$\LaTeX$ command declarations here.

EECS 545: Machine Learning
Lecture 6: Bayesian Linear Regression & Gaussian Processes

- Instructor: **Jacob Abernethy**
- Date: February 17, 2015

Lecture Exposition Credit: Ben & Valli

Outline for this Lecture

- More on Multivariate Gaussian
  - Partitioned Marginal in Multivariate Gaussian
  - Conditional of Multivariate Gaussian
  - Bayes' Theorem for Linear Gaussian System
- Bayesian Linear Regression
  - Bayesian Linear Regression
  - Sequential Bayesian Learning
- Gaussian Processes

Reading List

- Required:
  - **[PRML]**, §3.3: Bayesian Linear Regression
  - **[PRML]**, §6.4: Gaussian Processes
- Optional:
  - **[MLAPP]**, §7.6.1-7.6.2: Bayesian Linear Regression
  - **[MLAPP]**, §4.3: Inference in Jointly Gaussian Distributions
  - **[CS229]** Ng, Andrew. CS 229: Machine Learning (http://cs229.stanford.edu/). Autumn 2015.
    - Gaussian Processes (http://cs229.stanford.edu/section/cs229-gaussian_processes.pdf)
    - More on Gaussians (http://cs229.stanford.edu/section/more_on_gaussians.pdf)

> In this lecture, our main goal is to introduce linear regression from a Bayesian perspective which is an extension to last lecture. One important application of this is to let linear regression work in a *online* fashion.
>
> Specifically, we will first cover some basics of multivariate Gaussians and derive the results of Bayes' Theorem for linear Gaussian system which will be used later. Next we will show how to do linear regression in Bayesian setting. Unlike finding a deterministic estimate of coefficients $\mathbf{w}$ in previous lectures, we will find a distribution of $\mathbf{w}$. Then, we will show how to apply Bayesian linear regression to streaming data. Streaming scenarios include realtime measurements, large amount of data that exceeds memory limit, etc.. Finally, we will introduce Gaussian processes and show Bayesian linear regression is actually a Gaussian process.

More on Multivariate Gaussians

> Taken from **[PRML]** §2.3, **[MLAPP]** §4.3, 4.4, and **[CS229]**

Multivariate Gaussians: Basics

- For **Multivariate Gaussian** distribution $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$, we have
  - **mean** $\mu \in \mathbb{R}^D$
  - **covariance matrix** $\Sigma \in \mathbb{R}^{D \times D}$
  - **PDF**

$$\mathcal{N}(\mathbf{x}|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^D \det(\Sigma)}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)\right]$$

Partitioned Gaussian Distributions

- Partition $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$ as

$$\mathbf{x} = \begin{bmatrix} \mathbf{x_a} \\ \mathbf{x_b} \end{bmatrix} \quad \mu = \begin{bmatrix} \mu_\mathbf{a} \\ \mu_\mathbf{b} \end{bmatrix} \quad \Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}$$

  such that $\mathbf{x_a}, \mu_\mathbf{a} \in \mathbb{R}^{D_a}$, $\mathbf{x_b}, \mu_\mathbf{b} \in \mathbb{R}^{D_b}$, $\Sigma_{aa} \in \mathbb{R}^{D_a \times D_a}$ and $\Sigma_{bb} \in \mathbb{R}^{D_b \times D_b}$ for some D_a and D_b ($D_a + D_b = D$).
- Then we could have marginals

$$\boxed{\begin{aligned} \mathbf{x_a} &\sim \mathcal{N}(\mu_\mathbf{a}, \Sigma_{aa}) \\ \mathbf{x_b} &\sim \mathcal{N}(\mu_\mathbf{b}, \Sigma_{bb}) \end{aligned}}$$

- Proof is in the **notes**

Remark—Proof for partitioned Gaussian Distribution

- For covariance matrix $\Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}$, define $\Lambda = \Sigma^{-1} = \begin{bmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{bmatrix}$
- **Fact 1** Based on blockwise inversion of matrix (https://en.wikipedia.org/wiki/Invertible_matrix#Blockwise_inversion), we have

$$\begin{aligned} \Lambda_{bb} &= (\Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab})^{-1} \\ \Lambda_{ba} &= -\Lambda_{bb}\Sigma_{ba}\Sigma_{aa}^{-1} \\ \Sigma_{aa} &= (\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})^{-1} \end{aligned}$$

- **Fact 2** Based on determinant of block matrix (https://en.wikipedia.org/wiki/Determinant#Block_matrices), we have

$$\begin{aligned} \det(\Sigma) &= \det\left(\begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}\right) \\ &= \det(\Sigma_{aa})\det(\Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}) \\ &= \det(\Sigma_{aa})\det(\Lambda_{bb}^{-1}) \end{aligned}$$

- The **joint PDF** is

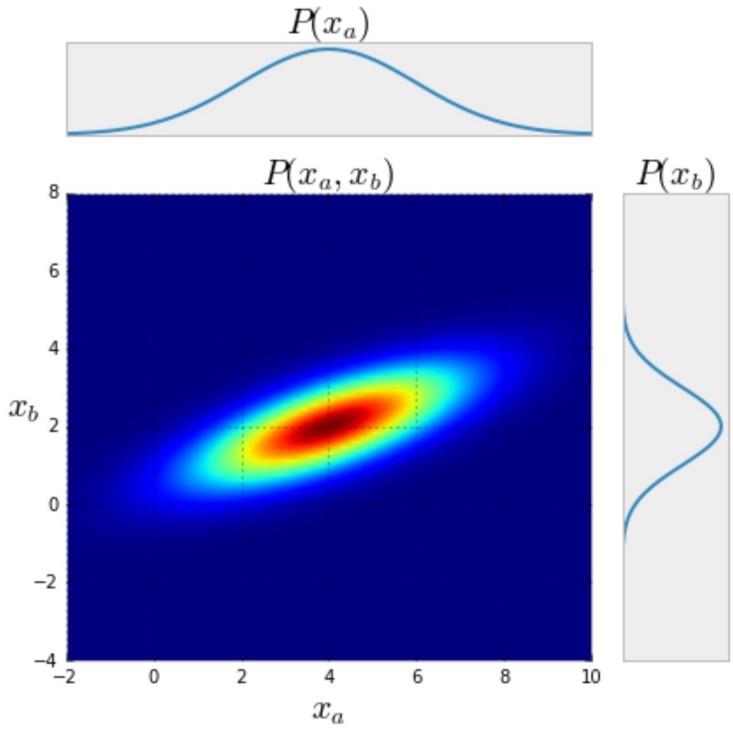
$$\begin{aligned} & P(\mathbf{x_a}, \mathbf{x_b}) \\ &= \frac{1}{(2\pi)^{(D_a+D_b)/2}} \frac{1}{(\det \Sigma)^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} \begin{bmatrix} (\mathbf{x_a} - \mu_\mathbf{a})^T & (\mathbf{x_b} - \mu_\mathbf{b})^T \end{bmatrix} \begin{bmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{bmatrix} \begin{bmatrix} \mathbf{x_a} - \mu_\mathbf{a} \\ \mathbf{x_b} - \mu_\mathbf{b} \end{bmatrix} \right\} \\ &= \frac{1}{(2\pi)^{D_a/2}} \frac{1}{\det(\Sigma_{aa})^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} (\mathbf{x_a} - \mu_\mathbf{a})^T \Sigma_{aa}^{-1} (\mathbf{x_a} - \mu_\mathbf{a}) \right\} \cdot \\ & \quad \frac{1}{(2\pi)^{D_b/2}} \frac{1}{\det(\Lambda_{bb}^{-1})^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} \Big[(\mathbf{x_a} - \mu_\mathbf{a})^T (\Lambda_{aa} - \Sigma_{aa}^{-1})(\mathbf{x_a} - \mu_\mathbf{a}) + \right. \\ & \quad \left. (\mathbf{x_b} - \mu_\mathbf{b})^T \Lambda_{bb} (\mathbf{x_b} - \mu_\mathbf{b}) + 2(\mathbf{x_a} - \mu_\mathbf{a})^T \Lambda_{ab} (\mathbf{x_b} - \mu_\mathbf{b}) \Big) \Big] \right\} \qquad (\text{According to } \mathbf{Fact\ 2}) \\ &= \mathcal{N}(\mathbf{x_a} | \mu_\mathbf{a}, \Sigma_{aa}) \cdot \frac{1}{(2\pi)^{D_b/2}} \frac{1}{\det(\Lambda_{bb}^{-1})^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} \Big[(\mathbf{x_a} - \mu_\mathbf{a})^T \Lambda_{ab} \Lambda_{bb}^{-1} \Lambda_{ba} (\mathbf{x_a} - \mu_\mathbf{a}) + \right. \\ & \quad \left. (\mathbf{x_b} - \mu_\mathbf{b})^T \Lambda_{bb} (\mathbf{x_b} - \mu_\mathbf{b}) + 2(\mathbf{x_a} - \mu_\mathbf{a})^T \Lambda_{ab} (\mathbf{x_b} - \mu_\mathbf{b}) \Big) \Big] \right\} \qquad (\text{According to } \mathbf{Fact\ 1}) \\ &= \mathcal{N}(\mathbf{x_a} | \mu_\mathbf{a}, \Sigma_{aa}) \cdot \frac{1}{(2\pi)^{D_b/2}} \frac{1}{\det(\Lambda_{bb}^{-1})^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} \Big[(\mathbf{x_a} - \mu_\mathbf{a})^T \Lambda_{ab} \Lambda_{bb}^{-1} \Lambda_{bb} \Lambda_{bb}^{-1} \Lambda_{ba} (\mathbf{x_a} - \mu_\mathbf{a}) + \right. \\ & \quad \left. (\mathbf{x_b} - \mu_\mathbf{b})^T \Lambda_{bb} (\mathbf{x_b} - \mu_\mathbf{b}) + 2(\mathbf{x_a} - \mu_\mathbf{a})^T \Lambda_{ab} \Lambda_{bb}^{-1} \Lambda_{bb} (\mathbf{x_b} - \mu_\mathbf{b}) \Big) \Big] \right\} \\ &= \mathcal{N}(\mathbf{x_a} | \mu_\mathbf{a}, \Sigma_{aa}) \cdot \\ & \quad \frac{1}{(2\pi)^{D_b/2}} \frac{1}{\det(\Lambda_{bb}^{-1})^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} [\mathbf{x_b} - \mu_\mathbf{b} + \Lambda_{bb}^{-1}\Lambda_{ba}(\mathbf{x_a} - \mu_\mathbf{a})]^T \Lambda_{bb} [\mathbf{x_b} - \mu_\mathbf{b} + \Lambda_{bb}^{-1}\Lambda_{ba}(\mathbf{x_a} - \mu_\mathbf{a})] \right\} \\ &= \boxed{\mathcal{N}(\mathbf{x_a} | \mu_\mathbf{a}, \Sigma_{aa}) \mathcal{N}(\mathbf{x_b} | \mu_\mathbf{b} - \Lambda_{bb}^{-1}\Lambda_{ba}(\mathbf{x_a} - \mu_\mathbf{a}), \Lambda_{bb}^{-1})} \end{aligned}$$

