Bayesian Quantile and Expectile Optimisation - Supplementary material

Victor Picheny[1] **Henry Moss**[1] **Léonard Torossian**[2] **Nicolas Durrande**[1]

[1]Secondmind Labs, Cambridge, UK
[2]Inria, Université Côte d'Azur, France

1 SUPPLEMENTARY MATERIAL: CALCULATION OF Q-GIBBON

We derive here the analytical form of our proposed Q-GIBBON acquisition function. For simplicity, we focus on the quantile setting, but the expectile case only requires a straightforward modification of the following derivation.

Recall that Q-GIBBON is defined as

$$\alpha_n^{\text{Q-GIBBON}} = \frac{1}{2}\log|C| - \frac{1}{2M}\sum_{g^*\in\mathcal{M}_n}\sum_{i=1}^{B}\log V_i(g^*),$$

where $|C|$ is the determinant of the $B \times B$ predictive covariance matrix with elements $C_{i,j} = \text{Cov}(y_{x_i}, y_{x_j}|\mathcal{D}_n)$ and $V(g^*)$ denotes the conditional variances $V_i(g^*) = \text{Var}(y_{x_i}|g^*, \mathcal{D}_n)$. Therefore, calculating Q-GIBBON boils down to being able to calculate $V_i(g^*)$ and $C_{i,j}$ across any candidate batch of points (i.e. for all $i, j \in \{1, .., B\}$). We now derive closed-form expressions for $V_i(g^*)$ and $C_{i,j}$.

1.1 REQUIRED PREDICTIVE QUANTITIES

For ease of notation, we will consider just a single pair of input values of x_1 and x_2 and show how to calculate $V_1(g^*)$ and $C_{1,2}$. Denote the quantiles, scales and (noisy) observations at these two location as $g_1 = g(x_1)|\mathcal{D}_n$, $g_2 = g(x_2)|\mathcal{D}_n$, $\sigma_1 = \sigma(x_1)|\mathcal{D}_n$, $\sigma_2 = \sigma(x_2)|\mathcal{D}_n$, $y_1 = y(x_1)|\mathcal{D}_n$ and $y_2 = y(x_2)|\mathcal{D}_n$, respectively. Then, from our underlying GP models we can extract our current beliefs about these random variables:

$$\begin{pmatrix} g_1 \\ g_2 \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_1^g \\ \mu_2^g \end{pmatrix}, \begin{pmatrix} (\sigma_1^g)^2 & \Sigma_{1,2}^g \\ \Sigma_{1,2}^g & (\sigma_2^g)^2 \end{pmatrix}\right],$$
$$\begin{pmatrix} \log(\sigma_1) \\ \log(\sigma_2) \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_1^\sigma \\ \mu_2^\sigma \end{pmatrix}, \begin{pmatrix} (\sigma_1^\sigma)^2 & \Sigma_{1,2}^\sigma \\ \Sigma_{1,2}^\sigma & (\sigma_2^\sigma)^2 \end{pmatrix}\right].$$

For closed form expressions of μ_1^g, σ_1^g, ... see any GP textbook, e.g. Rasmussen [2003].

Before deriving expressions for $V_1(g^*)$ and $C_{1,2}$, it is convenient to write the conditional mean and variance of our noisy observations y_1 and y_2. Following Yu and Moyeed [2001], we have

$$\mathbb{E}[y_1|g_1, \sigma_1] = g_1 + \frac{1-2\tau}{\tau(1-\tau)}\sigma_1, \quad (1)$$

$$\text{Var}(y_1|g_1, \sigma_1) = \frac{1-2\tau+2\tau^2}{\tau^2(1-\tau)^2}\sigma_1^2, \quad (2)$$

with similar expressions for the moments of $y_2|g_2, \sigma_2$

1.2 CALCULATING THE CONDITIONAL VARIANCE V

We now have all the quantities required to calculate $V_1(g^*) = \text{Var}(y|g^*)$. Recall that g^* denotes the maximal value obtained by the quantile (i.e. $g(x)$). First, we use the law of total variance to decompose V_1 into two terms:

$$\begin{aligned} V_1 =& \text{Var}_{g_1,\sigma|g^*}\left(\mathbb{E}[y_1|g_1, \sigma_1, g^*]\right) \\ &+ \mathbb{E}_{g_1,\sigma|g^*}\left[\text{Var}\left(y_1|g_1, \sigma_1, g^*\right)\right]. \end{aligned} \quad (3)$$

