Gaussian Processes

Maurizio Filippone

EURECOM, Sophia Antipolis, France

February, 9th, 2017

- Research Center in the French Riviera

Suggested readings

Gaussian Processes for Machine Learning

Carl E. Rasmussen and Christopher K. I. Williams

Pattern Recognition and Machine Learning

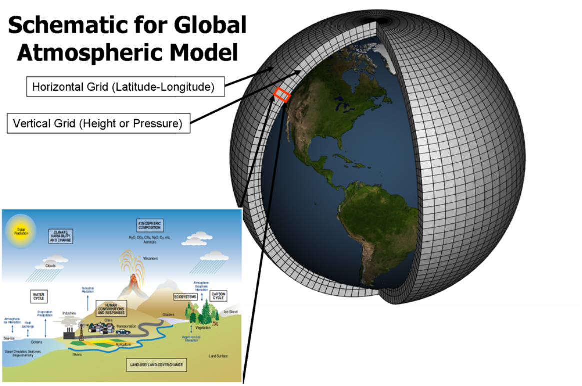
C. Bishop

Outline

- Motivation – Examples
- Introduction to Gaussian Processes
  - Weight space view
  - Function space view
- Challenges
- Modern Gaussian Processes

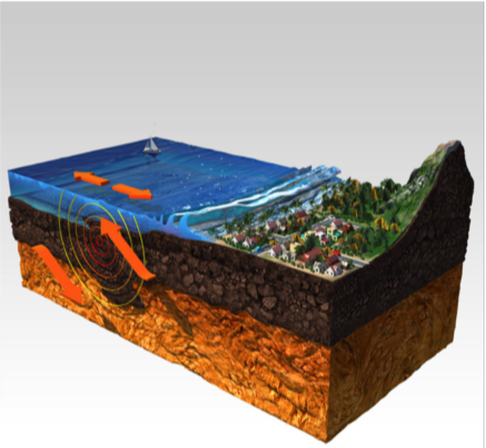
Motivation

- Climate modeling



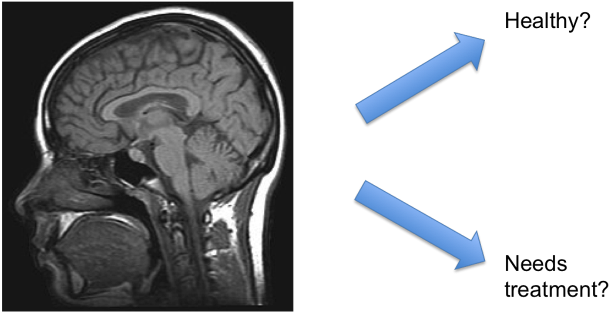
Kennedy and O'Hagan, *RSS-B*, 2001

- Earthquake modeling



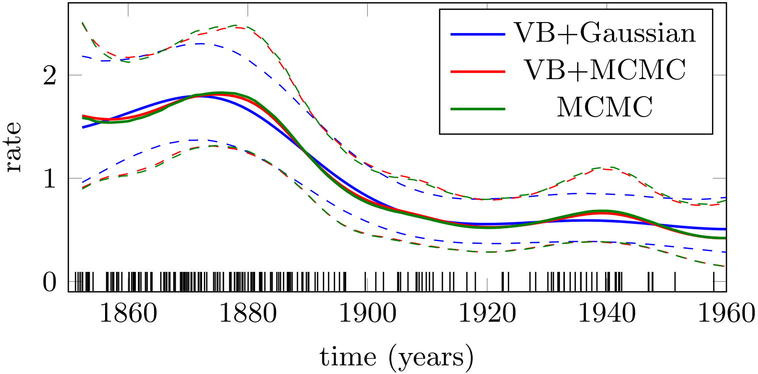
Kennedy and O'Hagan, *RSS-B*, 2001

- Classification of neurodegenerative diseases



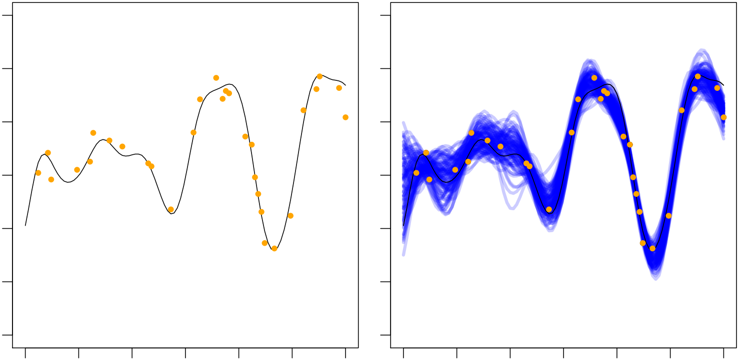
Filippone et al., *AoAS*, 2012

- Coal mining disaster data



Hensman, Matthews, Filippone, Ghahramani, *NIPS*, 2015

- Regression example



A Unified Framework

A model might be expensive to simulate/inaccurate

- Emulate model/discrepancy using a surrogate

A Unified Framework

A model might be expensive to simulate/inaccurate

- Emulate model/discrepancy using a surrogate

A model might not even be available

- Replace it with a flexible model

A Unified Framework

A model might be expensive to simulate/inaccurate

- Emulate model/discrepancy using a surrogate

A model might not even be available

- Replace it with a flexible model

Gaussian processes for Accurate Quantification of Uncertainty

Gaussian Processes

Gaussian Processes can be explained in two ways

- Weight Space View
    - Bayesian linear regression with infinite basis functions
- Function Space View
    - Defined as priors over functions

Gaussian Processes can be explained in two ways

- **Weight Space View**
  - **Bayesian linear regression with infinite basis functions**
- Function Space View
  - Defined as priors over functions

Bayesian Linear Regression

- Modeling observations as noisy realizations of a linear combination of the features:

$$p(\mathbf{y}|\mathbf{w}, \mathbf{X}, \sigma^2) = \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2 \mathbf{I})$$

Bayesian Linear Regression

- Modeling observations as noisy realizations of a linear combination of the features:

$$p(\mathbf{y}|\mathbf{w}, \mathbf{X}, \sigma^2) = \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I})$$

- Gaussian prior over model parameters:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \mathbf{S})$$

Bayesian Linear Regression

- Bayes rule:

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})}{p(\mathbf{y}|\mathbf{X})}$$

- Bayes rule:

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})}{p(\mathbf{y}|\mathbf{X})}$$

- **Posterior density**: $p(\mathbf{w}|\mathbf{X}, \mathbf{y})$
  - Distribution over parameters *after* observing data

- Bayes rule:

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})}{p(\mathbf{y}|\mathbf{X})}$$

- **Posterior density**: $p(\mathbf{w}|\mathbf{X}, \mathbf{y})$
  - Distribution over parameters *after* observing data
- **Likelihood** : $p(\mathbf{y}|\mathbf{X}, \mathbf{w})$
  - Measure of "fitness"

- Bayes rule:

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})}{p(\mathbf{y}|\mathbf{X})}$$

