Introduction to Probabilistic Machine Learning

Ralf Herbrich

Bayesian Regression

Overview

1. Bayesian Linear Regression
2. Bayesian Linear Regression via Message Passing
   - Normal Distribution Revisited
   - Posterior and Predictive Distribution
3. Fast Bayesian Linear Regression
4. Bayesian Linear Regression via Linear Algebra

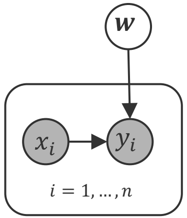
Overview

1. **Bayesian Linear Regression**
2. Bayesian Linear Regression via Message Passing
   - Normal Distribution Revisited
   - Posterior and Predictive Distribution
3. Fast Bayesian Linear Regression
4. Bayesian Linear Regression via Linear Algebra

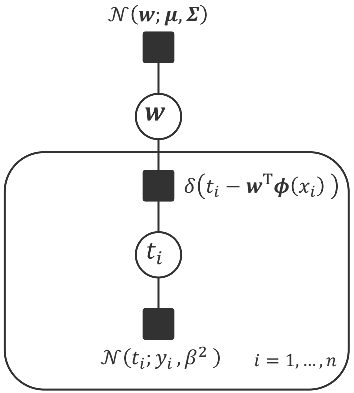
Bayesian Inference of Linear Basis Function Models

- **Given**:
  1. **Training Data**: $D \in (\mathcal{X} \times \mathbb{R})^n$ of n (labelled) examples (x_i, y_i)
  2. **Linear Basis Functions**: Basis function mapping $\boldsymbol{\phi}: \mathcal{X} \to \mathbb{R}^M$ and linear function model
     $$f(x; \mathbf{w}) := \underbrace{\mathbf{w}}_{\text{weight vector}}{}^{\mathrm{T}} \underbrace{\boldsymbol{\phi}(x)}_{\text{feature vector}}$$
  3. **Likelihood of functions**:
     $$p(D|f) = p(D|\mathbf{w}) = \prod_{i=1}^{n} \mathcal{N}\left(y_i; \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(x_i), \beta^2\right)$$
  4. **Prior belief over functions**:
     $$p(f) = p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Bayesian Network



Factor Graph



- **Bayesian Inference**:
  - **Posterior belief over functions**:

$$p(f|D) = p(\mathbf{w}|D) = \frac{\prod_{i=1}^{n} \mathcal{N}\left(y_i; \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(x_i), \beta^2\right) \cdot \mathcal{N}(\mathbf{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\int_{\mathbb{R}^M} \prod_{i=1}^{n} \mathcal{N}\left(y_i; \tilde{\mathbf{w}}^{\mathrm{T}} \boldsymbol{\phi}(x_i), \beta^2\right) \cdot \mathcal{N}(\tilde{\mathbf{w}}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\tilde{\mathbf{w}}}$$