Note that conditioning on g_1, σ, g^* is equivalent to conditioning on g_1, σ only, as knowing that $g^* = \max g(x)$ does not provide additional information over knowing g_1 itself. Therefore, we can insert our expressions for the moments of the asymmetric Laplace (1) and (2) into (3) which, after simple manipulation provides:

$$\begin{aligned} V_1(g^*) = \text{Var}_{g_1|g^*}(g_1) + \frac{3(1-2\tau)^2+1}{2\tau^2(1-\tau)^2}\mathrm{e}^{2(\mu_1^\sigma+(\sigma_1^\sigma)^2)} \\ + \frac{(1-2\tau)^2}{2\tau^2(1-\tau)^2}\mathrm{e}^{2\mu_1^\sigma+(\sigma_1^\sigma)^2}. \end{aligned} \quad (4)$$

All that remains for the calculation of $V(g^*)_1$ is an expression for $\text{Var}_{g_1|g^*}(g_1)$. Fortunately, as shown by Wang and Jegelka [2017], $g|g^*$ is simply an upper truncated Gaussian variable. Therefore, using the well-known expression for

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

the variance of a truncated Gaussian, we have

$$\mathrm{Var}_{g_1|g^*}(g_1) = (\sigma_1^g)^2 \left(1 + \frac{\phi(\gamma_{g^*})}{\Psi(\gamma_{g^*})}\left(\gamma_{g^*} - \frac{\phi(\gamma_{g^*})}{\Psi(\gamma_{g^*})}\right)\right), \tag{5}$$

where $\gamma_{g^*} = \frac{g^* - \mu_1^g}{\sigma_1^g}$, and ϕ and Ψ are the probability density functions and cumulative density functions of a standard Gaussian variable, respectively.

Finally, inserting (5) into (4) yields a closed form expression for $V_1(g^*)$.

1.3 CALCULATING THE PREDICTIVE COVARIANCE C

Just like when calculating the conditional variance V_1, we begin our decomposition of $C_{1,2} = Cov(y_1, y_2)$ by applying the law of total variance to get the following two term expansion:

$$\begin{aligned} C_{1,2} =& \mathrm{Cov}_{g_1,g_2,\sigma_1,\sigma_2}\left(\mathbb{E}\left[y_1|g_1,\sigma_1\right], \mathbb{E}\left[y_2,g_2,\sigma_2\right]\right) \\ &+ \mathbb{E}_{g_1,g_2,\sigma_1,\sigma_2}\left[\mathrm{Cov}(y_1,y_2|g_1,g_2,\sigma_1,\sigma_2)\right]. \end{aligned} \tag{6}$$

Now, as $y_1|g_1,\sigma_1$ and $y_2|g_2,\sigma_2$ are independent (all that remains after this conditioning is observation noise), the second term of (6) is in fact zero (at least for unique x_1 and x_2).

To calculate the first term of (6), we insert the expression for the first moment of $y|g,\sigma$ (i.e. Equation (1)) which, after recalling the independence of g and σ, yields

$$\begin{aligned} C_{1,2} =& \mathrm{Cov}_{g_1,g_2}(g_1,g_2) \\ &+ \frac{(1-2\tau)^2}{\tau^2(1-\tau)^2}\mathrm{Cov}_{\sigma_1,\sigma_2}(\sigma_1,\sigma_2). \end{aligned} \tag{7}$$

Finally, we can extract $\mathrm{Cov}(g_1, g_2)$ and $\mathrm{Cov}(\sigma_1, \sigma_2)$ from our underlying GP models as $\Sigma_{1,2}^g$ and $\mathrm{e}^{\mu_1^\sigma + \mu_2^\sigma + 0.5(\sigma_1^\sigma + \sigma_2^\sigma)}(\mathrm{e}^{\Sigma_{1,2}^\sigma} - 1)$ (using the formulae for the covariance of joint log Gaussian variables). Inserting these two covariances into (7) provides a closed-from expression for $C_{1,2}$.

