MHGP Technical Details

Asif J. Chowdhury

June 26, 2017

1 Bound Transform

Our un-normalized posterior is $p_X(x)$. Our Gaussian Process is based on $\log p_X(x)$. At first step, we get the mode of $\log p_X(x)$ by running Bayesian Optimization on the original space. Let this mode be at point x_{mode}. The Bayesian Optimization library returns us the points it traversed. The inputs of these points are X and the outputs are Y.

For a variable X bounded in the region $(0, b)$, the transformation is given by:

$$X^T = \log \frac{X}{b - X} \tag{1}$$

and the inverse transform is:

$$iT(X^T) = b * logit^{-1}(X^T) = \frac{b}{1 + e^{-X^T}} = \frac{be^{X^T}}{e^{X^T} + 1} \tag{2}$$

And the density of the transformed variable X^T is:

$$p_{X^T}(x^T) = p_X(x) \left| \frac{\partial iT(X^T)}{\partial X^T} \right| \tag{3}$$

Now the derivative above can be written as:

$$\begin{aligned} \frac{\partial iT(X^T)}{\partial X^T} = b\frac{\partial logit^{-1}(X^T)}{\partial X^T} &= b * logit^{-1}(X^T)(1 - logit^{-1}(X^T)) \\ &= b * (\frac{e^{X^T}}{e^{X^T} + 1}) * (\frac{1}{e^{X^T} + 1}) = \frac{be^{X^T}}{(e^{X^T} + 1)^2} \end{aligned} \tag{4}$$

If one point x has n dimensions, with the i-th dimension denoted by x_i, the upper bounds are $b_1, b_2, ..., b_n$ and the lower bounds are zeros, then taking log of equation (3) would give us:

$$\begin{aligned} &\log p_{X^T}(x^T) = \log p_X(x) + \log \frac{b_1 e^{x_1^T}}{(e^{x_1^T} + 1)^2} + \ldots + \log \frac{b_n e^{x_n^T}}{(e^{x_n^T} + 1)^2} \\ &= \log p_X(x) + \log b_1 + x_1^T - 2\log(e^{x_1^T} + 1) + \ldots + \log b_n + x_n^T - 2\log(e^{x_n^T} + 1) \end{aligned} \tag{5}$$

Now we need to transform the points X, Y that were obtained from Bayesian Optimization to the unbounded space. To convert each point of X, we perform the following conversion obtained from equation (1) for each of the dimension:

$$x_i^T = \log \frac{x_i}{b_i - x_i} \tag{6}$$

To convert Y, we follow equation (5):

$$y^T = y + \log b_1 + x_1^T - 2\log\left(e^{x_1^T} + 1\right) + \ldots + \log b_n + x_n^T - 2\log\left(e^{x_n^T} + 1\right) \tag{7}$$

Next a new Gaussian Process is trained with points X^T, Y^T. Then the inverse of the negative Hessian of this GP on the point x_{mode} is calculated. This gives our covariance matrix for the proposal distribution.

2 Hessian Calculation

Let X_T be the training set of size N. The i-th training point is denoted by $X_T^{(i)}$. Let each training point be of dimension D with the k-th dimension of point X denoted by X_k. So the matrix X_T is $N \times D$. We want to take the Hessian of the GP at the mode X_{mode}. If the mean of the GP is μ then the Hessian at mode is:

$$H(X)_{atX=X_{mode}} = \frac{\partial^2 \mu}{\partial X^2}_{X=X_{mode}} = \begin{bmatrix} \frac{\partial^2\mu}{\partial X_1^2} & \frac{\partial^2\mu}{\partial X_1 \partial X_2} & \cdots & \frac{\partial^2\mu}{\partial X_1 \partial X_D} \\ \frac{\partial^2\mu}{\partial X_2 \partial X_1} & \frac{\partial^2\mu}{\partial X_2^2} & \cdots & \frac{\partial^2\mu}{\partial X_2 \partial X_D} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2\mu}{\partial X_D \partial X_1} & \frac{\partial^2\mu}{\partial X_D \partial X_2} & \cdots & \frac{\partial^2\mu}{\partial X_D^2} \end{bmatrix} \tag{8}$$

