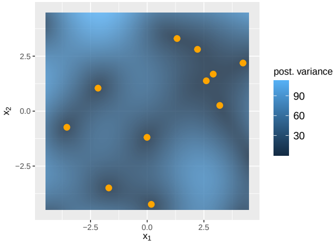
Introduction to Machine Learning

Gaussian Processes
Gaussian Posterior Process and Prediction



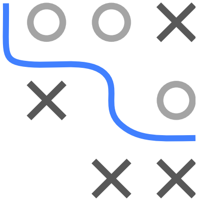
Learning goals

- Know how to derive the posterior process
- GPs are interpolating and spatial models
- Model noise via a nugget term

GP PREDICTION

- More interesting than drawing random samples from GP priors: predict at unseen test point $\mathbf{x}_*$ with $f_* = f(\mathbf{x}_*)$
- Given: training data with design matrix $\mathbf{X}$, observed values $\mathbf{f} = f(\mathbf{X}) = [f(\mathbf{x}^{(1)}), \dots, f(\mathbf{x}^{(n)})]^T$
- Goal: infer distribution of $f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{f}$
  $\Rightarrow$ update prior to posterior process

POSTERIOR PROCESS



- Again assuming $f \sim \mathcal{GP}(\mathbf{0}, k(\cdot,\cdot))$, we get

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

  with $\mathbf{k}_* = [k(\mathbf{x}_*, \mathbf{x}^{(1)}), \dots, k(\mathbf{x}_*, \mathbf{x}^{(n)})]^T$, $\mathbf{k}_{**} = k(\mathbf{x}_*, \mathbf{x}_*)$
- General rule for conditioning of Gaussian RVs
  - $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, partition $\boldsymbol{z} = (\boldsymbol{z}_1, \boldsymbol{z}_2)$ s.t. $\boldsymbol{z}_1 \in \mathbb{R}^{m_1}$, $\boldsymbol{z}_2 \in \mathbb{R}^{m_2}$,

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

  - Conditional distribution $\boldsymbol{z}_2 \mid \boldsymbol{z}_1 = \boldsymbol{a}$ is also Gaussian

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

- Apply to posterior process given $\mathbf{f}$ observed

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

- **Maximum-a-posteriori (MAP)** estimate: $\mathbf{k}_*^T\mathbf{K}^{-1}\mathbf{f}$

EXAMPLE: 2 POINTS

- Single training point $(\mathbf{x}, f(\mathbf{x})) = (-0.5, 1)$, test point $\mathbf{x}_* = 0.5$
- 0-mean GP with $k(\mathbf{x}, \tilde{\mathbf{x}}) = \exp(-\frac{1}{2}\|\mathbf{x} - \tilde{\mathbf{x}}\|^2)$ leads to

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

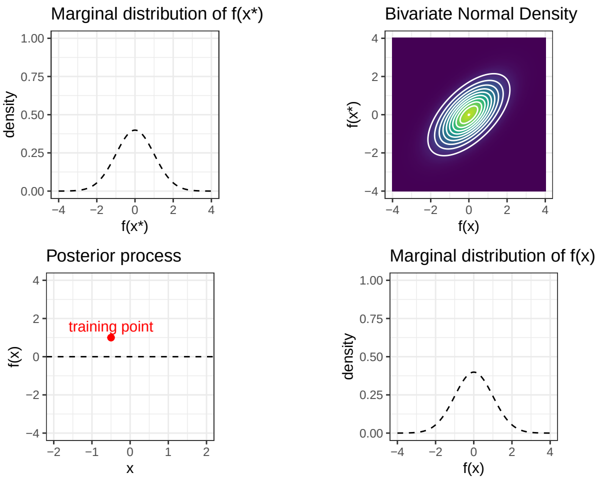
- Assuming we observe $f(\mathbf{x}) = 1$, compute posterior distribution

$$\begin{aligned} f_* \mid \mathbf{x}_*, \mathbf{x}, \mathbf{f} \quad &\sim \quad \mathcal{N}(\mathbf{k}_*^T \mathbf{K}^{-1} \mathbf{f}, \mathbf{k}_{**} - \mathbf{k}_*^T \mathbf{K}^{-1} \mathbf{k}_*) \\ &\sim \quad \mathcal{N}(0.61 \cdot 1 \cdot 1, 1 - 0.61 \cdot 1 \cdot 0.61) \\ &\sim \quad \mathcal{N}(0.61, 0.63) \end{aligned}$$

