Bayesian Theory and Computation

Lecture 17: Gaussian Processes

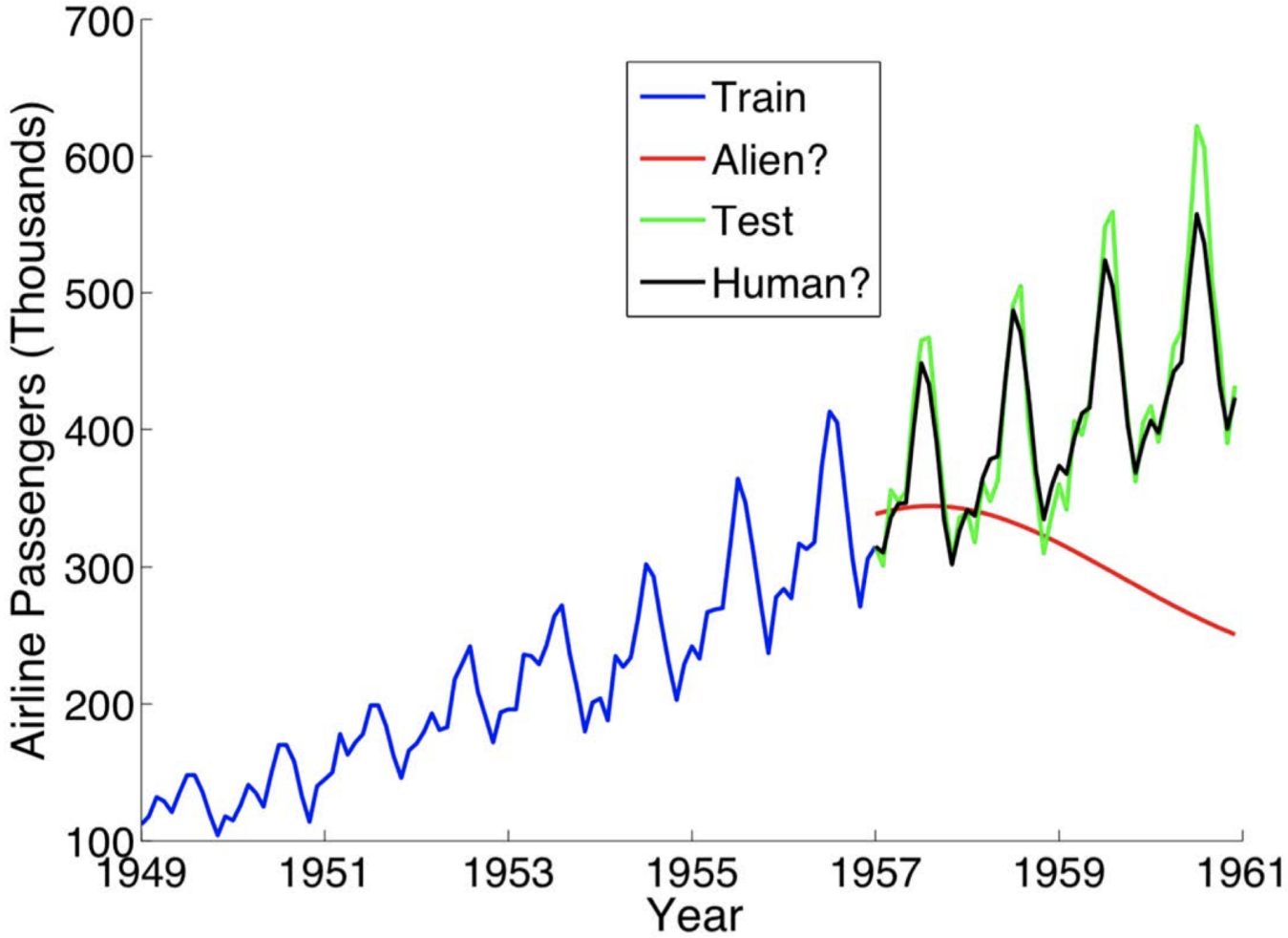
Cheng Zhang

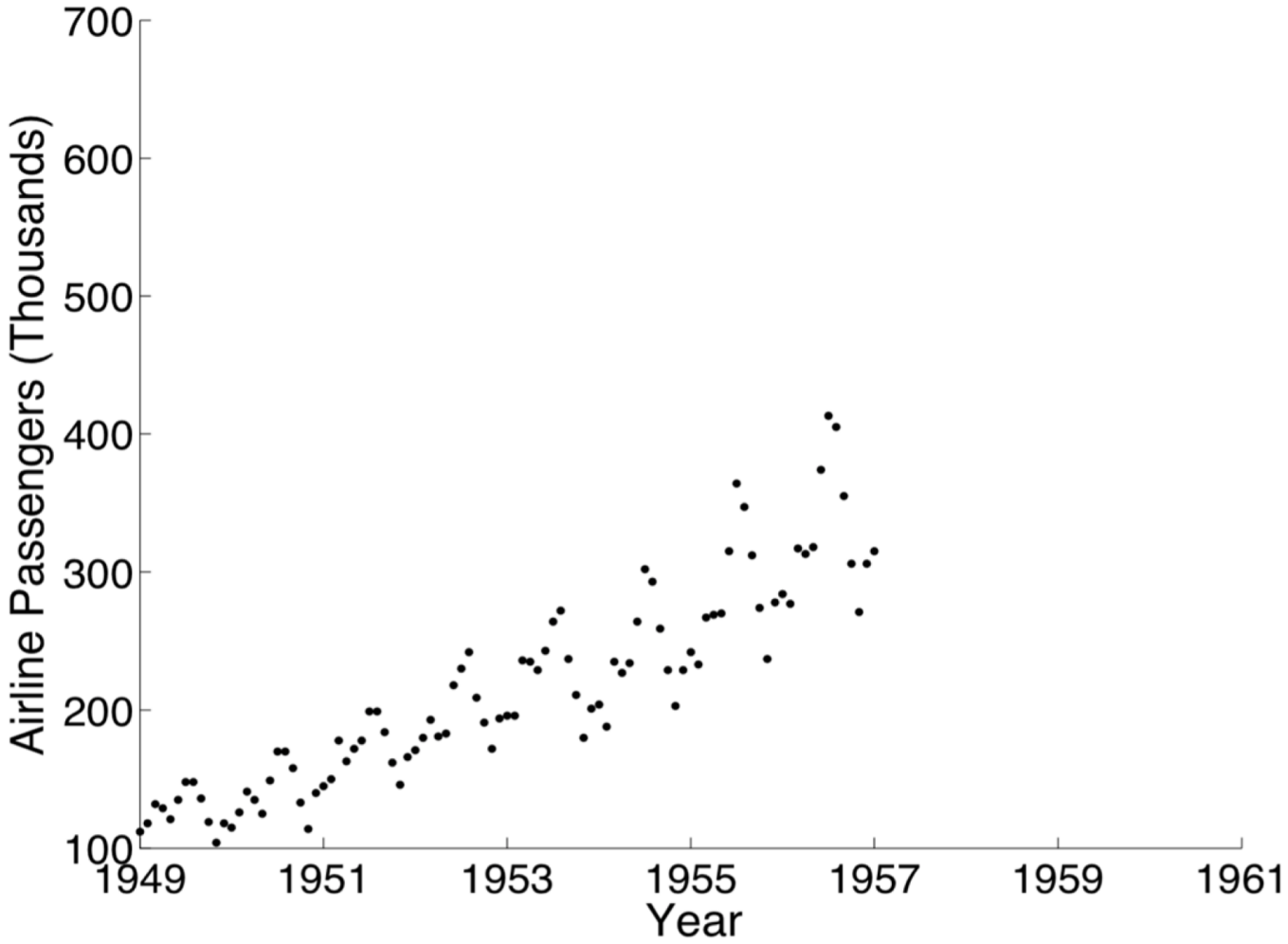
School of Mathematical Sciences, Peking University

May 18, 2022

- While parametric models can be powerful, choosing appropriate parametric models for certain data sets can be challenging.
- In the following lectures, we will discuss some Bayesian non-parametric models that are capable of dealing with data sets with extremely complicated structures.
- We start with Gaussian processes.

