Probabilistic Models for Regression

Piyush Rai

Probabilistic Machine Learning (CS772A)

Aug 5, 2017

Background: Linear Regression

- Supervised Learning. Training data as N i.i.d. examples $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$, $\boldsymbol{x}_n \in \mathbb{R}^D$, $y_n \in \mathbb{R}$

Background: Linear Regression

- Supervised Learning. Training data as N i.i.d. examples $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$, $\boldsymbol{x}_n \in \mathbb{R}^D$, $y_n \in \mathbb{R}$
- Assume input-output relationship to approximately follow a linear model with parameters $\boldsymbol{w} \in \mathbb{R}^D$

$$y_n \approx \boldsymbol{w}^\top \boldsymbol{x}_n \quad \Rightarrow \quad y_n \approx \sum_{d=1}^{D} w_d x_{nd}$$

Background: Linear Regression

- Supervised Learning. Training data as N i.i.d. examples $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$, $\boldsymbol{x}_n \in \mathbb{R}^D$, $y_n \in \mathbb{R}$
- Assume input-output relationship to approximately follow a linear model with parameters $\boldsymbol{w} \in \mathbb{R}^D$

$$y_n \approx \boldsymbol{w}^\top \boldsymbol{x}_n \quad \Rightarrow \quad y_n \approx \sum_{d=1}^{D} w_d x_{nd}$$

- $\boldsymbol{w} \in \mathbb{R}^D$ is also called the regression **weight vector** (the unknown we want to estimate)

Background: Linear Regression

- Supervised Learning. Training data as N i.i.d. examples $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$, $\boldsymbol{x}_n \in \mathbb{R}^D$, $y_n \in \mathbb{R}$
- Assume input-output relationship to approximately follow a linear model with parameters $\boldsymbol{w} \in \mathbb{R}^D$

$$y_n \approx \boldsymbol{w}^\top \boldsymbol{x}_n \quad \Rightarrow \quad y_n \approx \sum_{d=1}^{D} w_d x_{nd}$$

- $\boldsymbol{w} \in \mathbb{R}^D$ is also called the regression **weight vector** (the unknown we want to estimate)
  - Can think of the d^{th} component of $\boldsymbol{w}$, i.e. w_d, as weight/importance of d^{th} feature of the input

Background: Linear Regression

- Supervised Learning. Training data as N i.i.d. examples $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$, $\boldsymbol{x}_n \in \mathbb{R}^D$, $y_n \in \mathbb{R}$
- Assume input-output relationship to approximately follow a linear model with parameters $\boldsymbol{w} \in \mathbb{R}^D$

$$y_n \approx \boldsymbol{w}^\top \boldsymbol{x}_n \quad \Rightarrow \quad y_n \approx \sum_{d=1}^{D} w_d x_{nd}$$

- $\boldsymbol{w} \in \mathbb{R}^D$ is also called the regression **weight vector** (the unknown we want to estimate)
    - Can think of the d^{th} component of $\boldsymbol{w}$, i.e. w_d, as weight/importance of d^{th} feature of the input
- We usually learn $\boldsymbol{w}$ by optimizing a loss function or regularized loss function

Background: Linear Regression

- Supervised Learning. Training data as N i.i.d. examples $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$, $\boldsymbol{x}_n \in \mathbb{R}^D$, $y_n \in \mathbb{R}$
- Assume input-output relationship to approximately follow a linear model with parameters $\boldsymbol{w} \in \mathbb{R}^D$

$$y_n \approx \boldsymbol{w}^\top \boldsymbol{x}_n \quad \Rightarrow \quad y_n \approx \sum_{d=1}^{D} w_d x_{nd}$$

- $\boldsymbol{w} \in \mathbb{R}^D$ is also called the regression **weight vector** (the unknown we want to estimate)
    - Can think of the d^{th} component of $\boldsymbol{w}$, i.e. w_d, as weight/importance of d^{th} feature of the input
- We usually learn $\boldsymbol{w}$ by optimizing a loss function or regularized loss function

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2$$

Background: Linear Regression

- Supervised Learning. Training data as N i.i.d. examples $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$, $\boldsymbol{x}_n \in \mathbb{R}^D$, $y_n \in \mathbb{R}$
- Assume input-output relationship to approximately follow a linear model with parameters $\boldsymbol{w} \in \mathbb{R}^D$

$$y_n \approx \boldsymbol{w}^\top \boldsymbol{x}_n \quad \Rightarrow \quad y_n \approx \sum_{d=1}^{D} w_d x_{nd}$$

- $\boldsymbol{w} \in \mathbb{R}^D$ is also called the regression **weight vector** (the unknown we want to estimate)
  - Can think of the d^{th} component of $\boldsymbol{w}$, i.e. w_d, as weight/importance of d^{th} feature of the input
- We usually learn $\boldsymbol{w}$ by optimizing a loss function or regularized loss function

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2 = (\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top)^{-1} (\sum_{n=1}^{N} \boldsymbol{x}_n y_n)$$

Background: Linear Regression

- Supervised Learning. Training data as N i.i.d. examples $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$, $\boldsymbol{x}_n \in \mathbb{R}^D$, $y_n \in \mathbb{R}$
- Assume input-output relationship to approximately follow a linear model with parameters $\boldsymbol{w} \in \mathbb{R}^D$

$$y_n \approx \boldsymbol{w}^\top \boldsymbol{x}_n \quad \Rightarrow \quad y_n \approx \sum_{d=1}^{D} w_d x_{nd}$$

- $\boldsymbol{w} \in \mathbb{R}^D$ is also called the regression **weight vector** (the unknown we want to estimate)
  - Can think of the d^{th} component of $\boldsymbol{w}$, i.e. w_d, as weight/importance of d^{th} feature of the input
- We usually learn $\boldsymbol{w}$ by optimizing a loss function or regularized loss function

$$\hat{\boldsymbol{w}} \quad = \quad \arg\min_{\boldsymbol{w}} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2 = (\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top)^{-1} (\sum_{n=1}^{N} \boldsymbol{x}_n y_n) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{y}$$

Background: Linear Regression

- Supervised Learning. Training data as N i.i.d. examples $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$, $\boldsymbol{x}_n \in \mathbb{R}^D$, $y_n \in \mathbb{R}$
- Assume input-output relationship to approximately follow a linear model with parameters $\boldsymbol{w} \in \mathbb{R}^D$

$$y_n \approx \boldsymbol{w}^\top \boldsymbol{x}_n \quad \Rightarrow \quad y_n \approx \sum_{d=1}^{D} w_d x_{nd}$$

- $\boldsymbol{w} \in \mathbb{R}^D$ is also called the regression **weight vector** (the unknown we want to estimate)
  - Can think of the d^{th} component of $\boldsymbol{w}$, i.e. w_d, as weight/importance of d^{th} feature of the input
- We usually learn $\boldsymbol{w}$ by optimizing a loss function or regularized loss function

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2 = (\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top)^{-1} (\sum_{n=1}^{N} \boldsymbol{x}_n y_n) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{y}$$

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2 + \tau \boldsymbol{w}^\top \boldsymbol{w}$$

Background: Linear Regression

- Supervised Learning. Training data as N i.i.d. examples $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$, $\boldsymbol{x}_n \in \mathbb{R}^D$, $y_n \in \mathbb{R}$
- Assume input-output relationship to approximately follow a linear model with parameters $\boldsymbol{w} \in \mathbb{R}^D$

$$y_n \approx \boldsymbol{w}^\top \boldsymbol{x}_n \quad \Rightarrow \quad y_n \approx \sum_{d=1}^{D} w_d x_{nd}$$

- $\boldsymbol{w} \in \mathbb{R}^D$ is also called the regression **weight vector** (the unknown we want to estimate)
  - Can think of the d^{th} component of $\boldsymbol{w}$, i.e. w_d, as weight/importance of d^{th} feature of the input
- We usually learn $\boldsymbol{w}$ by optimizing a loss function or regularized loss function

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2 = (\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top)^{-1} (\sum_{n=1}^{N} \boldsymbol{x}_n y_n) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{y}$$

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2 + \tau \boldsymbol{w}^\top \boldsymbol{w} = (\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top + \tau \mathbf{I})^{-1} (\sum_{n=1}^{N} \boldsymbol{x}_n y_n)$$

Background: Linear Regression

- Supervised Learning. Training data as N i.i.d. examples $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$, $\boldsymbol{x}_n \in \mathbb{R}^D$, $y_n \in \mathbb{R}$
- Assume input-output relationship to approximately follow a linear model with parameters $\boldsymbol{w} \in \mathbb{R}^D$

$$y_n \approx \boldsymbol{w}^\top \boldsymbol{x}_n \quad \Rightarrow \quad y_n \approx \sum_{d=1}^{D} w_d x_{nd}$$

- $\boldsymbol{w} \in \mathbb{R}^D$ is also called the regression **weight vector** (the unknown we want to estimate)
  - Can think of the d^{th} component of $\boldsymbol{w}$, i.e. w_d, as weight/importance of d^{th} feature of the input
- We usually learn $\boldsymbol{w}$ by optimizing a loss function or regularized loss function

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2 = (\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top)^{-1} (\sum_{n=1}^{N} \boldsymbol{x}_n y_n) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{y}$$

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2 + \tau \boldsymbol{w}^\top \boldsymbol{w} = (\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top + \tau \mathbf{I})^{-1} (\sum_{n=1}^{N} \boldsymbol{x}_n y_n) = (\mathbf{X}^\top \mathbf{X} + \tau \mathbf{I})^{-1} \mathbf{X}^\top \boldsymbol{y}$$

Background: Linear Regression

- Supervised Learning. Training data as N i.i.d. examples $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$, $\boldsymbol{x}_n \in \mathbb{R}^D$, $y_n \in \mathbb{R}$
- Assume input-output relationship to approximately follow a linear model with parameters $\boldsymbol{w} \in \mathbb{R}^D$

$$y_n \approx \boldsymbol{w}^\top \boldsymbol{x}_n \quad \Rightarrow \quad y_n \approx \sum_{d=1}^{D} w_d x_{nd}$$

- $\boldsymbol{w} \in \mathbb{R}^D$ is also called the regression **weight vector** (the unknown we want to estimate)
  - Can think of the d^{th} component of $\boldsymbol{w}$, i.e. w_d, as weight/importance of d^{th} feature of the input
- We usually learn $\boldsymbol{w}$ by optimizing a loss function or regularized loss function

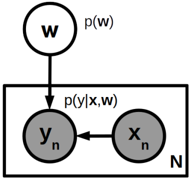
$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2 = (\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top)^{-1} (\sum_{n=1}^{N} \boldsymbol{x}_n y_n) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{y}$$

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2 + \tau \boldsymbol{w}^\top \boldsymbol{w} = (\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top + \tau \mathbf{I})^{-1} (\sum_{n=1}^{N} \boldsymbol{x}_n y_n) = (\mathbf{X}^\top \mathbf{X} + \tau \mathbf{I})^{-1} \mathbf{X}^\top \boldsymbol{y}$$

