Foundations of Machine Learning
AI2000 and AI5000

FoML-04
Maximum Likelihood Principle

Dr. Konda Reddy Mopuri
Department of AI, IIT Hyderabad

July-Nov 2025

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Data-driven Intelligence & Learning Lab

So far in FoML

- What is ML and the learning paradigms
- Probability refresher
    - Random variables, Bayes Theorem, Independence, Expectation, Variance

Maximum Likelihood Principle

Maximum Likelihood Principle

- Widely used technique for optimizing model parameters

Maximum Likelihood Principle

- Given - Dataset of N independent observations D $= \{x_1, x_2 \ldots x_N\}$

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Maximum Likelihood Principle

- Goal: recover the probability distribution that may have generated this dataset

Maximum Likelihood Principle

- Goal: recover the probability distribution that may have generated this dataset
- Likelihood of the dataset $p(D|w)$

Maximum Likelihood Principle

- The most likely 'explanation' of D is given by w_{ML} that maximizes the likelihood function

$$\mathbf{w}_{ML} = \arg\max_{w} P(D|w)$$

Maximum Likelihood Principle

- The iid assumption - each $x_i \epsilon D$ is independently distributed according to the same distribution conditioned on w

Maximum Likelihood Principle

- The iid assumption - each $x_i \epsilon D$ is independently distributed according to the same distribution conditioned on w

The joint distribution $P(D|w) = P(x_1, x_2 \ldots x_N|w) = \prod_{i=1}^{N} P(x_i|w)$

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

DiL
Data-driven Intelligence & Learning Lab

Maximum Likelihood Principle

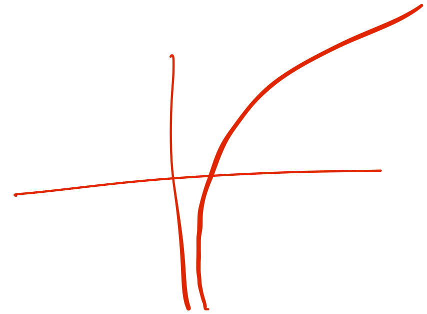
$$\mathbf{w}_{ML} = \arg\max_{\mathbf{w}} \ p(D|\mathbf{w}) = \arg\max_{\mathbf{w}} \ \prod_{i=1}^{N} p(x_i|\mathbf{w})$$

Maximum Likelihood Principle

$$\mathbf{w}_{ML} = \arg\max_{\mathbf{w}} \; p(D|\mathbf{w}) = \arg\max_{\mathbf{w}} \; \prod_{i=1}^{N} p(x_i|\mathbf{w})$$

- Numerical underflow

Maximum Likelihood Principle



$$\mathbf{w}_{ML} = \arg\max_{\mathbf{w}} \; p(D|\mathbf{w}) = \arg\max_{\mathbf{w}} \; \prod_{i=1}^{N} p(x_i|\mathbf{w})$$

- Numerical underflow
- Maximize the log-likelihood →

$$\log \prod_{i=1}^{N} p(x_i|w)$$

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

DiL
Data-driven Intelligence & Learning Lab

Maximum Likelihood Principle

$$\mathbf{w}_{ML} = \arg\max_{\mathbf{w}} \; p(D|\mathbf{w}) = \arg\max_{\mathbf{w}} \; \prod_{i=1}^{N} p(x_i|\mathbf{w})$$

x_1
x_2
$\vdots$
x_N

- Numerical underflow
- Maximize the log-likelihood

$$\mathbf{w}_{ML} = \arg\max_{\mathbf{w}} \; \log \prod_{i=1}^{N} p(x_i|\mathbf{w})$$

Error function:

$$E(D;\mathbf{w}) = -\log p(D|\mathbf{w}) = -\sum_{i=1}^{N} \log p(x_i|\mathbf{w})$$

NLL

MLE for Gaussian Distributions

- iid Gaussian distributed real variables D =

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

$$p(D|\mathbf{w}) = p(D|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

MLE for Gaussian Distributions

- iid Gaussian distributed real variables D =

$$p(x|\mathbf{w}) = \underline{\mathcal{N}}(x|\underline{\mu, \sigma^2}) \qquad p(D|\mathbf{w}) = p(D|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{N} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

log likelihood = $\underline{-\frac{N}{2}\log 2\pi\sigma^2} - \frac{1}{2\sigma^2}\underline{\sum_{i=1}^{N}(x_i-\mu)^2}$

