Hierarchically Partitioned Gaussian Process Approximation: Supplementary Material

October 13, 2016

1 Letter of Revision

We thank all the reviewers of NIPS 2016 for valuable feedbacks, and appreciate the resubmission chance we are given. The following only includes the list of revisions made to the main text by reviewers' suggestions; refer to our author feedback of NIPS 2016 for our responses on reviews.

Reviewer 1

- We added training time plots and the comparisons with VSGP (variable sigma Gaussian process) algorithm to complement the experiments.
- We clarified the confusions in the text.

Reviewer 3

- We added FITC result of figure 2 experiment from main text in the supplementary material. See figure 1.
- We added the comparisons with one of the special GP implementations on time series [3].

Reviewer 4

- We modified our potentially misleading statement about the Fourier trick.

Reviewer 5

- We added the visualization of higher layers in supplementary material.

2 Hierarchically Partitioned GP Inference Algorithm

Both N input points and M inducing points are both partitioned into K blocks with possibly different sizes denoted by $\{\mathbf{x}_i\}_{i=1}^{K_1}$ and $\{\mathbf{u}_i^{(1)}\}_{i=1}^{K_1}$ respectively (the superscript (1) explicitly denotes that the inducing points are at the lowest level). It is also naturally assumed that function values are divided into blocks $\{\mathbf{f}_i\}_{i=1}^{K_1}$. They are conditionally independent to function values in other blocks given the inducing points in the corresponding block. Further, our model introduces hierarchical dependencies among blocks to represent dependency among blocks that are far apart, i.e. a hierarchy of inducing-point blocks over inducing-point blocks. In the tree structure representing the hierarchy of inducing-point blocks, the input region corresponding to a block at a specific level is defined to be the union of input regions corresponding to its children. The full joint distribution over $\{\{\mathbf{u}_i^{(h)}\}_{i=1}^{K_h}\}_{h=1}^{H}$ and $\{\mathbf{f}_i\}_{i=1}^{K_1}$ is factored, given by

$$q(\{\{\mathbf{u}_i^{(h)}\}_{i=1}^{K_h}\}_{h=1}^{H}, \{\mathbf{f}_i\}_{i=1}^{K}) = q(\{\{\mathbf{u}_i^{(h)}\}_{i=1}^{K_h}\}_{h=1}^{H}) \prod_{i=1}^{K_1} q(\mathbf{f}_i|\mathbf{u}_i^{(1)}) \tag{1}$$

$$q(\{\{\mathbf{u}_i^{(h)}\}_{i=1}^{K_h}\}_{h=1}^{H}) = q(\mathbf{u}^{(H)}) \prod_{i=1}^{K_{H-1}} q(\mathbf{u}_i^{(H-1)}|\mathbf{u}^{(H)}) \prod_{l \in children(i)} q(\mathbf{u}_l^{(H-2)}|\mathbf{u}_i^{(H-1)})... \tag{2}$$

The following is the pseudocode of inferencing hierarchically partitioned GP approximation, achieving the marginal likelihood and predictive distribution. The blue lines with '//' at the front indicate that corresponding line is a comment, giving how the black codes below is originated.

$$\begin{aligned} // \; p(\boldsymbol{u}_i^{(h)}|\boldsymbol{u}_{p=par(i)}^{(h+1)}) &= \mathcal{N}(\boldsymbol{K}_{\boldsymbol{u}_i,\boldsymbol{u}_p}\boldsymbol{K}_{\boldsymbol{u}_p,\boldsymbol{u}_p}^{-1}\boldsymbol{u}_p^{(h+1)}, \boldsymbol{K}_{\boldsymbol{u}_i,\boldsymbol{u}_i} - \boldsymbol{K}_{\boldsymbol{u}_i,\boldsymbol{u}_p}\boldsymbol{K}_{\boldsymbol{u}_p,\boldsymbol{u}_p}^{-1}\boldsymbol{K}_{\boldsymbol{u}_p,\boldsymbol{u}_i}) \\ &= \mathcal{N}(\boldsymbol{A}_i^{(h)}\boldsymbol{u}_p^{(h+1)}, \boldsymbol{Q}_i^{(h)}) \end{aligned} \tag{3}$$

$$\begin{aligned} // \; q(\mathbf{y}_i|\boldsymbol{u}_i^{(1)}) &= \mathcal{N}(\boldsymbol{K}_{\mathbf{f}_i,\boldsymbol{u}_i}\boldsymbol{K}_{\boldsymbol{u}_i,\boldsymbol{u}_i}^{-1}\boldsymbol{u}_i^{(1)}, \boldsymbol{K}_{\mathbf{f}_i,\mathbf{f}_i} - \boldsymbol{K}_{\mathbf{f}_i,\boldsymbol{u}_i}\boldsymbol{K}_{\boldsymbol{u}_i,\boldsymbol{u}_i}^{-1}\boldsymbol{K}_{\boldsymbol{u}_i,\mathbf{f}_i} + \sigma_n^2\boldsymbol{I}) \\ &= p(\mathbf{y}_i|\boldsymbol{u}_i^{(1)}) = \mathcal{N}(\boldsymbol{C}_i\boldsymbol{u}_i^{(1)}, \boldsymbol{R}_i) \end{aligned} \tag{4}$$

$$\boldsymbol{A}_i^{(h)} := \boldsymbol{K}_{\boldsymbol{u}_i^{(h)}\boldsymbol{u}_{par(i)}^{(h)}}\boldsymbol{K}_{\boldsymbol{u}_{par(i)}^{(h)}\boldsymbol{u}_{par(i)}^{(h)}}^{-1} \tag{5}$$

$$\boldsymbol{Q}_i^{(h)} := \boldsymbol{K}_{\boldsymbol{u}_i^{(h)}\boldsymbol{u}_i^{(h)}} - \boldsymbol{K}_{\boldsymbol{u}_i^{(h)}\boldsymbol{u}_{par(i)}^{(h)}}\boldsymbol{K}_{\boldsymbol{u}_{par(i)}^{(h)}\boldsymbol{u}_{par(i)}^{(h)}}^{-1}\boldsymbol{K}_{\boldsymbol{u}_{par(i)}^{(h)}\boldsymbol{u}_i^{(h)}} \tag{6}$$

