Introduction to Machine Learning

Gaussian Processes
Stochastic Processes and Distributions on Functions

$$f(x) \sim \mathcal{N}(\mathbf{m}, \mathbf{K})$$

Learning goals

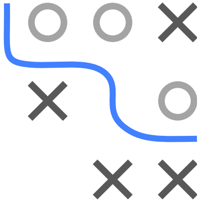
- GPs = distributions over functions
- Marginalization property
- Mean and covariance function

WEIGHT-SPACE VIEW

- Until now: hypothesis space $\mathcal{H}$ of parameterized functions $f(\mathbf{x} \mid \boldsymbol{\theta})$
- ERM: find risk-minimal parameters (weights) $\boldsymbol{\theta}$
- Bayesian paradigm: distribution over $\boldsymbol{\theta} \Rightarrow$ update prior to posterior belief after observing data according to Bayes' rule

$$p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}) = \frac{\text{likelihood} \cdot \text{prior}}{\text{marginal likelihood}} = \frac{p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) \cdot q(\boldsymbol{\theta})}{p(\mathbf{y}|\mathbf{X})}$$

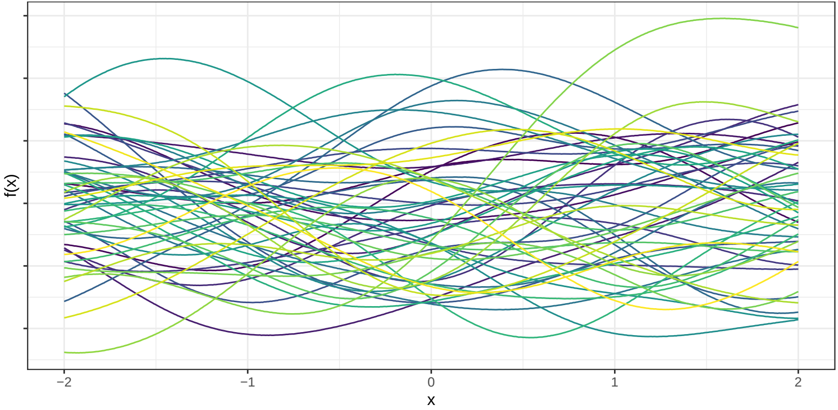
FUNCTION-SPACE VIEW



- New POV: rather than finding θ which parameterizes $f(\mathbf{x} \mid \theta)$, search in space of admissible functions directly
- Sticking to Bayesian inference, specify prior distribution **over functions** and update according to observed data points

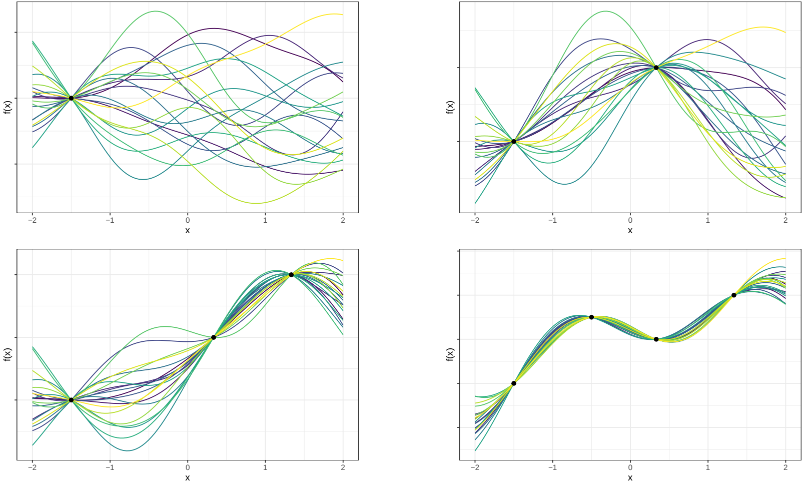
DRAWING FROM FUNCTION PRIORS

- Imagine we could draw functions from some prior distribution



DRAWING FROM FUNCTION PRIORS

- Restrict sampling to functions consistent with observed data



- Variety of admissible functions shrinks with seeing more data
- Intuitively: distributions over functions have “mean” & “variance”

