

# POSTERIOR PROCESS

- Let us now distinguish between observed training inputs, also denote by a design matrix  $\mathbf{X}$ , and the corresponding observed values

$$\mathbf{f} = \left[ f(\mathbf{x}^{(1)}), \dots, f(\mathbf{x}^{(n)}) \right]$$

and one single **unobserved test point**  $\mathbf{x}_*$  with  $f_* = f(\mathbf{x}_*)$ .

- We now want to infer the distribution of  $f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{f}$ .

$$f_* = f(\mathbf{x}_*)$$

- Assuming a zero-mean GP prior  $\mathcal{GP}(\mathbf{0}, k(\mathbf{x}, \mathbf{x}'))$  we know

$$\begin{bmatrix} \mathbf{f} \\ f_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K} & \mathbf{k}_* \\ \mathbf{k}_*^T & k_{**} \end{bmatrix}\right).$$

Here,  $\mathbf{K} = (k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}))_{i,j}$ ,  $\mathbf{k}_* = [k(\mathbf{x}_*, \mathbf{x}^{(1)}), \dots, k(\mathbf{x}_*, \mathbf{x}^{(n)})]$   
and  $k_{**} = k(\mathbf{x}_*, \mathbf{x}_*)$ .



## POSTERIOR PROCESS / 2

- Given that  $\mathbf{f}$  is observed, we can apply the general rule for condition (\*) of Gaussian random variables and obtain the following formula:

$$f_* \mid \mathbf{x}_*, \mathbf{X}, \mathbf{f} \sim \mathcal{N}(\mathbf{k}_*^T \mathbf{K}^{-1} \mathbf{f}, \mathbf{k}_{**} - \mathbf{k}_*^T \mathbf{K}^{-1} \mathbf{k}_*).$$

- As the posterior is a Gaussian, the maximum a-posteriori estimate, i.e. the mode of the posterior distribution, is  $\mathbf{k}_*^T \mathbf{K}^{-1} \mathbf{f}$ .



## POSTERIOR PROCESS / 3

(\*) General rule for condition of Gaussian random variables:

If the  $m$ -dimensional Gaussian vector  $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  can be partitioned with  $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2)$  where  $\mathbf{z}_1$  is  $m_1$ -dimensional and  $\mathbf{z}_2$  is  $m_2$ -dimensional,  $m_2$ -dimensional, and:

$$\begin{aligned} & (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2), \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \\ & (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2), \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \end{aligned}$$

then the conditioned distribution of  $\mathbf{z}_2 \mid \mathbf{z}_1 = \mathbf{a}$  is a multivariate normal

then the conditioned distribution of  $\mathbf{z}_2 \mid \mathbf{z}_1 = \mathbf{a}$  is a multivariate normal

$$\mathcal{N}(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} (\mathbf{a} - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12})$$

$$\mathcal{N}(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} (\mathbf{a} - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12})$$



## GP PREDICTION: TWO POINTS

Let us visualize this by a simple example:

- Assume we observed a single training point  $\mathbf{x} = -0.5$ , and want to make a prediction at a test point  $\mathbf{x}_* = 0.5$ .
- Under a zero-mean GP with  $k(\mathbf{x}, \mathbf{x}') = \exp(-\frac{1}{2}\|\mathbf{x} - \mathbf{x}'\|^2)$ , we compute the cov-matrix:

$$\begin{bmatrix} f \\ f_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 1 & 0.61 \\ 0.61 & 1 \end{bmatrix}\right).$$

- Assume that we observe the point  $f(\mathbf{x}) = 1$ .
- We compute the posterior distribution:

$$\begin{aligned} f_* \mid \mathbf{x}_*, \mathbf{x}, f &\sim \mathcal{N}(\mathbf{k}_*^T \mathbf{K}^{-1} f, k_{**} - \mathbf{k}_*^T \mathbf{K}^{-1} \mathbf{k}_*) \\ &\sim \mathcal{N}(0.61 \cdot 1 \cdot 1, 1 - 0.61 \cdot 1 \cdot 0.61) \\ &\sim \mathcal{N}(0.61, 0.6279) \end{aligned}$$

- The MAP-estimate for  $\mathbf{x}_*$  is  $f(\mathbf{x}_*) = 0.61$ , and the uncertainty estimate is 0.6279.