Function Learning Example




Airline Passengers (Thousands)
700
600
500
400
300
200
100
1949
1951
1953
1955
1957
1959
1961
Year

- Guess the parametric form of a function that could fit the data

$$f_w(x) = \begin{cases} w^T x, & \text{Linear model} \\ w^T \phi(x), & \text{Linear model with some basis functions} \\ g(w^T \phi(x)), & \text{Nonlinear model} \end{cases}$$

- For the real data, we could explicitly account for noise in our model

$$y(x) = f_w(x) + \epsilon(x)$$

- When taking $\epsilon(x) = \mathcal{N}(0, \sigma^2)$ for i.i.d. additive Gaussian noise, we have

$$p(y(x)|x, w, \sigma^2) = \mathcal{N}(y(x)|f_w(x), \sigma^2)$$

Therefore, we can find MLE for w and σ^2.

Parametric models

- Assume that all data can be represented using a fixed, finite number of parameters.
- Examples: Mixture of Gaussians, linear/polynomial regression, neural nets, etc.

Nonparametric models

- Number of parameters can grow with sample size.
- Number of parameters may be random.
- Examples: kernel density estimation.

Bayesian nonparameterics

- Allow for an infinite number of parameters a priori.
- Models of finite datasets will have only finite number of parameters.
- Other parameters are integrated out.

- A parametric likelihood: $x \sim p(\cdot|\theta)$
- Prior on θ: $\pi(\theta)$
- Posterior distribution

$$p(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{\int p(x|\theta)\pi(\theta)d\theta} \propto p(x|\theta)\pi(\theta)$$

Examples:

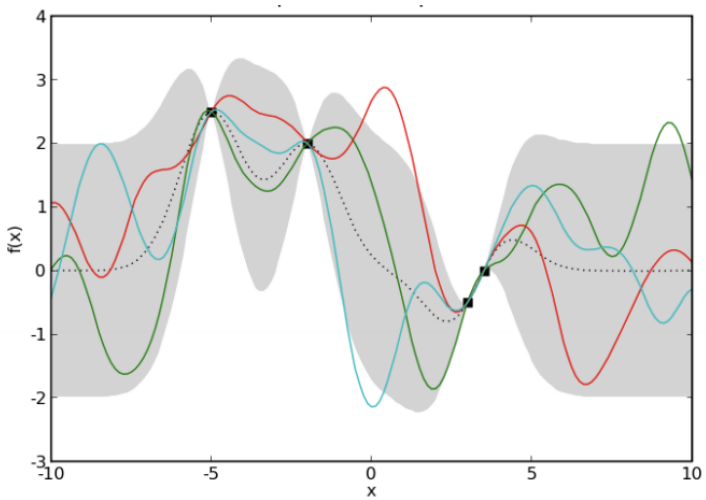
- Gaussian distribution prior + Gaussian likelihood $\rightarrow$ Gaussian posterior distribution
- Dirichlet distribution prior + Multinomial likelihood $\rightarrow$ Dirichlet posterior distribution
- Sparsity-inducing prior + some likelihood models $\rightarrow$ Sparse Bayesian inference

- A nonparametric likelihood: $x \sim p(\cdot|\mathcal{M})$
- Prior on $\mathcal{M}$: $\pi(\mathcal{M})$
- Posterior distribution

$$p(\mathcal{M}|x) = \frac{p(x|\mathcal{M})\pi(\mathcal{M})}{\int p(x|\mathcal{M})\pi(\mathcal{M})d\mathcal{M}} \propto p(x|\mathcal{M})\pi(\mathcal{M})$$

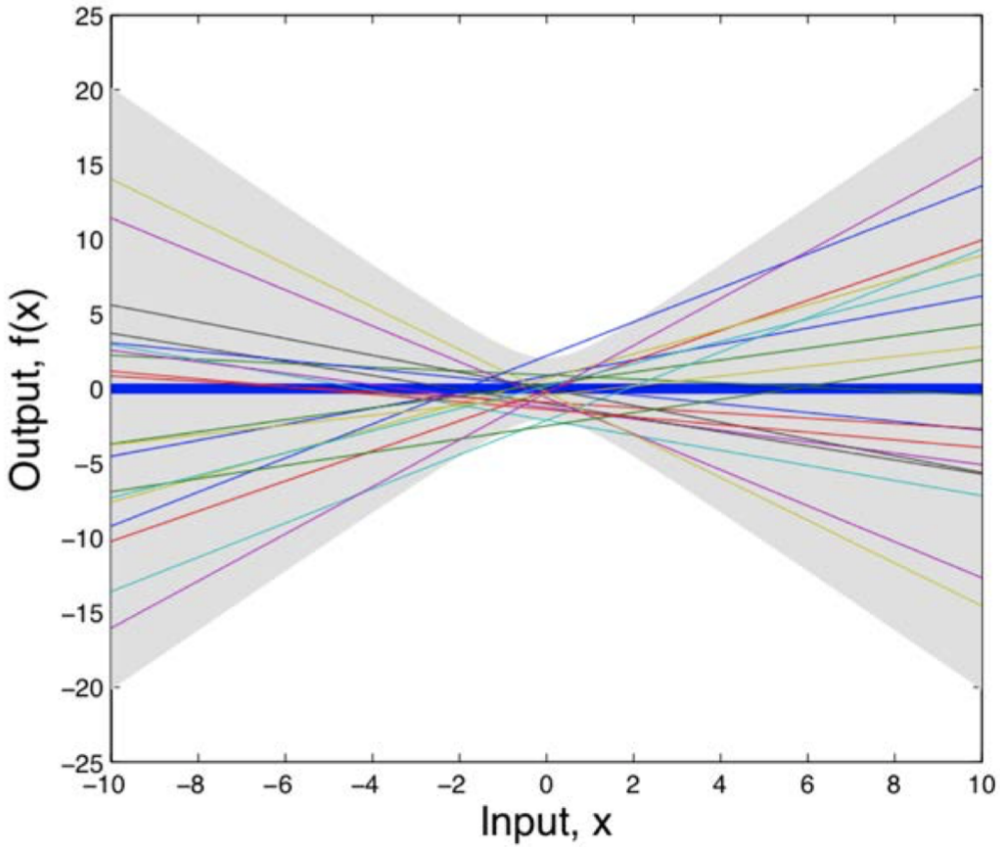
Examples:

- Gaussian Processes
- Dirichlet Processes
- Indian Buffet Processes



- Consider a simple linear model

$$f(x) = a_0 + a_1 x, \quad a_0, a_1 \sim \mathcal{N}(0, 1)$$



- We are interested in the distribution over functions induced by the distribution over parameters.
- In fact, we can characterize the properties of these functions directly.

$$f(x|a_0, a_1) = a_0 + a_1 x, \quad a_0, a_1 \sim \mathcal{N}(0, 1)$$

$$\mathbb{E}(f(x)) = \mathbb{E}(a_0) + \mathbb{E}(a_1)x = 0$$

$$\begin{aligned}\mathbb{C}\text{ov}(f(u), f(v)) &= \mathbb{E}(f(u)f(v)) - \mathbb{E}(f(u))\mathbb{E}(f(v)) \\ &= \mathbb{E}(a_0^2 + a_0 a_1(u+v) + a_1^2 uv) - 0 \\ &= 1 + uv.\end{aligned}$$

- Therefore, any collection of values has a joint Gaussian distribution. Note that the randomness comes from the function f, not from x.

$$f(x_1),\ f(x_2),\ \ldots,\ f(x_N) \sim \mathcal{N}(0, K)$$

$$K_{ij} = \mathbb{C}\text{ov}(f(x_i), f(x_j)) = k(x_i, x_j) = 1 + x_i x_j$$

- **Definition**: A **Gaussian process** (GP) is a collection of random variables, any finite number of which have a joint Gaussian distribution. We write $f(x) \sim \mathcal{GP}(m, k)$ to mean

$$f(x_1),\ f(x_2),\ \ldots,\ f(x_N) \sim \mathcal{N}(\mu, K)$$

$$\mu_i = m(x_i), \quad K_{ij} = k(x_i, x_j)$$

for any collection of input values $x_1, \ldots, x_N$. In other words, f is a GP with *mean function* $m(x)$ and *covariance kernel* $k(x_i, x_j)$.

Example: Linear Basis Function Models

- Consider a linear model with some basis function

$$f_w(x) = w^T \phi(x), \quad p(w) = \mathcal{N}(0, \Sigma_w)$$

- Moments of the induced distribution over functions

$$\mathbb{E}(f_w(x)) = m(x) = \mathbb{E}(w^T)\phi(x) = 0$$

$$\begin{aligned}
\mathbb{C}\text{ov}(f(x_i), f(x_j)) &= k(x_i, x_j) \\
&= \mathbb{E}(f_w(x_i) f_w(x_j)) - \mathbb{E}(f_w(x_i))\mathbb{E}(f_w(x_j)) \\
&= \phi(x_i)^T \mathbb{E}(ww^T)\phi(x_j) - 0 \\
&= \phi(x_i)^T \Sigma_w \phi(x_j)
\end{aligned}$$

- $f_w(x)$ is a Gaussian process, $f(x) \sim \mathcal{N}(m, k)$ with mean function $m(x) = 0$ and covariance kernel $k(x_i, x_j) = \phi(x_i)^T \Sigma_w \phi(x_j)$.

- Setting up the model this way, we are putting prior directly on the relationship between x_i and x_j as opposed to on some paremeters that represent this relationship (i.e., we cut out the middleman).
- This is specially useful when we are ultimately more interested in, and having strong intuition about, the functions that model our data and their correlations. We can express these intuitions using a mean function and a covariance kernel.
- Note that the prior here is implicit and reflects our choice of the functional form.
- In the above example, we are assuming the relationship is linear, In general, we could use other covariance functions to create nonlinear relationship.

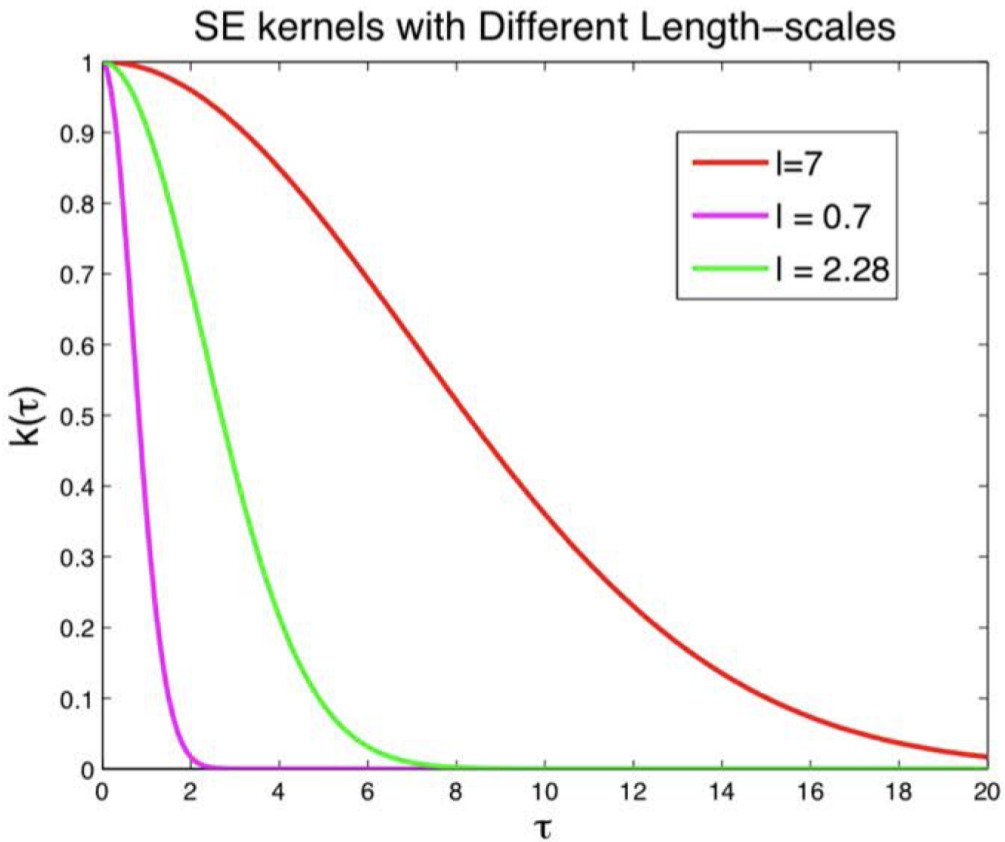
Example: Radial Basis Function Kernel

$$k_{\mathrm{RBF}}(x, x') = \mathbb{C}\mathrm{ov}(f(x), f(x')) = a^2 \exp\left(-\frac{\|x - x'\|^2}{2\ell^2}\right)$$

- One of the most popular kernels, also known as squared exponential kernel.
- Expresses the intuition that function values at nearby inputs are more correlated than function values at far away inputs.
- The kernel *hyparameters* a and ℓ control amplitudes and wiggliness of these functioins.
- GPs with an RBF kernel have large support and are *universal approximators*.