- So the **marginal PDF** is

$$\begin{aligned} P(\mathbf{x_a}) &= \int P(\mathbf{x_a}, \mathbf{x_b})\, \mathrm{d}\mathbf{x_b} \\ &= \mathcal{N}(\mathbf{x_a}|\mu_\mathbf{a}, \Sigma_{aa}) \int \mathcal{N}(\mathbf{x_b}|\mu_\mathbf{b} - \Lambda_{bb}^{-1}\Lambda_{ba}(\mathbf{x_a} - \mu_\mathbf{a}), \Lambda_{bb}^{-1})\, \mathrm{d}\mathbf{x_b} \\ &= \mathcal{N}(\mathbf{x_a}|\mu_\mathbf{a}, \Sigma_{aa}) \cdot 1 \\ &= \boxed{\mathcal{N}(\mathbf{x_a}|\mu_\mathbf{a}, \Sigma_{aa})} \end{aligned}$$

- Similarly, we could prove $\boxed{P(\mathbf{x_a}) = \mathcal{N}(\mathbf{x_b}|\mu_\mathbf{b}, \Sigma_{bb})}$ $\mathbf{Q.E.D.}$

Partitioned Marginals: Bivariate Examples



Partitioned Conditionals

- Given the setting of partition defined above, we have the **conditionals**

$$\boxed{\begin{aligned} P(\mathbf{x_b}|\mathbf{x_a}) &= \mathcal{N}(\mathbf{x_b}|\mu_{b|a}, \Sigma_{b|a}) \\ \Sigma_{b|a} &= \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab} \\ \mu_{b|a} &= \mu_b + \Sigma_{ba}\Sigma_{aa}^{-1}(\mathbf{x_a} - \mu_a) \end{aligned}} \qquad \boxed{\begin{aligned} P(\mathbf{x_a}|\mathbf{x_b}) &= \mathcal{N}(\mathbf{x_a}|\mu_{a|b}, \Sigma_{a|b}) \\ \Sigma_{a|b} &= \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba} \\ \mu_{a|b} &= \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(\mathbf{x_b} - \mu_b) \end{aligned}}$$

- Proof is in the **notes**

Remark

- The **conditional** distribution of $\mathbf{x_a}|\mathbf{x_b}$ means the distribution of $\mathbf{x_a}$ given the value of $\mathbf{x_b}$.
- For the following notes, we will use $\mathbf{x_a}|\mathbf{x_b}$ to denote "**random variable $\mathbf{x_a}$ given the value of $\mathbf{x_b}$**"
- For $\mathbf{x_a}|\mathbf{x_b}$,
  - Its **Mean** $\mu_{a|b}$ is a linear function w.r.t. $\mathbf{x_b}$.
  - Its **Covariance** $\Sigma_{a|b}$ is constant no matter what $\mathbf{x_b}$ is.

- **Proof** for partitioned conditionals
  - From previous **Remark**, we already showed

$$P(\mathbf{x_a}, \mathbf{x_b}) = \mathcal{N}(\mathbf{x_a}|\mu_\mathbf{a}, \Sigma_{aa})\mathcal{N}(\mathbf{x_b}|\mu_\mathbf{b} - \Lambda_{bb}^{-1}\Lambda_{ba}(\mathbf{x_a} - \mu_\mathbf{a}), \Lambda_{bb}^{-1})$$
$$P(\mathbf{x_a}) = \mathcal{N}(\mathbf{x_a}|\mu_\mathbf{a}, \Sigma_{aa})$$

  So the conditional is

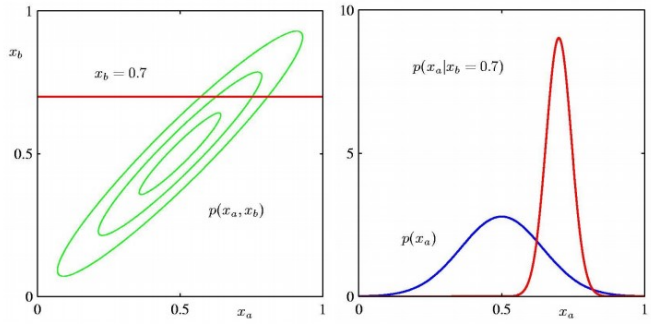
$$P(\mathbf{x_b}|\mathbf{x_a}) = \frac{P(\mathbf{x_a}, \mathbf{x_b})}{P(\mathbf{x_a})} = \mathcal{N}(\mathbf{x_b}|\mu_\mathbf{b} - \Lambda_{bb}^{-1}\Lambda_{ba}(\mathbf{x_a} - \mu_\mathbf{a}), \Lambda_{bb}^{-1})$$

  Based on results from **Fact 1** in previous **Remark**, we have

$$P(\mathbf{x_b}|\mathbf{x_a}) = \mathcal{N}(\mathbf{x_b}|\underbrace{\mu_b + \Sigma_{ba}\Sigma_{aa}^{-1}(\mathbf{x_a} - \mu_a)}_{\mu_{b|a}}, \underbrace{\Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}}_{\Sigma_{b|a}})$$

  Similarly we could prove $P(\mathbf{x_a}|\mathbf{x_b})$ case. $\mathbf{Q.E.D.}$

- Illustration of conditionals



- The conditional is obtained by "slicing" the joint PDF.