MLE for Gaussian Distributions

- Estimate the model parameters $\mu_{ML}, \sigma^2_{ML} = \underset{\mu,\sigma^2}{\arg\max} \ \log p(D|\mu, \sigma^2)$

$$\frac{\partial}{\partial \mu}\left\{ \quad \right\} = 0 \qquad = -\frac{1}{\cancel{2}\sigma^2}(\cancel{2}) \sum_{i=1}^{N} (x_i - \mu)$$

$$\sum_{i=1}^{N} (x_i - \mu) = 0$$

$$\mu_{ML} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

MLE for Gaussian Distributions

- Estimate the model parameters $\mu_{ML}, \sigma^2_{ML} = \arg\max_{\mu,\sigma^2} \log p(D|\mu, \sigma^2)$

$$\frac{\partial}{\partial \sigma^2}\left(\quad \right) = 0$$

$$-\frac{N}{2} \frac{1}{\cancel{2\pi}\sigma^2}(\cancel{2\pi}) + \frac{1}{2\sigma^4} \sum_{i=1}^{N} (x_i - \mu)^2 = 0$$

$$-\frac{N}{2}\sigma^2 + \frac{1}{2} \sum_{i=1}^{N} (x_i - \mu)^2 = 0$$

$$\sigma^2_{ML} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

DiL
Data-driven Intelligence & Learning Lab

MLE for Gaussian Distributions

$$\mu_{ML}, \sigma^2_{ML} = \arg\max_{\mu, \sigma^2} \; \log p(D|\mu, \sigma^2)$$

- How well do these estimates represent the true parameters?

MLE for Gaussian Distributions

$$\mu_{ML}, \sigma^2_{ML} = \arg\max_{\mu,\sigma^2} \ \log p(D|\mu, \sigma^2)$$

- How well do these estimates represent the true parameters?
- Note that these are functions of the data sample

$D_1 = \{x_1 \cdots x_N\}^1 \longrightarrow {}^1\mu_{ML} \quad {}^1\sigma^2_{ML}$

$D_2 = \{x_1 \cdots x_N\}^2 \longrightarrow {}^2\mu_{ML} \quad {}^2\sigma^2_{ML}$

$\vdots$

$D_M = \{x_1 \cdots x_N\}^M \longrightarrow {}^M\mu_{ML} \quad {}^M\sigma^2_{ML}$

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

DiL
Data-driven Intelligence & Learning Lab

MLE for Gaussian Distributions

$$\mu_{ML}, \sigma^2_{ML} = \arg\max_{\mu, \sigma^2} \ \log p(D|\mu, \sigma^2)$$

- How well do these estimates represent the true parameters?
- Note that these are functions of the data sample
  - → expected values of these estimates

$\mathbb{E}_D[\mu_{ML}]$ — ?

$\mathbb{E}_D[\sigma^2_{M}]$ — ?

$D \sim p(D|\mu, \sigma^2)$

MLE for Gaussian Distributions

- ML estimate of the mean

$$\mathbb{E}_{D\sim p(D|\mu,\sigma^2)}\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{D\sim p(D|\mu,\sigma^2)}\, x_i$$

$$= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{x_i\sim p(x|\mu,\sigma^2)}(x_i) = \frac{1}{N}\sum_{i=1}^{N}\mu = \mu$$

MLE for Gaussian Distributions

- Bias of the "ML estimate of the mean"

$x_i \sim p(x \mid \mu, \sigma^2)$

$\mathbb{E}_D[\mu_{ML}] - \mu$

$\mu \qquad - \mu$

$= 0$

MLE for Gaussian Distributions

- ML estimate of the variance

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(x_i-\frac{1}{N}\sum_{j=1}^{N}x_j\right)^2\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left(x_i-\frac{1}{N}\sum_{j=1}^{N}x_j\right)^2\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[x_i^2-\frac{2}{N}x_i\sum_{j=1}^{N}x_j+\frac{1}{N^2}\sum_{l=1}^{N}\sum_{k=1}^{N}x_l x_k\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left[\mu^2+\sigma^2-\frac{2}{N}\left(\sigma^2+\mu^2+(N-1)\mu^2\right)+\frac{1}{N^2}\left(N(\sigma^2+\mu^2)+N(N-1)\mu^2\right)\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left[\mu^2+\sigma^2-\frac{2}{N}\left(\sigma^2+N\mu^2\right)+\frac{1}{N^2}\left(N\sigma^2+N^2\mu^2\right)\right]$$

$$= \mu^2\left[1-2+1\right]+\sigma^2\left[1-\frac{2}{N}+\frac{1}{N}\right] = \underline{\frac{N-1}{N}\sigma^2}$$

