

Gaussian Process

- 1. 贝叶斯方法 (参数化模型) → 高斯过程 GP (非参数化模型)
- 2. **Noise-free GP regression**: 直接使用“高斯分布 + 核函数”描述预测值 f 的分布.
- 3. **Noisy GP regression**: 预测值 f “不是观测数据” → 含隐变量 f 的高斯过程 → 需要推断 f 的后验分布 $p(f|t)$.
- 4. **GP classification**: 同样是含隐变量 f 的高斯过程 → 需要推断 f 的后验分布 $p(f|t)$.



问题: 在分类任务中, “似然”与“先验”不共轭, 无法直接得到后验分布 $p(f|t)$ 的解析式.

- 4.4 近似方法 ①: 基于 Newton–Raphson 方法获得 $p(f|t)$ 的 MAP 点估计 $\hat{f}$.
- 4.5 近似方法 ②: 基于 Laplace approximation 方法获得 $p(f|t)$ 的近似分布 $\mathcal{N}(f; \mu_f, \Sigma_f)$.

1. 贝叶斯线性回归 (参数化模型) → 高斯过程 (非参数化模型)

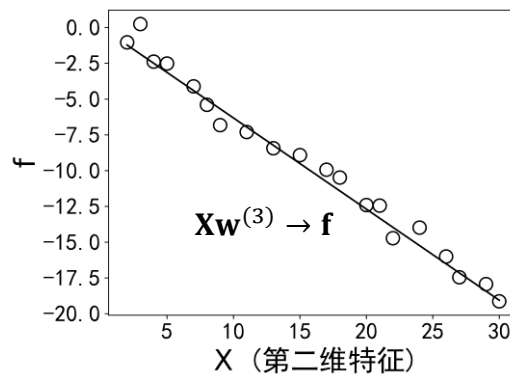
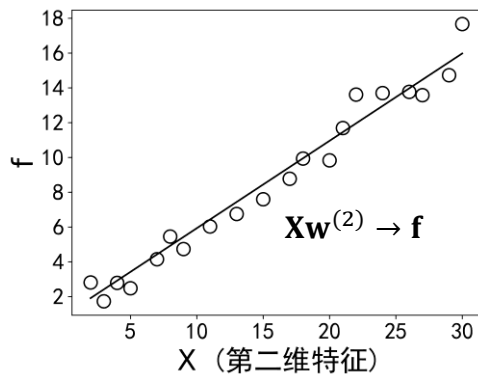
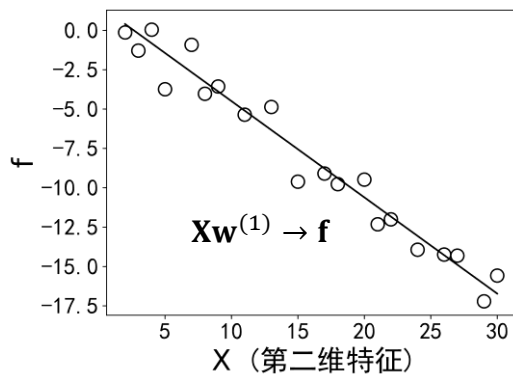
- 考虑一维贝叶斯线性回归例子: $f_i = \mathbf{w}^\top \mathbf{x}_i$

- 先验: $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \mathbf{0}, \mathbf{I})$. 似然: $p(\mathbf{f}|\mathbf{X}, \mathbf{w}; \sigma^2) = \mathcal{N}(\mathbf{f}; \mathbf{X}\mathbf{w}, \sigma^2\mathbf{I})$

↓ 采样 $\mathbf{w}^{(i)}$

- 预测方法①: 采样模型 $\mathbf{w}_i \rightarrow \mathbf{f}$ 的预测分布, 即 $p(\mathbf{f}|\mathbf{w}_i, \mathbf{x}; \sigma^2) = \mathcal{N}(\mathbf{f}; \mathbf{X}\mathbf{w}^{(i)}, \sigma^2\mathbf{I})$.

↓ 采样 $\mathbf{f}$

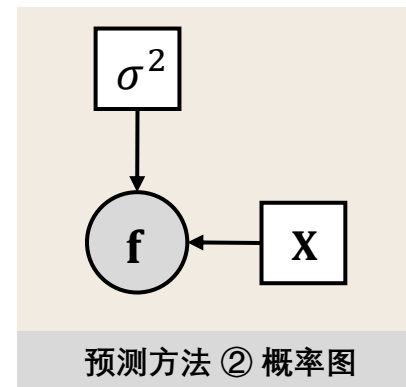
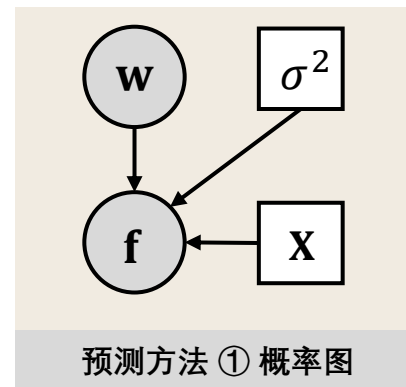


$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} \quad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \quad \mathbf{f} = \begin{bmatrix} f_1 \\ \vdots \\ f_N \end{bmatrix}$$

输入特征

模型

预测标签



- 预测方法②: 直接得到 $\mathbf{f}$ 的预测分布, 即 $p(\mathbf{f}|\mathbf{X}; \sigma^2) = \int p(\mathbf{f}|\mathbf{X}, \mathbf{w}; \sigma^2) p(\mathbf{w}) d\mathbf{w} = \mathcal{N}(\mathbf{f}; \mathbf{0}, \sigma^2\mathbf{I} + \mathbf{X}\mathbf{X}^\top)$



1. 均值向量为0向量
2. 协方差矩阵为“非对角矩阵”
3. 预测分布与模型 $\mathbf{w}$ 无关



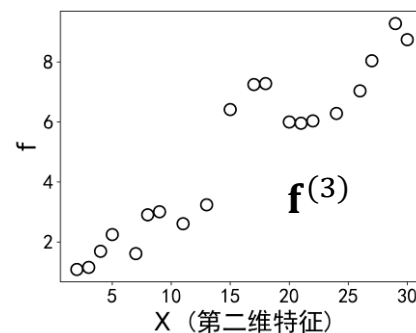
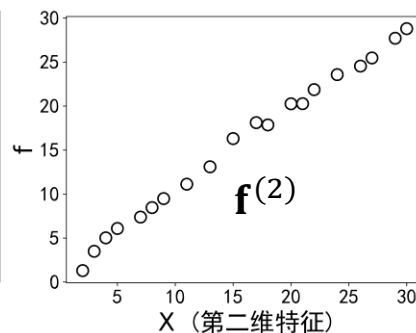
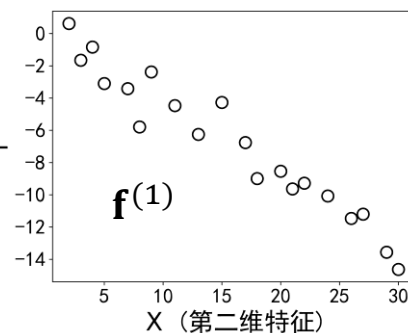
能否忽略模型参数 $\mathbf{w}$ 的表示,
直接描述“预测标签的分布”?



Gaussian Process (GP)

↓ 采样 $\mathbf{f}^{(1)}, \mathbf{f}^{(2)}, \mathbf{f}^{(3)}$

虽然均值为0,
但是协方差矩阵不是对角矩阵, 因此, $f_i^{(1)}$ 采样的结果会影响 $f_{i+1}^{(1)}$ 的结果。
例如: $f_i^{(1)} = 5$, 则 $f_{i+1}^{(1)} \approx 5$



2.1 Noise-free GP regression – 直接描述标签 f, f^* 的“联合分布” $p(f, f^*; \mathbf{x}, \mathbf{x}^*)$

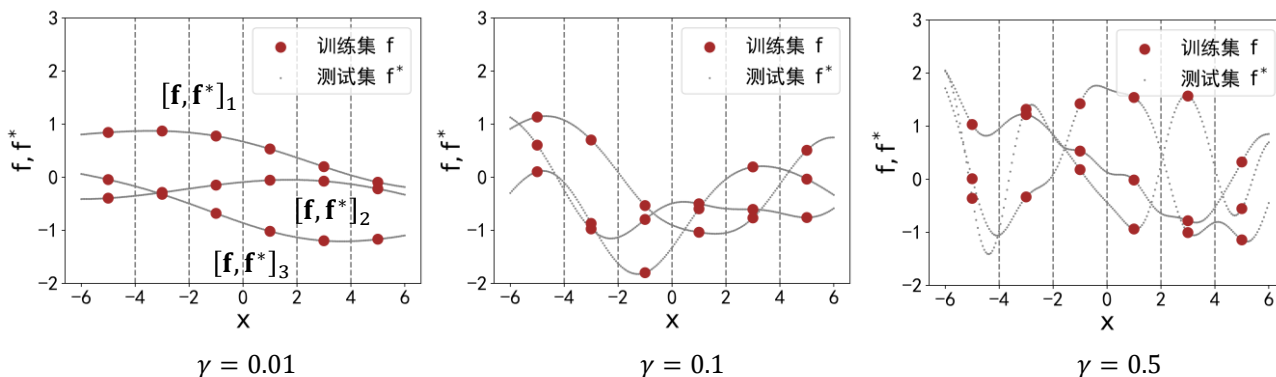
- Noise-free GP regression 训练集: $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$, $\mathbf{f} = \begin{bmatrix} f_1 \\ \vdots \\ f_N \end{bmatrix}$, 测试集: $\mathbf{x}^* = \begin{bmatrix} x_1^* \\ \vdots \\ x_L^* \end{bmatrix}$, $\mathbf{f}^* = \begin{bmatrix} f_1^* \\ \vdots \\ f_L^* \end{bmatrix}$.

- 定义训练、预测标签 f, f^* 的联合分布: $p(\mathbf{f}, \mathbf{f}^*; \mathbf{x}, \mathbf{x}^*) = \mathcal{N}(\mathbf{f}, \mathbf{f}^*; \mathbf{0}, \begin{bmatrix} \mathbf{C} & \mathbf{R} \\ \mathbf{R}^\top & \mathbf{C}^* \end{bmatrix})$.

其中, $\mathbf{C} = \begin{bmatrix} c(x_1, x_1) & \cdots & c(x_1, x_N) \\ \vdots & \ddots & \vdots \\ c(x_N, x_1) & \cdots & c(x_N, x_N) \end{bmatrix}$, $\mathbf{C}^* = \begin{bmatrix} c(x_1^*, x_1^*) & \cdots & c(x_1^*, x_L^*) \\ \vdots & \ddots & \vdots \\ c(x_L^*, x_1^*) & \cdots & c(x_L^*, x_L^*) \end{bmatrix}$, $\mathbf{R} = \begin{bmatrix} c(x_1, x_1^*) & \cdots & c(x_1, x_L^*) \\ \vdots & \ddots & \vdots \\ c(x_N, x_1^*) & \cdots & c(x_N, x_L^*) \end{bmatrix}$.

- 核函数RBF: $c(x_n, x_m) = \alpha \exp(-\gamma(x_n - x_m)^2)$; 衡量 x_n 和 x_m 的距离, 距离越近, 协方差越大, 说明“关联性越强”; 反之, “关联性越弱”.