$$\boldsymbol{C}_i := \boldsymbol{K}_{\mathbf{f}_i\boldsymbol{u}_i^{(1)}}\boldsymbol{K}_{\boldsymbol{u}_i^{(1)}\boldsymbol{u}_i^{(1)}}^{-1} \tag{7}$$

$$\boldsymbol{R}_i := \boldsymbol{K}_{\mathbf{f}_i\mathbf{f}_i} - \boldsymbol{K}_{\mathbf{f}_i\boldsymbol{u}_i^{(1)}}\boldsymbol{K}_{\boldsymbol{u}_i^{(1)}\boldsymbol{u}_i^{(1)}}^{-1}\boldsymbol{K}_{\boldsymbol{u}_i^{(1)}\mathbf{f}_i} + \sigma_n^2\boldsymbol{I} \tag{8}$$

```
input  : $\{\boldsymbol{C}_i, \boldsymbol{R}_i\}_i, \{\boldsymbol{A}_i^{(h)}, \boldsymbol{Q}_i^{(h)}\}_{i,h}, \{\mathbf{y}_i\}_i$
output : $\{\boldsymbol{\Sigma}_i^{(h)}, \boldsymbol{\mu}_{1,i}^{(h)}, \boldsymbol{\mu}_{2,i}^{(h)}\}_{i,h}, crc^{(H)}, cry^{(H)}, yry^{(H)}, logdetR$

Initialize $\{crc_i^{(h)}, cry_i^{(h)}, yry_i^{(h)}\}_{i,h} := \mathbf{0}, logdetR = 0$
for $i$ in 1:$K_1$ (number of blocks in level 1) do
    $crc_i^{(1)} = \boldsymbol{C}_i^\top \boldsymbol{R}_i^{-1} \boldsymbol{C}_i$
    $cry_i^{(1)} = \boldsymbol{C}_i^\top \boldsymbol{R}_i^{-1} \mathbf{y}_i$
    $yry_i^{(1)} = \mathbf{y}_i^\top \boldsymbol{R}_i^{-1} \mathbf{y}_i$
    $logdetR = logdetR + \ln |\boldsymbol{R}_i|$
end
for $h$ in 1:H-1 do
    for $i$ in 1:$K_h$ (number of blocks in level $h$) do
        Let $p$ as a parent of $i$.
        // $p(\boldsymbol{u}_i | \boldsymbol{u}_p, \mathbf{y}) = p(\boldsymbol{u}_i | \boldsymbol{u}_p, \mathbf{y}_i) = \frac{p(\mathbf{y}_i | \boldsymbol{u}_i) p(\boldsymbol{u}_i | \boldsymbol{u}_p)}{p(\mathbf{y}_i | \boldsymbol{u}_p)}$
        // $p(\boldsymbol{u}_i, \mathbf{y}_i | \boldsymbol{u}_p) = \mathcal{N}\left( \begin{bmatrix} \boldsymbol{A}_i \\ \boldsymbol{C}_i \boldsymbol{A}_i \end{bmatrix} \boldsymbol{u}_p, \begin{bmatrix} \boldsymbol{Q}_i & \boldsymbol{Q}_i \boldsymbol{C}_i^\top \\ \boldsymbol{C}_i \boldsymbol{Q}_i & \boldsymbol{R}_i + \boldsymbol{C}_i \boldsymbol{Q}_i \boldsymbol{C}_i^\top \end{bmatrix} \right)$
        // $p(\boldsymbol{u}_i | \mathbf{y}_i, \boldsymbol{u}_p) =$
        // $\mathcal{N}((\boldsymbol{Q}_i^{-1} + \boldsymbol{C}_i^\top \boldsymbol{R}_i^{-1} \boldsymbol{C}_i)^{-1} (\boldsymbol{C}_i^\top \boldsymbol{R}_i^{-1} \mathbf{y}_i + \boldsymbol{Q}_i^{-1} \boldsymbol{A}_i \boldsymbol{u}_p), (\boldsymbol{Q}_i^{-1} + \boldsymbol{C}_i^\top \boldsymbol{R}_i^{-1} \boldsymbol{C}_i)^{-1})$
        $\boldsymbol{\Sigma}_i^{(h)} = ((\boldsymbol{Q}_i^{(h)})^{-1} + crc_i^{(h)})^{-1}$
        $\boldsymbol{\mu}_{1,i}^{(h)} = \boldsymbol{\Sigma}_i^{(h)} cry_i^{(h)}$
        $\boldsymbol{\mu}_{2,i}^{(h)} = \boldsymbol{\Sigma}_i^{(h)} (\boldsymbol{Q}_i^{(h)})^{-1} \boldsymbol{A}_i^{(h)}$
        $logdetR = logdetR + \ln |\boldsymbol{\Sigma}_i^{(h)}| + \ln |\boldsymbol{Q}_i^{(h)}|$
        // $p(\mathbf{y}_p | \boldsymbol{u}_p) = \prod_{i=1}^{K_h} \int p(\boldsymbol{u}_i | \boldsymbol{u}_p) p(\mathbf{y}_i | \boldsymbol{u}_i) d\boldsymbol{u}_i$
        // $p(\mathbf{y}_p | \boldsymbol{u}_p) = \mathcal{N}\left( \begin{bmatrix} \boldsymbol{C}_1 \boldsymbol{A}_1 \\ \vdots \\ \boldsymbol{C}_{K_h} \boldsymbol{A}_{K_h} \end{bmatrix} \boldsymbol{u}_p, bkdiag(\{\boldsymbol{R}_i + \boldsymbol{C}_i \boldsymbol{Q}_i \boldsymbol{C}_i^\top\}_{i=1}^{K_h}) \right)$
        // Save intermediate values in form of $crc_p := \boldsymbol{C}_p^\top \boldsymbol{R}_p^{-1} \boldsymbol{C}_p$ and $cry_p := \boldsymbol{C}_p^\top \boldsymbol{R}_p^{-1} \mathbf{y}_p$
        // $yry_p := \mathbf{y}_p^\top \boldsymbol{R}_p^{-1} \mathbf{y}_p$ is needed in computing marginal likelihood
        $crc_p^{(h)} = crc_p^{(h)} + (\boldsymbol{A}_i^{(h)})^\top crc_i^{(h)} \boldsymbol{A}_i^{(h)} - (\boldsymbol{A}_i^{(h)})^\top crc_i^{(h)} \boldsymbol{\Sigma}_i^{(h)} crc_i^{(h)} \boldsymbol{A}_i^{(h)}$
        $cry_p^{(h)} = cry_p^{(h)} + (\boldsymbol{A}_i^{(h)})^\top cry_i^{(h)} - (\boldsymbol{A}_i^{(h)})^\top crc_i^{(h)} \boldsymbol{\Sigma}_i^{(h)} cry_i^{(h)}$
        $yry_p^{(h)} = yry_p^{(h)} + yry_i^{(h)} - (cry_i^{(h)})^\top \boldsymbol{\Sigma}_i^{(h)} cry_i^{(h)}$
    end
end
```