- **Posterior density**: $p(\mathbf{w}|\mathbf{X}, \mathbf{y})$
  - Distribution over parameters *after* observing data
- **Likelihood** : $p(\mathbf{y}|\mathbf{X}, \mathbf{w})$
  - Measure of “fitness”
- **Prior density**: $p(\mathbf{w})$
  - Anything we know about parameters *before* we see any data

Bayesian Linear Regression

- Bayes rule:

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})}{p(\mathbf{y}|\mathbf{X})}$$

- **Posterior density**: $p(\mathbf{w}|\mathbf{X}, \mathbf{y})$
    - Distribution over parameters *after* observing data
- **Likelihood** : $p(\mathbf{y}|\mathbf{X}, \mathbf{w})$
    - Measure of "fitness"
- **Prior density**: $p(\mathbf{w})$
    - Anything we know about parameters *before* we see any data
- **Marginal likelihood**: $p(\mathbf{y}|\mathbf{X})$
    - It is a normalization constant – ensures $\int p(\mathbf{w}|\mathbf{X}, \mathbf{y})\, d\mathbf{w} = 1$.

- Ignoring normalizing constants, the posterior is:

$$
\begin{aligned}
p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \sigma^2) &\propto \exp\left\{-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{w}-\boldsymbol{\mu})\right\} \\
&= \exp\left\{-\frac{1}{2}(\mathbf{w}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{w} - 2\mathbf{w}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right\} \\
&\propto \exp\left\{-\frac{1}{2}(\mathbf{w}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{w} - 2\mathbf{w}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right\}
\end{aligned}
$$

- Ignoring non-**w** terms, the prior multiplied by the likelihood is:

$$
\begin{aligned}
& p(\mathbf{y}|\mathbf{w}, \mathbf{X}, \sigma^2) \\
\propto \quad & \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\mathbf{w})^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\mathbf{w})\right\} \exp\left\{-\frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{S}^{-1}\mathbf{w}\right\} \\
\propto \quad & \exp\left\{-\frac{1}{2}\left(\mathbf{w}^{\mathsf{T}}\left[\frac{1}{\sigma^2}\mathbf{X}^{\mathsf{T}}\mathbf{X} + \mathbf{S}^{-1}\right]\mathbf{w} - \frac{2}{\sigma^2}\mathbf{w}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{y}\right)\right\}
\end{aligned}
$$

- Posterior (from previous slide):

$$
\propto \exp\left\{-\frac{1}{2}(\mathbf{w}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{w} - 2\mathbf{w}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right\}
$$

- Equate individual terms on each side.
- Covariance:

$$
\begin{aligned}
\mathbf{w}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{w} &= \mathbf{w}^{\mathsf{T}}\left[\frac{1}{\sigma^2}\mathbf{X}^{\mathsf{T}}\mathbf{X} + \mathbf{S}^{-1}\right]\mathbf{w} \\
\boldsymbol{\Sigma} &= \left(\frac{1}{\sigma^2}\mathbf{X}^{\mathsf{T}}\mathbf{X} + \mathbf{S}^{-1}\right)^{-1}
\end{aligned}
$$

- Mean:

$$
\begin{aligned}
2\mathbf{w}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} &= \frac{2}{\sigma^2}\mathbf{w}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{y} \\
\boldsymbol{\mu} &= \frac{1}{\sigma^2}\boldsymbol{\Sigma}\mathbf{X}^{\mathsf{T}}\mathbf{y}
\end{aligned}
$$

- Posterior **must be** Gaussian

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \sigma^2) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- Covariance:

$$\boldsymbol{\Sigma} = \left(\frac{1}{\sigma^2} \mathbf{X}^{\mathsf{T}} \mathbf{X} + \mathbf{S}^{-1} \right)^{-1}$$

- Mean:

$$\boldsymbol{\mu} = \frac{1}{\sigma^2} \boldsymbol{\Sigma} \mathbf{X}^{\mathsf{T}} \mathbf{y}$$

- Predictions – same tedious exercise as before:

$$p(y_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*, \sigma^2) = \mathcal{N}(\mathbf{x}_*^{\mathsf{T}} \boldsymbol{\mu}, \sigma^2 + \mathbf{x}_*^{\mathsf{T}} \boldsymbol{\Sigma} \mathbf{x}_*)$$

Introducing basis functions

- Imagine transforming the inputs using a set of D functions

$$\mathbf{x} \to \boldsymbol{\phi}(\mathbf{x}) = (\phi_1(\mathbf{x}), \ldots, \phi_D(\mathbf{x}))^\top$$

- The functions $\phi_1(\mathbf{x})$ are also known as *basis functions*
- Define:

$$\mathbf{\Phi} = \begin{bmatrix} \phi_1(\mathbf{x}_1) & \ldots & \phi_D(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_1(\mathbf{x}_N) & \ldots & \phi_D(\mathbf{x}_N) \end{bmatrix}$$

Introducing basis functions

- Applying Bayesian Linear Regression on the transformed features gives

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \sigma^2) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- Covariance:

$$\boldsymbol{\Sigma} = \left(\frac{1}{\sigma^2}\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi} + \mathbf{S}^{-1}\right)^{-1}$$

- Mean:

$$\boldsymbol{\mu} = \frac{1}{\sigma^2}\boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathsf{T}}\mathbf{y}$$

- Predictions:

$$p(y_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*, \sigma^2) = \mathcal{N}(\phi_*^{\mathsf{T}}\boldsymbol{\mu}, \sigma^2 + \phi_*^{\mathsf{T}}\boldsymbol{\Sigma}\phi_*)$$

- Linear models require specifying a set of basis functions
  - Polynomials, Trigonometric, . . .??

- Linear models require specifying a set of basis functions
  - Polynomials, Trigonometric, ...??
- Can we use Bayesian inference to let data tell this to us?

- Linear models require specifying a set of basis functions
  - Polynomials, Trigonometric, ...??
- Can we use Bayesian inference to let data tell this to us?
- Gaussian Processes work implicitly with an infinite set of basis functions and learn a probabilistic combination of these

Bayesian Linear Regression as a Kernel Machine

- We are going to show that predictions can be expressed exclusively in terms of scalar products as follows

$$k(\mathbf{x}, \mathbf{x}') = \psi(\mathbf{x})^\top \psi(\mathbf{x}')$$

- This allows us to work with either $k(\cdot, \cdot)$ or $\psi(\cdot)$
- Why is this useful??

- Working with $\psi(\cdot)$ costs $O(D^2)$ storage, $O(D^3)$ time
- Working with $k(\cdot, \cdot)$ costs $O(N^2)$ storage, $O(N^3)$ time

Bayesian Linear Regression as a Kernel Machine

- Working with $\psi(\cdot)$ costs $O(D^2)$ storage, $O(D^3)$ time
- Working with $k(\cdot, \cdot)$ costs $O(N^2)$ storage, $O(N^3)$ time
- Pick the one that makes computations faster … or

Bayesian Linear Regression as a Kernel Machine

- Working with $\psi(\cdot)$ costs $O(D^2)$ storage, $O(D^3)$ time
- Working with $k(\cdot, \cdot)$ costs $O(N^2)$ storage, $O(N^3)$ time
- Pick the one that makes computations faster ... or
- What if we could pick $k(\cdot, \cdot)$ so that $\psi(\cdot)$ is infinite dimensional?