Let the kernel variance is σ^2, kernel length scale is l and noise variance is σ^2_{noise}. The mean of the GP is:

$$\mu(X) = K(X, X_T)[K(X_T, X_T) + \sigma^2_{noise} I]^{-1} Y_T \tag{9}$$

where Y_T is the $N \times 1$ vector of training set outputs, $K(X_T, X_T)$ is the $N \times N$ matrix containing pairwise kernel evaluations of the training points, and $K(X, X_T)$ is the $1 \times N$ matrix containing kernel evaluations of each of the training points with X:

$$K(X, X_T) = \left[K(X, X_T^{(1)})K(X, X_T^{(2)}) \ldots K(X, X_T^{(N)})\right]$$

where the kernel value between X and the j-th training point is given by:

$$K(X, X_T^{(j)}) = \sigma^2 exp\left[-\frac{1}{2} * (X - X_T^{(j)})^T \Lambda^{-1} (X - X_T^{(j)})\right] \tag{10}$$

where Λ is a $D \times D$ matrix given by

$$\Lambda = \begin{bmatrix} l^2 & 0 & \ldots & 0 \\ 0 & l^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & l^2 \end{bmatrix}$$

and hence:

$$\Lambda^{-1} = \begin{bmatrix} \frac{1}{l^2} & 0 & \ldots & 0 \\ 0 & \frac{1}{l^2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \frac{1}{l^2} \end{bmatrix}$$

From equation (10) we get:

$$\frac{\partial K(X, X_T^{(j)})}{\partial X} = -\Lambda^{-1}(X - X_T^{(j)})K(X, X_T^{(j)})$$
$$= \begin{bmatrix} \frac{X_{T(1)}^{(j)} - X_1}{l^2} K(X, X_T^{(j)}) \\ \frac{X_{T(2)}^{(j)} - X_2}{l^2} K(X, X_T^{(j)}) \\ \vdots \\ \frac{X_{T(D)}^{(j)} - X_D}{l^2} K(X, X_T^{(j)}) \end{bmatrix} = \begin{bmatrix} \frac{\partial K(X, X_T^{(j)})}{\partial X_1} \\ \frac{\partial K(X, X_T^{(j)})}{\partial X_2} \\ \vdots \\ \frac{\partial K(X, X_T^{(j)})}{\partial X_D} \end{bmatrix} \quad (11)$$

Now from equation (9) we get:

$$\frac{\partial \mu}{\partial X} = \frac{\partial K(X, X_T)}{\partial X}[K(X_T, X_T) + \sigma_{noise}^2 I]^{-1} Y_T$$
$$= \begin{bmatrix} \frac{\partial K^{(1)}}{\partial X_1} & \frac{\partial K^{(2)}}{\partial X_1} & \cdots & \frac{\partial K^{(N)}}{\partial X_1} \\ \frac{\partial K^{(1)}}{\partial X_2} & \frac{\partial K^{(2)}}{\partial X_2} & \cdots & \frac{\partial K^{(N)}}{\partial X_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial K^{(1)}}{\partial X_D} & \frac{\partial K^{(2)}}{\partial X_D} & \cdots & \frac{\partial K^{(N)}}{\partial X_D} \end{bmatrix} K^{-1} Y_T \quad (12)$$

where $K^{(i)}$ represents $K(X, X_T^{(i)})$ and $K = K(X_T, X_T) + \sigma_{noise}^2 I$. Considering only the i-th dimension gives:

$$\frac{\partial \mu}{\partial X_i} = \begin{bmatrix} \frac{\partial K^{(1)}}{\partial X_i} & \frac{\partial K^{(2)}}{\partial X_i} & \cdots & \frac{\partial K^{(N)}}{\partial X_i} \end{bmatrix} K^{-1} Y_T$$
$$= \begin{bmatrix} \cdots & \frac{X_{T(i)}^{(j)} - X_i}{l^2} K(X, X_T^{(j)}) & \cdots \end{bmatrix} K^{-1} Y_T \quad (13)$$

where we reached the last line using equations (10) and (11). Here $X_{T(i)}^{(j)}$ denotes the value of the i-th dimension of the j-th training point. Now to obtain the Hessian we need the second order derivatives. For the diagonals (i=k) we have:

$$\frac{\partial^2 \mu}{\partial X_i^2} = \begin{bmatrix} \cdots & \left[\frac{X_{T(i)}^{(j)} - X_i}{l^2}\right]^2 K(X, X_T^{(j)}) - \frac{K(X, X_T^{(j)})}{l^2} & \cdots \end{bmatrix} K^{-1} Y_T \quad (14)$$

And for the off-diagonal items we have:

$$\frac{\partial^2 \mu}{\partial X_i \partial X_k} = \begin{bmatrix} \cdots & \left[\frac{X_{T(i)}^{(j)} - X_i}{l^2}\right] \left[\frac{X_{T(k)}^{(j)} - X_k}{l^2}\right] K(X, X_T^{(j)}) & \cdots \end{bmatrix} K^{-1} Y_T \quad (15)$$

Now putting the results from equation (14) and (15) into equation (8) completes the Hessian.