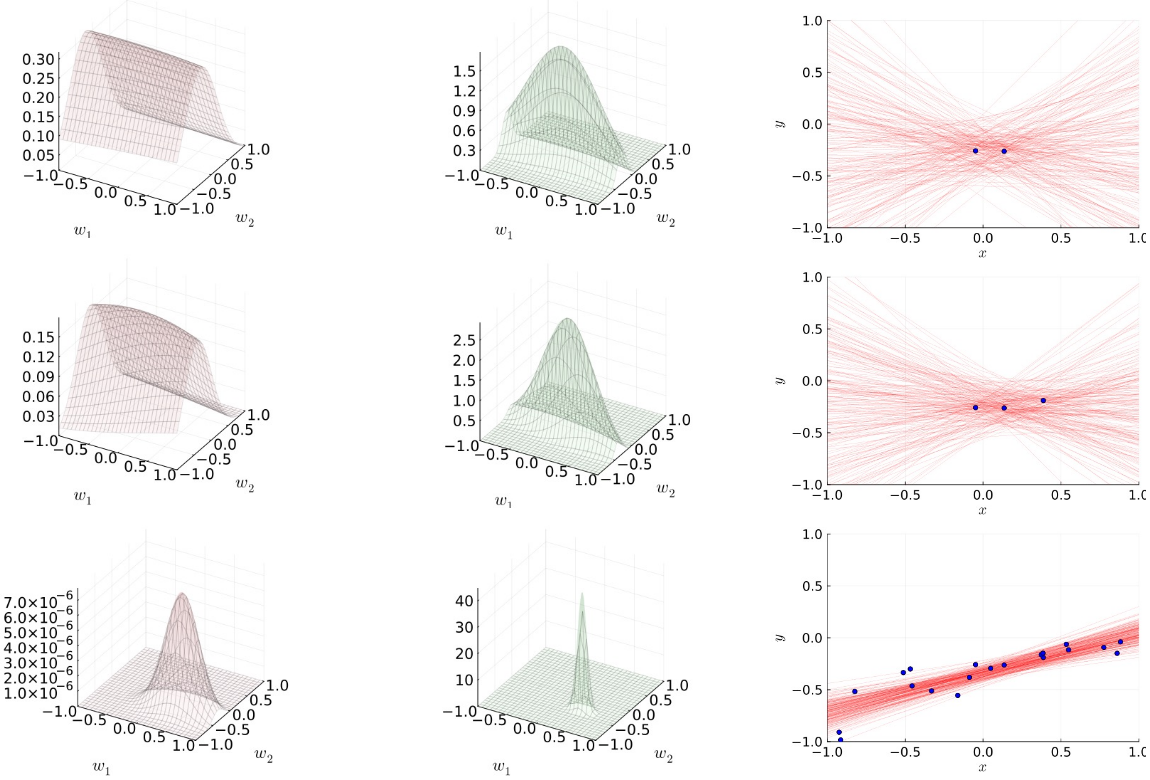
Bayesian Inference in Pictures

Likelihood **Posterior** **Input Space**

$n = 2$

$n = 3$

$n = 20$



$f(x) = w_1 x + w_2$

$P(y|x) = \mathcal{N}(y; f(x), 0.2^2)$

$P(w_j) = \mathcal{N}(w_j; 0,0.5)$

Introduction to Probabilistic Machine Learning

Unit 7 – Bayesian Regression

Overview

1. Bayesian Linear Regression
2. **Bayesian Linear Regression via Message Passing**
   - **Normal Distribution Revisited**
   - Posterior and Predictive Distribution
3. Fast Bayesian Linear Regression
4. Bayesian Linear Regression via Linear Algebra

Multivariate Normal Distribution

- **Multivariate Normal Distribution**. *A continuous random variable $\boldsymbol{X} \in \mathbb{R}^M$ is said to have a multivariate normal distribution if the density is given by*

$$p(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^M |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right)$$

  where $\boldsymbol{\Sigma}$ must be a positive definite $M \times M$ matrix.

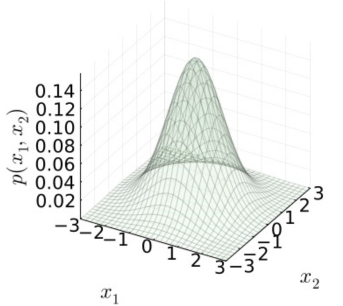
- **Properties**:

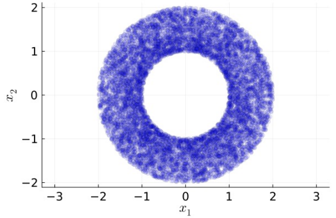
$$E[\boldsymbol{X}] = \boldsymbol{\mu}$$
$$\mathrm{cov}[\boldsymbol{X}] = \boldsymbol{\Sigma}$$

- **Covariance**. *For any two random variables X_1 and X_2 the covariance expresses the extent to which X_1 and X_2 vary together* **linearly** *and is given by*

$$\mathrm{cov}[X_1, X_2] = E_{X_1 X_2}[(X_1 - E[X_1]) \cdot (X_2 - E[X_2])] = E_{X_1 X_2}[X_1 X_2] - E[X_1] \cdot E[X_2]$$

  - Generalization of the variance to two random variables: $\mathrm{var}[X] = \mathrm{cov}[X, X]$
  - **Theorem**. *If two random variables X_1 and X_2 are independent, then $\mathrm{cov}[X_1, X_2] = 0$. The converse is not true!*





Multivariate Normal Distribution: Representations

- **Two Parameterizations (for different purposes):**
  - **Scale-Location Parameters**

$$\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-\frac{M}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}} \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