WEIGHT-SPACE VS. FUNCTION-SPACE VIEW

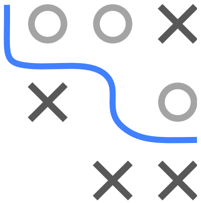
Weight-Space View

- Parameterize functions (e.g., $f(\mathbf{x} \mid \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathbf{x}$)
- Define distributions on $\boldsymbol{\theta}$
- Inference in param space Θ

Function-Space View

- Work on functions directly
- Define distributions on f
- Inference in fun space $\mathcal{H}$

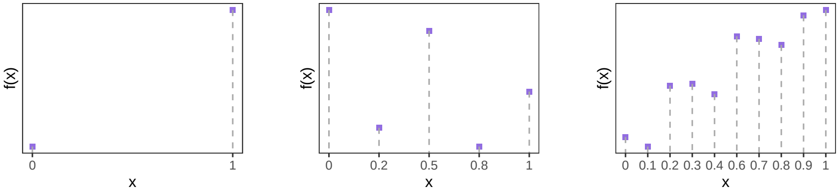
DISCRETE FUNCTIONS



- Let $\mathcal{X} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$, $\mathcal{H} = \{f \mid f : \mathcal{X} \to \mathbb{R}\}$
- Any $f \in \mathcal{H}$ has finite domain with $n < \infty$ elements
  $\Rightarrow$ neat representation with n-dim vector

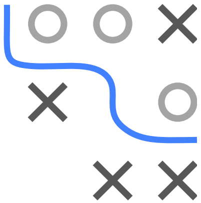
$$\mathbf{f} = [f(\mathbf{x}^{(1)}), \ldots, f(\mathbf{x}^{(n)})]^T$$

- Example functions living in this space for $|\mathcal{X}| \in \{2, 5, 10\}$



- NB: The $\mathbf{x}^{(i)}$ in the above are not really training points, we don't even consider training here. They are the points where we measure our (here: 1D) discrete functions. However, to avoid inventing too many symbols, and since the whole notation leads nicely into what follows next, we accept this "abuse" here.

DISTRIBUTIONS ON DISCRETE FUNCTIONS



- Specify density on vectors / functions with finite domain $f \in \mathcal{H}$
- Natural way: vector representation as n-dim RV, e.g.,

$$\mathbf{f} = [f(\mathbf{x}^{(1)}), \dots, f(\mathbf{x}^{(n)})]^T \sim \mathcal{N}(\mathbf{m}, \mathbf{K})$$

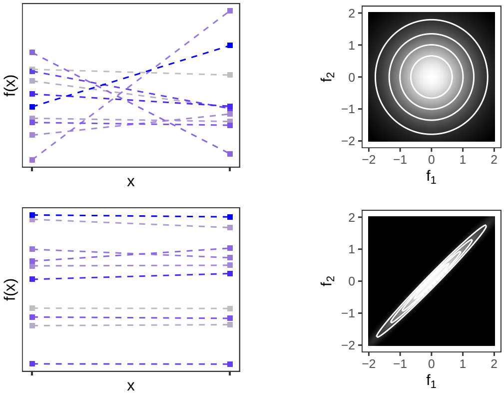
- For now: set $\mathbf{m} = \mathbf{0}$, assume $\mathbf{K}$ to be given

EXAMPLE: RANDOM DISCRETE FUNCTIONS

- Example ctd: **f** on 2 points
- Sample representatives by sampling from a 2-dim Gaussian

$$\mathbf{f} = [f(\mathbf{x}^{(1)}), f(\mathbf{x}^{(2)})]^T \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$$

- Where points are not (top) or strongly (bottom) correlated
- RHS shows 2D density

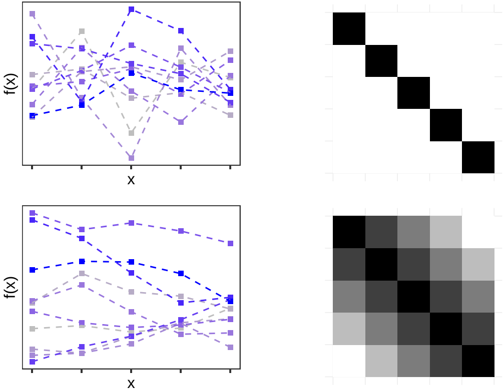


EXAMPLE: RANDOM DISCRETE FUNCTIONS

- Example ctd: **f** on 5 points
- Sample representatives by sampling from a 5-dim Gaussian

$$\mathbf{f} = [f(\mathbf{x}^{(1)}), \ldots, f(\mathbf{x}^{(5)})]^T \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$$