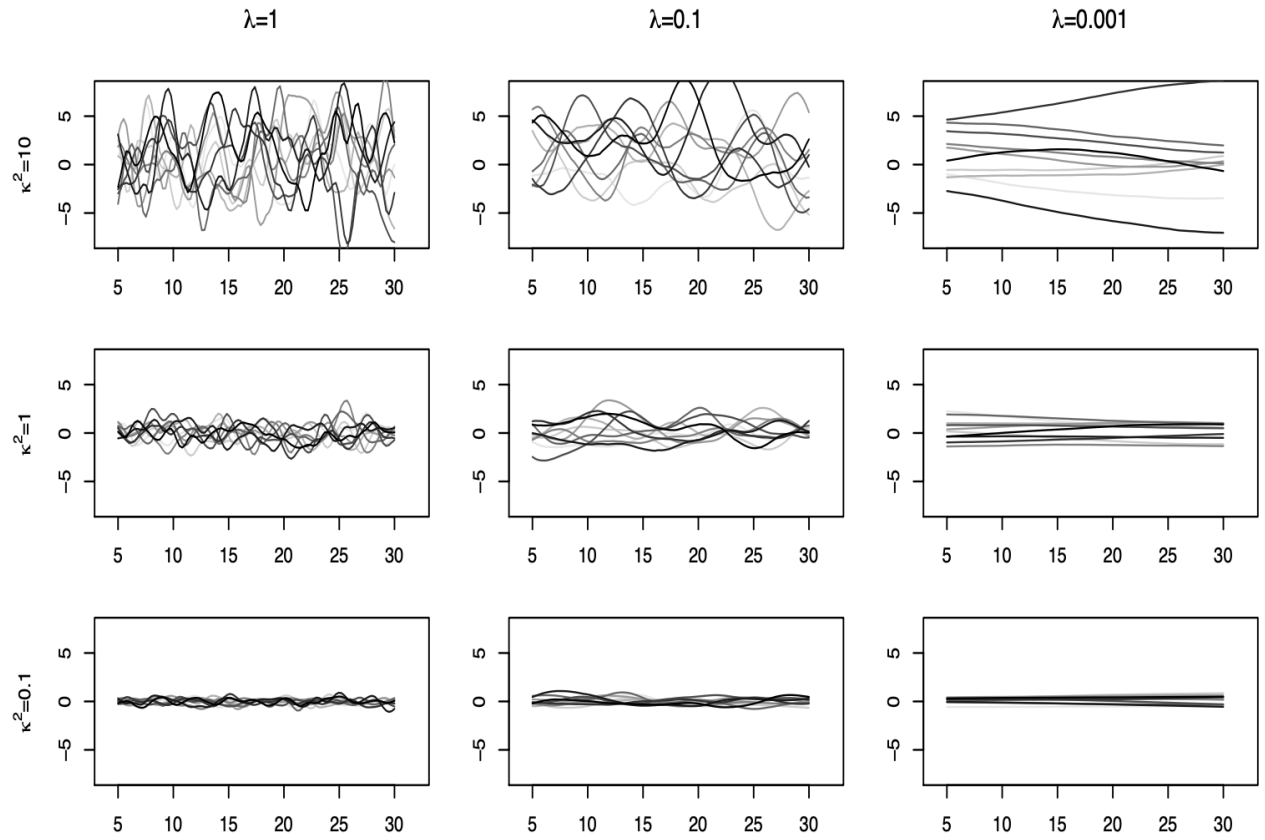
Example: Radial Basis Function Kernel

$$k_{\mathrm{RBF}}(x, x') = \mathbb{C}\mathrm{ov}(f(x), f(x')) = a^2 \exp\left(-\frac{\|x - x'\|^2}{2\ell^2}\right)$$



Example: Radial Basis Function Kernel

$$k_{\mathrm{RBF}}(x, x') = \mathbb{C}\mathrm{ov}(f(x), f(x')) = a^2 \exp\left(-\frac{\|x - x'\|^2}{2\ell^2}\right)$$



Here $\kappa^2 = a^2,\ \lambda = 1/2\ell^2$.

Other Covariance Functions

- Exponential

$$k(x_i, x_j) = \tau^2 \exp\left(-\frac{\|x_i - x_j\|}{2\ell}\right)$$

- Spherical

$$k(x_i, x_j) = \tau^2 \left(1 - \frac{3\|x_i - x_j\|}{2\theta} + \frac{\|x_i - x_j\|^3}{2\theta^3}\right) \mathbf{1}_{\|x_i - x_j\| \le \theta}$$

- Matérn

$$k(x_i, x_j) = \frac{\tau^2}{\Gamma(\nu)} \left(\frac{\|x_i - x_j\|}{2\phi}\right)^{\nu} B_\nu(\phi\|x_i - x_j\|)$$

where B_ν is the modified Bessel function.

- Linear

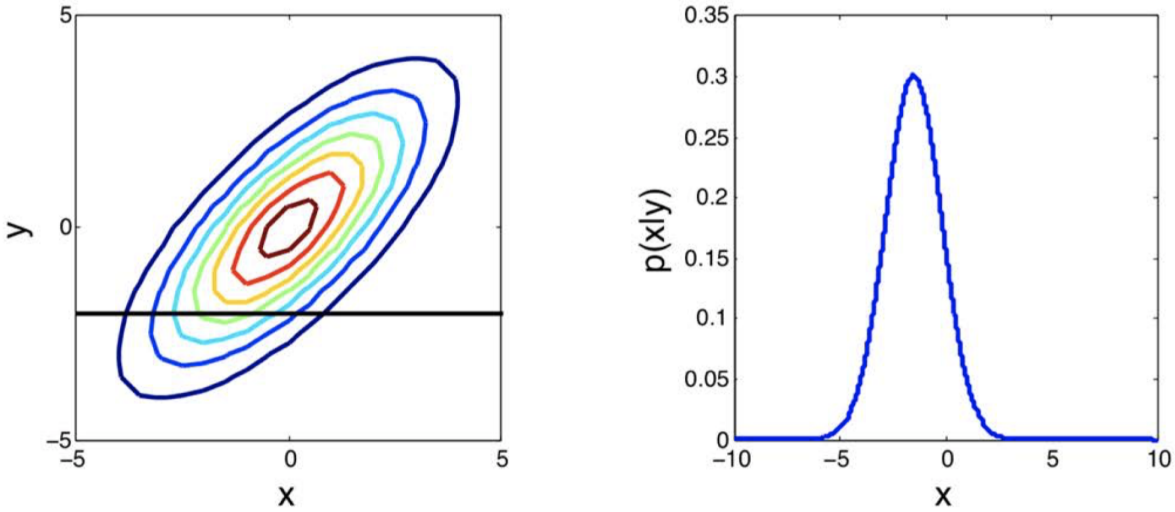
$$k(x_i, x_j) = \sigma^2 + \tau^2 (x_i - c)^T (x_j - c)$$

- Covariance functions must be positive semi-definite.
- Isotrophy/stationary. Covariance may only depends on distance
$$k(x_i, x_j) = c(\|x_i - x_j\|).$$
- Differentiability. Sample paths $f \sim \mathcal{GP}(0, k)$ may be m times differentiable. Can you find an example of non-differentiable Gaussian Process?
- Compact support. For any x_1, $\{x_2 : k(x_1, x_2) \neq 0\}$ is compact. This provides sparsity in covariance matrix. See the spherical covariance function for an example.
- Different convariance functions can be combined to form new convariance functions.

- Observed noisy data $y = (y(x_1), y(x_2), \ldots, y(x_N))^T$ at input locations X.
- Assume independent Gaussian noises and place a Gaussian process distribution over noise free functions $f(x) \sim \mathcal{GP}(0, k_\theta)$:

$$p(y|f) = \mathcal{N}(y|f, \sigma^2 I), \quad p(f|X) = \mathcal{N}(0, K_\theta(X, X))$$

- We want to infer $p(f_*|y, X, X_*)$ for the noise free function f evaluated at test points X_*.



