# CS489/698
# Lecture 11: Feb 8, 2017

Gaussian Processes

[B] Section 6.4 [M] Chap. 15 [HTF] Sec. 8.3

# Gaussian Process Regression

- Idea: distribution over functions

# Bayesian Linear Regression

- Setting: $f(\boldsymbol{x}) = \underset{\substack{\downarrow \\ \text{unknown}}}{\boldsymbol{w}}^T \phi(\boldsymbol{x})$ and $y = f(\boldsymbol{x}) + \underset{\substack{\downarrow \\ N(0,\sigma^2)}}{\epsilon}$

- Weight space view:
  - Prior: $\Pr(\boldsymbol{w})$
  - Posterior: $\underset{\substack{\downarrow \\ \text{Gaussian}}}{\Pr(\boldsymbol{w}|\boldsymbol{X},\boldsymbol{y})} = k \underset{\substack{\downarrow \\ \text{Gaussian}}}{\Pr(\boldsymbol{w})} \underset{\substack{\downarrow \\ \text{Gaussian}}}{\Pr(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{X})}$

# Bayesian Linear Regression

- Setting: $f(\boldsymbol{x}) = \boldsymbol{w}^T \phi(\boldsymbol{x})$ and $y = f(\boldsymbol{x}) + \epsilon$

  $f$: unknown; $\epsilon$: $N(0, \sigma^2)$

- Function space view:
  - Prior: $\Pr\big(f(\boldsymbol{x}_*)\big) = \int_{\boldsymbol{w}} \Pr(f|\boldsymbol{w}, \boldsymbol{x}_*) \Pr(\boldsymbol{w})\, d\boldsymbol{w}$

    $\Pr\big(f(\boldsymbol{x}_*)\big)$: Gaussian; $\Pr(f|\boldsymbol{w}, \boldsymbol{x}_*)$: Deterministic; $\Pr(\boldsymbol{w})$: Gaussian

  - Posterior: $\Pr(f(\boldsymbol{x}_*)|\boldsymbol{X}, \boldsymbol{y}) = \int_{\boldsymbol{w}} \Pr(f|\boldsymbol{w}, \boldsymbol{x}_*) \Pr(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y})\, d\boldsymbol{w}$

    $\Pr(f(\boldsymbol{x}_*)|\boldsymbol{X}, \boldsymbol{y})$: Gaussian; $\Pr(f|\boldsymbol{w}, \boldsymbol{x}_*)$: Deterministic; $\Pr(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y})$: Gaussian

# Gaussian Process

- According to the function view, there is a Gaussian at $f(\boldsymbol{x}_*)$ for every $\boldsymbol{x}_*$. Those Gaussians are correlated through $w$.

- What is the general form of $\Pr(f)$ (i.e., distribution over functions)?

- Answer: **Gaussian Process** (infinite dimensional Gaussian distribution)

# Gaussian Process

- Distribution over functions:

$$f(\boldsymbol{x}) \sim GP\big(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')\big)\ \forall \boldsymbol{x}, \boldsymbol{x}'$$

- Where $m(\boldsymbol{x}) = E(f(\boldsymbol{x}))$ is the mean
  and $k(\boldsymbol{x}, \boldsymbol{x}') = E((f(\boldsymbol{x}) - m(\boldsymbol{x}))(f(\boldsymbol{x}') - m(\boldsymbol{x}'))$ is
  the kernel covariance function

# Mean function $m(\boldsymbol{x})$

- Compute the mean function $m(x)$ as follows:
- Let $f(\boldsymbol{x}) = \phi(\boldsymbol{x})^T \boldsymbol{w}$
  with $\boldsymbol{w} \sim N(\boldsymbol{0}, \alpha^{-1}\boldsymbol{I})$
- Then $m(\boldsymbol{x}) = E(f(\boldsymbol{x}))$
  $= E(\boldsymbol{w})^T \phi(\boldsymbol{x})$
  $= \boldsymbol{0}$

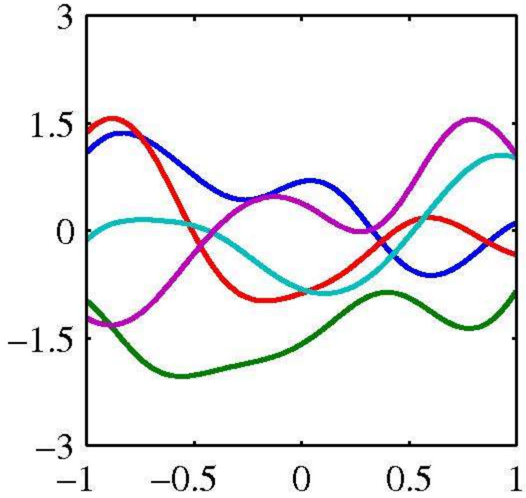
# Kernel covariance function $k(\boldsymbol{x}, \boldsymbol{x}')$

- Compute kernel covariance $k(\boldsymbol{x}, \boldsymbol{x}')$ as follows:

- $k(\boldsymbol{x}, \boldsymbol{x}') = E(f(\boldsymbol{x})f(\boldsymbol{x}'))$
  $= \phi(\boldsymbol{x})^T E(\boldsymbol{w}\boldsymbol{w}^{\boldsymbol{T}})\phi(\boldsymbol{x}')$
  $= \phi(\boldsymbol{x})^T \frac{I}{\alpha} \phi(\boldsymbol{x}')$
  $= \frac{\phi(\boldsymbol{x})^T \phi(\boldsymbol{x}')}{\alpha}$
- In some cases we can use domain knowledge to specify $k$ directly.

