star}\right]^{-1} \right\} \mathbf{k}^k_{\text{final}}(\mathbf{x}) \,, \end{aligned}$$

and

- $\mathbf{m}^{\mathbf{f}}$ and $\mathbf{V}^{\mathbf{f}}$ are the posterior mean and posterior covariance matrix of $\mathbf{f}$ given by $\tilde{q}$.

- $\mathbf{k}^f_{\text{final}}(\mathbf{x})$ is the $(N+1)$-dimensional vector with the prior cross-covariances between $f(\mathbf{x})$ and the elements of $\mathbf{f}$ given by $f(\mathbf{x}_\star), f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N)$.
- $\mathbf{K}^f_{\star,\star}$ is an $(N+1) \times (N+1)$ matrix with the prior covariances between elements of $\mathbf{f}$.
- $\mathbf{k}^f_{\text{final}}(\mathbf{x}_\star)$ is the $(N+1)$-dimensional vector with the prior cross-covariances between $f(\mathbf{x}_\star)$ and the elements of $\mathbf{f}$.
- $\mathbf{k}^k_{\text{final}}(\mathbf{x}^k)$ is an $(N+1)$ vector with the cross-covariances between $c_k(\mathbf{x})$ and the elements of $\mathbf{c}_k$ given by $c_k(\mathbf{x}_\star), c_k(\mathbf{x}_1), \ldots, c_k(\mathbf{x}_n)$.
- $\mathbf{K}^k_{\star,\star}$ is an $(N+1) \times (N+1)$ covariance matrix with the prior covariances between the elements of $\mathbf{c}_k$.

We have now computed the inner integral in Eq. (31), leaving us with our next approximation of the NFCPD:

$$
\begin{aligned}
p(f(\mathbf{x}), c_1(\mathbf{x}), \ldots, c_K(\mathbf{x}) \,|\, \mathcal{D}^f, \mathcal{D}^1, \ldots, \mathcal{D}^K, \mathbf{x}, \mathbf{x}_\star) \approx \\
\frac{1}{Z} \int \left(\left\{ \prod_{k=1}^K \Theta[c_k(\mathbf{x})] \right\} \Theta\left[f(\mathbf{x}) - f_0\right] + \left\{ 1 - \prod_{k=1}^K \Theta\left[c_k(\mathbf{x})\right] \right\} \right) \\
\left\{ \prod_{nk=1}^K \mathcal{N}\left(c_k(\mathbf{x}) \,|\, m'_k, v'_k\right) \right\} \mathcal{N}\left(\mathbf{f}' \,|\, \mathbf{m}'_\mathbf{f}, \mathbf{V}'_\mathbf{f}\right) \, df_0 \,. \qquad (35)
\end{aligned}
$$

Eq. (35) is the same as Eq. (15) in the main text. Note that the computations performed to obtain m'_k, v'_k, $\mathbf{m}'_\mathbf{f}$, and $\mathbf{V}'_\mathbf{f}$ do not depend on $\mathbf{x}$. In the following section we make a final Gaussian approximation to Eq. (35), which must be performed for every $\mathbf{x}$.

5 Final Gaussian approximation to the NFCPD, for each $\mathbf{x}$

Because the right-hand side of Eq. (35) is not tractable, we approximate it with a product of Gaussians that have the same marginal means and variances. In particular, we approximate it as

$$
p\left(\mathbf{z} \,|\, \mathcal{D}, \mathbf{x}, \mathbf{x}_\star\right) \approx \mathcal{N}\left(f(\mathbf{x}) \,|\, \mu_f^{\text{NFCPD}}(\mathbf{x}), v_f^{\text{NFCPD}}(\mathbf{x})\right) \prod_{k=1}^K \mathcal{N}\left(c_k(\mathbf{x}) \,|\, \mu_k^{\text{NFCPD}}(\mathbf{x}), v_k^{\text{NFCPD}}(\mathbf{x})\right) \,,
$$

where $\mu^f_{\text{NFCPD}}(\mathbf{x})$, $v^f_{\text{NFCPD}}(\mathbf{x})$ and $\mu^k_{\text{NFCPD}}(\mathbf{x})$ and $v^k_{\text{NFCPD}}(\mathbf{x})$ are the marginal means and marginal variances of $f(\mathbf{x})$ and $c_k(\mathbf{x})$ according to the right-hand-side of Eq. (35). Using Eqs. (5.12) and (5.13) in Minka (2001) we can compute these means and variances in terms of the derivatives of the normalization constant Z in Eq. (35), which is given by

$$
Z = \left\{ \prod_{k=1}^K \Phi[\alpha_k] \right\} \Phi[\alpha] + \left\{ 1 - \prod_{k=1}^K \Phi[\alpha_k] \right\} \qquad (36)
$$

where

$$
\begin{aligned}
s &= \left[\mathbf{V}'_\mathbf{f}\right]_{1,1} + \left[\mathbf{V}'_\mathbf{f}\right]_{2,2} - 2\left[\mathbf{V}'_\mathbf{f}\right]_{1,2} \\
\alpha &= \frac{[1, -1]\mathbf{m}'_\mathbf{f}}{\sqrt{s}} \\
\alpha_k &= \frac{m'_k}{\sqrt{v'_k}} \,.
\end{aligned}
$$