- Where points are not (top) or strongly (bottom) correlated
- RHS shows correlation matrix / structure

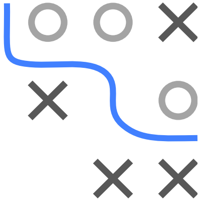


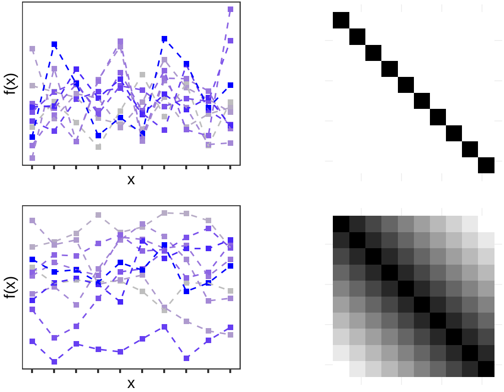
EXAMPLE: RANDOM DISCRETE FUNCTIONS

- Example ctd: **f** on 10 points
- Sample representatives by sampling from a 10-dim Gaussian

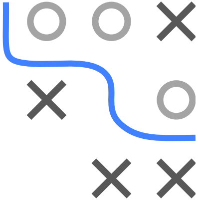
$$\mathbf{f} = [f(\mathbf{x}^{(1)}), \ldots, f(\mathbf{x}^{(10)})]^T \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$$

- Where points are not (top) or strongly (bottom) correlated
- RHS shows correlation matrix / structure





SPATIAL CORRELATION



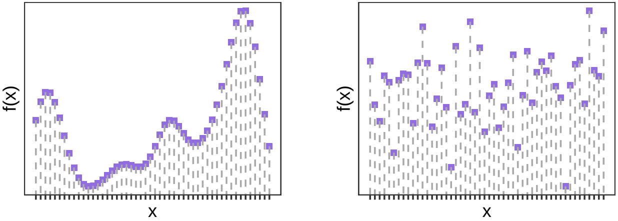
- "Meaningful" functions (on numeric $\mathcal{X}$) often have spatial property:

  $\mathbf{x}^{(i)}, \mathbf{x}^{(j)}$ close in $\mathcal{X} \Rightarrow f(\mathbf{x}^{(i)}), f(\mathbf{x}^{(j)})$ close / strongly correlated in $\mathcal{Y}$

- In other words: fun. values of nearby points should be correlated
- Enforce this by choosing dist.-based covariance function

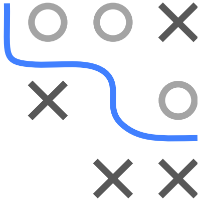
$$\mathbf{x}^{(i)}, \mathbf{x}^{(j)} \text{ close in } \mathcal{X} \Leftrightarrow \mathbf{K}_{ij} \text{ high}$$

- E.g., $\mathbf{K}_{ij} = k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \left(-\frac{1}{2}\|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|^2\right)$ vs identity cov.



- More on covariance function, or **kernel**, $k(\cdot, \cdot)$ later on

FROM DISCRETE TO CONTINUOUS FUNCTIONS



- So far: Multivar Gaussians to model outputs of discrete functions

$$\mathbf{f} = [f(\mathbf{x}^{(1)}), \dots, f(\mathbf{x}^{(n)})]^T \sim \mathcal{N}(\mathbf{m}, \mathbf{K})$$

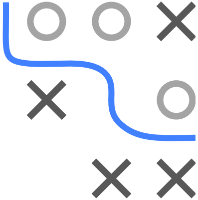
- Can we simply extend our distribution def to **continuous**-domain functions by taking $n \to \infty$?

- Unclear how to obtain “infinitely” long (Gaussian) random vectors

- Observation: random vectors $\mathbf{f}$ are collections of RVs enumerated by $\{1, \dots, n\} \Rightarrow$ **indexed family**

- Can we use more general, infinite index sets?