Algorithm 1: upward pass

After the upward pass, the following is calculated:

$$
\begin{aligned}
// \; q(\boldsymbol{u}^{(H)}) &= p(\boldsymbol{u}^{(H)}) = \mathcal{N}(\mathbf{0}, \boldsymbol{K}_{\boldsymbol{u}^{(H)},\boldsymbol{u}^{(H)}} = \boldsymbol{K}_r) && (9)\\
// \; p(\boldsymbol{u}^{(H)}|\mathbf{y}) &= \mathcal{N}((\boldsymbol{K}_r^{-1} + crc^{(H)})^{-1} cry^{(H)}, (\boldsymbol{K}_r^{-1} + crc^{(H)})^{-1}) && (10)\\
&= \mathcal{N}(\boldsymbol{m}^{(H)}, \boldsymbol{P}^{(H)}) && (11)\\
\boldsymbol{P}^{(H)} &:= (\boldsymbol{K}_r^{-1} + crc^{(H)})^{-1} && (12)\\
\boldsymbol{m}^{(H)} &:= \boldsymbol{P}^{(H)} cry^{(H)} && (13)\\
// \; p(\mathbf{y}) &= \int p(\mathbf{y}|\boldsymbol{u}^{(H)}) p(\boldsymbol{u}^{(H)}) d\boldsymbol{u}^{(H)} && (14)\\
&= \mathcal{N}(\mathbf{0}, \boldsymbol{R}^{(H)} + crc^{(H)}) && (15)\\
\ln p(\mathbf{y}) &= -\frac{N}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{K}_r^{-1} + crc^{(H)}| - \frac{1}{2}\ln|\boldsymbol{K}_r| - \frac{1}{2} logdetR && (16)\\
&\quad - \frac{1}{2}\left(yry^{(H)} - \{cry^{(H)}\}^\top (\boldsymbol{K}_r^{-1} + crc^{(H)})^{-1} cry^{(H)} \right) && (17)
\end{aligned}
$$

```
input  : $\{\boldsymbol{\Sigma}_i^{(h)}, \boldsymbol{\mu}_{1,i}^{(h)}, \boldsymbol{\mu}_{2,i}^{(h)}\}_{i,h}, \boldsymbol{m}^{(H)}, \boldsymbol{P}^{(H)}$
output : $\{\boldsymbol{m}_i^{(1)}, \boldsymbol{P}_i^{(1)}\}_i$
for $h$ in H-1:-1:1 do
    for $i$ in 1:$K_h$ do
        Let $p$ as a parent of $i$.
        // $p(\boldsymbol{u}_i|\mathbf{y}) = \int p(\boldsymbol{u}_p|\mathbf{y}) p(\boldsymbol{u}_i|\boldsymbol{u}_p, \mathbf{y}) d\boldsymbol{u}_p$
        // $p(\boldsymbol{u}_i|\mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}_{2,i}\boldsymbol{m}_p + \boldsymbol{\mu}_{1,i}, \boldsymbol{\Sigma}_i + \boldsymbol{\mu}_{2,i}\boldsymbol{P}_p\boldsymbol{\mu}_{2,i}^\top)$
        $\boldsymbol{m}_i^{(h)} = \boldsymbol{\mu}_{2,i}^{(h)} \boldsymbol{m}_p^{(h+1)} + \boldsymbol{\mu}_{1,i}^{(h)}$
        $\boldsymbol{P}_i^{(h)} = \boldsymbol{\Sigma}_i^{(h)} + \boldsymbol{\mu}_{2,i}^{(h)} \boldsymbol{P}_p^{(h+1)} (\boldsymbol{\mu}_{2,i}^{(h)})^\top$
        $\boldsymbol{P}_{ip}^{(h)} = \boldsymbol{\mu}_{2,i}^{(h)} \boldsymbol{P}_p^{(h+1)}$
    end
end
```

Algorithm 2: downward pass

Predictive distribution is then:

$$
\begin{aligned}
p(\mathbf{f}_i^*|\mathbf{y}) &= \int p(\mathbf{f}_i^*|\boldsymbol{u}_i^{(1)}) p(\boldsymbol{u}_i^{(1)}|\mathbf{y}) d\boldsymbol{u}_i^{(1)} && (18)\\
&= \mathcal{N}(\boldsymbol{C}_i \boldsymbol{m}_i^{(1)}, \boldsymbol{R}_i + \boldsymbol{C}_i \boldsymbol{P}_i^{(1)} \boldsymbol{C}_i^\top) && (19)
\end{aligned}
$$

2.1 Complexity analysis

Let J denote the average number of inducing points per block, L the average number of observations per block, K the number of blocks at the lowest level. The bottleneck of the algorithm is in computing the inverse of $\boldsymbol{R}_i$ and $\boldsymbol{Q}_i$, which have the asymptotic complexity of $\mathcal{O}(L^3K)$ and $\mathcal{O}(J^3K)$. Overall algorithm, therefore, has the complexity of $\mathcal{O}(L^3K + J^3K)$. In general, we let J be less than or equal to L so that $\mathcal{O}(L^3K)$ becomes dominant.

