ISTA 421/521
Introduction to Machine Learning

Lecture 7:
Continuous Prob., Gaussians, Maximum Likelihood

Clay Morrison

clayton@sista.arizona.edu

Gould-Simpson 927B

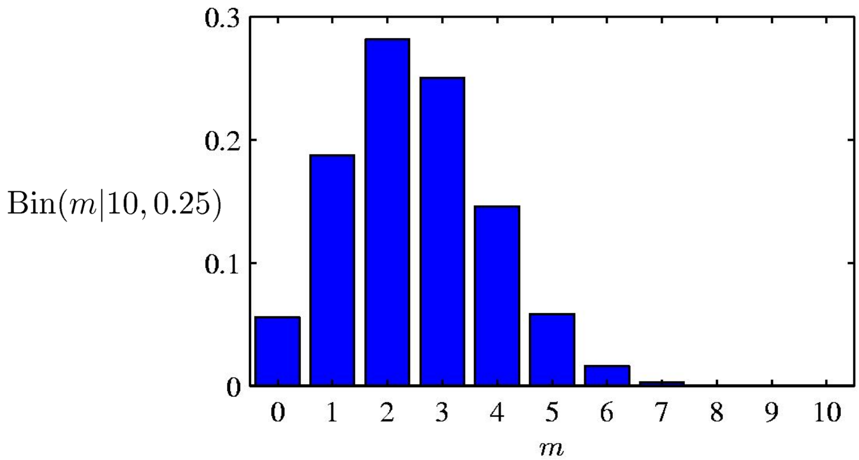
Phone 621-6609

16 September 2012

Next Topics

- Probability Basics
- Expectation and Random Vectors
- Discrete Probability (examples)
- Continuous Probability
- Gaussian Distribution
- Maximum Likelihood Estimation

Binomial Distribution



$$\mathrm{Bin}(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m}$$

Continuous Random Variables:

- Unlike a discrete space, you can't assign probability to a point in a continuous space
- Instead, we assign probabilities to ***regions*** (within some range or interval).
  - E.g., continuous random variable X

    $P(x_1 \leq X \leq x_2)$ but not $P(X = x)$

Continuous Random Variables:

- The continuous analog to a probability distribution: *probability density function* (pdf)
  - And to compute the probability that X lies in some range, we compute the definite integral of the fn:

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} p(x)\, dx$$

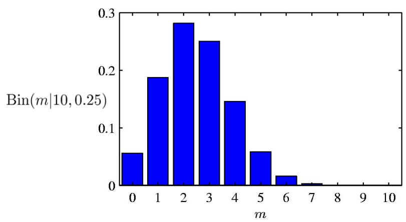
mass density "volume"

density mass

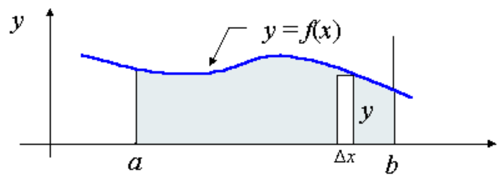
$$\rho = \frac{m}{V},$$

volume

Probability **Mass** Function (**pmf**)



Probability **Density** Function (**pdf**)



Continuous Random Variables:

- The continuous analog to a probability distribution: *probability density function* (pdf)
  - And to compute the probability that X lies in some range, we compute the definite integral of the fn:

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} p(x)\, dx$$

- Joint and conditional continuous densities

$$p(\mathbf{w}) = p(w_0, w_1, \ldots, w_k)$$ Probability vector is just a joint probability!

Joint $$P(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2) = \int_{x=x_1}^{x_2} \int_{y=y_1}^{y_2} p(x, y)\, dx\, dy$$

Conditional $$P(x_1 \leq X \leq x_2, Y = y) = \int_{x=x_1}^{x_2} p(x|Y = y)\, dx$$

Continuous Random Variables:

- Marginalization (summing out)

$$P(y) = \int_{x=x_1}^{x_2} p(x,y)\,dx$$

(where $x_1 \leq X \leq x_2$ describes the sample space of X)

- Expectations

$$\mathbf{E}_{p(x)}\{f(x)\} = \int f(x)p(x)\,dx$$

In many practical situations, not able to compute the integral (don't know the exact form of p(x), or it is impossible to integrate)