  - **Natural Parameters**

$$\mathcal{G}(\boldsymbol{x};\boldsymbol{\tau},\boldsymbol{P}) = (2\pi)^{-\frac{M}{2}}|\boldsymbol{P}|^{\frac{1}{2}} \cdot \exp\left(-\frac{1}{2}\boldsymbol{\tau}^{\mathrm{T}}\boldsymbol{P}^{-1}\boldsymbol{\tau}\right) \cdot \exp\left(\boldsymbol{\tau}^{\mathrm{T}}\boldsymbol{x}-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{x}\right)$$

- **Conversions**

$$\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathcal{G}(\boldsymbol{x};\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu},\boldsymbol{\Sigma}^{-1})$$

Matrix inverse

$$\mathcal{G}(\boldsymbol{x};\boldsymbol{\tau},\boldsymbol{P}) = \mathcal{N}(\boldsymbol{x};\boldsymbol{P}^{-1}\boldsymbol{\tau},\boldsymbol{P}^{-1})$$

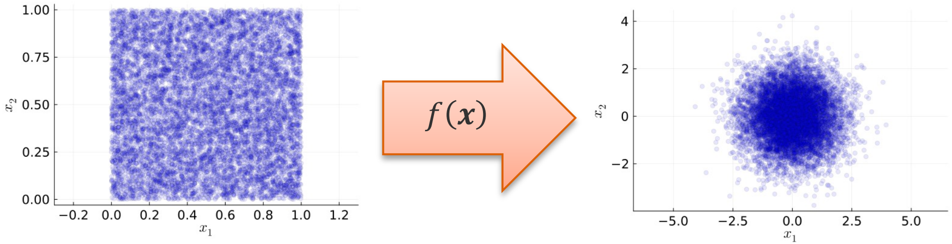
Sampling Multivariate Normal Distribution

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

- **Assumption**: We have access to a random number generator $x \sim \mathrm{Unif}([0,1])$
- **Box-Mueller**: If $x_1 \sim \mathrm{Unif}([0,1])$ and $x_2 \sim \mathrm{Unif}([0,1])$ then $f(\boldsymbol{x}) \sim N(\cdot; \mathbf{0}, \boldsymbol{I})$ for

$$f\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} \sqrt{-2\ln(x_1)} \cdot \cos(2\pi x_2) \\ \sqrt{-2\ln(x_1)} \cdot \sin(2\pi x_2) \end{bmatrix}$$

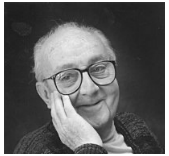
  - **In pictures:**



- **Sampling a multivariate Gaussian**. If $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ then for $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$

$$\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{y}; \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathrm{T}}\right)$$

  - For sampling a multivariate distribution, we require either the SVD or Cholesky decomposition of the covariance matrix, $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}$ (see exercises)
  - Can be easily proven from the properties of expectation and covariance



George Box
(1919 – 2013)

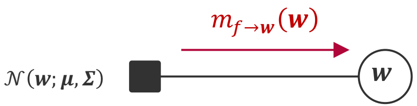
Mervin Mueller
(1928 – 2018)

Overview

1. Bayesian Linear Regression
2. **Bayesian Linear Regression via Message Passing**
   - Normal Distribution Revisited
   - **Posterior and Predictive Distribution**
3. Fast Bayesian Linear Regression
4. Bayesian Linear Regression via Linear Algebra

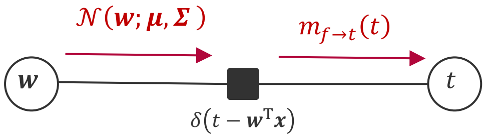
Multivariate Message Update Equations

Gaussian Factor

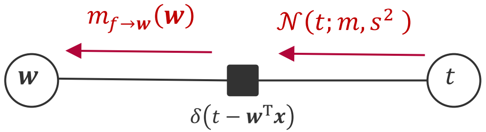


$$m_{f\to w}(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w};\boldsymbol{\mu},\boldsymbol{\Sigma})$$

