Supplementary Material for: Max-value Entropy Search for Multi-objective Bayesian Optimization with Unkown Constraints

Daniel Fernández-Sánchez Universidad Autónoma de Madrid
Eduardo C. Garrido-Merchán Universidad Autónoma de Madrid
Daniel Hernández-Lobato Universidad Autónoma de Madrid

1 Obtaining the approximate truncated Gaussians by ADF

1.1 Introduction to ADF

In this section, we will explain how Assumed Density Filtering (ADF) is used to approximate the predictive distributions conditioned to the Pareto $\mathcal{Y}^\star$ front by truncated Gaussians. ADF is a technique that is often used to calculate approximate posteriors (Boyen and Koller, 1998). When ADF is used to approximate a distribution of interest to $p(\mathbf{a})$, a distribution $q(\mathbf{a})$ is first chosen from a family of distributions that it is convenient for us to use. This $q(\mathbf{a})$ distribution is adjusted to approximate the target distribution $p(\mathbf{a})$. To adjust $q(\mathbf{a})$, ADF minimizes the Kullback-Leibler divergence between $p(\mathbf{a})$ and $q(\mathbf{a})$, i.e. it minimizes $KL(p(\mathbf{a})||q(\mathbf{a}))$. We have chosen that $q(\mathbf{a})$ is a truncated Gaussian, thus it belongs to the exponential family. Minimizing the divergence of Kullback-Leibler when we are approaching one distribution by another that belongs to the exponential family is equivalent to matching moments between the two distributions. Namely, we are going to adjust the means and variances of several truncated Gaussian distributions to approximate the predictive distributions conditioned to $\mathcal{Y}^\star$. This adjustment of means and variances is made while processing the points of a $\mathcal{Y}^\star$ sample.

1.2 ADF update equations

In this section, we will obtain the equations that allow us to update the means and variances. The expression of the predictive distribution conditioned to $\mathcal{Y}^\star$ is given by:

$$p(\mathbf{f}, \mathbf{c}|\mathcal{D}, \mathbf{x}, \mathcal{Y}^\star) = Z^{-1} p(\mathbf{f}, \mathbf{c}|\mathcal{D}, \mathbf{x}) p(\mathcal{Y}^\star|\mathbf{f}, \mathbf{c}) \tag{1}$$

since this distribution can be expressed as:

$$Z = \int t(\mathbf{a}) \mathcal{N}(\mathbf{a}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{a}, \qquad p(\mathbf{a}) = Z^{-1} t(\mathbf{a}) \mathcal{N}(\mathbf{a}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

since the factor $\Omega(\mathbf{f}^\star, \mathbf{f}, \mathbf{c})$ of $p(\mathcal{Y}^\star|\mathbf{f}, \mathbf{c})$ is $t(\mathbf{a})$ and the predictive distributions of the Gaussian processes are $\mathcal{N}(\mathbf{a}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ are the vector of means and the covariance matrix of a Gaussian, therefore we can use the following equations to iteratively adjust $q(\mathbf{a})$:

$$\begin{aligned} \mathbb{E}_{\hat{p}(\mathbf{a})}[\mathbf{a}] &= \boldsymbol{\mu} + \boldsymbol{\Sigma} \frac{\partial \log(Z)}{\partial \boldsymbol{\mu}} \\ \mathbb{E}_{\hat{p}(\mathbf{a})}[\mathbf{a}\mathbf{a}^{\mathrm{T}}] - \mathbb{E}_{\hat{p}(\mathbf{a})}[\mathbf{a}] \mathbb{E}_{\hat{p}(\mathbf{a})}[\mathbf{a}]^{\mathrm{T}} &= \boldsymbol{\Sigma} - \boldsymbol{\Sigma} \left(\frac{\partial \log(Z)}{\partial \boldsymbol{\mu}} \left(\frac{\partial \log(Z)}{\partial \boldsymbol{\mu}} \right)^{\mathrm{T}} - 2 \frac{\partial \log(Z)}{\partial \boldsymbol{\Sigma}} \right) \boldsymbol{\Sigma}. \end{aligned} \tag{2}$$

Thus, to use ADF, we must calculate Z and the partial derivatives of $\log(Z)$ with respect to the means and variances of the objectives and constraints. The calculation of Z is the following:

$$
\begin{aligned}
Z &= \int p(\mathbf{f}, \mathbf{c}|\mathcal{D}, \mathbf{x}) \Omega(\mathbf{f}^{\star}, \mathbf{f}, \mathbf{c}) d\mathbf{f} d\mathbf{c} \\
&= \int p(\mathbf{f}, \mathbf{c}|\mathcal{D}, \mathbf{x}) \left(1 - \prod_{j=0}^{C} \Theta(c_j(\mathbf{x})) \prod_{k=0}^{K} \Theta\left(f_k^{\star} - f_k(\mathbf{x})\right) \right) d\mathbf{f} d\mathbf{c} \\
&= 1 - \prod_{j=0}^{C} \Phi(\gamma_j^c) \prod_{k=0}^{K} \Phi(\gamma_k^f)
\end{aligned} \tag{3}
$$

where $\Phi(\cdot)$ is the cumulative probability distribution of Gaussian, and

$$
\gamma_k^f = \frac{f_k^{\star} - m_k^f(\mathbf{x})}{(v_k^f(\mathbf{x}))^{1/2}} \qquad\qquad \gamma_j^c = \frac{m_j^c(\mathbf{x})}{(v_j^c(\mathbf{x}))^{1/2}} \tag{4}
$$