$$\mathbb{E}[x_i x_j] = \mathrm{Cov}(x_i,x_j)+\mathbb{E}[x_i]\,\mathbb{E}[x_j]$$

$i=j \longrightarrow \sigma^2+\mu^2$ $\quad$ $i\neq j \quad \mu^2$

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

DiL
Data-driven Intelligence & Learning Lab

MLE for Gaussian Distributions

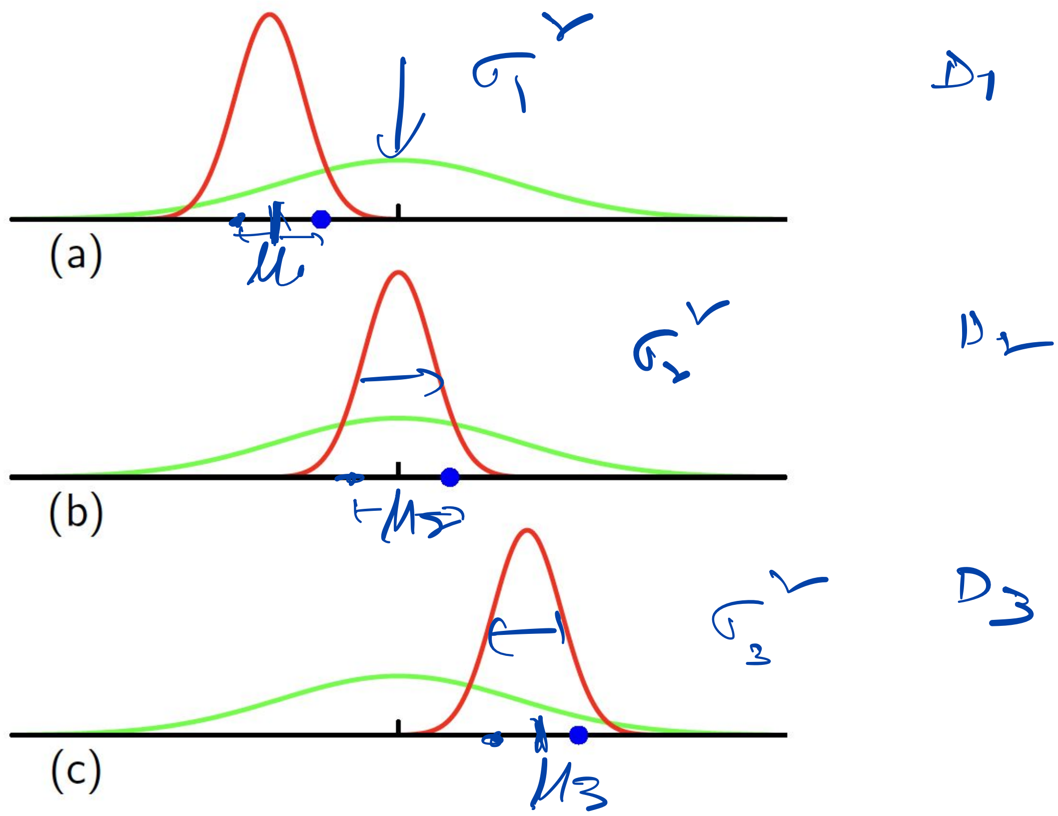
- Bias of the "ML estimate of the variance"

$$\underset{w}{\operatorname{argmax}}\ p(D/w)$$

σ^2_{ML} is biased $\quad \frac{N-1}{N}\sigma^2$

$$\underline{\sigma^2_{ML\text{-}UB}} = \underline{\frac{N}{N-1}}\ \underline{\sigma^2_{ML}}$$

Bias in variance estimate



Regression example

MLE for Regression (curve fitting)

$\hat{t} = \sum_{i=1}^{m-1} w_i x^i + w_0$

- Given data D

$$D = \{(x_1, t_1), (x_2, t_2), \dots (x_N, t_N)\} = \{\mathbf{x}, \mathbf{t}\}$$