- Nice, convex objective with unique solution

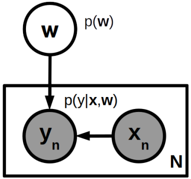
Probabilistic Linear Regression

- We would like to set up linear regression as a probabilistic model



Probabilistic Linear Regression

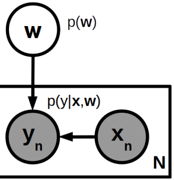
- We would like to set up linear regression as a probabilistic model



- We would specify the appropriate likelihood and prior distributions for this model

Probabilistic Linear Regression

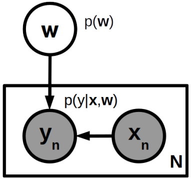
- We would like to set up linear regression as a probabilistic model



- We would specify the appropriate likelihood and prior distributions for this model
- We will then estimate the model parameters using MLE/MAP and fully Bayesian inference

Probabilistic Linear Regression

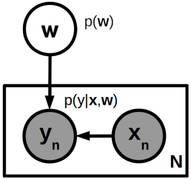
- We would like to set up linear regression as a probabilistic model



- We would specify the appropriate likelihood and prior distributions for this model
- We will then estimate the model parameters using MLE/MAP and fully Bayesian inference
- We will only be modeling the outputs y; the inputs $\boldsymbol{x}$ will not be modeled (assumed fixed)

Probabilistic Linear Regression

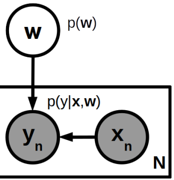
- We would like to set up linear regression as a probabilistic model



- We would specify the appropriate likelihood and prior distributions for this model
- We will then estimate the model parameters using MLE/MAP and fully Bayesian inference
- We will only be modeling the outputs y; the inputs $\boldsymbol{x}$ will not be modeled (assumed fixed)
    - Note: In some supervised learning problems (especially classification), inputs are also modeled (more on this when we discuss "generative classification")

Probabilistic Linear Regression

- We would like to set up linear regression as a probabilistic model

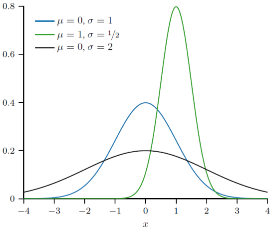


- We would specify the appropriate likelihood and prior distributions for this model
- We will then estimate the model parameters using MLE/MAP and fully Bayesian inference
- We will only be modeling the outputs y; the inputs $\boldsymbol{x}$ will not be modeled (assumed fixed)
  - Note: In some supervised learning problems (especially classification), inputs are also modeled (more on this when we discuss "generative classification")
- Only $\boldsymbol{w}$ and y_n $(n = 1, \ldots, N)$ are the random variables in this model (not $\boldsymbol{x}_n$'s, which are fixed)

Background: Univariate Gaussian Distribution

- Distribution over real-valued scalar r.v. x
- Defined by a scalar **mean** μ and a scalar **variance** $\sigma^2 > 0$
- Mean: $\mathbb{E}[x] = \mu$, Variance: $\text{var}[x] = \sigma^2$
- Precision (inverse variance) $\beta = 1/\sigma^2$
- Distribution defined as

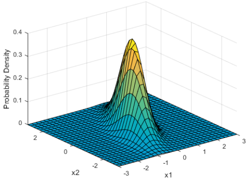
$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \sqrt{\frac{\beta}{2\pi}} e^{-\frac{\beta}{2}(x-\mu)^2}$$



Background: Multivariate Gaussian Distribution

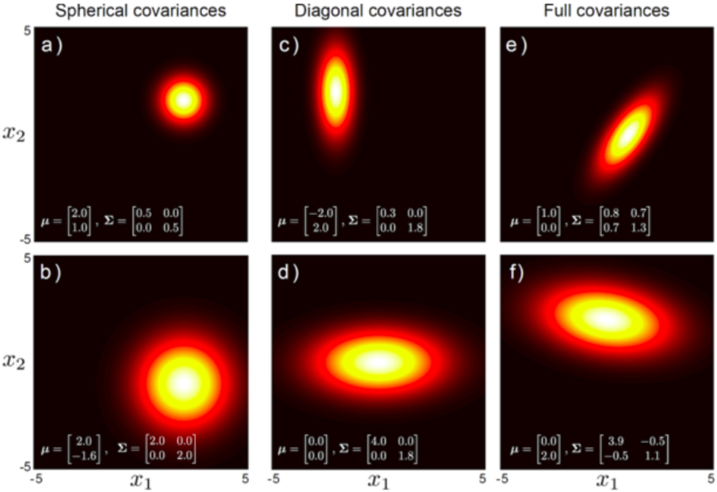
- Distribution over a multivariate r.v. vector $\mathbf{x} \in \mathbb{R}^D$ of real numbers
- Defined by a **mean vector** $\boldsymbol{\mu} \in \mathbb{R}^D$ and a $D \times D$ **cov. matrix** $\boldsymbol{\Sigma}$ (its inverse is precision matrix)

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$



- The covariance matrix $\boldsymbol{\Sigma}$ must be symmetric and positive definite
  - All eigenvalues are positive
  - $\boldsymbol{z}^\top \boldsymbol{\Sigma} \boldsymbol{z} > 0$ for any real vector $\boldsymbol{z}$
- Inverse of the covariance matrix is known as the **precision matrix** $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ (also positive definite)

The covariance matrix can be spherical, diagonal, or full



Picture courtesy: Computer vision: models, learning and inference (Simon Price)

Probabilistic Linear Regression: Likelihood Model

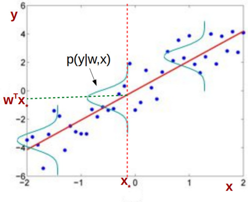
- Assume a model of the form $y_n = \boldsymbol{w}^\top \boldsymbol{x}_n + \epsilon_n$, where $\epsilon_n \sim \mathcal{N}(0, \beta^{-1})$ is Gaussian noise

Probabilistic Linear Regression: Likelihood Model

- Assume a model of the form $y_n = \boldsymbol{w}^\top \boldsymbol{x}_n + \epsilon_n$, where $\epsilon_n \sim \mathcal{N}(0, \beta^{-1})$ is Gaussian noise
- Thus each response y_n also has a Gaussian distribution: $p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1})$

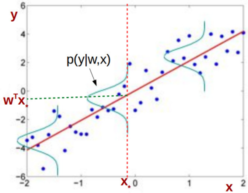
Probabilistic Linear Regression: Likelihood Model

- Assume a model of the form $y_n = \boldsymbol{w}^\top \boldsymbol{x}_n + \epsilon_n$, where $\epsilon_n \sim \mathcal{N}(0, \beta^{-1})$ is Gaussian noise
- Thus each response y_n also has a Gaussian distribution: $p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1})$



Probabilistic Linear Regression: Likelihood Model

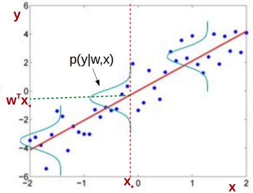
- Assume a model of the form $y_n = \boldsymbol{w}^\top \boldsymbol{x}_n + \epsilon_n$, where $\epsilon_n \sim \mathcal{N}(0, \beta^{-1})$ is Gaussian noise
- Thus each response y_n also has a Gaussian distribution: $p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1})$



- $p(y_n|\boldsymbol{x}_n, \boldsymbol{w})$ is our likelihood model - Gaussian distribution in this case

Probabilistic Linear Regression: Likelihood Model

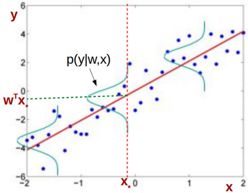
- Assume a model of the form $y_n = \boldsymbol{w}^\top \boldsymbol{x}_n + \epsilon_n$, where $\epsilon_n \sim \mathcal{N}(0, \beta^{-1})$ is Gaussian noise
- Thus each response y_n also has a Gaussian distribution: $p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1})$



- $p(y_n|\boldsymbol{x}_n, \boldsymbol{w})$ is our likelihood model - Gaussian distribution in this case
- Note: We are only modeling the outputs y; the inputs $\boldsymbol{x}$ are not being modeled (fixed)

Probabilistic Linear Regression: Likelihood Model

- Assume a model of the form $y_n = \boldsymbol{w}^\top \boldsymbol{x}_n + \epsilon_n$, where $\epsilon_n \sim \mathcal{N}(0, \beta^{-1})$ is Gaussian noise
- Thus each response y_n also has a Gaussian distribution: $p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1})$

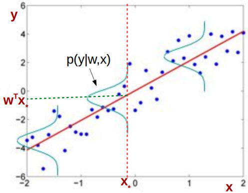


- $p(y_n|\boldsymbol{x}_n, \boldsymbol{w})$ is our likelihood model - Gaussian distribution in this case
- Note: We are only modeling the outputs y; the inputs $\boldsymbol{x}$ are not being modeled (fixed)
- Denoting $\mathbf{X} = [\boldsymbol{x}_1 \dots \boldsymbol{x}_N]^\top$ $(N \times D)$, $\boldsymbol{y} = [y_1 \dots y_N]^\top$ $(N \times 1)$, the overall likelihood for i.i.d. data

$$p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}) = \prod_{n=1}^{N} p(y_n|\boldsymbol{x}_n, \boldsymbol{w})$$

Probabilistic Linear Regression: Likelihood Model

- Assume a model of the form $y_n = \boldsymbol{w}^\top \boldsymbol{x}_n + \epsilon_n$, where $\epsilon_n \sim \mathcal{N}(0, \beta^{-1})$ is Gaussian noise
- Thus each response y_n also has a Gaussian distribution: $p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1})$

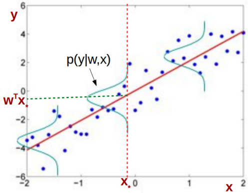


- $p(y_n|\boldsymbol{x}_n, \boldsymbol{w})$ is our likelihood model - Gaussian distribution in this case
- Note: We are only modeling the outputs y; the inputs $\boldsymbol{x}$ are not being modeled (fixed)
- Denoting $\mathbf{X} = [\boldsymbol{x}_1 \ldots \boldsymbol{x}_N]^\top$ $(N \times D)$, $\boldsymbol{y} = [y_1 \ldots y_N]^\top$ $(N \times 1)$, the overall likelihood for i.i.d. data

$$p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}) = \prod_{n=1}^{N} p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \prod_{n=1}^{N} \sqrt{\frac{\beta}{2\pi}} e^{-\frac{\beta}{2}(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2}$$

Probabilistic Linear Regression: Likelihood Model

- Assume a model of the form $y_n = \boldsymbol{w}^\top \boldsymbol{x}_n + \epsilon_n$, where $\epsilon_n \sim \mathcal{N}(0, \beta^{-1})$ is Gaussian noise
- Thus each response y_n also has a Gaussian distribution: $p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1})$