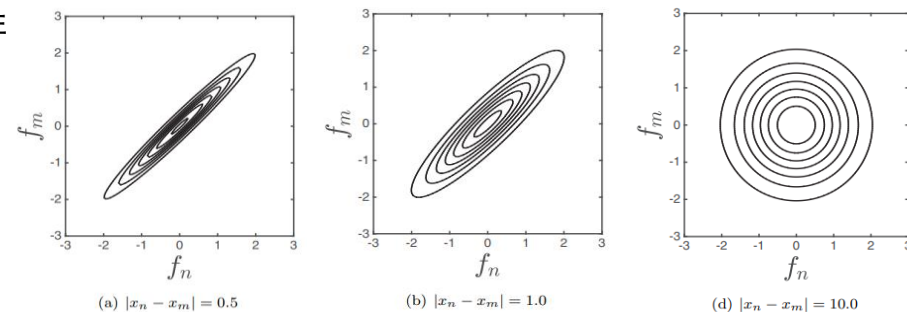
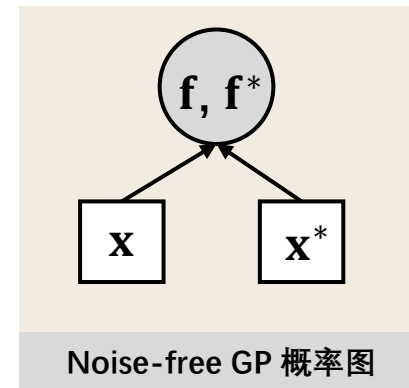
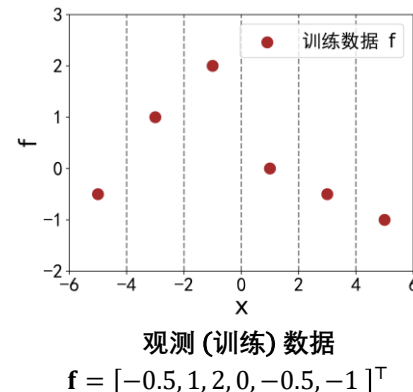
- 从联合分布 $p(\mathbf{f}, \mathbf{f}^*; \mathbf{x}, \mathbf{x}^*)$ 中采样 $[\mathbf{f}, \mathbf{f}^*]_1, [\mathbf{f}, \mathbf{f}^*]_2, [\mathbf{f}, \mathbf{f}^*]_3, \dots$, 得到预测结果如下:



事实上, 在分布 $p(\mathbf{f}, \mathbf{f}^*; \mathbf{x}, \mathbf{x}^*)$ 中, $\mathbf{f}$ 的值是已知的, 即“观测变量”.

如该例中, $\mathbf{f} = [-0.5, 1, 2, 0, -0.5, -1]^\top$

- 条件预测分布 $p(\mathbf{f}^* | \mathbf{f}; \mathbf{x}, \mathbf{x}^*)$: 能否由已知“观测变量 $\mathbf{f}$ ” 确定“条件预测分布” $p(\mathbf{f}^* | \mathbf{f}; \mathbf{x}, \mathbf{x}^*)$, 从而得到更准确的预测值?



随着 x_n, x_m 距离增大, f_n, f_m 组成二维高斯分布的协方差矩阵不断趋于对角矩阵

Comment 8.2 – Gaussian Conditionals: Consider a random vector $\mathbf{z}$ consisting of variables that can be split into two sets, $\mathbf{z}_1$ and $\mathbf{z}_2$:

$$\mathbf{z} = \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{bmatrix} \sim \begin{bmatrix} \mathbf{f} \\ \mathbf{f}^* \end{bmatrix}$$

Assume that $\mathbf{z}$ follows a Gaussian distribution:

$$p(\mathbf{z}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

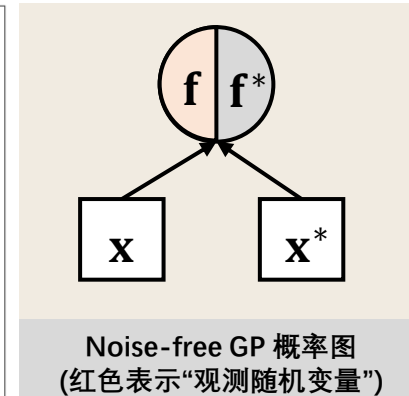
$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \sim \begin{bmatrix} \mathbf{C} & \mathbf{R} \\ \mathbf{R}^\top & \mathbf{C}^* \end{bmatrix}$$

(Note that $\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}^\top$ as covariance matrices must be symmetric.) If we observe the value of one of the sets of variables (say $\mathbf{z}_1$), the conditional density of the other set is also Gaussian:

$$p(\mathbf{z}_2 | \mathbf{z}_1) = \mathcal{N}(\mathbf{a}, \mathbf{B})$$

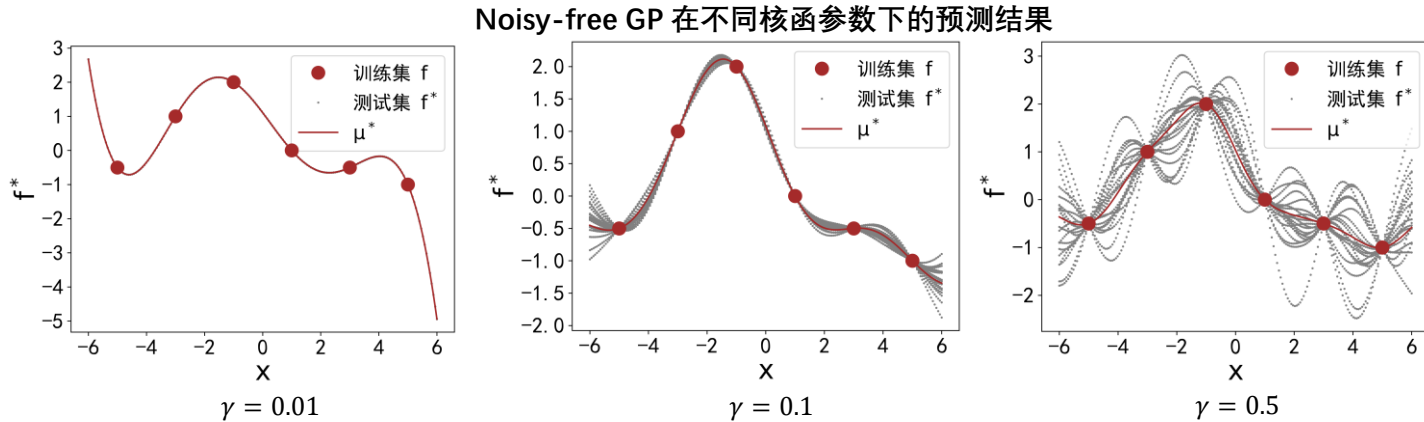
$$\mathbf{a} = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} (\mathbf{z}_1 - \boldsymbol{\mu}_1), \quad \mathbf{B} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12}$$

$\begin{matrix} 0 & \mathbf{R}^\top & \mathbf{C} & \mathbf{f} & 0 \end{matrix}$



2.2 Noise-free GP regression – 由观测变量 $\mathbf{f}$ 确定“条件分布” $p(\mathbf{f}|\mathbf{f}^*; \mathbf{x}, \mathbf{x}^*)$

- **Noise-free GP 条件预测分布:** $p(\mathbf{f}^*|\mathbf{f}, \mathbf{x}, \mathbf{x}^*) = \mathcal{N}(\mathbf{f}^*; \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$. $\boldsymbol{\mu}^* = \mathbf{R}^\top \mathbf{C}^{-1} \mathbf{f}$, $\boldsymbol{\Sigma}^* = \mathbf{C}^* - \mathbf{R}^\top \mathbf{C}^{-1} \mathbf{R}$ 代入 $\mathbf{f} = [-0.5, 1, 2, 0, -0.5, -1]^\top$



Comment 8.2 – Gaussian Conditionals: Consider a random vector $\mathbf{z}$ consisting of variables that can be split into two sets, $\mathbf{z}_1$ and $\mathbf{z}_2$:

$$\mathbf{z} = \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{bmatrix}, \quad \begin{bmatrix} \mathbf{f} \\ \mathbf{f}^* \end{bmatrix}$$

Assume that $\mathbf{z}$ follows a Gaussian distribution:

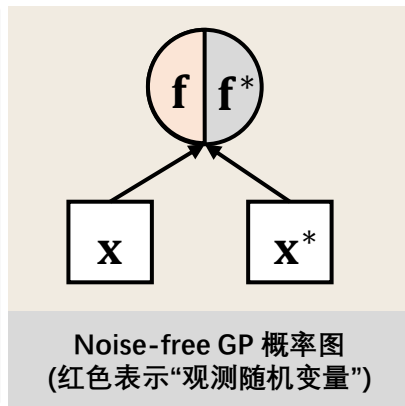
$$p(\mathbf{z}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{C} & \mathbf{R} \\ \mathbf{R}^\top & \mathbf{C}^* \end{bmatrix}$$

(Note that $\Sigma_{21} = \Sigma_{12}^T$ as covariance matrices must be symmetric.) If we observe the value of one of the sets of variables (say $\mathbf{z}_1$), the conditional density of the other set is also Gaussian:

$$p(\mathbf{z}_2|\mathbf{z}_1) = \mathcal{N}(\mathbf{a}, \mathbf{B})$$

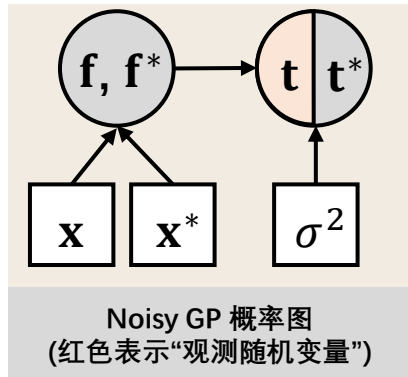
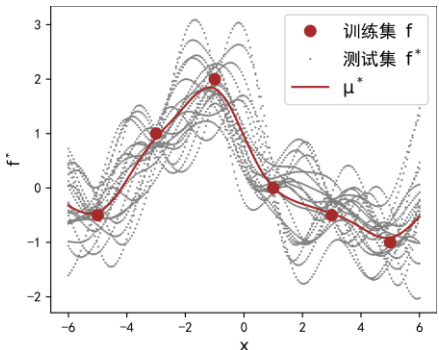
$$\mathbf{a} = \underbrace{\boldsymbol{\mu}_2}_{\mathbf{0}} + \underbrace{\boldsymbol{\Sigma}_{12}^\top}_{\mathbf{R}^\top} \underbrace{\boldsymbol{\Sigma}_{11}^{-1}}_{\mathbf{C}} (\underbrace{\mathbf{z}_1}_{\mathbf{f}} - \underbrace{\boldsymbol{\mu}_1}_{\mathbf{0}}), \quad \mathbf{B} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12}.$$



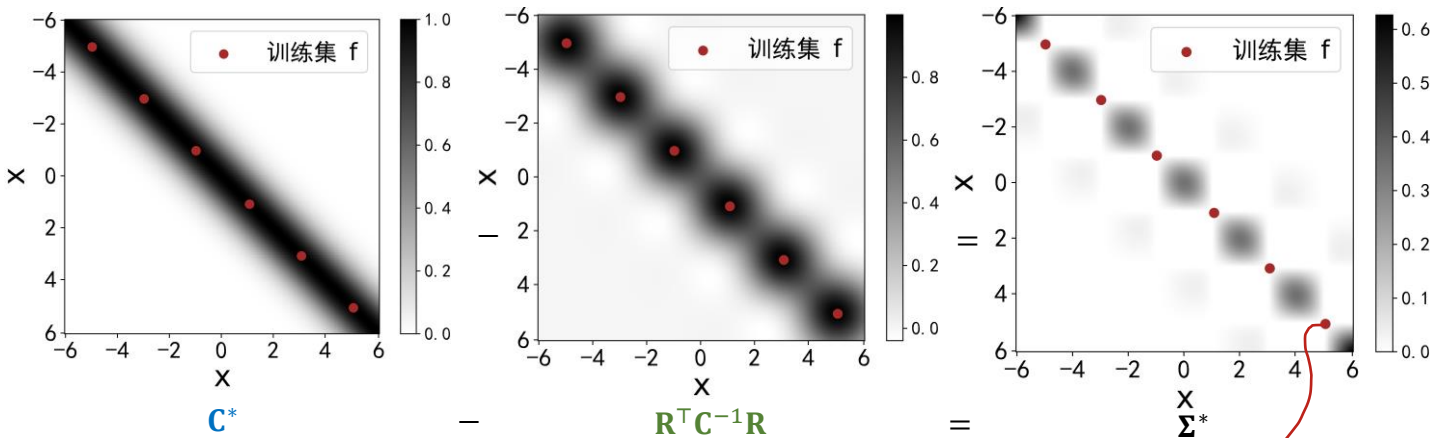
- ① 预测均值 $\mu^* = R^T C^{-1} f$: 将不同位置 f^* 的按照一定权重 (R^T) 靠向观测数据 f .
- ② 预测方差 $\Sigma^* = C^* - R^T C^{-1} R$: 降低“训练样本点周围”的方差.

