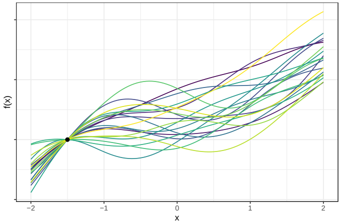
# Introduction to Machine Learning

## Gaussian Processes
## Mean functions for GPs



### Learning goals

- Trends can be modeled via specification of the mean function

# ZERO-MEAN FUNCTIONS

- Previously: common assumption of zero-mean prior

$$m(\mathbf{x}) \equiv 0$$

- Prior knowledge + inference solely handled via $k(\cdot, \cdot)$

- Implication: $m(\cdot)$ not relevant for posterior process

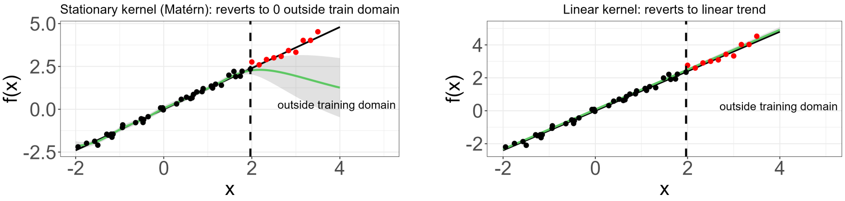
$$\mathbf{m}_{\text{post}} = \mathbb{E}(\mathbf{f}_* | \mathbf{X}_*, \mathbf{X}, \mathbf{y}_*) = \mathbf{K}_* \mathbf{K}_y^{-1} \mathbf{y}, \quad \mathbf{K}_{\text{post}} = \mathbf{K}_{**} - \mathbf{K}_*^T \mathbf{K}_y^{-1} \mathbf{K}_*$$

- Not necessarily drastic limitation: **posterior** mean generally $\neq 0$

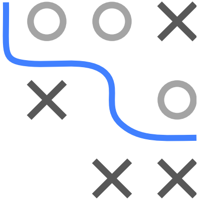
- If data follow some trend $m(\boldsymbol{X})$, we can always center them by subtracting $m(\boldsymbol{X}) \Rightarrow \mathcal{GP}(\mathbf{0}, k(\cdot, \cdot))$ applicable again

# TREND VIA COVARIANCE STRUCTURE

- For zero-mean GPs with stationary kernels, posterior mean reverts to the prior further outside the training domain (no extrapolation)
- But trend-like behaviour could be directly encoded in $k(\cdot, \cdot)$:
  - Linear kernel: $k(\mathbf{x}, \mathbf{x}') = \sigma^2 \mathbf{x}^\top \mathbf{x}'$
  - Polynomial kernels for global polynomial trends
  - Composite kernels: $k = k_{\text{long}} + k_{\text{short}}$
- Produces non-reverting priors even with $m(\mathbf{x}) = 0$, but lower interpretability and kernel-dependent extrapolation
- Consider GP for DGP with linear trend:



# WHY MODEL A TREND EXPLICITLY?



- Still: can make sense to model $m(\cdot)$ explicitly as potentially nonzero
  - **Efficiency:** kernel $k(\cdot,\cdot)$ need not mimic global structure via very long lengthscales
  - **Extrapolation:** outside data range, $\mathcal{GP}(\mathbf{0}, k(\cdot,\cdot))$ reverts to flat mean
    $\Rightarrow$ often unrealistic
  - **Interpretability:** clear separation between systematic trend and stochastic fluctuations
  - **Prior knowledge:** encode known effects (linear, seasonal, additive)

- Assuming $\mathcal{GP}(m(\cdot), k(\cdot,\cdot))$, posterior mean with $m(\cdot)$ becomes

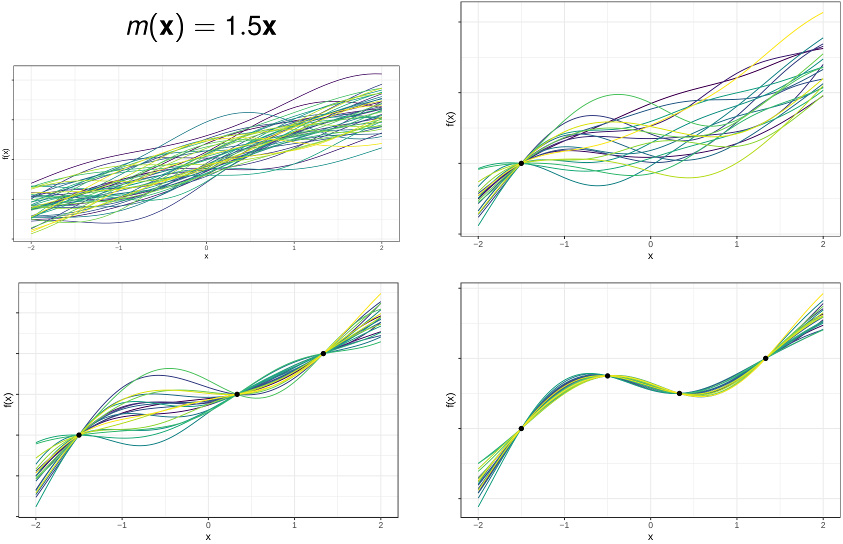
$$\mathbf{m}_{\text{post}}(\mathbf{X}_*) = m(\mathbf{X}_*) + \mathbf{K}_*\mathbf{K}_y^{-1}(\mathbf{y} - m(\mathbf{X}))$$