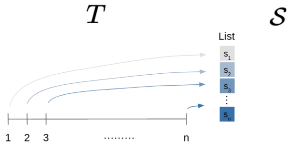
DEFINITION: INDEXED FAMILY

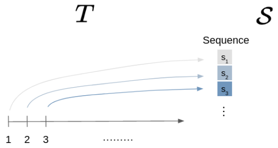


- Index T allows us to identify objects in arbitrary sets $\mathcal{S}$

$$s : T \to \mathcal{S}, \quad t \mapsto s_t = s(t)$$

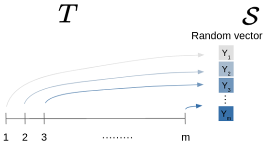
- This mapping is the formal definition of notation $\{s_t : t \in T\}$
- Example: real-valued $\mathcal{S}$
  - $\mathcal{S} = \mathbb{R}$, $t \mapsto s_t$
  - Finite index set, e.g., $T = \{1, \ldots, n\} \Rightarrow$ vector
  - Countable, infinite index set, e.g., $T = \mathbb{N} \Rightarrow$ sequence
  - Uncountable index set, e.g., $T = \mathbb{R} \Rightarrow$ function

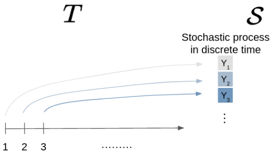




DEFINITION: STOCHASTIC PROCESS

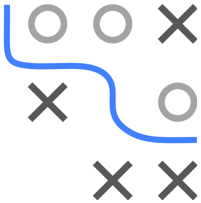
- Collection (potentially infinite) of RVs as indexed family $\{Y_t : t \in T\}$; further distributional assumptions give rise to important subclasses
- Intuition: probability distributions describe random vectors, SP describe random functions
- Examples
  - $\mathcal{S}$: space of RVs, $t \mapsto Y_t$
  - Finite index set, e.g., $T = \{1, \dots, m\}$ $\Rightarrow$ random vector
  - Countable, infinite index set, e.g., $T = \mathbb{N} \Rightarrow$ discrete-time SP
  - Uncountable index set, e.g., $T = \mathbb{R} \Rightarrow$ continuous-time SP





DEF.: GAUSSIAN PROCESS

▸ Rasmussen and Williams 2006   ▸ Snelson 2001



- Special kind of SP with index set $\mathcal{X}$; often $\mathcal{X} = \mathbb{R}^p$, but as in SVMs, feature vectors only enter the model via the kernel, so we can work on arbitrary spaces
- We write formally $f \sim \mathcal{GP}(m(\cdot), k(\cdot, \cdot))$
- Defining marginalization property: we have a GP iff for any finite set of inputs $\boldsymbol{X} \subset \mathcal{X}$,

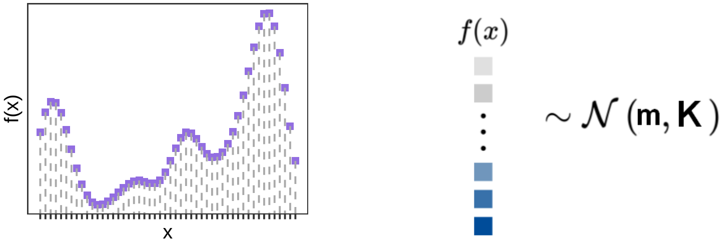
$$f(\boldsymbol{X}) \sim \mathcal{N}(m(\boldsymbol{X}), k(\boldsymbol{X}, \boldsymbol{X}))$$

- With **mean function** $m : \mathcal{X} \to \mathbb{R}$ and **cov function** $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_0^+$
- With slight abuse of notation, we allow matrix args and write:
  - $\mathbf{m} = m(\boldsymbol{X}) = [m(\mathbf{x}^{(1)}), \ldots, m(\mathbf{x}^{(n)})]^T$
  - $\mathbf{K} = k(\boldsymbol{X}, \boldsymbol{X}) = (k(\mathbf{x}, \tilde{\mathbf{x}}))_{\mathbf{x}, \tilde{\mathbf{x}} \in \boldsymbol{X}}$

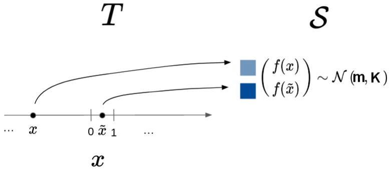
MARGINALIZATION PROPERTY

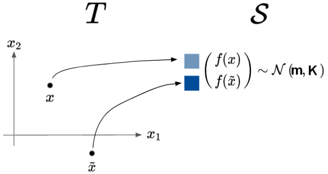
- For **any** finite set of inputs $\boldsymbol{X} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\} \subset \mathcal{X}$:

$$\mathbf{f} = f(\boldsymbol{X}) = [f(\mathbf{x}^{(1)}), \dots, f(\mathbf{x}^{(n)})]^T \sim \mathcal{N}(\mathbf{m}, \mathbf{K})$$



