# ISTA 421 + INFO 521
Introduction to Machine Learning

**Lecture 8b:**
**Linear Model**
**with Guassian Noise**
**Maximum Likelihood**

**Clay Morrison**
claytonm@email.arizona.edu
Harvill 437A
Phone 621-6609

18 September 2017

1

## Next Topics

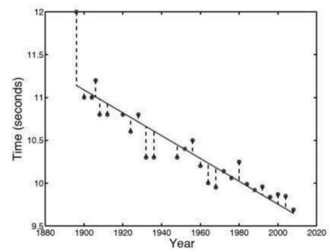
- Return to the Linear Model, with Noise!
- Likelihood Function
- Maximum Likelihood Estimation
- Uncertainty in parameters
- Uncertainty in predictions

2

# Augmenting our Linear Model

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

- Add "noise" to prediction



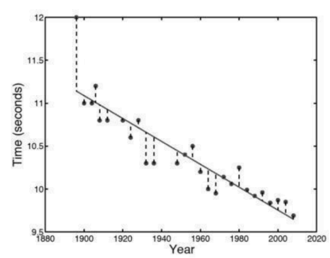

SISTA 3

# Augmenting our Linear Model

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

- Add "noise" to prediction

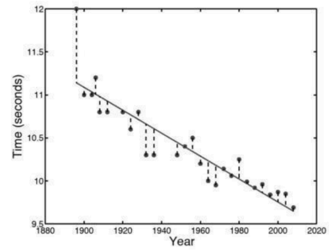
$$t_n = \mathbf{w}^\top \mathbf{x}_n + \epsilon_n$$




SISTA 4

## Augmenting our Linear Model

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




- Add “noise” to prediction

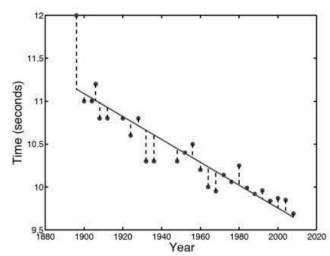
$$t_n = \mathbf{w}^\top \mathbf{x}_n + \epsilon_n$$

- $\varepsilon$ should be continuous

SISTA 5

## Augmenting our Linear Model

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




- Add “noise” to prediction

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

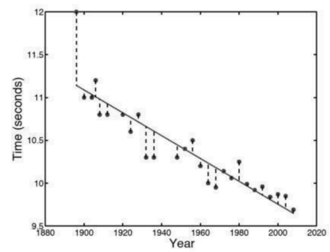
- $\varepsilon$ should be continuous
- Noise on each data point is

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

SISTA 6

# Augmenting our Linear Model

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




- Add “noise” to prediction

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

- $\varepsilon$ should be continuous
- Noise on each data point is

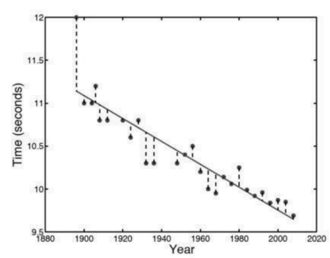
*identical* and *independent* (i.i.d)

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

7

# Augmenting our Linear Model

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




- Add “noise” to prediction

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

- $\varepsilon$ should be continuous
- Noise on each data point is

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

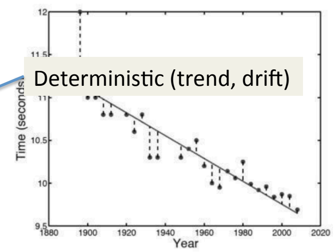
$$p(\epsilon_1, ..., \epsilon_N) = \prod_{n=1}^{N} p(\epsilon_n)$$

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

8

# Augmenting our Linear Model

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




- Add "noise" to prediction

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

- $\varepsilon$ should be continuous
- Noise on each data point is

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

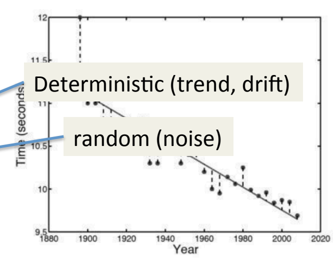
$$p(\epsilon_1, ..., \epsilon_N) = \prod_{n=1}^{N} p(\epsilon_n)$$

