Bézier Gaussian Processes for Tall and Wide Data

Martin Jørgensen
Department of Engineering Science
University of Oxford
martinj@robots.ox.ac.uk

Michael A. Osborne
Department of Engineering Science
University of Oxford
mosb@robots.ox.ac.uk

1 Computations in the Bézier Buttress

This section seeks to explain how the KL-divergence is computed using the Bézier buttress. It further explains more detailed parametrisation in the architecture. For completeness we here give a forward pass to compute \( \text{Var}(f(x)) \).

\[
\text{Var}(f(x)) = 1^\top \nu_1 + 1 \tilde{w}_1 \tilde{B}^2_{1,1} \cdots \tilde{w}_d \tilde{B}^2_{x_d,1} \nu_{d+1},
\]

(1)

Here we make the choice that \( \{ \tilde{w}_r \}_{i,j} := \exp(v_{i,j}) \). This ensures positive weights and hence a positive output for the variance of \( f \). \( \varsigma \) comes from the inverse squared Bernstein adjusted prior (see Section 2). \( v \) are free parameters to be inferred in the variational posterior. This parametrisation makes computing the KL terms easier.

We remark all the following calculation are only for one Bézier buttress. Are there multiple Bézier buttresses, with different orderings of layers, the computations are equivalent for all of them.

For computing the KL we first recall from the paper

\[
\text{KL}(q(P)||p(P)) = \nu_1 + \nu_2 + \cdots + \nu_d \sum_{i_1=0}^{\nu_1} \sum_{i_2=0}^{\nu_2} \cdots \sum_{i_d=0}^{\nu_d} \text{KL}(q(P_{i_1,\ldots,i_d})||p(P_{i_1,\ldots,i_d}))
\]

(2)

\[
= \nu_1 + \nu_2 + \cdots + \nu_d \left\{ \frac{\hat{\Sigma}_{i_1,\ldots,i_d}}{\Sigma_{i_1,\ldots,i_d}} - 1 + \frac{\hat{\vartheta}^2_{i_1,\ldots,i_d}}{\Sigma_{i_1,\ldots,i_d}} + \log \frac{\hat{\Sigma}_{i_1,\ldots,i_d}}{\Sigma_{i_1,\ldots,i_d}} \right\}.
\]

(3)

For easier reference we declare

\[
S_1 := \nu_1 + \nu_2 + \cdots + \nu_d \sum_{i_1=0}^{\nu_1} \sum_{i_2=0}^{\nu_2} \cdots \sum_{i_d=0}^{\nu_d} \hat{\Sigma}_{i_1,\ldots,i_d},
\]

(4)

\[
S_2 := \nu_1 + \nu_2 + \cdots + \nu_d \sum_{i_1=0}^{\nu_1} \sum_{i_2=0}^{\nu_2} \cdots \sum_{i_d=0}^{\nu_d} \frac{\hat{\vartheta}^2_{i_1,\ldots,i_d}}{\Sigma_{i_1,\ldots,i_d}},
\]

(5)

\[
S_3 := \nu_1 + \nu_2 + \cdots + \nu_d \sum_{i_1=0}^{\nu_1} \sum_{i_2=0}^{\nu_2} \cdots \sum_{i_d=0}^{\nu_d} \log \frac{\hat{\Sigma}_{i_1,\ldots,i_d}}{\Sigma_{i_1,\ldots,i_d}}.
\]

(6)

We remind again that “hat” notation refers to parameters from variational posterior \( q \). We also remind that the prior variance is given \( \Sigma_{i_1,\ldots,i_d} = \prod_{r=1}^d \varsigma_r \). Because we have included \( \varsigma \) in the parametrisation in the posterior \( \hat{\Sigma}_{i_1,\ldots,i_d} \), they are cancelling out in the expression in \( S_1 \) and \( S_3 \). We get

\[
S_1 = 1^\top \nu_1 + 1 \exp v_1 \cdots \exp v_d 1_{\nu_d+1}.
\]

(7)

where \( \exp \) is element-wise on the matrices.
For $S_3$ we make the observation, based again on $\varsigma$ cancelling out in the fraction, that

$$\log \frac{\Sigma_{i_1,i_2,\ldots,i_d}}{\Sigma_{i_1,i_2,\ldots,i_d}} = -\log \prod_{\gamma=1}^{d} \exp v_{i_{\gamma}-1,i_{\gamma},\gamma} = \sum_{\gamma=1}^{d} v_{i_{\gamma}-1,i_{\gamma},\gamma}, \quad (8)$$

That is, summing over $\log \hat{\Sigma}_{i_1,i_2,\ldots,i_d}$ is basically counting how many paths (i.e. control points) use $v_{i_{\gamma}-1,i_{\gamma},\gamma}$. That is determined as $\psi_{\gamma} = \frac{\nu_{\gamma}+1}{(\nu_{\gamma}+1)(\nu_{\gamma}+1)}$. Here $\nu_0 := 0$. Hence,

$$-S_3 = \sum_{\gamma=1}^{d} \psi_{\gamma} \bigoplus v_{\gamma}, \quad (9)$$

where $\bigoplus$ denotes summing the elements in the matrix.

Notice how the variational parametrisation of $\{w_{\gamma}\}_{i,j} := \exp(v_{i,j})\varsigma_{\gamma i}$ was carefully chosen for easily computing $S_1$ and $S_3$.

$S_2$ is more close to what described in main paper. We simply just need to square all the weights and correct with the prior variance. That is, correct with $1/\varsigma_{\gamma}$. Hence,

$$S_2 = 1_{\nu_{d}+1}^T w_{1}^2 s_{1}^{-1} \cdots w_{d}^2 s_{d}^{-1} 1_{\nu_{d}+1}, \quad (10)$$

where here $\varsigma_{\gamma}$ is the diagonal matrix with $\varsigma_{\gamma i}$ along its diagonal, for $i = 1, \ldots, \nu_{\gamma}$. Notice further here $w$ are the weights in the mean Bézier buttress.

Now

$$\text{KL}(q(P)||p(P)) = S_1 - \tau + S_2 + S_3, \quad (11)$$

all of which are computed in a single forward pass in the Bézier buttress. $\tau$ is the number of all control points (in one buttress).

## 2 Numerical results

For reproducibility we give the values used to generate Figure 4. These are given in Table 1.

| year | buzz | houseelectric | slice |
|------|------|---------------|-------|
| B20: $-3.6209 \pm 0.00$ | B20: $-0.0832 \pm 0.01$ | B20: $1.5987 \pm 0.00$ | B3: $-0.5321 \pm 1.36$ | B5: $-2.7831 \pm 1.49$ |
| B20: $9.0461 \pm 0.01$ | B20: $0.2629 \pm 0.00$ | B20: $0.0489 \pm 0.00$ | B3: $0.0761 \pm 0.01$ | B5: $0.0880 \pm 0.02$ |

| Test log-likelihood |
| Test RMSE |

Table 1: Numerical values used create Figure 4. Here is listed average and standard deviation over 3 splits. On year the test-set was not standardised to compare with baselines there.