- Trend $m(\mathbf{X}_*)$ = interpretable global component; Correction = GP adjustment around this trend; Variance stays = $\mathbf{K}_{**} - \mathbf{K}_*^\top \mathbf{K}_y^{-1}\mathbf{K}_*$

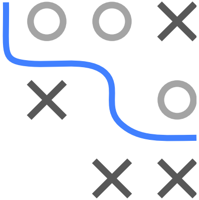
# NON-ZERO-MEAN FUNCTIONS

- GPs with **trend**

$$m(\mathbf{x}) = 1.5\mathbf{x}$$



# SEMI-PARAMETRIC GP



- (Deterministic) mean functions $m(\cdot)$ often hard to specify

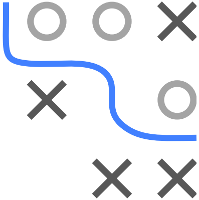
- Solution: **semi-parametric** GPs combining global (often linear) model + zero-mean GP for residuals

$$g(\mathbf{x}) = m_{\boldsymbol{\beta}}(\mathbf{x}) + f(\mathbf{x}), \quad f \sim \mathcal{GP}(\mathbf{0}, k(\cdot, \cdot))$$

- In principle: **any model** $m(\cdot)$ can be used
  - Fixed parametric: $m_{\boldsymbol{\beta}}(\mathbf{x}) = \beta_0 + \mathbf{x}^\top \boldsymbol{\beta}$
  - Basis expansions: $m_{\boldsymbol{\beta}}(\mathbf{x}) = b(\mathbf{x})^\top \boldsymbol{\beta}$
  - Flexible ML models: GLMs, boosting, neural nets, . . .

# ESTIMATION APPROACHES

▸ Rasmussen and Williams 2006



- Log marginal likelihood:

  $$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) = -\tfrac{1}{2}\boldsymbol{r}^\top \mathbf{K}_y^{-1}\boldsymbol{r} - \tfrac{1}{2}\log|\mathbf{K}_y| - \tfrac{n}{2}\log(2\pi),$$

  with $\boldsymbol{r} = \mathbf{y} - m_{\boldsymbol{\beta}}(\mathbf{X})$

- **Joint estimation:** maximize $\ell$ over all parameters

- **Sequential:** fit $m(\cdot)$ first, GP on residuals
  $\Rightarrow$ ignores uncertainty from first stage, variance underestimated

- **Fully Bayesian:** priors on $(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2)$, posterior inference via MCMC or VI

  $$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2 \mid \mathbf{y}, \mathbf{X}) \;\propto\; p(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2, \mathbf{X})\, p(\boldsymbol{\beta})\, p(\boldsymbol{\theta})\, p(\sigma^2)$$

- For complex $m(\cdot)$, estimation by full Bayesian inference or joint likelihood becomes computationally difficult

# SEPARABILITY OF GRADIENTS

- Gradients of $\ell$ decompose neatly into:

$$\nabla_{\boldsymbol{\beta}}\ell = \left(\frac{\partial m_{\boldsymbol{\beta}}(\mathbf{X})}{\partial\boldsymbol{\beta}}\right)^{\top}\mathbf{K}_y^{-1}\boldsymbol{r},$$

$$\nabla_{\boldsymbol{\theta}}\ell = \tfrac{1}{2}\boldsymbol{r}^{\top}\mathbf{K}_y^{-1}\frac{\partial\mathbf{K}_y}{\partial\boldsymbol{\theta}}\mathbf{K}_y^{-1}\boldsymbol{r} - \tfrac{1}{2}\mathrm{tr}\Big(\mathbf{K}_y^{-1}\frac{\partial\mathbf{K}_y}{\partial\boldsymbol{\theta}}\Big)$$

- Trend parameters $\boldsymbol{\beta}$ enter only via $\boldsymbol{r}$ and the design/basis functions

- Kernel hyperparameters $\boldsymbol{\theta}$ and noise $\sigma^2$ enter only via $\mathbf{K}_y$ and its derivatives

- Consequence: updates are **decoupled in form**, though they interact through $\boldsymbol{r}$ and $\mathbf{K}_y^{-1}$