Gaussian Projection Factor

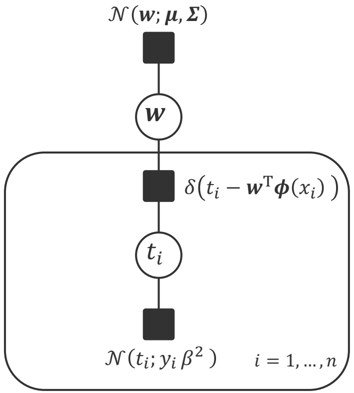


$$m_{f\to t}(t) = \int \delta\left(t-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}\right)\cdot\mathcal{N}(\boldsymbol{w};\boldsymbol{\mu},\boldsymbol{\Sigma})\ d\boldsymbol{w} = \mathcal{N}\left(t;\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{x},\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{x}\right)$$



$$m_{f\to w}(\boldsymbol{w}) = \int \delta\left(t-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}\right)\cdot\mathcal{N}(t;m,s^2)\ dt = \mathcal{G}\left(\boldsymbol{w};\frac{m}{s^2}\boldsymbol{x},\frac{1}{s^2}\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right)$$

Factor Graph



Bayesian Linear Regression by Message Passing

- **Message**: Simple factor tree where each training example is summarized in an M-dimensional message
  - Prior Message $m_{1,0}(\boldsymbol{w}) = \mathcal{G}(\boldsymbol{w}; \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}) = p(\boldsymbol{w})$
  - Target Message $m_{2,i}(t_i) = \mathcal{N}(t_i; y_i, \beta^2) = p(y_i|t_i)$
  - Data Message $m_{1,i}(\boldsymbol{w}) = \mathcal{G}\left(\boldsymbol{w}; \beta^{-2} y_i \boldsymbol{\phi}(x_i), \beta^{-2}\boldsymbol{\phi}(x_i)\boldsymbol{\phi}^{\mathrm{T}}(x_i)\right) = p(y_i|\boldsymbol{w})$
- **Posterior**: Multiplying prior and data messages we have

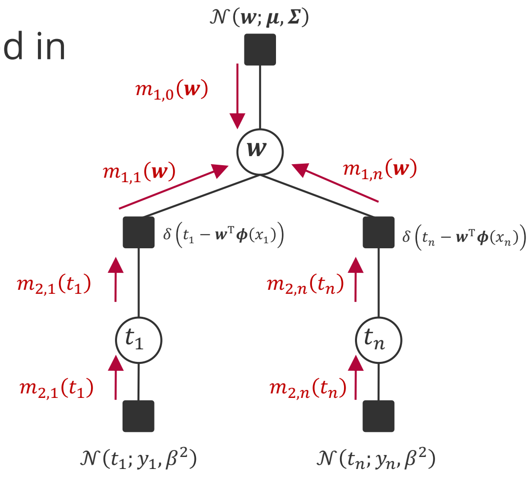
$$p(\boldsymbol{w}|D) = \mathcal{G}\left(\boldsymbol{w}; \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \beta^{-2}\boxed{\sum_{i=1}^{n} y_i \boldsymbol{\phi}(x_i)}, \boldsymbol{\Sigma}^{-1} + \beta^{-2}\boxed{\sum_{i=1}^{n} \boldsymbol{\phi}(x_i)\boldsymbol{\phi}^{\mathrm{T}}(x_i)}\right)$$

- **Feature Matrix**: All feature vectors are stacked on top of each other in a *feature matrix*

feature vector

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi_1(x_1) & \cdots & \phi_M(x_1) \\ \vdots & \ddots & \vdots \\ \phi_1(x_n) & \cdots & \phi_M(x_n) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\phi}^{\mathrm{T}}(x_1) \\ \vdots \\ \boldsymbol{\phi}^{\mathrm{T}}(x_n) \end{bmatrix}$$

$$\boxed{\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{y}} = [\boldsymbol{\phi}(x_1) \quad \cdots \quad \boldsymbol{\phi}(x_n)] \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^{n} y_i \boldsymbol{\phi}(x_i) \qquad \boxed{\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}} = [\boldsymbol{\phi}(x_1) \quad \cdots \quad \boldsymbol{\phi}(x_n)] \begin{bmatrix} \boldsymbol{\phi}^{\mathrm{T}}(x_1) \\ \vdots \\ \boldsymbol{\phi}^{\mathrm{T}}(x_n) \end{bmatrix} = \sum_{i=1}^{n} \boldsymbol{\phi}(x_i)\boldsymbol{\phi}^{\mathrm{T}}(x_i)$$