能否对标签 f, f^* 添加噪声 σ^2 ,
使“观测数据 f ”周围的预测值也具备一定随机性

- **Noise GP regression:** 观测数据 $t = \text{模型输出 } f + \text{高斯噪声 } \sigma^2$



下图表示: 当核函数参数 $\alpha = 1, \gamma = 0.5$ 时, 获得条件预测分布 Σ^* 所涉及的矩阵 $(\mathbf{C}^*, \mathbf{R}^\top \mathbf{C}^{-1} \mathbf{R}, \Sigma^*)$



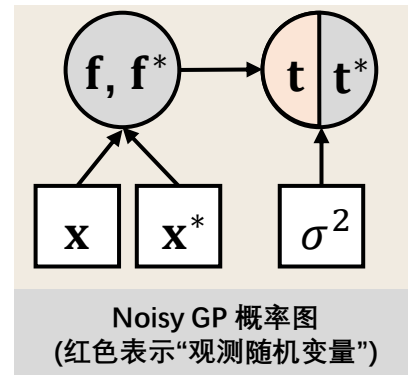
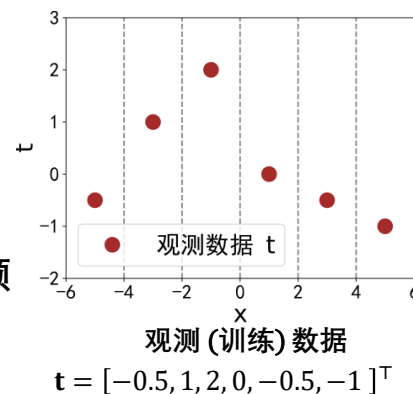
在 Noise-free GP 中, 观测数据 f 周围颜色较浅 (方差较小), 因此, 在训练点周围的采样曲线都经过同一位置

$$\mathbf{C} = \begin{bmatrix} c(x_1, x_1) & \cdots & c(x_1, x_N) \\ \vdots & \ddots & \vdots \\ c(x_N, x_1) & \cdots & c(x_N, x_N) \end{bmatrix}, \mathbf{C}^* = \begin{bmatrix} c(x_1^*, x_1^*) & \cdots & c(x_1^*, x_L^*) \\ \vdots & \ddots & \vdots \\ c(x_L^*, x_1^*) & \cdots & c(x_L^*, x_L^*) \end{bmatrix}, \mathbf{R} = \begin{bmatrix} c(x_1, x_1^*) & \cdots & c(x_1, x_L^*) \\ \vdots & \ddots & \vdots \\ c(x_N, x_1^*) & \cdots & c(x_N, x_L^*) \end{bmatrix}$$

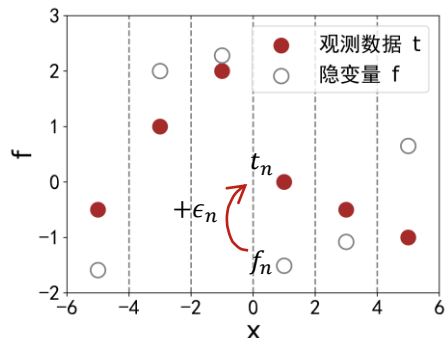
3.1 Noisy GP regression – 观测数据 $\mathbf{t}$ = 模型输出 $\mathbf{f}$ + 高斯噪声 σ^2 (含隐变量的 GP)

- 训练集: $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$, $\mathbf{f} = \begin{bmatrix} f_1 \\ \dots \\ f_N \end{bmatrix}$, $\mathbf{t} = \begin{bmatrix} t_1 \\ \dots \\ t_N \end{bmatrix}$, 测试集: $\mathbf{x}^* = \begin{bmatrix} x_1^* \\ \vdots \\ x_L^* \end{bmatrix}$, $\mathbf{f}^* = \begin{bmatrix} f_1^* \\ \dots \\ f_L^* \end{bmatrix}$, $\mathbf{t}^* = \begin{bmatrix} t_1^* \\ \dots \\ t_L^* \end{bmatrix}$

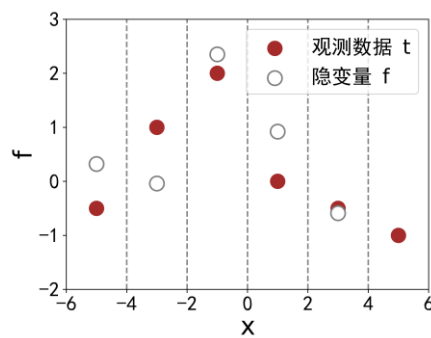
- Noisy GP 预测过程:** $t_n = f_n + \epsilon_n$, 其中, $\epsilon_n \sim \mathcal{N}(\epsilon_n; 0, \sigma^2)$. 即假定: “观测值 t_n ” 为 “模型输出 $\mathbf{f}$ ” + “额外高斯噪声 σ^2 ”.



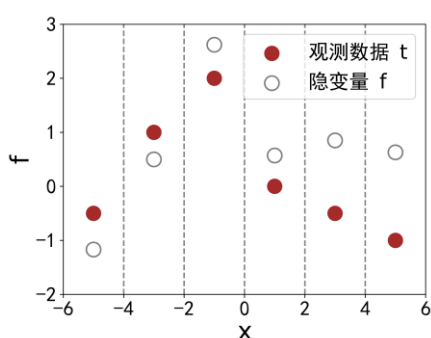
- 存在问题:** 此时无法直接将 $\mathbf{f}$ 代入 GP 条件预测分布 $p(\mathbf{f}^* | \mathbf{f}, \mathbf{x}, \mathbf{x}^*) = \mathcal{N}(\mathbf{f}^*; \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$. $\boldsymbol{\mu}^* = \mathbf{R}^\top \mathbf{C}^{-1} \mathbf{f}$, $\boldsymbol{\Sigma}^* = \mathbf{C}^* - \mathbf{R}^\top \mathbf{C}^{-1} \mathbf{R}$. $\leftarrow \mathbf{f} = ??$
 代入
- 原因: 所需的模型输出 $\mathbf{f}$ 是“未知”的, 此时能够利用的信息仅为观测标签数据 $\mathbf{t} = [-0.5, 1, 2, 0, -0.5, -1]^\top$.
- 然而, 生成上述“标签数据 $\mathbf{t}$ ”的“模型输出 $\mathbf{f}$ ”可能有多个取值, 例如: $\mathbf{f} = [-1.59, 2, 2.28, -1.51, -1.08, 0.65]^\top$, $\mathbf{f} = [0.32, -0.04, 2.35, 0.92, -0.59, -4.13]^\top \dots$



$$\mathbf{f} = [-1.59, 2, 2.28, -1.51, -1.08, 0.65]^\top$$



$$\mathbf{f} = [-1.59, 2, 2.28, -1.51, -1.08, 0.65]^\top$$



$$\mathbf{f} = [-1.17, 0.5, 2.62, 0.57, 0.85, 0.63]^\top$$

上图表示生成观测数据 $\mathbf{t} = [-0.5, 1, 2, 0, -0.5, -1]^\top$ 的隐变量 $\mathbf{f}$ 有很多种取值可能

在给定观测数据 $\mathbf{t}$ 时, 隐变量 $\mathbf{f}$ 可以有很多个“有意义的取值”

因此, 能否得到 $\mathbf{f}$ 的分布 $p(\mathbf{f} | \mathbf{t}, \sigma^2)$
 利用 $p(\mathbf{f} | \mathbf{t}; \mathbf{x}, \sigma^2)$ 获得 $\mathbf{f}^*$ 的 GP 条件预测分布?

- 解决方法:** 利用“观测数据 $\mathbf{t}$ ”得到后验分布 $p(\mathbf{f} | \mathbf{t})$, 从而边缘化条件预测分布 $p(\mathbf{f}^* | \mathbf{t}, \mathbf{x}, \mathbf{x}^*, \sigma^2) = \int p(\mathbf{f}^* | \mathbf{f}; \mathbf{x}^*) p(\mathbf{f} | \mathbf{t}; \mathbf{x}, \sigma^2) d\mathbf{f}$

- 如何获得隐变量 $\mathbf{f}$ 的后验分布 $p(\mathbf{f} | \mathbf{t}, \sigma^2)$? $\rightarrow p(\mathbf{f} | \mathbf{t}; \mathbf{x}, \sigma^2) = \frac{p(\mathbf{t} | \mathbf{f}, \sigma^2) p(\mathbf{f}; \mathbf{x})}{\int p(\mathbf{t} | \mathbf{f}, \sigma^2) p(\mathbf{f}; \mathbf{x}) d\mathbf{f}}$
 高斯分布 $\rightarrow$ 高斯分布

3.2 Noisy GP regression – 获得隐变量 $\mathbf{f}$ 的后验分布 $p(\mathbf{f}|\mathbf{t}; \mathbf{x}, \sigma^2)$

- 如何获得隐变量 $\mathbf{f}$ 的后验分布 $p(\mathbf{f}|\mathbf{t}; \mathbf{x}, \sigma^2)$? $\rightarrow p(\mathbf{f}|\mathbf{t}; \mathbf{x}, \sigma^2) = \frac{p(\mathbf{t}|\mathbf{f}, \sigma^2)p(\mathbf{f}; \mathbf{x})}{\int p(\mathbf{t}|\mathbf{f}, \sigma^2)p(\mathbf{f}; \mathbf{x})d\mathbf{f}}$

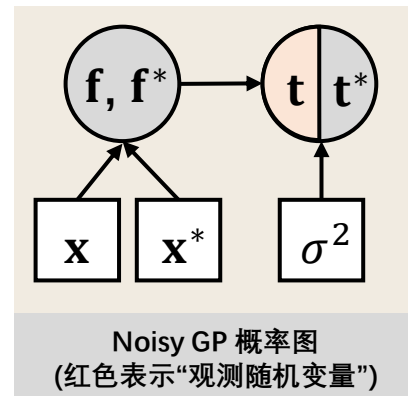
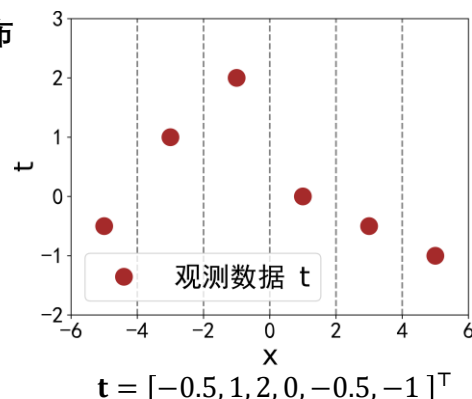
高斯分布 $\rightarrow$

$\leftarrow$ 高斯分布

- ① 似然: $p(\mathbf{t}|\mathbf{f}; \sigma^2) = \mathcal{N}(\mathbf{t}; \mathbf{f}, \sigma^2 \mathbf{I})$. 因为 $t_n = f_n + \epsilon_n, \epsilon_n \sim \mathcal{N}(0, \sigma^2)$.

- ② 先验: $p(\mathbf{f}; \mathbf{x}) = \mathcal{N}(\mathbf{f}; \mathbf{0}, \mathbf{C})$, $\mathbf{C} = \begin{bmatrix} c(x_1, x_1) & \cdots & c(x_1, x_N) \\ \vdots & \ddots & \vdots \\ c(x_N, x_1) & \cdots & c(x_N, x_N) \end{bmatrix}$.

- 由于“似然”与“先验”为共轭分布, 因此 $p(\mathbf{f}|\mathbf{t}, \sigma^2)$ 为高斯分布, 且 $p(\mathbf{f}|\mathbf{t}, \sigma^2) \propto p(\mathbf{t}|\mathbf{f}, \sigma^2)p(\mathbf{f})$.