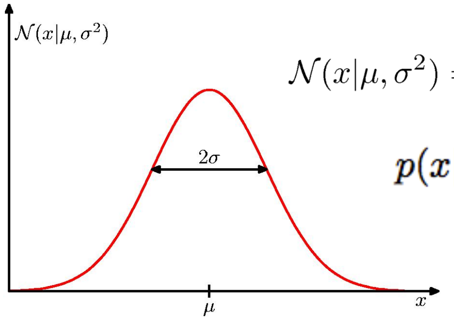
Monte Carlo estimation (if we can draw samples from $p(x)$)

$$\mathbf{E}_{p(x)}\{f(x)\} \approx \frac{1}{S}\sum_{s=1}^{S} f(x)$$

Continuous Density Functions

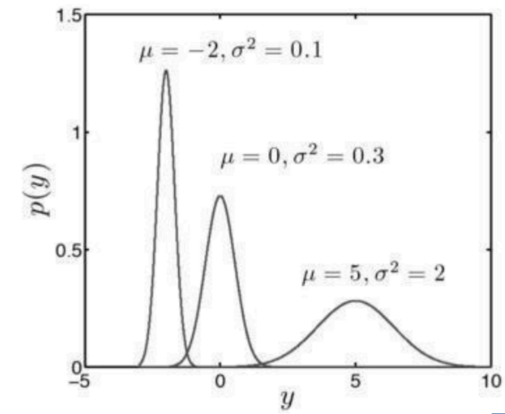
- Uniform
- Beta
- Gaussian (and Multivariate version)

The Gaussian Distribution

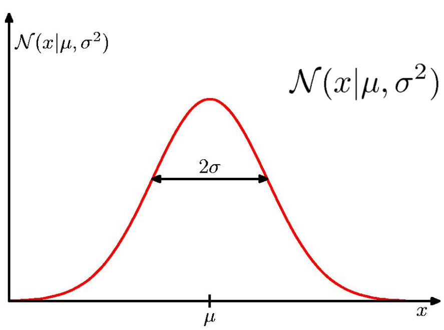


$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

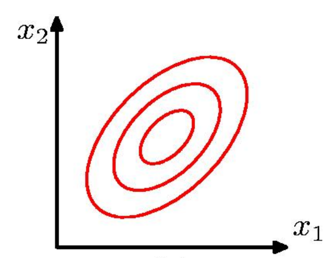
$$p(x|\mu,\sigma^2) = \mathcal{N}(x|\mu,\sigma^2) = \mathcal{N}(\mu,\sigma^2)$$



The Gaussian Distribution



$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

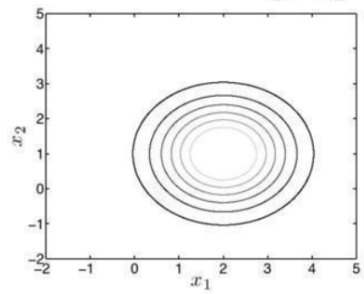


$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

Covariance in Gaussian Distribution

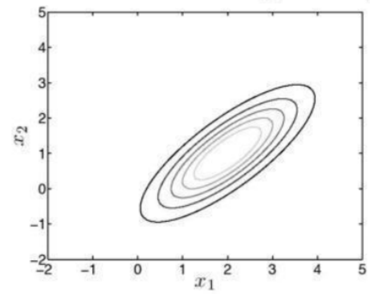
(An intuition for the Covariance Matrix)

$\boldsymbol{\mu} = [2,\ 1]^{\mathsf{T}},\ \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

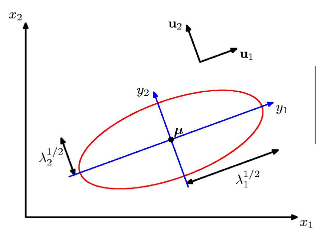


(b)

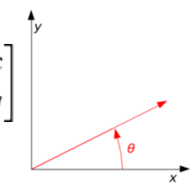
$\boldsymbol{\mu} = [2,\ 1]^{\mathsf{T}},\ \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$



(d)



$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$



Also Note: **precision** (beta) used as inverse of variance: $\beta = \Sigma^{-1}$