Linear Gaussian Systems: Model

- **Just then** we have just showed if we partition a multivariate Gaussian into two vectors

$$\begin{bmatrix}\mathbf{x_a} \\ \mathbf{x_b}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mu_\mathbf{a} \\ \mu_\mathbf{b}\end{bmatrix}, \begin{bmatrix}\Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb}\end{bmatrix}\right)$$

  we have the following conditional with mean a **linear** function w.r.t. $\mathbf{x_b}$

$$P(\mathbf{x_a}|\mathbf{x_b}) = \mathcal{N}(\mathbf{x_a}|\mu_{a|b}, \Sigma_{a|b}) \quad \mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(\mathbf{x_b} - \mu_b) \quad \Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$$

- **Now** let's consider a **different** but **similar** setting:
  - We have a **Gaussian** $\mathbf{x} \in \mathbb{R}^{D_X}$ and **some** $\mathbf{y} \in \mathbb{R}^{D_Y}$ of which

$$\mathbf{x} \sim \mathcal{N}(\mu_x, \Sigma_x) \quad (\mu_\mathbf{x} \text{ and } \Sigma_\mathbf{x} \text{ are } \mathbf{known})$$

    And the conditional relation between $\mathbf{x}$ and $\mathbf{y}$ is

$$\mathbf{y}|\mathbf{x} \sim \mathcal{N}(A\mathbf{x} + \mathbf{b}, \Sigma_{\mathbf{y}|\mathbf{x}}) \quad (A, b \text{ and } \Sigma_{\mathbf{y}|\mathbf{x}} \text{ are } \mathbf{known})$$

    which says $\mathbf{y}|\mathbf{x}$ is also a Gaussian with mean $\mu_{\mathbf{y}|\mathbf{x}}$ a **linear** function of $\mathbf{x}$
- **Then**, what is the distribution of $\mathbf{y}$ and what is the distribution of $\mathbf{x}|\mathbf{y}$?
  - Are they also **Gaussian**?
  - If so, what are the mean and covariance?

Linear Gaussian Systems: Bayes' Theorem

- Actually, given

$$\mathbf{x} \sim \mathcal{N}(\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}}) \quad \text{and} \quad \mathbf{y}|\mathbf{x} \sim \mathcal{N}(A\mathbf{x} + \mathbf{b}, \Sigma_{\mathbf{y}|\mathbf{x}})$$

  we could show

$$\boxed{\mathbf{y} \sim \mathcal{N}(A\mu_{\mathbf{x}} + \mathbf{b}, \Sigma_{\mathbf{y}|\mathbf{x}} + A\Sigma_{\mathbf{x}}A^T)}$$

  and

$$\boxed{\mathbf{x}|\mathbf{y} \sim \mathcal{N}(\mu_{\mathbf{x}|\mathbf{y}}, \Sigma_{\mathbf{y}|\mathbf{x}})}$$

  of which

$$\Sigma_{\mathbf{x}|\mathbf{y}} = (\Sigma_{\mathbf{x}}^{-1} + A^T\Sigma_{\mathbf{y}|\mathbf{x}}^{-1}A)^{-1}$$
$$\mu_{\mathbf{x}|\mathbf{y}} = \Sigma_{\mathbf{x}|\mathbf{y}}\left[A^T\Sigma_{\mathbf{y}|\mathbf{x}}^{-1}(\mathbf{y} - b) + \Sigma_{\mathbf{x}}^{-1}\mu_{\mathbf{x}}\right]$$

- Proof is in the **notes**
- This theorem will play a **key role** throughout this lecture.

Remark

- **Proof for $\mathbf{y}$ (*Less* rigorous one)**
  - Let's assume (less rigorous because here we directly assume $\mathbf{y}$ to have Gaussian distribution)

$$\mathbf{y} \sim \mathcal{N}(\mu_{\mathbf{y}}, \Sigma_{\mathbf{y}})$$

    So based on conditional of partitioned Gaussian, we have

$$\mathbf{y}|\mathbf{x} \sim \mathcal{N}(\mu_{\mathbf{y}} + \Sigma_{\mathbf{yx}}\Sigma_{\mathbf{x}}^{-1}(\mathbf{x} - \mu_{\mathbf{x}}), \Sigma_{\mathbf{y}} - \Sigma_{\mathbf{yx}}\Sigma_{\mathbf{x}}^{-1}\Sigma_{\mathbf{xy}})$$

    Since we have known $\mathbf{y}|\mathbf{x} \sim \mathcal{N}(A\mathbf{x} + \mathbf{b}, \Sigma_{\mathbf{y}|\mathbf{x}})$, we could have the following equations

$$\begin{cases} \Sigma_{\mathbf{yx}}\Sigma_{\mathbf{x}}^{-1} = A \\ \mu_{\mathbf{y}} - \Sigma_{\mathbf{yx}}\Sigma_{\mathbf{x}}^{-1}\mu_{\mathbf{x}} = \mathbf{b} \\ \Sigma_{\mathbf{y}} - \Sigma_{\mathbf{yx}}\Sigma_{\mathbf{x}}^{-1}\Sigma_{\mathbf{xy}} = \Sigma_{\mathbf{y}|\mathbf{x}} \end{cases} \Rightarrow \begin{cases} \Sigma_{\mathbf{yx}} = A\Sigma_{\mathbf{x}},\ \Sigma_{\mathbf{xy}} = \Sigma_{\mathbf{x}}A^T \\ \mu_{\mathbf{y}} = A\mu_{\mathbf{x}} + \mathbf{b} \\ \Sigma_{\mathbf{y}} = \Sigma_{\mathbf{y}|\mathbf{x}} + A\Sigma_{\mathbf{x}}A^T \end{cases} \Rightarrow \mathbf{y} \sim \mathcal{N}(A\mu_{\mathbf{x}} + \mathbf{b}, \Sigma_{\mathbf{y}|\mathbf{x}} + A\Sigma_{\mathbf{x}}A^T)$$

- **Proof for $\mathbf{x}|\mathbf{y}$**
  - We already have

$$\begin{cases} \mathbf{x} \sim \mathcal{N}(\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}}) \\ \mathbf{y} \sim \mathcal{N}(\underbrace{A\mu_{\mathbf{x}} + \mathbf{b}}_{\mu_{\mathbf{y}}}, \underbrace{\Sigma_{\mathbf{y}|\mathbf{x}} + A\Sigma_{\mathbf{x}}A^T}_{\Sigma_{\mathbf{y}}}) \\ \Sigma_{\mathbf{yx}} = A\Sigma_{\mathbf{x}},\ \Sigma_{\mathbf{xy}} = \Sigma_{\mathbf{x}}A^T \end{cases}$$

    According to conditional of partitioned Gaussian, we know $\mathbf{x}|\mathbf{y}$ has Gaussian distribution, so we could assume

$$\mathbf{x}|\mathbf{y} \sim \mathcal{N}(\mu_{\mathbf{x}|\mathbf{y}}, \Sigma_{\mathbf{x}|\mathbf{y}})$$

    Applying the result of conditional of partitioned Gaussian, we have covariance:

$$\begin{aligned} \Sigma_{\mathbf{x}|\mathbf{y}} &= \Sigma_{\mathbf{x}} - \Sigma_{\mathbf{xy}}\Sigma_{\mathbf{y}}^{-1}\Sigma_{\mathbf{yx}} \\ &= \Sigma_{\mathbf{x}} - \Sigma_{\mathbf{x}}A^T(\Sigma_{\mathbf{y}|\mathbf{x}} + A\Sigma_{\mathbf{x}}A^T)^{-1}A\Sigma_{\mathbf{x}} \\ &= \boxed{(\Sigma_{\mathbf{x}}^{-1} + A^T\Sigma_{\mathbf{y}|\mathbf{x}}^{-1}A)^{-1}} \end{aligned}$$

    of which the last equation is based on matrix inversion lemma.