- 隐变量 $\mathbf{f}$ 的后验分布:

$$p(\mathbf{f}|\mathbf{t}; \mathbf{x}, \sigma^2) = \mathcal{N}(\mathbf{f}; \boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f), \text{ 其中, } \boldsymbol{\mu}_f = \frac{1}{\sigma^2} \boldsymbol{\Sigma}_f \mathbf{t}, \boldsymbol{\Sigma}_f = \left(\mathbf{C}^{-1} + \frac{1}{\sigma^2} \mathbf{I} \right)^{-1}$$

$\leftarrow \mathbf{t} = [-0.5, 1, 2, 0, -0.5, -1]^T$
代入

- 预测: $p(\mathbf{f}^*|\mathbf{t}; \mathbf{x}, \mathbf{x}^*, \sigma^2) = \int p(\mathbf{f}^*|\mathbf{f}; \mathbf{x}^*)p(\mathbf{f}|\mathbf{t}; \mathbf{x}, \sigma^2)d\mathbf{f} = \int \mathcal{N}(\mathbf{f}^*; \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)\mathcal{N}(\mathbf{f}; \boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f)d\mathbf{f}$.

- 其中, $\boldsymbol{\mu}^* = \mathbf{R}^T \mathbf{C}^{-1} \mathbf{f}$, $\boldsymbol{\Sigma}^* = \mathbf{C}^* - \mathbf{R}^T \mathbf{C}^{-1} \mathbf{R}$, $\boldsymbol{\mu}_f = \frac{1}{\sigma^2} \boldsymbol{\Sigma}_f \mathbf{t}$, $\boldsymbol{\Sigma}_f = \left(\mathbf{C}^{-1} + \frac{1}{\sigma^2} \mathbf{I} \right)^{-1}$

结论 ①: $p(\mathbf{f}^*|\mathbf{f}; \mathbf{x}, \mathbf{x}^*) = \mathcal{N}(\mathbf{f}^*; \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$. $\boldsymbol{\mu}^* = \mathbf{R}^T \mathbf{C}^{-1} \mathbf{f}$, $\boldsymbol{\Sigma}^* = \mathbf{C}^* - \mathbf{R}^T \mathbf{C}^{-1} \mathbf{R}$

结论 ②: $\int \mathcal{N}(\mathbf{a}; \mathbf{B}\mathbf{c}, \mathbf{D})\mathcal{N}(\mathbf{c}; \mathbf{e}, \mathbf{F})d\mathbf{c} = \mathcal{N}(\mathbf{a}; \mathbf{B}\mathbf{e}, \mathbf{D} + \mathbf{B}\mathbf{F}\mathbf{B}^T)$

- Noisy GP regression 条件预测分布:

$$p(\mathbf{f}^*|\mathbf{t}; \mathbf{x}, \mathbf{x}^*, \sigma^2) = \mathcal{N}(\mathbf{f}^*; \underbrace{\mathbf{R}^T (\mathbf{C} + \sigma^2 \mathbf{I})^{-1} \mathbf{t}}_{\text{wavy}}, \underbrace{\mathbf{C}^* - \mathbf{R}^T \mathbf{C}^{-1} \mathbf{R} + \mathbf{R}^T \mathbf{C}^{-1} \boldsymbol{\Sigma}_f \mathbf{C}^{-1} \mathbf{R}}_{\text{wavy}})$$

$\leftarrow \mathbf{t} = [-0.5, 1, 2, 0, -0.5, -1]^T$
代入

- 对比 Noisy-free GP regression 条件预测分布:

$$p(\mathbf{f}^*|\mathbf{t}; \mathbf{x}, \mathbf{x}^*, \sigma^2) = \mathcal{N}(\mathbf{f}^*; \mathbf{R}^T (\mathbf{C})^{-1} \mathbf{t}, \mathbf{C}^* - \mathbf{R}^T \mathbf{C}^{-1} \mathbf{R})$$

Noisy GP
的预测包含额外信息

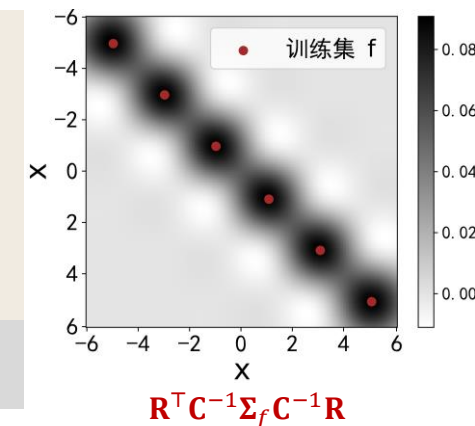
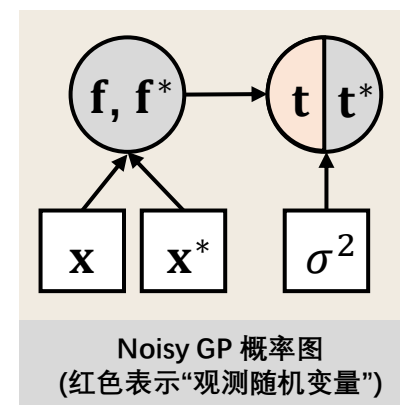
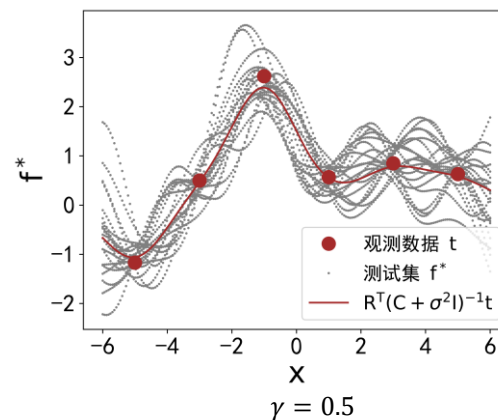
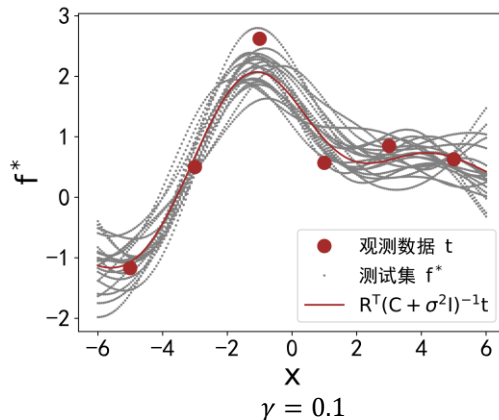
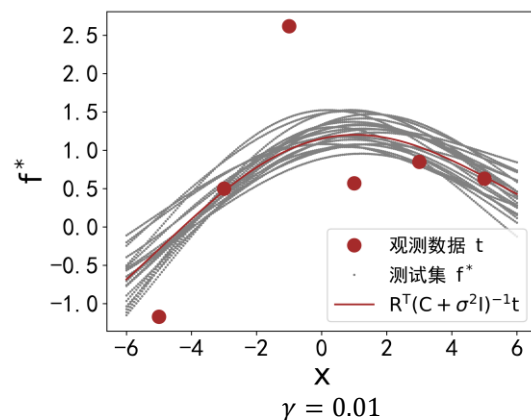
$\leftarrow \mathbf{f} = [-0.5, 1, 2, 0, -0.5, -1]^T$
代入

3.3 Noisy GP regression – 获得测试集 $\mathbf{f}^*$ 和 $\mathbf{t}^*$ 的条件预测分布

● Noisy GP regression $\mathbf{f}^*$ 的条件预测分布:

$$p(\mathbf{f}^*|\mathbf{t}; \mathbf{x}, \mathbf{x}^*, \sigma^2) = \mathcal{N}(\mathbf{f}^*; \mathbf{R}^\top(\mathbf{C} + \sigma^2\mathbf{I})^{-1}\mathbf{t}, \mathbf{C}^* - \mathbf{R}^\top\mathbf{C}^{-1}\mathbf{R} + \mathbf{R}^\top\mathbf{C}^{-1}\Sigma_f\mathbf{C}^{-1}\mathbf{R})$$

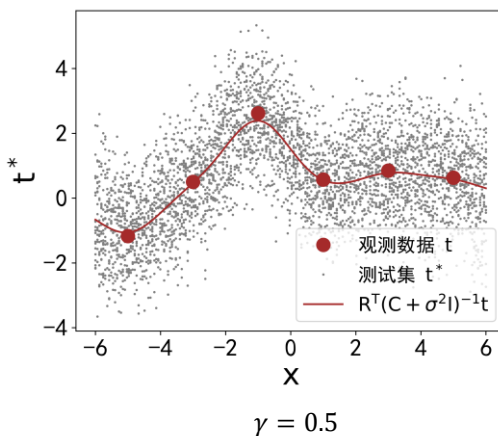
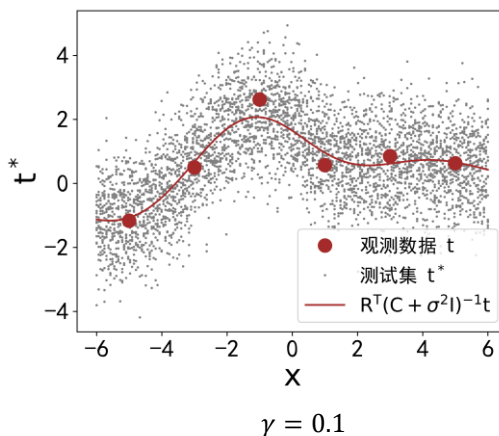
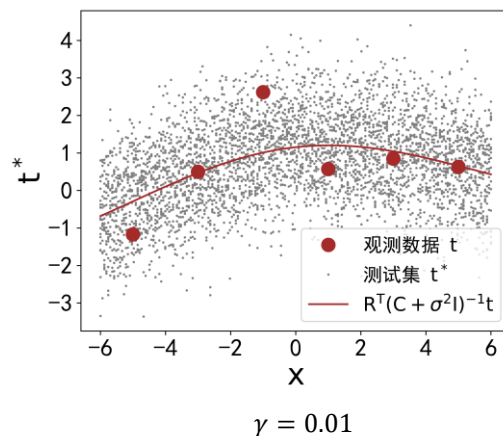
← $\mathbf{t} = [-0.5, 1, 2, 0, -0.5, -1]^\top$
代入



● Noisy GP regression $\mathbf{t}^*$ 的条件预测分布:

$$p(\mathbf{t}^*|\mathbf{f}^*; \sigma^2) = \mathcal{N}(\mathbf{t}^*; \mathbf{f}^*, \sigma^2\mathbf{I})$$

$$t_n = f_n + \epsilon_n, \text{ 其中, } \epsilon_n \sim \mathcal{N}(\epsilon_n; 0, \sigma^2)$$



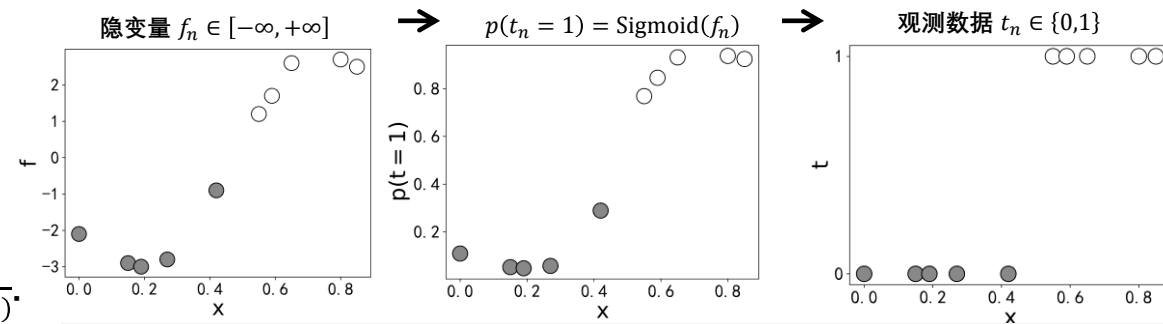
预测 $\mathbf{t}^*$ 的过程

$$t_n = f_n + \epsilon_n, \text{ 其中, } \epsilon_n \sim \mathcal{N}(\epsilon_n; 0, \sigma^2)$$