Kernels

- It is possible to show that for

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

  there exists a corresponding $\psi(\cdot)$ that is infinite dimensional!!!
- There are other kernels satisfying this property

- For simplicity consider one dimensional inputs x, y
- Expand the Gaussian kernel $k(x, y)$ as

$$\exp\left(-\frac{(x-y)^2}{2}\right) = \exp\left(-\frac{x^2}{2}\right)\exp\left(-\frac{y^2}{2}\right)\exp(xy)$$

- Focusing on the last term and applying the Taylor expansion of the $\exp(\cdot)$ function

$$\exp(xy) = 1 + (xy) + \frac{(xy)^2}{2!} + \frac{(xy)^3}{3!} + \frac{(xy)^4}{4!} + \dots$$

- Define the infinite dimensional mapping

$$\psi(x) = \exp\left(-\frac{x^2}{2}\right)\left(1, x, \frac{x^2}{\sqrt{2!}}, \frac{x^3}{\sqrt{3!}}, \frac{x^4}{\sqrt{4!}}, \dots\right)^\top$$

- It is easy to verify that

$$k(x, y) = \exp\left(-\frac{(x-y)^2}{2}\right) = \psi(x)^\top \psi(y)$$

- To show that Bayesian Linear Regression can be formulated through scalar products only, we need Woodbury identity:

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$$

- Do not memorize this!

- Woodbury identity:

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$$

- We can rewrite:

$$\begin{aligned}\mathbf{\Sigma} &= \left(\frac{1}{\sigma^2}\mathbf{\Phi}^{\mathsf{T}}\mathbf{\Phi} + \mathbf{S}^{-1}\right)^{-1} \\ &= \mathbf{S} - \mathbf{S}\mathbf{\Phi}^{\mathsf{T}}\left(\sigma^2\mathbf{I} + \mathbf{\Phi}\mathbf{S}\mathbf{\Phi}^{\mathsf{T}}\right)^{-1}\mathbf{\Phi}\mathbf{S}\end{aligned}$$

- We set $A = \mathbf{S}$, $U = V^{\top} = \mathbf{\Phi}^{\mathsf{T}}$, and $C = \frac{1}{\sigma^2}\mathbf{I}$

- Mean and variance of the predictions:

$$p(y_* | \mathbf{X}, \mathbf{y}, \mathbf{x}_*, \sigma^2) = \mathcal{N}(\phi_*^{\mathsf{T}} \boldsymbol{\mu}, \sigma^2 + \phi_*^{\mathsf{T}} \boldsymbol{\Sigma} \phi_*)$$

- Rewrite the variance:

$$\begin{aligned} \sigma^2 \quad &+ \quad \phi_*^{\mathsf{T}} \boldsymbol{\Sigma} \phi_* = \\ \sigma^2 \quad &+ \quad \phi_*^{\mathsf{T}} \mathbf{S} \phi_* - \phi_*^{\mathsf{T}} \mathbf{S} \boldsymbol{\Phi}^{\mathsf{T}} \left(\sigma^2 \mathbf{I} + \boldsymbol{\Phi} \mathbf{S} \boldsymbol{\Phi}^{\mathsf{T}} \right)^{-1} \boldsymbol{\Phi} \mathbf{S} \phi_* \end{aligned}$$

. . . continued

- Mean and variance of the predictions:

$$p(y_* | \mathbf{X}, \mathbf{y}, \mathbf{x}_*, \sigma^2) = \mathcal{N}(\phi_*^\mathsf{T} \boldsymbol{\mu}, \sigma^2 + \phi_*^\mathsf{T} \boldsymbol{\Sigma} \phi_*)$$

- Rewrite the variance:

$$\begin{aligned} \sigma^2 &+ \phi_*^\mathsf{T} \mathbf{S} \phi_* - \phi_*^\mathsf{T} \mathbf{S} \boldsymbol{\Phi}^\mathsf{T} \left(\sigma^2 \mathbf{I} + \boldsymbol{\Phi} \mathbf{S} \boldsymbol{\Phi}^\mathsf{T} \right)^{-1} \boldsymbol{\Phi} \mathbf{S} \phi_* = \\ \sigma^2 &+ k_{**} - \mathbf{k}_*^\top \left(\sigma^2 \mathbf{I} + \mathbf{K} \right)^{-1} \mathbf{k}_* \end{aligned}$$

- Where the mapping defining the kernel is

$$\psi(\mathbf{x}) = \mathbf{S}^{1/2} \phi(\mathbf{x})$$

and

$$\begin{aligned} k_{**} &= k(\mathbf{x}_*, \mathbf{x}_*) = \psi(\mathbf{x}_*)^\mathsf{T} \psi(\mathbf{x}_*) \\ (\mathbf{k}_*)_i &= k(\mathbf{x}_*, \mathbf{x}_i) = \psi(\mathbf{x}_*)^\mathsf{T} \psi(\mathbf{x}_i) \\ (\mathbf{K})_{ij} &= k(\mathbf{x}_i, \mathbf{x}_j) = \psi(\mathbf{x}_i)^\mathsf{T} \psi(\mathbf{x}_j) \end{aligned}$$

- Mean and variance of the predictions:

$$p(y_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*, \sigma^2) = \mathcal{N}(\phi_*^\mathsf{T}\boldsymbol{\mu}, \sigma^2 + \phi_*^\mathsf{T}\boldsymbol{\Sigma}\phi_*)$$

- Rewrite the mean:

$$\begin{aligned}
\phi_*^\mathsf{T}\boldsymbol{\mu} &= \frac{1}{\sigma^2}\phi_*^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\Phi}^\mathsf{T}\mathbf{y} \\
&= \frac{1}{\sigma^2}\phi_*^\mathsf{T}\left(\mathbf{S} - \mathbf{S}\boldsymbol{\Phi}^\mathsf{T}\left(\sigma^2\mathbf{I} + \boldsymbol{\Phi}\mathbf{S}\boldsymbol{\Phi}^\mathsf{T}\right)^{-1}\boldsymbol{\Phi}\mathbf{S}\right)\boldsymbol{\Phi}^\mathsf{T}\mathbf{y} \\
&= \frac{1}{\sigma^2}\phi_*^\mathsf{T}\mathbf{S}\boldsymbol{\Phi}^\mathsf{T}\left(\mathbf{I} - \left(\sigma^2\mathbf{I} + \boldsymbol{\Phi}\mathbf{S}\boldsymbol{\Phi}^\mathsf{T}\right)^{-1}\boldsymbol{\Phi}\mathbf{S}\boldsymbol{\Phi}^\mathsf{T}\right)\mathbf{y} \\
&= \frac{1}{\sigma^2}\phi_*^\mathsf{T}\mathbf{S}\boldsymbol{\Phi}^\mathsf{T}\left(\mathbf{I} - \left(\mathbf{I} + \frac{\boldsymbol{\Phi}\mathbf{S}\boldsymbol{\Phi}^\mathsf{T}}{\sigma^2}\right)^{-1}\frac{\boldsymbol{\Phi}\mathbf{S}\boldsymbol{\Phi}^\mathsf{T}}{\sigma^2}\right)\mathbf{y}
\end{aligned}$$