Doing so yields

$$
\begin{aligned}
v_f^{\text{NFCPD}}(\mathbf{x}) &= [\mathbf{V}'_{\mathbf{f}}]_{1,1} - \frac{\beta}{s}\left(\beta + \alpha\right)\left([\mathbf{V}'_{\mathbf{f}}]_{1,1} - [\mathbf{V}'_{\mathbf{f}}]_{1,2}\right)^2 \\
\mu_f^{\text{NFCPD}}(\mathbf{x}) &= [\mathbf{m}'_{\mathbf{f}}]_1 + \left([\mathbf{V}'_{\mathbf{f}}]_{1,1} - [\mathbf{V}'_{\mathbf{f}}]_{1,2}\right)\frac{\beta}{\sqrt{s}} \\
v_k^{\text{NFCPD}}(\mathbf{x}) &= \left\{[v'_k]^{-1} + \tilde{a}\right\}^{-1} \\
\mu_k^{\text{NFCPD}}(\mathbf{x}) &= v_{N,\text{const}}^k(\mathbf{x})\left\{[m'_k]^{-1}[v'_k]^{-1} + \tilde{b}\right\}, \qquad (37)
\end{aligned}
$$

where

$$
\begin{aligned}
\beta &= \frac{\left\{\prod_{k=1}^K \Phi[\alpha_k]\right\}\phi(\alpha)}{Z} \\
\tilde{a} &= -\left\{\frac{\partial^2 \log Z}{\partial\left[m'_k\right]^2} + v'_k\right\}^{-1} \\
\tilde{b} &= \tilde{a}\left\{m'_k + \frac{\sqrt{v'_k}}{\alpha_k + \beta_k}\right\} \\
\beta_k &= \frac{\phi(\alpha_n)}{Z\Phi(\alpha_n)}(Z-1), \\
\frac{\partial^2 \log Z}{\partial\left[m'_k\right]^2} &= -\frac{\beta_k\left\{\alpha_k + \beta_k\right\}}{v'_k}.
\end{aligned}
$$

5.1 Initialization and convergence of EP

Initially EP sets the parameters of all the approximate factors to be zero. We use at the convergence criterion that the absolute change in all parameters should be below 10^{-4}.

5.2 EP with damping

To improve convergence we use damping (Minka & Lafferty, 2002). If $\tilde{h}_n^{\text{new}}$ is the minimizer of the KL-divergence, damping entails using instead $\tilde{h}_n^{\text{damped}}$ as the new factor, as defined below:

$$
\tilde{h}_n^{\text{damped}} = [\tilde{h}_n^{\text{new}}]^{\epsilon} + \tilde{h}_n^{1-\epsilon}, \qquad (38)
$$

where $\tilde{h}_n$ is the factor at the previous iteration. We do the same for $\tilde{g}_k$. The parameter ϵ controls the amount of damping, with $\epsilon = 1$ corresponding to no damping. We initialize ϵ to 1 and multiply it by a factor of 0.99 at each iteration. Furthermore, the factors $\tilde{h}_n$ and $\tilde{g}_k$ are updated in parallel (i.e. without updating $\tilde{q}$ in between) in order to speed up convergence (Gerven et al., 2009). During the execution of EP, some covariance matrices may become non positive definite due to an excessively large step size (i.e. large ϵ). If this issue is encountered during an EP iteration, the damping parameter is reduced by half and the iteration is repeated. The EP algorithm is terminated when the absolute change in all the parameters in $\tilde{q}$ is less than 10^{-4}.

References

Gerven, Marcel V, Cseke, Botond, Oostenveld, Robert, and Heskes, Tom. Bayesian source localization with the multivariate Laplace prior. pp. 1901–1909, 2009.

Minka, Thomas and Lafferty, John. Expectation-propagation for the generative aspect model. pp. 352–359, 2002.

Minka, Thomas P. *A family of algorithms for approximate Bayesian inference*. PhD thesis, Massachusetts Institute of Technology, 2001.

Rasmussen, C. and Williams, C. *Gaussian Processes for Machine Learning*. MIT Press, 2006.