- Suppose that

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$$

then

$$y_1|y_2 \sim \mathcal{N}(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}^T)$$

- When combined with test points X_*, the joint distribution of y and f_* is

$$\begin{pmatrix} y \\ f_* \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} K_\theta(X,X) + \sigma^2 I & K_\theta(X, X_*) \\ K_\theta(X_*, X) & K_\theta(X_*, X_*) \end{pmatrix}\right)$$

- Therefore, the conditional predictive distribution is

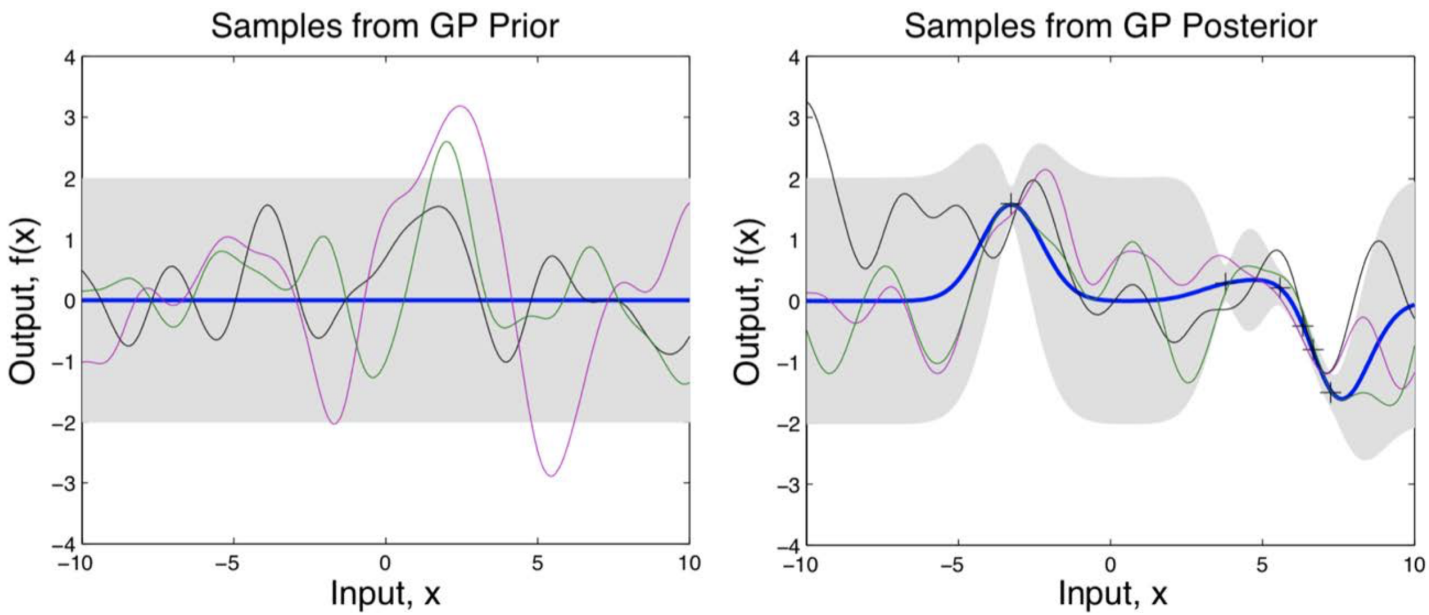
$$f_* | X_*, X, y, \theta \sim \mathcal{N}(\bar{f}_*, \mathbb{C}\mathrm{ov}(f_*))$$

where

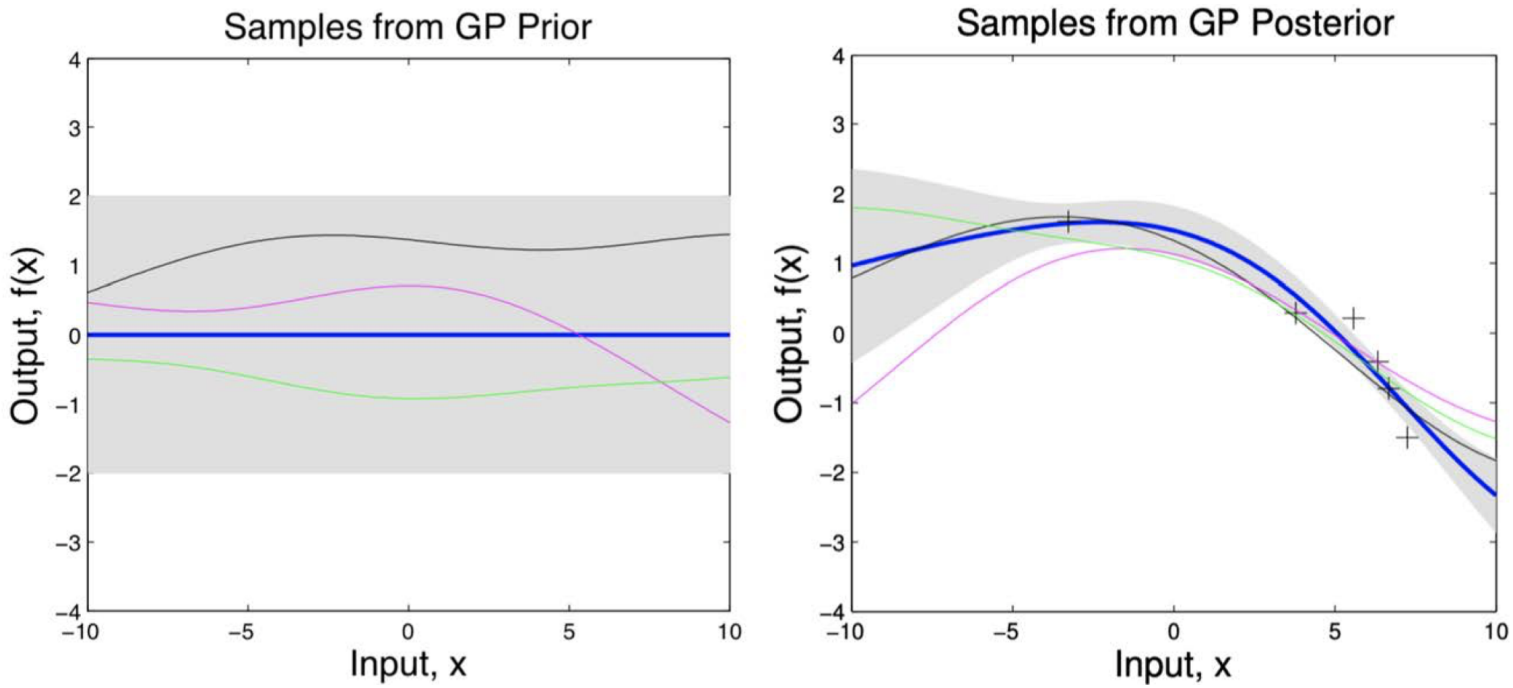
$$\bar{f}_* = K_\theta(X_*, X)(K_\theta(X,X) + \sigma^2 I)^{-1} y,$$

$$\mathbb{C}\mathrm{ov}(f_*) = K_\theta(X_*, X_*) - K_\theta(X_*, X)(K_\theta(X,X) + \sigma^2 I)^{-1} K_\theta(X, X_*).$$

- Specify $f(x) \sim \mathcal{GP}(0, k)$.
- Choose $k_{\text{RBF}}(x, x') = a_0^2 \exp\left(-\frac{\|x-x'\|^2}{2\ell_0^2}\right)$. Choose values for a_0 and ℓ_0.
- Observe data, look at the prior and posterior over functions.