- ① 从 $p(\mathbf{f}^*|\mathbf{t}, \mathbf{x}, \mathbf{x}^*, \sigma^2)$ 采样得到 $\mathbf{f}^*$
- ② 由 $\mathbf{f}^*$ 确定高斯分布 $p(\mathbf{t}^*|\mathbf{f}^*, \sigma^2)$
- ③ 从 $p(\mathbf{t}^*|\mathbf{f}^*, \sigma^2)$ 采样得到 $\mathbf{t}^*$

4.1 GP classification – 问题描述 (含隐变量 f 的GP)

- 训练集: $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}, \mathbf{f} = \begin{bmatrix} f_1 \\ \dots \\ f_N \end{bmatrix}, \mathbf{t} = \begin{bmatrix} t_1 \\ \dots \\ t_N \end{bmatrix}$, 测试集: $\mathbf{x}^* = \begin{bmatrix} x_1^* \\ \vdots \\ x_L^* \end{bmatrix}, \mathbf{f}^* = \begin{bmatrix} f_1^* \\ \dots \\ f_L^* \end{bmatrix}, \mathbf{t}^* = \begin{bmatrix} t_1^* \\ \dots \\ t_L^* \end{bmatrix}$
- 预测过程: $x_n \rightarrow f_n \rightarrow \text{sigmoid}(f_n) \rightarrow p(t_n = 1) \rightarrow t_n$, $\text{sigmoid}(f_n) = \frac{1}{1+\exp(-f_n)}$

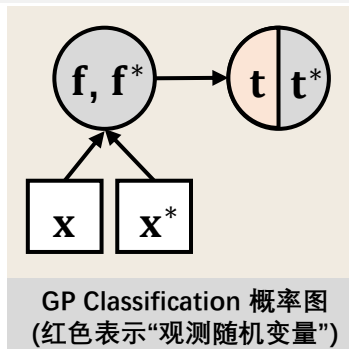


GP classification 预测过程

- **GP classification:** 定义 $\mathbf{f}$ 和 $\mathbf{f}^*$ 的联合高斯分布: $p(\mathbf{f}, \mathbf{f}^*; \mathbf{x}, \mathbf{x}^*) = \mathcal{N}(\mathbf{f}, \mathbf{f}^*; \mathbf{0}, \begin{bmatrix} \mathbf{C} & \mathbf{R} \\ \mathbf{R}^\top & \mathbf{C}^* \end{bmatrix})$. $\rightarrow f_n$ 和 f_l^* 的取值范围满足 $f_n, f_l^* \in [-\infty, +\infty]$

其中, $\mathbf{C} = \begin{bmatrix} c(x_1, x_1) & \dots & c(x_1, x_N) \\ \vdots & \ddots & \vdots \\ c(x_N, x_1) & \dots & c(x_N, x_N) \end{bmatrix}, \mathbf{C}^* = \begin{bmatrix} c(x_1^*, x_1^*) & \dots & c(x_1^*, x_L^*) \\ \vdots & \ddots & \vdots \\ c(x_L^*, x_1^*) & \dots & c(x_L^*, x_L^*) \end{bmatrix}, \mathbf{R} = \begin{bmatrix} c(x_1, x_1^*) & \dots & c(x_1, x_L^*) \\ \vdots & \ddots & \vdots \\ c(x_N, x_1^*) & \dots & c(x_N, x_L^*) \end{bmatrix}$

- **问题:** 为什么不直接定义 $\mathbf{t}$ 和 $\mathbf{t}^*$ 的联合高斯分布 $p(\mathbf{t}, \mathbf{t}^*; \mathbf{x}, \mathbf{x}^*) = \mathcal{N}(\mathbf{t}, \mathbf{t}^*; \mathbf{0}, \begin{bmatrix} \mathbf{C} & \mathbf{R} \\ \mathbf{R}^\top & \mathbf{C}^* \end{bmatrix})$? $\rightarrow$ 注意到, t_n 和 t_l^* 的取值范围为 $t_n, t_l^* \in \{0, 1\}$ 而服从于高斯分布的随机变量取值范围应为 $[-\infty, +\infty]$



- **GP 条件预测分布:** $p(\mathbf{f}^* | \mathbf{f}, \mathbf{x}, \mathbf{x}^*) = \mathcal{N}(\mathbf{f}^*; \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$. $\boldsymbol{\mu}^* = \mathbf{R}^\top \mathbf{C}^{-1} \mathbf{f}, \boldsymbol{\Sigma}^* = \mathbf{C}^* - \mathbf{R}^\top \mathbf{C}^{-1} \mathbf{R}$ $\leftarrow \mathbf{f} = ??$
代入

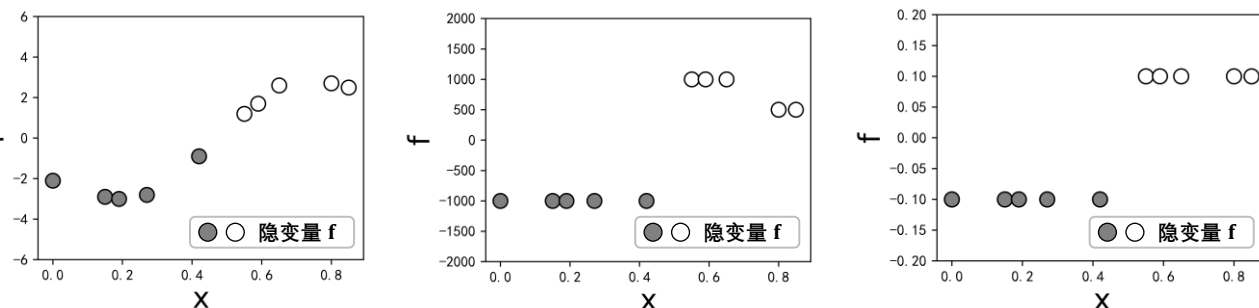
- **存在问题:** 与 Noisy GP regression 中存在相同的问题, “观测数据”是样本标签 $\mathbf{t}$ (0 或 1), 而 $\mathbf{f}$ 是“未知”的, 因此, 无法直接代入上述条件分布中求解。

- 例如, 确定样本标签 $\mathbf{t} = [0, 0, 0, 0, 0, 1, 1, 1, 1, 1]^\top$ 的 $\mathbf{f}$ 可能的取值包括:

$$\mathbf{f} = [-2.1, -2.9, -3, -2.8, -0.9, 1.2, 1.7, 2.6, 2.7, 2.5]^\top$$

$$\mathbf{f} = [-1000, -1000, -1000, -1000, -1000, 1000, 1000, 1000, 1000, 500, 500]^\top$$

$$\mathbf{f} = [-0.1, -0.1, -0.1, -0.1, -0.1, 0.1, 0.1, 0.1, 0.1, 0.1]^\top$$



现假定 $\mathbf{f}$ 已知, 如何利用 $\mathbf{f}$ 对测试集 $\mathbf{f}^*$ 和 $\mathbf{t}^*$ 进行预测?

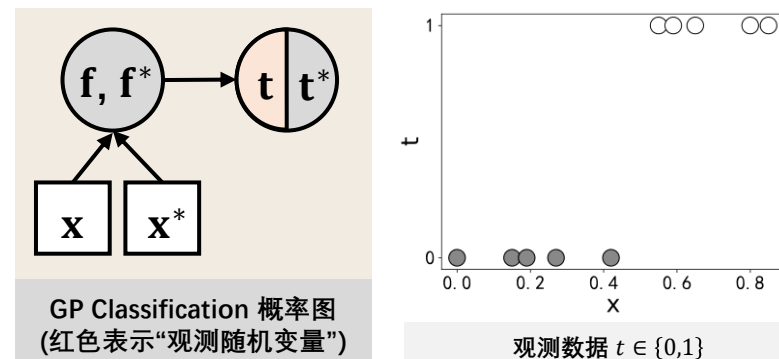
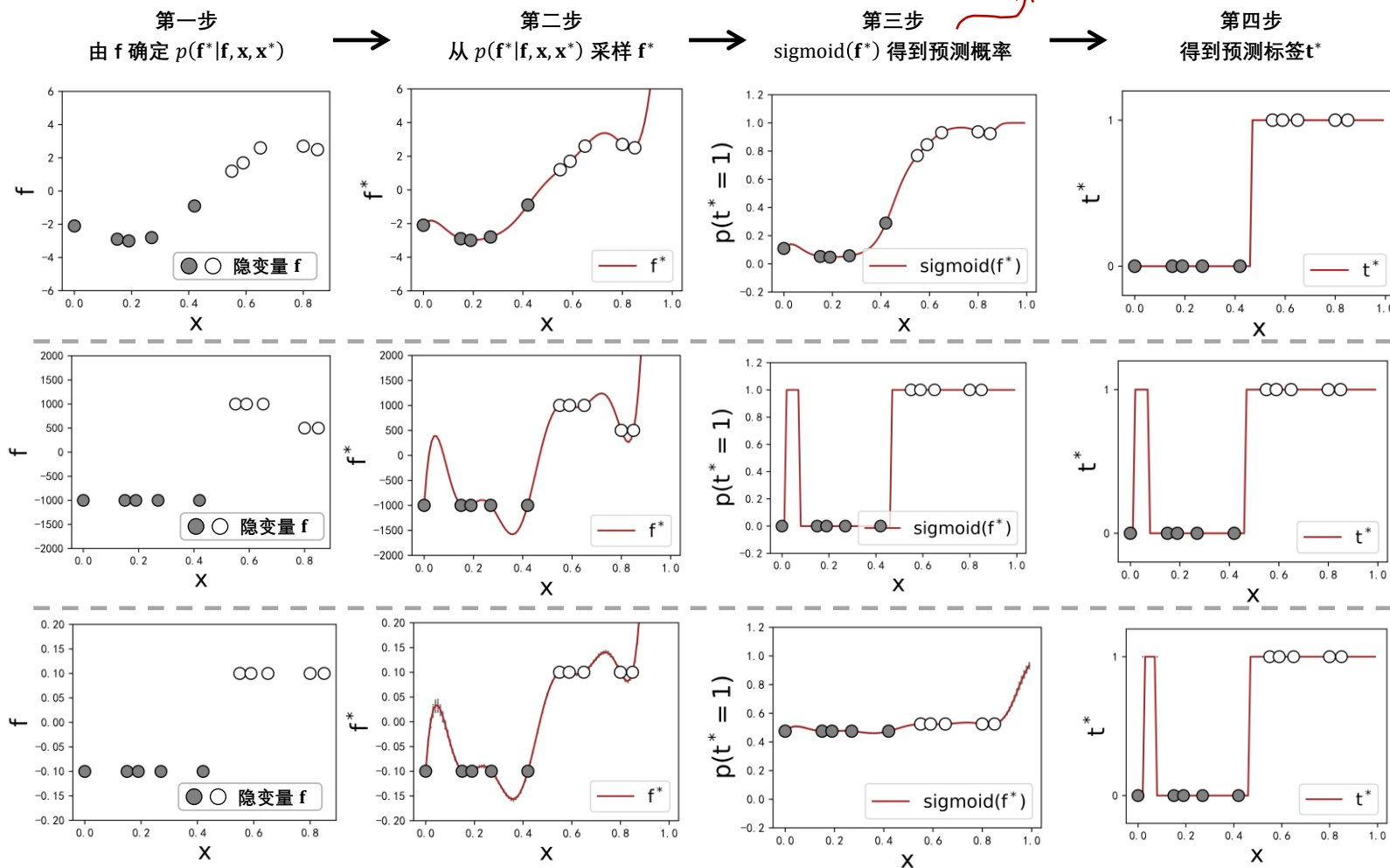
上述3个不同的 $\mathbf{f}$ 取值均可生成“观测数据” $\mathbf{t} = [0, 0, 0, 0, 0, 1, 1, 1, 1, 1]^\top$

4.2 GP classification – 假定隐变量 f 已知, $p(f^*|f; x, x^*)$ 预测过程

- GP 条件预测分布: $p(f^*|f; x, x^*) = \mathcal{N}(f^*; \mu^*, \Sigma^*)$. $\mu^* = R^T C^{-1} f$, $\Sigma^* = C^* - R^T C^{-1} R$

现假定 f 已知, 如何利用 f 对测试集 f^* 和 t^* 进行预测?