3 Derivative of Log Marginal Likelihood (from [1])

The derivative of the log marginal likelihood is used in optimizing hyperparameters using the BFGS algorithm. The intermediate values of upward-downward algorithm is used in the calculation. The level superscipts are omitted when there is no confusion.

$$
\begin{aligned}
\frac{d}{d\theta}\log p(\mathbf{y}|\theta) &= \frac{1}{p(\mathbf{y}|\theta)}\frac{d}{d\theta}p(\mathbf{y}|\theta) \qquad (20)\\
&= \frac{1}{p(\mathbf{y}|\theta)}\frac{d}{d\theta}\int p(\mathbf{y},\boldsymbol{u}|\theta)d\boldsymbol{u} \qquad (21)\\
&= \int d\boldsymbol{u}\frac{1}{p(\mathbf{y}|\theta)}\frac{d}{d\theta}p(\mathbf{y},\boldsymbol{u}|\theta) \qquad (22)\\
&= \int d\boldsymbol{u}\frac{1}{p(\mathbf{y}|\theta)}p(\mathbf{y},\boldsymbol{u}|\theta)\frac{d}{d\theta}\log p(\mathbf{y},\boldsymbol{u}|\theta) \qquad (23)\\
&= \int d\boldsymbol{u}p(\boldsymbol{u}|\mathbf{y},\theta)\frac{d}{d\theta}\log\left[p(\boldsymbol{u}^{(H)}|\theta)\left(\prod_{h=1}^{H}\prod_{i=1}^{K_h}p(\boldsymbol{u}_i^{(h)}|\boldsymbol{u}_{par(i)}^{(h+1)},\theta)\right)\prod_{i=1}^{K_1}p(\mathbf{y}_i|\boldsymbol{u}_i^{(1)},\theta)\right] \qquad (24)\\
&= \underbrace{\int d\boldsymbol{u}^{(H)}p(\boldsymbol{u}^{(H)}|\mathbf{y},\theta)\frac{d}{d\theta}\log p(\boldsymbol{u}^{(H)}|\theta)}_{=L_1} \qquad (25)\\
&\quad+\sum_{h=1}^{H}\sum_{i=1}^{K_h}\underbrace{\int d\boldsymbol{u}_i^{(h)}d\boldsymbol{u}_{par(i)}^{(h+1)}p(\boldsymbol{u}_{par(i)}^{(h+1)},\boldsymbol{u}_i^{(h)}|\mathbf{y},\theta)\frac{d}{d\theta}\log p(\boldsymbol{u}_i^{(h)}|\boldsymbol{u}_{par(i)}^{(h+1)},\theta)}_{=L_2} \qquad (26)\\
&\quad+\sum_{i=1}^{K_1}\underbrace{\int d\boldsymbol{u}_i^{(1)}p(\boldsymbol{u}_i^{(1)}|\mathbf{y},\theta)\frac{d}{d\theta}\log p(\mathbf{y}_i|\boldsymbol{u}_i^{(1)},\theta)}_{=L_3} \qquad (27)
\end{aligned}
$$

$$
\begin{aligned}
\frac{d}{d\theta}\log p(\boldsymbol{u}^{(H)}|\theta) &= -\frac{1}{2}\frac{d}{d\theta}\log|\boldsymbol{K}_r| - \frac{1}{2}(\boldsymbol{u}^{(H)})^\top\frac{d\boldsymbol{K}_r^{-1}}{d\theta}\boldsymbol{u}^{(H)} \qquad (28)\\
p(\boldsymbol{u}^{(H)}|\mathbf{y},\theta) &= \mathcal{N}(\boldsymbol{m}^{(H)},\boldsymbol{P}^{(H)}) \qquad (29)\\
L_1 &= -\frac{1}{2}\frac{d}{d\theta}\log|\boldsymbol{K}_r| - \frac{1}{2}Tr(\frac{d\boldsymbol{K}_r^{-1}}{d\theta}\boldsymbol{P}^{(H)}) - \frac{1}{2}(\boldsymbol{m}^{(H)})^\top\frac{d\boldsymbol{K}_r^{-1}}{d\theta}\boldsymbol{m}^{(H)} \qquad (30)
\end{aligned}
$$

$$
\begin{aligned}
\frac{d}{d\theta}\log p(\boldsymbol{u}_i^{(h)}|\boldsymbol{u}_p^{(h+1)},\theta) &= -\frac{1}{2}\frac{d}{d\theta}\left[\log|\boldsymbol{Q}_i| + \boldsymbol{u}_i^\top\boldsymbol{Q}_i^{-1}\boldsymbol{u}_i - 2\boldsymbol{u}_i^\top\boldsymbol{Q}_i^{-1}\boldsymbol{A}_i\boldsymbol{u}_p + \boldsymbol{u}_p^\top\boldsymbol{A}_i^\top\boldsymbol{Q}_i^{-1}\boldsymbol{A}_i\boldsymbol{u}_p\right] \qquad (31)\\
p(\boldsymbol{u}_i^{(h)},\boldsymbol{u}_p^{(h+1)}|\mathbf{y},\theta) &= \mathcal{N}\left(\begin{bmatrix}\boldsymbol{m}_i^{(h)}\\ \boldsymbol{m}_p^{(h+1)}\end{bmatrix},\begin{bmatrix}\boldsymbol{P}_i^{(h)} & \boldsymbol{P}_{ip}^{(h)}\\ (\boldsymbol{P}_{ip}^{(h)})^\top & \boldsymbol{P}_p^{(h+1)}\end{bmatrix}\right) \qquad (32)\\
L_2 &= L_{21}+L_{22}+L_{23}+L_{24} \qquad (33)\\
L_{21} &= -\frac{1}{2}\frac{d}{d\theta}\log|\boldsymbol{Q}_i| \qquad (34)\\
L_{22} &= -\frac{1}{2}Tr(\frac{d\boldsymbol{Q}_i^{-1}}{d\theta}\boldsymbol{P}_i) - \frac{1}{2}\boldsymbol{m}_i^\top\frac{d\boldsymbol{Q}_i^{-1}}{d\theta}\boldsymbol{m}_i \qquad (35)\\
L_{23} &= -\frac{1}{2}Tr(\frac{d\boldsymbol{A}_i^\top\boldsymbol{Q}_i^{-1}\boldsymbol{A}_i}{d\theta}\boldsymbol{P}_p) - \frac{1}{2}\boldsymbol{m}_p^\top\frac{d\boldsymbol{A}_i^\top\boldsymbol{Q}_i^{-1}\boldsymbol{A}_i}{d\theta}\boldsymbol{m}_p \qquad (36)\\
L_{24} &= Tr(\boldsymbol{P}_{ip}^\top\frac{d\boldsymbol{Q}_i^{-1}\boldsymbol{A}_i}{d\theta}) + \boldsymbol{m}_i^\top\frac{d\boldsymbol{Q}_i^{-1}\boldsymbol{A}_i}{d\theta}\boldsymbol{m}_p \qquad (37)\\
& \qquad (38)
\end{aligned}
$$