. . . continued

- Define $\mathbf{H} = \frac{\mathbf{\Phi S \Phi}^{\mathsf{T}}}{\sigma^2}$
- The term in the parenthesis

$$\left(\mathbf{I} - \left(\mathbf{I} + \frac{\mathbf{\Phi S \Phi}^{\mathsf{T}}}{\sigma^2}\right)^{-1} \frac{\mathbf{\Phi S \Phi}^{\mathsf{T}}}{\sigma^2}\right)$$

becomes

$$\left(\mathbf{I} - (\mathbf{I} + \mathbf{H})^{-1}\mathbf{H}\right) = \mathbf{I} - (\mathbf{H}^{-1} + \mathbf{I})^{-1}$$

- Using Woodbury ($A, U, V = \mathbf{I}$ and $C = \mathbf{H}^{-1}$)

$$\mathbf{I} - (\mathbf{H}^{-1} + \mathbf{I})^{-1} = (\mathbf{I} + \mathbf{H})^{-1}$$

- Substituting into the expression of the predictive mean

$$
\begin{aligned}
\phi_*^{\mathsf{T}} \boldsymbol{\mu} &= \frac{1}{\sigma^2} \phi_*^{\mathsf{T}} \mathbf{S} \boldsymbol{\Phi}^{\mathsf{T}} \left(\mathbf{I} - \left(\mathbf{I} + \frac{\boldsymbol{\Phi} \mathbf{S} \boldsymbol{\Phi}^{\mathsf{T}}}{\sigma^2} \right)^{-1} \frac{\boldsymbol{\Phi} \mathbf{S} \boldsymbol{\Phi}^{\mathsf{T}}}{\sigma^2} \right) \mathbf{y} \\
&= \frac{1}{\sigma^2} \phi_*^{\mathsf{T}} \mathbf{S} \boldsymbol{\Phi}^{\mathsf{T}} \left(\mathbf{I} + \frac{\boldsymbol{\Phi} \mathbf{S} \boldsymbol{\Phi}^{\mathsf{T}}}{\sigma^2} \right)^{-1} \mathbf{y} \\
&= \phi_*^{\mathsf{T}} \mathbf{S} \boldsymbol{\Phi}^{\mathsf{T}} \left(\sigma^2 \mathbf{I} + \boldsymbol{\Phi} \mathbf{S} \boldsymbol{\Phi}^{\mathsf{T}} \right)^{-1} \mathbf{y} \\
&= \mathbf{k}_*^{\top} \left(\sigma^2 \mathbf{I} + \mathbf{K} \right)^{-1} \mathbf{y}
\end{aligned}
$$

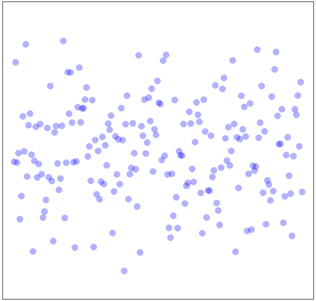
- All definitions as in the case of the variance

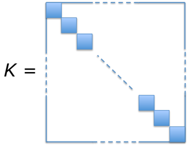
$$
\begin{aligned}
\psi(\mathbf{x}) &= \mathbf{S}^{1/2} \phi(\mathbf{x}) \\
(\mathbf{k}_*)_i &= k(\mathbf{x}_*, \mathbf{x}_i) = \psi(\mathbf{x}_*)^{\mathsf{T}} \psi(\mathbf{x}_i) \\
(\mathbf{K})_{ij} &= k(\mathbf{x}_i, \mathbf{x}_j) = \psi(\mathbf{x}_i)^{\mathsf{T}} \psi(\mathbf{x}_j)
\end{aligned}
$$

Gaussian Processes can be explained in two ways

- Weight Space View
    - Bayesian linear regression with infinite basis functions
- **Function Space View**
    - **Defined as priors over functions**

- Consider an infinite number of Gaussian random variables
- Think of them as indexed by the real line and as independent
- Denote them as $f(x)$





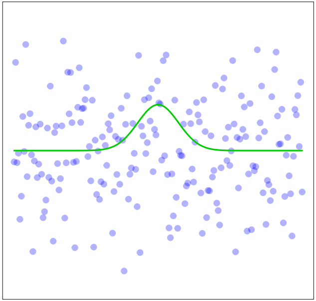
Kernel

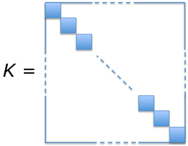
- Consider the Gaussian kernel again

$$k(\mathbf{x}, \mathbf{x}') = \alpha \exp(-\beta \|\mathbf{x} - \mathbf{x}'\|^2)$$

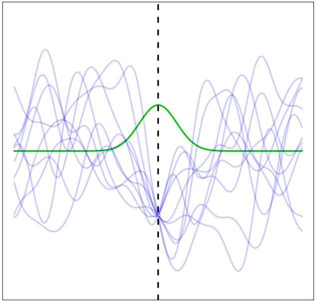
- We introduced some parameters for added flexibility

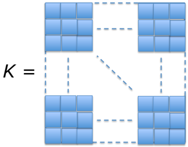
- Impose covariance using the kernel function



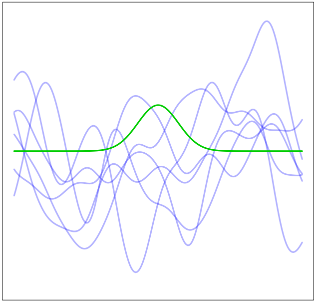


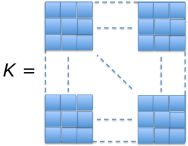
- Draw the infinite random variables again fixing one of them (the one at $x = 0$)



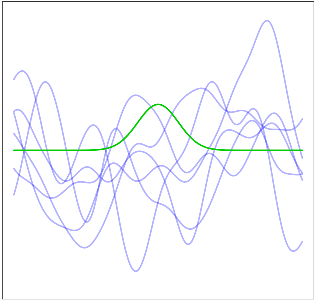


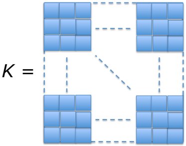
- Draw the infinite random variables again allowing the one at $x = 0$ to be random too





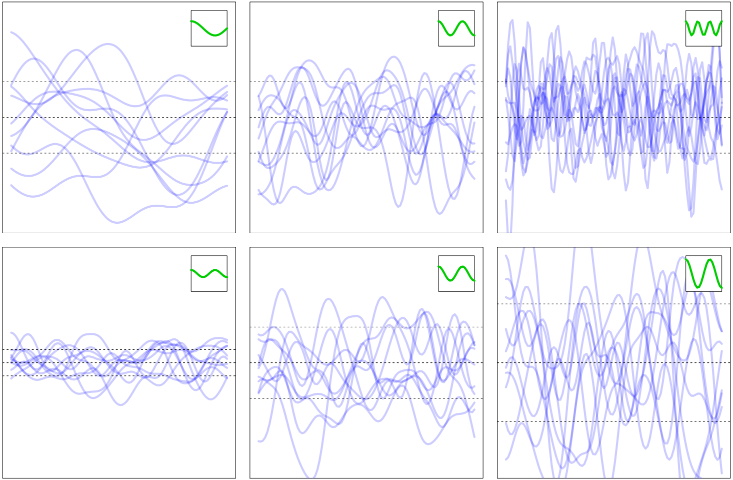
- This can be used as a prior over functions!