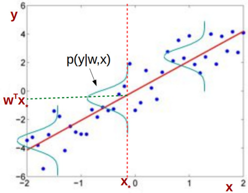


- $p(y_n|\boldsymbol{x}_n, \boldsymbol{w})$ is our likelihood model - Gaussian distribution in this case
- Note: We are only modeling the outputs y; the inputs $\boldsymbol{x}$ are not being modeled (fixed)
- Denoting $\mathbf{X} = [\boldsymbol{x}_1 \ldots \boldsymbol{x}_N]^\top$ $(N \times D)$, $\boldsymbol{y} = [y_1 \ldots y_N]^\top$ $(N \times 1)$, the overall likelihood for i.i.d. data

$$p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}) = \prod_{n=1}^{N} p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \prod_{n=1}^{N} \sqrt{\frac{\beta}{2\pi}} e^{-\frac{\beta}{2}(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2} \propto e^{-\frac{\beta}{2}\sum_{n=1}^{N}(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2}$$

Probabilistic Linear Regression: Likelihood Model

- Assume a model of the form $y_n = \boldsymbol{w}^\top \boldsymbol{x}_n + \epsilon_n$, where $\epsilon_n \sim \mathcal{N}(0, \beta^{-1})$ is Gaussian noise
- Thus each response y_n also has a Gaussian distribution: $p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1})$



- $p(y_n|\boldsymbol{x}_n, \boldsymbol{w})$ is our likelihood model - Gaussian distribution in this case
- Note: We are only modeling the outputs y; the inputs $\boldsymbol{x}$ are not being modeled (fixed)
- Denoting $\mathbf{X} = [\boldsymbol{x}_1 \dots \boldsymbol{x}_N]^\top$ $(N \times D)$, $\boldsymbol{y} = [y_1 \dots y_N]^\top$ $(N \times 1)$, the overall likelihood for i.i.d. data

$$p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}) = \prod_{n=1}^{N} p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \prod_{n=1}^{N} \sqrt{\frac{\beta}{2\pi}} e^{-\frac{\beta}{2}(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2} \propto e^{-\frac{\beta}{2}\sum_{n=1}^{N}(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2} \propto e^{-\frac{\beta}{2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{w})^\top(\boldsymbol{y} - \mathbf{X}\boldsymbol{w})}$$

Probabilistic Linear Regression: Prior

- Let's assume a zero mean multivariate Gaussian prior on the weight vector $\boldsymbol{w} \in \mathbb{R}^D$

$$P(\boldsymbol{w}|\lambda) = \mathcal{N}(\boldsymbol{w}|\mathbf{0}, \lambda^{-1}\mathbf{I}_D)$$

Probabilistic Linear Regression: Prior

- Let's assume a zero mean multivariate Gaussian prior on the weight vector $\boldsymbol{w} \in \mathbb{R}^D$

$$P(\boldsymbol{w}|\lambda) = \mathcal{N}(\boldsymbol{w}|\mathbf{0}, \lambda^{-1}\mathbf{I}_D) = \prod_{d=1}^{D} \mathcal{N}(w_d|0, \lambda^{-1})$$

Probabilistic Linear Regression: Prior

- Let's assume a zero mean multivariate Gaussian prior on the weight vector $\boldsymbol{w} \in \mathbb{R}^D$

$$P(\boldsymbol{w}|\lambda) = \mathcal{N}(\boldsymbol{w}|\mathbf{0}, \lambda^{-1}\mathbf{I}_D) = \prod_{d=1}^{D} \mathcal{N}(w_d|0, \lambda^{-1}) = \prod_{d=1}^{D} \sqrt{\frac{\lambda}{2\pi}} e^{-\frac{\lambda}{2} w_d^2}$$

Probabilistic Linear Regression: Prior

- Let's assume a zero mean multivariate Gaussian prior on the weight vector $\boldsymbol{w} \in \mathbb{R}^D$

$$P(\boldsymbol{w}|\lambda) = \mathcal{N}(\boldsymbol{w}|\mathbf{0}, \lambda^{-1}\mathbf{I}_D) = \prod_{d=1}^{D} \mathcal{N}(w_d|0, \lambda^{-1}) = \prod_{d=1}^{D} \sqrt{\frac{\lambda}{2\pi}} e^{-\frac{\lambda}{2} w_d^2} \propto e^{-\frac{\lambda}{2}\sum_{d=1}^{D} w_d^2}$$

Probabilistic Linear Regression: Prior

- Let's assume a zero mean multivariate Gaussian prior on the weight vector $\boldsymbol{w} \in \mathbb{R}^D$

$$P(\boldsymbol{w}|\lambda) = \mathcal{N}(\boldsymbol{w}|\mathbf{0}, \lambda^{-1}\mathbf{I}_D) = \prod_{d=1}^{D} \mathcal{N}(w_d|0, \lambda^{-1}) = \prod_{d=1}^{D} \sqrt{\frac{\lambda}{2\pi}} e^{-\frac{\lambda}{2} w_d^2} \propto e^{-\frac{\lambda}{2} \sum_{d=1}^{D} w_d^2} \propto e^{-\frac{\lambda}{2} \boldsymbol{w}^\top \boldsymbol{w}}$$

Probabilistic Linear Regression: Prior

- Let's assume a zero mean multivariate Gaussian prior on the weight vector $\boldsymbol{w} \in \mathbb{R}^D$

$$P(\boldsymbol{w}|\lambda) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \lambda^{-1}\mathbf{I}_D) = \prod_{d=1}^{D} \mathcal{N}(w_d|0, \lambda^{-1}) = \prod_{d=1}^{D} \sqrt{\frac{\lambda}{2\pi}} e^{-\frac{\lambda}{2} w_d^2} \propto e^{-\frac{\lambda}{2}\sum_{d=1}^{D} w_d^2} \propto e^{-\frac{\lambda}{2}\boldsymbol{w}^\top \boldsymbol{w}}$$

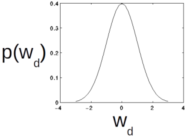
where λ denotes the precision of the Gaussian (how shrunk w_d is towards the mean of the Gaussian)

Probabilistic Linear Regression: Prior

- Let's assume a zero mean multivariate Gaussian prior on the weight vector $\boldsymbol{w} \in \mathbb{R}^D$

$$P(\boldsymbol{w}|\lambda) = \mathcal{N}(\boldsymbol{w}|\mathbf{0}, \lambda^{-1}\mathbf{I}_D) = \prod_{d=1}^{D} \mathcal{N}(w_d|0, \lambda^{-1}) = \prod_{d=1}^{D} \sqrt{\frac{\lambda}{2\pi}} e^{-\frac{\lambda}{2} w_d^2} \propto e^{-\frac{\lambda}{2} \sum_{d=1}^{D} w_d^2} \propto e^{-\frac{\lambda}{2} \boldsymbol{w}^\top \boldsymbol{w}}$$

where λ denotes the precision of the Gaussian (how shrunk w_d is towards the mean of the Gaussian)

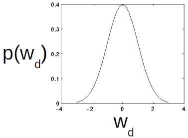


Probabilistic Linear Regression: Prior

- Let's assume a zero mean multivariate Gaussian prior on the weight vector $\boldsymbol{w} \in \mathbb{R}^D$

$$P(\boldsymbol{w}|\lambda) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \lambda^{-1}\mathbf{I}_D) = \prod_{d=1}^{D} \mathcal{N}(w_d|0, \lambda^{-1}) = \prod_{d=1}^{D} \sqrt{\frac{\lambda}{2\pi}} e^{-\frac{\lambda}{2} w_d^2} \propto e^{-\frac{\lambda}{2}\sum_{d=1}^{D} w_d^2} \propto e^{-\frac{\lambda}{2}\boldsymbol{w}^\top \boldsymbol{w}}$$

where λ denotes the precision of the Gaussian (how shrunk w_d is towards the mean of the Gaussian)



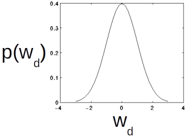
- The prior says that *a priori* we want the individual weights w_d to be small (close to zero)

Probabilistic Linear Regression: Prior

- Let's assume a zero mean multivariate Gaussian prior on the weight vector $\boldsymbol{w} \in \mathbb{R}^D$

$$P(\boldsymbol{w}|\lambda) = \mathcal{N}(\boldsymbol{w}|\mathbf{0}, \lambda^{-1}\mathbf{I}_D) = \prod_{d=1}^{D} \mathcal{N}(w_d|0, \lambda^{-1}) = \prod_{d=1}^{D} \sqrt{\frac{\lambda}{2\pi}} e^{-\frac{\lambda}{2} w_d^2} \propto e^{-\frac{\lambda}{2}\sum_{d=1}^{D} w_d^2} \propto e^{-\frac{\lambda}{2}\boldsymbol{w}^\top \boldsymbol{w}}$$

where λ denotes the precision of the Gaussian (how shrunk w_d is towards the mean of the Gaussian)



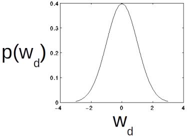
- The prior says that *a priori* we want the individual weights w_d to be small (close to zero)
  - Akin to imposing a regularizer on $\boldsymbol{w}$ that promotes its elements to take small values

Probabilistic Linear Regression: Prior

- Let's assume a zero mean multivariate Gaussian prior on the weight vector $\boldsymbol{w} \in \mathbb{R}^D$

$$P(\boldsymbol{w}|\lambda) = \mathcal{N}(\boldsymbol{w}|\mathbf{0}, \lambda^{-1}\mathbf{I}_D) = \prod_{d=1}^{D} \mathcal{N}(w_d|0, \lambda^{-1}) = \prod_{d=1}^{D} \sqrt{\frac{\lambda}{2\pi}} e^{-\frac{\lambda}{2} w_d^2} \propto e^{-\frac{\lambda}{2} \sum_{d=1}^{D} w_d^2} \propto e^{-\frac{\lambda}{2} \boldsymbol{w}^\top \boldsymbol{w}}$$

  where λ denotes the precision of the Gaussian (how shrunk w_d is towards the mean of the Gaussian)



- The prior says that *a priori* we want the individual weights w_d to be small (close to zero)
  - Akin to imposing a regularizer on $\boldsymbol{w}$ that promotes its elements to take small values
  - We'll see that the Gaussian prior exactly corresponds to having an ℓ_2 (squared Euclidean norm) regularizer on $\boldsymbol{w}$, of the form $\boldsymbol{w}^\top \boldsymbol{w} = \sum_{d=1}^{D} w_d^2$

MLE for Probabilistic Linear Regression

- The negative log-likelihood will be

$$NLL(\boldsymbol{w}) = -\log p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}) = -\log \prod_{n=1}^{N} \sqrt{\frac{\beta}{2\pi}} e^{-\frac{\beta}{2}(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2}$$