- **Proof for $\mathbf{x}|\mathbf{y}$**
  - And we have mean:

$$\begin{aligned}\mu_{\mathbf{x}|\mathbf{y}} &= \mu_{\mathbf{x}} + \Sigma_{\mathbf{xy}}\Sigma_{\mathbf{y}}^{-1}(\mathbf{y} - \mu_{\mathbf{y}}) \\ &= \mu_{\mathbf{x}} + \Sigma_{\mathbf{xy}}\Sigma_{\mathbf{y}}^{-1}(\mathbf{y} - A\mu_{\mathbf{x}} - \mathbf{b}) \\ &= \Sigma_{\mathbf{xy}}\Sigma_{\mathbf{y}}^{-1}(\mathbf{y} - \mathbf{b}) + (I - \Sigma_{\mathbf{xy}}\Sigma_{\mathbf{y}}^{-1}A)\mu_{\mathbf{x}}\end{aligned}$$

  of which

$$\begin{aligned}\Sigma_{\mathbf{xy}}\Sigma_{\mathbf{y}}^{-1} &= \Sigma_{\mathbf{x}}A^T(\Sigma_{\mathbf{y}|\mathbf{x}} + A\Sigma_{\mathbf{x}}A^T)^{-1} \\ &= \Sigma_{\mathbf{x}|\mathbf{y}}[\Sigma_{\mathbf{x}|\mathbf{y}}^{-1}\Sigma_{\mathbf{x}}A^T(\Sigma_{\mathbf{y}|\mathbf{x}} + A\Sigma_{\mathbf{x}}A^T)^{-1}] \\ &= \Sigma_{\mathbf{x}|\mathbf{y}}[(\Sigma_{\mathbf{x}} + A^T\Sigma_{\mathbf{y}|\mathbf{x}}^{-1}A)\Sigma_{\mathbf{x}}A^T(\Sigma_{\mathbf{y}|\mathbf{x}} + A\Sigma_{\mathbf{x}}A^T)^{-1}] \\ &= \Sigma_{\mathbf{x}|\mathbf{y}}[(A^T + A^T\Sigma_{\mathbf{y}|\mathbf{x}}^{-1}A\Sigma_{\mathbf{x}}A^T)(\Sigma_{\mathbf{y}|\mathbf{x}} + A\Sigma_{\mathbf{x}}A^T)^{-1}] \\ &= \Sigma_{\mathbf{x}|\mathbf{y}}A^T[(I + \Sigma_{\mathbf{y}|\mathbf{x}}^{-1}A\Sigma_{\mathbf{x}}A^T)(\Sigma_{\mathbf{y}|\mathbf{x}} + A\Sigma_{\mathbf{x}}A^T)^{-1}] \\ &= \Sigma_{\mathbf{x}|\mathbf{y}}A^T\Sigma_{\mathbf{y}|\mathbf{x}}^{-1}[(\Sigma_{\mathbf{y}|\mathbf{x}} + A\Sigma_{\mathbf{x}}A^T)(\Sigma_{\mathbf{y}|\mathbf{x}} + A\Sigma_{\mathbf{x}}A^T)^{-1}] \\ &= \Sigma_{\mathbf{x}|\mathbf{y}}A^T\Sigma_{\mathbf{y}|\mathbf{x}}^{-1}\end{aligned}$$

- **Proof for $\mathbf{x}|\mathbf{y}$**

$$\begin{aligned}I - \Sigma_{\mathbf{xy}}\Sigma_{\mathbf{y}}^{-1}A &= I - \Sigma_{\mathbf{x}|\mathbf{y}}A^T\Sigma_{\mathbf{y}|\mathbf{x}}^{-1}A \\ &= \Sigma_{\mathbf{x}|\mathbf{y}}(\Sigma_{\mathbf{x}|\mathbf{y}}^{-1} - A^T\Sigma_{\mathbf{y}|\mathbf{x}}^{-1}A) \\ &= \Sigma_{\mathbf{x}|\mathbf{y}}(\Sigma_{\mathbf{x}}^{-1} + A^T\Sigma_{\mathbf{y}|\mathbf{x}}^{-1}A - A^T\Sigma_{\mathbf{y}|\mathbf{x}}^{-1}A) \\ &= \Sigma_{\mathbf{x}|\mathbf{y}}\Sigma_{\mathbf{x}}^{-1}\end{aligned}$$

  Plug this back to the mean,

$$\begin{aligned}\mu_{\mathbf{x}|\mathbf{y}} &= \Sigma_{\mathbf{x}|\mathbf{y}}A^T\Sigma_{\mathbf{y}|\mathbf{x}}^{-1}(\mathbf{y} - \mathbf{b}) + \Sigma_{\mathbf{x}|\mathbf{y}}\Sigma_{\mathbf{x}}^{-1}\mu_{\mathbf{x}} \\ &= \boxed{\Sigma_{\mathbf{x}|\mathbf{y}}[A^T\Sigma_{\mathbf{y}|\mathbf{x}}^{-1}(\mathbf{y} - \mathbf{b}) + \Sigma_{\mathbf{x}}^{-1}\mu_{\mathbf{x}}]} \qquad \mathbf{Q.E.D.}\end{aligned}$$

Bayesian Linear Regression

Taken from **[PRML]** §3.3, **[MLAPP]** §7.6

Review: Regression

- In last lecture, we have the model for single data point $(\mathbf{x}, t)$

$$\begin{aligned}t &= \mathbf{w}^T\phi(\mathbf{x}) + \epsilon \\ \epsilon &\sim \mathcal{N}(0, \beta^{-1}) \\ t \,|\, \mathbf{x}, \mathbf{w} &\sim \mathcal{N}(\mathbf{w}^T\phi(\mathbf{x}), \beta^{-1})\end{aligned}$$

- So for inputs $\mathcal{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and target values $\mathbf{t} = (t_1, \ldots, t_n)$, we have

$$\mathbf{t} \,|\, \mathcal{X}, \mathbf{w} \sim \mathcal{N}(\Phi\mathbf{w}, \beta^{-1}I)$$

- Recall in last lecture, if we have **prior** $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \alpha^{-1}I)$, the **MAP Estimate** of $\mathbf{w}$ corresponds to solution to **ridge regression** (least squares with ℓ^2 regularization)

$$\mathbf{w}_{MAP} = (\Phi^T\Phi + \frac{\alpha}{\beta}I)^{-1}\Phi^T\mathbf{t}$$

- Next, we will see what $\mathbf{w}$'s **posterior** $P(\mathbf{w} \,|\, \mathcal{X}, \mathbf{t})$ is with more general prior.
- Note that $\mathbf{w}_{MAP}$ is just a single point estimate of **posterior** $P(\mathbf{w} \,|\, \mathcal{X}, \mathbf{t})$. More about this will come later.

Remark

- Recall **design matrix** Φ is

$$\Phi = \begin{bmatrix} \phi(\mathbf{x}_1)^T \\ \phi(\mathbf{x}_2)^T \\ \vdots \\ \phi(\mathbf{x}_N)^T \end{bmatrix} = \begin{bmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{bmatrix}$$

Bayesian Linear Regression

- We have **prior** for coefficients $\mathbf{w}$

$$P(\mathbf{w}) = \mathcal{N}(\mathbf{w} \,|\, \mathbf{m}_0, S_0)$$

- And from last slide, we have **likelihood**

$$P(\mathbf{t} \,|\, \mathbf{w}, \mathcal{X}) = \mathcal{N}(\mathbf{t} \,|\, \Phi\mathbf{w}, \beta^{-1} I)$$

- **Bayesian Linear Regression** computes the full **posterior** over the weights (prior),

$$\begin{aligned} P(\mathbf{w} \,|\, \mathbf{t}, \mathcal{X}) &= \frac{P(\mathbf{t} \,|\, \mathbf{w}, \mathcal{X}) P(\mathbf{w} \,|\, \mathcal{X})}{P(\mathbf{t} \,|\, \mathcal{X})} \\ &\propto P(\mathbf{t} \,|\, \mathbf{w}, \mathcal{X}) P(\mathbf{w} \,|\, \mathcal{X}) \\ &\propto \underbrace{P(\mathbf{t} \,|\, \mathbf{w}, \mathcal{X})}_{\text{Likelihood}} \underbrace{P(\mathbf{w})}_{\text{Prior}} \qquad (\text{Drop } \mathcal{X} \text{ because of independence}) \\ &\propto \mathcal{N}(\mathbf{t} \,|\, \Phi\mathbf{w}, \beta^{-1} I) \mathcal{N}(\mathbf{w} \,|\, \mathbf{m}_0, S_0) \\ &= \mathcal{N}(\mathbf{w} \,|\, \mathbf{m}_N, S_N) \qquad (\text{Can be obtained from some normalization}) \end{aligned}$$

  where

$$\boxed{S_N = (S_0^{-1} + \beta \Phi^T \Phi)^{-1} \qquad \mathbf{m}_N = S_N(\beta \Phi^T \mathbf{t} + S_0^{-1} \mathbf{m}_0)}$$

- We will derive this using results we just derived in Bayes' Theorem for Linear Gaussian

Remark

- Generally, we should follow

$$\text{Posterior} = \text{Likelihood} \times \text{Prior}$$

  to compute the posterior. But since we have the results from Bayes' Theorem for Linear Gaussian, we will use directly.
- **Posterior** $P(\mathbf{w} \,|\, \mathbf{t}, \mathcal{X})$ tells us our prior belief in $\mathbf{w}$ with **mean** $\mathbf{m}_0$ and **covariance** S_0 has been updated to **mean** $\mathbf{m}_N$ and **covariance** S_N.
- We will show this change in some plots later.