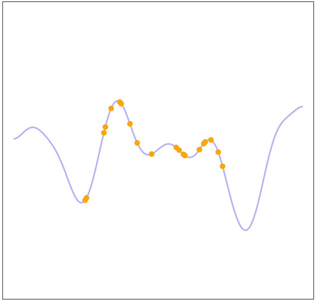


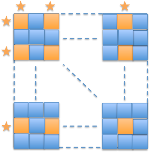
Gaussian Processes - Priors over Functions

- Infinite Gaussian random variables with parameterized and input-dependent covariance

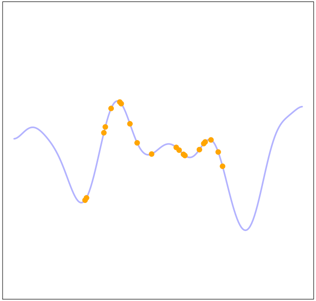


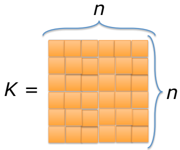
- The distribution of N random variables $f(x_1), \ldots, f(x_N)$ depends exclusively on the corresponding rows and columns of the infinite by infinite kernel matrix K



$K =$ 

- The distribution of N random variables $f(x_1), \ldots, f(x_N)$ depends exclusively on the corresponding rows and columns of the infinite by infinite kernel matrix K





- The marginal distribution of $\mathbf{f} = (f(x_1), \ldots, f(x_N))^\top$ is

$$p(\mathbf{f}|\mathbf{X}) = \mathcal{N}(\mathbf{0}, \mathbf{K})$$

- The conditional distribution of f_* given $\mathbf{f}$

$$p(f_*|\mathbf{f}, \mathbf{x}_*, \mathbf{X}) = \mathcal{N}(\bar{m}, \bar{s}^2)$$

with

$$\bar{m} = \mathbf{k}_*^\top \mathbf{K}^{-1} \mathbf{f}$$

$$\bar{s}^2 = k_{**} - \mathbf{k}_*^\top \mathbf{K}^{-1} \mathbf{k}_*$$

Gaussian Processes - Priors over Functions

- Remember that when we modeled labels $\mathbf{y}$ in the linear model we assumed noise with variance σ around $\mathbf{w}^{\mathsf{T}}\mathbf{x}$
- We can do the same in Gaussian processes

$$p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{N} p(y_i|f_i)$$

with

$$p(y_i|f_i) = \mathcal{N}(y_i|f_i, \sigma^2)$$

- Likelihood and prior are both Gaussian - conjugate!

Gaussian Processes - Priors over Functions

- Remember that when we modeled labels $\mathbf{y}$ in the linear model we assumed noise with variance σ around $\mathbf{w}^\mathsf{T}\mathbf{x}$
- We can do the same in Gaussian processes

$$p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{N} p(y_i|f_i)$$

with

$$p(y_i|f_i) = \mathcal{N}(y_i|f_i, \sigma^2)$$

- Likelihood and prior are both Gaussian - conjugate!
- We can integrate out Gaussian process prior on $\mathbf{f}$

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X})d\mathbf{f}$$

- This gives

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

Gaussian Processes - Priors over Functions

- We can derive the predictive distribution of the function:

$$p(f_* | \mathbf{y}, \mathbf{x}_* \mathbf{X}) = \int p(f_* | \mathbf{f}, \mathbf{x}_*, \mathbf{X}) p(\mathbf{f} | \mathbf{y}, \mathbf{X}) d\mathbf{f} df_* = \mathcal{N}(m, s^2)$$

with

$$m = \mathbf{k}_*^\top \left(\mathbf{K} + \sigma^2 \mathbf{I}\right)^{-1} \mathbf{y}$$

$$s^2 = k_{**} - \mathbf{k}_*^\top \left(\mathbf{K} + \sigma^2 \mathbf{I}\right)^{-1} \mathbf{k}_*$$

- Same expression as in the "Weight-Space View" section

- We can also make predictions as follows:

$$\begin{aligned} p(y_* | \mathbf{y}, \mathbf{x}_* \mathbf{X}) &= \int p(t_* | f_*) p(f_* | \mathbf{f}, \mathbf{x}_*, \mathbf{X}) p(\mathbf{f} | \mathbf{y}, \mathbf{X}) d\mathbf{f} \, df_* \\ &= \mathcal{N}(m_{\mathbf{y}}, s_{\mathbf{y}}^2) \end{aligned}$$

with

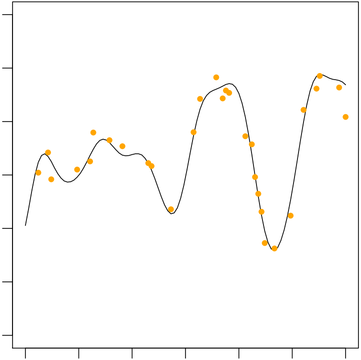
$$m_{\mathbf{y}} = \mathbf{k}_*^\top \left(\mathbf{K} + \sigma^2 \mathbf{I} \right)^{-1} \mathbf{y}$$

$$s_{\mathbf{y}}^2 = \sigma^2 + k_{**} - \mathbf{k}_*^\top \left(\mathbf{K} + \sigma^2 \mathbf{I} \right)^{-1} \mathbf{k}_*$$

- Same expression as in the "Weight-Space View" section

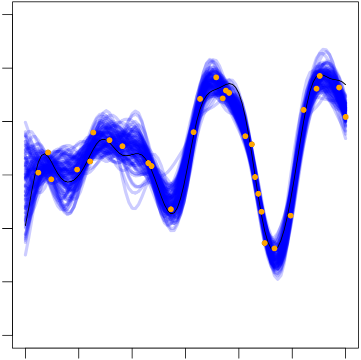
Gaussian Processes - Regression example

- Some data generated as a noisy version of some function



Gaussian Processes - Regression example

- Draws from the posterior distribution over f_* on the real line



- The kernel has parameters that have to be tuned

$$k(\mathbf{x}, \mathbf{x}') = \alpha \exp(-\beta \|\mathbf{x} - \mathbf{x}'\|^2)$$

  ... and there is also the noise parameter σ^2.
- Define $\boldsymbol{\theta} = (\alpha, \beta, \sigma^2)$
- How should we tune them?