MLE for Probabilistic Linear Regression

- The negative log-likelihood will be

$$NLL(\boldsymbol{w}) = -\log p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}) = -\log \prod_{n=1}^{N} \sqrt{\frac{\beta}{2\pi}} e^{-\frac{\beta}{2}(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2}$$

- Simplifying further and ignoring the constants w.r.t. $\boldsymbol{w}$, we get

$$NLL(\boldsymbol{w}) = \frac{\beta}{2} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2$$

MLE for Probabilistic Linear Regression

- The negative log-likelihood will be

$$NLL(\boldsymbol{w}) = -\log p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}) = -\log \prod_{n=1}^{N} \sqrt{\frac{\beta}{2\pi}} e^{-\frac{\beta}{2}(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2}$$

- Simplifying further and ignoring the constants w.r.t. $\boldsymbol{w}$, we get

$$NLL(\boldsymbol{w}) = \frac{\beta}{2} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2$$

- Note that this objective has the same form as unregularized linear regression

MLE for Probabilistic Linear Regression

- The negative log-likelihood will be

$$NLL(\boldsymbol{w}) = -\log p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}) = -\log \prod_{n=1}^{N} \sqrt{\frac{\beta}{2\pi}} e^{-\frac{\beta}{2}(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2}$$

- Simplifying further and ignoring the constants w.r.t. $\boldsymbol{w}$, we get

$$NLL(\boldsymbol{w}) = \frac{\beta}{2} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2$$

- Note that this objective has the same form as unregularized linear regression
- Therefore the MLE solution will be $\hat{w}_{MLE} = \arg\min_{\boldsymbol{w}} NLL(\boldsymbol{w}) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{y}$

MAP Estimation for Probabilistic Linear Regression

- The negative log-prior will be (ignoring constants w.r.t. $\boldsymbol{w}$)

$$-\log p(\boldsymbol{w}|\lambda) = -\log \prod_{d=1}^{D} \sqrt{\frac{\lambda}{2\pi}} e^{-\frac{\lambda}{2} w_d^2} = \frac{\lambda}{2} \sum_{d=1}^{D} w_d^2 = \frac{\lambda}{2} \boldsymbol{w}^\top \boldsymbol{w}$$

MAP Estimation for Probabilistic Linear Regression

- The negative log-prior will be (ignoring constants w.r.t. $\boldsymbol{w}$)

$$-\log p(\boldsymbol{w}|\lambda) = -\log \prod_{d=1}^{D} \sqrt{\frac{\lambda}{2\pi}} e^{-\frac{\lambda}{2} w_d^2} = \frac{\lambda}{2} \sum_{d=1}^{D} w_d^2 = \frac{\lambda}{2} \boldsymbol{w}^\top \boldsymbol{w}$$

- Therefore the MAP solution $\hat{w}_{MAP} = \arg\min_{\boldsymbol{w}} NLL(\boldsymbol{w}) - \log p(\boldsymbol{w}|\lambda)$ will be

$$\hat{w}_{MAP} = \arg\min_{\boldsymbol{w}} \frac{\beta}{2} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2 + \frac{\lambda}{2} \sum_{d=1}^{D} w_d^2$$

MAP Estimation for Probabilistic Linear Regression

- The negative log-prior will be (ignoring constants w.r.t. $\boldsymbol{w}$)

$$-\log p(\boldsymbol{w}|\lambda) = -\log \prod_{d=1}^{D} \sqrt{\frac{\lambda}{2\pi}} e^{-\frac{\lambda}{2} w_d^2} = \frac{\lambda}{2} \sum_{d=1}^{D} w_d^2 = \frac{\lambda}{2} \boldsymbol{w}^\top \boldsymbol{w}$$

- Therefore the MAP solution $\hat{w}_{MAP} = \arg\min_{\boldsymbol{w}} NLL(\boldsymbol{w}) - \log p(\boldsymbol{w}|\lambda)$ will be

$$\hat{w}_{MAP} = \arg\min_{\boldsymbol{w}} \frac{\beta}{2} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2 + \frac{\lambda}{2} \sum_{d=1}^{D} w_d^2$$

- Can also write the above in the following form

$$\hat{w}_{MAP} = \arg\min_{\boldsymbol{w}} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2 + \frac{\lambda}{\beta} \boldsymbol{w}^\top \boldsymbol{w}$$

MAP Estimation for Probabilistic Linear Regression

- The negative log-prior will be (ignoring constants w.r.t. $\boldsymbol{w}$)

$$-\log p(\boldsymbol{w}|\lambda) = -\log \prod_{d=1}^{D} \sqrt{\frac{\lambda}{2\pi}} e^{-\frac{\lambda}{2} w_d^2} = \frac{\lambda}{2} \sum_{d=1}^{D} w_d^2 = \frac{\lambda}{2} \boldsymbol{w}^\top \boldsymbol{w}$$

- Therefore the MAP solution $\hat{w}_{MAP} = \arg\min_{\boldsymbol{w}} NLL(\boldsymbol{w}) - \log p(\boldsymbol{w}|\lambda)$ will be

$$\hat{w}_{MAP} = \arg\min_{\boldsymbol{w}} \frac{\beta}{2} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2 + \frac{\lambda}{2} \sum_{d=1}^{D} w_d^2$$

- Can also write the above in the following form

$$\hat{w}_{MAP} = \arg\min_{\boldsymbol{w}} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2 + \frac{\lambda}{\beta} \boldsymbol{w}^\top \boldsymbol{w}$$

- Note that this objective has the same form as regularized ("ridge") linear regression with $\tau = \frac{\lambda}{\beta}$

MAP Estimation for Probabilistic Linear Regression

- The negative log-prior will be (ignoring constants w.r.t. $\boldsymbol{w}$)

$$-\log p(\boldsymbol{w}|\lambda) = -\log \prod_{d=1}^{D} \sqrt{\frac{\lambda}{2\pi}} e^{-\frac{\lambda}{2} w_d^2} = \frac{\lambda}{2} \sum_{d=1}^{D} w_d^2 = \frac{\lambda}{2} \boldsymbol{w}^\top \boldsymbol{w}$$

- Therefore the MAP solution $\hat{w}_{MAP} = \arg\min_{\boldsymbol{w}} NLL(\boldsymbol{w}) - \log p(\boldsymbol{w}|\lambda)$ will be

$$\hat{w}_{MAP} = \arg\min_{\boldsymbol{w}} \frac{\beta}{2} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2 + \frac{\lambda}{2} \sum_{d=1}^{D} w_d^2$$

- Can also write the above in the following form

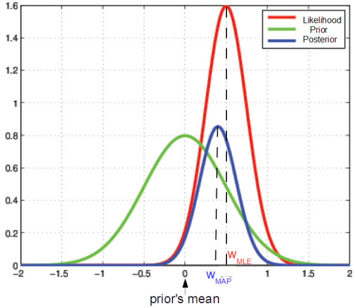
$$\hat{w}_{MAP} = \arg\min_{\boldsymbol{w}} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2 + \frac{\lambda}{\beta} \boldsymbol{w}^\top \boldsymbol{w}$$

- Note that this objective has the same form as regularized ("ridge") linear regression with $\tau = \frac{\lambda}{\beta}$

- Therefore the MAP solution will be $\hat{w}_{MAP} = \arg\min_{\boldsymbol{w}} NLL(\boldsymbol{w}) = (\mathbf{X}^\top \mathbf{X} + \frac{\lambda}{\beta} \mathbf{I})^{-1} \mathbf{X}^\top \boldsymbol{y}$

MLE vs MAP: An Illustration

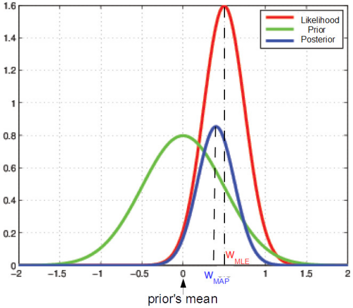
- $\boldsymbol{w}_{MAP}$ is a compromise between prior's mean (zero in this case) and $\boldsymbol{w}_{MLE}$



- Doing MAP shrinks the estimate of $\boldsymbol{w}$ towards the prior's mean

MLE vs MAP: An Illustration

- $\boldsymbol{w}_{MAP}$ is a compromise between prior's mean (zero in this case) and $\boldsymbol{w}_{MLE}$



- Doing MAP shrinks the estimate of $\boldsymbol{w}$ towards the prior's mean
- Note that a "peaked" prior (large precision) will pull the solution more towards the prior mean

MLE vs MAP for Probabilistic Linear Regression

- The MLE and MAP solutions are

$$\hat{w}_{MLE} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \boldsymbol{y}$$

$$\hat{w}_{MAP} = (\mathbf{X}^\top \mathbf{X} + \frac{\lambda}{\beta}\mathbf{I})^{-1}\mathbf{X}^\top \boldsymbol{y}$$

MLE vs MAP for Probabilistic Linear Regression

- The MLE and MAP solutions are

$$
\begin{aligned}
\hat{w}_{MLE} &= (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \boldsymbol{y} \\
\hat{w}_{MAP} &= (\mathbf{X}^\top \mathbf{X} + \frac{\lambda}{\beta}\mathbf{I})^{-1}\mathbf{X}^\top \boldsymbol{y}
\end{aligned}
$$

- MLE doesn't use a prior on $\boldsymbol{w}$ (akin to unregularized linear regression)

MLE vs MAP for Probabilistic Linear Regression

- The MLE and MAP solutions are

$$\begin{aligned}\hat{w}_{MLE} &= (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \boldsymbol{y} \\ \hat{w}_{MAP} &= (\mathbf{X}^\top \mathbf{X} + \frac{\lambda}{\beta}\mathbf{I})^{-1}\mathbf{X}^\top \boldsymbol{y}\end{aligned}$$

- MLE doesn't use a prior on $\boldsymbol{w}$ (akin to unregularized linear regression)
- MAP uses a Gaussian prior on $\boldsymbol{w}$ (which acts as a regularizer on $\boldsymbol{w}$)

MLE vs MAP for Probabilistic Linear Regression

- The MLE and MAP solutions are

$$
\begin{aligned}
\hat{w}_{MLE} &= (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \boldsymbol{y} \\
\hat{w}_{MAP} &= (\mathbf{X}^\top \mathbf{X} + \frac{\lambda}{\beta}\mathbf{I})^{-1}\mathbf{X}^\top \boldsymbol{y}
\end{aligned}
$$

- MLE doesn't use a prior on $\boldsymbol{w}$ (akin to unregularized linear regression)
- MAP uses a Gaussian prior on $\boldsymbol{w}$ (which acts as a regularizer on $\boldsymbol{w}$)
- As the prior's precision goes to zero, $\hat{w}_{MAP}$ approaches $\hat{w}_{MLE}$

MLE vs MAP for Probabilistic Linear Regression

- The MLE and MAP solutions are

$$\begin{aligned}\hat{w}_{MLE} &= (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \boldsymbol{y} \\ \hat{w}_{MAP} &= (\mathbf{X}^\top \mathbf{X} + \frac{\lambda}{\beta}\mathbf{I})^{-1}\mathbf{X}^\top \boldsymbol{y}\end{aligned}$$