Bayesian Linear Regression: Derivation

- Following the results in Bayes' Theorem for Linear Gaussian, we have

$$\boxed{\begin{aligned} &\text{Bayes' Theorem of Linear Gaussian} \\ &\mathbf{x} \sim \mathcal{N}(\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}}) \\ &\mathbf{y}|\mathbf{x} \sim \mathcal{N}(A\mathbf{x} + \mathbf{b}, \Sigma_{\mathbf{y}|\mathbf{x}}) \\ &\qquad \Downarrow \\ &\mathbf{x}|\mathbf{y} \sim \mathcal{N}(\mu_{\mathbf{x}|\mathbf{y}}, \Sigma_{\mathbf{y}|\mathbf{x}}) \\ &\Sigma_{\mathbf{x}|\mathbf{y}} = (\Sigma_{\mathbf{x}}^{-1} + A^T \Sigma_{\mathbf{y}|\mathbf{x}}^{-1} A)^{-1} \\ &\mu_{\mathbf{x}|\mathbf{y}} = \Sigma_{\mathbf{x}|\mathbf{y}} \left[A^T \Sigma_{\mathbf{y}|\mathbf{x}}^{-1} (\mathbf{y} - b) + \Sigma_{\mathbf{x}}^{-1} \mu_{\mathbf{x}} \right] \end{aligned}} \Rightarrow \boxed{\begin{aligned} &\text{Posterior of } \mathbf{w} \\ &\mathbf{w}|\mathcal{X} \sim \mathcal{N}(\mathbf{m}_0, S_0) \\ &\mathbf{t}|\mathbf{w}, \mathcal{X} \sim \mathcal{N}(\Phi\mathbf{w}, \beta^{-1} I) \\ &\qquad \Downarrow \\ &\mathbf{w}|\mathbf{t}, \mathcal{X} \sim \mathcal{N}(\mathbf{m}_N, S_N) \\ &S_N = (S_0^{-1} + \beta \Phi^T \Phi)^{-1} \\ &\mathbf{m}_N = S_N(\beta \Phi^T \mathbf{t} + S_0^{-1} \mathbf{m}_0) \end{aligned}}$$

- Done!
- Don't get confused by $\mathcal{X}$ in the condition position. We are free to drop it without any impact.
- Think like a Bayesian. Sometimes we usually drop some priors like "Coin is fair" in $P(\text{Head} \,|\, \text{Coin is fair})$

BLR: Simplifying the Prior

- Assume $\mathbf{m}_0 = 0$ and $S_0 = \alpha^{-1} I$ (like what we did in ridge regression), we have
$$P(\mathbf{w} \mid \mathbf{t}, \mathcal{X}) = \mathcal{N}(\mathbf{w} \mid \mathbf{m}_N, S_N)$$
where
$$S_N = (\alpha I + \beta \Phi^T \Phi)^{-1}$$
$$\mathbf{m}_N = \beta S_N \Phi^T \mathbf{t} = (\Phi^T \Phi + \frac{\alpha}{\beta} I)^{-1} \Phi^T \mathbf{t}$$
- Note that the **posterior mean** $\mathbf{m}_N$ is the **MAP estimate** of $\mathbf{w}$.
- This makes sense because **posterior mean** is exactly the point that maximizes the posterior P(\vw|\vt, \mathcal{X}, \beta).
- **Posterior Covariance** S_N tells us how confident we are in our prediction $\mathbf{w}_{MAP}$ because it tells us how far $\mathbf{w}$ are spread out from its mean $\mathbf{w}_{MAP}$.

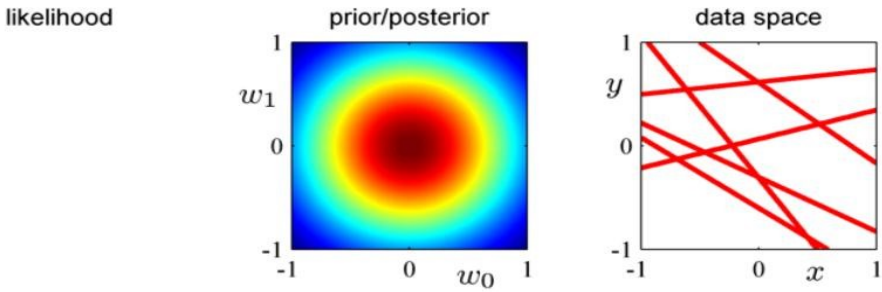
Sequential Bayesian Learning

- Sometimes, our data is *streaming*, and we want to learn $\mathbf{w}$ in a online fashion.
- **Note:** The posterior $P(\mathbf{w} \mid \mathbf{t}, \mathcal{X})$ and prior $P(\mathbf{w})$ are both Gaussians.
  - We can use the posterior for one set of observations as a prior for the next set of observations.
  - And following exactly the same procedures as above, we will get a new posterior .
  - Starting from a fixed prior, sequentially update our beliefs as new data arrives.
- Here are the details
  - **Initialize** $\mathbf{m}_0$ and S_0 of $\mathcal{N}(\mathbf{w} \mid m_0, S_0)$
  - **Repeat** when the ith observation $\{\mathcal{X}_i, \mathbf{t}_i\}$ arrives
    - $S_i = (S_{i-1}^{-1} + \beta \Phi_{\mathcal{X}_i}^T \Phi_{\mathcal{X}_i})^{-1}$
    - $\mathbf{m}_i = S_i(\beta \Phi_{\mathcal{X}_i}^T \mathbf{t}_i + S_{i-1}^{-1} \mathbf{m}_{i-1})$
  - **End**
- This is **Bayesian Updating**.

Sequential Bayesian Learning: Example

- Assume observation y is generated by
$$y = w_1 x + w_0 + \epsilon$$
of which
$$w_1 = 0.5,\ w_0 = -0.3,\ \epsilon \sim \mathcal{N}(0, \beta^{-1})$$
(We use both y and t to denote observations)
- We want to learn the coefficients w_1 and w_0 using sequential Bayesian learning.

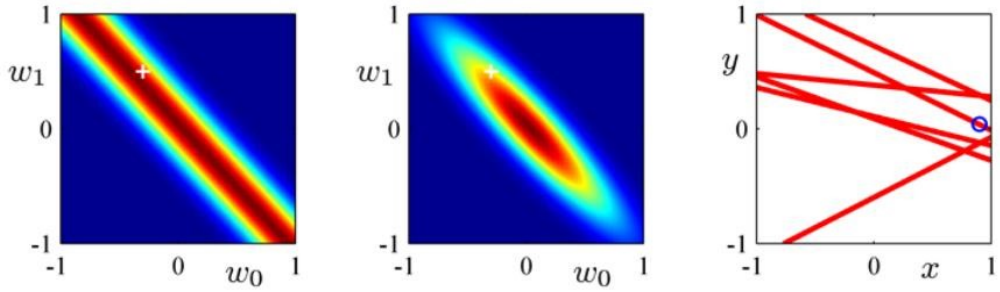
Sequential Bayesian Learning: Example—No Data Comes Yet



- **First Plot**—Nothing
- **Second Plot**—Prior $P(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \mathbf{m}_0, S_0)$
- **Third Plot**—Sample lines of $y = w_1 x + w_0$
  - We draw 6 samples $\mathbf{w}$ based on distribution $\mathcal{N}(\mathbf{w} \mid \mathbf{m}_0, S_0)$ and plot the lines $y = w_1 x + w_0$
  - We could see they are highly scattered and far from the true line $y = 0.5x - 0.3$