$$\frac{d}{d\theta}\log p(\mathbf{y}_i|\boldsymbol{u}_i^{(1)},\theta) = -\frac{1}{2}\frac{d}{d\theta}\left[\log|\boldsymbol{R}_i| + \mathbf{y}_i^\top \boldsymbol{R}_i^{-1}\mathbf{y}_i - 2\mathbf{y}_i^\top \boldsymbol{R}_i^{-1}\boldsymbol{C}_i\boldsymbol{u}_i + \boldsymbol{u}_i^\top \boldsymbol{C}_i^\top \boldsymbol{R}_i^{-1}\boldsymbol{C}_i\boldsymbol{u}_i\right] \tag{39}$$

$$p(\boldsymbol{u}_i^{(1)}|\mathbf{y},\theta) = \mathcal{N}\left(\boldsymbol{m}_i^{(1)},\boldsymbol{P}_i^{(1)}\right) \tag{40}$$

$$L_3 = L_{31} + L_{32} + L_{33} + L_{34} \tag{41}$$

$$L_{31} = -\frac{1}{2}\frac{d}{d\theta}\log|\boldsymbol{R}_i| \tag{42}$$

$$L_{32} = -\frac{1}{2}\mathbf{y}_i^\top \frac{d\boldsymbol{R}_i^{-1}}{d\theta}\mathbf{y}_i \tag{43}$$

$$L_{33} = -\frac{1}{2}Tr(\frac{d\boldsymbol{C}_i\boldsymbol{R}_i^{-1}\boldsymbol{C}_i}{d\theta}\boldsymbol{P}_l) - \frac{1}{2}\boldsymbol{m}_i^\top \frac{d\boldsymbol{C}_i^\top \boldsymbol{R}_i^{-1}\boldsymbol{C}_i}{d\theta}\boldsymbol{m}_i \tag{44}$$

$$L_{34} = \mathbf{y}_i^\top \frac{d\boldsymbol{R}_i^{-1}\boldsymbol{C}_i}{d\theta}\boldsymbol{m}_i \tag{45}$$

$$\tag{46}$$

4 Calculating Cross-covariance (from [2])

In case of the using stationary covariance functions that are represented in the following form:

$$k_1(\boldsymbol{x},\boldsymbol{x'}) = \int_{-\infty}^{\infty} g_1(\boldsymbol{x}-\boldsymbol{u})g_1(\boldsymbol{x'}-\boldsymbol{u})d\boldsymbol{u} \tag{47}$$

$$k_2(\boldsymbol{x},\boldsymbol{x'}) = \int_{-\infty}^{\infty} g_2(\boldsymbol{x}-\boldsymbol{u})g_2(\boldsymbol{x'}-\boldsymbol{u})d\boldsymbol{u} \tag{48}$$

where $g(\boldsymbol{u}) = g(-\boldsymbol{u})$. We can define cross-covariance as:

$$k((\boldsymbol{x},1),(\boldsymbol{x}',2)) = \int_{-\infty}^{\infty} g_1(\boldsymbol{x}-\boldsymbol{u})g_2(\boldsymbol{x}'-\boldsymbol{u})d\boldsymbol{u} \tag{49}$$

$$\tag{50}$$

so that the overall covariance matrix can remain positive semi-definite.

Since we assumed $g(\boldsymbol{u}) = g(-\boldsymbol{u})$,

$$k(\boldsymbol{x},\boldsymbol{x'}) = \int_{\boldsymbol{R^D}} g(\boldsymbol{x}-\boldsymbol{u})g(\boldsymbol{u}-\boldsymbol{x'})d\boldsymbol{u} \tag{51}$$

$$= \int_{\boldsymbol{R^D}} g(\boldsymbol{x}-\boldsymbol{x'}-\boldsymbol{u})g(\boldsymbol{u})d\boldsymbol{u} \tag{52}$$

$$= \int_{\boldsymbol{R^D}} g(\boldsymbol{\tau}-\boldsymbol{u})g(\boldsymbol{u})d\boldsymbol{u} \tag{53}$$

$$= (g * g)(\boldsymbol{\tau}) \tag{54}$$

where * stands for convolution. Here we can use the fact that the Fourier transform of the convolution of two functions can be expressed by the product of the Fourier transforms of the functions being convoluted. Defining the Fourier transform as:

$$h^*(\boldsymbol{s}) = \boldsymbol{F}_{\boldsymbol{x}\to\boldsymbol{s}}[h(\boldsymbol{x})] = (2\pi)^{-\frac{D}{2}}\int_{\boldsymbol{R^D}} h(\boldsymbol{x})\exp(i\boldsymbol{s}\cdot\boldsymbol{x})d\boldsymbol{x} \tag{55}$$

$$h(\boldsymbol{x}) = \boldsymbol{F}_{\boldsymbol{s}\to\boldsymbol{x}}^{-1}[h^*(\boldsymbol{s})] = (2\pi)^{-\frac{D}{2}}\int_{\boldsymbol{R^D}} h^*(\boldsymbol{s})\exp(-i\boldsymbol{s}\cdot\boldsymbol{x})d\boldsymbol{s} \tag{56}$$

we have that

$$(g_1 * g_2)^*(\boldsymbol{s}) = (2\pi)^{\frac{D}{2}} g_1^*(\boldsymbol{s}) g_2^*(\boldsymbol{s}) \tag{57}$$

$$k^*(\boldsymbol{s}) = (2\pi)^{\frac{D}{2}} (g^*(\boldsymbol{s}))^2 \tag{58}$$

and we can compute the basis function using the covariance function as:

$$g(\boldsymbol{\tau}) = (2\pi)^{-\frac{D}{4}} \boldsymbol{F}^{-1}_{\boldsymbol{s}\to\boldsymbol{\tau}} \left[\sqrt{\boldsymbol{F}_{\boldsymbol{\tau}\to\boldsymbol{s}}[k(\boldsymbol{\tau})]} \right] \tag{59}$$