- MLE doesn't use a prior on $\boldsymbol{w}$ (akin to unregularized linear regression)
- MAP uses a Gaussian prior on $\boldsymbol{w}$ (which acts as a regularizer on $\boldsymbol{w}$)
- As the prior's precision goes to zero, $\hat{w}_{MAP}$ approaches $\hat{w}_{MLE}$
  - Likewise, the higher the prior's precision λ is, the more *regularized* the solution will be

MLE vs MAP for Probabilistic Linear Regression

- The MLE and MAP solutions are

$$\begin{aligned}\hat{w}_{MLE} &= (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \boldsymbol{y} \\ \hat{w}_{MAP} &= (\mathbf{X}^\top \mathbf{X} + \frac{\lambda}{\beta}\mathbf{I})^{-1}\mathbf{X}^\top \boldsymbol{y}\end{aligned}$$

- MLE doesn't use a prior on $\boldsymbol{w}$ (akin to unregularized linear regression)
- MAP uses a Gaussian prior on $\boldsymbol{w}$ (which acts as a regularizer on $\boldsymbol{w}$)
- As the prior's precision goes to zero, $\hat{w}_{MAP}$ approaches $\hat{w}_{MLE}$
  - Likewise, the higher the prior's precision λ is, the more *regularized* the solution will be
- Prediction for a new test input $\boldsymbol{x}_*$ will be

$$\begin{aligned}\text{MLE} &: \quad p(y_*|\boldsymbol{w}_{MLE}, \boldsymbol{x}_*) = \mathcal{N}(y_*|\boldsymbol{w}_{MLE}^\top \boldsymbol{x}_*, \beta^{-1}) \\ \text{MAP} &: \quad p(y_*|\boldsymbol{w}_{MAP}, \boldsymbol{x}_*) = \mathcal{N}(y_*|\boldsymbol{w}_{MAP}^\top \boldsymbol{x}_*, \beta^{-1})\end{aligned}$$

MLE vs MAP for Probabilistic Linear Regression

- The MLE and MAP solutions are

$$\hat{w}_{MLE} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \boldsymbol{y}$$
$$\hat{w}_{MAP} = (\mathbf{X}^\top \mathbf{X} + \frac{\lambda}{\beta}\mathbf{I})^{-1}\mathbf{X}^\top \boldsymbol{y}$$

- MLE doesn't use a prior on $\boldsymbol{w}$ (akin to unregularized linear regression)
- MAP uses a Gaussian prior on $\boldsymbol{w}$ (which acts as a regularizer on $\boldsymbol{w}$)
- As the prior's precision goes to zero, $\hat{w}_{MAP}$ approaches $\hat{w}_{MLE}$
  - Likewise, the higher the prior's precision λ is, the more *regularized* the solution will be
- Prediction for a new test input $\boldsymbol{x}_*$ will be

$$\text{MLE} : p(y_*|\boldsymbol{w}_{MLE}, \boldsymbol{x}_*) = \mathcal{N}(y_*|\boldsymbol{w}_{MLE}^\top \boldsymbol{x}_*, \beta^{-1})$$
$$\text{MAP} : p(y_*|\boldsymbol{w}_{MAP}, \boldsymbol{x}_*) = \mathcal{N}(y_*|\boldsymbol{w}_{MAP}^\top \boldsymbol{x}_*, \beta^{-1})$$

- Predictive means are different but predictive variances are the same (and constant for all $\boldsymbol{x}$!)

Inferring the Posterior Distribution (fully Bayesian Inference)

- Inferring the full posterior is straightforward if the hyperparams β and λ to be known/fixed

Inferring the Posterior Distribution (fully Bayesian Inference)

- Inferring the full posterior is straightforward if the hyperparams β and λ to be known/fixed
  - Basically, the conjugacy helps here (Gaussian prior is conjugate to Gaussian likelihood)

Inferring the Posterior Distribution (fully Bayesian Inference)

- Inferring the full posterior is straightforward if the hyperparams β and λ to be known/fixed
  - Basically, the conjugacy helps here (Gaussian prior is conjugate to Gaussian likelihood)
- The posterior over the weight vector $\boldsymbol{w}$ (with β and λ known)

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

Inferring the Posterior Distribution (fully Bayesian Inference)

- Inferring the full posterior is straightforward if the hyperparams β and λ to be known/fixed
  - Basically, the conjugacy helps here (Gaussian prior is conjugate to Gaussian likelihood)
- The posterior over the weight vector $\boldsymbol{w}$ (with β and λ known)

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

- Computing $P(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda)$ (like Bernoulli-Beta case, doing it only upto proportionality constant)

$$P(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) \propto P(\boldsymbol{w}|\lambda)P(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}, \beta)$$

Inferring the Posterior Distribution (fully Bayesian Inference)

- Inferring the full posterior is straightforward if the hyperparams β and λ to be known/fixed
  - Basically, the conjugacy helps here (Gaussian prior is conjugate to Gaussian likelihood)
- The posterior over the weight vector $\boldsymbol{w}$ (with β and λ known)

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

- Computing $P(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda)$ (like Bernoulli-Beta case, doing it only upto proportionality constant)

$$P(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) \propto P(\boldsymbol{w}|\lambda)P(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}, \beta)$$

- After some algebra, this gets simplified into the following (proof on the next two slides)

$$P(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) \quad = \quad \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \text{(The posterior must be Gaussian due to conjugacy)}$$

Inferring the Posterior Distribution (fully Bayesian Inference)

- Inferring the full posterior is straightforward if the hyperparams β and λ to be known/fixed
  - Basically, the conjugacy helps here (Gaussian prior is conjugate to Gaussian likelihood)
- The posterior over the weight vector $\boldsymbol{w}$ (with β and λ known)

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

- Computing $P(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda)$ (like Bernoulli-Beta case, doing it only upto proportionality constant)

$$P(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) \propto P(\boldsymbol{w}|\lambda)P(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}, \beta)$$

- After some algebra, this gets simplified into the following (proof on the next two slides)

$$\begin{aligned} P(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) &= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \text{(The posterior must be Gaussian due to conjugacy)} \\ \text{where} \quad \boldsymbol{\Sigma} &= (\beta \sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top + \lambda \mathbf{I}_D)^{-1} = (\beta \mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_D)^{-1} \end{aligned}$$

Inferring the Posterior Distribution (fully Bayesian Inference)

- Inferring the full posterior is straightforward if the hyperparams β and λ to be known/fixed
  - Basically, the conjugacy helps here (Gaussian prior is conjugate to Gaussian likelihood)
- The posterior over the weight vector $\boldsymbol{w}$ (with β and λ known)

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

- Computing $P(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda)$ (like Bernoulli-Beta case, doing it only upto proportionality constant)

$$P(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) \propto P(\boldsymbol{w}|\lambda)P(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}, \beta)$$

- After some algebra, this gets simplified into the following (proof on the next two slides)

$$\begin{aligned}
P(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) &= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \text{(The posterior must be Gaussian due to conjugacy)} \\
\text{where} \quad \boldsymbol{\Sigma} &= (\beta \sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top + \lambda \mathbf{I}_D)^{-1} = (\beta \mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_D)^{-1} \\
\boldsymbol{\mu} &= \boldsymbol{\Sigma}(\beta \sum_{n=1}^{N} y_n \boldsymbol{x}_n) = \boldsymbol{\Sigma}(\beta \mathbf{X}^\top \boldsymbol{y}) = (\mathbf{X}^\top \mathbf{X} + \frac{\lambda}{\beta} \mathbf{I}_D)^{-1} \mathbf{X}^\top \boldsymbol{y}
\end{aligned}$$

The "Completing The Square" Trick for Gaussian Posterior

- Plugging in the respective distributions for $p(\boldsymbol{w}|\lambda)$ and $p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}, \beta)$, we will get

$$p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) \propto p(\boldsymbol{w}|\lambda)p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}, \beta) \quad = \quad \mathcal{N}(\boldsymbol{w}|\mathbf{0}, \lambda^{-1}\mathbf{I}_D)\mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{w}, \beta^{-1}\mathbf{I}_N)$$

The "Completing The Square" Trick for Gaussian Posterior

- Plugging in the respective distributions for $p(\boldsymbol{w}|\lambda)$ and $p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}, \beta)$, we will get

$$\begin{aligned} p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) \propto p(\boldsymbol{w}|\lambda)p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}, \beta) &= \mathcal{N}(\boldsymbol{w}|\mathbf{0}, \lambda^{-1}\mathbf{I}_D)\mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{w}, \beta^{-1}\mathbf{I}_N) \\ &\propto \exp\left(-\frac{\lambda}{2}\boldsymbol{w}^\top\boldsymbol{w}\right)\exp\left(-\frac{\beta}{2}(\boldsymbol{y}-\mathbf{X}\boldsymbol{w})^\top(\boldsymbol{y}-\mathbf{X}\boldsymbol{w})\right) \end{aligned}$$

The "Completing The Square" Trick for Gaussian Posterior

- Plugging in the respective distributions for $p(\boldsymbol{w}|\lambda)$ and $p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}, \beta)$, we will get

$$
\begin{aligned}
p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) \propto p(\boldsymbol{w}|\lambda)p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}, \beta) &= \mathcal{N}(\boldsymbol{w}|\mathbf{0}, \lambda^{-1}\mathbf{I}_D)\mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{w}, \beta^{-1}\mathbf{I}_N) \\
&\propto \exp\left(-\frac{\lambda}{2}\boldsymbol{w}^\top\boldsymbol{w}\right)\exp\left(-\frac{\beta}{2}(\boldsymbol{y}-\mathbf{X}\boldsymbol{w})^\top(\boldsymbol{y}-\mathbf{X}\boldsymbol{w})\right) \\
&= \exp\left[-\frac{\lambda}{2}\boldsymbol{w}^\top\boldsymbol{w} - \frac{\beta}{2}(\boldsymbol{y}^\top\boldsymbol{y} + \boldsymbol{w}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{w} - 2\boldsymbol{w}^\top\mathbf{X}^\top\boldsymbol{y})\right]
\end{aligned}
$$

The "Completing The Square" Trick for Gaussian Posterior

- Plugging in the respective distributions for $p(\boldsymbol{w}|\lambda)$ and $p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}, \beta)$, we will get

$$\begin{aligned}
p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) \propto p(\boldsymbol{w}|\lambda)p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}, \beta) &= \mathcal{N}(\boldsymbol{w}|\mathbf{0}, \lambda^{-1}\mathbf{I}_D)\mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{w}, \beta^{-1}\mathbf{I}_N) \\
&\propto \exp\left(-\frac{\lambda}{2}\boldsymbol{w}^\top\boldsymbol{w}\right)\exp\left(-\frac{\beta}{2}(\boldsymbol{y}-\mathbf{X}\boldsymbol{w})^\top(\boldsymbol{y}-\mathbf{X}\boldsymbol{w})\right) \\
&= \exp\left[-\frac{\lambda}{2}\boldsymbol{w}^\top\boldsymbol{w} - \frac{\beta}{2}(\boldsymbol{y}^\top\boldsymbol{y} + \boldsymbol{w}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{w} - 2\boldsymbol{w}^\top\mathbf{X}^\top\boldsymbol{y})\right] \\
&\propto \exp\left[-\frac{\lambda}{2}\boldsymbol{w}^\top\boldsymbol{w} - \frac{\beta}{2}(\boldsymbol{w}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{w} - 2\boldsymbol{w}^\top\mathbf{X}^\top\boldsymbol{y})\right]
\end{aligned}$$