- MAP-estimate: $f(\mathbf{x}_*) = 0.61$, uncertainty estimate: 0.63

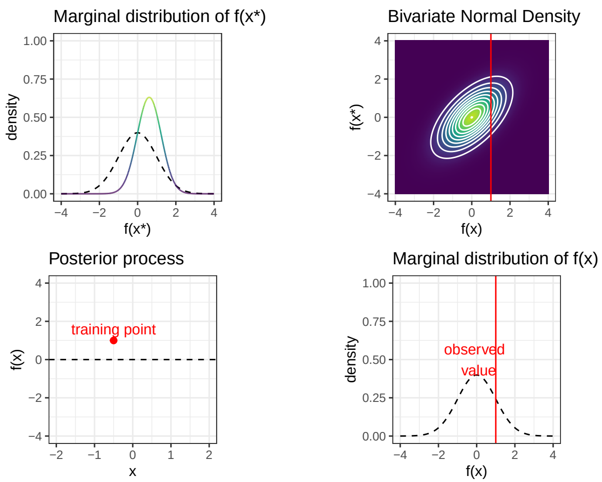
EXAMPLE: 2 POINTS

- Bivariate normal density + marginals for joint distribution of f, f_*



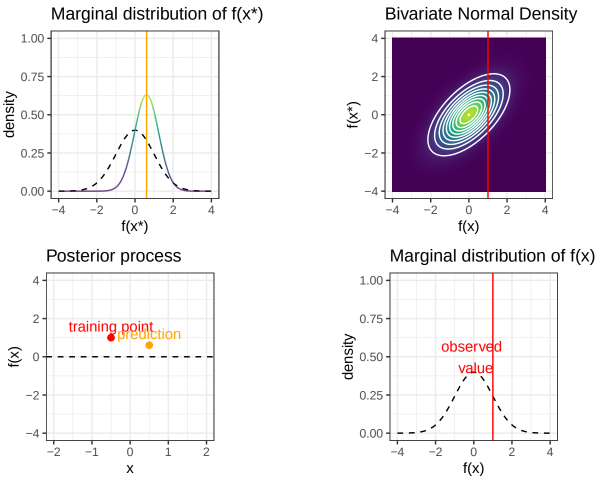
EXAMPLE: 2 POINTS

- Update posterior distribution, conditioning on observed value



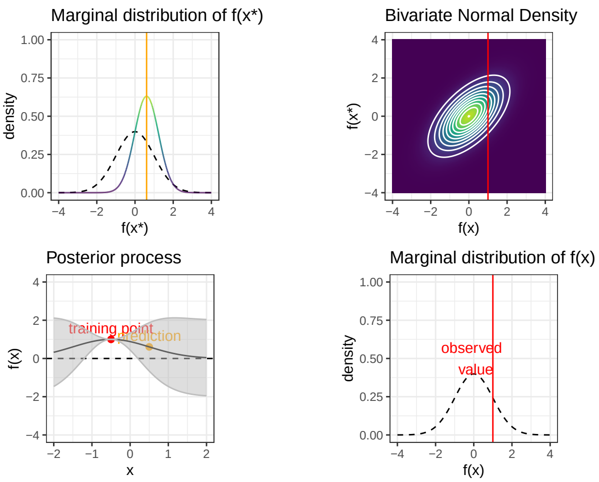
EXAMPLE: 2 POINTS

- Posterior mean: MAP estimate



EXAMPLE: 2 POINTS

- Posterior mean: MAP estimate
- Posterior uncertainty: ± 2 posterior SD (grey)