What happened if we increase the length-scale ℓ?



- We can integrate away the entire Gaussian process $f(x)$ to obtain the marginal likelihood, as a function of kernel hyperparameters θ alone

$$p(y|\theta, X) = \int p(y|f, X)p(f|\theta, X)df = \mathcal{N}(y|0, K_\theta + \sigma^2 I)$$

- Maximum likelihood estimate (MLE)

$$\hat{\theta}_{\text{MLE}} = \arg\max_\theta \log p(y|\theta, X)$$

- Posterior Inference for $p(\theta|y, X)$

$$p(\theta|y, X) \propto p(y|\theta, X)p(\theta)$$

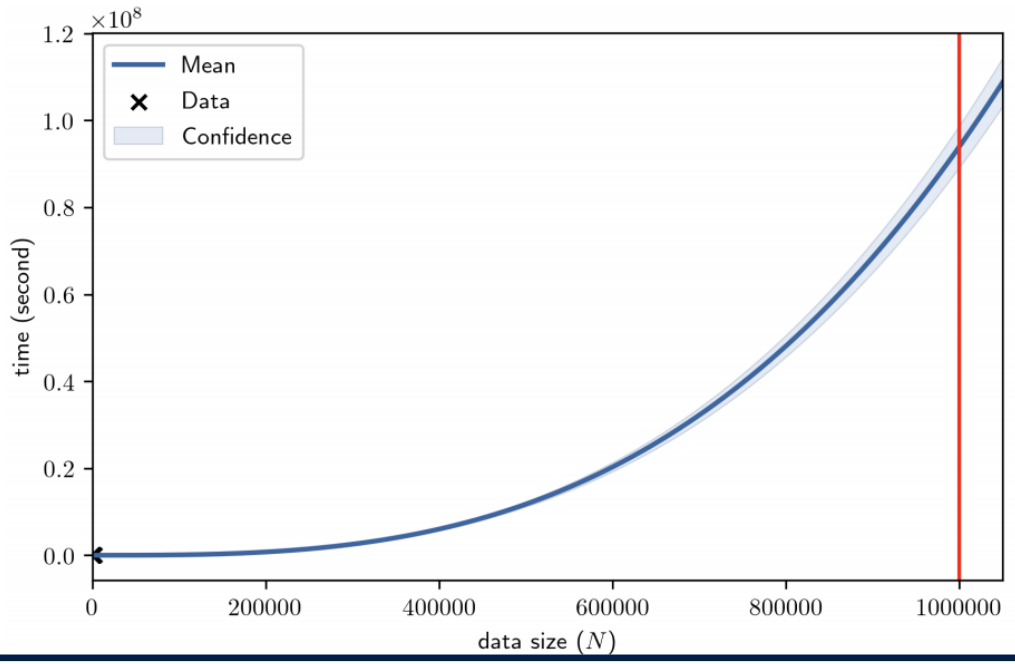
We could use VI/MCMC.

The Scalability Issue

- The log marginal likelihood is

$$\log p(y|\theta, X) = \log \mathcal{N}(y|0, K_\theta + \sigma^2 I)$$
$$= -\frac{1}{2} y^T (K_\theta + \sigma^2 I)^{-1} y - \frac{1}{2} \log |K_\theta + \sigma^2 I| - \frac{N}{2} \log(2\pi)$$

- The slowest components are the inversion $(K_\theta + \sigma^2 I)^{-1}$ and determinant $|K_\theta + \sigma^2 I|$, both are $\mathcal{O}(N^3)$.



Three Families of Approaches

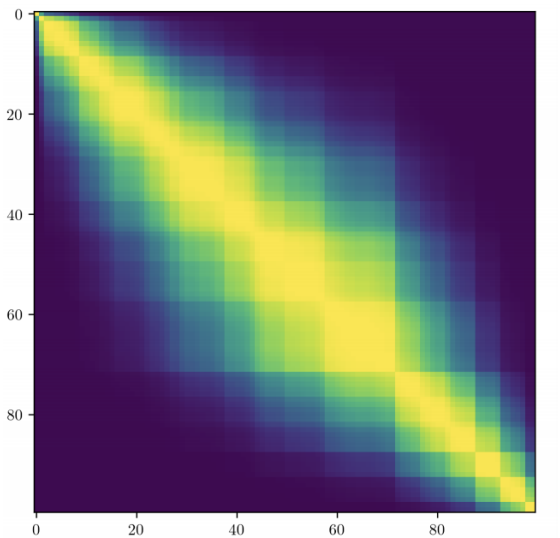
- Approximate non-parametric kernels in a finite basis 'dual space'. Requires $\mathcal{O}(m^2 n)$ computations and $\mathcal{O}(m)$ storage for m basis functions. Examples: SSGP, Random Kitchen Sinks, Fastfood.
- Inducing point based sparse approximations. Examples: SoR, FITC, KISS-GP.
- Exploit existing structure in K to quickly (and exactly) solve linear systems and log determinants. Examples: Toeplitz and Kronecker methods.

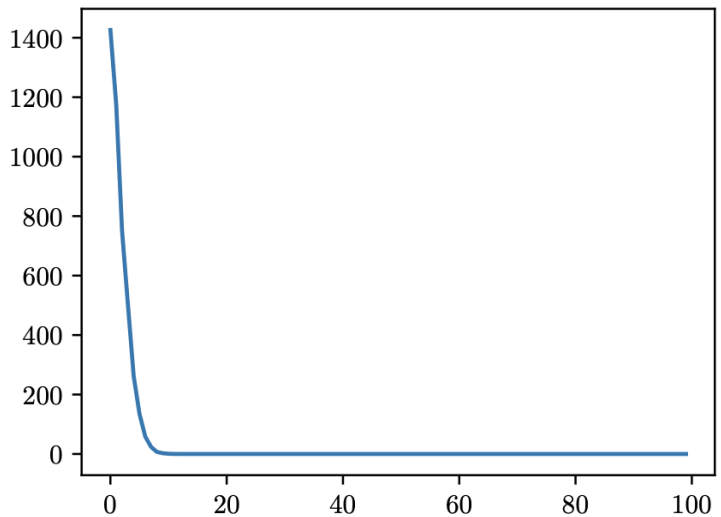
Low Rank Approximation

- Let's recall the log-likelihood of GP

$$\log p(y|X) = \log \mathcal{N}(y|0, K_\theta + \sigma^2 I)$$

- With redundant data, the covariance matrix K_θ becomes low rank.
- What about low rank approximation?