The "Completing The Square" Trick for Gaussian Posterior

- Plugging in the respective distributions for $p(\boldsymbol{w}|\lambda)$ and $p(\boldsymbol{y}|\mathbf{X},\boldsymbol{w},\beta)$, we will get

$$
\begin{aligned}
p(\boldsymbol{w}|\mathbf{X},\boldsymbol{y},\beta,\lambda) \propto p(\boldsymbol{w}|\lambda)p(\boldsymbol{y}|\mathbf{X},\boldsymbol{w},\beta) &= \mathcal{N}(\boldsymbol{w}|\mathbf{0},\lambda^{-1}\mathbf{I}_D)\mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{w},\beta^{-1}\mathbf{I}_N) \\
&\propto \exp\left(-\frac{\lambda}{2}\boldsymbol{w}^\top\boldsymbol{w}\right)\exp\left(-\frac{\beta}{2}(\boldsymbol{y}-\mathbf{X}\boldsymbol{w})^\top(\boldsymbol{y}-\mathbf{X}\boldsymbol{w})\right) \\
&= \exp\left[-\frac{\lambda}{2}\boldsymbol{w}^\top\boldsymbol{w}-\frac{\beta}{2}(\boldsymbol{y}^\top\boldsymbol{y}+\boldsymbol{w}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{w}-2\boldsymbol{w}^\top\mathbf{X}^\top\boldsymbol{y})\right] \\
&\propto \exp\left[-\frac{\lambda}{2}\boldsymbol{w}^\top\boldsymbol{w}-\frac{\beta}{2}(\boldsymbol{w}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{w}-2\boldsymbol{w}^\top\mathbf{X}^\top\boldsymbol{y})\right] \\
&= \exp\left[-\frac{1}{2}\left(\boldsymbol{w}^\top(\lambda\mathbf{I}_D+\beta\mathbf{X}^\top\mathbf{X})\boldsymbol{w}-2\beta\boldsymbol{w}^\top\mathbf{X}^\top\boldsymbol{y}\right)\right]
\end{aligned}
$$

The "Completing The Square" Trick for Gaussian Posterior

- Plugging in the respective distributions for $p(\boldsymbol{w}|\lambda)$ and $p(\boldsymbol{y}|\mathbf{X},\boldsymbol{w},\beta)$, we will get

$$\begin{aligned}
p(\boldsymbol{w}|\mathbf{X},\boldsymbol{y},\beta,\lambda) \propto p(\boldsymbol{w}|\lambda)p(\boldsymbol{y}|\mathbf{X},\boldsymbol{w},\beta) &= \mathcal{N}(\boldsymbol{w}|\mathbf{0},\lambda^{-1}\mathbf{I}_D)\mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{w},\beta^{-1}\mathbf{I}_N) \\
&\propto \exp\left(-\frac{\lambda}{2}\boldsymbol{w}^\top\boldsymbol{w}\right)\exp\left(-\frac{\beta}{2}(\boldsymbol{y}-\mathbf{X}\boldsymbol{w})^\top(\boldsymbol{y}-\mathbf{X}\boldsymbol{w})\right) \\
&= \exp\left[-\frac{\lambda}{2}\boldsymbol{w}^\top\boldsymbol{w}-\frac{\beta}{2}(\boldsymbol{y}^\top\boldsymbol{y}+\boldsymbol{w}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{w}-2\boldsymbol{w}^\top\mathbf{X}^\top\boldsymbol{y})\right] \\
&\propto \exp\left[-\frac{\lambda}{2}\boldsymbol{w}^\top\boldsymbol{w}-\frac{\beta}{2}(\boldsymbol{w}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{w}-2\boldsymbol{w}^\top\mathbf{X}^\top\boldsymbol{y})\right] \\
&= \exp\left[-\frac{1}{2}\left(\boldsymbol{w}^\top(\lambda\mathbf{I}_D+\beta\mathbf{X}^\top\mathbf{X})\boldsymbol{w}-2\beta\boldsymbol{w}^\top\mathbf{X}^\top\boldsymbol{y}\right)\right]
\end{aligned}$$

- We will now try to bring the exponent into a quadratic form to see if it corresponds to some Gaussian. So basically, we will use the "complete the square" trick

The “Completing The Square” Trick for Gaussian Posterior

- So we had.. $p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{w}^\top(\lambda \mathbf{I}_D + \beta \mathbf{X}^\top \mathbf{X})\boldsymbol{w} - 2\beta \boldsymbol{w}^\top \mathbf{X}^\top \boldsymbol{y}\right)\right]$

The "Completing The Square" Trick for Gaussian Posterior

- So we had.. $p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{w}^\top(\lambda \mathbf{I}_D + \beta \mathbf{X}^\top \mathbf{X})\boldsymbol{w} - 2\beta \boldsymbol{w}^\top \mathbf{X}^\top \boldsymbol{y}\right)\right]$
- Let's see if we can bring the above posterior into the form of the following Gaussian

$$\mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{w} - \boldsymbol{\mu})\right] = \exp\left[-\frac{1}{2}(\boldsymbol{w}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{w} - 2\boldsymbol{w}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu})\right]$$

The "Completing The Square" Trick for Gaussian Posterior

- So we had.. $p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{w}^\top(\lambda \mathbf{I}_D + \beta \mathbf{X}^\top \mathbf{X})\boldsymbol{w} - 2\beta \boldsymbol{w}^\top \mathbf{X}^\top \boldsymbol{y}\right)\right]$
- Let's see if we can bring the above posterior into the form of the following Gaussian

$$\mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{w} - \boldsymbol{\mu})\right] = \exp\left[-\frac{1}{2}(\boldsymbol{w}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{w} - 2\boldsymbol{w}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right]$$

- Let's multiply and divide $p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{w}^\top(\lambda \mathbf{I}_D + \beta \mathbf{X}^\top \mathbf{X})\boldsymbol{w} - 2\beta \boldsymbol{w}^\top \mathbf{X}^\top \boldsymbol{y}\right)\right]$ by $\exp\left[-\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right]$

The "Completing The Square" Trick for Gaussian Posterior

- So we had.. $p(\boldsymbol{w}|\mathbf{X},\boldsymbol{y},\beta,\lambda) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{w}^\top(\lambda\mathbf{I}_D + \beta\mathbf{X}^\top\mathbf{X})\boldsymbol{w} - 2\beta\boldsymbol{w}^\top\mathbf{X}^\top\boldsymbol{y}\right)\right]$
- Let's see if we can bring the above posterior into the form of the following Gaussian

$$\mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{w}-\boldsymbol{\mu})\right] = \exp\left[-\frac{1}{2}(\boldsymbol{w}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{w} - 2\boldsymbol{w}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right]$$

- Let's multiply and divide $p(\boldsymbol{w}|\mathbf{X},\boldsymbol{y},\beta,\lambda) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{w}^\top(\lambda\mathbf{I}_D + \beta\mathbf{X}^\top\mathbf{X})\boldsymbol{w} - 2\beta\boldsymbol{w}^\top\mathbf{X}^\top\boldsymbol{y}\right)\right]$ by $\exp\left[-\frac{1}{2}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right]$
- This gives the following up to a prop. constant (remember $\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ is constant w.r.t. $\boldsymbol{w}$):

$$p(\boldsymbol{w}|\mathbf{X},\boldsymbol{y},\beta,\lambda) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{w}^\top(\lambda\mathbf{I}_D + \beta\mathbf{X}^\top\mathbf{X})\boldsymbol{w} - 2\beta\boldsymbol{w}^\top\mathbf{X}^\top\boldsymbol{y} + \boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right]$$

The "Completing The Square" Trick for Gaussian Posterior

- So we had.. $p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{w}^\top(\lambda \mathbf{I}_D + \beta \mathbf{X}^\top \mathbf{X})\boldsymbol{w} - 2\beta \boldsymbol{w}^\top \mathbf{X}^\top \boldsymbol{y}\right)\right]$
- Let's see if we can bring the above posterior into the form of the following Gaussian

$$\mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{w} - \boldsymbol{\mu})\right] = \exp\left[-\frac{1}{2}(\boldsymbol{w}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{w} - 2\boldsymbol{w}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right]$$

- Let's multiply and divide $p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{w}^\top(\lambda \mathbf{I}_D + \beta \mathbf{X}^\top \mathbf{X})\boldsymbol{w} - 2\beta \boldsymbol{w}^\top \mathbf{X}^\top \boldsymbol{y}\right)\right]$ by $\exp\left[-\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right]$
- This gives the following up to a prop. constant (remember $\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ is constant w.r.t. $\boldsymbol{w}$):

$$p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{w}^\top(\lambda \mathbf{I}_D + \beta \mathbf{X}^\top \mathbf{X})\boldsymbol{w} - 2\beta \boldsymbol{w}^\top \mathbf{X}^\top \boldsymbol{y} + \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right]$$

- Finally comparing with the expression of $\mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ we can see that

$$\begin{aligned} \boldsymbol{\Sigma} &= (\lambda \mathbf{I}_D + \beta \mathbf{X}^\top \mathbf{X})^{-1} \\ \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} &= \beta \mathbf{X}^\top \boldsymbol{y} \quad \Rightarrow \quad \boldsymbol{\mu} = \boldsymbol{\Sigma}(\beta \mathbf{X}^\top \boldsymbol{y}) \end{aligned}$$

The "Completing The Square" Trick for Gaussian Posterior

- So we had.. $p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{w}^\top(\lambda \mathbf{I}_D + \beta \mathbf{X}^\top \mathbf{X})\boldsymbol{w} - 2\beta \boldsymbol{w}^\top \mathbf{X}^\top \boldsymbol{y}\right)\right]$
- Let's see if we can bring the above posterior into the form of the following Gaussian

$$\mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{w}-\boldsymbol{\mu})\right] = \exp\left[-\frac{1}{2}(\boldsymbol{w}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{w} - 2\boldsymbol{w}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right]$$

- Let's multiply and divide $p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{w}^\top(\lambda \mathbf{I}_D + \beta \mathbf{X}^\top \mathbf{X})\boldsymbol{w} - 2\beta \boldsymbol{w}^\top \mathbf{X}^\top \boldsymbol{y}\right)\right]$ by $\exp\left[-\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right]$
- This gives the following up to a prop. constant (remember $\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ is constant w.r.t. $\boldsymbol{w}$):

$$p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{w}^\top(\lambda \mathbf{I}_D + \beta \mathbf{X}^\top \mathbf{X})\boldsymbol{w} - 2\beta \boldsymbol{w}^\top \mathbf{X}^\top \boldsymbol{y} + \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right]$$