不同的隐变量 f 会输出不同的预测结果



然而, “观测数据”是样本标签 t (0 或 1), 而 f 是“未知”的, 因此, 无法直接代入上述条件分布中求解

- 解决方法: 利用“观测数据 t ”和“训练特征 x ”得到后验分布 $p(f|t; x)$, 从而边缘化条件预测分布 $p(f^*|t; x, x^*) = \int p(f^*|f; x, x^*) p(f|t; x) df$.

4.3 GP classification – 获得隐变量 $\mathbf{f}$ 的后验分布 $p(\mathbf{f}|\mathbf{t}; \mathbf{x})$

- GP 条件预测分布: $p(\mathbf{f}^*|\mathbf{f}; \mathbf{x}, \mathbf{x}^*) = \mathcal{N}(\mathbf{f}^*; \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$. $\boldsymbol{\mu}^* = \mathbf{R}^\top \mathbf{C}^{-1} \mathbf{f}$, $\boldsymbol{\Sigma}^* = \mathbf{C}^* - \mathbf{R}^\top \mathbf{C}^{-1} \mathbf{R}$.

↓
存在问题: “观测数据”是样本标签 $\mathbf{t}$ (0 或 1), 而 $\mathbf{f}$ 是“未知”的, 因此, 无法直接代入上述条件分布中求解

- 边缘化GP 条件预测分布: 得到后验分布 $p(\mathbf{f}|\mathbf{t}; \mathbf{x})$, 从而边缘化条件预测分布 $p(\mathbf{f}^*|\mathbf{t}; \mathbf{x}, \mathbf{x}^*) = \int p(\mathbf{f}^*|\mathbf{f}, \mathbf{x}^*) p(\mathbf{f}|\mathbf{t}; \mathbf{x}) d\mathbf{f}$.

- 如何获得隐变量 $\mathbf{f}$ 的后验分布 $p(\mathbf{f}|\mathbf{t}; \mathbf{x})$? $\xrightarrow{\text{Bernoulli分布}} p(\mathbf{f}|\mathbf{t}; \mathbf{x}) = \frac{p(\mathbf{t}|\mathbf{f})p(\mathbf{f}; \mathbf{x})}{\int p(\mathbf{t}|\mathbf{f})p(\mathbf{f}; \mathbf{x}) d\mathbf{f}} \xleftarrow{\text{高斯分布}}$

- ① 似然: $p(\mathbf{t}|\mathbf{f}) = \prod_{n=1}^N p(t_n = 1)^{t_n} (1 - p(t_n = 1))^{1-t_n}$, 其中, $p(t_n = 1) = \text{sigmoid}(f_n)$.

Logistic likelihood

- ② 先验: $p(\mathbf{f}; \mathbf{x}) = \mathcal{N}(\mathbf{f}; \mathbf{0}, \mathbf{C})$, $\mathbf{C} = \begin{bmatrix} c(x_1, x_1) & \cdots & c(x_1, x_N) \\ \vdots & \ddots & \vdots \\ c(x_N, x_1) & \cdots & c(x_N, x_N) \end{bmatrix}$.

- 此时, “似然”与“先验”不共轭, 因此, 后验分布 $p(\mathbf{f}|\mathbf{t})$ 的形式未知, 因此, 无法直接求解其解析式.

↓
可以利用 $p(\mathbf{f}|\mathbf{t}; \mathbf{x}) \propto p(\mathbf{t}|\mathbf{f})p(\mathbf{f}; \mathbf{x})$

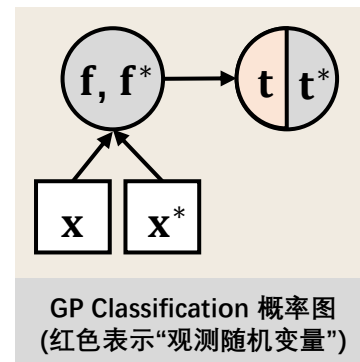
- ① 似然: $p(\mathbf{t}|\mathbf{f}) = \prod_{n=1}^N (t_n)^{p(t_n=1)} (1 - t_n)^{1-p(t_n=1)}$, 其中, $p(t_n = 1) = \text{sigmoid}(f_n)$.

- 方法 ①: 基于 Newton-Raphson 方法获得 $p(\mathbf{f}|\mathbf{t}; \mathbf{x})$ 的 MAP 点估计 $\hat{\mathbf{f}}$.

- 该方法预测方式: 基于 $\hat{\mathbf{f}}$ 获得“近似”条件预测分布: $p(\mathbf{f}^*|\mathbf{t}, \mathbf{x}, \mathbf{x}^*) = \int p(\mathbf{f}^*|\mathbf{f}, \mathbf{x}^*) p(\mathbf{f}|\mathbf{t}; \mathbf{x}) d\mathbf{f} \approx p(\mathbf{f}^*; \hat{\mathbf{f}}, \mathbf{x}^*)$.

- 方法 ②: 基于 Laplace approximation 方法获得 $p(\mathbf{f}|\mathbf{t}; \mathbf{x})$ 的近似分布 $\mathcal{N}(\mathbf{f}; \boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f)$.

- 该方法预测方式: 基于 $p(\mathbf{f}|\mathbf{t}; \mathbf{x})$ 获得“边缘化”条件预测分布: $p(\mathbf{f}^*|\mathbf{t}, \mathbf{x}, \mathbf{x}^*) = \int p(\mathbf{f}^*|\mathbf{f}, \mathbf{x}^*) p(\mathbf{f}|\mathbf{t}; \mathbf{x}) d\mathbf{f}$.



4.4 GP classification – 获得隐变量 $\mathbf{f}$ 的后验分布 $p(\mathbf{f}|\mathbf{t}; \mathbf{x})$ – 方法 ①: Newton–Raphson 方法 (MAP点估计)

- 如何获得隐变量 $\mathbf{f}$ 的后验分布 $p(\mathbf{f}|\mathbf{t}; \mathbf{x})$? $\rightarrow p(\mathbf{f}|\mathbf{t}; \mathbf{x}) = \frac{p(\mathbf{t}|\mathbf{f})p(\mathbf{f}; \mathbf{x})}{\int p(\mathbf{t}|\mathbf{f})p(\mathbf{f}; \mathbf{x})d\mathbf{f}} \rightarrow p(\mathbf{f}|\mathbf{t}; \mathbf{x}) \propto p(\mathbf{t}|\mathbf{f})p(\mathbf{f}; \mathbf{x})$

- Newton-Raphson 方法 (MAP点估计):** 通过多次迭代, 求解使得“似然” $\times$ “先验”最大的 $\hat{\mathbf{f}}$, 即

$$\hat{\mathbf{f}} = \operatorname{argmax}_{\mathbf{f}} \log p(\mathbf{t}|\mathbf{f}) + \log p(\mathbf{f}; \mathbf{x})$$

- 第 ① 步: 随机初始化 $\hat{\mathbf{f}}$, 令 $g(\mathbf{f}) = \log p(\mathbf{t}|\mathbf{f}) + \log p(\mathbf{f}; \mathbf{x})$.

能否令 $\frac{\partial g(\mathbf{f})}{\partial \mathbf{f}} = 0$ 求得 $\mathbf{f}$?

似然: $p(\mathbf{t}|\mathbf{f}) = \prod_{n=1}^N (t_n)^{p(t_n=1)} (1 - t_n)^{1-p(t_n=1)}$

先验: $p(\mathbf{f}; \mathbf{x}) = \mathcal{N}(\mathbf{f}; \mathbf{0}, \mathbf{C})$

- 第 ② 步: 求 $g(\mathbf{f})$ 对于 $\mathbf{f}$ 的一阶、二阶导数矩阵.

$\mathbf{q}$ 中所包含的 $\mathbf{f}$ 无法分离, 无法直接求解 $\hat{\mathbf{f}}$

$$\text{一阶导数: } \frac{\partial g(\mathbf{f})}{\partial \mathbf{f}} = \mathbf{q} - \mathbf{C}^{-1}\mathbf{f}$$

其中, $\mathbf{q} = [q_1, q_2, \dots, q_N]^\top$, $q_n = t_n - P_n$,
 $P_n = p(t_n = 1) = \text{sigmoid}(f_n)$.

$$\text{二阶导数: } \frac{\partial^2 g(\mathbf{f})}{\partial \mathbf{f} \partial \mathbf{f}^\top} = -\mathbf{P} - \mathbf{C}^{-1} \quad \text{其中, } \mathbf{P} = \begin{bmatrix} P_1(1 - P_1) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & P_N(1 - P_N) \end{bmatrix}.$$

- 第 ③ 步: 将当前 $\hat{\mathbf{f}}$ 代入一阶、二阶导数矩阵.

- 第 ④ 步: 更新 $\hat{\mathbf{f}}$.

$$\hat{\mathbf{f}} \leftarrow \hat{\mathbf{f}} - \left(\frac{\partial^2 g(\mathbf{f})}{\partial \mathbf{f} \partial \mathbf{f}^\top} \right)^{-1} \bigg|_{\hat{\mathbf{f}}} \cdot \frac{\partial g(\mathbf{f})}{\partial \mathbf{f}} \bigg|_{\hat{\mathbf{f}}}$$

- 第 ⑤ 步: 重复 ② ~ ④, 直至收敛.

- 第 ⑥ 步: 输出 $\hat{\mathbf{f}}$.

- 预测:** $p(\mathbf{f}^*; \hat{\mathbf{f}}, \mathbf{x}^*) = \mathcal{N}(\mathbf{f}^*; \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$. $\boldsymbol{\mu}^* = \mathbf{R}^\top \mathbf{C}^{-1} \hat{\mathbf{f}}$, $\boldsymbol{\Sigma}^* = \mathbf{C}^* - \mathbf{R}^\top \mathbf{C}^{-1} \mathbf{R}$.

Comment 4.1 – The Newton–Raphson method: The Newton–Raphson method (also known as the Newton method) is a general method for finding points where functions are equal to zero, i.e., finding points where the function $f(x) = 0$. Given a current estimate of the zero point, x_n , we update it by moving to the point where the tangent to the function at x_n passes through the x -axis. This point can be computed by approximating the gradient as a change in $f(x)$ divided by a change in x . Defining $\partial f(x)/\partial x$ as $f'(x)$,

$$\begin{aligned} f'(x_n) &= \frac{f(x_n) - 0}{x_n - x_{n+1}} \\ (x_n - x_{n+1})f'(x_n) &= f(x_n) \\ x_{n+1} &= x_n - \frac{f(x_n)}{f'(x_n)}. \end{aligned}$$

The method can also be used to find minima and maxima, as these are simply points where the gradient passes through zero. Therefore, we simply replace $f(x)$ with its derivative $f'(x)$ and $f'(x)$ with its derivative $f''(x)$:

任意函数都可用该方式迭代求得
 导数 $f'(x)$ 为 0 的点
 即最大(小)值点

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)} \quad \begin{aligned} f''(x_n) > 0 &\Rightarrow f(x_n) \text{ 为最大值} \\ f''(x_n) < 0 &\Rightarrow f(x_n) \text{ 为最小值} \end{aligned}$$

This is readily extendable to functions of a vector – say $\mathbf{x}$. In this instance, $f'(x_n)$ is replaced by the vector of partial derivatives evaluated at $\mathbf{x}_n$ and $1/f''(x_n)$ is replaced by the inverse of the Hessian matrix (see Comment 2.6) – $\partial^2 f(\mathbf{x})/\partial \mathbf{x} \partial \mathbf{x}^\top$ – evaluated at $\mathbf{x} = \mathbf{x}_n$.