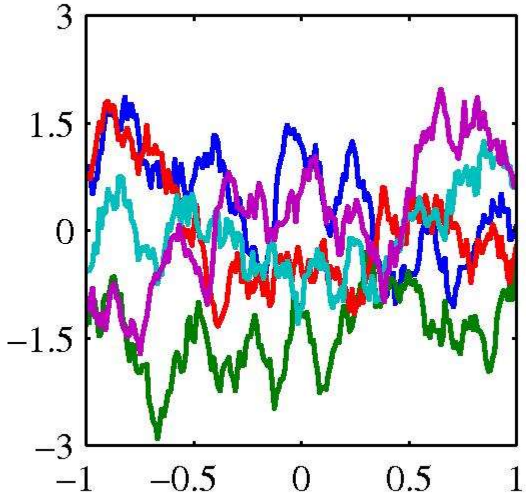
# Examples

- Sampled functions from a Gaussian Process




Gaussian kernel

$$k(\boldsymbol{x}, \boldsymbol{x}') = e^{-\frac{\left|\left|\boldsymbol{x}-\boldsymbol{x}'\right|\right|^2}{2\sigma^2}}$$




Exponential kernel (Brownian motion)

$$k(\boldsymbol{x}, \boldsymbol{x}') = e^{-\theta|\boldsymbol{x}-\boldsymbol{x}'|}$$

# Gaussian Process Regression

- Gaussian Process Regression corresponds to kernelized Bayesian Linear Regression
- Bayesian Linear Regression:
  - Weight space view
  - Goal: $\Pr(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y})$ (posterior over $\boldsymbol{w}$)
  - Complexity: cubic in # of basis functions
- Gaussian Process Regression:
  - Function space view
  - Goal: $\Pr(f|\boldsymbol{X}, \boldsymbol{y})$ (posterior over $f$)
  - Complexity: cubic in # of training points

# Recap: Bayesian Linear Regression

- Prior: $\Pr(\boldsymbol{w}) = N(\mathbf{0}, \boldsymbol{\Sigma})$
- Likelihood: $\Pr(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) = N\left(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{\Phi}, \sigma^2 \boldsymbol{I}\right)$

- Posterior: $\Pr(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}) = N\left(\bar{\boldsymbol{w}}, \boldsymbol{A}^{-\mathbf{1}}\right)$
  where $\bar{\boldsymbol{w}} = \sigma^{-2}\boldsymbol{A}^{-\mathbf{1}}\boldsymbol{\Phi}\boldsymbol{y}$ and $\boldsymbol{A} = \sigma^{-2}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\boldsymbol{T}} + \boldsymbol{\Sigma}^{-1}$
- Prediction: $\Pr(y_*|\boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y}) =$
  $N(\sigma^{-2}\phi(\boldsymbol{x}_*)^T \boldsymbol{A}^{-1}\boldsymbol{\Phi}\boldsymbol{y}, \phi(\boldsymbol{x}_*)^T\boldsymbol{A}^{-1}\phi(\boldsymbol{x}_*))$

- Complexity: inversion of $\boldsymbol{A}$ is cubic in # of basis functions

# Gaussian Process Regression

- Prior: $\Pr(f(\cdot)) = N(m(\cdot), k(\cdot,\cdot))$
- Likelihood: $\Pr(\boldsymbol{y}|\boldsymbol{X}, f) = N(f(\cdot), \sigma^2\boldsymbol{I})$

- Posterior: $\Pr(f(\cdot)|\boldsymbol{X}, \boldsymbol{y}) = N\left(\bar{f}(\cdot), k'(\cdot,\cdot)\right)$

  where $\bar{f}(\cdot) = k(\cdot, \boldsymbol{X})(\boldsymbol{K} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{y}$ and

  $k'(\cdot,\cdot) = k(\cdot,\cdot) - k(\cdot, \boldsymbol{X})(\boldsymbol{K} + \sigma^2\boldsymbol{I})^{-1}k(\boldsymbol{X},\cdot)$
- Prediction: $\Pr(y_*|\boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y}) = N\left(\bar{f}(\boldsymbol{x}_*), k'(\boldsymbol{x}_*, \boldsymbol{x}_*)\right)$

- Complexity: inversion of $\boldsymbol{K} + \sigma^2\boldsymbol{I}$ is cubic in # of training points

# Case Study: AIBO Gait Optimization

# Gait Optimization

- Problem: find best parameter setting of the gait controller to maximize walking speed
  - Why?: Fast robots have a better chance of winning in robotic soccer

- Solutions:
  - Stochastic hill climbing
  - **Gaussian Processes**
    - Lizotte, Wang, Bowling, Schuurmans (2007) Automatic Gait Optimization with Gaussian Processes, *International Joint Conferences on Artificial Intelligence (IJCAI)*.