- Finally comparing with the expression of $\mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ we can see that

$$\begin{aligned} \boldsymbol{\Sigma} &= (\lambda \mathbf{I}_D + \beta \mathbf{X}^\top \mathbf{X})^{-1} \\ \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} &= \beta \mathbf{X}^\top \boldsymbol{y} \quad \Rightarrow \quad \boldsymbol{\mu} = \boldsymbol{\Sigma}(\beta \mathbf{X}^\top \boldsymbol{y}) \end{aligned}$$

- Note: The above expression for the posterior can also be directly obtained using properties of Gaussian distributions (we will see those in the coming lectures)

Information within the Posterior

- As we saw, the posterior over the weight vector $\boldsymbol{w}$

$$
\begin{aligned}
P(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) &= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
\text{where} \quad \boldsymbol{\Sigma} &= (\beta \sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top + \lambda \mathbf{I}_D)^{-1} = (\beta \mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_D)^{-1} \\
\boldsymbol{\mu} &= \boldsymbol{\Sigma}(\beta \sum_{n=1}^{N} y_n \boldsymbol{x}_n) = \boldsymbol{\Sigma}(\beta \mathbf{X}^\top \boldsymbol{y}) = (\mathbf{X}^\top \mathbf{X} + \frac{\lambda}{\beta} \mathbf{I}_D)^{-1} \mathbf{X}^\top \boldsymbol{y}
\end{aligned}
$$

Information within the Posterior

- As we saw, the posterior over the weight vector $\boldsymbol{w}$

$$
\begin{aligned}
P(\boldsymbol{w}|\mathbf{X},\boldsymbol{y},\beta,\lambda) &= \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}) \\
\text{where} \quad \boldsymbol{\Sigma} &= (\beta\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^\top + \lambda\mathbf{I}_D)^{-1} = (\beta\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_D)^{-1} \\
\boldsymbol{\mu} &= \boldsymbol{\Sigma}(\beta\sum_{n=1}^{N}y_n\boldsymbol{x}_n) = \boldsymbol{\Sigma}(\beta\mathbf{X}^\top\boldsymbol{y}) = (\mathbf{X}^\top\mathbf{X} + \frac{\lambda}{\beta}\mathbf{I}_D)^{-1}\mathbf{X}^\top\boldsymbol{y}
\end{aligned}
$$

- Note the form of the mean $\boldsymbol{\mu}$ of the posterior. Same as the MAP estimate $\boldsymbol{w}_{MAP}$

Information within the Posterior

- As we saw, the posterior over the weight vector $\boldsymbol{w}$

$$\begin{aligned} P(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) &= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \text{where} \quad \boldsymbol{\Sigma} &= (\beta \sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top + \lambda \mathbf{I}_D)^{-1} = (\beta \mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_D)^{-1} \\ \boldsymbol{\mu} &= \boldsymbol{\Sigma}(\beta \sum_{n=1}^{N} y_n \boldsymbol{x}_n) = \boldsymbol{\Sigma}(\beta \mathbf{X}^\top \boldsymbol{y}) = (\mathbf{X}^\top \mathbf{X} + \frac{\lambda}{\beta} \mathbf{I}_D)^{-1} \mathbf{X}^\top \boldsymbol{y} \end{aligned}$$

- Note the form of the mean $\boldsymbol{\mu}$ of the posterior. Same as the MAP estimate $\boldsymbol{w}_{MAP}$
  - Reason: For a Gaussian distribution, mean is the same as the mode (maxima of the distribution)

Information within the Posterior

- As we saw, the posterior over the weight vector $\boldsymbol{w}$

$$
\begin{aligned}
P(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) &= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
\text{where} \quad \boldsymbol{\Sigma} &= (\beta \sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top + \lambda \mathbf{I}_D)^{-1} = (\beta \mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_D)^{-1} \\
\boldsymbol{\mu} &= \boldsymbol{\Sigma}(\beta \sum_{n=1}^{N} y_n \boldsymbol{x}_n) = \boldsymbol{\Sigma}(\beta \mathbf{X}^\top \boldsymbol{y}) = (\mathbf{X}^\top \mathbf{X} + \frac{\lambda}{\beta} \mathbf{I}_D)^{-1} \mathbf{X}^\top \boldsymbol{y}
\end{aligned}
$$

- Note the form of the mean $\boldsymbol{\mu}$ of the posterior. Same as the MAP estimate $\boldsymbol{w}_{MAP}$
  - Reason: For a Gaussian distribution, mean is the same as the mode (maxima of the distribution)
- Setting λ to 0 makes μ equivalent to the MLE solution

Information within the Posterior

- As we saw, the posterior over the weight vector $\boldsymbol{w}$

$$
\begin{aligned}
P(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) &= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
\text{where} \quad \boldsymbol{\Sigma} &= (\beta \sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top + \lambda \mathbf{I}_D)^{-1} = (\beta \mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_D)^{-1} \\
\boldsymbol{\mu} &= \boldsymbol{\Sigma}(\beta \sum_{n=1}^{N} y_n \boldsymbol{x}_n) = \boldsymbol{\Sigma}(\beta \mathbf{X}^\top \boldsymbol{y}) = (\mathbf{X}^\top \mathbf{X} + \frac{\lambda}{\beta} \mathbf{I}_D)^{-1} \mathbf{X}^\top \boldsymbol{y}
\end{aligned}
$$

- Note the form of the mean $\boldsymbol{\mu}$ of the posterior. Same as the MAP estimate $\boldsymbol{w}_{MAP}$
  - Reason: For a Gaussian distribution, mean is the same as the mode (maxima of the distribution)
- Setting λ to 0 makes μ equivalent to the MLE solution
- However, MLE and MAP are only point estimates whereas we now have the full posterior over $\boldsymbol{w}$

Information within the Posterior

- As we saw, the posterior over the weight vector $\boldsymbol{w}$

$$
\begin{aligned}
P(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) &= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
\text{where} \quad \boldsymbol{\Sigma} &= (\beta \sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top + \lambda \mathbf{I}_D)^{-1} = (\beta \mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_D)^{-1} \\
\boldsymbol{\mu} &= \boldsymbol{\Sigma}(\beta \sum_{n=1}^{N} y_n \boldsymbol{x}_n) = \boldsymbol{\Sigma}(\beta \mathbf{X}^\top \boldsymbol{y}) = (\mathbf{X}^\top \mathbf{X} + \frac{\lambda}{\beta} \mathbf{I}_D)^{-1} \mathbf{X}^\top \boldsymbol{y}
\end{aligned}
$$

- Note the form of the mean $\boldsymbol{\mu}$ of the posterior. Same as the MAP estimate $\boldsymbol{w}_{MAP}$
  - Reason: For a Gaussian distribution, mean is the same as the mode (maxima of the distribution)
- Setting λ to 0 makes μ equivalent to the MLE solution
- However, MLE and MAP are only point estimates whereas we now have the full posterior over $\boldsymbol{w}$
- Therefore, the solution given by the full posterior in a way subsumes MAP/MLE solutions

Posterior Predictive Distribution

- Now we want to use the posterior over $\boldsymbol{w}$ to make predictions (response y_* for a new input $\boldsymbol{x}_*$)

Posterior Predictive Distribution

- Now we want to use the posterior over $\boldsymbol{w}$ to make predictions (response y_* for a new input $\boldsymbol{x}_*$)
- The posterior predictive distribution in this case will be

$$p(y_*|\boldsymbol{x}_*, \mathbf{X}, \boldsymbol{y}, , \beta, \lambda) = \int p(y_*|\boldsymbol{x}_*, \boldsymbol{w}, , \beta)p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, , \beta, \lambda)d\boldsymbol{w}$$

Posterior Predictive Distribution

- Now we want to use the posterior over $\boldsymbol{w}$ to make predictions (response y_* for a new input $\boldsymbol{x}_*$)
- The posterior predictive distribution in this case will be

$$p(y_*|\boldsymbol{x}_*, \mathbf{X}, \boldsymbol{y}, , \beta, \lambda) = \int p(y_*|\boldsymbol{x}_*, \boldsymbol{w}, , \beta) p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, , \beta, \lambda) d\boldsymbol{w}$$

- Both terms on RHS are Gaussians: $p(y_*|\boldsymbol{x}_*, \boldsymbol{w}, , \beta) = \mathcal{N}(y_*|\boldsymbol{w}^\top \boldsymbol{x}_*, \beta^{-1})$ and $p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Posterior Predictive Distribution

- Now we want to use the posterior over $\boldsymbol{w}$ to make predictions (response y_* for a new input $\boldsymbol{x}_*$)
- The posterior predictive distribution in this case will be

$$p(y_*|\boldsymbol{x}_*, \mathbf{X}, \boldsymbol{y}, , \beta, \lambda) = \int p(y_*|\boldsymbol{x}_*, \boldsymbol{w}, , \beta) p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, , \beta, \lambda) d\boldsymbol{w}$$

- Both terms on RHS are Gaussians: $p(y_*|\boldsymbol{x}_*, \boldsymbol{w}, , \beta) = \mathcal{N}(y_*|\boldsymbol{w}^\top \boldsymbol{x}_*, \beta^{-1})$ and $p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- The result and therefore the posterior predictive will be another Gaussian (we will see proof later)

$$\boxed{p(y_*|\boldsymbol{x}_*, \mathbf{X}, \boldsymbol{y}, , \beta, \lambda) = \mathcal{N}(\boldsymbol{\mu}^\top \boldsymbol{x}_*, \beta^{-1} + \boldsymbol{x}_*^\top \boldsymbol{\Sigma} \boldsymbol{x}_*)}$$

Posterior Predictive Distribution

- Now we want to use the posterior over $\boldsymbol{w}$ to make predictions (response y_* for a new input $\boldsymbol{x}_*$)
- The posterior predictive distribution in this case will be

$$p(y_*|\boldsymbol{x}_*, \mathbf{X}, \boldsymbol{y}, , \beta, \lambda) = \int p(y_*|\boldsymbol{x}_*, \boldsymbol{w}, , \beta) p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, , \beta, \lambda) d\boldsymbol{w}$$

- Both terms on RHS are Gaussians: $p(y_*|\boldsymbol{x}_*, \boldsymbol{w}, , \beta) = \mathcal{N}(y_*|\boldsymbol{w}^\top \boldsymbol{x}_*, \beta^{-1})$ and $p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- The result and therefore the posterior predictive will be another Gaussian (we will see proof later)

$$\boxed{p(y_*|\boldsymbol{x}_*, \mathbf{X}, \boldsymbol{y}, , \beta, \lambda) = \mathcal{N}(\boldsymbol{\mu}^\top \boldsymbol{x}_*, \beta^{-1} + \boldsymbol{x}_*^\top \boldsymbol{\Sigma} \boldsymbol{x}_*)}$$