Predictions

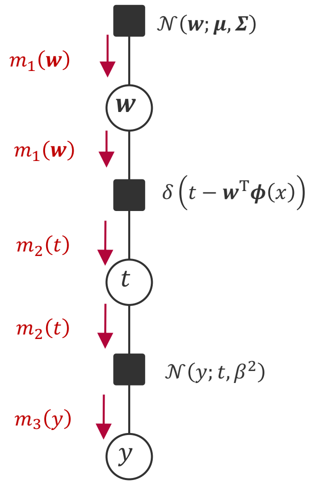
- **Predicition Tree**: Simple factor chain given posterior $p(\boldsymbol{w}|x, D) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - Posterior Message $m_1(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = p(\boldsymbol{w}|x, D)$
  - Projection Message $m_2(t) = \mathcal{N}\left(t; \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\phi}(x), \boldsymbol{\phi}^{\mathrm{T}}(x)\boldsymbol{\Sigma}\boldsymbol{\phi}(x)\right) = p(t|x, D)$
  - Prediction Message $m_3(y) = \mathcal{N}\left(y; \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\phi}(x), \beta^2 + \boldsymbol{\phi}^{\mathrm{T}}(x)\boldsymbol{\Sigma}\boldsymbol{\phi}(x)\right) = p(y|x, D)$
- **Bayesian Linear Regression in Matrix Notation**

$$p(\boldsymbol{w}|D) = \mathcal{N}\left(\boldsymbol{w}; \underbrace{\boldsymbol{S}_D\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \beta^{-2}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{y}\right)}_{\boldsymbol{m}}, \boldsymbol{S}_D\right), \qquad \boldsymbol{S}_D = \left(\boldsymbol{\Sigma}^{-1} + \beta^{-2}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}$$

$$p(y|x, D) = \mathcal{N}\left(y; \boldsymbol{m}^{\mathrm{T}}\boldsymbol{\phi}(x), \beta^2 + \boldsymbol{\phi}^{\mathrm{T}}(x)\boldsymbol{S}_D\boldsymbol{\phi}(x)\right)$$

data uncertainty

model uncertainty



Bayesian Linear Regression: Example

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{0}, \lambda^2 \boldsymbol{I})$$