4.4 GP classification – 获得隐变量 f 的后验分布 $p(f|t; \mathbf{x})$ – 方法 ①: Newton-Raphson 方法 (MAP点估计)

● Newton-Raphson 方法 (MAP点估计):

$$\hat{\mathbf{f}} = \underset{\mathbf{f}}{\operatorname{argmax}} \log p(\mathbf{t}|\mathbf{f}) + \log p(\mathbf{f}; \mathbf{x})$$

• 第 ① 步: 随机初始化 $\hat{\mathbf{f}}$, 令 $g(\mathbf{f}) = \log p(\mathbf{t}|\mathbf{f}) + \log p(\mathbf{f}; \mathbf{x})$.

• 第 ② 步: 求 $g(\mathbf{f})$ 对于 $\mathbf{f}$ 的一阶、二阶导数矩阵.

$$\text{一阶导数: } \frac{\partial g(\mathbf{f})}{\partial \mathbf{f}} = \mathbf{q} - \mathbf{C}^{-1}\mathbf{f}$$

其中, $\mathbf{q} = [q_1, q_2, \dots, q_N]^\top$, $q_n = t_n - P_n$,

$$P_n = p(t_n = 1).$$

$$\text{二阶导数: } \frac{\partial^2 g(\mathbf{f})}{\partial \mathbf{f} \partial \mathbf{f}^\top} = -\mathbf{P} - \mathbf{C}^{-1}$$

$$\text{其中, } \mathbf{P} = \begin{bmatrix} P_1(1-P_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & P_N(1-P_N) \end{bmatrix}.$$

• 第 ③ 步: 将当前 $\hat{\mathbf{f}}$ 代入一阶、二阶导数矩阵.

• 第 ④ 步: 更新 $\hat{\mathbf{f}}$.

$$\hat{\mathbf{f}} \leftarrow \hat{\mathbf{f}} - \left(\frac{\partial^2 g(\mathbf{f})}{\partial \mathbf{f} \partial \mathbf{f}^\top} \right)^{-1} \bigg|_{\hat{\mathbf{f}}} \cdot \frac{\partial g(\mathbf{f})}{\partial \mathbf{f}} \bigg|_{\hat{\mathbf{f}}}$$

• 第 ⑤ 步: 重复 ② ~ ④, 直至收敛.

• 第 ⑥ 步: 输出 $\hat{\mathbf{f}}$.

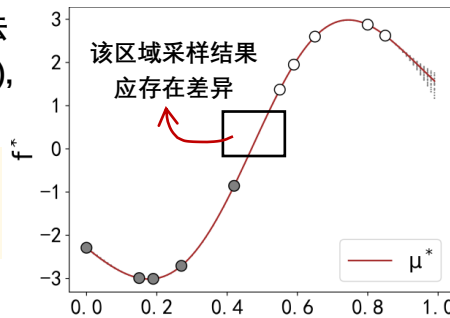
$$\hat{\mathbf{f}} = [-2.3, -3.0, -3.0, -2.7, -0.9, 1.4, 1.9, 2.6, 2.9, 2.6]^\top$$

● 预测: $p(\mathbf{f}^*; \hat{\mathbf{f}}, \mathbf{x}^*) = \mathcal{N}(\mathbf{f}^*; \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$. $\boldsymbol{\mu}^* = \mathbf{R}^\top \mathbf{C}^{-1} \hat{\mathbf{f}}$, $\boldsymbol{\Sigma}^* = \mathbf{C}^* - \mathbf{R}^\top \mathbf{C}^{-1} \mathbf{R}$

• **存在的问题:** 如右图所示, Newton-Raphson方法虽能够准确地分类 (均值 $\boldsymbol{\mu}^*$ 曲线穿过所有观测点), 但采样结果单一 (方差 $\boldsymbol{\Sigma}^*$ 的评估不够合理).

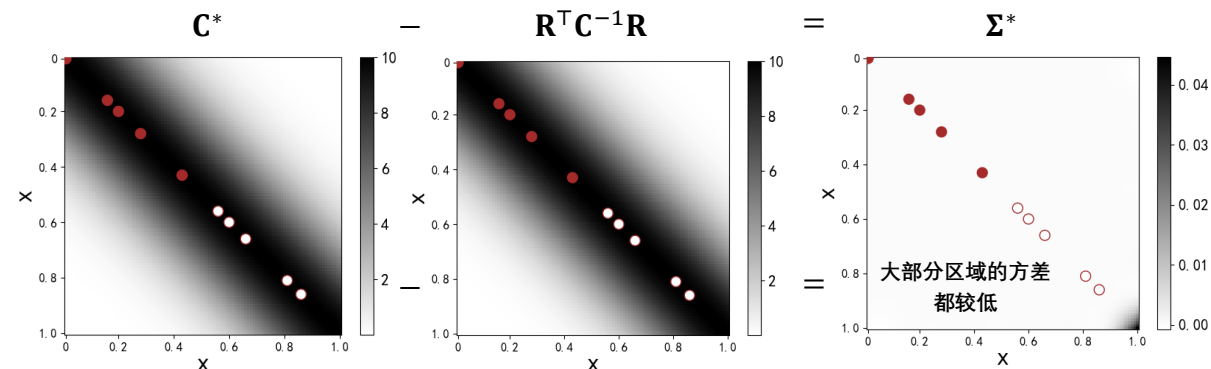
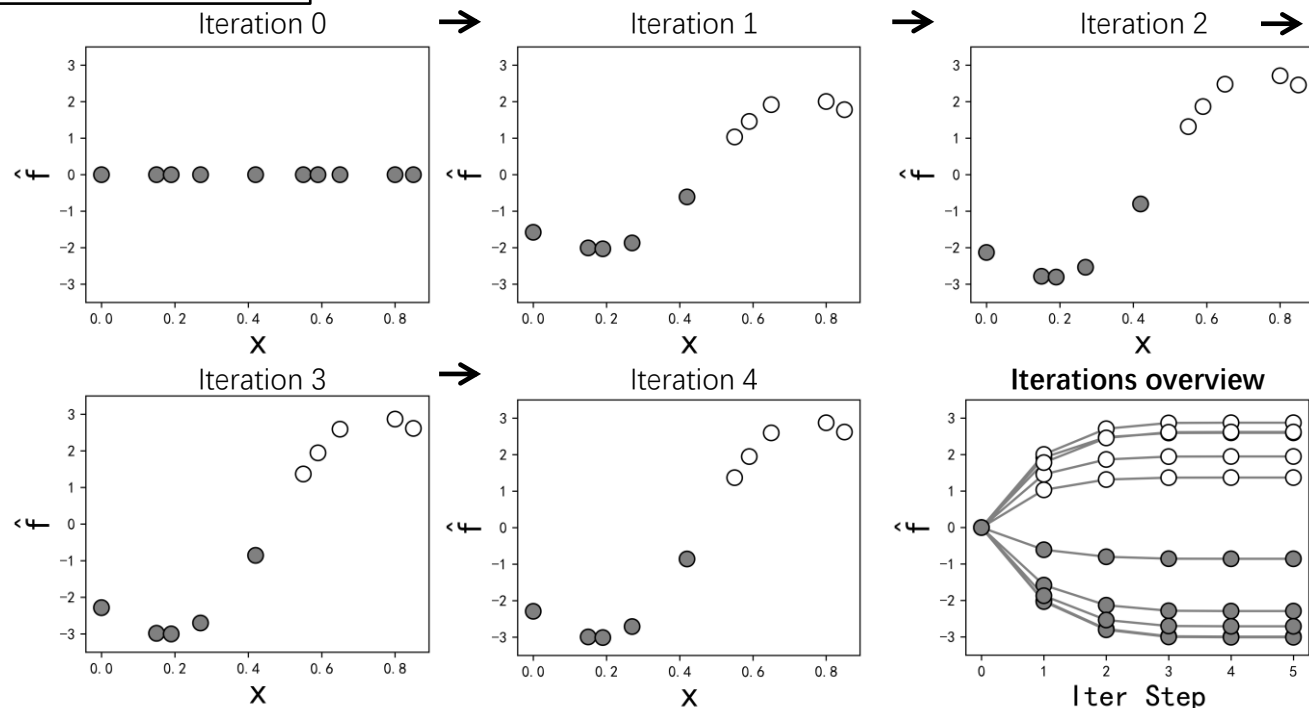
• **原因:** 直接使用点估计 $\hat{\mathbf{f}}$ 来预测, 而并未考虑隐变量 $\mathbf{f}$ 的不确定性, 即 $p(\mathbf{f}|\mathbf{t}; \mathbf{x})$.

● 方法 ② Laplace approximation 方法: 获得 $p(\mathbf{f}|\mathbf{t}; \mathbf{x})$ 的近似分布, 更准确地度量预测方差.



预测均值和采样结果 (阴影部分)

上图表示在 Newton-Raphson 方法的每次迭代步 (第 ④ 步) 中, 隐变量 $\hat{\mathbf{f}}$ 的变化



上图为当RBF核函数参数 $\alpha = 10$, $\gamma = 10$ 时, 获得条件预测分布 $\boldsymbol{\Sigma}^*$ 所涉及各矩阵 ($\mathbf{C}^*$, $\mathbf{R}^\top \mathbf{C}^{-1} \mathbf{R}$, $\boldsymbol{\Sigma}^*$)

4.5 GP classification – 获得隐变量 $\mathbf{f}$ 的后验分布 $p(\mathbf{f}|\mathbf{t}; \mathbf{x})$ – 方法 ②: Laplace approximation 方法 (近似后验分布)

- 如何获得隐变量 $\mathbf{f}$ 的后验分布 $p(\mathbf{f}|\mathbf{t}; \mathbf{x})$? $\rightarrow p(\mathbf{f}|\mathbf{t}; \mathbf{x}) = \frac{p(\mathbf{t}|\mathbf{f})p(\mathbf{f}; \mathbf{x})}{\int p(\mathbf{t}|\mathbf{f})p(\mathbf{f}; \mathbf{x})d\mathbf{f}} = \frac{g(\mathbf{f}; \mathbf{x})}{Z}$, 令 $g(\mathbf{f}) = p(\mathbf{t}|\mathbf{f})p(\mathbf{f}; \mathbf{x})$, $Z = \int p(\mathbf{t}|\mathbf{f})p(\mathbf{f}; \mathbf{x})d\mathbf{f}$.

- Laplace approximation 方法:** 利用高斯分布 $\mathcal{N}(\mathbf{f}; \boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f)$, 近似后验分布 $p(\mathbf{f}|\mathbf{t})$, 即

$$p(\mathbf{f}|\mathbf{t}; \mathbf{x}) \approx \mathcal{N}(\mathbf{f}; \boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f)$$

- 第 ① 步: 利用 Newton-Raphson 方法获得 MAP 点估计 $\hat{\mathbf{f}}$.
- 第 ② 步: 将 $\log p(\mathbf{f}|\mathbf{t}; \mathbf{x}) = \log \frac{g(\mathbf{f})}{Z} = \log g(\mathbf{f}) - \log Z$ 在 $\mathbf{f} = \hat{\mathbf{f}}$ 处进行泰勒展开

$$\log p(\mathbf{f}|\mathbf{t}) = \log g(\hat{\mathbf{f}}) - \log Z + (\mathbf{f} - \hat{\mathbf{f}}) \cdot \left. \frac{\partial g(\mathbf{f})}{\partial \mathbf{f}} \right|_{\hat{\mathbf{f}}} + \frac{1}{2} (\mathbf{f} - \hat{\mathbf{f}})^\top \cdot \left. \left(\frac{\partial^2 g(\mathbf{f})}{\partial \mathbf{f} \partial \mathbf{f}^\top} \right) \right|_{\hat{\mathbf{f}}} \cdot (\mathbf{f} - \hat{\mathbf{f}}) + \dots$$