- In contrast, MLE and MAP can only give us "plug-in" predictions (using the point estimate of $\boldsymbol{w}$)

$$\begin{aligned} p(y_*|\boldsymbol{x}_*, \boldsymbol{w}_{MLE}) &= \mathcal{N}(\boldsymbol{w}_{MLE}^\top \boldsymbol{x}_*, \beta^{-1}) && \text{- MLE prediction} \\ p(y_*|\boldsymbol{x}_*, \boldsymbol{w}_{MAP}) &= \mathcal{N}(\boldsymbol{w}_{MAP}^\top \boldsymbol{x}_*, \beta^{-1}) && \text{- MAP prediction} \end{aligned}$$

Posterior Predictive Distribution

- Now we want to use the posterior over $\boldsymbol{w}$ to make predictions (response y_* for a new input $\boldsymbol{x}_*$)
- The posterior predictive distribution in this case will be

$$p(y_*|\boldsymbol{x}_*, \mathbf{X}, \boldsymbol{y}, , \beta, \lambda) = \int p(y_*|\boldsymbol{x}_*, \boldsymbol{w}, , \beta) p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, , \beta, \lambda) d\boldsymbol{w}$$

- Both terms on RHS are Gaussians: $p(y_*|\boldsymbol{x}_*, \boldsymbol{w}, , \beta) = \mathcal{N}(y_*|\boldsymbol{w}^\top \boldsymbol{x}_*, \beta^{-1})$ and $p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- The result and therefore the posterior predictive will be another Gaussian (we will see proof later)

$$\boxed{p(y_*|\boldsymbol{x}_*, \mathbf{X}, \boldsymbol{y}, , \beta, \lambda) = \mathcal{N}(\boldsymbol{\mu}^\top \boldsymbol{x}_*, \beta^{-1} + \boldsymbol{x}_*^\top \boldsymbol{\Sigma} \boldsymbol{x}_*)}$$

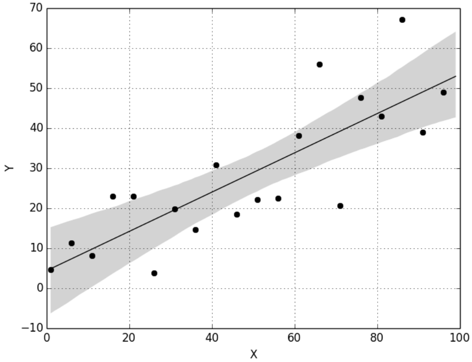
- In contrast, MLE and MAP can only give us "plug-in" predictions (using the point estimate of $\boldsymbol{w}$)

$$\begin{aligned} p(y_*|\boldsymbol{x}_*, \boldsymbol{w}_{MLE}) &= \mathcal{N}(\boldsymbol{w}_{MLE}^\top \boldsymbol{x}_*, \beta^{-1}) \quad \text{- MLE prediction} \\ p(y_*|\boldsymbol{x}_*, \boldsymbol{w}_{MAP}) &= \mathcal{N}(\boldsymbol{w}_{MAP}^\top \boldsymbol{x}_*, \beta^{-1}) \quad \text{- MAP prediction} \end{aligned}$$

- Important: Unlike MLE/MAP, the variance of y_* also depends on the input $\boldsymbol{x}_*$ (this, as we will see later, will be very useful in sequential decision-making problems such as active learning)

Posterior Predictive Distribution: An Illustration

Black dots are training examples



Regions with more training examples have smaller predictive variance

Hyperparameter Estimation

- So far we assumed that only $\boldsymbol{w}$ is the unknown and the hyperparameters λ and β are known

Hyperparameter Estimation

- So far we assumed that only $\boldsymbol{w}$ is the unknown and the hyperparameters λ and β are known
- One way to choose the "optimal" values of the hyperparams is cross-validation

Hyperparameter Estimation

- So far we assumed that only $\boldsymbol{w}$ is the unknown and the hyperparameters λ and β are known
- One way to choose the "optimal" values of the hyperparams is cross-validation
- However, we can also learn these hyperparameters

Hyperparameter Estimation

- So far we assumed that only $\boldsymbol{w}$ is the unknown and the hyperparameters λ and β are known
- One way to choose the "optimal" values of the hyperparams is cross-validation
- However, we can also learn these hyperparameters
  - Marginal likelihood maximization (also called MLE-II) is one such approach

$$\hat{\lambda}, \hat{\beta} = \arg\max_{\lambda,\beta} \; p(\boldsymbol{y}|\lambda, \beta, \mathbf{X})$$

where $p(\boldsymbol{y}|\lambda, \beta, \mathbf{X}) = \int_{\boldsymbol{w}} p(\boldsymbol{y}|\boldsymbol{w}, \lambda, \beta, \mathbf{X}) p(\boldsymbol{w}|\lambda) d\boldsymbol{w}$ is the marginal likelihood

Hyperparameter Estimation

- So far we assumed that only $\boldsymbol{w}$ is the unknown and the hyperparameters λ and β are known
- One way to choose the "optimal" values of the hyperparams is cross-validation
- However, we can also learn these hyperparameters
  - Marginal likelihood maximization (also called MLE-II) is one such approach

$$\hat{\lambda}, \hat{\beta} = \arg\max_{\lambda, \beta} \; p(\boldsymbol{y}|\lambda, \beta, \mathbf{X})$$

    where $p(\boldsymbol{y}|\lambda, \beta, \mathbf{X}) = \int_{\boldsymbol{w}} p(\boldsymbol{y}|\boldsymbol{w}, \lambda, \beta, \mathbf{X}) p(\boldsymbol{w}|\lambda) d\boldsymbol{w}$ is the marginal likelihood

  - Note: This optimization problem may not have a closed-form solution (iterative methods can be used)

Hyperparameter Estimation

- So far we assumed that only $\boldsymbol{w}$ is the unknown and the hyperparameters λ and β are known
- One way to choose the "optimal" values of the hyperparams is cross-validation
- However, we can also learn these hyperparameters
  - Marginal likelihood maximization (also called MLE-II) is one such approach

$$\hat{\lambda}, \hat{\beta} = \arg\max_{\lambda,\beta} \; p(\boldsymbol{y}|\lambda, \beta, \mathbf{X})$$

  where $p(\boldsymbol{y}|\lambda, \beta, \mathbf{X}) = \int_{\boldsymbol{w}} p(\boldsymbol{y}|\boldsymbol{w}, \lambda, \beta, \mathbf{X}) p(\boldsymbol{w}|\lambda) d\boldsymbol{w}$ is the marginal likelihood

  - Note: This optimization problem may not have a closed-form solution (iterative methods can be used)
  - Marginal likelihood is also known as "evidence"

Hyperparameter Estimation

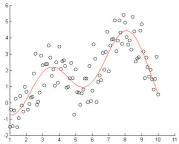
- So far we assumed that only $\boldsymbol{w}$ is the unknown and the hyperparameters λ and β are known
- One way to choose the "optimal" values of the hyperparams is cross-validation
- However, we can also learn these hyperparameters
  - Marginal likelihood maximization (also called MLE-II) is one such approach

$$\hat{\lambda}, \hat{\beta} = \arg\max_{\lambda,\beta} \; p(\boldsymbol{y}|\lambda, \beta, \mathbf{X})$$

    where $p(\boldsymbol{y}|\lambda, \beta, \mathbf{X}) = \int_{\boldsymbol{w}} p(\boldsymbol{y}|\boldsymbol{w}, \lambda, \beta, \mathbf{X}) p(\boldsymbol{w}|\lambda) d\boldsymbol{w}$ is the marginal likelihood

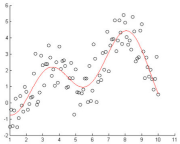
  - Note: This optimization problem may not have a closed-form solution (iterative methods can be used)
  - Marginal likelihood is also known as "evidence"
  - This procedure of hyperparameter estimation is also called "evidence maximization"

Probabilistic (Nonlinear) Regression



- Many ways to extend the probabilistic linear regression model to nonlinear regression.

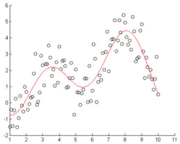
Probabilistic (Nonlinear) Regression



- Many ways to extend the probabilistic linear regression model to nonlinear regression. E.g.,
  - Replace $\boldsymbol{x}$ by a pre-defined higher-dim. feature mapping $\phi(\boldsymbol{x})$ (e.g. $x \to (1, x, x^2)$) and apply the linear model on the new feature mapped version of the inputs

$$\text{Likelihood model:} \quad p(y|\boldsymbol{x}, \boldsymbol{w}) = \mathcal{N}(y|\boldsymbol{w}^\top \phi(\boldsymbol{x}), \beta^{-1})$$

Probabilistic (Nonlinear) Regression



- Many ways to extend the probabilistic linear regression model to nonlinear regression. E.g.,
  - Replace $\boldsymbol{x}$ by a pre-defined higher-dim. feature mapping $\phi(\boldsymbol{x})$ (e.g. $x \rightarrow (1, x, x^2)$) and apply the linear model on the new feature mapped version of the inputs

$$\text{Likelihood model:} \quad p(y|\boldsymbol{x}, \boldsymbol{w}) = \mathcal{N}(y|\boldsymbol{w}^\top \phi(\boldsymbol{x}), \beta^{-1})$$

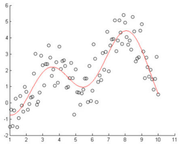
  .. also related to kernel methods

  - Replace the linear model $\boldsymbol{w}^\top \boldsymbol{x}$ by a nonlinear function f (e.g., a neural network)

$$p(y|\boldsymbol{x}, \boldsymbol{w}) = \mathcal{N}(y|f(\boldsymbol{x}), \beta^{-1})$$

  .. more on this when we discuss deep probabilistic models

Probabilistic (Nonlinear) Regression



- Many ways to extend the probabilistic linear regression model to nonlinear regression. E.g.,
  - Replace $\boldsymbol{x}$ by a pre-defined higher-dim. feature mapping $\phi(\boldsymbol{x})$ (e.g. $x \to (1, x, x^2)$) and apply the linear model on the new feature mapped version of the inputs

$$\text{Likelihood model:} \quad p(y|\boldsymbol{x}, \boldsymbol{w}) = \mathcal{N}(y|\boldsymbol{w}^\top \phi(\boldsymbol{x}), \beta^{-1})$$

    .. also related to kernel methods
  - Replace the linear model $\boldsymbol{w}^\top \boldsymbol{x}$ by a nonlinear function f (e.g., a neural network)

$$p(y|\boldsymbol{x}, \boldsymbol{w}) = \mathcal{N}(y|f(\boldsymbol{x}), \beta^{-1})$$

    .. more on this when we discuss deep probabilistic models
  - Use Gaussian Processes