We are computing the cross-covariance between two squared exponential kernels:

$$k_{SE_1}(\boldsymbol{x}, \boldsymbol{x}'; \sigma_1, \boldsymbol{P}_1) = \sigma_1^2 \exp\left[-\frac{(\boldsymbol{x}-\boldsymbol{x}')^\top \boldsymbol{P}_1^{-1} (\boldsymbol{x}-\boldsymbol{x}')}{2} \right] \tag{60}$$

$$k_{SE_2}(\boldsymbol{x}, \boldsymbol{x}'; \sigma_2, \boldsymbol{P}_2) = \sigma_2^2 \exp\left[-\frac{(\boldsymbol{x}-\boldsymbol{x}')^\top \boldsymbol{P}_2^{-1} (\boldsymbol{x}-\boldsymbol{x}')}{2} \right] \tag{61}$$

Appying Fourier transformation:

$$\begin{aligned}
\boldsymbol{F}_{\boldsymbol{\tau}\to\boldsymbol{s}}[k_{SE_1}(\boldsymbol{\tau})] &= (2\pi)^{-\frac{D}{2}} \int_{\boldsymbol{R^D}} \sigma_1^2 \exp\left[-\frac{\boldsymbol{\tau}^\top \boldsymbol{P}_1^{-1} \boldsymbol{\tau}}{2} \right] \exp(i\boldsymbol{s}\cdot\boldsymbol{\tau}) d\boldsymbol{\tau} && (62)\\
&= (2\pi)^{-\frac{D}{2}} \sigma_1^2 \int_{\boldsymbol{R^D}} \exp\left[-\frac{\boldsymbol{\tau}^\top \boldsymbol{P}_1^{-1} \boldsymbol{\tau}}{2} + i\boldsymbol{s}^\top \boldsymbol{\tau} \right] d\boldsymbol{\tau} && (63)\\
&= (2\pi)^{-\frac{D}{2}} \sigma_1^2 \int_{\boldsymbol{R^D}} \exp\left[-\frac{1}{2}(\boldsymbol{\tau} - i\boldsymbol{P}_1\boldsymbol{s})^\top \boldsymbol{P}_1^{-1} (\boldsymbol{\tau} - i\boldsymbol{P}_1\boldsymbol{s}) - \frac{1}{2}\boldsymbol{s}^\top \boldsymbol{P}_1 \boldsymbol{s} \right] d\boldsymbol{\tau} && (64)\\
&= \sigma_1^2 |\boldsymbol{P}_1|^{\frac{1}{2}} \exp\left[-\frac{1}{2}\boldsymbol{s}^\top \boldsymbol{P}_1 \boldsymbol{s} \right] && (65)\\
g^*_{SE_1}(\boldsymbol{s}) &= \left((2\pi)^{-\frac{D}{2}} k^*_{SE_1}(\boldsymbol{s}) \right)^{\frac{1}{2}} && (66)\\
&= (2\pi)^{-\frac{D}{4}} \sigma_1 |\boldsymbol{P}_1|^{\frac{1}{4}} \exp\left[-\frac{1}{4}\boldsymbol{s}^\top \boldsymbol{P}_1 \boldsymbol{s} \right] && (67)
\end{aligned}$$

Then we apply the inverse Fourier transformation:

$$\begin{aligned}
\boldsymbol{F}^{-1}_{\boldsymbol{s}\to\boldsymbol{\tau}}[g^*_{SE_1}(\boldsymbol{s})] &= (2\pi)^{-\frac{3D}{4}} \sigma_1 |\boldsymbol{P}_1|^{\frac{1}{4}} \int_{\boldsymbol{R^D}} \exp\left[-\frac{1}{4}\boldsymbol{s}^\top \boldsymbol{P}_1 \boldsymbol{s} \right] \exp(-i\boldsymbol{s}\cdot\boldsymbol{\tau}) d\boldsymbol{s} && (68)\\
&= (2\pi)^{-\frac{3D}{4}} \sigma_1 |\boldsymbol{P}_1|^{\frac{1}{4}} \int_{\boldsymbol{R^D}} \exp\left[-\frac{1}{2}(\boldsymbol{s}^\top \frac{\boldsymbol{P}_1}{2} \boldsymbol{s} + 2i\boldsymbol{s}^\top \boldsymbol{\tau}) \right] d\boldsymbol{s} && (69)\\
&= (2\pi)^{-\frac{3D}{4}} \sigma_1 |\boldsymbol{P}_1|^{\frac{1}{4}} \int_{\boldsymbol{R^D}} \exp\left[-\frac{1}{2}(\boldsymbol{s} + 2i\boldsymbol{P}_1^{-1}\boldsymbol{\tau})^\top \frac{\boldsymbol{P}_1}{2} (\boldsymbol{s} + 2i\boldsymbol{P}_1^{-1}\boldsymbol{\tau}) - \boldsymbol{\tau}^\top \boldsymbol{P}_1^{-1} \boldsymbol{\tau} \right] d\boldsymbol{s} && (70)\\
&= (2\pi)^{-\frac{3D}{4}} \sigma_1 |\boldsymbol{P}_1|^{\frac{1}{4}} (2\pi)^{\frac{D}{2}} 2^{\frac{D}{2}} |\boldsymbol{P}_1|^{-\frac{1}{2}} \exp\left[-\boldsymbol{\tau}^\top \boldsymbol{P}_1^{-1} \boldsymbol{\tau} \right] && (71)\\
&= 2^{\frac{D}{4}} \pi^{-\frac{D}{4}} \sigma_1 |\boldsymbol{P}_1|^{-\frac{1}{4}} \exp\left[-\boldsymbol{\tau}^\top \boldsymbol{P}_1^{-1} \boldsymbol{\tau} \right] = g_{SE_1}(\boldsymbol{\tau}) && (72)
\end{aligned}$$