Optimization of Gaussian Process parameters

- Define $\mathbf{K_y} = \mathbf{K} + \sigma^2 \mathbf{I}$
- Maximize the logarithm of the likelihood

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, \mathbf{K_y})$$

  that is

$$-\frac{1}{2} \log |\mathbf{K_y}| - \frac{1}{2} \mathbf{y}^{\mathsf{T}} \mathbf{K_y}^{-1} \mathbf{y} + \mathrm{const.}$$

- Derivatives can be useful for gradient-based optimization

$$\frac{\partial \log[p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})]}{\partial \boldsymbol{\theta}_i}$$

Optimization of Gaussian Process parameters

- Log-likelihood

$$-\frac{1}{2}\log|\mathbf{K_y}| - \frac{1}{2}\mathbf{y}^{\mathsf{T}}\mathbf{K_y}^{-1}\mathbf{y} + \text{const.}$$

- Derivatives can be useful for gradient-based optimization:

$$\frac{\partial \log[p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}_i} = -\frac{1}{2}\mathrm{Tr}\left(\mathbf{K_y}^{-1}\frac{\partial \mathbf{K_y}}{\partial \boldsymbol{\theta}_i}\right) + \frac{1}{2}\mathbf{y}^{\mathsf{T}}\mathbf{K_y}^{-1}\frac{\partial \mathbf{K_y}}{\partial \boldsymbol{\theta}_i}\mathbf{K_y}^{-1}\mathbf{y}$$

Challenges

Challenges

- Non-Gaussian Likelihoods?
- Scalability?

Marginal likelihood of GP models - non-Gaussian case

- Marginal likelihood

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta}) d\mathbf{f}$$

  can only be computed if $p(\mathbf{y}|\mathbf{f})$ is Gaussian
- What if $p(\mathbf{y}|\mathbf{f})$ is **not** Gaussian?

Tackling non-Gaussian case

- Approximation options:
  - Local variational bounds (classification only)
    - Gibbs and MacKay, *IEEE TNN*, 2000

Tackling non-Gaussian case

- Approximation options:
  - Local variational bounds (classification only)
    - Gibbs and MacKay, *IEEE TNN*, 2000
  - Laplace Approximation
    - Williams and Barber, *IEEE TPAMI*, 1998

Tackling non-Gaussian case

- Approximation options:
  - Local variational bounds (classification only)
    - Gibbs and MacKay, *IEEE TNN*, 2000
  - Laplace Approximation
    - Williams and Barber, *IEEE TPAMI*, 1998
  - Expectation Propagation
    - Minka, *PhD thesis*, 2001

Tackling non-Gaussian case

- Approximation options:
  - Local variational bounds (classification only)
    - Gibbs and MacKay, *IEEE TNN*, 2000
  - Laplace Approximation
    - Williams and Barber, *IEEE TPAMI*, 1998
  - Expectation Propagation
    - Minka, *PhD thesis*, 2001
  - Variational Bayes
    - Nickisch and Rasmussen, *JMLR*, 2008
    - Opper and Archambeau, *Neural Comp*, 2009

Tackling non-Gaussian case

- Approximation options:
  - Local variational bounds (classification only)
    - Gibbs and MacKay, *IEEE TNN*, 2000
  - Laplace Approximation
    - Williams and Barber, *IEEE TPAMI*, 1998
  - Expectation Propagation
    - Minka, *PhD thesis*, 2001
  - Variational Bayes
    - Nickisch and Rasmussen, *JMLR*, 2008
    - Opper and Archambeau, *Neural Comp*, 2009
  - Markov chain Monte Carlo
    - Murray and Adams, *NIPS*, 2010
    - Filippone and Girolami, *IEEE TPAMI*, 2014

- Marginal likelihood

$$p(\mathbf{y}|\mathbf{X},\theta) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X},\theta)d\mathbf{f}$$

can only be computed if $p(\mathbf{y}|\mathbf{X},\mathbf{f})$ is Gaussian

- ... even then

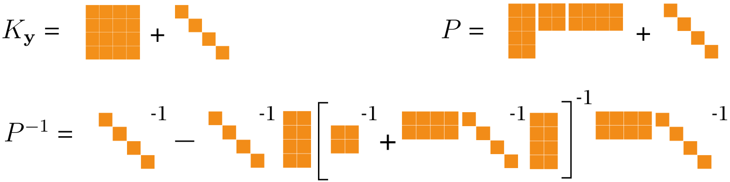
$$\log[p(\mathbf{y}|\mathbf{X},\theta)] = -\frac{1}{2}\log|\mathbf{K}_{\mathbf{y}}| - \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}_{\mathbf{y}}^{-1}\mathbf{y} + \mathrm{const.}$$

where $\mathbf{K}_{\mathbf{y}} = K(\mathbf{X},\theta)$ is a $n \times n$ dense matrix!

- Complexity of exact method is $\mathcal{O}(n^3)$ time and $\mathcal{O}(n^2)$ space!

Tackling Gaussian case and $n \gg$

- Low-Rank Approximation options - $\mathcal{O}(nm^2)$
- Call P as a low rank approximation to $\mathbf{K_y}$
- Woodbury identity exploits low rank structure of P



Tackling Gaussian case and $n \gg$

- Low-Rank Approximation options - $\mathcal{O}(nm^2)$
  - Subset-of-data 'sparse' methods
    - Smola and Bartlett, *NIPS*, 2001
    - Seeger and Williams, *AISTATS*, 2003

- Low-Rank Approximation options - $\mathcal{O}(nm^2)$
  - Subset-of-data 'sparse' methods
    - Smola and Bartlett, *NIPS*, 2001
    - Seeger and Williams, *AISTATS*, 2003
  - Pseudo-inputs introduced
    - Snelson and Ghahramani, *NIPS*, 2005

Tackling Gaussian case and $n \gg$

- Low-Rank Approximation options - $\mathcal{O}(nm^2)$
  - Subset-of-data 'sparse' methods
    - Smola and Bartlett, *NIPS*, 2001
    - Seeger and Williams, *AISTATS*, 2003
  - Pseudo-inputs introduced
    - Snelson and Ghahramani, *NIPS*, 2005
  - A unifying view brings several ideas together
    - Quiñonero-Candela and Rasmussen, *JMLR*, 2005

Tackling Gaussian case and $n \gg$

- Low-Rank Approximation options - $\mathcal{O}(nm^2)$
  - Subset-of-data 'sparse' methods
    - Smola and Bartlett, *NIPS*, 2001
    - Seeger and Williams, *AISTATS*, 2003
  - Pseudo-inputs introduced
    - Snelson and Ghahramani, *NIPS*, 2005
  - A unifying view brings several ideas together
    - Quiñonero-Candela and Rasmussen, *JMLR*, 2005
  - Variational approach for better placement of pseudo points
    - Titsias, *AISTATS*, 2009