$\lambda = 10$

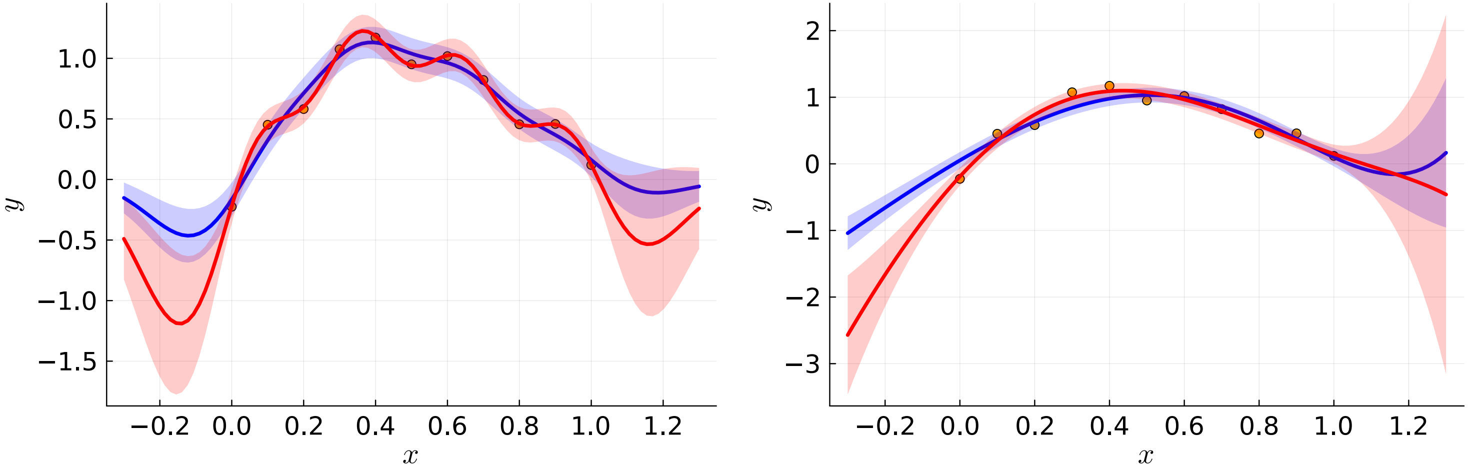
$\lambda = 1$

Gaussian Basis

$\phi_j(x) = \mathcal{N}(x; j, 0.15^2)$

Polynomial Basis

$\phi_j(x) = x^j$



Overview

1. Bayesian Linear Regression
2. Bayesian Linear Regression via Message Passing
   - Normal Distribution Revisited
   - Posterior and Predictive Distribution
3. **Fast Bayesian Linear Regression**
4. Bayesian Linear Regression via Linear Algebra

Fast Bayesian Linear Regression

- **Speeding up Bayesian Linear Regression**: Factorize the prior **and** posterior over the weight vector and then use message passing
  - Since x is fixed, we used $\boldsymbol{\phi} \coloneqq \boldsymbol{\phi}(x)$
  - Message $m_{1,i}(w_i) = \mathcal{N}\left(w_i; \mu_i, \sigma_i^2\right)$
  - Message $m_3(t) = \mathcal{N}(t; y, \beta^2)$
  - Message $m_{2,i}(w_i) = \mathcal{N}\left(w_i; \phi_i^{-1} \cdot \left(y - \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\phi} + \mu_i\phi_i\right), \phi_i^{-2} \cdot \left(\beta^2 + \sum_{j=1}^{M} \phi_j^2\sigma_j^2 - \phi_i^2\sigma_i^2\right)\right)$
- One can show that the product of $m_{1,i}(w_i)$ and $m_{2,i}(w_i)$ gives

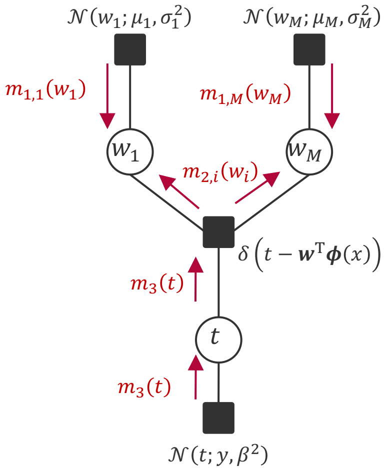
$$\mu_i \leftarrow \mu_i + \frac{y - \boldsymbol{\mu}^T\boldsymbol{\phi}(x)}{\phi_i(x)} \cdot \left[\frac{\phi_i^2(x)\sigma_i^2}{\beta^2 + \sum_{j=1}^{M}\phi_j^2(x)\sigma_j^2}\right]$$

target mismatch is measured in units of $\phi_i(x)$

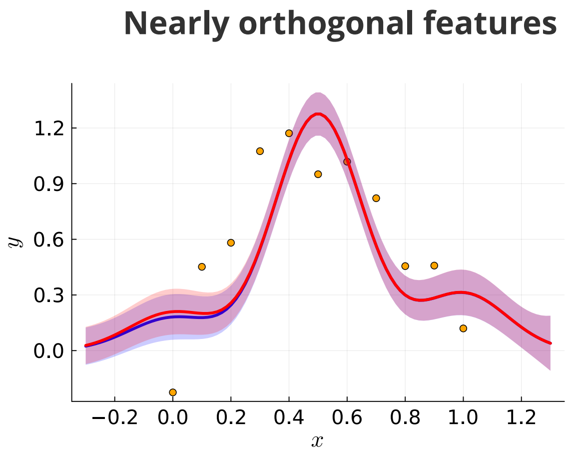
$$\sigma_i^2 \leftarrow \sigma_i^2 \cdot \left[1 - \frac{\phi_i^2(x)\sigma_i^2}{\beta^2 + \sum_{j=1}^{M}\phi_j^2(x)\sigma_j^2}\right]$$