where $m_k^f(\mathbf{x})$, $v_k^f(\mathbf{x})$, $m_k^c(\mathbf{x})$ and $v_k^c(\mathbf{x})$ are the mean and variance values of the k-th and j-th predictive distributions of the objectives and constraints at $\mathbf{x}$. On the other hand, these are the values of the derivatives $\frac{\partial \log(Z)}{\partial m_k^f}$, $\frac{\partial \log(Z)}{\partial v_k^f}$, $\frac{\partial \log(Z)}{\partial m_j^c}$ and $\frac{\partial \log(Z)}{\partial v_j^c}$:

$$
\frac{\partial \log(Z)}{\partial m_k^f} = \frac{(Z-1)}{Z\Phi(\gamma_k^f)} \left(\frac{-\mathcal{N}(\gamma_k^f|0,1)}{(v_k^f)^{1/2}} \right) \qquad \frac{\partial \log(Z)}{\partial v_k^f} = \frac{(Z-1)}{Z\Phi(\gamma_k^f)} \left(\mathcal{N}(\gamma_k^f|0,1) \frac{-\gamma_k^f}{2v_k^f} \right) \tag{5}
$$

$$
\frac{\partial \log(Z)}{\partial m_j^c} = \frac{(Z-1)}{Z\Phi(\gamma_j^c)} \left(\frac{\mathcal{N}(\gamma_j^c|0,1)}{(v_j^c)^{1/2}} \right) \qquad \frac{\partial \log(Z)}{\partial v_j^c} = \frac{(Z-1)}{Z\Phi(\gamma_j^c)} \left(\mathcal{N}(\gamma_j^f|0,1) \frac{-\gamma_j^c}{2v_j^c} \right) \tag{6}
$$

where $v_k^f = v_k^f(\mathbf{x})$ and $v_j^c = v_j^c(\mathbf{x})$.

We show the ADF algorithm in the Algorithm 1. We can see that in each iteration the values of $\tilde{\mathbf{m}}^f$, $\tilde{\mathbf{v}}^f$, $\tilde{\mathbf{m}}^c$ and $\tilde{\mathbf{v}}^c$ are updated using the values calculated in the derivatives of (5) and (6). We can also see that the order of processing the $\mathbf{f}^{\star}$ points of the Pareto front sample will influence the result of the means and variances. For this reason, we have established this order as random.

Algorithm 1: ADF Algorithm

Input: $\mathbf{m}^f$, $\mathbf{v}^f$, $\mathbf{m}^c$ and $\mathbf{v}^c$

Initialize: $\tilde{\mathbf{m}}^f = \mathbf{m}^f$, $\tilde{\mathbf{v}}^f = \mathbf{v}^f$, $\tilde{\mathbf{m}}^c = \mathbf{m}^c$ and $\tilde{\mathbf{v}}^c = \mathbf{v}^c$

for each $\mathbf{f}^{\star}$ **in** $\mathcal{Y}^{\star}$ **do**

- $\boldsymbol{\gamma}^f = (\mathbf{f}^{\star} - \mathbf{m}^f)/\sqrt{\mathbf{v}^f}$
- $\boldsymbol{\gamma}^c = \mathbf{m}^c/\sqrt{\mathbf{v}^c}$
- $Z = 1 - \prod_{j=0}^{C} \Phi(\gamma_j^c) \prod_{k=0}^{K} \Phi(\gamma_k^f)$
- $\tilde{\mathbf{m}}^f = \tilde{\mathbf{m}}^f + \tilde{\mathbf{v}}^f \frac{\partial \log(Z)}{\partial \tilde{\mathbf{m}}^f}$
- $\tilde{\mathbf{v}}^f = \tilde{\mathbf{v}}^f - \tilde{\mathbf{v}}^f \left(\frac{\partial \log(Z)}{\partial \tilde{\mathbf{m}}^f} \left(\frac{\partial \log(Z)}{\partial \tilde{\mathbf{m}}^f} \right)^{\mathrm{T}} - 2 \frac{\partial \log(Z)}{\partial \tilde{\mathbf{v}}^f} \right) \tilde{\mathbf{v}}^f$
- $\tilde{\mathbf{m}}^c = \tilde{\mathbf{m}}^c + \tilde{\mathbf{v}}^c \frac{\partial \log(Z)}{\partial \tilde{\mathbf{m}}^f}$
- $\tilde{\mathbf{v}}^c = \tilde{\mathbf{v}}^c - \tilde{\mathbf{v}}^c \left(\frac{\partial \log(Z)}{\partial \tilde{\mathbf{m}}^c} \left(\frac{\partial \log(Z)}{\partial \tilde{\mathbf{m}}^c} \right)^{\mathrm{T}} - 2 \frac{\partial \log(Z)}{\partial \tilde{\mathbf{v}}^c} \right) \tilde{\mathbf{v}}^c$

return $\tilde{\mathbf{m}}^f$, $\tilde{\mathbf{v}}^f$, $\tilde{\mathbf{m}}^c$ and $\tilde{\mathbf{v}}^c$;

References

Boyen, X. and Koller, D. (1998). Tractable inference for complex stochastic processes. *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pages 33–42.