# Explaining SPGP (4)

| Bishop | (article) SPGP | What |
|---|---|---|
| $x_n$ | $x_n$ | input values |
| $y_n$ | $f_n$ | GP function output |
| $t_n$ | $y_n$ | observation / target value |

By GP def: • $p(\mathbf{f}|X) = \mathcal{N}(\mathbf{m}, K)$ were $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ is always the case.

• No prior knowledge of $f_n$ ⇒ can take $\mathbf{m} = \mathbf{0}$. (bishop p 305) ⇒ $\boxed{p(\mathbf{f}|X) = \mathcal{N}(\mathbf{0}, K)}$ $p(x)$

Note in GP for regression (SPGP p2, Bishop p 306):

Assume a Gaussian noise model: $y_n = f_n + \varepsilon_n$ where $\varepsilon_n = \mathcal{N}(\varepsilon_n | 0, \sigma^2)$

GP function values

observation / tgt value

$\boxed{p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma^2 I)}$ $p(y|x)$

then $p(\mathbf{y}|X, \sigma^2) = \int_{\mathbf{f}} p(\mathbf{y}|\mathbf{f})\, p(\mathbf{f}|X)\, d\mathbf{f}$ = {Bishop 2.115 p 93}

$$= \mathcal{N}(I0 + 0, \sigma^2 I + IKI^T) = \mathcal{N}(\mathbf{0}, K + \sigma^2 I)$$

Let's consider pseudo examples only, $\bar{x}$, $\bar{\mathbf{f}}$ generated from $X$, $\mathbf{f}$. Since pseudo inputs are constructed from the input data (they are 'artificial' training examples) we can simply impose for the pseudo inputs that their observations $\bar{y}$ are fully explained by the GP (no noise addition from GP output $\bar{\mathbf{f}}$ to 'observed' value $\bar{\mathbf{y}}$ for the pseudo data)

$p(y|x)$ $\boxed{p(\bar{\mathbf{y}}|\bar{\mathbf{f}}) = \mathcal{N}(\bar{\mathbf{y}}|\bar{\mathbf{f}}, 0 \cdot I)}$ where

$p(x)$ $\boxed{p(\bar{\mathbf{f}}|\bar{X}) = \mathcal{N}(\bar{\mathbf{f}}|\mathbf{0}, K_M)}$ } ⇒ {from Bishop 2.115 p93}

$$\Rightarrow p(\bar{\mathbf{y}}|\bar{X}) = \int_{\bar{\mathbf{f}}} p(\bar{\mathbf{y}}|\bar{\mathbf{f}})\, p(\bar{\mathbf{f}}|\bar{X})\, d\bar{\mathbf{f}} = \mathcal{N}(I0 + 0, 0I + IK_M I^T) = \mathcal{N}(\mathbf{0}, K_M)$$

predictive distribution

Now we observe the input values of one of the real training examples $\mathbf{x}$. We want to find $p(y)$ conditioned on the pseudo inputs $\bar{X}, \bar{\mathbf{f}}$ and $\mathbf{x}$ (and any hyperparams $\theta$), $p(y|\mathbf{x}, \bar{X}, \bar{\mathbf{f}})$. Note that the new example is noisy $y = f + \varepsilon \Rightarrow p(f|x) = \mathcal{N}(0, k(x,x))$. $p(y|f) = \mathcal{N}(f, \sigma^2) \Rightarrow p(y|x) = \mathcal{N}(0, \sigma^2 + k(x,x))$