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

SISTA 9

# Augmenting our Linear Model

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




- Add "noise" to prediction

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

- $\varepsilon$ should be continuous
- Noise on each data point is

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

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

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

SISTA 10

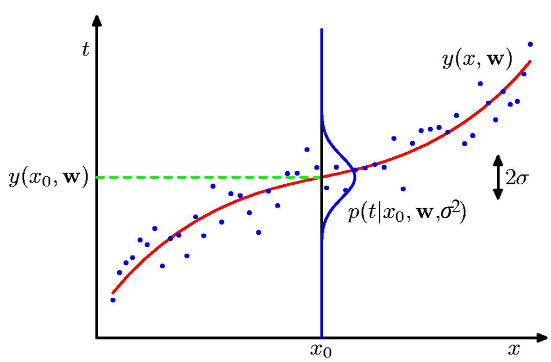
# Defining the Likelihood

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

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

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

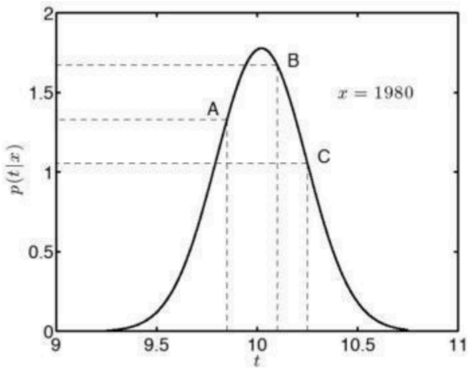
SISTA 11




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

SISTA 12

# Defining the Likelihood




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

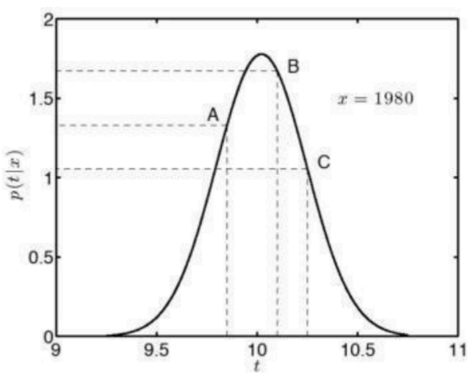
$t_{1980}$ = 10.25 (C)

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

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

13

# Defining the Likelihood




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

$t_{1980}$ = 10.25 (C)

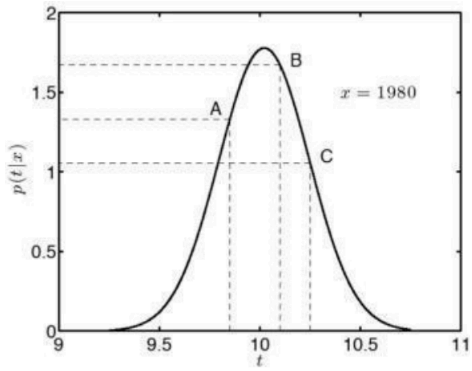
$$p(t_n|\mathbf{x}_n = [1,\ 1980]^\mathsf{T}, \mathbf{w} = [36.416, -0.0133]^\mathsf{T}, \sigma^2 = 0.05)$$

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

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

14

# Defining the Likelihood




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

$t_{1980}$ = 10.25 (C)

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

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

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

SISTA 15

# 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*) )

SISTA 16

# **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*) )

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

17


# **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*) )


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

18

# 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*) )

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

SISTA 19

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

# Maximize the Likelihood: w

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

SISTA 20

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

# Maximize the Likelihood: w

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

21

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

# Maximize the Likelihood: w

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

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

22

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

# Maximize the Likelihood: w

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

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

SISTA 23

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

# Maximize the Likelihood: w

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

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

Recall:

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

SISTA 24

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

# Maximize the Likelihood: w

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

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

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

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

SISTA 25

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

# Maximize the Likelihood: w

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

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

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

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

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

# Maximize the Likelihood: σ

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

SISTA 27

# Maximize the Likelihood: σ

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

$$\frac{\partial \log L}{\partial \sigma} =$$

SISTA 28

# Maximize the Likelihood: σ

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

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

29

# Maximize the Likelihood: σ

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

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

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

30

# Maximize the Likelihood: σ

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

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

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

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

SISTA 31

# Maximize the Likelihood: σ

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

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

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

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

Simplify further by plugging in

$$\widehat{\mathrm{w}} = (\mathrm{X}^\mathsf{T}\mathrm{X})^{-1}\,\mathrm{X}^\mathsf{T}\mathrm{t}$$

SISTA 32

# Maximize the Likelihood: σ

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

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

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

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

$$\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t}^{\mathsf{T}}\mathbf{t} - 2\mathbf{t}^{\mathsf{T}}\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t} + \mathbf{t}^{\mathsf{T}}\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t})$$
$$= \frac{1}{N}(\mathbf{t}^{\mathsf{T}}\mathbf{t} - 2\mathbf{t}^{\mathsf{T}}\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t} + \mathbf{t}^{\mathsf{T}}\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t})$$
$$= \frac{1}{N}(\mathbf{t}^{\mathsf{T}}\mathbf{t} - \mathbf{t}^{\mathsf{T}}\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t})$$

Simplify further by plugging in

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

SISTA 33


# Maximize the Likelihood: σ

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

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

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

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

$$\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t}^{\mathsf{T}}\mathbf{t} - 2\mathbf{t}^{\mathsf{T}}\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t} + \mathbf{t}^{\mathsf{T}}\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t})$$
$$= \frac{1}{N}(\mathbf{t}^{\mathsf{T}}\mathbf{t} - 2\mathbf{t}^{\mathsf{T}}\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t} + \mathbf{t}^{\mathsf{T}}\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t})$$
$$= \frac{1}{N}(\mathbf{t}^{\mathsf{T}}\mathbf{t} - \mathbf{t}^{\mathsf{T}}\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t})$$


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

Simplify further by plugging in

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

SISTA 34