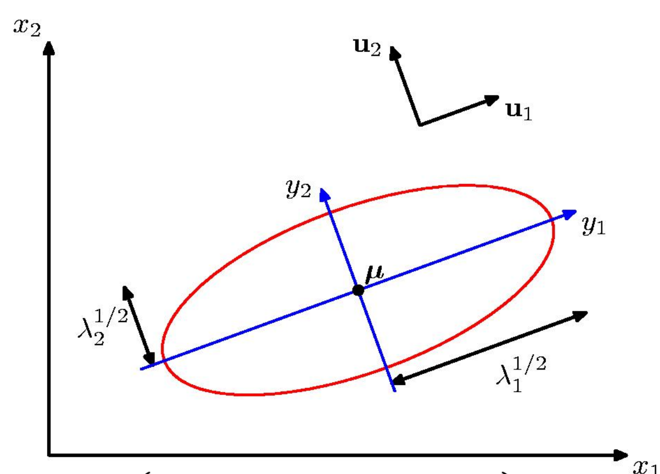
Geometry of the Multivariate Gaussian

$$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

$$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^{\mathrm{T}}$$

$$\Delta^2 = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i}$$

$$y_i = \mathbf{u}_i^{\mathrm{T}} (\mathbf{x} - \boldsymbol{\mu})$$



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

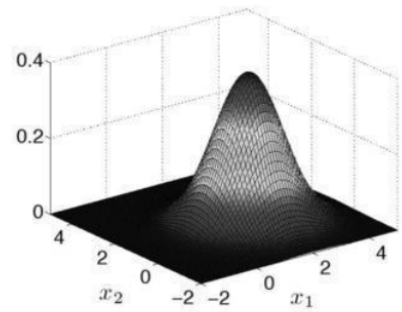
The Multivariate Gaussian

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

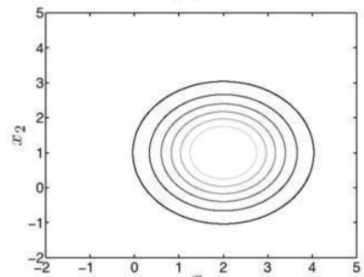
Example: $\boldsymbol{\mu} = [2,\ 1]^{\mathsf{T}},\ \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

$$\begin{aligned}
p(\mathbf{x}) &= \frac{1}{(2\pi)^{D/2}|\mathbf{I}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\mathbf{I}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \\
&= \frac{1}{(2\pi)^{D/2}|\mathbf{I}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\mathbf{I}(\mathbf{x}-\boldsymbol{\mu})\right\} \\
&= \frac{1}{(2\pi)^{D/2}|\mathbf{I}|^{1/2}}\exp\left\{-\frac{1}{2}\sum_{d=1}^{D}(x_d-\mu_d)^2\right\} \\
&= \frac{1}{(2\pi)^{D/2}|\mathbf{I}|^{1/2}}\prod_{d=1}^{D}\exp\left\{-\frac{1}{2}(x_d-\mu_d)^2\right\} \\
&= \prod_{d=1}^{D}\frac{1}{(2\pi)^{1/2}}\exp\left\{-\frac{1}{2}(x_d-\mu_d)^2\right\}
\end{aligned}$$

Each term in the product is a univariate Gaussian!



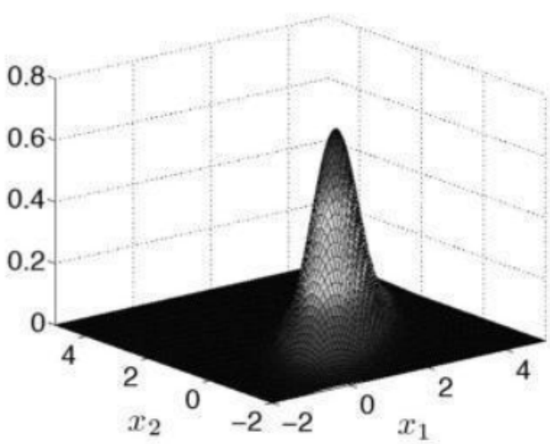
(a)



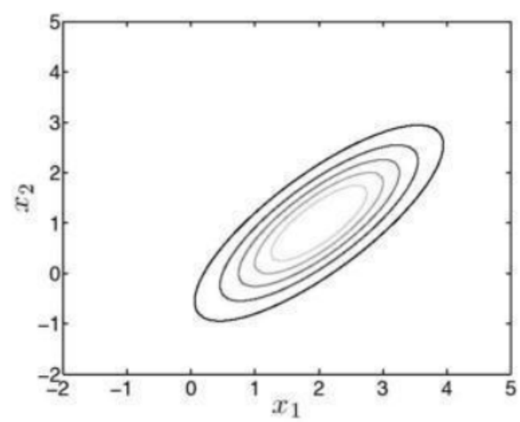
(b)