$\bar{\mathbf{f}} = \bar{\mathbf{y}}$

Form: • $y_{M+1} = [y \; y^1 \; y^2 \ldots y^M]^T$

• $K_{M+1} = \begin{bmatrix} k(\mathbf{x},\mathbf{x}) + \sigma^2 & k(\mathbf{x}, \mathbf{x}^{1:M})^T \\ k(\mathbf{x}^{1:M}, \mathbf{x}) & K_M \end{bmatrix} = \left\{ \begin{array}{l} K_{xx} = k(\mathbf{x},\mathbf{x}) \\ \text{using SPGP notation } k_x = k(\mathbf{x}^{1:M}, \mathbf{x}) \\ \text{and noting } k(\mathbf{x}, \mathbf{x}^{1:M}) = k(\mathbf{x}^{1:M}, \mathbf{x}) \\ \text{(since covar matrix must be symmetric} \end{array} \right\} = \begin{array}{cc} \Sigma_{aa} & \Sigma_{ab} \\ \multicolumn{2}{c}{\begin{bmatrix} K_{xx} + \sigma^2 & k_{\mathbf{x}}^T \\ k_{\mathbf{x}} & K_M \end{bmatrix}} \\ \Sigma_{ba} & \Sigma_{bb} \end{array}$

then from Bishop p 85-87:

$p(y|\mathbf{x}, \bar{X}, \bar{\mathbf{f}}) = \mathcal{N}(\mu_{a|b}, \Sigma_{a|b})$ where $\mu_{a|b} = 0 + k_{\mathbf{x}}^T K_M^{-1}(\bar{\mathbf{f}} - 0)$

$$\Sigma_{a|b} = K_{xx} + \sigma^2 - k_{\mathbf{x}}^T K_M^{-1} k_{\mathbf{x}}$$

$$\Rightarrow p(y|\mathbf{x}, \bar{X}, \bar{\mathbf{f}}) = \mathcal{N}(y | k_{\mathbf{x}}^T K_M^{-1} \bar{\mathbf{f}}, K_{xx} - k_{\mathbf{x}}^T K_M^{-1} k_{\mathbf{x}} + \sigma^2)$$ ie SPGP (4)

$$p(y|x,\bar{X},\bar{f}) = \mathcal{N}(k_x^T K_M^{-1} \bar{f},\ K_{xx} - k_x^T K_M^{-1} k_x + \sigma^2) \qquad (4)$$

The idea now is to express the joint probability of all observed values $y_n$ for all training data $n = 1, \ldots, N$, conditioned on their resp input values $x_n$ and the pseudo data points $\bar{X}, \bar{f}$. If we can maximize the joint probability of all training data observations wrt the pseudo points, then we know that the pseudo points selected are the optimal for this particular training data $X, f$ (and n.o. pseudo data points $M$).

If example data is i.i.d: (where we now denote pseudo inputs $\bar{x}_m$, $m = 1, \ldots, M$. Training data inputs $x_n$, $n = 1, \ldots, N$):

$$p(y|X,\bar{X},\bar{f}) = \prod_{n=1}^{N} p(y_n|x_n,\bar{X},\bar{f}) = \{\text{expressing as multivariate Gaussian}\} = p(\mathbf{y}|\mu_y, \Sigma_y) \text{ where}$$

$$\mu_y = \begin{bmatrix} k_{x_1}^T K_M^{-1} \bar{f} \\ k_{x_2}^T K_M^{-1} \bar{f} \\ \cdots \\ k_{x_N}^T K_M^{-1} \bar{f} \end{bmatrix} = \overbrace{\begin{bmatrix} - k_{x_1}^T - \\ - k_{x_2}^T - \\ \cdots \\ - k_{x_N}^T - \end{bmatrix}}^{\in \mathbb{R}^{N \times M}} \overbrace{K_M^{-1} \bar{f}}^{\in \mathbb{R}^{M \times 1}} = \underbrace{\begin{bmatrix} k(x_1,\bar{x}_1) & k(x_1,\bar{x}_2) & \cdots & k(x_1,\bar{x}_M) \\ k(x_2,\bar{x}_1) & k(x_2,\bar{x}_2) & \cdots & k(x_2,\bar{x}_M) \\ & \cdots & & \\ k(x_N,\bar{x}_1) & k(x_N,\bar{x}_2) & \cdots & k(x_N,\bar{x}_M) \end{bmatrix}}_{K_{NM}} K_M^{-1} \bar{f} = K_{NM} K_M^{-1} \bar{f}$$