## POSTERIOR PROCESS

- We can generalize the formula for the posterior process for multiple unobserved test points:

$$\mathbf{f}_* = \left[ f(\mathbf{x}_*^{(1)}), \dots, f(\mathbf{x}_*^{(m)}) \right].$$

- Under a zero-mean Gaussian process, we have

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K} & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} \end{bmatrix}\right),$$

$$\text{with } \mathbf{K}_* = \left( k(\mathbf{x}^{(i)}, \mathbf{x}_*^{(j)}) \right)_{i,j}, \mathbf{K}_{**} = \left( k(\mathbf{x}_*^{(i)}, \mathbf{x}_*^{(j)}) \right)_{i,j}.$$



## POSTERIOR PROCESS / 2

- Similar to the single test point situation, to get the posterior distribution, we exploit the general rule of conditioning for Gaussians:

$$\mathbf{f}_* | \mathbf{X}_*, \mathbf{X}, \mathbf{f} \sim \mathcal{N}(\mathbf{K}_*^T \mathbf{K}^{-1} \mathbf{f}, \mathbf{K}_{**} - \mathbf{K}_*^T \mathbf{K}^{-1} \mathbf{K}_*).$$

- This formula enables us to talk about correlations among different test points and sample functions from the posterior process.

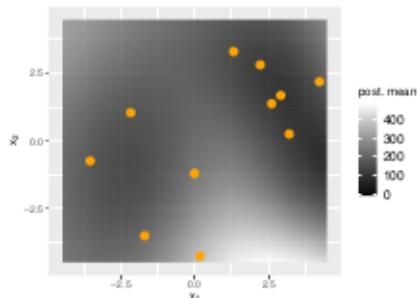


# GP AS A SPATIAL MODEL

- The correlation among two outputs depends on distance of the corresponding input points  $\mathbf{x}$  and  $\mathbf{x}'$  (e.g. Gaussian covariance kernel)

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\ell^2}\right)$$

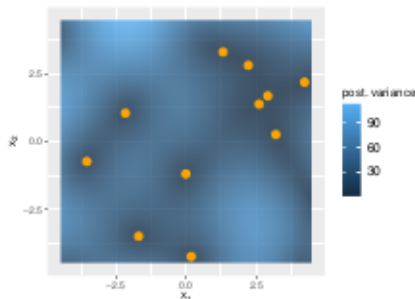
- Hence, close data points with high spatial similarity  $k(\mathbf{x}, \mathbf{x}')$  enter into more strongly correlated predictions:  $\mathbf{k}_*^\top \mathbf{K}^{-1} \mathbf{f}$  ( $\mathbf{k}_* := (k(\mathbf{x}, \mathbf{x}^{(1)}), \dots, k(\mathbf{x}, \mathbf{x}^{(n)}))$ ).



Example: Posterior mean of a GP that was fitted with the Gaussian covariance kernel with  $\ell = 1$ .

## GP AS A SPATIAL MODEL / 2

- Posterior uncertainty increases if the new data points are far from the design points.
- The uncertainty is minimal at the design points, since the posterior variance is zero at these points.



Example (continued): Posterior variance.





## NOISY GAUSSIAN PROCESS / 2

- In reality, however, this is often not the case.
- We often only have access to a noisy version of the true function value

$$y = f(\mathbf{x}) + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2).$$

- Let us still assume that  $f(\mathbf{x})$  is a Gaussian process.
- Then,

$$\begin{aligned} \text{Cov}(y^{(i)}, y^{(j)}) &= \text{Cov}(f(\mathbf{x}^{(i)}) + \epsilon^{(i)}, f(\mathbf{x}^{(j)}) + \epsilon^{(j)}) \\ &= \text{Cov}(f(\mathbf{x}^{(i)}), f(\mathbf{x}^{(j)})) + 2 \cdot \text{Cov}(f(\mathbf{x}^{(i)}), \epsilon^{(j)}) + \text{Cov}(\epsilon^{(i)}, \epsilon^{(j)}) \\ &= k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) + \sigma^2 \delta_j. \end{aligned}$$

- $\sigma^2$  is called **nugget**.