Sequential Bayesian Learning: Example—First Observation Arrives

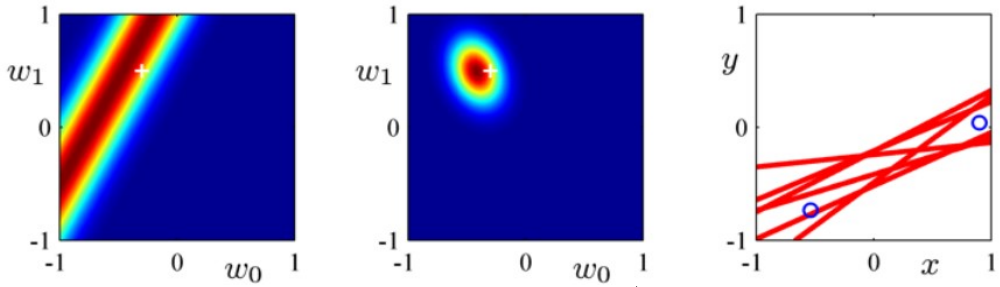
- First observation $\{y_1, x_1\}$ (has been marked with blue circle in third plot)



- **First Plot**—Likelihood $P(y = y_1 \,|\, \mathbf{w}, x_1) = P(y = y_1 \,|\, w_1 x_1 + w_0, \beta^{-1})$
  - Shows the probability of observing y_1 from distribution $\mathcal{N}(\, w_1 x_1 + w_0, \beta^{-1})$ given different values of $\mathbf{w}$.
  - $\mathbf{w}$ from *redder* region are more likely to produce y_1 given x_1
  - True $\mathbf{w} = [-0.3, 0.5]$ (marked with white cross) falls in the orange zone.
- **Second Plot**—Posterior $P(\mathbf{w} \,|\, y_1, x_1) = \mathcal{N}(\mathbf{w} \,|\, m_1, S_1)$
  - Computed from prior $\mathcal{N}(\mathbf{w}|m_0, S_0)$ and the first observation
  - Variance is smaller compared with our prior belief
  - The mean is still far from true $\mathbf{w} = [-0.3, 0.5]$
- **Third Plot**—Sample lines of $y = w_1 x + w_0$ where $\mathbf{w}$ is drawn from $\mathcal{N}(\mathbf{w} \,|\, m_1, S_1)$
  - Still scattered and far from true line $y = 0.5x - 0.3$

Sequential Bayesian Learning: Example—Second Observation Arrives

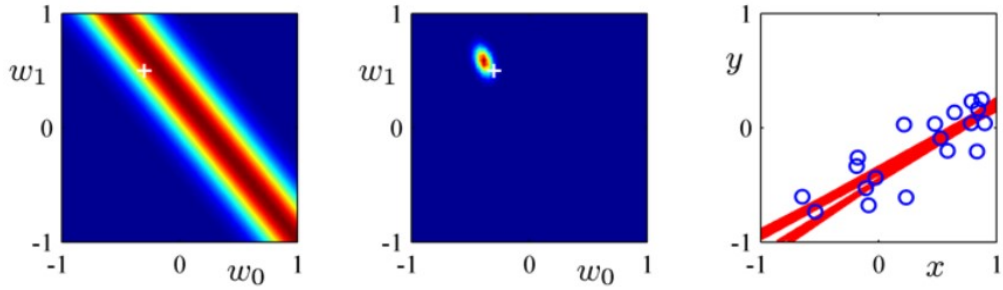
- Second observation $\{y_2, x_2\}$ (has been marked with blue circle in the left corner third plot)



- **First Plot**—Likelihood $P(y = y_2 \,|\, \mathbf{w}, x_2) = P(y = y_2 \,|\, w_1 x_2 + w_0, \beta^{-1})$
  - True $\mathbf{w} = [-0.3, 0.5]$ (marked with white cross) still falls in the orange zone.
- **Second Plot**—Posterior $P(\mathbf{w} \,|\, y_1, x_1, y_2, x_2) = \mathcal{N}(\mathbf{w} \,|\, m_2, S_2)$
  - Computed from posterior $\mathcal{N}(\mathbf{w} \,|\, m_1, S_1)$ and the second observation
  - Variance is even smaller
  - The mean has moved much closer to $\mathbf{w} = [-0.3, 0.5]$
- **Third Plot**—Sample lines of $y = w_1 x + w_0$ where $\mathbf{w}$ is drawn from $\mathcal{N}(\mathbf{w} \,|\, m_2, S_2)$
  - Less scattered and slope is closer to $y = 0.5x - 0.3$

Sequential Bayesian Learning: Example—More Data Arrives

- More observations $\{y_i, x_i\}_{i=3}^N$ arrive (have been marked with blue circle third plot)



- **First Plot**—Likelihood $P(y = y_N \,|\, \mathbf{w}, x_N) = P(y = y_N \,|\, w_1 x_N + w_0, \beta^{-1})$
  - Although for different $\{y_i, x_i\}$, the likelihood plots have different slope, true $\mathbf{w} = [-0.3, 0.5]$ always falls in the highly likely region.
- **Second Plot**—Posterior $P(\mathbf{w} \,|\, \{y_i, x_i\}_{i=1}^N) = \mathcal{N}(\mathbf{w} \,|\, m_N, S_N)$
  - Computed from posterior $\mathcal{N}(\mathbf{w} \,|\, m_{N-1}, S_{N-1})$ and the second observation
  - The iridescent ring almost shrinks to a *point* locates at $[-0.3, 0.5]$
  - This indicates the variance is really small and mean is really close to true value.
- **Third Plot**—Sample lines of $y = w_1 x + w_0$ where $\mathbf{w}$ is drawn from $\mathcal{N}(\mathbf{w} \,|\, m_N, S_N)$
  - The lines almost converge to true $y = 0.5x - 0.3$
- For details of these plots, please refer to **[PRML]**-P154.

Predictive Distribution

- Recall our ultimate goal in linear regression is to predict t for new data $\mathbf{x}$
  - We could take the mean of posterior $\mathbf{m}_N$ as the estimate of $\mathbf{w}$ and get a prediction
$$t = \mathbf{m}_N^T \phi(\mathbf{x})$$
  - **But**, this could be a waste of the posterior of $\mathbf{w}$. We want to know more about prediction t.
  - Actually, we could obtain distribution of t based on the full posterior of $\mathbf{w}$
- Let $\mathcal{D} = \{(t_i, \mathbf{x}_i)\}_{i=1}^N$ denote the training data. Then the **predictive distribution** for some **new data** $\mathbf{x}$ is

$$\begin{aligned} P(t \,|\, \mathbf{x}, \mathcal{D}) &= \int_{\mathbf{w}} P(t, \mathbf{w} \,|\, \mathbf{x}, \mathcal{D}) \; \mathrm{d}\mathbf{w} \\ &= \int_{\mathbf{w}} P(t \,|\, \mathbf{w}, \mathbf{x}, \mathcal{D}) P(\mathbf{w} \,|\, \mathbf{x}, \mathcal{D}) \; \mathrm{d}\mathbf{w} \\ &= \int_{\mathbf{w}} P(t \,|\, \mathbf{w}, \mathbf{x}) P(\mathbf{w} \,|\, \mathcal{D}) \; \mathrm{d}\mathbf{w} \\ &= \int_{\mathbf{w}} \mathcal{N}(t \,|\, \mathbf{w}^T \phi(\mathbf{x}), \beta^{-1}) \mathcal{N}(\mathbf{w} \,|\, \mathbf{m}_N, S_N) \; \mathrm{d}\mathbf{w} \end{aligned}$$

  The third quality holds because t is independent of $\mathcal{D}$ and $\mathbf{w}$ is independent of $\mathbf{x}$.

Predictive Distribution: Derivation

- We don't have to take the pains to do integration!
- The results we derived in Bayes' theorem for linear Gaussian will save us!