Another Example:

$$\boldsymbol{\mu} = [2,\ 1]^{\mathsf{T}},\ \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$



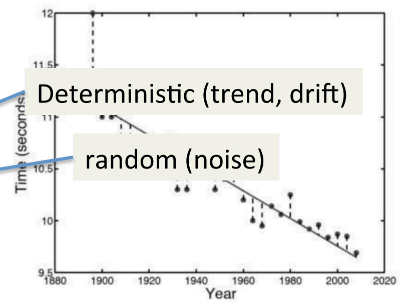
(c)



(d)

Augmenting our Linear Model

$$t_n = \mathbf{w}^\top \mathbf{x}_n$$



- Add "noise" to prediction

$$t_n = \mathbf{w}^\top \mathbf{x}_n + \epsilon_n$$

- *ε* should be continuous
- Noise on each data point is

identical and *independent* (i.i.d)

$$p(\epsilon_1, ..., \epsilon_N) = \prod_{n=1}^{N} p(\epsilon_n)$$

$$\mathcal{N}(0, \sigma^2)$$

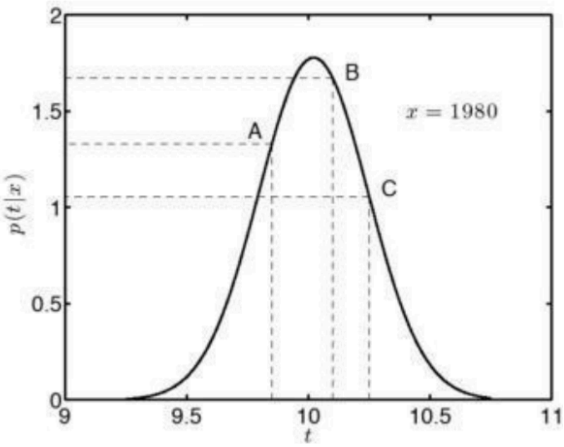
Defining the Likelihood

$$t_n = f(\mathbf{x}_n; \mathbf{w}) + \epsilon_n, \;\; \epsilon_n \sim \mathcal{N}(0, \sigma^2)$$

$$\begin{aligned} y &= a + z \\ p(z) &= \mathcal{N}(m, s) \\ p(y) &= \mathcal{N}(m + a, s) \end{aligned}$$

$$p(t_n | \mathbf{x}_n, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{w}^\top \mathbf{x}_n, \sigma^2)$$

Defining the Likelihood



$\hat{t}_{1980}$ = 10 (pred)

t_{1980} = 10.25 (C)

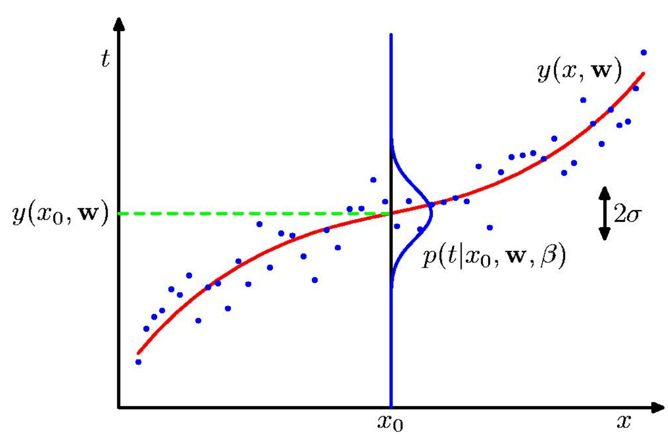
$$p(t_n|\mathbf{x}_n = [1,\ 1980]^\mathsf{T}, \mathbf{w} = [36.416, -0.0133]^\mathsf{T}, \sigma^2 = 0.05)$$

$$p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{x}_n, \sigma^2)$$

$$L = p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2)$$

Given **w**, data are ***independent***

$$L = p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{x}_n, \sigma^2)$$



$$p(t_{1960}|\mathbf{x}_{1960}, \mathbf{X}, \mathbf{t}) = \frac{p(t_{1960}|\mathbf{x}_{1960}) \prod_n p(t_n|\mathbf{x}_n)}{\prod_n p(t_n|\mathbf{x}_n)} = p(t_{1960}|\mathbf{x}_{1960})$$