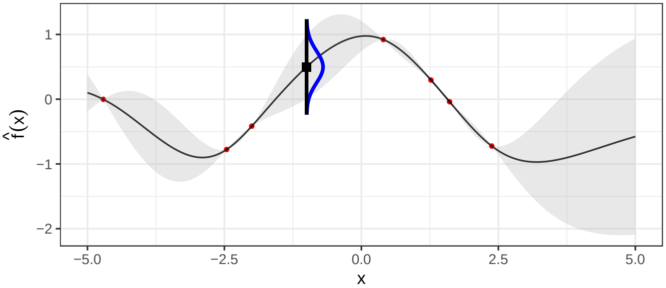
GP IS DISTRIBUTIONAL REGRESSION

- We have $\mathbf{f}$ observed

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

- Defines full **distributional regression** approach

- At each $\mathbf{x}_*$ get predictive distrib. instead of only point-wise preds



At each x, posterior defines a full distributional approach instead of only point−wise predictions.
(k(x,x') is Matèrn with nu = 2.5, the default for DiceKriging::km)

MULTIPLE TEST POINTS

- Now consider multiple test points

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

- Joint distribution (under zero-mean GP) becomes

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

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

- Again, employ rule of conditioning for Gaussians to get **posterior**

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

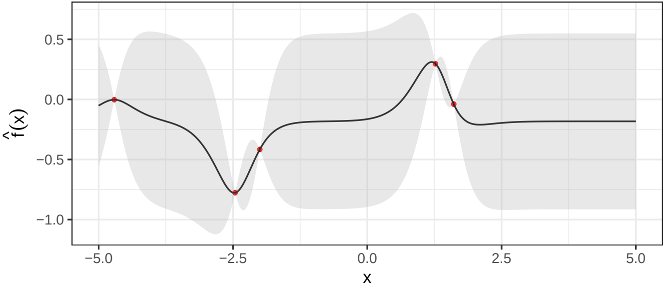
- Allows to compute correlations between test points + draw samples from posterior process

GP AS INTERPOLATOR

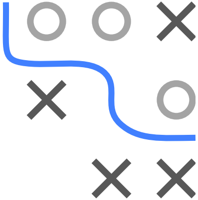
- MAP “prediction” for training point is exact function value

$$\begin{aligned}\mathbf{f} \mid \mathbf{X}, \mathbf{f} &\sim \mathcal{N}(\mathbf{K}\mathbf{K}^{-1}\mathbf{f}, \mathbf{K} - \mathbf{K}^T\mathbf{K}^{-1}\mathbf{K}) \\ &\sim \mathcal{N}(\mathbf{f}, \mathbf{0})\end{aligned}$$

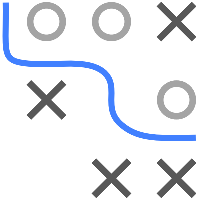
- Implication: GP is function **interpolator** (if unique points in $\mathbf{X}$)



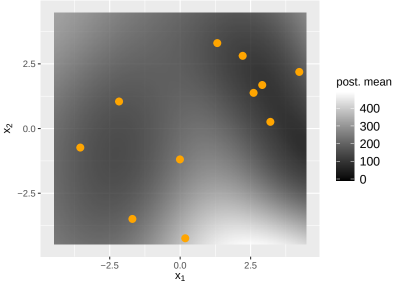
After observing the training points (red), the posterior process (black) interpolates the training points. (k(x,x') is Matèrn with nu = 2.5, the default for DiceKriging::km)