Tackling Gaussian case and $n \gg$

- Low-Rank Approximation options - $\mathcal{O}(nm^2)$
  - Subset-of-data 'sparse' methods
    - Smola and Bartlett, *NIPS*, 2001
    - Seeger and Williams, *AISTATS*, 2003
  - Pseudo-inputs introduced
    - Snelson and Ghahramani, *NIPS*, 2005
  - A unifying view brings several ideas together
    - Quiñonero-Candela and Rasmussen, *JMLR*, 2005
  - Variational approach for better placement of pseudo points
    - Titsias, *AISTATS*, 2009
  - Random feature expansions
    - Rahimi and Recht, *NIPS*, 2008
    - Lazaro-Gredilla et al., *JMLR*, 2010

Tackling Gaussian case and $n \gg$

- Approximation options:
  - Structured approximations based on Toeplitz/circulant matrices - $\mathcal{O}(dn^{\frac{d+1}{d}})$ time
    - Wilson and Nickisch, *ICML*, 2015
    - Gilboa et al., *IEEE TPAMI*, 2015

Tackling Gaussian case and $n \gg$

- Approximation options:
  - Structured approximations based on Toeplitz/circulant matrices - $\mathcal{O}(dn^{\frac{d+1}{d}})$ time
    - Wilson and Nickisch, *ICML*, 2015
    - Gilboa et al., *IEEE TPAMI*, 2015
  - Stochastic-gradient optimization/inference **without** model approximations - $\mathcal{O}(n^2)$ time and $\mathcal{O}(n)$ space
    - Filippone and Engler, *ICML*, 2015
    - Cutajar, Osborne, Cunnningham, Filippone, *ICML*, 2016

Teaser - Stochastic Gradients in GP regression

- Marginal likelihood

$$\log[p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})] = -\frac{1}{2}\log|\mathbf{K_y}| - \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K_y}^{-1}\mathbf{y} + \mathrm{const.}$$

- Derivatives wrt $\boldsymbol{\theta}$

$$\frac{\partial \log[p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})]}{\partial \boldsymbol{\theta}_i} = -\frac{1}{2}\mathrm{Tr}\left(\mathbf{K_y}^{-1}\frac{\partial \mathbf{K_y}}{\partial \boldsymbol{\theta}_i}\right) + \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K_y}^{-1}\frac{\partial \mathbf{K_y}}{\partial \boldsymbol{\theta}_i}\mathbf{K_y}^{-1}\mathbf{y}$$

Filippone and Engler, *ICML*, 2015 - Cutajar, Osborne, Cunningham, Filippone, *ICML*, 2016

Teaser - Stochastic Gradients in GP regression

- Stochastic estimate of the trace

$$\mathrm{Tr}\left(K^{-1}\frac{\partial \mathbf{K_y}}{\partial \boldsymbol{\theta}_i}\right) = \mathrm{Tr}\left(\mathbf{K_y}^{-1}\frac{\partial \mathbf{K_y}}{\partial \boldsymbol{\theta}_i}\mathrm{E}[\mathbf{r}\mathbf{r}^{\mathrm{T}}]\right) = \mathrm{E}\left[\mathbf{r}^{\mathrm{T}}\mathbf{K_y}^{-1}\frac{\partial \mathbf{K_y}}{\partial \boldsymbol{\theta}_i}\mathbf{r}\right]$$

with $\mathrm{E}[\mathbf{r}\mathbf{r}^{\mathrm{T}}] = I$

- Stochastic estimate of the trace

$$\mathrm{Tr}\left(K^{-1}\frac{\partial \mathbf{K_y}}{\partial \boldsymbol{\theta}_i}\right) = \mathrm{Tr}\left(\mathbf{K_y}^{-1}\frac{\partial \mathbf{K_y}}{\partial \boldsymbol{\theta}_i}\mathrm{E}[\mathbf{r}\mathbf{r}^{\mathrm{T}}]\right) = \mathrm{E}\left[\mathbf{r}^{\mathrm{T}}\mathbf{K_y}^{-1}\frac{\partial \mathbf{K_y}}{\partial \boldsymbol{\theta}_i}\mathbf{r}\right]$$

with $\mathrm{E}[\mathbf{r}\mathbf{r}^{\mathrm{T}}] = I$

- Stochastic gradient

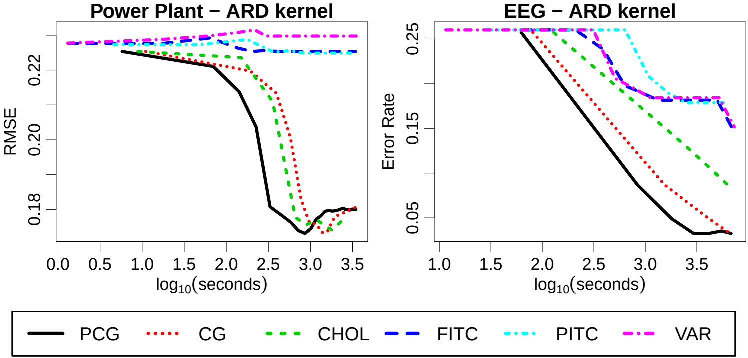
$$-\frac{1}{2N_{\mathbf{r}}}\sum_{i=1}^{N_{\mathbf{r}}}\mathbf{r}^{(i)\mathrm{T}}\mathbf{K_y}^{-1}\frac{\partial \mathbf{K_y}}{\partial \boldsymbol{\theta}_i}\mathbf{r}^{(i)} + \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K_y}^{-1}\frac{\partial \mathbf{K_y}}{\partial \boldsymbol{\theta}_i}\mathbf{K_y}^{-1}\mathbf{y}$$

- Linear systems only!

Filippone and Engler, *ICML*, 2015 - Cutajar, Osborne, Cunningham, Filippone, *ICML*, 2016

Teaser - Preconditioning Kernel Matrices

- Stochastic Gradient Optimization



Cutajar, Osborne, Cunningham, Filippone, *ICML*, 2016

Challenges

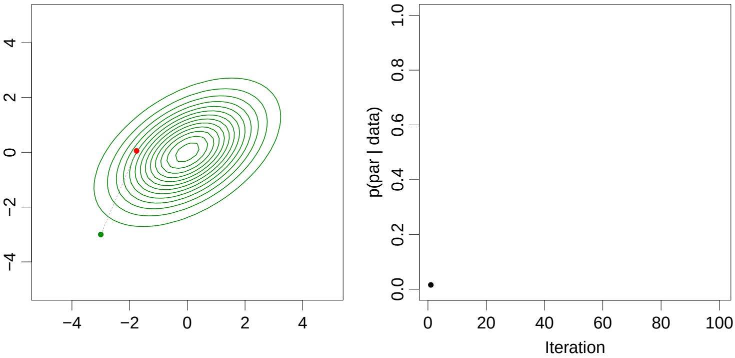
- Non-Gaussian Likelihoods?
- Scalability?

