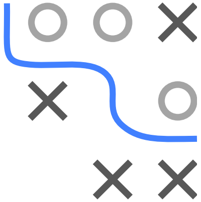
TRAINING OF A GAUSSIAN PROCESS



- To make predictions for a regression task by a Gaussian process, one simply needs to perform matrix computations.
- But for this to work out, we assume that the covariance functions is fully given, including all of its hyperparameters.
- A very nice property of GPs is that we can learn the numerical hyperparameters of a selected covariance function directly during GP training.

TRAINING A GP VIA MAXIMUM LIKELIHOOD

Let us assume

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

where $f(\mathbf{x}) \sim \mathcal{GP}\left(\mathbf{0}, k\left(\mathbf{x}, \mathbf{x}' | \theta\right)\right)$.

Observing $\boldsymbol{y} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{K} + \sigma^2 \boldsymbol{I}\right)$, the marginal log-likelihood (or evidence) is

$$\begin{aligned} \log p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\theta}) &= \log\left[(2\pi)^{-n/2} |\boldsymbol{K}_y|^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{y}^\top \boldsymbol{K}_y^{-1} \boldsymbol{y}\right)\right] \\ &= -\frac{1}{2}\boldsymbol{y}^T \boldsymbol{K}_y^{-1} \boldsymbol{y} - \frac{1}{2}\log|\boldsymbol{K}_y| - \frac{n}{2}\log 2\pi. \end{aligned}$$

with $\boldsymbol{K}_y := \boldsymbol{K} + \sigma^2 \boldsymbol{I}$ and $\boldsymbol{\theta}$ denoting the hyperparameters (the parameters of the covariance function).

TRAINING A GP: EXAMPLE

To visualize this, we consider a zero-mean Gaussian process with squared exponential kernel

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2\ell^2}\|\mathbf{x} - \mathbf{x}'\|^2\right),$$

- Recall, the model is smoother and less complex for higher length-scale ℓ.
- We show how the
  - data fit $-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{K}_y^{-1}\boldsymbol{y}$,
  - the complexity penalty $-\frac{1}{2}\log|\boldsymbol{K}_y|$, and
  - the overall value of the marginal likelihood $\log p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\theta})$

  behave for increasing value of ℓ.

TRAINING A GP VIA MAXIMUM LIKELIHOOD

To set the hyperparameters by maximizing the marginal likelihood, we seek the partial derivatives w.r.t. the hyperparameters

$$
\begin{aligned}
\frac{\partial}{\partial \theta_j} \log p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\theta}) &= \frac{\partial}{\partial \theta_j} \left(-\frac{1}{2} \boldsymbol{y}^T \boldsymbol{K}_y^{-1} \boldsymbol{y} - \frac{1}{2} \log |\boldsymbol{K}_y| - \frac{n}{2} \log 2\pi \right) \\
&= \frac{1}{2} \mathbf{y}^\top \boldsymbol{K}^{-1} \frac{\partial \boldsymbol{K}}{\partial \theta_j} \boldsymbol{K}^{-1} \mathbf{y} - \frac{1}{2} \mathrm{tr} \left(\boldsymbol{K}^{-1} \frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}} \right) \\
&= \frac{1}{2} \mathrm{tr} \left((\boldsymbol{K}^{-1} \boldsymbol{y} \boldsymbol{y}^T \boldsymbol{K}^{-1} - \boldsymbol{K}^{-1}) \frac{\partial \boldsymbol{K}}{\partial \theta_j} \right)
\end{aligned}
$$

using $\frac{\partial}{\partial \theta_j} \boldsymbol{K}^{-1} = -\boldsymbol{K}^{-1} \frac{\partial \boldsymbol{K}}{\partial \theta_j} \boldsymbol{K}^{-1}$ and $\frac{\partial}{\partial \boldsymbol{\theta}} \log |\boldsymbol{K}| = \mathrm{tr} \left(\boldsymbol{K}^{-1} \frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}} \right)$.