$$MSE = \frac{1}{N} \sum_{i=1}^{N} \left(t_i - \hat{t}_i\right)^2$$

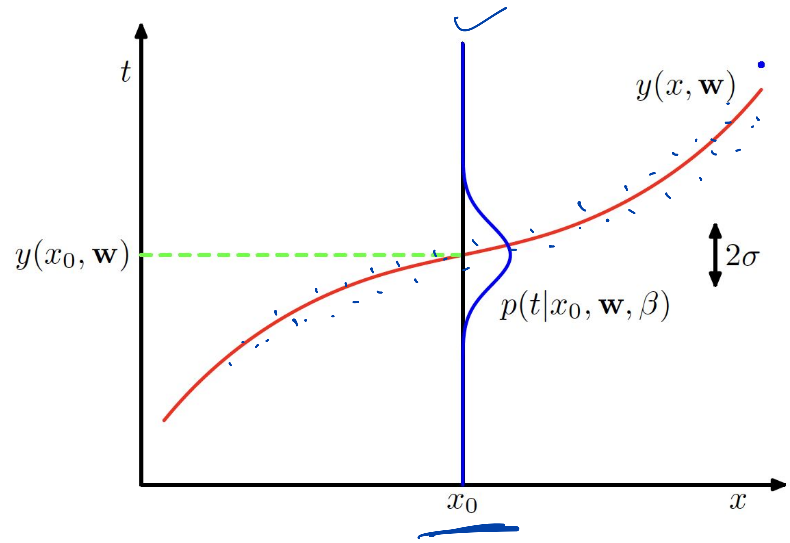
MLE for Regression (curve fitting)

$D = \{(x_i, t_i)\}$

- Given data D $\quad D = \{(x_1, t_1), (x_2, t_2), \dots (x_N, t_N)\} = \{\mathbf{x}, \mathbf{t}\}$
- Assume the data is generated by

$$t = y(x, \mathbf{w}) + \sigma \cdot \epsilon, \quad \epsilon \in \mathcal{N}(0, 1)$$

measurement error/noise



MLE for Regression (curve fitting)

- Target distribution $$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1})$$

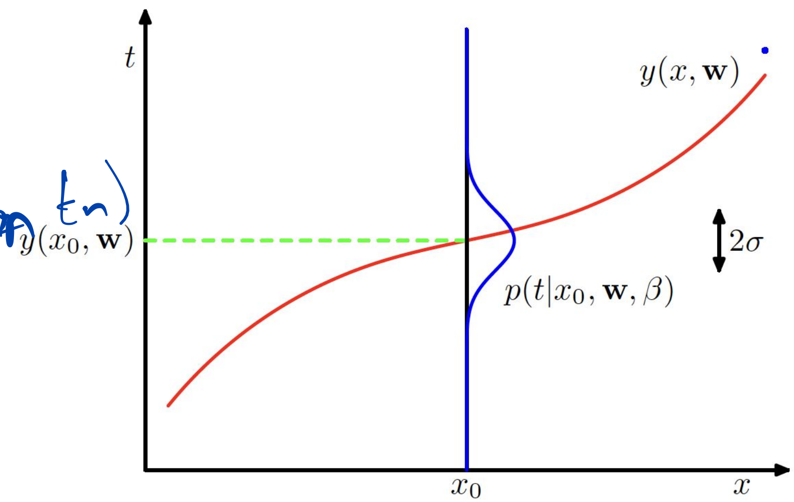
$x_1 \quad t_1$
$\vdots$
$x_n \quad t_n$
$(t_1 - \hat{t}_1)$

$\beta \uparrow \quad \downarrow$

$(x_1, t_1) \ldots (x_n, t_n)$

$p(\underline{D} \mid \underline{w}, \underline{\beta})$

$y = \underline{w^T} x + \epsilon$



MLE for Regression (curve fitting)

$\{(x, t)\}$

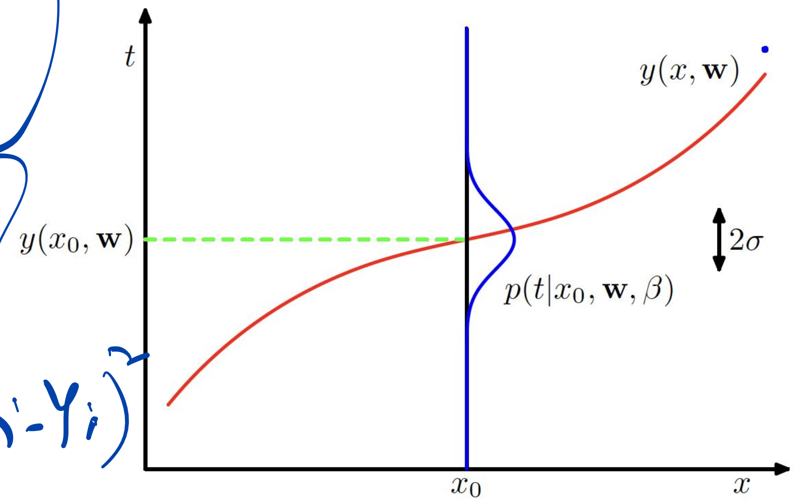
- Target distribution $\quad p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1})$
- log likelihood

$$\log p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta^{-1})$$

$$= \log \left(\prod_{i=1}^{N} \sqrt{\frac{\beta}{2\pi}}\, e^{-\frac{\beta}{2}(t_i - y_i)^2} \right)$$

$$= \sum_{i=1}^{N} \log \left(\sqrt{\frac{\beta}{2\pi}} \cdot e^{-\frac{\beta}{2}(t_i - y_i)^2} \right)$$

$$= N \log \sqrt{\beta/2\pi} - \frac{\beta}{2} \sum_{i=1}^{N} (t_i - y_i)^2$$



MLE for Regression (curve fitting)