# Search Problem

- Let $\boldsymbol{x} \in \Re^{15}$, be a vector of 15 parameters that defines a controller for gait
- Let $f: \boldsymbol{x} \rightarrow \Re$ be a mapping from controller parameters to gait speed

- Problem: find parameters $\boldsymbol{x}^*$ that yield highest speed.

$$\boldsymbol{x}^* \leftarrow argmax_{\boldsymbol{x}} \, f(\boldsymbol{x})$$

But $f$ is unknown…

# Approach

- Picture

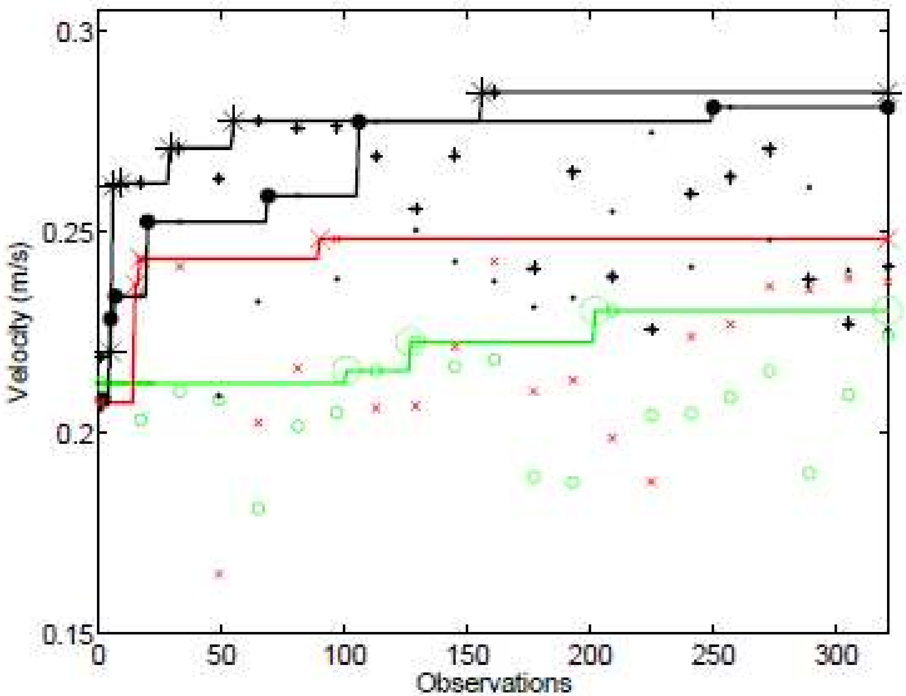
# Approach

- Initialize $f(\cdot) \sim GP(m(\cdot), k(\cdot,\cdot))$
- Repeat:
  - Select new $\boldsymbol{x}$:

    $$\boldsymbol{x}_{new} \leftarrow argmax_{\boldsymbol{x}} \frac{k(\boldsymbol{x},\boldsymbol{x})}{\max_{\boldsymbol{x}' \in X} f(\boldsymbol{x}') - m(\boldsymbol{x})}$$

  - Evaluate $f(\boldsymbol{x}_{\boldsymbol{new}})$ by observing speed of robot with parameters set to $\boldsymbol{x}_{new}$
  - Update Gaussian process:
    - $\boldsymbol{X} \leftarrow \boldsymbol{X} \cup \{\boldsymbol{x}_{\boldsymbol{new}}\}$ and $\boldsymbol{y} \leftarrow \boldsymbol{y} \cup f(\boldsymbol{x}_{\boldsymbol{new}})$
    - $m(\cdot) \leftarrow k(\cdot, \boldsymbol{X})(\boldsymbol{K} + \sigma^{-2}\boldsymbol{I})^{-1}\boldsymbol{y}$
    - $k(\cdot,\cdot) \leftarrow k(\cdot,\cdot) - k(\cdot, \boldsymbol{X})(\boldsymbol{K} + \sigma^{2}\boldsymbol{I})^{-1}k(\boldsymbol{X},\cdot)$

# Results




(•) GP w/MPI 0.281 m/s $\sigma_f^2 = 0.06$ | (∗) GP w/MPI 0.285 m/s $\sigma_f^2 = 0.6$ | (×) H.Clmb 0.248 m/s | (○) U.Rand 0.230 m/s

Gaussian kernel:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma_f^2 e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}')^T S(\boldsymbol{x}-\boldsymbol{x}')}$$