Approximation by Subset

- Let's randomly pick a subset from training data $\{z_1, z_2, \ldots, z_M\}$.
- Approximation the covariance matrix $K_\theta(X, X)$ by $\tilde{K}$

$$\tilde{K} = K_{xz} K_{zz}^{-1} K_{zx}$$

  where $K_{xz} = K_\theta(X, Z)$ and $K_{zz} = K_\theta(Z, Z)$.
- Note that $\tilde{K} \in \mathbb{R}^{N \times N}$, $K_{xz} \in \mathbb{R}^{N \times M}$ and $K_{zz} \in \mathbb{R}^{M \times M}$.
- The log-likelihood is approximated by

$$\log p(y|X, \theta) \approx \log \mathcal{N}(y|0, K_{xz} K_{zz}^{-1} K_{zx} + \sigma^2 I).$$

- The computation computational complexity reduces to $\mathcal{O}(NM^2)$ via the Woodbury formula

$$(K_{xz}K_{zz}^{-1}K_{zx}+\sigma^2 I)^{-1} = \sigma^{-2}I-\sigma^{-4}K_{xz}(K_{zz}+\sigma^{-2}K_{zx}K_{xz})^{-1}K_{zx}$$

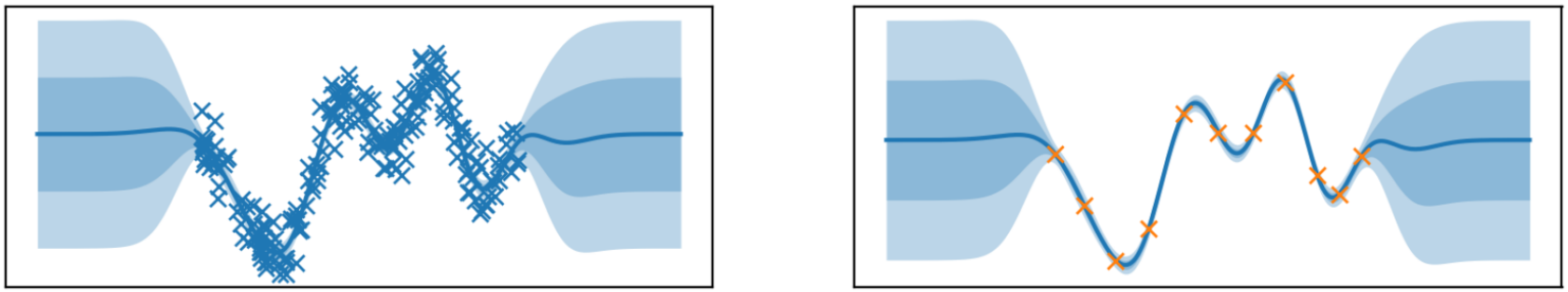
- The above approach is called Nyström approximation by Williams and Seeger (2001).
- Note that the approximation is directly done on the covariance matrix without the concept of inducing points and becomes exact if the whole data set is taken

$$K_{xx}K_{xx}^{-1}K_{xx} = K_{xx}$$

- The subset selection is done randomly.

Inducing Point Methods: Motivation

An example of a posterior obtained from many noisy observations (left) and from very frew noiseless observations (right).



- The GP prior places strong constraints on what values neighouring output can take, making it possible to obtain good approximations from only a few observations at appropriate positions, i.e., inducing points.

- We can approximate GP through $M < N$ inducing points X_u to obtain low rank approximations to the joint prior

$$p(f, f_*) = \int p(f, f_*, f_u) df_u = \int p(f, f_*|f_u) p(f_u) df_u$$
$$\approx \int q(f|f_u) q(f_*|f_u) p(f_u) df_u$$

  where $p(f_u) = \mathcal{N}(0, K_{uu}),\ K_{uu} = K_\theta(X_u, X_u)$.
- Now that f and f_* are conditionally independent given f_u, they can only communicate through f_u, and f_u therefore induces the dependencies between f and f_*.
- Different inducing point methods correspond to different additional assumptions about the two inducing conditionals $q(f|f_u), q(f_*|f_u)$.

- The Subset of Regressor (SoR) algorithm was given by Silverman (1985), and later on adapted by Smola and Bartlett (2001) for sparse Gaussian process regression.
- In SoR, f and f_* are assumed to deterministically depend on f_u

$$f = K_{xu}K_{uu}^{-1}f_u,\ f_* = K_{*u}K_{uu}^{-1}f_u,\ f_u \sim \mathcal{N}(0, K_{uu}).$$

- The effective prior implied by the SoR approximation is

$$q_{\mathrm{SoR}}(f, f_*) = \mathcal{N}\left(0, \begin{pmatrix} Q_{xx} & Q_{x*} \\ Q_{*x} & Q_{**} \end{pmatrix}\right)$$

where $Q_{ab} = K_{au}K_{uu}^{-1}K_{ub}$.

- Note that the exact conditionals are

$$p(f|f_u) = \mathcal{N}(K_{xu}K_{uu}^{-1}f_u, K_{xx} - Q_{xx})$$
$$p(f_*|f_u) = \mathcal{N}(K_{*u}K_{uu}^{-1}f_u, K_{**} - Q_{**})$$

- The approximate conditionals in SoR can be viewed as

$$q_{\text{SoR}}(f|f_u) = \mathcal{N}(K_{xu}K_{uu}^{-1}f_u, 0), \quad q_{\text{SoR}}(f_*|f_u) = \mathcal{N}(K_{*u}K_{uu}^{-1}f_u, 0)$$

- **Remark**: The SoR approximation is equivalent to exact inference in the degenerated Gaussian process with covariance function

$$k_{\text{SoR}}(x_i, x_j) = k(x_i, u)K_{uu}^{-1}k(u, x_i).$$

Fully Independent Training Conditional (FITC)

- Snelson and Ghahramani (2006) proposed the Sparse Pseudo-input Gaussian process (SPGP), later referred to as Fully independent training conditional (FITC).
- Augment the training data (X, y) with pseudo data f_u at location X_u.

$$p\left(\begin{pmatrix} y \\ f_u \end{pmatrix}\right) = \mathcal{N}\left(0, \begin{pmatrix} K_{xx} + \sigma^2 I & K_{xu} \\ K_{ux} & K_{uu} \end{pmatrix}\right)$$

- We can rewrite the joint probability as

$$p(y, f_u | X, X_u) = p(y | f_u, X, X_u) p(f_u | X_u)$$

where $p(f_u|X_u) = \mathcal{N}(0, K_{uu})$ and

$$p(y|f_u, X, X_u) = \mathcal{N}(K_{xu}K_{uu}^{-1}f_u, K_{xx} - K_{xu}K_{uu}^{-1}K_{ux} + \sigma^2 I).$$

Fully Independent Training Conditional (FITC)

- So far, no approximation has been made. $p(y|X)$ would still be expensive to evaluate.
- The FITC approximation assumes

$$q(y|f_u, X, X_u) = \mathcal{N}(K_{xu}K_{uu}^{-1}f_u, \Lambda + \sigma^2 I),$$

  where $\Lambda = \text{diag}\{K_{xx} - K_{xu}K_{uu}^{-1}K_{ux}\}$.
- Note that this implies the fully conditional independence of training data $\{y_i\}_{i=1}^N$ given f_u

$$q(y|f_u, X, X_u) = \prod_{i=1}^{N} p(y_i|f_u, X, X_u)$$

- Integrate out f_u we can get an approximate marginal likelihood

$$\tilde{p}(y|X, X_u) = \mathcal{N}(0, K_{xu}K_{uu}^{-1}K_{ux} + \Lambda + \sigma^2 I)$$