We can check whether we can reconstruct the original SE kernel:

$$
\begin{aligned}
k(\boldsymbol{\tau}) &= \int_{\boldsymbol{R^D}} g_{SE_1}(\boldsymbol{\tau} - \boldsymbol{u}) g_{SE_1}(\boldsymbol{u}) d\boldsymbol{u} && (73)\\
&= 2^{\frac{D}{2}} \pi^{-\frac{D}{2}} |\boldsymbol{P}_1|^{-\frac{1}{2}} \sigma_1^2 \int_{\boldsymbol{R^D}} \exp\left[-(\boldsymbol{\tau} - \boldsymbol{u})^\top \boldsymbol{P}_1^{-1} (\boldsymbol{\tau} - \boldsymbol{u}) - \boldsymbol{u}^\top \boldsymbol{P}_1^{-1} \boldsymbol{u}\right] d\boldsymbol{u} && (74)\\
&= 2^{\frac{D}{2}} \pi^{-\frac{D}{2}} |\boldsymbol{P}_1|^{-\frac{1}{2}} \sigma_1^2 \int_{\boldsymbol{R^D}} \exp\left[-2\boldsymbol{u}^\top \boldsymbol{P}_1^{-1} \boldsymbol{u} + 2\boldsymbol{\tau}^\top \boldsymbol{P}_1^{-1} \boldsymbol{u} - \boldsymbol{\tau}^\top \boldsymbol{P}_1^{-1} \boldsymbol{\tau}\right] d\boldsymbol{u} && (75)\\
&= 2^{\frac{D}{2}} \pi^{-\frac{D}{2}} |\boldsymbol{P}_1|^{-\frac{1}{2}} \sigma_1^2 \int_{\boldsymbol{R^D}} \exp\left[-\frac{1}{2}\left(\boldsymbol{u} - \frac{1}{2}\boldsymbol{\tau}\right)^\top \left(\frac{\boldsymbol{P}_1}{4}\right)^{-1} \left(\boldsymbol{u} - \frac{1}{2}\boldsymbol{\tau}\right) - \frac{1}{2}\boldsymbol{\tau}^\top \boldsymbol{P}_1^{-1} \boldsymbol{\tau}\right] d\boldsymbol{u} && (76)\\
&= 2^{\frac{D}{2}} \pi^{-\frac{D}{2}} |\boldsymbol{P}_1|^{-\frac{1}{2}} \sigma_1^2 (2\pi)^{\frac{D}{2}} 4^{-\frac{D}{2}} |\boldsymbol{P}_1|^{\frac{1}{2}} \exp\left[-\frac{1}{2}\boldsymbol{\tau}^\top \boldsymbol{P}_1^{-1} \boldsymbol{\tau}\right] && (77)\\
&= \sigma_1^2 \exp\left[-\frac{1}{2}\boldsymbol{\tau}^\top \boldsymbol{P}_1^{-1} \boldsymbol{\tau}\right] && (78)
\end{aligned}
$$

Finally, the cross-covariance would be:

$$
\begin{aligned}
&k((\boldsymbol{x}, 1), (\boldsymbol{x}', 2)) && (79)\\
&\quad = \int_{-\infty}^{\infty} g_1(\boldsymbol{\tau} - \boldsymbol{u}) g_2(\boldsymbol{u}) d\boldsymbol{u} && (80)\\
&\quad = 2^{\frac{D}{2}} \pi^{-\frac{D}{2}} |\boldsymbol{P}_1|^{-\frac{1}{4}} |\boldsymbol{P}_2|^{-\frac{1}{4}} \sigma_1 \sigma_2 \\
&\qquad\qquad \cdot \int_{\boldsymbol{R^D}} \exp\left[-(\boldsymbol{\tau} - \boldsymbol{u})^\top \boldsymbol{P}_1^{-1} (\boldsymbol{\tau} - \boldsymbol{u}) - \boldsymbol{u}^\top \boldsymbol{P}_2^{-1} \boldsymbol{u}\right] d\boldsymbol{u} && (81)
\end{aligned}
$$

$$
\begin{aligned}
&-(\boldsymbol{\tau} - \boldsymbol{u})^\top \boldsymbol{P}_1^{-1} (\boldsymbol{\tau} - \boldsymbol{u}) - \boldsymbol{u}^\top \boldsymbol{P}_2^{-1} \boldsymbol{u} && (82)\\
&\quad = -\frac{1}{2}(\boldsymbol{u} - (\boldsymbol{P}_1^{-1} + \boldsymbol{P}_2^{-1})^{-1} \boldsymbol{P}_1^{-1} \boldsymbol{\tau})^\top 2(\boldsymbol{P}_1^{-1} + \boldsymbol{P}_2^{-1})(\boldsymbol{u} - (\boldsymbol{P}_1^{-1} + \boldsymbol{P}_2^{-1})^{-1} \boldsymbol{P}_1^{-1} \boldsymbol{\tau})\\
&\qquad + \boldsymbol{\tau}^\top \boldsymbol{P}_1^{-1} (\boldsymbol{P}_1^{-1} + \boldsymbol{P}_2^{-1})^{-1} \boldsymbol{P}_1^{-1} \boldsymbol{\tau} - \boldsymbol{\tau}^\top \boldsymbol{P}_1^{-1} \boldsymbol{\tau} && (83)\\
&\quad = -\frac{1}{2}(\boldsymbol{u} - (\boldsymbol{P}_1^{-1} + \boldsymbol{P}_2^{-1})^{-1} \boldsymbol{P}_1^{-1} \boldsymbol{\tau})^\top 2(\boldsymbol{P}_1^{-1} + \boldsymbol{P}_2^{-1})(\boldsymbol{u} - (\boldsymbol{P}_1^{-1} + \boldsymbol{P}_2^{-1})^{-1} \boldsymbol{P}_1^{-1} \boldsymbol{\tau})\\
&\qquad - \boldsymbol{\tau}^\top (\boldsymbol{P}_1 + \boldsymbol{P}_2)^{-1} \boldsymbol{\tau} && (84)
\end{aligned}
$$