2 SUPPLEMENTARY MATERIAL: RFF FOR MATERN KERNELS

We present in this section how to use RFFs to generate samples from d-dimensional Matern kernels with regularity ν, variance σ^2 and lengthscales $\theta \in \mathbb{R}^d$. First of all, we start from the spectral density of a Matérn kernel:

$$s(w) = \sigma^2 |\Lambda|^{1/2} \frac{\Gamma(\frac{d}{2}+\nu)}{\Gamma(\nu)} \frac{(2\sqrt{\pi})^d}{(1 + w^T \Lambda w)^{\frac{d}{2}+\nu}},$$

where $\Lambda = \mathrm{diag}(\theta_1, \cdots, \theta_d)$ is the diagonal matrix containing the length scale hyperparameters. Using the change of variable $\Lambda' = 2\nu \times \Lambda$ and introducing rescaling factor $\sigma^2(\sqrt{2}\pi)^d$, one can recognise here the probability density function of the *multivariate t-distribution*:

$$p(w) = |\Lambda|^{1/2} \frac{\Gamma(\frac{d}{2}+\nu)}{\Gamma(\nu)\pi^{d/2}\nu^{d/2}} \frac{1}{(1 + \frac{1}{2\nu} w^T \Lambda w)^{\frac{d}{2}+\nu}}.$$

As a consequence, prior samples can be generated by computing

$$g(x) = \sigma \sqrt{2(\sqrt{2}\pi)^d / m} \sum_{i=1}^{m} \omega_i \cos(w_i^T x + b_i)$$

where $\omega_i \sim \mathcal{N}(0,1)$, $w_i \sim p$, $b_i \sim \mathcal{U}(0, 2\pi)$, and m is the number of features.

3 SUPPLEMENTARY MATERIAL: DESCRIPTION OF THE GLD SYNTHETIC CASE

Several formulations of the GLD exist, we use here the parameterisation of Freimer et al. [1988]. The GLD is defined by its quantile function:

$$Q(u) = \lambda_0 + \lambda_1 (T_1 - T_2), \tag{8}$$

with:

$$T_1 = \begin{cases} \frac{u^{\lambda_2}-1}{\lambda_2} & \textit{if } \lambda_2 \neq 0 \\ \log(u) & \textit{if } \lambda_2 = 0 \end{cases}$$

$$T_2 = \begin{cases} \frac{(1-u)^{\lambda_3}-1}{\lambda_3} & \textit{if } \lambda_3 \neq 0 \\ \log(1-u) & \textit{if } \lambda_3 = 0 \end{cases}.$$

Here, the only constraint for the parameter values is $\lambda_1 > 0$.

To define an experiment, each λ_j is a realisation of a GP, except for λ_1 for which we use a softplus transform to ensure positivity:

$$\begin{aligned} \lambda_j(x) &\sim \mathcal{GP}\big(0, k(\cdot,\cdot)\big), \quad j \in \{0,2,3\}, \\ \phi(\lambda_1(x)) &\sim \mathcal{GP}\big(0, k(\cdot,\cdot)\big), \end{aligned}$$

with $\phi^{-1}(w) = \log(1 + e^w)$. All GPs have a Matern 5/2 kernel k with unit variance. We add to $\lambda_0(x)$ a small quadratic mean function to avoid having the optimum located on the edges of the domain. We use a lengthscale of 0.5 in dimension 3 and 1.0 in dimension 6. These settings ensure that the 6-dimensional test cases do not have too many local optima.

References

Marshall Freimer, Georgia Kollia, Govind S Mudholkar, and C Thomas Lin. A study of the generalized Tukey lambda family. Communications in Statistics-Theory and Methods, 1988.

Carl Edward Rasmussen. Gaussian processes in machine learning. In Summer School on Machine Learning. Springer, 2003.

Zi Wang and Stefanie Jegelka. Max-value entropy search for efficient bayesian optimization. In International Conference on Machine Learning, 2017.

Keming Yu and Rana A Moyeed. Bayesian quantile regression. Statistics & Probability Letters, 2001.