Previous Results		Prediction of $\mathbf{w}$
$\mathbf{x} \sim \mathcal{N}(\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}})$		$\mathbf{w}\|\mathbf{x}, \mathcal{D} \sim \mathcal{N}(\mathbf{m}_N, S_N)$
$\mathbf{y}\|\mathbf{x} \sim \mathcal{N}(A\mathbf{x} + \mathbf{b}, \Sigma_{\mathbf{y}\|\mathbf{x}})$	$\Rightarrow$	$t\|\mathbf{w}, \mathbf{x}, \mathcal{D} \sim \mathcal{N}(\phi^T \mathbf{w}, \beta^{-1})$
$\Downarrow$		$\Downarrow$
$\mathbf{y} \sim \mathcal{N}(A\mu_{\mathbf{x}} + b, \Sigma_{\mathbf{y}\|\mathbf{x}} + A\Sigma_{\mathbf{x}} A^T)$		$t \,\|\, \mathbf{x}, \mathcal{D} \sim \mathcal{N}(\, \phi(\mathbf{x})^T \mathbf{m}_n \,,\, \beta^{-1} + \phi(\mathbf{x})^T S_N \phi(\mathbf{x}) \,)$

- So, we have

$$\boxed{P(t \,|\, \mathbf{x}, \mathcal{D}) = \mathcal{N}(t \,|\, \phi(\mathbf{x})^T \mathbf{m}_n \,,\, \beta^{-1} + \phi(\mathbf{x})^T S_N \phi(\mathbf{x}) \,)}$$

- Done!

Predictive Distribution: Derivation

- We don't have to take the pains to do integration!
- The results we derived in Bayes' theorem for linear Gaussian will save us!

$$\boxed{\begin{array}{c}\text{Previous Results}\\ \mathbf{x}\sim\mathcal{N}(\mu_{\mathbf{x}},\Sigma_{\mathbf{x}})\\ \mathbf{y}|\mathbf{x}\sim\mathcal{N}(A\mathbf{x}+\mathbf{b},\Sigma_{\mathbf{y}|\mathbf{x}})\\ \Downarrow\\ \mathbf{y}\sim\mathcal{N}(A\mu_{\mathbf{x}}+b,\Sigma_{\mathbf{y}|\mathbf{x}}+A\Sigma_{\mathbf{x}}A^T)\end{array}}\Rightarrow\boxed{\begin{array}{c}\text{Prediction of }\mathbf{w}\\ \mathbf{w}|\mathbf{x},\mathcal{D}\sim\mathcal{N}(\mathbf{m}_N,S_N)\\ t|\mathbf{w},\mathbf{x},\mathcal{D}\sim\mathcal{N}(\phi^T\mathbf{w},\beta^{-1})\\ \Downarrow\\ t\,|\,\mathbf{x},\mathcal{D}\sim\mathcal{N}(\,\phi(\mathbf{x})^T\mathbf{m}_n\,,\,\beta^{-1}+\phi(\mathbf{x})^TS_N\phi(\mathbf{x})\,)\end{array}}$$

- Done!

Predictive Distribution: Example

- Here the underlying true model we will use is

$$t=\sin(2\pi x)+\epsilon$$

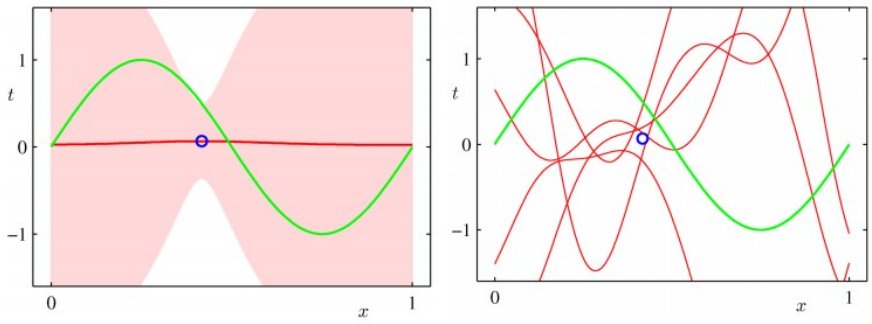
  of which ϵ represents some noise.
- In regression, rather than polynomial basis function we used in previous lectures, we will use 9 Gaussian basis function

$$\phi_j(x)=\exp\left\{-\frac{(x-\mu_j)}{2s^2}\right\}$$

- We will still use streaming fashion.

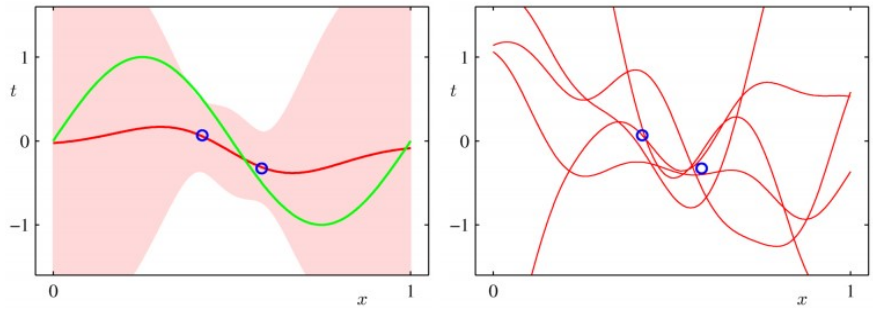
Predictive Distribution: Example—One Observation

- The observations are marked with blue circles in the plots.
- Green curve is the true model $t=\sin(2\pi x)$



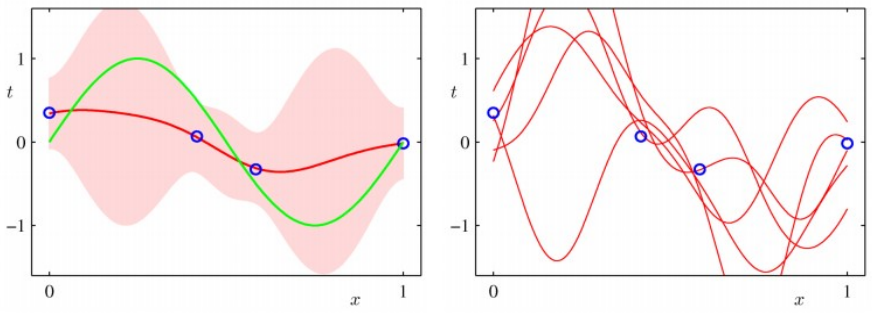
- **Left Plot**
  - Red curve gives the mean value of prediction distribution $P(t|x,\mathcal{D}_1)$, which is $\phi(x)^Tm_1$
  - Red shaded region spans one standard deviation on either side of the mean
  - Note that smallest deviation/variance occurs near observation point
- **Right Plot**
  - We first draw samples of $\mathbf{w}$ from $P(\mathbf{w}|\mathcal{D}_1)$
  - Then plot sample curves $y(x,\mathbf{w})=\mathbf{w}^T\phi(x)$
  - If we draw infinite sample cures, their mean is just the red mean curve in left plot.
- Here $\mathcal{D}_i$ represents the first i observations

Predictive Distribution: Example—Two Observations



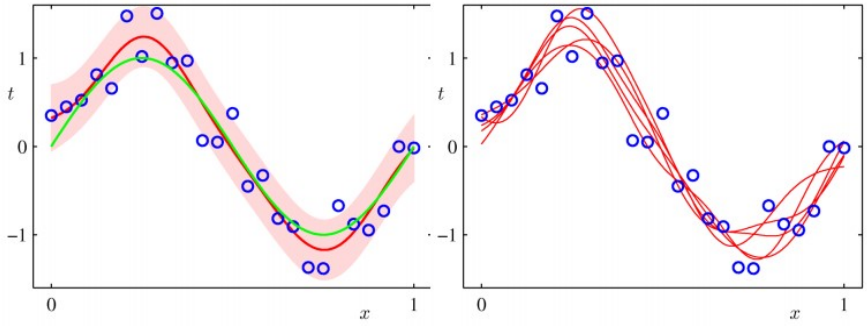
- **Left Plot**
  - Red curve gives the mean value of prediction distribution $P(t|x, \mathcal{D}_2)$, which is $\phi(x)^T m_2$
  - Red Mean Curve starts to resemble the true curve.
  - Red shaded region has shrunk indicating variance becomes smaller
- **Right Plot**
  - We first draw samples of $\mathbf{w}$ from $P(\mathbf{w}|\mathcal{D}_2)$
  - Then plot sample curves $y(x, \mathbf{w}) = \mathbf{w}^T \phi(x)$
  - The sample curves are still quite scattered