- For the predictive distribution of f_* at x_*, we can integrate with the posterior

$$q(f_*|x_*, y, X, X_u) = \int p(f_*|x_*, f_u, X_u)q(f_u|y, X, X_u)df_u$$

- The approximate conditionals in FITC is

$$q_{\text{FITC}}(f|f_u) = \prod_{i=1}^{N} p(f_i|f_u) = \mathcal{N}(K_{xu}K_{uu}^{-1}f_u, \text{diag}(K_{xx} - Q_{xx}))$$

$$q_{\text{FITC}}(f_*|f_u) = p(f_*|f_u)$$

- The effective prior implied by the FITC approximation is

$$q_{\mathrm{FITC}}(f, f_*) = \mathcal{N}\left(0, \begin{pmatrix} Q_{xx} - \mathrm{diag}(Q_{xx} - K_{xx}) & Q_{x*} \\ Q_{*x} & K_{**} \end{pmatrix}\right)$$

- **Remark**: If the assumption of fully independence is extended to the test conditional, the FITC approximation is equivalent to exact inference in a non-degenerated Gaussian process with covariance function

$$k_{\mathrm{FIC}}(x_i, x_j) = k_{\mathrm{SoR}}(x_i, x_j) + \delta_{i,j}(k(x_i, x_j) - k_{\mathrm{SoR}}(x_i, x_j)).$$

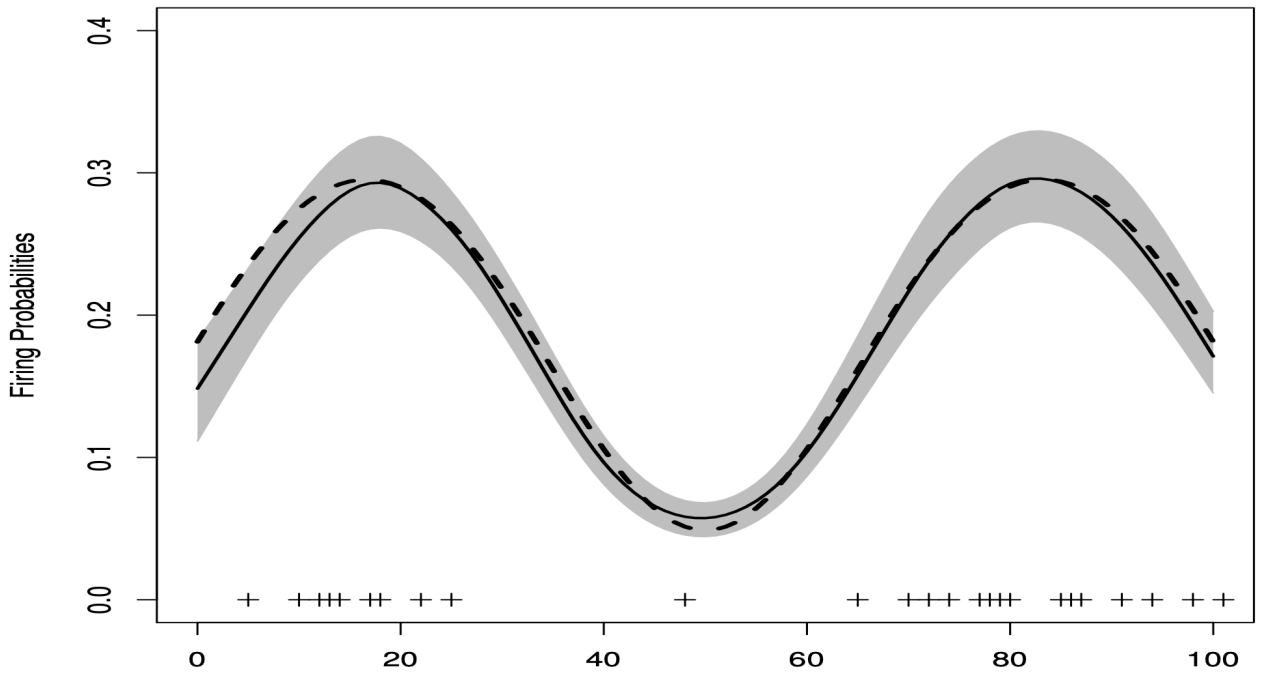
- The inducing points X_u can be optimized via gradient optimization.

- In case of non-Gaussian likelihoods, we can use the Gaussian process prior over a continuous latent function u which determines the likelihood $p(y|u, \phi)$ through a link function, just as in generalized linear models.
- For example, if the outcome variable y is binary, we can use the following logistic model

$$p(y_i = 1|u(x_i)) = \frac{1}{1 + \exp(-u(x_i))}$$

- We can use a multinomial logit model for outcome variables with multiple categories.

- Here, we are using a GP model to estimate the underlying firing rates of a neuron (i.e., $y_t = 1$ when the neuron fires, $y_t = 0$ otherwise).
- The dashed line shows the true firing probability and the plus signs show the firing time.



When $p(y|u, X)$ is not Gaussian, we lack closed-form expression for posterior and need approximation approaches.

$$p(u|y, X) \propto p(y|u, X)p(u|X)$$

- MCMC. Use GP-specific samplers (e.g., elliptic slice sampler). Due to the dependency between the latent values and the hyper-parameters, mixing can be slow.
- Laplace approximation.
  - Find posterior mode $\hat{u}$ using any gradient-based optimizer.
  - Use normal approximation to posterior

$$p(u|y, X) \approx \mathcal{N}(\hat{u}, \hat{\Sigma})$$

- When the likelihood contribution is heavily skewed (e.g., logistic model), expectation propagation (EP)/variational inference (VI) can be used.

Summary

- Gaussian processes (GPs) are Bayesian nonparametric models that can represent distributions over smooth functions.
- Using expressive covariance kernel functions, GPs can model a variety of data (scalar, vector, sequential, structured, etc.).
- Inference can be done fully analytically (in case of Gaussian likelihood).
- Inference and learning are very computationally costly since exact methods require large matrix inversions.
- There has been a variety of approximation methods to scale up GPs to large data sets.

References

- Bernhard W. Silverman. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. J. Roy. Stat. Soc. B, 47(1):1–52, 1985. (with discussion).
- Alexander J. Smola and Peter L. Bartlett. Sparse greedy Gaussian process regression. In Advances in Neural Information Processing Systems 13, pages 619–625, Cambridge, Massachussetts, 2001. The MIT Press.
- Ali Rahimi and Benjamin Recht. Random Features for Large-Scale Kernel Machines. In Advances in Neural Information Processing Systems, 2007.
- Ali Rahimi and Benjamin Recht. Weighted Sums of Random Kitchen Sinks: Replacing minimization with randomization in Learning. In Advances in Neural Information Processing Systems, 2008.

References

- Carl Edward Rasmussen and Christopher K. I. Williams. Gaussian Processes for Machine Learning. The MIT press, 2006.
- Williams, C. K. I. and Seeger, M. (2001). Using the nyström method to speed up kernel machines. In Advances in Neural Information Processing Systems, pages 682–688.
- Quinonero-Candela, J. and Rasmussen, C. E. (2005). A unifying view of sparse approximate gaussian process regression. Journal of Machine Learning Research, 6:1939—1959.
- Snelson, E. and Ghahramani, Z. (2006). Sparse gaussian processes using pseudo-inputs. In Advances in Neural Information Processing Systems, pages 1257–1264.