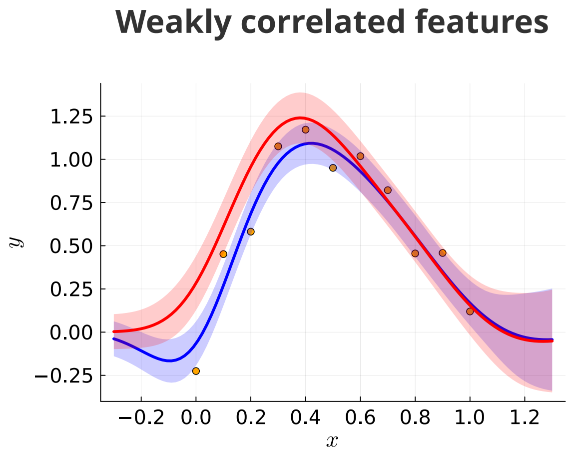
largest for parameter with largest uncertainty so far

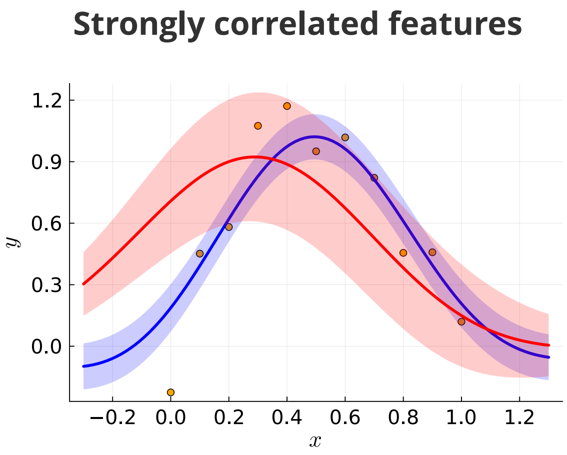
multiplicative update



Speeding up Bayesian Linear Regression







Overview

Bayes' Theorem for Normal Distributions

- **Conjugate Gaussians.** *Given a normally distributed variable*

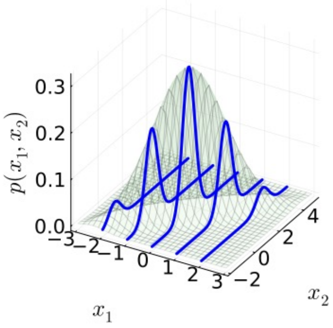
$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

and a conditional distribution for $\boldsymbol{y}$ *given* $\boldsymbol{x}$ *such that* $\boldsymbol{y}|\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{y}; \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{S})$ *we have the following for the marginal* $p(\boldsymbol{y})$ *and the "inverse" conditional* $p(\boldsymbol{x}|\boldsymbol{y})$

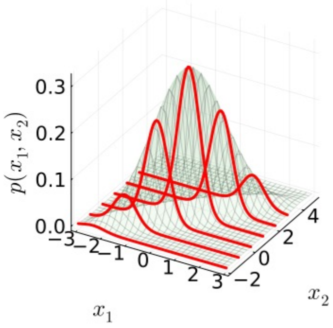
$$p(\boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{y}; \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{S} + \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathrm{T}}\right)$$

$$p(\boldsymbol{x}|\boldsymbol{y}) = \mathcal{G}\left(\boldsymbol{x}; \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{S}^{-1}(\boldsymbol{y} - \boldsymbol{b}), \boldsymbol{\Sigma}^{-1} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{S}^{-1}\boldsymbol{A}\right),$$

$$p(x_1) = \mathcal{N}(x_1; 0,1)$$



$$p(x_2|x_1) = \mathcal{N}\left(x_2; x_1 + 1, \frac{1}{2}\right)$$



$$p(x_1|x_2) = \mathcal{N}\left(x_1; \frac{2}{3}(x_2 - 1), \frac{1}{3}\right)$$

Introduction to Probabilistic Machine Learning

Unit 7 – Bayesian Regression

Conjugate Gaussians: Derivation

- **Main Ideas**:
  1. **Representation**: Represent the Gaussian distribution via natural parameters and introduce a log-normalization constant to capture the marginal
  2. **Multiplication**: Derive the update of the multiplication of two Gaussians over $\boldsymbol{x}$
  3. **Linear Mapping**: Derive a relation between Gaussian in $\boldsymbol{x}$ and in $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$
- **1D Warm-Up**
  - **Theorem (Multiplication)**. *Given two non-normalized one-dimensional Gaussian distributions $\mathcal{G}(x;\tau_1,\rho_1)$ and $\mathcal{G}(x;\tau_2,\rho_2)$ we have*

$$\mathcal{G}(x;\tau_1,\rho_1)\cdot\mathcal{G}(x;\tau_2,\rho_2)=\mathcal{G}(x;\tau_1+\tau_2,\rho_1+\rho_2)\cdot\mathcal{N}(\mu_1;\mu_2,\sigma_1^2+\sigma_2^2)$$