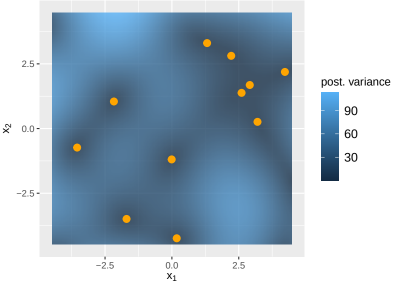


GP AS SPATIAL MODEL



- Spatial property: output correlation depends on input distance
- E.g., squared exponential kernel $k(\mathbf{x}, \tilde{\mathbf{x}}) = \exp\left(-\frac{\|\mathbf{x}-\tilde{\mathbf{x}}\|^2}{2\ell^2}\right)$
- Strongly correlated predictions for points with spatial proximity
- High posterior uncertainty for far-away points (0 at training locs)





NOISY GP

- GP as interpolator: implicitly assumed access to true function value $f(\mathbf{x}) \Rightarrow 0$ uncertainty at training points
- Reality: noisy version $y = f(\mathbf{x}) + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2)$
- Covariance becomes

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

- δ_{ij} = Kronecker delta
- σ^2 often called **nugget** $\Rightarrow$ estimate during training

PREDICTIVE DISTRIBUTION FOR NOISY GP

- Let $f \sim \mathcal{GP}(\mathbf{0}, k(\cdot, \cdot))$, $\mathbf{X} = \left\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\right\}$
- Prior predictive distribution for $\mathbf{y}$

$$\mathbf{y} = \left(y^{(1)}, \dots, y^{(n)}\right)^T \sim \mathcal{N}\left(\mathbf{0}, \mathbf{K} + \sigma^2 \boldsymbol{I}_n\right),$$

  with $\mathbf{m} = \mathbf{0}, \mathbf{K} = k(\mathbf{X}, \mathbf{X})$

- Consider joint distribution of training $(\mathbf{X}, \mathbf{y})$ and test points $(\mathbf{X}_*, \mathbf{f}_*)$

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

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

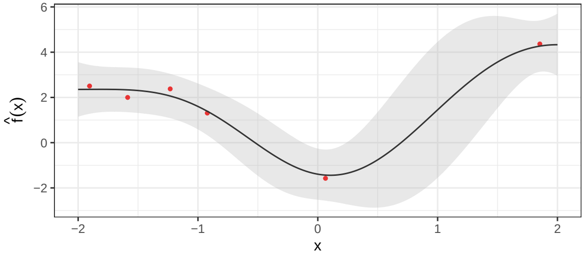
- **NB**: Since we work with $\mathbf{f}_*$ and not $\mathbf{y}_*$ there is no σ in $\mathbf{K}_{**}$

PREDICTIVE DISTRIBUTION FOR NOISY GP

- Again, employ rule of conditioning for Gaussians

$$\mathbf{f}_* \mid \mathbf{X}_*, \mathbf{X}, \mathbf{y} \sim \mathcal{N}(\mathbf{m}_{\text{post}}, \mathbf{K}_{\text{post}})$$

  with $\mathbf{m}_{\text{post}} = \mathbf{K}_*^T(\mathbf{K} + \sigma^2 \cdot \boldsymbol{I})^{-1}\mathbf{y}$, $\mathbf{K}_{\text{post}} = \mathbf{K}_{**} - \mathbf{K}_*^T(\mathbf{K} + \sigma^2 \cdot \boldsymbol{I})^{-1}\mathbf{K}_*$
- Recovers noise-free case for $\sigma^2 = 0$
- Noisy GP: no longer an interpolator
- Posterior uncertainty increases with nugget (wider “band”)



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

RISK MINIMIZATION FOR GP

- Recall: theoretical risk for unseen obs based on loss function L

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

- No access to $\mathbb{P}_{xy} \Rightarrow$ compute empirical risk over training data

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

- For GPs, make use of posterior predictive distribution over y

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

- Intuition: expected loss weighted by posterior probability of each $\tilde{y}_*$ given observed data
- Optimal prediction wrt loss function

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