$\Sigma_y$ is a diagonal $N \times N$ matrix $= \Lambda + \sigma^2 I$ where $\Lambda_{nn} = K_{x_n x_n} - k_{x_n}^T K_M^{-1} k_{x_n}$ $\Rightarrow$

$$p(y|X,\bar{X},\bar{f}) = \mathcal{N}(y|K_{NM} K_M^{-1} \bar{f},\ \Lambda + \sigma^2 I) \quad \text{ie SPGP (5)}$$

## SPGP (5)-(7)

If we put a prior on $\bar{f}$ it turns out (which will be shown) that we can integrate out $\bar{f}$, thereby reducing the n.o. parameters for which to optimize the joint probability.

$$p(\bar{f}\,|\,\bar{X}) = \mathcal{N}(\bar{f}\,|\,0, K_M) \qquad (6)$$

In fact, this was already used in the derivation of (4). For motivation, see corresponding text.

Summarizing, we have:

Bishop notation

$$p(\bar{f}\,|\,\bar{X}) = \mathcal{N}(\bar{f}\,|\,\overset{\mu}{0}, \overset{\Lambda^{-1}}{K_M})$$

$$p(y\,|\,\bar{f}, X, \bar{X}) = \mathcal{N}(y\,|\,\overbrace{K_{NM}K_M^{-1}}^{A}\bar{f} \overset{b=0}{}, \overbrace{\Lambda + \sigma^2 I}^{L^{-1}})$$

We can use Bayes' rule to find $p(\bar{f}\,|\,y, X, \bar{X}) = p(\bar{f}\,|\,\mathcal{D}, \bar{X})$. Note $p(\bar{f}\,|\,y, X, \bar{X}) = \dfrac{p(y\,|\,\bar{f}, X, \bar{X})\, p(\bar{f}\,|\,\bar{X}, X)}{p(y\,|\,\bar{X}, X)}$

$$= \frac{p(y\,|\,\bar{f}, X, \bar{X})\, p(\bar{f}\,|\,\bar{X})}{Z}$$ where $Z$ is a constant normalization factor (does not depend on $\bar{f}$).

Using Bishop 2.116 p93:

Bishop notation

Note $K_M$ is symmetric $\Rightarrow K_M^{-1}$ is symmetric.

posterior covariance

$$\Sigma = \Big(\underbrace{\overset{\Lambda}{K_M^{-1}}}_{\mathbb{R}^{M\times M}} + \underbrace{\overset{A^T}{(K_{NM}K_M^{-1})^T}}_{\mathbb{R}^{M\times N}}\ \underbrace{\overset{L}{(\Lambda + \sigma^2 I)^{-1}}}_{\mathbb{R}^{N\times N}}\ \underbrace{\overset{A}{(K_{NM}K_M^{-1})}}_{\mathbb{R}^{N\times M}}\Big)^{-1} = \left(K_M^{-1} + K_M^{-1}K_{MN}(\Lambda + \sigma^2 I)^{-1}K_{NM}K_M^{-1}\right)^{-1}$$

Consider $P = K_M(K_M + K_{MN}(\Lambda + \sigma^2 I)^{-1}K_{NM})^{-1}K_M$ (the covariance expression in SPGP (7))

Note
$$P^{-1} = K_M^{-1}(K_M + K_{MN}(\Lambda + \sigma^2 I)^{-1}K_{NM})K_M^{-1}$$
$$= K_M^{-1}K_M K_M^{-1} + K_M^{-1}K_{MN}(\Lambda + \sigma^2 I)^{-1}K_{NM}K_M^{-1}$$
$$= K_M^{-1} + K_M^{-1}K_{MN}(\Lambda + \sigma^2 I)^{-1}K_{NM}K_M^{-1} = \Sigma^{-1}$$

So $P$ and $\Sigma$ share the same inverse $\Rightarrow \underline{\Sigma = P}$ (since $P = PI = P(P^{-1}\Sigma) = I\Sigma = \Sigma$)

posterior mean

The mean is given by (Bishop notation): $\Sigma(A^T L(y - b)) = \Sigma K_M^{-1}K_{MN}(\Lambda + \sigma^2 I)^{-1}y$