  - **Theorem (Linearity)**. *Given a non-normalized one-dimensional Gaussian distribution $\mathcal{N}(y;aw+b,\beta^2)$ we have*

$$\mathcal{N}(y;aw+b,\beta^2)=\mathcal{N}(w;a^{-1}(y-b),a^{-2}\beta^2)\cdot\frac{1}{a}$$

  - These two theorems combined allow to both efficiently and robustly compute the posterior parameters and derive the conjugate Gaussian equations

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Bayesian Linear Regression

- **Bayesian Linear Regression**: *For the linear basis function model* $f(x; \boldsymbol{w}) := \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(x)$ *with likelihood* $p(D|\boldsymbol{w}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{\Phi w}, \beta^2 \boldsymbol{I})$ *and prior* $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$p(\boldsymbol{w}|D) = \mathcal{N}\left(\boldsymbol{w}; \boldsymbol{S}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \frac{1}{\beta^2}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{y}\right), \boldsymbol{S}\right), \qquad \boldsymbol{S}^{-1} = \boldsymbol{\Sigma}^{-1} + \frac{1}{\beta^2}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}$$

- **Special Case**: $\boldsymbol{\mu} = \boldsymbol{0}$ and $\boldsymbol{\Sigma} = \tau^2 \boldsymbol{I}$
  - For the posterior **mean** we have (Why?):

$$\boldsymbol{\mu}_{\text{posterior}} = \left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \frac{\beta^2}{\tau^2}\boldsymbol{I}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{w}_{\text{MAP}}$$

  - If the mean of the full Bayesian inference and the maximum-a-posteriori are the same, what's the difference?! The **variance** of the predictive distribution!

$$p(y|x, D) = \int p(y|x, \boldsymbol{w}) \cdot p(\boldsymbol{w}|D) d\boldsymbol{w} = \int \mathcal{N}\left(y; \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(x), \beta^2\right) \cdot p(\boldsymbol{w}|D) d\boldsymbol{w}$$

$$p(y|x, D) = \mathcal{N}\left(y; \left(\boldsymbol{S}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \frac{1}{\beta^2}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{y}\right)\right)^{\mathrm{T}}\boldsymbol{\phi}(x), \beta^2 + \boldsymbol{\phi}^{\mathrm{T}}(x)\boldsymbol{S}\boldsymbol{\phi}(x)\right)$$

Properties of Gaussians

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$p(\boldsymbol{v}|\boldsymbol{w}) = \mathcal{N}(\boldsymbol{v}; \boldsymbol{A}\boldsymbol{w}, \Xi)$$

$$p(\boldsymbol{v}) = \mathcal{N}\left(\boldsymbol{v}; \boldsymbol{A}\boldsymbol{\mu}, \Xi + \boldsymbol{A}\Sigma\mathrm{A}^{\mathrm{T}}\right)$$
$$p(\boldsymbol{w}|\boldsymbol{v}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{m}, \boldsymbol{S})$$
$$\boldsymbol{m} = \boldsymbol{S}\left(\boldsymbol{A}^{\mathrm{T}}\Xi^{-1}\boldsymbol{v} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)$$
$$\boldsymbol{S}^{-1} = \boldsymbol{\Sigma}^{-1} + \mathrm{A}^{\mathrm{T}}\Xi^{-1}\boldsymbol{A}$$

Summary

1. **Bayesian Linear Regression**
   - Averaging over all functions weighting them by their posterior probability gives both a smoother mean and confidence intervals for each prediction (predictive distribution)
   - Marginals and conditionals for multivariate Normals are linearly transformed Normals!
   - Message passing on the Bayesian Regression factor graph involves no loops and is exact
   - For linear basis function models with Normal noise, the posterior can be computed closed form
   - Mean of Bayesian regression equals MAP solution but variance accounts for model uncertainty
2. **Fast Bayesian Linear Regression**
   - The Bayesian linear regression algorithm is of cubic complexity in the features and quadratic in the training set size
   - By factorizing *both* the prior and posterior distribution over the weight vector, we get a completely linear-complexity algorithm!

See you next week!