## NOISY GAUSSIAN PROCESS / 3

- Let us now derive the predictive distribution for the case of noisy observations.
- The prior distribution of  $y$ , assuming that  $f$  is modeled by a Gaussian process is then

$$\mathbf{y} = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{pmatrix} \sim \mathcal{N}(\mathbf{m}, \mathbf{K} + \sigma^2 \mathbf{I}_n),$$

with

$$\mathbf{m} := \left( m(\mathbf{x}^{(i)}) \right)_i, \quad \mathbf{K} := \left( k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \right)_{i,j}.$$



## NOISY GAUSSIAN PROCESS / 4

- We distinguish again between
  - observed training points  $\mathbf{X}$ ,  $\mathbf{y}$ , and
  - unobserved test inputs  $\mathbf{X}_*$  with unobserved values  $\mathbf{f}_*$

and get

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K} + \sigma^2 \mathbf{I}_n & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} \end{bmatrix}\right).$$



## NOISY GAUSSIAN PROCESS / 5

- Similarly to the noise-free case, we condition according to the rule of conditioning for Gaussians to get the posterior distribution for the test outputs  $f_*$  at  $\mathbf{X}_*$ :

$$f_* | \mathbf{X}_*, \mathbf{X}, \mathbf{y} \sim \mathcal{N}(\mathbf{m}_{\text{post}}, \mathbf{K}_{\text{post}}).$$

with

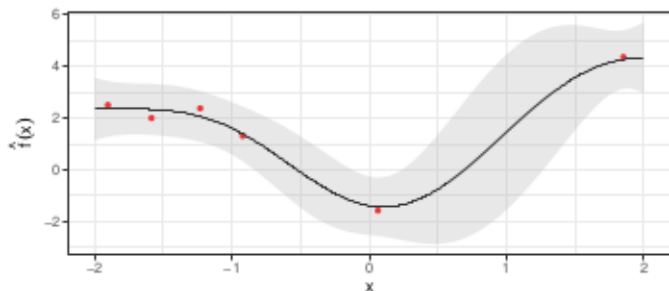
$$\begin{aligned} \mathbf{m}_{\text{post}} &= \mathbf{K}_*^T (\mathbf{K} + \sigma^2 \cdot \mathbf{I})^{-1} \mathbf{y} \\ \mathbf{K}_{\text{post}} &= \mathbf{K}_{**} - \mathbf{K}_*^T (\mathbf{K} + \sigma^2 \cdot \mathbf{I})^{-1} \mathbf{K}_*, \end{aligned}$$

- This converts back to the noise-free formula if  $\sigma^2 = 0$ .



## NOISY GAUSSIAN PROCESS / 6

- The noisy Gaussian process is not an interpolator any more.
- A larger nugget term leads to a wider “band” around the observed training points.
- The nugget term is estimated during training.



After observing the training points (red), we have a nugget-band around the observed points.  
( $k(x,x')$  is the squared exponential)



# RISK MINIMIZATION FOR GAUSSIAN PROCESSES

In machine learning, we learned about risk minimization. We usually choose a loss function and minimize the empirical risk

$$\mathcal{R}_{\text{emp}}(f) := \sum_{i=1}^n L(y^{(i)}, f(\mathbf{x}^{(i)}))$$

as an approximation to the theoretical risk

$$\mathcal{R}(f) := \mathbb{E}_{xy}[L(y, f(\mathbf{x}))] = \int L(y, f(\mathbf{x})) d\mathbb{P}_{xy}.$$

- How does the theory of Gaussian processes fit into this theory?
- What if we want to make a prediction which is optimal w.r.t. a certain loss function?



# RISK MINIMIZATION FOR GAUSSIAN PROCESSES

/ 2

- The theory of Gaussian process gives us a posterior distribution

$$p(y | \mathcal{D})$$

- If we now want to make a prediction at a test point  $\mathbf{x}_*$ , we approximate the theoretical risk in a different way, by using the posterior distribution:

$$\mathcal{R}(y_* | \mathbf{x}_*) \approx \int L(\tilde{y}_*, y_*) p(\tilde{y}_* | \mathbf{x}_*, \mathcal{D}) d\tilde{y}_*.$$

- The optimal prediction w.r.t the loss function is then:

$$\hat{y}_* | \mathbf{x}_* = \arg \min_{y_*} \mathcal{R}(y_* | \mathbf{x}_*).$$