Let $Q_M = K_M + K_{MN}(\Lambda + \sigma^2 I)^{-1}K_{NM} \Rightarrow \Sigma = K_M Q_M^{-1} K_M$

mean: $K_M Q_M^{-1} K_{MN}(\Lambda + \sigma^2 I)^{-1} y$

Thus $p(\bar{f}\,|\,y, X, \bar{X}) = p(\bar{f}\,|\,\mathcal{D}, \bar{X}) = \mathcal{N}(\bar{f}\,|\,K_M Q_M^{-1} K_{MN}(\Lambda + \sigma^2 I)^{-1} y,\ K_M Q_M^{-1} K_M)$ ie SPGP (7)

SPGP (4) + (7) => (8)

Finding the predictive distribution $y_*$ for a new point $\mathbf{x}_*$ given the training data $X, \mathbf{y}$ and pseudo inputs $\bar{X}$.

From (4) we have the predictive distribution over the target value for a new point given the pseudo inputs $\bar{X}$ and pseudo targets $\bar{\mathbf{f}}$ ($= \bar{\mathbf{y}}$ since no noise for pseudo points)

$$(4): \quad p(y_* | \mathbf{x}_*, \bar{X}, \bar{\mathbf{f}}) = \mathcal{N}(y_* | \overbrace{\mathbf{k}_{\mathbf{x}_*}^T K_M^{-1}}^{A} \bar{\mathbf{f}}, \overbrace{K_{\mathbf{x}_*\mathbf{x}_*} - \mathbf{k}_{\mathbf{x}_*}^T K_M^{-1} \mathbf{k}_{\mathbf{x}_*} + \sigma^2}^{L^{-1}}) \quad b = 0$$

← Bishop's notation 2.114

From (7) we have the distribution over the pseudo targets $\bar{\mathbf{f}}$ given the training data $\underbrace{X, \mathbf{y}}_{\mathcal{D}}$ and pseudo inputs $\bar{X}$

$$(7): \quad p(\bar{\mathbf{f}} | \mathcal{D}, \bar{X}) = \mathcal{N}(\bar{\mathbf{f}} | \overbrace{K_M Q_M^{-1} K_{MN} (\Lambda + \sigma^2 I)^{-1} \mathbf{y}}^{\mu}, \overbrace{K_M Q_M^{-1} K_M}^{\Lambda^{-1}})$$

← Bishop's notation 2.113

Note that
$$p(y_* | \mathbf{x}_*, \mathcal{D}, \bar{X}) = \int_{\bar{\mathbf{f}}} p(y_*, \bar{\mathbf{f}} | \mathbf{x}_*, \mathcal{D}, \bar{X}) \, d\bar{\mathbf{f}}$$
$$= \int_{\bar{\mathbf{f}}} p(y_* | \mathbf{x}_*, \bar{X}, \bar{\mathbf{f}}) \, p(\bar{\mathbf{f}} | \mathcal{D}, \bar{X}) \, d\bar{\mathbf{f}}$$

Note:
$\mathbf{k}_{\mathbf{x}_*}$ here is $\mathbf{k}_*$ in SPGP
$K_{\mathbf{x}_*\mathbf{x}_*}$ " $K_{**}$ "

{Using Bishop 2.115 p. 93,}

$$p(y_* | \mathbf{x}_*, \mathcal{D}, \bar{X}) = \mathcal{N}(y_* | \mu_*, \sigma_*^2)$$ where