Maximize the Likelihood

$$L = p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{x}_n, \sigma^2)$$

Since we are working with a product of Gaussians, which in turn include The exponential function (*e*), take the natural log (often just represented Generically as log(*L*))

$$\log L = \sum_{n=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(t_n - f(\mathbf{x}_n;\mathbf{w}))^2\right\}\right)$$

$$\frac{df}{dx} = \frac{df}{dg} \cdot \frac{dg}{dx}.$$

Maximize the Likelihood: w

$$\begin{aligned}\log L &= -\frac{N}{2}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(t_n - f(\mathbf{x}_n;\mathbf{w}))^2 \\ &= -\frac{N}{2}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(t_n - \mathbf{w}^\mathsf{T}\mathbf{x}_n)^2\end{aligned}$$

$$\begin{aligned}\frac{\partial \log L}{\partial \mathbf{w}} &= \frac{1}{\sigma^2}\sum_{n=1}^{N}\mathbf{x}_n(t_n - \mathbf{x}_n^\mathsf{T}\mathbf{w}) \\ &= \frac{1}{\sigma^2}\sum_{n=1}^{N}\mathbf{x}_n t_n - \mathbf{x}_n\mathbf{x}_n^\mathsf{T}\mathbf{w} = \mathbf{0}\end{aligned}$$

$$\mathbf{X} = \begin{bmatrix}\mathbf{x}_1^\mathsf{T} \\ \mathbf{x}_2^\mathsf{T} \\ \vdots \\ \mathbf{x}_N^\mathsf{T}\end{bmatrix} = \begin{bmatrix}1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N\end{bmatrix}, \quad \mathbf{t} = \begin{bmatrix}t_1 \\ t_2 \\ \vdots \\ t_N\end{bmatrix}$$

$$\frac{\partial \log L}{\partial \mathbf{w}} = \frac{1}{\sigma^2}(\mathbf{X}^\mathsf{T}\mathbf{t} - \mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w}) = \mathbf{0}$$

$$\begin{aligned}\frac{1}{\sigma^2}(\mathbf{X}^\mathsf{T}\mathbf{t} - \mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w}) &= 0 \\ \mathbf{X}^\mathsf{T}\mathbf{t} - \mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w} &= 0 \\ \mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w} &= \mathbf{X}^\mathsf{T}\mathbf{t} \\ \mathbf{w} &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}\end{aligned}$$

Maximize the Likelihood: σ

$$\log L = -\frac{N}{2}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(t_n - \mathbf{w}^\top \mathbf{x}_n)^2$$

$$\frac{\partial L}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{n=1}^{N}(t_n - \mathbf{x}^\top \widehat{\mathbf{w}})^2 = 0$$

$$\widehat{\sigma^2} = \frac{1}{N}\sum_{n=1}^{N}(t_n - \mathbf{x}^\top \widehat{\mathbf{w}})^2$$

$$\begin{aligned}\sigma^2 &= \frac{1}{N}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})^\top(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}}) \\ &= \frac{1}{N}(\mathbf{t}^\top\mathbf{t} - 2\mathbf{t}^\top\mathbf{X}\widehat{\mathbf{w}} + \widehat{\mathbf{w}}^\top\mathbf{X}^\top\mathbf{X}\widehat{\mathbf{w}})\end{aligned}$$

Simplify further by plugging in

$\widehat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{t}$

$$\begin{aligned}\widehat{\sigma^2} &= \frac{1}{N}(\mathbf{t}^\top\mathbf{t} - 2\mathbf{t}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{t} + \mathbf{t}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{t}) \\ &= \frac{1}{N}(\mathbf{t}^\top\mathbf{t} - 2\mathbf{t}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{t} + \mathbf{t}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{t}) \\ &= \frac{1}{N}(\mathbf{t}^\top\mathbf{t} - \mathbf{t}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{t})\end{aligned}$$

$$\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t}^\top\mathbf{t} - \mathbf{t}^\top\mathbf{X}\widehat{\mathbf{w}})$$

Maximum Likelihood

$$L = p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N}\mathcal{N}(\mathbf{w}^\top\mathbf{x}_n, \sigma^2)$$

$$\widehat{\mathbf{w}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{t}$$

$$\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t}^\top\mathbf{t} - \mathbf{t}^\top\mathbf{X}\widehat{\mathbf{w}})$$

Predictive distribution