Predictive Distribution: Example—Four Observations



- **Left Plot**
  - Red curve gives the mean value of prediction distribution $P(t|x, \mathcal{D}_3)$, which is $\phi(x)^T m_3$
  - Red shaded region has further shrunk indicating smaller variance/uncertainty
  - Red shaded region has small height near observation points
  - Our observations have the effect of reducing the variance at their positions.
- **Right Plot**
  - We first draw samples of $\mathbf{w}$ from $P(\mathbf{w}|\mathcal{D}_2)$
  - Then plot sample curves $y(x, \mathbf{w}) = \mathbf{w}^T \phi(x)$
  - The sample curves are still quite scattered

Predictive Distribution: Example—Twenty Five Observations



- **Left Plot**
  - Red curve gives the mean value of prediction distribution $P(t|x, \mathcal{D}_{25})$, which is $\phi(x)^T m_{25}$
  - Red curve is already very close to the true curve!
  - Red shaded region has shrunk greatly and variance has reduced greatly.
- **Right Plot**
  - We first draw samples of $\mathbf{w}$ from $P(\mathbf{w}|\mathcal{D}_{25})$
  - Then plot sample curves $y(x, \mathbf{w}) = \mathbf{w}^T \phi(x)$
  - The sample curves almost converge to the true curve!
- For details of these plots, please refer to **[PRML]**-P157.

Gaussian Processes

> Taken from **[CS229]** and **[MLAPP]**

- Now we will show you Bayesian linear regression is actually a Gaussian process.

Gaussian Processes

- **Motivation**: Here are some data points. What function did they come from?
  - GPs are a nice way of expressing “priors on functions”
  - Applications: Regression and Classification

Gaussian Processes: Motivation

- **Multivariate Gaussians** are useful for modeling *finite* collections of real-valued variables.
  - Nice analytical properties
  - Distribution over **random vectors**
  - Easily model *correlations* between variables
- **Gaussian Processes** extend Multivariate Gaussians to *infinite-sized* collections of real-valued variables.
  - Distribution over **random functions**

Distributions over Functions: Finite Domain

- How can we parameterize probability distributions over functions?
- Consider the following simple example:
  - Let $\mathcal{X} = \{x_1, \ldots, x_m\}$ be any finite set.
  - Let $\mathcal{H}$ be the set of all functions $h : \mathcal{X} \mapsto \mathbb{R}$.
- For example, one function $h_0 \in \mathcal{H}$ is

$$h_0(x_1) = 5 \quad h_0(x_2) = 2.3 \quad \cdots \quad h_0(x_{m-1}) = -\pi \quad h_0(x_m) = 8$$

- Then h_0 could be represented as a vector $\mathbf{h_0} = [5, 2.3, \ldots, \pi, 8]$
- Any function $h \in \mathcal{H}$ can be represented as a vector.

Distributions over Functions: Finite Domain

- To specify a distribution over $\mathcal{H}$, exploit the one-one mapping to $\mathbb{R}^m$
  - Assume a distribution over vectors, $\mathbf{h} \sim \mathcal{N}(\mu, \sigma^2 I)$.
- This **induces** a distribution over $\mathcal{H}$ given by likelihoods at each *"sample point"*:

$$P(h) = P(\mathbf{h}) = \prod_{k=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(f(x_k) - \mu_k)^2\right]$$

Distributions over Functions: Infinite Domain

- A **stochastic process** is a collection of random variables, $f = \{f(x)\}_{x \in \mathcal{X}}$ with index set $\mathcal{X}$, e.g.
  - Dirichlet Processes, Poisson Processes, etc.
- A **Gaussian Process** is a stochastic process such that any finite subcollection of random variables has a multivariate Gaussian distribution

Gaussian Processes: Definition

- We say $f(\cdot) \sim \mathcal{GP}(m, k)$ is drawn from a Gaussian process with
  - mean function $m(\cdot) : \mathcal{X} \mapsto \mathbb{R}$
  - covariance or kernel function $k(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$

  provided that for any finite set $\{x_1, \ldots, x_m\} \subset \mathcal{X}$, the associated random variables have distribution

$$\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_m) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(x_1) \\ \vdots \\ m(x_m) \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_m) \\ \vdots & \ddots & \vdots \\ k(x_m, x_1) & \cdots & k(x_m, x_m) \end{bmatrix} \right)$$

Gaussian Processes: Interpretation

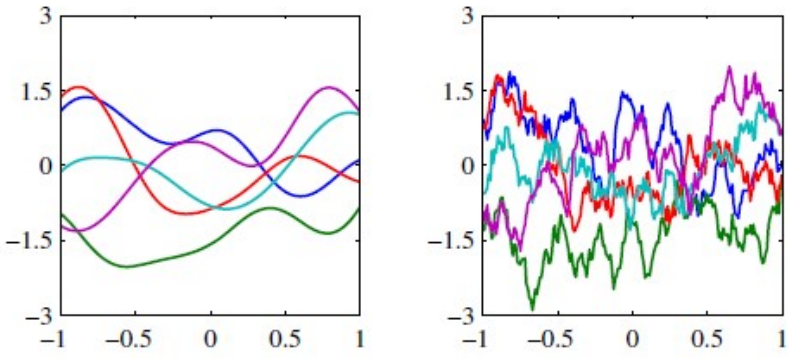
- Intuitively, $f \sim \mathcal{GP}(m, k)$ is an
  - extremely high-dimensional vector
  - drawn from an extremely high-dimensional Gaussian
- Each dimension corresponds to an element $x \in \mathcal{X}$,
  - the corresponding component of the vector represents $f(x)$

Gaussian Processes: Mean and Covariance

- The **mean function** $m(\cdot) : \mathcal{X} \mapsto \mathbb{R}$ can be any function.
  - For most applications, we set $m \equiv 0$.
- The **covariance function** $k(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ must be a valid kernel.
  - $f(x)$ and $f(x')$ will have high covariance if x and x' are "nearby"
  - Therefore, kernel controls **smoothness**

Gaussian Processes: Example

- When $m \equiv 0$, the choice of kernel defines the prior.
  - (Left) Gaussian Kernel $k(x, x') = \exp(-\theta ||x - x'||_2^2)$
  - (Right) Exponential Kernel $k(x, x') = \exp(-\theta ||x - x'||_1)$
- Samples from a Gaussian Process:



Linear Regression Revisited

- **Model:** Assume $y \approx w^T \phi(x)$ is a combination of M fixed basis functions.

$$w \sim \mathcal{N}(w_0, S_0)$$
$$y | x, w, \beta \sim \mathcal{N}(w^T \phi(x_n), \beta^{-1})$$

- Given training points $x_1, \ldots, x_N$, what is the joint distribution $P(\mathbf{y})$ of $y(x_1), \ldots, y(x_N)$?

$$\mathbf{y} = \Phi w = \begin{bmatrix} y(x_1) & \cdots & y(x_N) \end{bmatrix}^T$$

Linear Regression Revisited

- Note $\mathbf{y} = \Phi w$ is a linear combination of Gaussians w, so is itself Gaussian!

$$\mathrm{E}[\mathbf{y}] = \Phi E[w] = 0$$

$$\mathrm{Cov}[\mathbf{y}] = \mathrm{E}[yy^T] = \Phi \mathrm{E}[ww^T]\Phi^T = \frac{1}{\alpha}\Phi\Phi^T = K$$

- where $K = [k(x_i, x_j)]_{i,j} \in \mathbb{R}^{N \times N}$ is the **Gram Matrix** over the training data with kernel

$$k(x_i, x_j) = \frac{1}{\alpha}\phi(x_i)^T\phi(x_j)$$

Bayesian Linear Regression

- So, Bayesian Kernel Linear Regression is a Gaussian Process!
  - Kernel $k(\cdot, \cdot)$ is dot product in feature space.

$$y = f(x) + \epsilon$$
$$f \sim \mathcal{GP}(0, k(\cdot, \cdot))$$
$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

> Features in BLR $\iff$ Kernel functions for GPs

- In general, $k(\cdot, \cdot)$ can be any valid kernel,
- See the book for more details.