$D = \{x_i, t_i\}$

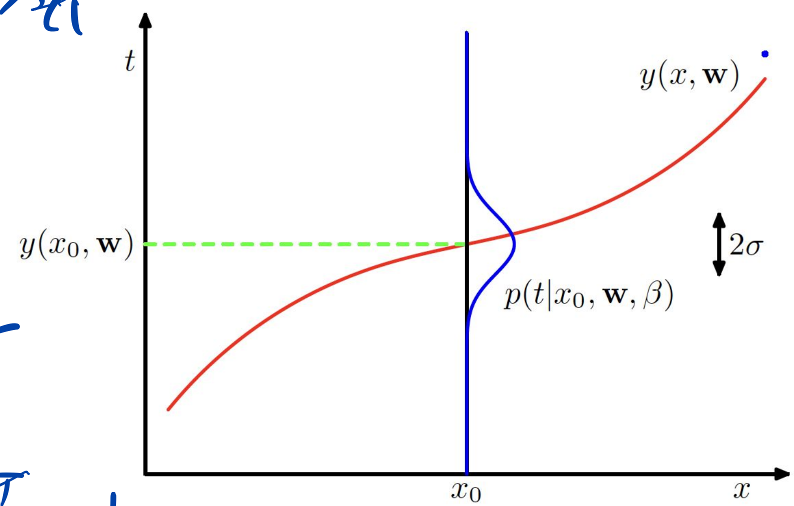
- Minimize the NLL w.r.t the parameters w and β

$$NLL = \frac{\beta}{2}\sum_{i=1}^{N}\left[t_i - y(x_i, w)\right]^2 - \frac{N}{2}\log\frac{\beta}{2\pi}$$

$$\frac{\partial(\cdot)}{\partial w} = 0 \qquad \underset{w}{\operatorname{argmin}} \quad MSE$$

$$\frac{\partial}{\partial\beta}(\;) = \frac{1}{2}\sum_{i=1}^{N}\left(t_i - y(x_i, w)\right)^2 - \frac{N}{2}\,\frac{2\pi}{\beta}\,\frac{1}{2\pi} = 0$$

$$\Rightarrow \frac{N}{\beta} = \sum_{i=1}^{N}\left(t_i - y_i\right)^2$$

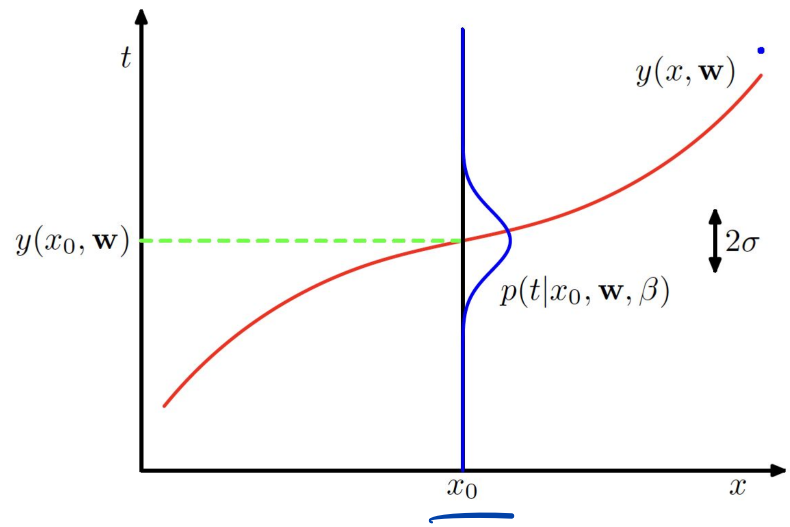


MLE for Regression (curve fitting)

- The predictive distribution

for any test sample x'

$$p(t' \mid y(x', w_{ML}), \beta_{ML}) = \mathcal{N}(t \mid y(x', w_{ML}), \beta_{ML}^{-1})$$



if one wants the point estimate,

$y(x', w_{ML})$

$\Rightarrow$ expected value

Rough work

Next
MAP