Modern GP works tackle both

Modern Gaussian Processes

Desirable properties

- Mini-batch-based learning - $\mathcal{O}(1)$ time for each iteration!
- Exploit GPU and distributed computing
- Automatic differentiation
- Application-specific representations (e.g., convolutional)



Robbins and Monro, *AoMS*, 1951

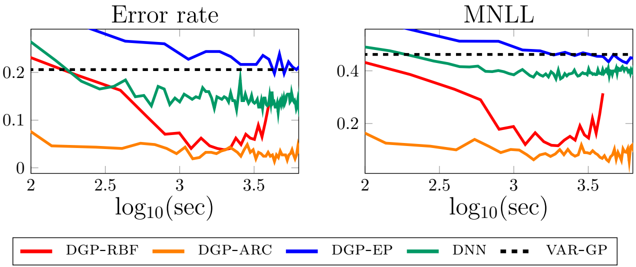
- Approximation options:
    - Scalable Expectation Propagation
        - Bui et al., *ICML*, 2016

- Approximation options:
  - Scalable Expectation Propagation
    - Bui et al., *ICML*, 2016
  - Inducing points methods
    - Hensman et al., *AISTATS*, 2013
    - Hensman, Matthews, Ghahramani, Filippone, *NIPS*, 2015

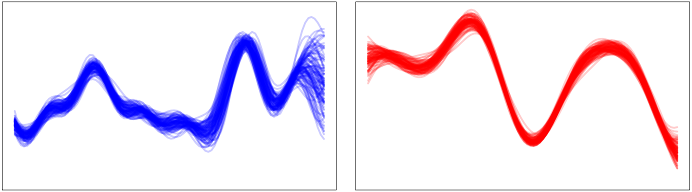
Modern GPs - Any likelihood and $n \gg$

- Approximation options:
  - Scalable Expectation Propagation
    - Bui et al., *ICML*, 2016
  - Inducing points methods
    - Hensman et al., *AISTATS*, 2013
    - Hensman, Matthews, Ghahramani, Filippone, *NIPS*, 2015
  - Random feature expansions
    - Gal, Ghahramani, *ICML*, 2016
    - Cutajar, Bonilla, Michiardi, Filippone, *ICML*, 2017

EEG dataset
$(n = 14979,\ d = 14)$



- Composition of processes - Deep Gaussian Processes

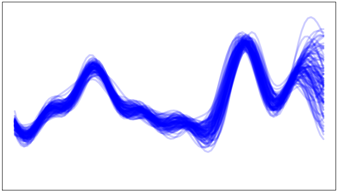


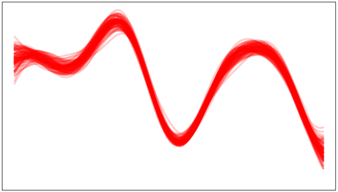
$$(f \circ g)(x)??$$

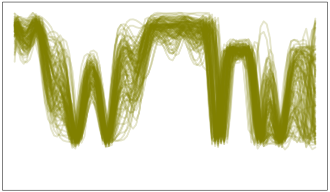
Damianou and Lawrence, *AISTATS*, 2013 – Cutajar, Bonilla, Michiardi, Filippone, *ICML*, 2017

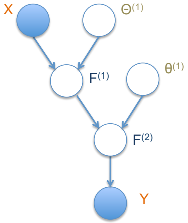
Teaser - Modern GPs - Any likelihood and $n \gg$

- Composition of processes





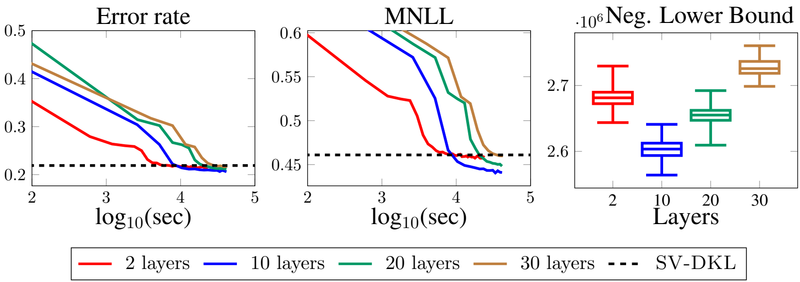




Damianou and Lawrence, *AISTATS*, 2013 – Cutajar, Bonilla, Michiardi, Filippone, *ICML*, 2017

Airline dataset

$(n = 5\mathrm{M}+,\ d = 8)$



Other interesting topics

- Bayesian Optimization
  - Jones et al., *JoGO*, 1998

Other interesting topics

- Bayesian Optimization
  - Jones et al., *JoGO*, 1998
- Unsupervised Learning
  - Lawrence, *NIPS*, 2004

Other interesting topics

- Bayesian Optimization
  - Jones et al., *JoGO*, 1998
- Unsupervised Learning
  - Lawrence, *NIPS*, 2004
- Deep Gaussian Processes
  - Damianou and Lawrence, *AISTATS*, 2013
  - Cutajar, Bonilla, Michiardi, and Filippone, *ICML*, 2017

Other interesting topics

- Bayesian Optimization
  - Jones et al., *JoGO*, 1998
- Unsupervised Learning
  - Lawrence, *NIPS*, 2004
- Deep Gaussian Processes
  - Damianou and Lawrence, *AISTATS*, 2013
  - Cutajar, Bonilla, Michiardi, and Filippone, *ICML*, 2017
- Convolutional Gaussian Processes
  - Wilson et al., *AISTATS*, 2015
  - Wilson et al., *NIPS*, 2016
  - van der Wilk et al., *NIPS*, 2017

Other interesting topics

- Bayesian Optimization
  - Jones et al., *JoGO*, 1998
- Unsupervised Learning
  - Lawrence, *NIPS*, 2004
- Deep Gaussian Processes
  - Damianou and Lawrence, *AISTATS*, 2013
  - Cutajar, Bonilla, Michiardi, and Filippone, *ICML*, 2017
- Convolutional Gaussian Processes
  - Wilson et al., *AISTATS*, 2015
  - Wilson et al., *NIPS*, 2016
  - van der Wilk et al., *NIPS*, 2017
- Structured output
  - Galliani et al., *AISTATS*, 2017

Other interesting topics

- Bayesian Optimization
  - Jones et al., *JoGO*, 1998
- Unsupervised Learning
  - Lawrence, *NIPS*, 2004
- Deep Gaussian Processes
  - Damianou and Lawrence, *AISTATS*, 2013
  - Cutajar, Bonilla, Michiardi, and Filippone, *ICML*, 2017
- Convolutional Gaussian Processes
  - Wilson et al., *AISTATS*, 2015
  - Wilson et al., *NIPS*, 2016
  - van der Wilk et al., *NIPS*, 2017
- Structured output
  - Galliani et al., *AISTATS*, 2017
- Probabilistic Numerics
  - Fitzsimons, Cutajar, Osborne, Roberts, Filippone, *UAI*, 2017

Acknowledgments

Thank you!