Bishop's notation

$$\mu_* = A\mu + b = \mathbf{k}_{\mathbf{x}_*}^T K_M^{-1} K_M Q_M^{-1} K_{MN} (\Lambda + \sigma^2 I)^{-1} \mathbf{y} = \underline{\mathbf{k}_{\mathbf{x}_*}^T Q_M^{-1} K_{MN} (\Lambda + \sigma^2 I)^{-1} \mathbf{y}}$$

$$\sigma_*^2 = L^{-1} + A \Lambda^{-1} A^T = K_{\mathbf{x}_*\mathbf{x}_*} - \mathbf{k}_{\mathbf{x}_*}^T K_M^{-1} \mathbf{k}_{\mathbf{x}_*} + \sigma^2 + \mathbf{k}_{\mathbf{x}_*}^T K_M^{-1} K_M Q_M^{-1} K_M K_M^{-1} \mathbf{k}_{\mathbf{x}_*}$$

$$= K_{\mathbf{x}_*\mathbf{x}_*} - \mathbf{k}_{\mathbf{x}_*}^T K_M^{-1} \mathbf{k}_{\mathbf{x}_*} + \mathbf{k}_{\mathbf{x}_*}^T Q_M^{-1} \mathbf{k}_{\mathbf{x}_*} + \sigma^2$$

$$= \underline{K_{\mathbf{x}_*\mathbf{x}_*} - \mathbf{k}_{\mathbf{x}_*}^T (K_M^{-1} + Q_M^{-1}) \mathbf{k}_{\mathbf{x}_*} + \sigma^2}$$

ie SPGP (8)

The predictive distribution is conditioned on the pseudo input locations $\bar{X}$ (and hyper-parameters $\Theta$). To find them we use (9).

# SPGP (5) + (6) => (9)

We can find Maximum Likelihood estimates for $\Theta = \{\underbrace{c, \mathbf{b}, \sigma}_{\text{used in Kernel}}\}$ and the pseudo input locations $\bar{X}$.

The ML estimations are done using the marginal probability distribution of training data target values (observations) $\mathbf{y}$, given training data inputs $X$, pseudo inputs $\bar{X}$ and hyperparameters $\Theta$.

From (5): $p(\mathbf{y}|X,\bar{X},\bar{\mathbf{f}}) = \mathcal{N}(\mathbf{y}|\overbrace{K_{NM}K_M^{-1}}^{A}\bar{\mathbf{f}} \overset{b:0}{,} \overbrace{\Lambda + \sigma^2 I}^{L^{-1}})$ ← Bishop's notation 2.114

(6): $p(\bar{\mathbf{f}}|\bar{X}) = \mathcal{N}(\bar{\mathbf{f}}|\overset{\mu}{\mathbf{0}}, \overset{\Lambda^{-1}}{K_M})$ ← Bishop's notation 2.113

Recall that all probability distributions so far have been implicitly conditioned on $\Theta$

Then $p(\mathbf{y}|X,\bar{X},\Theta) = \int_{\bar{\mathbf{f}}} p(\mathbf{y},\bar{\mathbf{f}}|X,\bar{X},\Theta)\, d\bar{\mathbf{f}}$

$$= \int_{\mathbf{f}} p(\mathbf{y}|X,\bar{X},\bar{\mathbf{f}},\Theta)\, p(\bar{\mathbf{f}}|\bar{X},\Theta)\, d\bar{\mathbf{f}}$$

{Using Bishop 2.115 p 93}

$p(\mathbf{y}|X,\bar{X},\Theta) = \mathcal{N}(\mathbf{y}|\mathbf{m}, S)$ where

$$\mathbf{m} = \underbrace{A\mu + b}_{\text{Bishop's notation}} = K_{NM}K_M^{-1}\mathbf{0} + \mathbf{0} = \mathbf{0}$$

$$S = L^{-1} + A\Lambda^{-1}A^T = \Lambda + \sigma^2 I + K_{NM}K_M^{-1}K_M K_M^{-1}K_{MN} = \underline{K_{NM}K_M^{-1}K_{MN} + \Lambda + \sigma^2 I}$$

so $p(\mathbf{y}|X,\bar{X},\Theta) = \mathcal{N}(\mathbf{y}|\mathbf{0}, K_{NM}K_M^{-1}K_{MN} + \Lambda + \sigma^2 I)$ ie SPGP (9)