- Example with 1D (left) and 2D (right) index set $\mathcal{X}$: Dimension of $\mathbf{f}$ depends on n, not on dimension of $\mathcal{X}$:





GP EXISTENCE THEOREM

- For **any**
  - state space $\mathcal{X}$,
  - mean function $m : \mathcal{X} \to \mathbb{R}$,
  - covariance function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_0^+$,

there **exists** $f \sim \mathcal{GP}(m(\cdot), k(\cdot, \cdot))$ s.t. $\forall \mathbf{x}, \tilde{\mathbf{x}} \in \mathcal{X}$

$$\begin{aligned} \mathbb{E}(f(\mathbf{x})) &= m(\mathbf{x}) \\ \mathrm{Cov}(f(\mathbf{x}), f(\tilde{\mathbf{x}})) &= k(\mathbf{x}, \tilde{\mathbf{x}}) \end{aligned}$$

and $f(\boldsymbol{X}) \sim \mathcal{N}(m(\boldsymbol{X}), k(\boldsymbol{X}, \boldsymbol{X}))$ for any $\boldsymbol{X} \subset \mathcal{X}$

- Version of Kolmogorov consistency theorem
  $\Rightarrow$ proof ▸ Grimmett and Stirzaker 2001 (Thm. 8.6.3)

IMPLICATIONS OF EXISTENCE THEOREM

- GPs completely specified by their mean & cov function

$$
\begin{aligned}
m(\mathbf{x}) &= \mathbb{E}[f(\mathbf{x})] \\
k(\mathbf{x}, \tilde{\mathbf{x}}) &= \mathrm{Cov}(f(\mathbf{x}), f(\tilde{\mathbf{x}})) = \mathbb{E}[(f(\mathbf{x}) - \mathbb{E}[f(\mathbf{x})])(f(\tilde{\mathbf{x}}) - \mathbb{E}[f(\tilde{\mathbf{x}})])]
\end{aligned}
$$

- For now, we consider zero-mean GPs with $m(\mathbf{x}) \equiv \mathbf{0}$
  $\Rightarrow$ common, not necessarily drastic assumption

- Denote by $\mathcal{GP}(\mathbf{0}, k(\cdot, \cdot)) \Rightarrow$ properties mainly governed by $k(\cdot, \cdot)$

- By virtue of existence thm: sampling from GP priors gives us random functions with our properties of choice

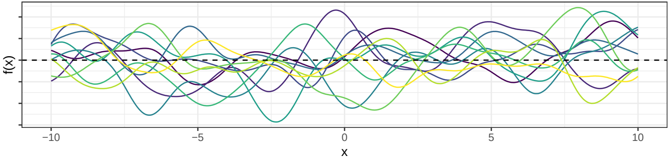
SAMPLING FROM GAUSSIAN PROCESS PRIORS

- Example: $f \sim \mathcal{GP}(\mathbf{0}, k(\cdot, \cdot))$ with cov function

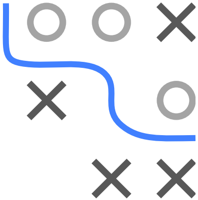
$$k(\mathbf{x}, \tilde{\mathbf{x}}) = \exp\left(-\frac{1}{2}\|\mathbf{x} - \tilde{\mathbf{x}}\|^2\right)$$

- To visualize sample functions,
  - choose high number n of points $\boldsymbol{X} = \left\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\right\}$
  - compute $\mathbf{K} = k(\boldsymbol{X}, \boldsymbol{X})$ from all pairs $\mathbf{x}^{(i)}, \mathbf{x}^{(j)} \in \boldsymbol{X}$
  - draw $\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$

- 10 randomly drawn functions (note 0 mean)



FURTHER READING



- Will go through many details now, but some general refs already
- The standard book: ▸ Rasmussen and Williams 2006
- Good videos can be found here: ▸ Monk 2011 ▸ Freitas 2020