$\downarrow$ 常数项 一次项=0 二次项

$$\log p(\mathbf{f}|\mathbf{t}) \approx \log K - \frac{1}{2} (\mathbf{f} - \hat{\mathbf{f}})^\top \cdot \mathbf{v} \cdot (\mathbf{f} - \hat{\mathbf{f}})$$

$$\text{省略3次及其之后项, 并令 } \mathbf{v} = - \left. \left(\frac{\partial^2 g(\mathbf{f})}{\partial \mathbf{f} \partial \mathbf{f}^\top} \right) \right|_{\hat{\mathbf{f}}}$$

$$\log \mathcal{N}(\mathbf{f}; \boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f) = \log K - \frac{1}{2} (\mathbf{f} - \boldsymbol{\mu}_f)^\top \cdot \boldsymbol{\Sigma}_f^{-1} \cdot (\mathbf{f} - \boldsymbol{\mu}_f)$$

$$\text{比较可得: } \boldsymbol{\mu}_f = \hat{\mathbf{f}}, \boldsymbol{\Sigma}_f = \mathbf{v}^{-1}$$

- 第 ③ 步: 输出

$$p(\mathbf{f}|\mathbf{t}) \approx \mathcal{N}(\mathbf{f}; \boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f), \text{ 其中 } \boldsymbol{\mu}_f = \hat{\mathbf{f}}, \boldsymbol{\Sigma}_f = - \left. \left(\frac{\partial^2 g(\mathbf{f})}{\partial \mathbf{f} \partial \mathbf{f}^\top} \right) \right|_{\hat{\mathbf{f}}}^{-1}$$

- 预测:** $p(\mathbf{f}^*|\mathbf{t}, \mathbf{x}, \mathbf{x}^*) = \int p(\mathbf{f}^*|\mathbf{f}, \mathbf{x}^*)p(\mathbf{f}|\mathbf{t}; \mathbf{x})d\mathbf{f}$

Comment 4.3 – Taylor expansions: The Taylor expansion is a way of approximating a function. The approximation is always made ‘about’ some value – the approximation will tend to diverge from the true function as we move away from that value. The definition of the Taylor series of $f(w)$ about $\hat{w}$ is

$$f(w) = \sum_{n=0}^{\infty} \frac{(w - \hat{w})^n}{n!} \left. \frac{\partial^n f(w)}{\partial w^n} \right|_{\hat{w}}$$

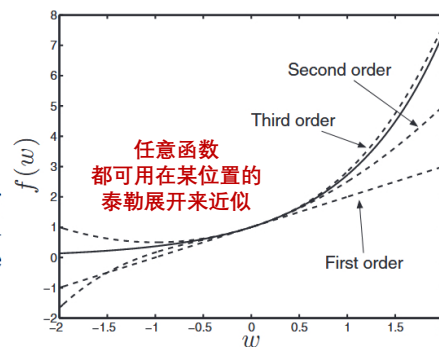
where $\left. \frac{\partial^n f(w)}{\partial w^n} \right|_{\hat{w}}$ is the n th derivative of $f(w)$ with respect to w , evaluated at $\hat{w}$. When $n = 0$, this derivative is simply the function $f(w)$. If we only compute a finite number of terms, we will have an approximation to the function. A first-order approximation would just include terms $n = 0$ and $n = 1$ – an n th-order approximation includes all terms up to and including term n . For example, we can approximate $f(w) = \exp(w)$ at $\hat{w} = 0$:

$$\exp(w) = \exp(\hat{w}) + \frac{w}{1!} \exp(\hat{w}) + \frac{w^2}{2!} \exp(\hat{w}) + \dots$$

Now, $\exp(\hat{w}) = 1$, so

$$\exp(w) = 1 + \frac{w}{1!} + \frac{w^2}{2!} + \frac{w^3}{3!} + \dots$$

The approximation will get better and better as we add more and more terms. This can be seen in the figure on the right.

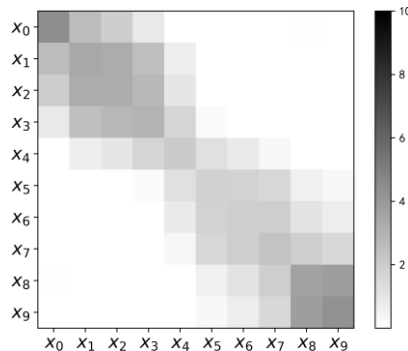


4.5 GP classification – 获得隐变量 $\mathbf{f}$ 的后验分布 $p(\mathbf{f}|\mathbf{t}; \mathbf{x})$ – 方法 ②：Laplace approximation 方法 (近似后验分布)

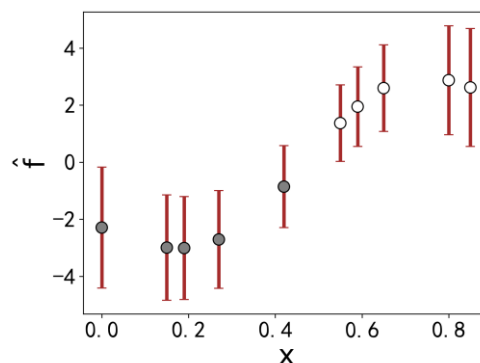
- Laplace approximation 方法:** 利用高斯分布 $\mathcal{N}(\mathbf{f}; \boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f)$, 近似后验分布 $p(\mathbf{f}|\mathbf{t})$, 即

$$p(\mathbf{f}|\mathbf{t}; \mathbf{x}) \approx \mathcal{N}(\mathbf{f}; \boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f), \text{ 其中, } \boldsymbol{\mu}_f = \hat{\mathbf{f}}, \boldsymbol{\Sigma}_f = - \left(\frac{\partial^2 g(\mathbf{f})}{\partial \mathbf{f} \partial \mathbf{f}^\top} \right)^{-1} \Big|_{\hat{\mathbf{f}}}$$

通过 $\boldsymbol{\Sigma}_f$
可进一步影响
预测值 $\mathbf{f}^*$ 的采样结果



近似后验分布 $p(\mathbf{f}|\mathbf{t})$ 的方差 $\boldsymbol{\Sigma}_f$



$\boldsymbol{\Sigma}_f$ 的对角线元素

- 预测:** $p(\mathbf{f}^*|\mathbf{t}, \mathbf{x}, \mathbf{x}^*) = \int p(\mathbf{f}^*|\mathbf{f}, \mathbf{x}^*)p(\mathbf{f}|\mathbf{t}; \mathbf{x})d\mathbf{f} \approx \int \mathcal{N}(\mathbf{f}^*|\mathbf{f}; \mathbf{x}^*)\mathcal{N}(\mathbf{f}; \boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f)d\mathbf{f}$, 即

利用结论: $\int \mathcal{N}(\mathbf{a}; \mathbf{B}\mathbf{c}, \mathbf{D})\mathcal{N}(\mathbf{c}; \mathbf{e}, \mathbf{F})d\mathbf{c} = \mathcal{N}(\mathbf{a}; \mathbf{B}\mathbf{e}, \mathbf{D} + \mathbf{B}\mathbf{F}\mathbf{B}^\top)$

方法 ①：Newton-Raphson 预测结果

$$p(\mathbf{f}^*|\hat{\mathbf{f}}, \mathbf{t}, \mathbf{x}, \mathbf{x}^*) = \mathcal{N}(\mathbf{f}^*; \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*).$$

$$\boldsymbol{\mu}^* = \mathbf{R}^\top \mathbf{C}^{-1} \hat{\mathbf{f}}, \boldsymbol{\Sigma}^* = \mathbf{C}^* - \mathbf{R}^\top \mathbf{C}^{-1} \mathbf{R}$$

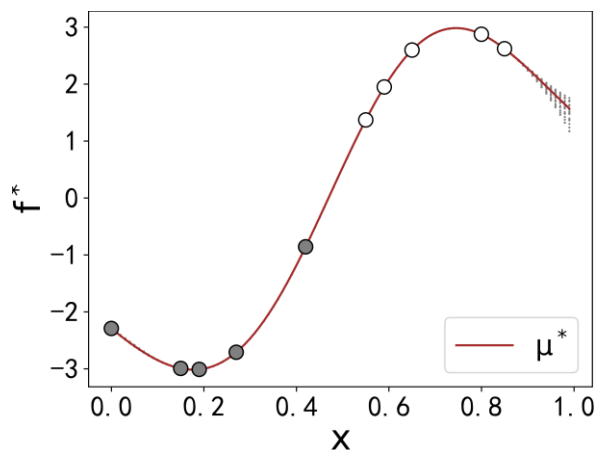
预测结果区别仅存在于方差 $\boldsymbol{\Sigma}^*$

方法 ②：Laplace approximation 预测结果

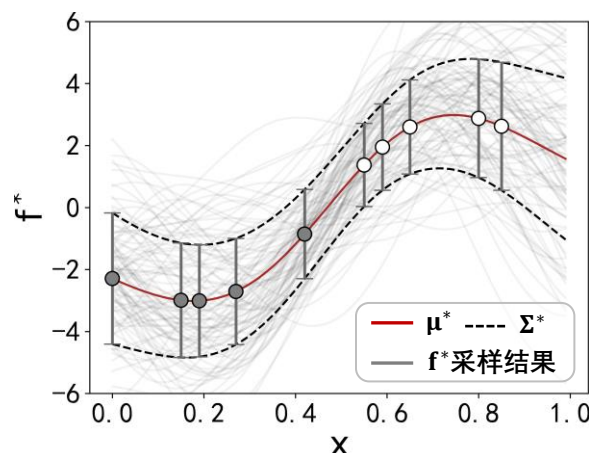
$$p(\mathbf{f}^*|\mathbf{t}, \mathbf{x}, \mathbf{x}^*) = \mathcal{N}(\mathbf{f}^*; \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$$

$$\boldsymbol{\mu}^* = \mathbf{R}^\top \mathbf{C}^{-1} \hat{\mathbf{f}}, \boldsymbol{\Sigma}^* = \mathbf{C}^* - \mathbf{R}^\top \mathbf{C}^{-1} \mathbf{R} + \underbrace{\mathbf{R}^\top \mathbf{C}^{-1} \boldsymbol{\Sigma}_f \mathbf{C}^{-1} \mathbf{R}}_{\text{Laplace approximation term}}$$

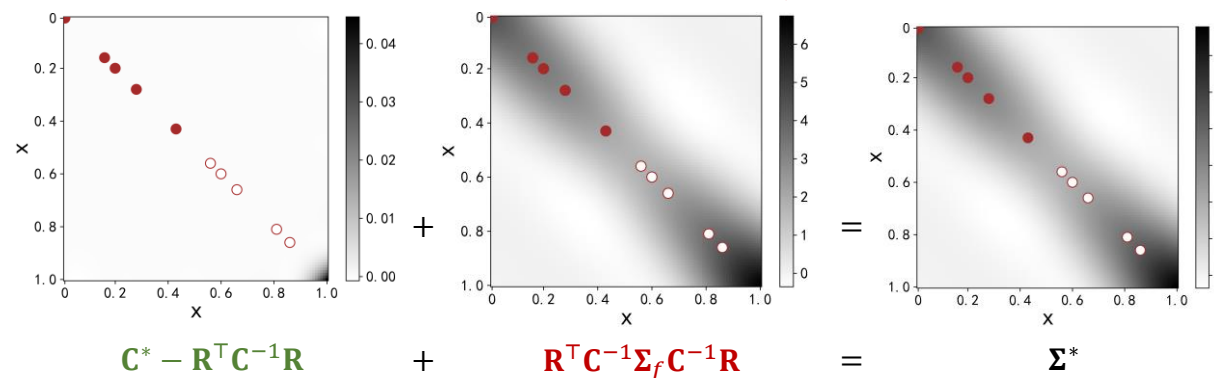
近似分布的方差 $\boldsymbol{\Sigma}_f$ 会影响预测值 $\mathbf{f}^*$ 的采样结果



方法 ①: Newton-Raphson 预测结果



方法 ②: Laplace approximation 预测结果



上图为当RBF核函数参数 $\alpha = 10$, $\gamma = 10$ 时, 获得条件预测分布 $\boldsymbol{\Sigma}^*$ 所涉及的各矩阵 ($\mathbf{C}^*$, $\mathbf{R}^\top \mathbf{C}^{-1} \mathbf{R}$, $\boldsymbol{\Sigma}^*$)

因此, Laplace approximation 通过近似后验分布 $p(\mathbf{f}|\mathbf{t}; \mathbf{x})$ 能够提高模型的不确定性 (方差 $\boldsymbol{\Sigma}^*$) 评估能力