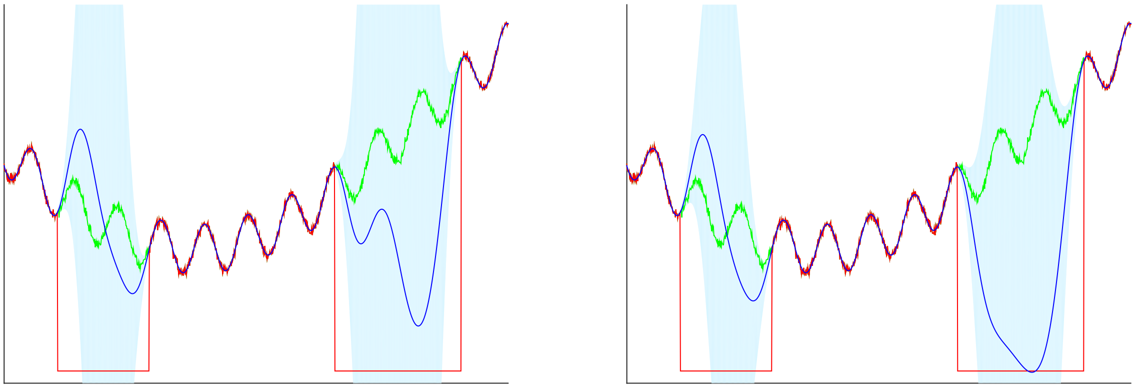
$$
\begin{aligned}
&k((\boldsymbol{x}, 1), (\boldsymbol{x}', 2)) && (85)\\
&\quad = 2^{\frac{D}{2}} \pi^{-\frac{D}{2}} |\boldsymbol{P}_1|^{-\frac{1}{4}} |\boldsymbol{P}_2|^{-\frac{1}{4}} \sigma_1 \sigma_2 (2\pi)^{\frac{D}{2}} 2^{-\frac{D}{2}} |\boldsymbol{P}_1^{-1} + \boldsymbol{P}_2^{-1}|^{-\frac{1}{2}} && (86)\\
&\qquad\qquad \cdot \exp\left[-\boldsymbol{\tau}^\top (\boldsymbol{P}_1 + \boldsymbol{P}_2)^{-1} \boldsymbol{\tau}\right] && (87)\\
&\quad = 2^{\frac{D}{2}} \sigma_1 \sigma_2 \frac{|\boldsymbol{P}_1|^{\frac{1}{4}} |\boldsymbol{P}_2|^{\frac{1}{4}}}{|\boldsymbol{P}_1 + \boldsymbol{P}_2|^{\frac{1}{2}}} \exp\left[-(\boldsymbol{x} - \boldsymbol{x}')^\top (\boldsymbol{P}_1 + \boldsymbol{P}_2)^{-1} (\boldsymbol{x} - \boldsymbol{x}')\right] && (88)
\end{aligned}
$$

References

[1] Thang D Bui and Richard E Turner. Tree-structured Gaussian process approximations. In *Advances in Neural Information Processing Systems*, pages 2213–2221, 2014.

[2] Arman Melkumyan and Fabio Ramos. Multi-kernel Gaussian processes. In *International Joint Conference on Artificial Intelligence*, volume 22, 2011.

[3] Simo Sarkka, Arno Solin, and Jouni Hartikainen. Spatiotemporal learning via infinite-dimensional bayesian filtering and smoothing: A look at gaussian process regression through kalman filtering. *IEEE Signal Processing Magazine*, 30(4):51–61, 2013.



(a) Full GP
(smse: 2.215)

(b) FITC
(smse: 3.245)

Figure 1: GP algorithms on synthetic data omitted in main text.

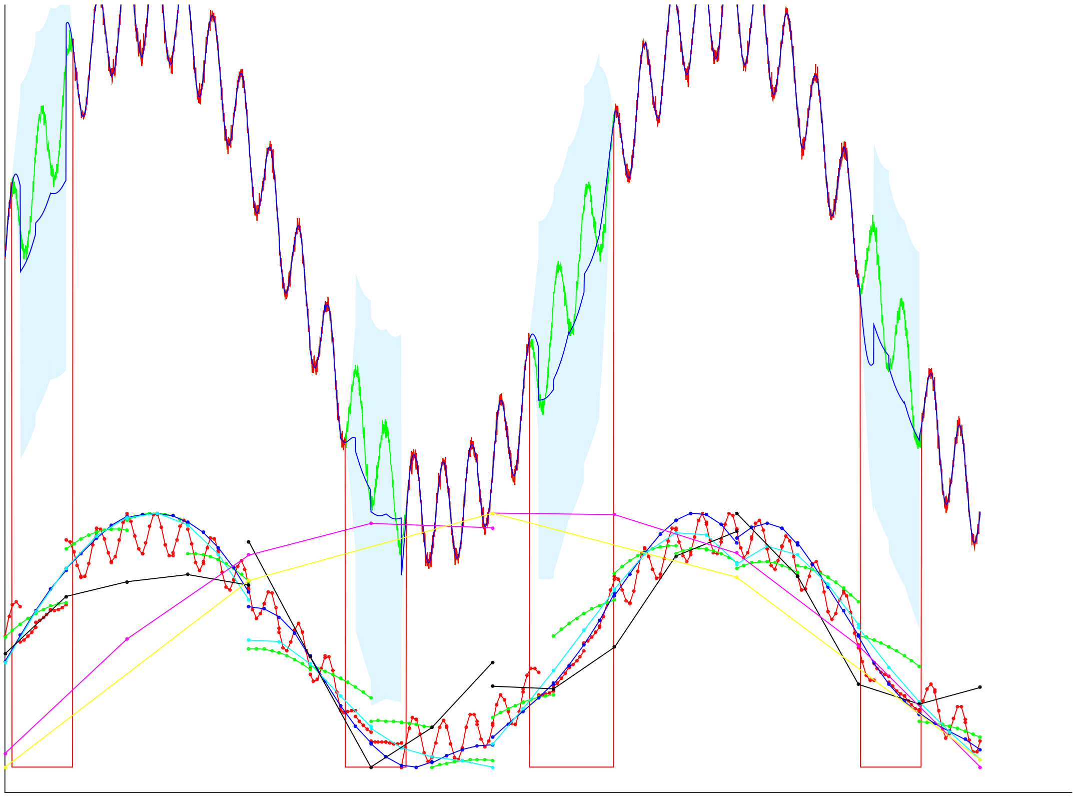


Figure 2: HPGPA algorithm on synthetic data with inducing points visualized. Note that the values of inducing points in each layer are normalized for clarity. Starting from top layer of yellow points, followed by magenta, black, cyan, blue, green, and red points at its lowest level.